Automated Network Optimisation Using Data Mining as Support for Economic Decision Systems

This Thesis is submitted for the degree of

Doctor of Philosophy

ELENI ROZAKI

April 2019

Cardiff University
School of Computer Science and Informatics

APPENDIX 1 - STATEMENTS AND DECLARATIONS TO BE SIGNED BY THE CANDIDATE AND INCLUDED IN THE THESIS

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of ... (*insert PhD, MD, MPhil, etc., as appropriate*)

Signed

Date 30/04/2019

STATEMENT 2

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is it being submitted concurrently for any other degree or award (outside of any formal collaboration agreement between the University and a partner organisation)

Signed

Date 30/04/2019

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available in the University's Open Access repository (or, where approved, to be available in the University's library and for inter-library loan), and for the title and summary to be made available to outside organisations, subject to the expiry of a University-approved bar on access if applicable.

Signed

Date 30/04./2019

DECLARATION

This thesis is the result of my own independent work, except where otherwise stated, and the views expressed are my own. Other sources are acknowledged by explicit references. The thesis has not been edited by a third party beyond what is permitted by Cardiff University's Use of Third Party Editors by Research Degree Students Procedure.

Signed

Date 30/04/2019

WORD COUNT 35,471

(Excluding summary, acknowledgements, declarations, contents pages, appendices, tables, diagrams and figures, references, bibliography, footnotes and endnotes)

ABSTRACT

The evolution from wired voice communications to wireless and cloud computing services has led to the rapid growth of wireless communication companies attempting to meet consumer needs. While these companies have generally been able to achieve quality of service (QoS) high enough to meet most consumer demands, the recent growth in data hungry services in addition to wireless voice communication, has placed significant stress on the infrastructure and begun to translate into increased QoS issues. As a result, wireless providers are finding difficulty to meet demand and dealing with an overwhelming volume of mobile data. Many telecommunication service providers have turned to data analytics techniques to discover hidden insights for fraud detection, customer churn detection and credit risk analysis. However, most are ill-equipped to prioritise expansion decisions and optimise network faults and costs to ensure customer satisfaction and optimal profitability. The contribution of this thesis in the decision-making process is significant as it initially proposes a network optimisation scheme using data mining algorithms to develop a monitoring framework capable of troubleshooting network faults while optimising costs based on financial evaluations. All the data mining experiments contribute to the development of a super –framework that has been tested using real-data to demonstrate that data mining techniques play a crucial role in the prediction of network optimisation actions. Finally, the insights extracted from the super-framework demonstrate that machine learning mechanisms can draw out promising solutions for network optimisation decisions, customer segmentation, customers churn prediction and also in revenue management. The outputs of the thesis seek to help wireless providers to determine the QoS factors that should be addressed for an efficient network optimisation plan and also presents the academic contribution of this research.

Acknowledgements

Firstly, I would like to express my appreciation to my supervisor Professor Stuart Allen for his continuous guidance, motivation and constant feedback.

Secondly, I would like to thank Cormac Dullaghan for contributing to this thesis with his knowledge and expertise from the telecommunication industry. Cormac also provided us with the opportunity to access essential information for our research. This is greatly appreciated.

Finally, I would like to also say thank you to Pauline Kildunne who provided insight and expertise in R programming language that greatly assisted the final part of this thesis.

Table of Contents

List of Figures

List of Tables

Acronyms

ANNs	Artificial Neural Networks
Arff	Attribute-Relation File Format
CDR	Call Drop Rate
CSSR	Call Setup Success Rate
CN	Core Network
GSM	Global System for Mobile Communications
HOF	Handover Failures
HSR	Handover Success Rate
HSUPA	High Speed Uplink Packet Access
ID3	Iterative Dichotomiser 3
IMT-2000	International Mobile Telecommunications-2000
ITU	International Telecommunication Union
KPI	Key Performance Indicator
LTE	Long Term Evolution
MANETs	Mobile ad hoc networks
MAP	Mobile Application Part
MOA	Massive Online Analysis
MLP	Multilayer perceptron
OECD	Organisation for Economic Cooperation and Development

PART	Pruning rule based classification tree
QoS	Quality of Service
RAB_FR	Radio Access bearer failure radio
RAN	Radio Access Network
RBF	Radial Basis Function
RF	Radio Frequency
RIPPER	Produce Error Reduction
ROC	Receiver Operating Curve
RRM	Radio Resource Management
SDCHAR	Standalone Dedicated Control Channel Access Rate
SDCCHCR	Standalone Dedicated Control Channel Congestion Rate
SDCCHDropsExcessiveTA	Standalone Dedicated Control Channel Drops due to Excessive Timing Advance
SDCCHCRSuddLostCon	Standalone Dedicated Control Channel Suddenly lost connection
SIF	Signal Information Field
SIM	Subscriber Identity Module
SIO	Service Information Octets
TCH	Traffic Channel
TCHCR	Traffic channel congestion rate

TM	Traffic Model
TD-CDMA	Time Division CDMA
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
TR	Traffic Rate
TCH	Traffic Channel
W-CDMA	Wideband CDMA
Weka	Waikato Environment for Knowledge Analysis
GLM	Generalized Linear Model
DRF	Distributed Random Forest
GBM	Gradient Boost Machine

PART 1

Part 1 of this thesis includes Chapter 1 that gives an overview of the research and Chapter 2 that discuss similar research on the challenge of fault diagnosis in cellular networks associated with cost optinisation and revenue analysis. Chapter 3 provides an overview of common Key Performance Indicators (KPIs) and the direct dependencies between the causes and symptoms of network faults using alarm correlation, which can be considered the first step in the diagnosis of faults and alarms. Chapter 4 describes our proposed approach to automated troubleshooting and network fault diagnosis using data mining classifiers and clustering algorithms.

Chapter 1. Introduction

Big data management technologies and analytics will revolutionize the systems and processes that companies use to obtain value from their data. Moreover, challenges in optimisation processes will continue to arise as more relevant data may be integrated with the use of 5G, IoT as a system that connects smart devices, people or animals, objects and digital machines, cloud computing technologies and other emerging network applications that offer better data collection, storage and delivery mechanisms and drive the network optimisation methods to be more flexible and dynamic. (Gupta et al 2019) The explosion in demand for smartphones, tablets and other Internet-enabled portable devices, and subsequent growth of mobile Internet services is placing increasing pressure on mobile operators and networks. Recent statistics from the OECD (Organisation for Economic Co-operation and Development) broadband portal show that high-speed mobile Internet subscriptions grew by 7.7 million, or 7.6%, in the 12 months leading to June 2017, taking mobile broadband penetration to over 100% in the OECD area for the first time. As of end December 2018, there were 1.484 billion mobile broadband subscriptions out of a total OECD population of 1.352 billion. “Mobile broadband penetration is highest in Japan, Finland, Estonia, the United States and Denmark, with subscriptions per 100 inhabitants at 172%, 157%, 149%, 144% and 136% respectively.

Overall, fixed broadband subscriptions in OECD countries totalled 418 million as of December 2018, up from 406 million a year earlier and averaged 30.9 subscriptions per 100 inhabitants. Switzerland still leads the pack with a penetration rate of 46.8 subscriptions per 100 people, followed by Denmark (43.3%), France (43.3%), The Netherlands (43%) and Norway (41.5%)" (OECD 2018).

One of the most noticeable effects of this growth in network demand is an increase in technical quality of service (QoS) issues. One well-known example is network congestion, which occurs when user demand outstrips the assigned network capacity (Rodrigues et al., 2009). Capacity in networks is assigned by *radio resource management* (RRM) algorithms and requires sufficient infrastructure to deal with average demand. RRM is designed to be flexible in its performance and applicable to various user cases through the use of Radio Frequency (RF) profiles. However, the operation of RRM algorithms is applied to network usage, monitoring and network configuration data while data analytics algorithms may be able to connect and handle additional data such as network costs and customers' preferences and willingness to pay. Thus, data analytics techniques can improve the mobile network performance and also maximize the revenue of the telecommunication operators. (He et al. 2016).

There is a limit to the number of radios that can operate for a given infrastructure and with many new portable devices entering the market, exceeding an RF designs capacity is becoming much more common (Cisco 2016). Furthermore, if average demand is significantly exceeded over a short period for some reason (such as daily changes in utilisation patterns or an increase in mobile data usage due to temporary events), QoS can be compromised, leading to service outage or degradation.

Network congestion is highly noticeable to users, particularly users of quality-sensitive applications such as voice or video calls.

Congestion control algorithms, which continuously monitor networks for signs of congestion are commonly used in mobile telecommunications systems, but they do not work perfectly as they can only manage the resources that they have to work with (Rodrigues et al., 2009), (Cisco 2016).

Thus, a major impact of increasing mobile cellular use (including mobile data) is network congestion and its consequences.

The design process for mobile network upgrades and implementation is ordinarily a complex and long operation for engineers with hardware infrastructure and mobile operations. While more efficient technologies may become available (e.g., LTE, 5G, IoT), congestion control mechanisms cannot make up for the deficiencies in network planning and resource optimisation in reducing congestion on mobile networks. There are many issues that delay network planning and resource optimisation deployment such as the need to develop and support a maximum network bandwidth, define the use of new bands, design and produce infrastructure equipment, install new equipment at cell sites, deal with traffic and interference that may arise with neighboring bands. Consequently, making the network design process time-efficient is non-trivial, and there is an opportunity for planning/troubleshooting tools and algorithms to significantly increase network performance, operation and maintenance costs (Katsha & Ramli 2016).

In this thesis, experimental research has been conducted on the role that data analytics can play in network troubleshooting considering costs and revenue optimisation decisions. Integrating the latest advances that employ data mining techniques applied in networks' data including costs' considerations and customers' demographics might be the best way to build robust optimisation frameworks for managing network performance while conducting intelligent customer profiling and segmentation.

1.1 Purpose of the study

The volume of low-level optimisation data produced from cellular network infrastructure, and smartphones/tablets themselves, is growing at an exponential level. The data analytics techniques proposed in this thesis are demonstrated using real data from iD Ireland's datasets, and contribute to the use of advanced data mining methods to uncover hidden patterns, unknown correlations, financial trends, telecommunication customer profiles and other useful business and technical information from our data sets.

The contribution of iD Mobile Ireland in providing data is very important to support the experimental work in revenue management performed in this thesis. They are a start-up telecommunications provider in the Republic of Ireland, which differentiates itself in the competitive Irish market by separating the mobile tariff from the handset. This allows customers the flexibility to enter or leave a 12, 18 or 24-month contract without penalty, and to purchase a new handset every three months (should they wish to do so) once the previous handset cost is fully paid off.

The company has access to a wide range of data, with prospects to capture even more growing at a rapid rate. The data that the company can access are currently not being used to their full potential as a means of understanding the customers that are served, their sale patterns, potential fraud risks, churn patterns and revenue management issues (C.Dullaghan & E.Rozaki 2017).Furthermore, the EU General Data Protection Regulation (GDPR), set to come into force in May 2018, may provide obligatory guidance to reach a fair deal between the interests of mobile providers and mobile users (Wachter, 2018).

The customers' data used for this thesis are anonymised as personal data are encrypted and mobile customers' identifiable information was removed from the iD mobile Irelands' data sets. Hence, GDPR implementation is not required for anonymous data (Štarchoň , Pikulík 2019).

Network optimisation is the process of configuring infrastructure and allocating resources to improve performance, evaluated through technical measures (for example, traffic rates) and financial metrics (for example, the cost of required hardware or software). In this thesis, the problem of *fault detection* in cellular networks was taken in consideration which aims to utilize KPIs to identify and repair/upgrade anomalous cells while minimizing downtime. As such, the optimisation problem was consider to define rules that, given a stream of network performance KPIs, are capable of identifying faults and malfunctions, leading to improved traffic handling capability.

In contrast with many existing studies, this thesis considers not just the technical concerns of implementing recommendations, but also the costs and revenues that real-world network operators and service providers must face as they compete in a saturated market with rapidly growing bandwidth demands. The findings show that using data mining and business analytics techniques can lead to more effective network optimisation decisions, new revenue opportunities, dynamic customer profiling, improved operational efficiency and churn prediction and prevention.

1.2 Scope and structure of the thesis

The research is arranged into three broad parts, structured as follows. The scope of Part 1 that includes chapters 1, 2, 3 and 4 is limited to technical resource management and implementation of technical optimisation mechanisms, including the relevant aspects of network planning and resource optimisation using data mining techniques. It explores the use of data mining models and algorithms concerning the causes and symptoms of network faults as a means of automating defect management. It concludes with a summary of recommendations based on the network optimisation and resource management outcomes. The goal is also to find and cluster relations between the several sets of key performance indicators (symptoms) and additional variables (attributes) to discover the causes that affect the optimisation procedure.

In *Part 2*, that contains chapter 5, the scope of the research widens to also consider financial evaluation. This section includes a comprehensive financial analysis of the network optimisation approach as defined above, as well as addressing the issue of costs associated with remedying the network faults presented in Part 1. This section is related to the cost and revenue considerations using meta cost classifiers to raise the accuracy of the results. Part 2 also discuss meta -cost classifiers running with Bayesian and decision trees algorithms, together with the previous work of network fault detection using key performance indicators. Bayesian networks are used to show the cost probability of the distribution of alarms and the cost sensitive meta-learning classification rules using C.5 algorithms and finally neural networks to determine the minimum network cost conditions, bandwidth costs and benefits that are possible with network optimisation.

Finally, in Part 3 that includes chapter 6, 7 and 8 the scope of the research widens still further to integrate an end to end analytics framework for decision making. The aggregation of available information about clients, their demographics, service use and network cost enables the prioritisation of location-based network optimisaiton actions, thus enabling to maintain the required end to end quality of services. Furthermore, the availability of location-based information facilitates data-driven management decisions as well as reactions to client-data in the appropriate time using the right deal, which will be offered to the proper target market. The focus in this part is also on revenue management, and in particular identifying consumers' preferences.

The concern in this section is not merely the revenue analysis itself, but also comparing the performance of revenue optimisation to cost optimisation using business analytics modelling and making recommendations to providers about network predictive modelling.

The last part of the thesis will combine a library of different algorithms that provides the ability to enhance performance, as well as to create the best combination of algorithms to improve the performance of a network framework and define an efficient end to end optimisation super framework. A stacked learner model utilised which was run using V-fold cross-validation in order to create the optimal weighted combination of predictions from a selected library of algorithms to avoid over-fitting. Finally, the last step was to compare ensemble methods to combine multiple base learners, including general linear models, gradient boost model, random forest, and deep learning.

1.3 Significance and contribution

The aim of this thesis is to demonstrate a working system that will be able to assist engineers in the decision-making process for optimising an existing mobile communications network based on network faults, cost considerations and customers' needs and willingness to pay. Although the technical work will assume a UMTS (Universal Mobile Telecommunications System) network, it would be straightforward to apply the findings to LTE networks due to significant overlaps and technical complementarities. The question of optimisation is designed to take on a number of different dimensions, including resource availability and allocation, cost management, and revenue management.

This holistic approach will allow companies using the optimisation model generated in this research to select the level of involvement required, as well as clearly demonstrating the importance of interplay between the technical and cost objectives of optimisation.

The main contributions that will allow for the accomplishment of this aim are:

I. Establishing an automated defect management framework using data mining methods, which will enable telecommunication network operators to engage in more accurate and rapid network planning and resource optimisation. This will also include automated network optimisation processes, derived from the KPI alarm classification and localisation processes. The elements generated during completion of this objective will include technical models and specifications based on decision trees, rules, and Bayesian network and Clustering algorithms intended to implement the suggested optimisation solutions.

II. Establishing a cost sensitive neural network classification model intended to define cost optimisation and budgeting, address annual budgeting and planning efforts, provide support for management planning processes, and allow for the production of pricing and valuation information for customers' profiles. This model will balance costs of service provision and new investments with expected revenues using neural network models.

III. Establishing a cost optimisation model using data mining modelling. These revenue management algorithms define an appropriate technique for valuation of fault detection and budgeting to improve the overall quality of services, especially for customers who are willing to pay more for a higher quality of services.

Generating recommendations for customer segmentation, technical and cost management optimisation for use in existing networks, in order to assist in the optimisation of existing resources for network carriers.

IV. Propose an end to end optimisation super- framework capable of reducing network performance costs while improving the customer experience. The results showed the most efficient data mining methods for each purpose and the optimal combinations of the most important variables of our datasets. Several data mining algorithms tested individually, or as base learners combined with the super leader and cost learner models that gave a superior performance to predict revenue decisions, define customer's priorities, provide support for revenue management planning processes, and allow for the production of pricing and valuation information for customers. In addition, an important contribution of this part of the thesis is that the multi- class classification models that has been proposed that ads to the existing academic literature, which is limited regarding the relation to the multinomial classifiers based on cost learners and super learners.

Published papers related to the research in this thesis can be seen below:

Published contributions:

Journals:

C. Dullaghan, E. Rozaki, “Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers” International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 7 No 1, 2017.

doi: 10.5121/ijdkp.2017.7102

E. Rozaki, "Financial Predictions Using Cost Sensitive Neural Networks for Multi-Class Learning", Advanced Engineering Forum, Vol. 16, pp. 104-116, 2016.

doi:10.4028/www.scientific.net/AEF.16.104

E. Rozaki, "Design and Implementation for Automated Network Troubleshooting Using Data Mining", International Journal of Data Mining & Knowledge Management Process (IJDKP). Volume 5 pp.9-27, 2015. doi:10.5121/ijdkp.2015.5302

E. Rozaki, "Clustering optimisation techniques in mobile networks", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC). Volume 4, 2, 22 – 29, 2016.

Conferences:

E. Rozaki, “Network Fault Diagnosis Using Data Mining Classifiers”, Third International Conference on Database and Data Mining, pp 29 40, 2015. doi:10.5121/csit.2015.50703

E. Rozaki, “Automated Network Optimization Using Data Mining Techniques”, ICSPDM International Conference on Signal Processing and Data Mining (ICSPDM) 2017 in Rome.

E. Rozaki, “Data mining modeling and cost sensitive analysis for network faults”, ICSPDM International Conference on Signal Processing and Data Mining (ICSPDM) 2017 in Rome.

E. Rozaki, “Business analytics applications in budget modelling to improve network performance”, 4th International Conference on Big Data Analysis and Data Mining September 07-08, 2017 Paris, France doi: 10.4172/2324-9307-C1-014

Chapter 2. Background

This chapter explores the existing literature on fault diagnosis and network troubleshooting in cellular networks, an area which includes network fault detection, diagnosis of network malfunctions and solving the problems that arise from network troubleshooting. The aim of this chapter is to examine the optimisation costs and end to end quality of services in mobile networks.

The mobile telecommunication industry is changing rapidly driven mainly by revolutionizing network technologies, evolving state of the art solutions and services and increasingly composite consumer demands. Thus, there is a need for generating new business models, requiring the network providers to be continuously updated. Economic pressures mean that operators need to decrease service costs while increasing the value to the consumer in order to remain competitive. At the same time, new technologies, social networking and mobile services are becoming increasingly more important and available to a wide base of customers. Consumers today require access to multimedia applications as their main entertainment is fostered by mobile communications and social media networks (Kryftis Y. et al. 2016). Consequently, the telecommunications industry is heading rapidly towards a model of heterogeneous wireless access to networks, with an increasing number of users who want high value services, but reduced costs, and ever-increasing numbers of applications that have high bandwidth requirements.

Maintaining mobile networks can be a particular challenge, particularly the process of detecting, interpreting and rectifying defects. Malfunctions, hardware and software failures can be exposed through a range of different symptoms, such as dropped calls, access failures, congestion or an excessive number of interference handovers, and can be located in any part of the network. The task of locating the root cause of defects can be very difficult since they can be located in any part of the distributed network . Poorly performing cells might not be immediately obvious from performance metrics. Tracking using support ticket systems can be difficult and time consuming in a situation where time is of the essence. Defects must be corrected before causing significant impact to users, because consumers who are dissatisfied with services might switch to a competitor (Lunn, Lyons 2018). Most defect management is related to finding and correcting difficulties in the radio access network which it resides between mobile devices, and provides connection with its core network (CN) therefore represents the largest part of the cellular system.

Barco, Díez, Wille, and Lázaro (2009) suggested that there are three primary steps to controlling defect management. The first is detection of defective cells that are defective, followed by determining or diagnosing why the malfunction occurred. The third step is to deploy a solution by planning and carrying out an action to correct the defect. Currently, these operations rely on the experience of skilled professionals. Increasing the quality of service by automating the process would decrease operational costs for the carriers and improve the value to end users.

The diagnosis step is the most complex part of defect management. There are currently very few automated approaches to diagnosing of Radio Accessibility Network (RAN) faults in cellular networks (Popoola et al., 2018a), (Popoola et al., 2018b) even though there is a great deal of interest in automating this step. However, similar automated processes have been used for diagnosis in other fields such as spine surgery (Han S.S., et al. 2019) or detecting defects in the core networks in communications networks (Musa 2016).

However, automated methods for detecting and diagnosing network faults are very early in the development process, time consuming, and contribute to increased customer dissatisfaction. (Widanapathirana et al. 2012). The fact that automated fault detection methods do not exist is an important gap within both the practical and academic literature. The goal of this investigation is to provide an efficient and effective means by which the process of detecting and diagnosing network faults can be automated. Thus, our initial aim in this thesis is to investigate the causes that affect the optimisation procedures and then discover how to automate the process for detecting and diagnosing network faults using intelligent data mining methods.

A monitoring scheme for mobile networks should be based on optimisation issues to investigate the relationship between causes and symptoms. The most critical alarms which are indicators of network issues for mobile network providers may be cells experiencing a high rate of dropped calls that can create negative impact on the QoS offered to the end user. False alarms of high dropped call rate can be grouped such as interference, handover and coverage (Barco et al. 2006).

This approach represents an improvement since it identifies cause and effect in the relationship and assists in determining specific issues which may exist and be located through the process of troubleshooting (Barco, et al. 2006). The selection of traffic data, handovers, dropped calls and utilisation data to be analysed involves several considerations, such as which data are representative of the KPIs (symptoms), data availability (inputs), the attributes of the data, and establishing limits of the network anomaly detection, the generation of alarms and the correction of the network issues (Nor Haizan. et al., 2009), (Shahzad, Asad, & Khan, 2013).

2.1 Automated Defect Management

The core aspect of this research in terms of network planning and resource optimisation is defect management automation. Defects in a mobile network can be understood to be any condition that creates a transmission failure (Rahnema, 2008). However, in some cases, such as conditions of excessive congestion or physical equipment failure, there is a need to identify and correlate defects and reroute transmissions around problematic areas. (Rahnema, 2008).

The process of identifying and reacting to defects or alarms was studied in the early development of mobile networks. Wietgrefe et al. (1997) considered the specific case of a single failure in a GSM network, which they indicated could generate as many as 100 alarms in various parts of the network. They noted that it was not just noticing the alarms that is important, but also correlating these alarms to identify the underlying condition.

There are several issues that can be identified in the use of alarm correlation that will need to be explored in order for it to be used effectively in the current research. One issue is similarity, which can be used to determine the likelihood that multiple alarms correspond to a single event such as link failure. For instance, an alarm generated by a high drop calls rate which may be caused by overloaded traffic or inefficient traffic channels infrastructure, but may be also related with poor hardware equipment such as antennas or software upgrades. Yamashita et al. (2010), tackling the problem of online recommendations, discussed a similarity computational approach based on Pearson's correlation coefficient and Particle Swarm Optimisation.

Okamoto et al. (2006) suggested a modification to the standard correlation process that approximated the relationship between null hypotheses rather than exactly calculating it, resulting in a substantial improvement in the computational performance of the correlation at the expense of a small amount of accuracy. A second approach also used the Neyman-Pearson framework to add stochastic resonance to differentiation equations (Bayram & Gezici, 2012). The remainder of this chapter is structured as follows: Section 2 outlines the main issues in fault diagnosis, and the significance of mobile network performance and resource optimisation in network troubleshooting. Then section 3 describes the importance of the End-to-End Network Optimisation techniques for providing quality in mobile services following by section 4 that discusses the financial analysis and a cost opimisation approach, and finally section 5 that presents how users revenue can influence network management decisions.

2.2 Fault diagnosis in cellular networks

Cusani, Inzerelli, and Valentini (2007) pointed out that the main problems arising in mobile networks is malfunction of cell elements leading to degradation of service. To identify these, network operators monitor Key Performance Indicators to trigger alarms, including the dropped call rate (DCR), Radio Access bearer failure (RAB_FR), that are channels that layer the data connections according to the type of data and facilitate the transfer data to the users or control network data, Call Set up Success Rate (CSSR), utilization of Handover Success (HS) that refers to the process of transferring an ongoing call or data from one channel to another in the mobile netwrok and finally the traffic model (TM), an indicator that measures the mobile netwrok traffic (Cusani et al., 2007; Seytnazarov, 2010).

Key performance indicators can be used to diagnose the cause of the problem in a cell. When a malfunction occurs, the KPIs in the cell change value, and an automated diagnosis system can analyse these variations to identify which cell or cells malfunctioned, and the cause of the malfunction. Due to the wide range of KPIs that are used by network operators, this thesis cannot provide an exhaustive classification.

The next important issue is the generation of alarms, as well as the relationship between the KPIs and alarms. Before discussing alarms, it is appropriate to briefly explain the difference between an alarm and a fault. An alarm is an indication of a problem or issue in a network. An alarm can be thought of as the outward symptom of a problem.

For instance, an alarm may be triggered by a KPI that shows a very high dropped call rate, such as DCR>=4. The Maximum Permissible Dropped Call Rates based on the Commission for Communications Regulation Ireland (Wireless Telegraphy - Liberalised Use and Preparatory Licences in the 800 MHz, 900 MHz and 1800 MHz bands) Regulations, 2012) .That KPI represents a symptom that could be caused by a high number of handovers due to interference (Barco et al. 2006). Although, the fault is the actual issue that needs to be corrected in order to restore a network to an optimised state of operation (Monacelli, & Francescangeli, 2011). Network operators typically generate alarms when the values of KPIs cross a pre-determined limit or threshold.

2.3 Performance of mobile networks

One of the specific challenges for mobile networks is the need to integrate capacity in existing networks, which provide the bulk of coverage outside high-demand urban areas (Rahnema, 2008). This is an issue that need to be dealt with because of the complexity of networks as well as consumer demands for service quality across a range of heterogeneous types of service, radio interface standards, and other complexities. Maximum capacity planning without compromising, quality of service is the basis for efficient network services, although routing and other issues also need to be considered. Clever spectrum reuse of the base stations location will be cost effective for users as well as for mobile operators. (Jha et al. 2017).

Although network capacity is clearly the core issue in network planning; it is not the only problem, given the structure and heterogeneous utilisation of mobile networks. Another issue that emerges in mobile network planning is cell performance (Beyer & Mao, 2012). Mobile networks can face QoS degradation from network congestion or attenuation or even service interruption, particularly if the network planning does not provide dense enough coverage for demand (Beyer & Mao, 2012). These types of problems are explored in more detail in Part I of the thesis, which addresses the various concerns of network optimisation of existing networks and identifies the network malfunctions that can arise.

There are a variety of approaches that can be used in network troubleshooting that includes fault detection, cause diagnosis and solution deployment. A statistical analysis approach can be used for network diagnosis and network management, which predicts many of the issues such as hardware or software failures caused by network congestion.

Ye and Fallah (2010) explored the application of KPIs to UMTS packet-switched networks in order to highlight the benefits of their statistical approach to capacity planning. They specifically considered UMTS, where packet switching is used to transfer data, including voice, video, music, and other data. The authors identified a service monitoring approach based on the IP network core, which they conceptualize as sitting between the packets being passed on the network between devices and nodes.

They then used statistical techniques including correlation, factor analysis, multidimensional scaling, correspondence analysis, and cluster analysis as a means of identifying issues such as hardware or software failures, technician issues in data flow and determining capacity gaps. One of these techniques, correlation, will be returned to in discussion of alarm correlation. Ye and Fallah (2010) identified a number of KPIs that could be used to track the performance of the network capacity. This is a useful approach that could be further leveraged to account for other resource needs, such as resource optimisation and defect management automation as discussed below.

2.4 Resource optimisation

In the widest sense, equipment in a network environment includes backbone or core network, base stations (which provide transmission and reception to the core network), and handsets or other user devices. By extension, the resources associated with this equipment generate transmission capacity, which can be understood in a number of different ways, such as usage time or transmission quality. In a situation where resources are unlimited, network planning is trivial, since it is possible to simply assume more resources can be made available to provide coverage.

However, there are few situations, if any, where resources can truly be said to be unlimited with no additional cost. Because of this, *resource optimisation*, or identification of the most efficient way to use individual resources and groups of resources, is required as part of the network planning process.

There are many different resources that could be optimised within the network in order to improve network efficiency. For example, Radio resources are targeted for optimisation in order to maximize QoS (Seytnazarov, 2010). A well-structured network requires less infrastructure and mainly fewer Base Station sites, hence more efficient use of radio resource, offering additional capacities under the same infrastructure. (Abrao 2012). Defect identification and management is also part of the process of resource optimisation. For instance, KPIs can be used to identify uplink performance issues that refers to the signal from mobile station (cellphone) to base station and RF failure. In their discussion of defect identification in mobile networks, Sánchez-González et al. (2008), used threshold levels for KPIs in order to identify potential RF failure conditions in the uplink. This approach can be used to identify issues, but not necessarily to automatically specify how to repair networks. This is one of the areas where the proposed research will build on existing solutions, not just by maximizing defect management, but also by optimising it in order to balance defect management and resource utilisation.

2.5 Financial analysis

Troubleshooting existing networks is not the only way to achieve improved utilisation. Analysis of costs suggests that there could be ways to improve the financial benefits from existing resources, such as determining what upgrades are cost-effective and how existing resources could be used more effectively (Stevens, 2012). The use of cost analysis and cost optimisation is one of the approaches that will be used to understand how the technical optimisation approaches identified in Part 1 can be further supported. Alternative approaches that can be used to evaluate cost-benefit analysis are simulations (Asche et al. 2018) or Cost-Sensitive Analysis with the use of Time Series predictions (Walgampaya & Kantardzic 2016). Cost-sensitive techniques allow for a focus on achieving high accuracy of sample classification into a set of predefined classes while reducing misclassification costs that can arise when predicting the cost of rectifying network errors (Charles et al. 2008). A problem, however, is that many machine learning approaches are binary in nature. Moreover, cost-sensitive aspects and measures defined by classification algorithms have not been widely studied yet as frequently indicate low efficiency and issues with scalability when dealing with large-data sets. (Wang et al. 2012). In order to improve how they function and to allow for efficient cost sensitivity analysis for network fault detection, the algorithms need to be generalized to deal with costly network issues such as hardware replacement that have many solutions (Charnay, Lachiche, & Braud, 2013).

The initial goal for this investigation was to present a monitoring scheme for mobile networks using data mining classifiers to upgrade fault detection and handling. The purpose is to extract optimisation rules that improve anomaly detection and support a monitoring scheme that relies on data mining techniques for the purpose of fault isolation and cost estimation. The reason for this is because network technicians and administrators need to know on regular basis where faults are occurring on a regular basis so that they can determine where network malfunctions are occurring, as well as the cost of optimization actions required in a specific location in a given period of time.

2.6 Cost optimisation

Cost optimisation is part of a holistic approach to network resource allocation and capacity sizing applied at the network planning stages (Santoyo-Gomzalez & Cervello-Pastor 2018). Here, the expression of cost optimization is more general than financial analysis as it also refers to the telecommunication costs reduction such as traffic cost, infrastructure costs, transmission costs and maintenance cost. Different cost optimisation measures are required due to the increasing and conflicting demands of video and data applications, network saturation, and other concerns that mean that networks need to be managed effectively in order to ensure that capital expenditures are focused where they will be most effective (Liu et al. 2014). Chu et al. (2011) note that failures in network planning can result in operational network failures generated from natural disasters, human interference, or technical issues included in infrastructure costs.

However, not all network issues stem from total failure or lack of network survivability; instead, there is the issue of capacity management, or optimisation of system capacity in response to changes of demand and revenue of users that needs to be included in operations and maintenance costs.

As the authors point out, the NP-complete nature of network optimisation makes a complete solution to this problem highly difficult, particularly for complex or integrated networks. Furthermore, “full optimisation” can be risky because, when a network is optimised very tightly to achieve one goal, typically profit, it might be expensive to adapt the network to changing needs. Therefore, instead of trying to find optimal solutions, planning models should allow the user to understand a network’s behaviour and trade-offs better. This suggests that it is not necessarily ideal to simply focus on technical optimisation of the network, since this could make it difficult to address changing strategic priorities for a reliable cost benefit analysis (Liu et al., 2014). It is also likely that the use of strict technical optimisation will generate conditions that make the network financially inviable for the network operator, due to excessive demand for capital investment and operational monitoring (Chu et al., 2011).

Cost optimisation models used in previous network types such as 2G are not appropriate for use in 3G and beyond, since there are substantial differences between these networks, especially data reliability concerns. Chu et al. (2011) make a strong case for including cost optimisation, as well as technical optimisation, in network planning and resource allocation.

Findings of Chu et al.'s (2011) analysis of network planning and cost optimisation in mobile networks show that accounting for survivability concerns, link and node constraints at the network planning stage significantly reduces costs.

They also indicate that network planning is not a one-time activity but needs to be a constant focus as network demand increases to account for changing impacts on cost. This strongly suggests that there is value for network planning in considering cost optimisation as well as technical optimisation, to keep service provisions both ready to meet the demands of consumers and within the bounds of financial resources of the providing firm. (Santoyo-Gonzalez, Cervello -Pastor, 2018).

Methods for achieving network performance have traditionally been used by network administrators and technicians as a means of finding network faults returning networks to conditions of optimisation. Although, network planning was not considered as the datasets used for this thesis do not contain networking hardware data known also as network equipment data, the plan was to follow the same approach in continuous optimisation of existing networks while incorporating cost into optimisation methods that have not been widely used by managers as a means of approaching financial predictions. Thus, the multiclass cost-sensitive classification problem that aim to achieve cost efficient network improvement does require predicting cost optimisation rules with regards to fault detection. Financial performance in the process of broader network fault detection and optimisation is possible (Kim, J., etal. 2012).

Chapter 3. Network Performance Measure

This chapter summarises research techniques for modeling network faults that can be classified based on their cause or their type. However, the most convenient network anomaly classification method is based on location. The reason for this is because network technicians and administrators need to know where faults are occurring so that they can determine where malfunctions are occurring, as well as the number of faults that occur in a specific location in a given period of time (Deljac, Mostak & Stjepanovic, 2010). In this chapter of the thesis the main scope is using statistical methods for selecting, analysing and drawing meaningful interpretations in relation to the network data and key variables of the optimisation plan. The results of this part of the research demonstrated the need to extend to the machine learning techniques in order to make most accurate predictions on network performance management.

3.1 Key Performance Indicators

Network performance measurement is essential for providing efficient quality of mobile services. The KPIs used in this research for QoS of Global System for Mobile Communications (GSM) networks include Call Setup Success Rate (CSSR), Drop Call Rate (DCR), Stand-alone DedicatedChannel (SDCCH) congestion, and Traffic Channel (TCH) congestion. (Popoola et al. 2018) The first set address issues related to retainability and the probability that calls drop.

The call setup success rate (CSSR) which is the fraction of the successful call connections divided by the total number of the attempts to make a call and call drop rate (CDR) are two of the primary KPIs used by operators (Seytnazarov, 2010). The second KPI in this category, CDR, is the rate at which calls are dropped when moving between radio channels that facilitate the call transmission associated with a core network (Barco, Wille, & Diez, 2005).

An additional KPI metric which is important for identifying the network performance is the handover failures (HOF) or assignments that present the ratio of attempts/handovers that have failed divided by the total handovers, (Panda and Padhy, 2009).

The HOF KPI is associated with the mobility of the network that includes channels control and it is also related with the KPI of SDCCHSR (Stand-Alone Dedicated Control Channel Success Rate).The network mobility include also traffic channel drops measurements such as Call Traffic Rate (TR) and transition (TCH congestion Rate) assignments that are both successful and unsuccessful (Seytnazarov, 2010).The second set of KPIs are measures of utilization and accessibility, relating to the extent to which the network optimises cost quality (Cusani et al. 2007). Over-dimensioning increases costs because it reduces utilisation below capacity.

When this happens, the costs to the provider (and ultimately the consumer) increase. In the same fashion, under-dimensioning results in congestion across the network, service delays, increased dropped calls, and other issues. The number of channel groups, or group of frequencies within a specified connection in the core network must be sufficiently large to prevent these issues from occurring (Cusani et al., 2007). Both of these issues affect customer satisfaction and customer retention over a longer term than CSSR and CDR. The Traffic model is used as the base measurement for utilization. This measurement monitors the traffic channels of the network, the lower call transition frequency, the better the utilization (Cusani et al., 2007).

End-user services are categorized into four groups by the 3rd Generation Mobile System (3GPP) according to their traffic characteristics and QoS demands. Four corresponding traffic classes defined for Radio Access Bearers (RABs) that transfer calls and data between the user equipment and the Universal Terrestrial Radio Access Network (UTRAN) that connects mobile handsets to the public telephone network or the Internet.

RABs values given by Flood 1997 are optimised so that the end-user service group corresponding to the RAB class is supported)with the four RAB classes being conversational, streaming, interactive class, and background.

The conversational RAB class relates to the end-to-end user service and is typical of human-to-human interaction. A cell phone is the typical example of this class, in which the appropriate transfer delay relates to the human perception of video and audio conversations. The streaming RAB class relates to a continuous audio/video stream transferred to the end-user. In this case the appropriate transfer delay relates to the capability of the time alignment function in the client where buffers can compensate for minute delays in the delivery of packets. The interactive class is the most common type of end-to-end service and is utilized in the internet. Transfer delay is given by the request response pattern in these client-service type applications. The final class is the background class, in which the end-to-end user service runs in the background. Examples of this type of service include downloading in the background, and messaging. There is no requirement on transfer delay in this type of background communication (Flood, 1997).

3.1.1 Measurement techniques for fault diagnosis using KPI

A diagnosis model and a method of interference constitute an automated system for diagnosis.Selected sets of KPIs are defined to categorize the causes in groups. The relationship that is assumed between the cause of the malfunction and the observed system represents the diagnosis model based on the results of the KPIs. Therefore, the relations and interdependencies of the KPIs are selected, followed by an investigation using statistical techniques for testing the false hypothesis of our samples in the selected datasets.

To implement the diagnosis model, a two pronged hypothesis test of the causes in a mobile network is devised, consisting of the correlation of a certain number of symptoms that can be investigated. Figure 3.1 shows the direct dependencies between the variables, the causes, and the symptoms, and shows how they can be related to the data using the KPIs selected.

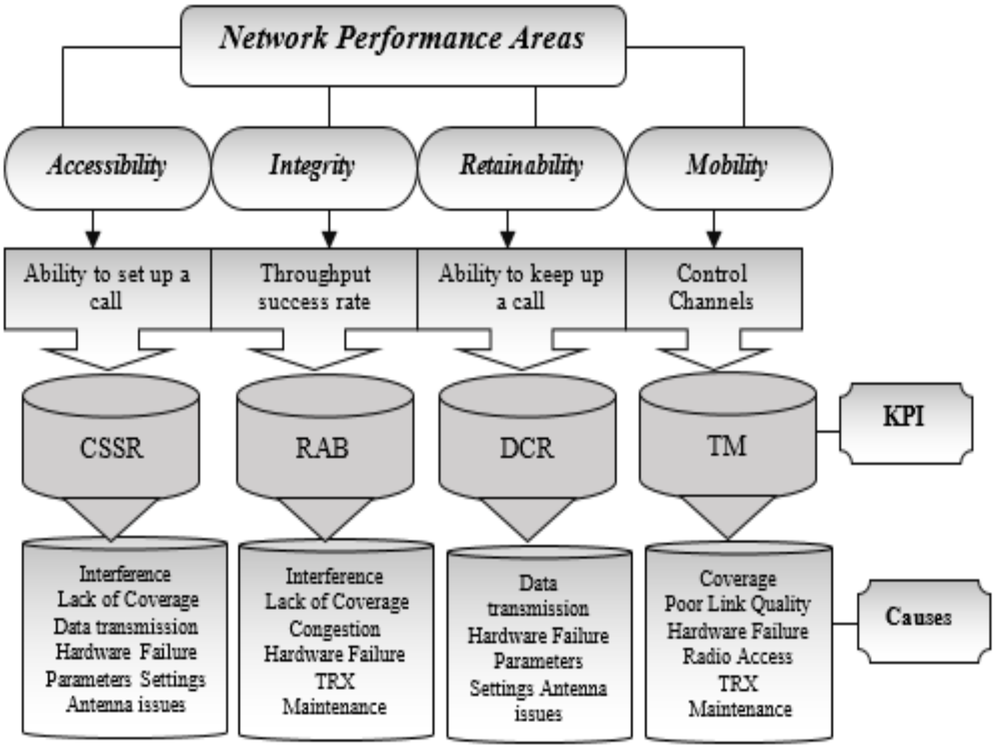

Network Performance Areas
Accessibility
Integrity
Retainability
Mobility
Ability to set up a call
Throughput success rate
Ability to keep up a call
Control Channels
CSSR
RAB
DCR
TM
KPI
Interference Lack of Coverage Data transmission Hardware Failure Parameters Settings Antenna issues
Interference Lack of Coverage Congestion Hardware Failure TRX Maintenance
Data transmission Hardware Failure Parameters Settings Antenna issues
Coverage Poor Link Quality Hardware Failure Radio Access TRX Maintenance
Causes

Figure 3.1 - Direct dependencies of KPIs and Causes (Source: Rozaki 2016, p. 24)

3.1.1.1. Accessibility

Call Success Rate

$$CSSR = \frac{Number\ of\ call\ Setup}{Number\ of\ call\ attempts} * 100$$

is an important KPI parameter for telecommunication companies to evaluate their performance and optimise QoS (Hammed and Fatimah 2018). CSSR can be defined as the number of successful calls divided by the total number of call attempts. The value that is acceptable for the CSSR indicator that indicate the percentage of calls successfully established (set up) should be greater than 95%.

3.1.1.2. Integrity

Integrity generally refers to a phenomenon in which a service once obtained provides low impairments. Due to the increased need of accessibility and mobility in telecommunication networks, integrity is also an important KPI.

The RAB_FR (radio access bearer failure ratio), which is

$$RAB\_FR = \left(\frac{N_{\underset{TOT}{RAB}} - N_{\underset{FLR}{RAB}}}{N_{\underset{TOT}{RAB}}} \right)$$

related to radio channel accessibility, which is a function of the number of control attempts required, congestion, and traffic levels (Cusani et al., 2007), and defined as:

where $N_{\underset{TOT}{RAB}}$ is the number RAB setup attempts that failed; and $N_{\underset{TOT}{RAB}}$ is the total number of RAB setup attempts (Cusani et al., 2007).

A Radio Access Bearer determines the QoS allocated to an application. It also defines a set of allowed frame sizes and wasted bandwidth due to padding bits that can be added, if needed, in order to fill the remaining bits of the framework. The main purpose of the RAB indicator is to provide a segment of the network that contains a limited number of base stations.

For a given application it is necessary to adjust the RAB to its requirements (Perez-Costa, et al., 2004). Too small bandwidth might result in bad quality while too large allocation is a waste of resources.

3.1.1.3. Mobility

Location management keeps track of the mobile terminal while travelling from place to place and is an important factor in the overall resource management of the network (Sawson et al, 2007). Location updating and paging are the two important operations of location management and it is necessary to provide a trade-off between them in order to reduce total cost of hardware equipment. Moreover, an important factor to consider in mobility is the standalone dedicated control channel success rate (SDCCHSR) that shows the rate of successful signaling channel assignments divided by the number of call attempts.

In order to understand the effect of how the channels impact the network's performance, TCH congestion rate.

$$\text{TCH congestion} = \frac{Number\ of\ unavailable\ (blocked\ )TCH\ requests}{Total\ number\ TCH\ requests}*100$$

and

SDCCH congestion rate

$$= \frac{Number\ of\ failed\ connections\ due\ to\ assignment\ failure}{Total\ number\ call\ attempts} * 100$$

were discussed in regard to the congestion in the network as it increases to a certain level (Hammed and Fatimah 2018).

Traffic model (TM) refers to the total traffic, which is the summary of the outgoing and the incoming traffic. The outgoing traffic is equal to the number of outgoing calls that a cell receives during a certain number of minutes multiplied by the number of minutes divided by 60 (min). Correspondingly, the incoming traffic is the number of the incoming calls that a cell receiving during a certain number of minutes multiplied by the minutes divided by 60 (min).

3.1.1.4. Retainability:

Retainability is the capability of a service to continue providing service under the given conditions for the requested period of time. It is also defined as the ratio of dropped voice calls over successfully established voice calls (Skianis et al., 2013). There may be multiple causes of low CDR

$$\text{CDR } = \frac{Number\ of\ dropped\ calls}{Number\ of\ succesfully\ completed\ call\ attempts}$$

within a network such as radio interference or hardware faults. Traffic Channel congestion is one of the issues relating to the KPI which should be resolved in order to improve QoS. It is defined as the ratio of blocked calls due to unavailable resources by the total number of call requests (Seytnazarov et al., 2010). Congestion in traffic channels occurs due to an increase in traffic at a specific site or an increase in the number of subscribers (Hammed and Fatimah 2018).

3.2 Proposed approach: Characteristics of optimal solution

The difficulties of network management are increased when devices and equipment from different manufacturers are used. "*Cisco reports that 50 billion devices and objects will be connected to the Internet by 2020. Also, the Internet of Things (IoT) will contribute $117 billion to the IoT-based healthcare industry and $1.9 trillion to the global economy according to Gartner and Forbes*" As a result the development of IoT technologies will introduce challenges in network performance, mobility and connectivity (Gad & Ahmed 2019). Thus with the use of a variety of devices on a network, the process of anomaly detection problems and issues becomes more difficult because the circuits in the network do not have a single console under which they operate. (Park et al., 2009). In this part of the thesis, an investigation is undertaken of a proposed automated scheme as a solution for network troubleshooting and fault detection based on data mining techniques to provide converged services with functionalities that facilitate its fault-handling and operational management. This section defines the values and limits of false alarm values based on KPI measures suggesting the optimal solution can be proposed.

3.3 Correlation of Key Performance Indicators and faults

In this study, evaluation is presented on the basis of four important KPIs used as an input to create a classification algorithm to analyse and identify network faults. Currently, in most cellular networks, components provide alerts when there is a malfunction. However, these alerts are generally very low-level, leading to a large number of alerts for every potential fault. Our approach is that grouping alerts together in a meaningful way could help to identify the potential causes of network issues more clearly. However, even after associating alerts, there is usually not enough information to uniquely identify the cause of malfunction, especially when malfunctions can be caused by something other than equipment defects or failure. Moreover, even after alert association, the resulting number of alerts resulting from a single fault can still be very high. In addition, similar alerts can also be triggered by different causes. In order to reach a conclusive diagnosis, some alarms need to be discarded because they do not provide additional information, and some causes need to be ruled out because they are not applicable to the specific situation being investigated. Due to the nature of investigation required for these operations, the causes and symptoms of the problems within the network are explored.

3.4 Applied Statistics for fault diagnosis

In this chapter the aim is to collect, clean, categorise, and gain insights from large datasets created from 2014- 2017 that contain approximately 26,717 rows that present the users or the cell sites/Base Stations and Base station controllers (BCS) in approximately 86 columns of attributes. The fields of the columns are customers' data (such as unique ids, age group, gender, day of purchasing the service, type of deals, payments data, base station & location data, service duration) and network performance data (dropped/successful calls, traffic data, traffic channels, Radio Access Barriers, etc). The first dataset tested in this part of the thesis contains network data from different telecommunication companies in Ireland including iD Mobile Ireland. The mobile data also contain an additional 11 columns comprised of formula derived values or alarm classes used to categorise the data in Key Performance Indicators. The primary aim of this effort is to better meet network troubleshooting needs, improve network optimisation, diagnose and identify the correlation of the different causes of network faults and KPI alarms.

The initial step in carrying out this effort was to acquire the relevant data to generate the various KPI reports, using the attributes available, and to find relations between our data and KPIs as a means of confirming the accuracy of the data. This verification step proved to be very important because as several tables of data where combined, additional derived variables occurred.The methods being utilized and applied to collected UMTS packet switched (PS) performance and service parameters (QoS) are discussed and suggested by the work of Ouyang, & Hosein Fallah, (2010).

KPI correlation measures the strength and direction of a relationship between two or more KPIs. Correlation is a standardized measurement that generates a value that is easy to interpret where the correlation coefficients range from -1.00 to +1.00, where +1 shows a perfect positive linear relationship between two KPIs metrics. The correlation is applied to analyse the relationship between QoS performance in a mobile network. The results of the correlation will may be utilized to identify service-level quality and KPIs. The correlation results reveal the impact of different types of service on the performance of the UMTS network (Ouyang & Fallah, 2010). The initial mobile data file used for this part of the thesis contains 11 columns comprised of formula derived values or alarm classes used to categorise the data in Key Performance Indicators. Figure 3.2 shows an example of the relationship between the KPIs that will be testing based on the additional 11 columns that contain the KPI values. The figure shows 12 different scatterplots of the KPI data display. Each scatterplot shows the relationship between two different KPI variables.

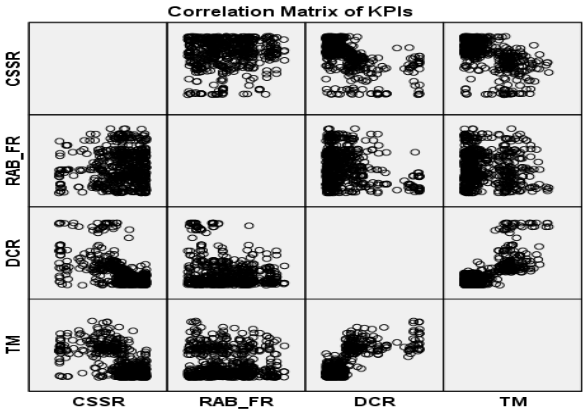


Figure 3.2 KPI alarms correlation matrix

Generally, it can be observed the correlation results to be verified at the population proportions of KPIs. For instance, the KPIs of DCR and TM values are usually both accepted as “optimised values” or rejected together as “not optimised values” in any network location (that verifies the correlation results) in contrast with the CSSR and RAB_FR indicators values that could be accepted or rejected independently from the other indicators.

Where the alerts of two or more locations are to be compared, the comparison is based on repeated observational faults at the same time and location, based on the results of several KPIs. Excluding sequential test situations from consideration, for each of these alerts related to the KPIs targets, a null hypothesis is postulated stating that for the particular time and base station, it would be beneficial to review the cell parameters for optimisation.

3.5 Define Null Hypothesis (H0) and Alternate Hypothesis (H1) for KPI Values

Consider a binary hypothesis testing optimisation issue with two hypotheses. The two hypotheses are testing every cell individually based on the KPIs parameters to check if it is an “Optimised Cell” which will not be considered such as a network faultor a “Non Optimised Cell” which is the cell that needs to be optimized accordingly to the QoS standard parameters (Segun et al. 2018),(Musa 2016). The “Non Optimised Cell” will be termed, H_0 and the individual null hypotheses at the N base station locations will be termed $H_0 i\ i = 1,2, \dots$, N locations. If H_0 is tested, at significance level, against the alternative: H_i"optimise the cell" the KPIs target agrees to take a risk of CDR metrics (for example when a=0.05) to reject H_0 when in fact it is true.

Consequently, hypothesis testing will facilitate the process of finding the KPI alarms while the Pearson correlation coefficient will contribute to finding the relations and associations between the problematic cells according to the KPI false alarms. Thus, $H_{j,j=0,1}$ was set up, where the 0 is the alert that requires the decision for not optimising the base station, and 1 is the alert for optimise. A two-level distributed optimisation system was proposed that consists of a number of base stations (locations) of a cellular network connected with a set of KPIs. Thus, each location (cell) employs a support decision KPI related to a data set of Signal and Service Information, Received/Transmitted/dropped calls.

If the difference between two KPIs in a cell is significant according to the statistical test, the difference may also be considered significant for the purposes of making practical judgments. However, if the data set of the locations is large enough, the optimisation decision will always become statistically significant (Okamoto et al., 2006). Each cell is categorised to an optimisation decision where the decision rule for H_0 versus H_1 is defined as follows:

$$u_i = \begin{cases} 0, \ if \ H_0 \ is \ accepted \\ 1, \ if \ H_i \ is \ accepted \end{cases}$$

In this case, 0 might be the decision for upgrade and 1 for optimise the base station. If H_{0i} is tested, at significance level against the alternative H_{1i} then when the null hypothesis is rejected the alternative hypothesis is accepted and in this case the cell should be optimised.

Next, the parameters of the KPI alarms need to be set up. Our purpose is testing hypotheses applied in KPI alarms metrics in order to automate the optimisation process. Specifically, it was used the Two-Tailed Test of KPI False Alarms Population Proportion.

3.5.1. Defining the levels/limits of the KPI false alarms.

The definition of KPI alarms metrics based on existing optimisation projects at global level using various resources such as Hammed, Aderinkola, and Fatimah 2018, (Popoola et al. 2018),(Musa 2016), (Nokia Siemens Networks 2014) to set up hypotheses testing in order to specify the paramaters for the KPI values. In many countries KPI metrics may be defined from the Commission for Communications Regulation (ComReg) as QoS parameters. ComReg is a statutory body responsible for the regulation of the electronic communications sector with the purpose to set up the QoS requirements that are compulsory for the telecommunication providers (Bangladesh Telecommunication Regulatory Comission 2017). (Commission for Communications Regulation Ireland 2012). Initially the limits need to be set up to determine whether a KPI variable X will be considered such as a KPI alarm or not. The limits of the KPI alarms can be modified based on the location and the customers' needs.

For this experiment the KPI false alarm values are defined as follows:

Call Setup Success Rate alarm (CSSR >=98%)

If CSSR(X)<98% then X_{CSSR} will be a Call Setup Success Rate alarm

If CSSR(X) >=98% then X_{CSSR} will not be a Call Setup Success Rate alarm

Radio Access Bearer Alarm rate (RAB FR >=30%)

If RAB_FR(X) >30% then X_{RAB} will be a Radio Access Bearer Failure Radio alarm

If RAB_ FR(X) <=30% then X_{RAB} will not be a Radio Access Bearer Failure Radio alarm

The KPI alarm values of CCR and RAB_FR are addressed based on a percentage rate (%) that the optimisation standards have to keep in a high level in order to provide high Quality of Services

Traffic Model Alarm (TM <70%)

If TM (x)>70% then X_{TM} will be a Traffic Model KPI alarm

If TM (X) < 70% X_{TM} then will not be a Traffic Model KPI alarm

Dropped Calls Rate alarm (DCR <=2)

If DCR(X) <2 then X_{DCR} will be a Dropped Calls Rate alarm

If DCR(X) >=2 then X_{DCR} will not be a Dropped Calls Rate alarm

The KPIs alarm values of RAB_FR are addressed based on a different range of 0.00-30% values. To provide acceptable QoS it needs to keep the range of the Dropped Calls Rate alarms (DCR) at a level lower than 2 and also the Traffic Model Alarms values lower than 70%. When the KPIs values are over the limits then the cells need to be checked in order to support optimisation decisions.

3.5.2 Two-Tailed Test of KPI False Alarms Population Proportion

Given the KPIs false alarms, while proceeding with the formulation of hypothesis testing to analyse the proportions of alarms in different regions of the network. The method of the Two-Tailed Test of KPI False Alarms Population Proportion was used. If the KPI False Alarms Population Proportion (H0: P > 0.20) in a specific network area is more than 20% of the cells (sample size) then the Base Station Controller in this area of the network has to be checked for optimisation issues.

To test the above hypothesis it needs to compare the KPIs correlations using proportions to find out if the values are out or in the rejected zone of hypotheses testing.

3.5.3 Null hypothesis and an alternative hypothesis of the optimisation decisions

Every BCS location contains a sample size from one to four cells in different areas of Dublin. The KPI False Alarms defined of the sample proportion is no greater than 20%. That means that the observed proportion is p =0.02 per base station (cell) so the alternative hypothesis suggests the way to perform the optimisation network as follows: If the observed proportion is p =0.02 per base station (cell) Then the population proportion was set up that is also π = 0.10

If H0 : π <=0.01 then the KPI fault alarm values would be less than 10% then the area does not need to be checked for optimisation (so then accept the hypothesis).

If H1: π >0.01 then the KPI fault alarm values per base station would be more than 10%

3.6 Applying alarm correlation techniques in mobile data

An important issue during the correlation process is the collaborative filtering, which is the process of data cleaning and finding patterns between different KPIs while calculating the degree of correlation between several locations based on the KPI targets. KPI metrics-based computation of the faults comparison is necessary in practical situations, since rating predictions are required to provide recommendations for a geographical area that might contain more than 10 cells/ base stations and two or more Base station controllers (BSC).

To compare the performance of any different area of a set of KPIs, 2000 locations test sets of different combination of KPIs individually obtained using the Pearson’s correlation coefficient (Bayram & Gezici, 2012) between the chosen (selected) KPIs and obtain the following results.

3.6.1 CSSR-DCR Correlation

In the present study, the correlation coefficient between the indicators of dropped and successful calls calculated as -0.58 and presented in Figure 3.3 below.

The results were calculated approximately; while were tested on approximately 600 locations. If the aim is to insert additional locations, the results might change but they will always be at the same rate.

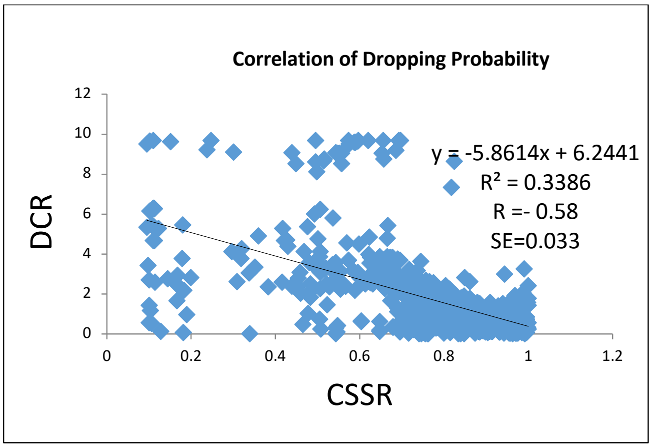


Figure 3.3- CSSR-DCR correlation graph

The correlation graph shows that the KPIs of dropped calls and call success set up rate that are also both related to the retainability and the frequency of dropped calls being negatively correlated. From a practical standpoint this suggests that locations with a high number of dropped calls are also be likely to be suffer a low call success set up rate.

These findings mirror those found in the literature for the most part. As noted by Lehtimäki and Raivio (2005), the literature suggests that in the case of call setup failures, most dropped calls are due to radio channel or air interface problems. Therefore, the same models explain the number of call setup failures due to bad signal quality or the call success set up rate also describe the number of dropped calls due to radio channel problems.

The literature also suggests that the framework for dropped calls may be more complex than can be illustrated with a simple correlation between the indicators for dropped and successful calls, as related to the call success set up rate. For example, most current networks are interference limited, due to the tight frequency reuse patterns used. There are also different ways to calculate DCR: dropped calls per erlangs[1], dropped calls over originated calls, dropped calls over all calls handled by each cell.

[1] An Erlang is a unit of telecommunications traffic measurement that presents the continuous use of one voice path. It is used to describe the total traffic volume of one hour.

These differences need to be mapped in order to address the issue in a cohesive way. Interference may be present in the uplink path (affecting the reception at the base station) or in the downlink path (affecting the reception of the mobile station).

Different symptoms may accompany the causes, including the frequency of transmission, degree of retainabiilty and accessibility, and the probability of dropped calls (Seytnazarov, 2010). The statistical characteristics of interference are different in each path because a mobile receives interference from a limited number of fixed locations (the base stations); whereas the base stations are interfered by a potentially large number of moving mobile stations (Barco et al., 2005).

Ultimately, wherever the call is dropped, the existence of dropped calls has a very negative impact on the quality of service perceived by the end user, and this is the reason why achieving a low number of dropped calls is a crucial objective of mobile network operators and the reason why the call success set up rate is a crucial factor.

Depending on the manufacturer of the equipment, dropped calls can be classified according to the cause of the drop. It should be pointed out that operators do not aim to eliminate dropped calls as the cost would be too high, but to achieve an acceptable percentage of dropped calls (as perceived by the end user) to mitigate dissatisfaction (Wang et al., 2011).

3.6.2. Radio Accessibility & Dropped Transition Correlation

RAB_FR - TM

The correlation coefficient between the indicators of traffic model and radio success bearer failure radio is around =-0.122, calculated approximately, as presented in Figure 3.4, below.

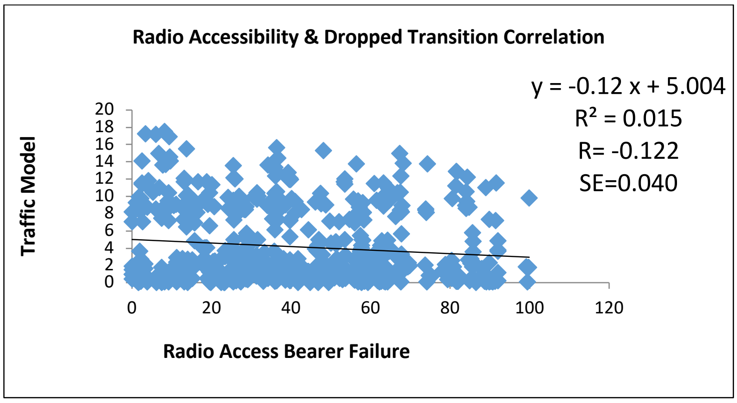


Figure 3.4. RAB - TM correlation

There is a negative correlation between these two sets of data. Even if this set of KPIs are negatively related with similar symptoms, are also evaluated under very similar circumstances, it is possible that they are still partly correlated and that the issues are caused by utilization issues.

The results from this study indicate that there is a relationship between the traffic model and the radio success bearer failure ratio, but it is not significant as the faults or malfunctions have also been affected by other factors. These findings are in line with the wider research literature. As an example, Ouyang and Fallah (2010) demonstrate that a mobile operator can add bandwidth to satisfy increased traffic, but this will not necessarily improve the way in which radio bearers interact with that traffic.

Instead, mobile operators need to reasonably allocate limited network resources to their clients with different priorities during peak hours but need to recognise the intrinsic difficulty in dealing with data traffic such as streaming video, which will result in problems with both network throughput and congestion. The results of a correlation between these factors may therefore provide evidence that there are more packets than supportable bandwidth and extra packets are denied and blocked through radio transfers, and therefore retransmission may burden the overloaded network again (Ouyang & Fallah, 2010).

3.6.3 Frequency of Dropped Calls & Transitions Correlation (DCR - TM)

The Pearson correlation between the indicators of the chosen traffic model and the number of dropped calls is equal to 0.76 calculated, as presented in Figure 3.5, below.

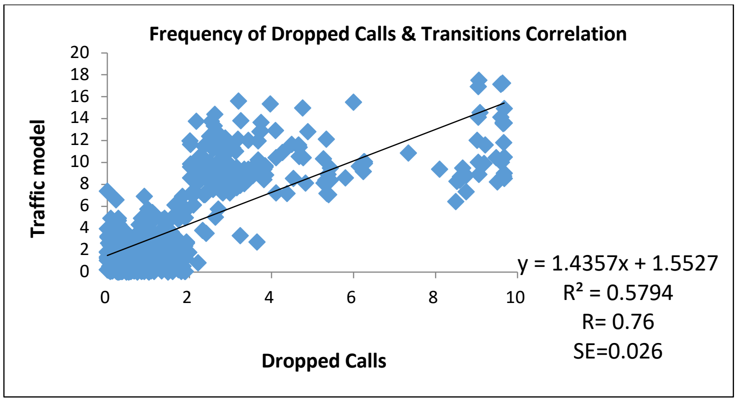


Figure 3.5. DCR- TM correlation

A positive correlation can be seen, that has to be taken in consideration as it shows that where increasing traffic is linked to an increase in the number of calls dropped. As per the previous research findings, there is significant support for this relationship in the literature. Border zones are especially affected by the number of calls dropped when traffic increases. The lack of coverage due to traffic congestion has been classified into two types in the literature, namely a lack of coverage in the borders of the cell and shadow zones within a cell, both of which can easily become overloaded. Lack of coverage in the borders usually shows up in rural areas, i.e. in areas with low density of population. In these areas, tele-traffic density is normally low and cells tend to have large coverage areas (Jain et al., 2011).

To this end, the ability of providers to manage handoffs in a mobile cellular communications environment have become an increasingly important issue accommodating an increasingly large demand for high traffic on an ongoing basis, with simultaneously reduced cell sizes (Bhowmik, Roy, Guha Thakurta, & Sarkar, 2011).

3.6.4 RAB and DCR Correlation

The correlation between the indicators of radio success bearer failure and dropped calls rate is equal to -0.22 calculated approximately, as presented in Figure 3.6, below.

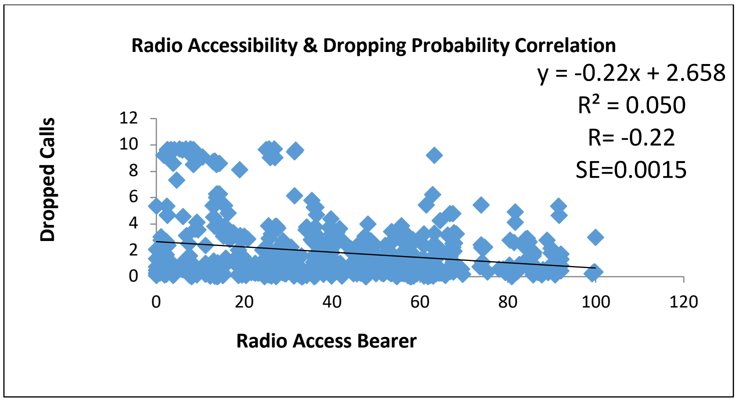


Figure 3.6 . RAB & DCR correlation

In this case the two KPIs are negatively correlated. There is not a complete disagreement (-1) between the radio success bearer failure radio and the dropped calls rate, though the difference is not significant, and must be seen as a negative correlation. As a result, the radio channel failure accessibility cannot be the only symptom of the frequency of dropped calls. For example, a call can be dropped because the distance between the mobile station and the base transceiver station is too great. This cause can be related with the synchronization method of the radio subsystem in 3G networks.

The solution in GSM is that the base transceiver station sends to each mobile station a timing advance parameter according to the perceived round trip propagation delay the route from the base transceiver station to the mobile station and back to the base transceiver station afterwards (Barreto, Mota, Souza, et al., 2004).

These types of challenges related to identifying the cause of a high frequency of dropped calls are reflected in the findings from the literature. For example, although efficient handoff algorithms to predict the frequency of dropped calls are a cost-effective way of enhancing the capacity and quality of service of cellular systems; this is something that is not always addressed in practice (Jain et al., 2011).

Ultimately, the most significant way to increase the system capacity of a wireless link is by getting the transmitter and receiver closer to each other, which creates the dual benefits of higher quality links and more spatial reuse. In heavy populated areas the need arises to shrink the cell sizes and scale the coverage pattern. The extension of the service into different domains such as railway stations, malls, pedestrian areas, markets and other hotspots further enhances this trend (Chandrasekhar & Andrews, 2009).

3.6.5 Retainability and Dropped Transitions Correlation (TM & CSSR)

The correlation that needs to be identified between the indicators of the traffic model and the call set up success rate is -0.59, calculated approximately, as presented in Figure 3.7, below.

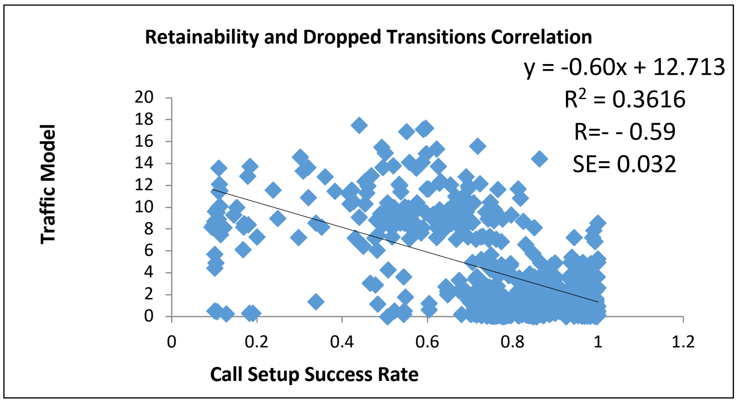


Figure 3.7- TM & CSSR correlation graph

The correlation of the indicators of retainability and the frequency of dropped transitions show that the traffic channel drops, and the successful call rate are negatively correlated. Thus, the results are significant as clearly show that there is a correlation indicating that the ability of the network to keep up a call is linked to transition frequency. Specifically, the literature reflects this finding in that it has been determined that over-dimensioning increases costs and reduces utilization below capacity, and in order to prevent these issues, the core network must have the appropriate number of channel groups (or group of frequencies within a specific connection) (Cusani et al, 2007).

To this end, there are proven correlations between the frequency of transmission, the degree of retainabiilty and accessibility, and the probability of dropped calls (Seytnazarov, 2010).

3.6.6 RAB_FR & CSSR Correlation

A correlation between the indicators of the radio success bearer failure radio and the call set up success rate results in a calculation of 0.080 calculated, as presented in Figure 3.8, below.

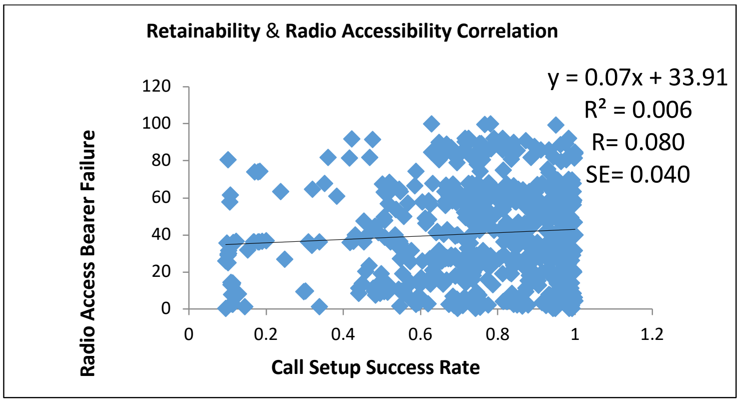


Figure 3.8 - CSSR & RAB_FR correlation graph

The KPIs of call success set up rate and radio bearer failure were shown to be uncorrelated. The correlation shows an almost flat relationship between the two variables from a practical standpoint. There is minimal literature concerning the correlation between these two variables, but some findings align with these results.

It may be feasible that user requests are not served due to problems in resource allocation or the call set-up phase of the transaction, as noted by Lehtimäki and Raivio (2005). As in the case of service blocking, a model describing the contributions of each base transceiver station to the total number of call setup failures in the network is defined.

Nonetheless, the most common reasons for failures during the call setup phase or actual service are the inadequate radio signal propagation conditions, which indicate a higher than normal level of problems in the air interface, in line with the slight positive correlation in the current study. Failures in the radio channel are usually due to bad signal quality, namely that the transmitted data includes too many bit errors.

3.7 Evaluation of applied statistical techniques

The results of the correlation study will be utilized and extended in the following Chapter to identify the levels of the KPI false alarms. The correlation results reveal the impact of different types of mobile services, however the KPI metrics need to be clearly defined and tested to demonstrate the false alarms and the acceptable scores as support for optimization decisions.

Alarm correlation consists of the conceptual interpretation of multiple alarms. It is a process that involves different tasks: reduction of multiple occurrences of an alarm into a single alarm, inhibition of low priority alarms in the presence of higher priority alarms, substitution of a specific set of correlated alarms by a new one, etc.

This way, alarm correlation systems are required to filter and condense the incoming alarms to meaningful high-level alarms in order to avoid overloading the operators.

Although alarm correlation can be considered a first step in the diagnosis of faults, alarms do not provide enough information to identify the cause of problems, especially if the possible causes are not only faults in pieces of equipment.

3.8 Results of technical optimisation

Generally, it can be observed that the correlation results to be verified at the population proportions of KPIs. For instance the KPIs of DCR and TM are usually both accepted or rejected together in any network location (that verifies the correlation results) to contrast with the CSSR and RAB_FR indicators values that could be accepted or rejected independently from the other indicators.The figure 3.9 presents the correlation results for a specific Base Station Controller that manages and monitors a number of base stations/cells. It can be seen that the TM and DCR correlation graph shows the results in ratio scale while the RAB & CSSR presentation of the results is in percentage scale according to the KPIs metrics defined in 3.1.

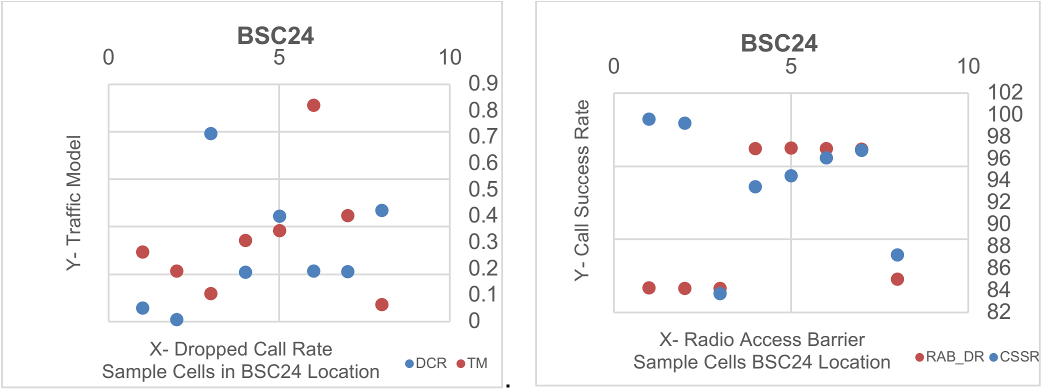


Figure 3.9 – KPI alarms correlation in BSC24 location

The graph of DCR and TM KPIs shows that the proportion of false alarms for both the indicators at the base station location cell BSC24 based on the values of correlation coefficient. The base station cells are selected randomly to present a general overview of the KPI alarms correlations.

The same results are also verified by the hypothesis testing procedure. Based on these results QoS can be provided since the Dropped Calls Rate alarms (DCR) is lower than 2 and also the Traffic Model Alarms is lower than 70%. Optimisation would be done by checking radio channel accessibility issues.

In the second graph it can be seen that the RAB KPI provides results lower than 30% in about 400 different base stations, so the area needs to be checked for Radio Channel accessibility issues. The same results are also approved by the hypothesis testing results. The KPIs alarm values of RAB_FR are looked at in regard to a percentage rate (%) that the optimisation standards have to keep in a high level in order to provide high QoS.

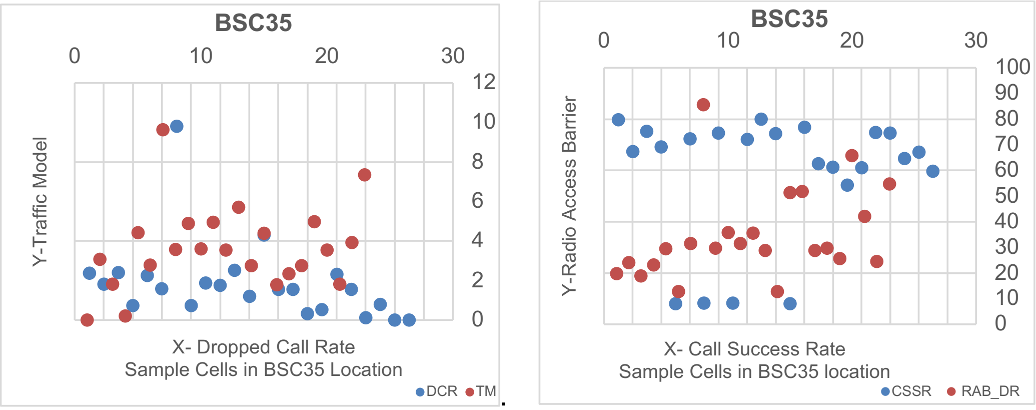


Figure 3.10 – KPI alarms in BSC35 location

The first graph in Figure 3.10 (BSC35 case) shows that even if the KPI values of DCR and TM are not all acceptable regarding the alarms limits (X_{TM}>70% and X_{DCR}>2), the test statistic (z) values are not rejected by the hypotheses testing. It is evident that the results depend on the amount of sample size, the rejected values distribution and also the population proportion. In order for optimisation to take place Radio Channel Accessibility issues should be checked.

In the second case (CSSR, RAB) the X_{CSSR} and X_{RAB_FR} values are also reasonable even though some of them are under the (70%) limit. Although, X_{RAB_FR} t statistics values are not acceptable by the hypothesis testing and the cells need to be checked regarding the Radio Channel accessibility issues, optimisation can still be achieved.

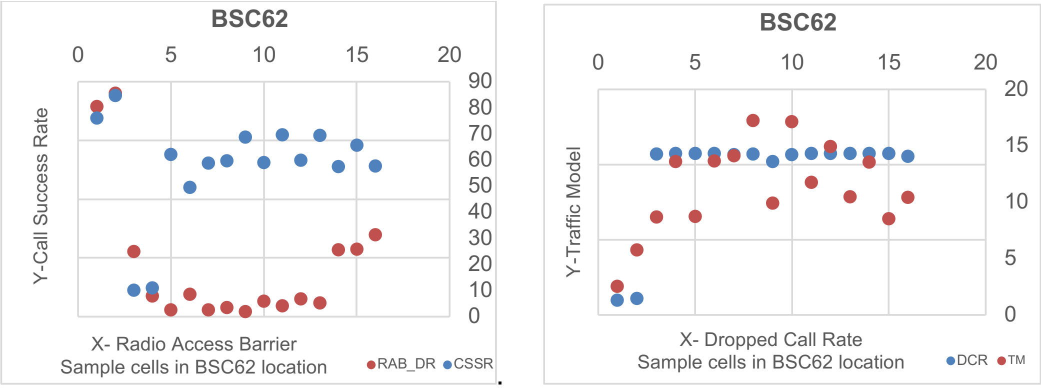


Figure 3.11 – KPI alarms in BSC62 location

In the KPI set shown in figure 3.11 almost all the KPIs values falls to the expected optimisation standard values. Also the test statistic values are all rejected so it can be clearly seen that the majority of the KPI alarms are in the critical area.

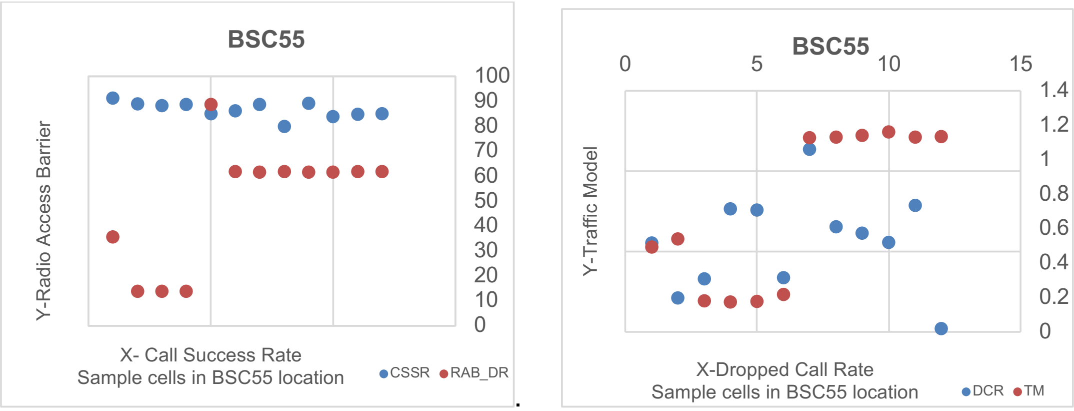


Figure 3.12 – KPI alarms in BSC65 location

Finally, in the KPI set shown in figure 3.12 (BSC55 area) it can be observed an optimised network area that also all the KPIs t testing values are accepted based on the hypothesis testing results.

3.9 Summary

This part of the thesis had the goal of supporting engineers in the decision making process for optimising an existing mobile communications network based on KPI alarms. Areas that require automation to offer a cost-effective implementation are those involving heavy calculations and/or the evaluation of several input parameters, according to the findings from this data and its assessment in comparison to the literature. In contrast to humans, computers can easily carry out analysis of complicated input data. In addition, computers are able to repeat a method many times and thus enable the application of the method to all cells in the network. The benefit of automation to these kinds of tasks is an increase in network performance since without the application of automation these tasks could not be carried out on a wide scale, i.e. a large number of cells.

Automation can also be applied to existing tasks, e.g. frequency planning. In this case, one positive side effect of increasing the level of automation is that the staff is freed from mundane and tedious tasks and can apply their skills in new areas, which bring additional value to the operator. In this sense, automation assures that the staff is working on challenging tasks, which in turn helps to motivate and retain staff.

In addition, the clearance of faults, which includes the detection, isolation and correction of faults, where a fault is a cause of malfunctioning, is also very important in order to ensure that the network operates exactly as designed. If a cell is temporarily non-operational, the performance of all cells in the vicinity is impaired.

The research studies reviewed herein that look at the automation of diagnosis in the radio access network of cellular networks have been focused on alarm correlation. In current cellular networks, most systems are semi-intelligent and generate alarm messages when errors occur. Next step of this work will be to perform the automation of KPI alarms using data mining techniques to improve network performance.

Chapter 4. Automated Network Troubleshooting Using Data Mining Techniques

Alarm correlation can be considered a first step in the diagnosis of faults, but does not provide enough information to identify the cause of network issues as our findings in chapter 3 demonstrated. However, defect management on telecommunication networks has been studied extensively utilising the alarm correlation techniques. In this way, data mining techniques help to move existing knowledge forward regarding fault diagnosis. Data mining algorithms contribute to identify hidden relations between network variables, multivariate dependencies, network faults while providing an output that is useful for fault analysis and predictions. Thus, there has been a limited amount of research regarding the degree of fault alarm under data mining techniques. This makes it much more reliable for alarm localisation and defect detection and management than would otherwise be presumed. This is one of the areas that this research will explore in much more detail in order to illuminate the possibilities for defect management and network optimisation

This chapter proposes an automated scheme for network troubleshooting and fault detection based on several data mining techniques. Both supervised and unsupervised techniques can be used to define the suggested network management scheme. Supervised techniques contribute to the learning process using labeled classes for KPI alarms in order to create an optimisation network algorithm to extract outputs.

This will indicate hidden patterns between the network data sets and additional rules for the network faults. Unsupervised learning can be used to unfold the distribution of the data and KPIs, to draw inferences and discover unknown patterns between the datasets consisting of input data without using the labeled data of KPI alarms. The approach and evaluation aim to show that it is possible to carry out effective and accurate monitoring of services in order to detect and classify network anomalies that may arise. Appropriate measures can then be taken to ensure correct fault-handling and compliance with the Quality of Service constraints established between service providers and end-users. Administrators need to recognize that a network communication failure is occurring at the earliest possible opportunity.

With current methods, network operators must conduct manual searches through large amounts of data in order to discover network malfunctions. Furthermore, mobile networks are often handled through different channels such as control channels and traffic channels, with each likely to have different traffic characteristics and anomalies, as well as different faults. The entire process is not only complicated but can require hours of work in order to diagnose and return a network to fully optimized operation.

The monitoring scheme is based on the use of rules extracted from decision tree algorithms, and identification of clusters that indicate the causes of network malfunctions to upgrade fault detection and handling. The rules were defined using weka 3.6 encoding in java code. In addition, a monitoring scheme that relies on Bayesian classifiers was also implemented for the purpose of fault isolation and localisation.

A fault localisation protocol monitors network data in order to isolate abnormalities such as high drop calls, high traffic rate and/or forgery on a certain cell (Zhang 2012).

4.1 Background of network troubleshooting system

During a network troubleshooting process, several applications and databases have to be queried to analyse performance indicators, cell configurations and alarms. The speed in identifying faults is dependent on the type and accuracy, the level of expertise of the troubleshooter and their ability to find and interpret relevant information. This means that in addition to knowledge of the possible causes of problems, troubleshooters require a very good understanding of the tools available to access the sources of information is also required.

The ability for a human to visualise all of the alarms in a specific area per day that may be generated is difficult, hence alarms generally need to be aggregated so that they can be examined more easily (Monacelli, & Francescangeli, 2011). It is also possible to implement an automatic fault identification module that can correlate observed alarms in order to indicate the root cause. With an automatic fault identification module, a human may not even be needed as an automatic recovery system can subsequently be used to quickly correct the fault (Monacelli, and Francescangeli, 2011).

The first steps in producing systems to automatically troubleshoot the radio access network of cellular networks have focused on performance visualization and fault detection. Thanks to methods to achieve efficiency in network performance, fault detection and diagnosis can be carried out more easily. Subsequently, several methods have been proposed for fault detection, which build models for the normal behavior of the system and define some form of anomaly detector.

Findings from the literature have shown that in the radio access network of cellular systems there are two main reasons why alarms are not the only indicators of malfunction. Firstly, most faults are not related to a physical component, but to poor network planning or an incorrect parameter setting. Thus, modelling the interactions among entities is not a requirement as in other communication networks because, in most cases, faults are not related to pieces of equipment which are physically connected.

Secondly, although alarms play a very important role in identifying faults in specific pieces of equipment, they do not provide the conclusive information needed in order to isolate configuration problems. The fact that performance indicators in the radio access network domain are continuous, rather than binary (like alarms), leads to difficulty in modelling, which is non-existent in the core network domain.

4.2 A Proposed troubleshooting management framework

In this chapter the use for decision trees for automated fault detection was considered. The proposed methodology is the inference engine for network troubleshooting as shown in Figure 4.1. The engine maintains a list of KPI values and potential causes of faults identified by experts from historical data. Based on the data and the interpretation of the user, the engine can perform calculations on the correlation between the metrics and the faults that are occurring, before fault optimisation rules are applied.

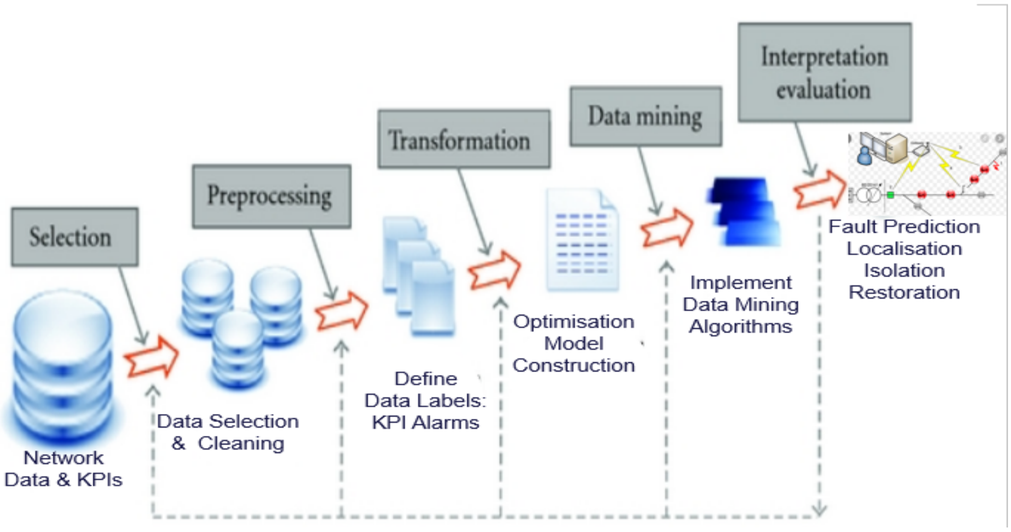

Selection
Preprocessing
Transformation
Data mining
Interpretation evaluation
Network Data & KPIs
Data Selection & Cleaning
Define Data Labels: KPI Alarms
Optimisation Model Construction
Implement Data Mining Algorithms
Fault Prediction Localisation Isolation Restoration

Figure 4.1 Troubleshooting management using data mining

The first step in the data mining process is *data cleaning*, in which mobile network data referring to different areas are examined and converted to KPIs to indicate the class attributes which define the relevant set of instances for the given task or purpose. Irrelevant and invalid data for network optimisation are removed.

Next, data integration occurs, in which data from multiple sources such as raw data of transmitted/received call seizures, traffic channels, MessageSignal Units (MSU); various KPIs shown in Table 4.1 and , and spreadsheets that show the fall percentage and cell performance. All the different sources are combined into a single source or dataset as the data were integrated from multiple sources such as Standalone Dedicated Control Channels stats.

The next step is data selection, in which further examination of the telecommunication data occurs in order to determine which KPIs should be taken from the dataset to be used for the task at hand. Once data selection has been completed, the selected data is transformed into appropriate formats for the mining procedure (Srimani, & Balaji, 2014).

Following this, the actual data mining takes place by applying a classifier algorithm for the learning process. In this step, various techniques for analysing data can be used (Patil, & Joshi 2009). Then, pattern evaluation is undertaken in which interesting or unique patterns in the data are identified in relation to the larger issue or problem that is attempting to be addressed (Srimani & Balaji, 2014).

With the use of data mining tools, the final step in the process is knowledge representation. Moreover, class attributes must be defined followed by the actual extraction of the features that will be used for classification. Consequently, a subset of features is selected to be used as part of the knowledge construction (Srimani & Balaji, 2014).

The following part of the knowledge representation using data mining tools is the investigation of any imbalances in the selected data and a determination of how those imbalances may be counteracted. For instance, an imbalance issue in a network optimisation plan will be when given a dataset consisting of data, KPIs and fault classification rules, the classifier will tend to categorise cells that are Critical so they need to be fixed as soon as possible as Normal. Consequently, a subset of instances is chosen. Finally, a testing method is chosen in order to estimate the performance of the algorithm (Srimani, & Balaji, 2014).

4.3 Multi-label classification algorithms

Classification algorithms are ubiquitous in the machine learning sector these days. However, the traditional single label classification is concerned that learning a group of attributes that are connected with an only single class (Tsoumakas & Katakis 2007). Consequently, binary classification methods have been developed that can be extended to multi-label classification techniques to evaluate the three different categories of network faults that were considered in table 4.2.

The ability to distinguish between different classes of faults can only occur if training data is available for all of the different issues that can occur. The system must learn the anomalies that have occurred in the past in order to be able to detect different types of issues in the present.

Providing data for different sources encourage us to examine different data mining models such as decision tree and rules classifiers that can learn differences between normal activities in a network and activities that are not normal and also extract valuable rules and hidden relations between the network data. The next step was observing similarities, trends and measuring probabilities found in our datasets using clustering algorithms and Bayesian models. With more historical fault data, the system can achieve higher rates of effectiveness at detecting real problems, as well as reducing the rate of false alarms (Li and Manikopoulos, 2002).

4.4 Decision Tree and Decision Rules Algorithms

A decision is defined as a structure that is used to classify data that have common attributes (Park et al., 2009). A decision tree represents a set of rules around which data are categorised. Within a decision tree, there are internal nodes, leaf nodes, and edges. Internal nodes are specific attributes around which data are partitioned. Leaf nodes are labels for categorization of the data. Edges are labels for possible conditions of the attributes contained in the parent node. Decision trees and decision rules algorithms follow a set of if-else conditions based on predefined optimisation rules with the purpose of extracting additional rules using the most important variables of a dataset. These algorithms have been chosen for this part of the thesis as they are easy to interpret and identify the most significant variables for the proposed optimisation model. Waikato Environment for Knowledge Analysis (Weka) software version 3.4.3 was used for this part of the thesis.

4.4.1 J48 tree

The C4.5 (J48) is an algorithm used to generate a decision tree developed by Ross Quinlan (1993). The decision trees generated by C4.5 can be used for classifying datasets while decision trees such as random forests can be also used for regression analysis. The J48 decision tree classifier is the implementation of the C4.5 algorithm developed by the Weka project team. The decision tree algorithm generates the rules for the prediction of the target attribute. Decision tree induction initiates with a dataset of attributes that is split into small subsets based on the optimisation rules that were set up for the target attribute. (Sharma, Jain 2013). To every rule is associated a class of label which identifies whether a cell in a particular location belongs to this class or not.

4.4.2 JRip

The JRip classifier, which was introduced in 1995 by W. W. Cohen, is an optimised version of an older classifier that Cohen created (Shahzad, Asad, and Khan, 2013). JRip was implemented with the prepositional rule learner known as Repeated Incremental Pruning to Produce Error Reduction (RIPPER. With JRip, it is possible to isolate some of the data being examined for training. At the same time, a decision can be made regarding the set of rules generated for selected attributes.

4.4.3 PART

The PART algorithm (Deljac, Mostak and Stjepanovic, 2010) is an algorithm that does not generate accurate rules through global optimisation. Instead, PART incrementally builds a rule and then removes the instances that it covers in order to create a recursive rule, until there are no instances that remain. Another way of thinking about the PART algorithm is that it uses the input labels and generates decision lists. With the decision lists, new data are compared to the rules in the decision list. Upon the first match of a rule, data are assigned to that rule. In this way, PART makes the best leaf into a rule rather than the optimised leaf (Nor Haizan et al. 2012).

4.4.4 Bayes Networks

Bayes Networks, or Bayesnet, are also known as belief probabilistic networks. It is the structure of a Bayes network that determines the relationships and dependencies between the variables in a dataset. Because of the fact that the anatomy of Bayesian networks stimulate conditional dependencies and relationships between variables, the use of such networks have been recommended for use in the diagnosis of faults in some cellular networks (Barco et al. 2006).

4.4.5 Naïve Bayes

The Naïve Bayes classifier (Ali, Shehzad, & Akram, 2010) operates in accordance with the Bayes rule of conditional probability. The classifier uses all of the attributes of a dataset by analysing them individually. The idea is that all of the attributes are equally important. No single attribute is more important than another attribute. At the same time, all of the attributes of the data are considered to be independent of each other (Srimani, & Balaji, 2014).

4.4.6 K-Mean Clustering

The* K*-means clustering algorithm uses iterative filtering to produce groups of data. The algorithm inputs are datasets having "n" data points that are partitioned into "k" groups or clusters. (Haraty Dimishkieh and Masud 2014). The algorithm runs while valuating the K centroids. It is usually unsupervised and can find unknown patterns from the data set without pre-existing labels. The k-mean clustering algorithm can also be supervised if the input of the dataset (X) are inserted with pre-defined labels (Y) that determine the desired clusters for unseen sets of items x. The algorithm has been widely used in the telecommunication sector for customers profiling and segmentation.

4.5 Preparation of input data

To compare the performance of any different area of a set of KPIs, 2,100 instances were used that represent the performance of Cell IDs and Base Station locations in the Dublin area, with 5 different sources of data shown in tables 4.1& 4.2. The data files showing on the database model contain raw data, cells with poor performance, KPIs, channels failures, correlation results from the previous chapter and KPIs alarms. Hence, 25 input attributes are selected from our databases to run the data mining models. The input data of our datasets are shown below:

TABLE 4.1 NETWORK DATA MODEL

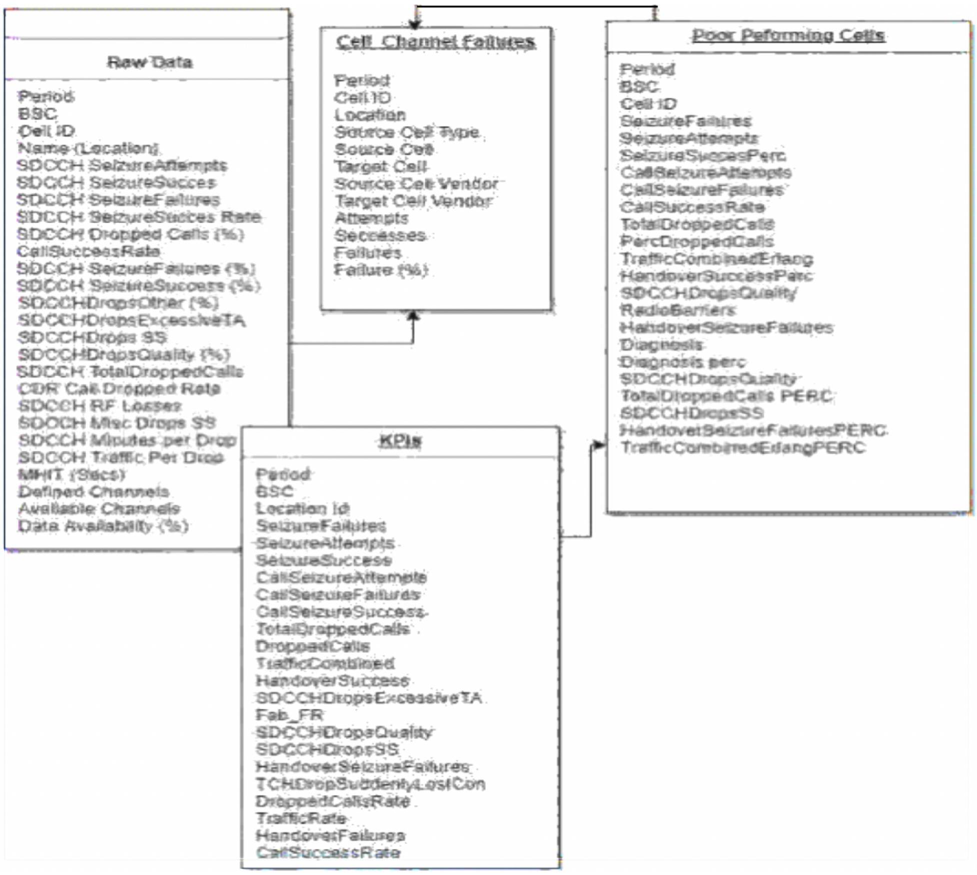
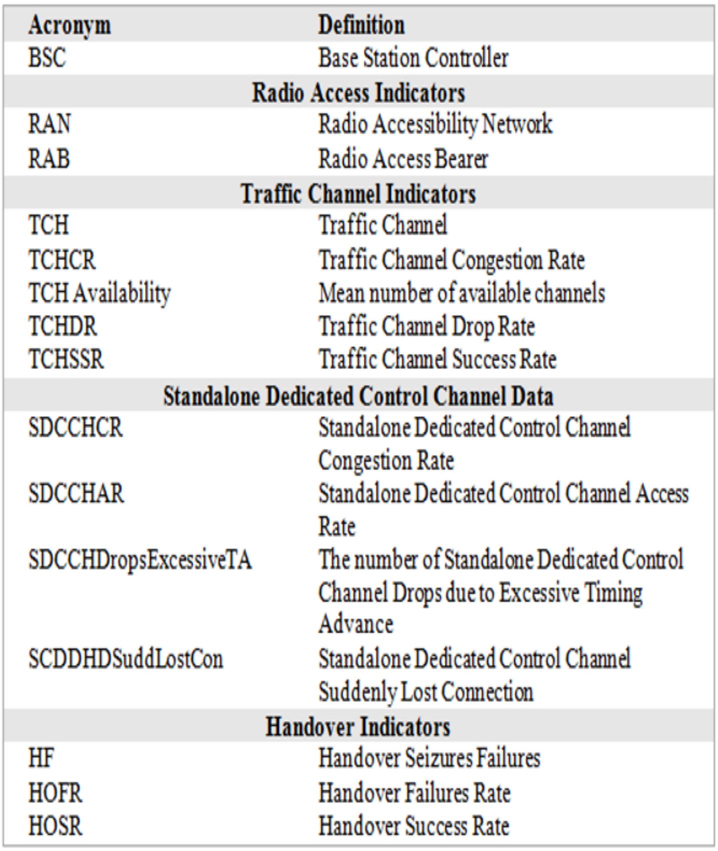


TABLE 4.2. INPUT DATA-RAW DATA AND KEY PERFORMANCE INDICATOR ABBREVIATIONS



Acronym	Definition
BSC	Base Station Controller
Radio Access Indicators	
RAN	Radio Accessibility Network
RAB	Radio Access Bearer
Traffic Channel Indicators	
TCH	Traffic Channel
TCHCR	Traffic Channel Congestion Rate
TCH Availability	Mean number of available channels
TCHDR	Traffic Channel Drop Rate
TCHSSR	Traffic Channel Success Rate
Standalone Dedicated Control Channel Data	
SDCCHCR	Standalone Dedicated Control Channel Congestion Rate
SDCCHAR	Standalone Dedicated Control Channel Access Rate
SDCCHDropsExcessiveTA	The number of Standalone Dedicated Control Channel Drops due to Excessive Timing Advance
SCDDHDSuddLostCon	Standalone Dedicated Control Channel Suddenly Lost Connection
Handover Indicators	
HF	Handover Seizures Failures
HOFR	Handover Failures Rate
HOSR	Handover Success Rate

Radio Access Indicators

The radio access indicators such as RAN (radio accessibility network performance audit) and RAB (Radio Access Barrier) show the level of network connectivity (Ali, Shehzad, & Akram 2010).

Traffic Channel Indicators

The traffic channels indicators are related to the The TCHSS shows the percentage of TCHs that are successfully seized. TCH Seizure Attempts is also the number of traffic channels that are allocated in a specific area. The TCH Availability is the rate of Traffic Channels availability in the network, while the TCH Drop is the total drop rate of traffic assignment failures in the network and also the TCHCR is the rate of TCH congestion rate related to call setup traffic in the network area (Shahzad, Asad, and Khan, 2013). The Traffic Channel Congestion Rate of the SDCCH drops measures the total number of RF losses as a percentage of the total number of call attempts for the SDCCH channels on the network. The SDCCD Availability Rate is the percentage of SCDDH channels that are available on a network at a given time (Shahzad, Asad, and Khan, 2013).

Standalone Dedicated Control Channel

The Standalone Dedicated Control Channel Congestion Rate (SDCCHCR) indicates the probability of accessing a stand-alone dedicated control channel available during call set up. The SDCCH Access Rate shows the percentage of call access success rate received in the base station location (Ali, Shehzad, & Akram 2010). The SDCCH Drops is the number of drops due to low signal strength or network congestion (Shahzad, Asad, and Khan, 2013). The SDCCH Drops Excessive TA presents the number of drops due to excessive timing advance (Ali, Shehzad, & Akram 2010). The SCDDHDSuddLostCon is the rate of connections that are suddenly lost on a network.

Handover Indicators

The Handover Success parameter shows the percentage of success handovers on the network of all handover attempts (Skianis 2013). Based on those relational nodes, the system would identify the optimised status of the TCH (traffic channel). The HOFR statistic is intended to give an indication of the rate as a percentage of handover failures in relation to total handovers (Shahzad, Asad, and Khan, 2013), (Hammed and Fatimah 2018). The Handover Success Rate is the percentage of rate of handover success which is the successful handovers divided by the total handover requests. Once classification is settled down, representation of data can use data visualization techniques of decision trees and rules classifiers.

4.5.1 Preparation of target variable

A crucial step of problem definition in a data mining experiment is defining the target variable. Thus, our work of data preparation continues with analysing the process that relates to the data and missing value implications,. The selection of the data to be analysed involves several considerations, such as which data are relevant for the KPI symptoms, selected row data (inputs), the types of attributes, and establishing limits of the network faults

The KPIs are derived with the help of measuring network performance using different formulations. A single counter helps to provide a very limited indication of the larger network. However, with the use of several counters, it is possible to gain a much broader view of the network (Barco, et, al, 2006).

In this study, our evaluation is presented on the basis of the four major KPIs identified in chapter 3 used as an input to create a classification algorithm perform on the data for the pre-processing process. The data pre-processing stage is very important in a data mining process as raw data will be transformed into a meaningful and useful format suitable for efficient data manipulation and data visualization. The correlation results demonstrated that KPIs of dropped calls and call success set up rate, which are related to both the retainability and the frequency of dropped calls, are negatively correlated. From a practical standpoint this may suggest that a high number of dropped calls would also be likely to be caused by a low call success set up rate. The following KPI metrics confirm those found in the literature, and explained in the previous chapter.

TABLE 4.3 KPI METRICS AND ALARMS CONDITIONS

State Alarm Indicator			
KPI	**Normal** ***(Norm)***	**Critical** ***(Cr)***	**Warning** ***(Warn)***
DCR	DCR<2	4>DCR>=2	DCR>=4
CSSR	CSSR>=98%	90%>CSSR>=98%	CSSR<=90%
TR	TR<60%	60%>TR>=70%	TR<=70%
HOF	HOFR<10%	10%>HOFR>=25%	HOFR>=25%

The causes and KPI alarms defined in chapter 3 are represented as a categorical variable with three states: Normal, Critical and Warning.

Consequently, the rules generated for the class labels of the target variable are based on the fault classification algorithm given below:

Input data:

Training set of KPIs = {(xi, yi,)},.n= (i=Base Station/cell)

X_i is a set of KPI objects: M = {(DCRi, CSSji, TRi, HOFRi,)}
Y_i is a set of KPI alarms attributes: N = {NRMi, CRi, WRNi,}

Get values for M and N, the size of the data set

```
do if DCR>2 AND CSSR>=90%
    do if DCR<=4 AND CSS<=98%
        then KPI alarm class =CR
  else if TR>60% AND HOF> 10%
     do if TR<70% AND HOF<25%
          then KPI alarm class =CR
  else if DCR>4% OR CSSR<90%
       do if HOFR>25% OR TR>70%
          then KPI alarm class =WARN
  else if KPI alarm class =NORM
        end if
```

4.5.2 Network fault classification algorithm

The performance of the optimisation system that derived versus the fault detection of the selected sets of the KPIs is the main scope of the fault classification algorithm. KPI metrics-based computation of the faults comparison is necessary in practical situations, since optimisation predictions are required to provide recommendations for a geographical area.

TABLE 4.4 DATA MODEL WITH INPUT TRAINING DATA

Input Variables (Network Metrics)							
Period	BSC	ID	SDCCHCR	TCHSeizureAttempts	SDCCHAccessRate	TCHCR	
RAN	SDCCHDR	SDCCHAvailabilityRate	HF	SDCCHDropSuddLostCon	HOFR	HOSR	
TCHTR	HandoverSuccess	TCHSS	TCHDrop	SDCCHDropsExcessiveTA			
TCHDropRate	TCHCongestionRate	SDCCH Drops	TCH Avalability	TCHAssignmentFailureRate			

Dependent variables (Network KPIs)			
CSSR	TM	HOFR	CDR

Class Labels of the Categorical Variable
KPI Alarms: “NORM”, “CRITICAL”, “WARN”

Fault Classification Algorithm

Aim: Derive additional optimisation rules and hidden patterns between the network metrics.

Define an efficient optimisation scheme according to the rules and decision trees classifications models.

Outputs:

Output 1: Learn the JRP and PART algorithms to obtain additional optimisation rules according to the class labels.

Output 2: Learn the C4.5 (J48) and algorithms to obtain hidden patterns between the input variables.

Initially, in order to select the parameters of the model, the data were divided into training and testing sets. The test data is comprised of 30% of our dataset, while the remaining 70% of data was used as training data. The training data was used for parameter estimation, while the test set was used for evaluation of the methodology (Lantz 2015). Additionally, cross-validation and percentage split test options were used. To ensure a different result each of the ten times, a new random seed was provided in the options for the percentage split each time.

For test networks, training data sets that contain normal and faulty data are available. This creates an obstacle for network technicians and administrators because the classifier for network troubleshooting will not be well trained for the real conditions of the network environment. This is certainly the case if the real network is greatly different from the test network on which the classifier was trained.

In addition, without training that includes a range of fault data for the real network, the classifier will be even more inefficient at detecting actual network faults. If these conditions are present, then additional training will need to occur so that the classifier can be effective at detecting faults on the production network (Li, & Manikopoulos, 2002).For this reason the dataset with network data and KPIs contains labeled training data with network faults categories to learn the mapping function that turns input variables (X-data, KPIs) into the output variable (Y-KPIs faults).

4.5.3 Rule sets extracted as an output from the classification models:

Weka tool was used for learning the decision trees and rule classification algorithms. Initially, the training data set was used with selected target attribute that gives the best split of our optimisation rules using JPR and RIPPER for showing the different optimisation rules and J48 and random forest algorithms to visualise the decision trees.

The six rules shown below were extracted regarding the JPR classifier showing the conditions of critical optimisation faults and the KPI metrics to show when a cell should be considered for optimisation. The last rule that was generated tells the system that if the cell is not classified as either normal (NORM) or critical (CR), then it needs to be checked for a warning (WARN) status.

TABLE 4.5 OUTPUTS OF JRP ALGORITHM

JRP Optimisation rules
Output variables taken in consideration from JRP Algorithm: TCHDropRate, TCHSeizureAttempts, SDCCHCR, TCHSeizureAttempts, TCHDropRate, RAB, TCHAvailabiity, TCHSS, HF
a) Output: Set of Rules I
(TCHDropRate<=1.96) AND (TCHSeizureAttempts <=59.78) and (SDCCHCR <= 5.57)=> KPI Alarms= NORM
b) Output: Set of Rules II
(TCHSeizureAttempts <=12.93) AND (TCHDropRate<=5) => KPI Alarms= NORM
c) **Output: Set of Rules III**
(TCHDropRate<=4.99) AND (RAB >=51.09) AND (TCHAvailabiity <=99,22)=> KPI Alarms= CR
d) **Output: Set of Rules IV**
(TCHSeizureAttempts <=59.14) AND (RAB >=49.52) AND (TCHSS > 90.01)=> KPI Alarms= CR
e) **Output: Set of Rules V**
(RAB >=15.98) AND (HF>=714) AND (SDCCHCR <= 7.97) => KPIAlarms =CR

The Part algorithm output rules shown below provide the attributes of TCHDropRate, RAB, HOFR, SDCCRDR, TCHAvailability TCHCR, TCHDrop, TCHSS in order to define the optimisation decision outputs. A total of five rules were extracted from Weka regarding the PART classifier.

While JRip was implemented with the prepositional rule learner known as Repeated Incremental Pruning to Produce Error Reduction (RIPPER), meaning that it has the ability to replace or revise our optimisation rules. Consequently, a decision can be made regarding the set of rules generated for selected attributes (Rozaki 2015).

TABLE 4.6 OUTPUTS OF PART ALGORITHM

Optimisation Decision List Rules
Attributes considered for the optimisation rules: TCHDropRate, RAB, HOFR, SDCCRDR, CHAvailability TCHCR, TCHDrop, TCHSS
a) Output: Set of Rules I
TCHDropRate<=1.96 AND RAB < =35.91 AND HOFR <= 6.76 : NORM
b) **Output: Set of Rules II**
TCHDropRate> 9.98: WARN
c) **Output: Set of Rules III**
RAB > 51.11 : WARN

d) **Output: Set of Rules IV**
SDCCRDR > 193 AND TCHDropRate> 3.01: WARN
e) **Output: Set of Rules V**
TCHCR < = 4.16 AND TCHAvailability > 90.66 AND TCHDropRate <= 5 AND TCHDropRate > 2.35 AND TCHDrop > 1.6: CR
f) Output: Set of Rules VI
HOFR > 7.99 AND TCHAvailability > 90.66 AND TCHSS > 89.98 AND TCHDropRate> 5: WARN
g) Output: Set of Rules VII
TCHCR <= 3.46 AND TCHAvailability > 98.99: CR RAB <= 30.32: NORM WARM

Moreover, PART functions build a rule and then remove the instances that it covers in order to create a recursive rule until there are no instances that remain. Another way of thinking about the PART algorithm is that creates decision lists and generates different rules sets that show specifically the types of alarms that are captured by the classifier. With the decision lists, new data are compared to the rules in our optimisation decision list. (Rozaki 2015). As a result, the J48 algorithm examined the variations of additional attributes such as TCH Seizure Attempts and captured extra alarm cases while the PART algorithm rules have shown a variety of combinations with less attributes trying to keep the models simpler.

As before, the rules that were generated illustrate the conditions of critical optimisation faults and the KPI metrics to show when a cell can be considered to be optimised. The rule based detection system now expects an alarm to occur after other alarms have been triggered. The rules that indicate the normal (NORM) classification shows the records that need to be changed or adjusted so that the network can be returned to an optimised state.

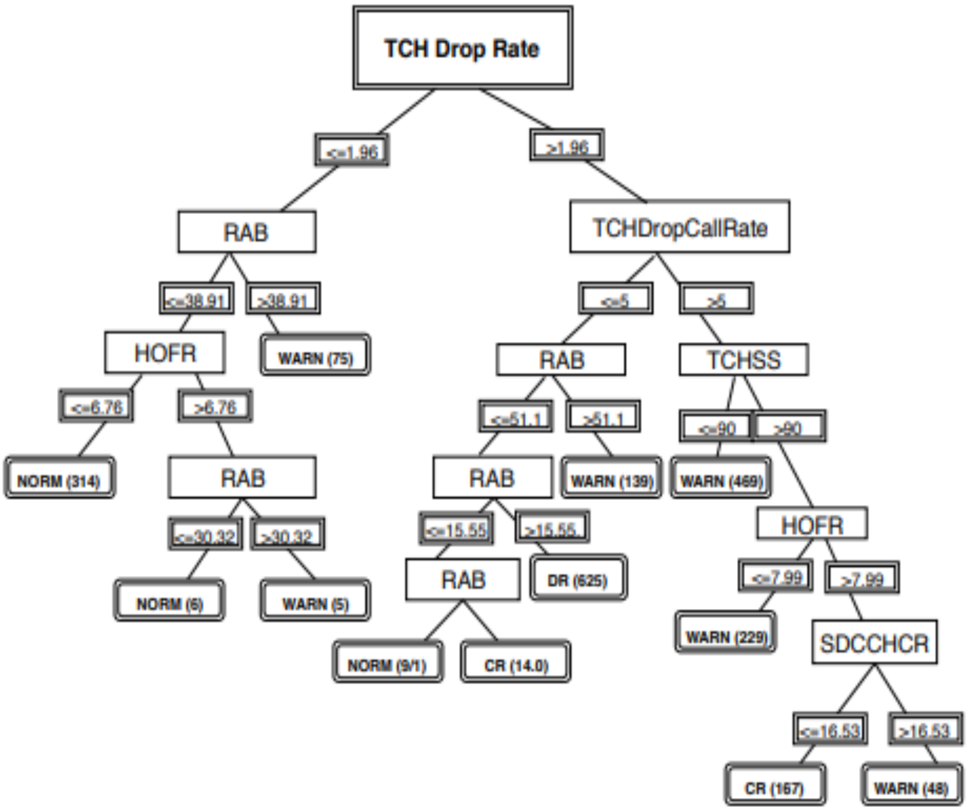


Figure 4.2 Optimisation decision tree

Figure 4.3 shows the optimisation decision trees using J48 Tree models. The number of leaves is 12 for J48 class and the size of the tree is 23. The attribute selection from the J48 classifier is based on the variables related to the utilization of the network such as Traffic Channels and Handover success and Number.

The tree models discovered rules and results of key performance indicators alarm predictions based on those rules respectively. Just as importantly, the decision trees demonstrate the ability to provide a visual representation of the faults that can be understood by a human operator whose job it might be to analyse the data and determine the cause of the lack of network optimisation. By examining the decision trees, it is possible to quickly determine which cells are indicated as being normal (NORM) and which cells are indicated as being critical (CR). By looking at the specific classifiers, the human operator can make a quick decision about the fault in the network to correct that fault and return the network to maximum optimisation.

4.5.4 Fault diagnosis using clustering techniques

The advantage of using decision tree techniques is that the data contained within the tree is easy to read and interpret, and is meaningful to the user. Interpreting data in a decision tree is easier than in other approaches. However, the disadvantage of using decision tree classifiers is that the causes of the optimisation diagnosis are not shown.

As a result, further increases in accuracy might be achieved through the investigation of the decision trees outputs that influence performance of the network. The significance of this part of the thesis is also the use of clustering techniques in order to investigate optimisation of the mobile network. The use of the clustering techniques allows for the use of the multi-objective approach that has been missing in much of the previous research on this topic.

The resulting study will improve upon and enhance the decision tree optimisation rules, while also providing new information and techniques that mobile network operators can use in actual practice. The clustering techniques used in this part of the thesis will contribute to segregate groups of cells and base stations that present similar optimisation faults according to the KPI alarms. The main reason that a clustering framework was used in this part of the thesis is that it is an unsupervised method where similar instances such as network cells are clustered based on main KPI optimisation rules. So, an unsupervised method will boost the optimisation algorithm to provide hidden rules that are not based on the predefined optimisation rules. While decision trees is a supervised method used to define additional optimisation rules based on predefined labels according to KPI alarms. The tools used for running and visualising the decision tree models and the clustering algorithm are Weka (versions 3.8 & 3,9) and Massive Online Analysis (MOA) version 9.0.1.

The proposed clustering algorithm shown below will identify the areas requiring optimisation so that the results can be viewed in detail.

Clustering Algorithm:

Input data:

Training set of KPIs = {(xi, yi,)} n where ,i=cell id which is also the primary key that connects our datasets in our data model shown in table 4.2. Any learning algorithm that define an optimisation scheme based on the efficient KPI metrics according to the decision tree classification.

L to train the decision tree classifier h : X → Y ,

Step 1: Estimate the different metrics that contribute to an optimisation scheme according to the decision tree classification.

Step 2: Define the group of causes C_a, C_b, C_c, C_d, present the KPI values of each network performance areas:

A_i (T_N, D_N, H_N, CS_N...H_N).

- ● Traffic= $\{T_1, \cdots, T_N\}$.,
- ● Dropped Calls = $\{D_1, \cdots, D_N\}$,
- ● Handover Success = $\{H_1, \cdots, H_N\}$,
- ● Call Successful Rate = $\{C_1, \cdots, C_N\}$.

Step 3: Estimate Group of Causes (GC_i) using the rules of the classification tree J48 algorithm, Step 4: Identify the population of cells that used to determine the optimal set of group causes

Step 5: For each KPI value = 1, 2, 3,... ,i that carry out a clustering area and compute Attributes (C_i = 0, 1, 2...i);

Output: Calculate the optimised cells for each cluster partition that shows the clustered network performance areas.

The clustering optimisation approach was based on the results of the decision tree classifiers in which the limits of the different KPIs were identified and the values were used to set up the cause group. Our clustering model identifies the areas of the optimisation issues based on the value of the symptoms and causes of network malfunctions.

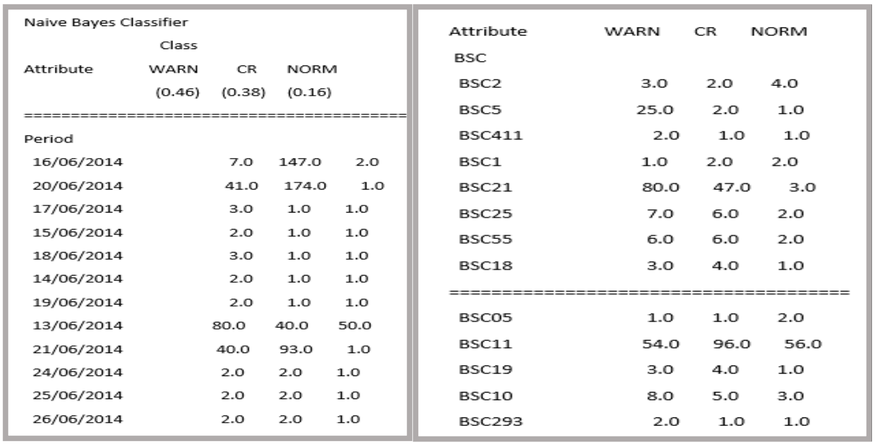
The model can be run with a different number of attributes based on the characteristics of the mobile network and the multidimensional queries that network operators request. In addition, the number of symptoms and the number of cells can also be changed based on the technical areas of concern of the network operators. The number of symptoms and the number of cells can also be changed based on the geographical areas of the network that are being examined, and the level of analysis that is desired by the network operators (Wu, Chiang &Fu 2014).

4.5.5 Fault localisation evaluation based on Bayesian Networks

As has already been noted, network anomaly detection can be classified in different ways, but the most convenient and useful classification for network technicians and administrators is by location. By grouping the faults based on location it is possible for the technicians and administrators to not only know where faults are occurring regularly, but how many issues are occurring in specific locations at given points in time and over specific time intervals (Deljac, Mostak & Stjepanovic, 2010).

Furthermore, grouping the causes of faults s is of interest in this investigation. In this regard, the effort has been made to match all available data that were explicitly collected or that were received based on sensed signals in a way that created a logical cause and effect relationship. Then, the same data sets and relationships were input into Weka in order to run Bayesian network classifiers to provide important information about fault isolation and localisation. This Bayesian model considers the overall fault probability distribution of each alarm in individual base station locations.

TABLE 4.7 FAULT LOCALISATION



```
Naive Bayes Classifier
                  Class
Attribute         WARN     CR      NORM
                  (0.46)   (0.38)  (0.16)
========================================
Period
  16/06/2014           7.0    147.0     2.0
  20/06/2014          41.0    174.0     1.0
  17/06/2014           3.0      1.0     1.0
  15/06/2014           2.0      1.0     1.0
  18/06/2014           3.0      1.0     1.0
  14/06/2014           2.0      1.0     1.0
  19/06/2014           2.0      1.0     1.0
  13/06/2014          80.0     40.0    50.0
  21/06/2014          40.0     93.0     1.0
  24/06/2014           2.0      2.0     1.0
  25/06/2014           2.0      2.0     1.0
  26/06/2014           2.0      2.0     1.0
```

```
Attribute          WARN     CR     NORM
 BSC
  BSC2              3.0     2.0     4.0
  BSC5             25.0     2.0     1.0
  BSC411            2.0     1.0     1.0
  BSC1              1.0     2.0     2.0
  BSC21            80.0    47.0     3.0
  BSC25             7.0     6.0     2.0
  BSC55             6.0     6.0     2.0
  BSC18             3.0     4.0     1.0
=====================================
  BSC05             1.0     1.0     2.0
  BSC11            54.0    96.0    56.0
  BSC19             3.0     4.0     1.0
  BSC10             8.0     5.0     3.0
  BSC293            2.0     1.0     1.0
```

Table 4.7 shows the probability of fault localisation by time and base station controller using Bayesian classifiers. The Naïve Bayes classifier shows the location and probability of the fault alarm distribution by base station location and day. It is also possible to achieve fault detection based on the fault probabilities. However, one-to-one mapping between the nodes is not always possible. In order to carry out simultaneous detection of KPI faults, additional locations have to be used in the model, which requires knowledge of the faults and their impacts on the process variables.

4.6 Fault diagnosis results discussion

The results of the experiments show that an average of 860 out of 2100 cells of the data base were classified in a state alarm of “Warning”, 785 found to be “Critical” and only 320 cells were considered as “Normal”. The most common KPI metrics to form our model basic rules extracted by the system are based on TCH Availability, Congestion and Success Rate, Handover Success and Failure Rate, SDCCH and TCH Drop Rate and final TCH attempts.

TABLE 4.8 PERFORMANCE MEASURE RESULTS

Classifiers	Correctly classified instances	Incorrectly classified instances	Kappa statistics
J48	2099	1	0.99
JRIP	2097	3	0.99
Part	2099	1	0.99
BayesNet	1839	261	0.80
NaiveBayes	1701	399	0.69
Classifiers	**Mean absolute error**	**Root mean squared Error**	**Accuracy**
J48	0.0006	0.01	99.9%
JRIP	0.0019	0.03	99.8%
Part	0.0006	0.01	99.9%
BayesNet	0.08	0.28	87.5%
NaiveBayes	0.13	0.34	81%

Table 4.8 also shows the accuracy alarm prediction by the proposed technique. The table shows that the performance measure results of J48 and PART classifiers have the same performance in terms of accuracy in classification. Both methods achieved an accuracy rate of 99.9%. With the rules and decision tree classifiers, the accuracy rate is actually more than 99.9%. However, the Bayesian classifiers showed a decline in their accuracy levels as compared to the other methods

In order to measure the accuracy of the decision trees and the rules that were generated for the network fault classification results, three measures were used that are widely used with the data set of testing data to evaluate the data mining classifiers, which are precision, recall, and the F1 metrics.

Precision and recall are calculated based on the relative number of true positives (TP), the number of objects that are correctly classified, false positives (FP), the number of classifiers that are not correctly classified, and false negatives (FN), the number of positive objects that are incorrectly classified. Precision (TP/TP+FP) is calculated as the ratio of true positive (TP) to the sum of true positive (TP) and false positive (FP). Recall (TP/(TP+FP)) is the ratio of true positive (TP) to the sum of true positive and false negative (FN). The F1 metric is the harmonic mean of recall and precision (Park et al., 2009).

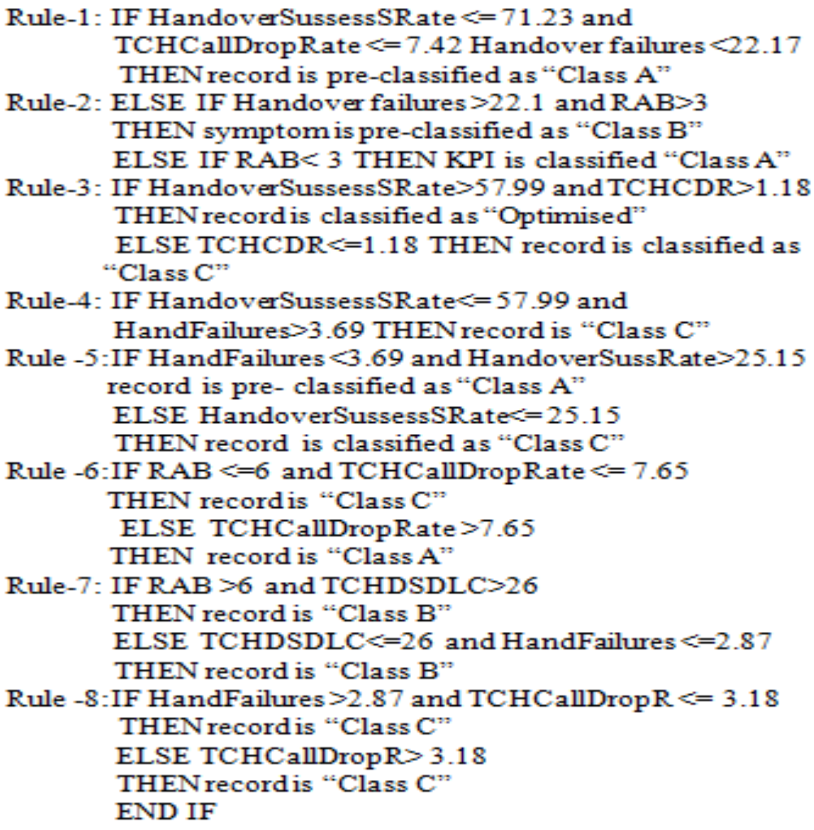
TABLE 4.9 COMPARISON OF FINAL STATISTICS WEIGHTED AVERAGE

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
J48	0.99	0.00	0.99	0.99	0.99	0.99	0.99	0.99
JRIP	0.99	0.00	0.99	0.99	0.99	0.99	0.99	0.99
Part decision list	1.00	0.00	1.00	1.00	1.00	0.99	1.00	1.00
BayesNet	0.87	0.03	0.91	0.87	0.88	0.83	0.99	0.98
NaiveBayes	0.81	0.10	0.84	0.81	0.80	0.70	0.93	0.90

Table 4.9 shows the evaluation of final statistics. The scale is based on the level of accuracy of the algorithm with regards to correctly identifying faults as compared to false results (Shahzad, Asad, and Khan, 2013).

4.7 Experimental results using k-means clustering techniques

Retroactive data of 550 Base Stations (BS) were selected for this experiment. From the 550 BSC that were available, 440 cells were defined as training data and 110 were defined as test data. The data and KPIs shown in table 4.1 were used as an input. Rule sets generated from the decision tree classifier and extracted as an output from the clustering model:



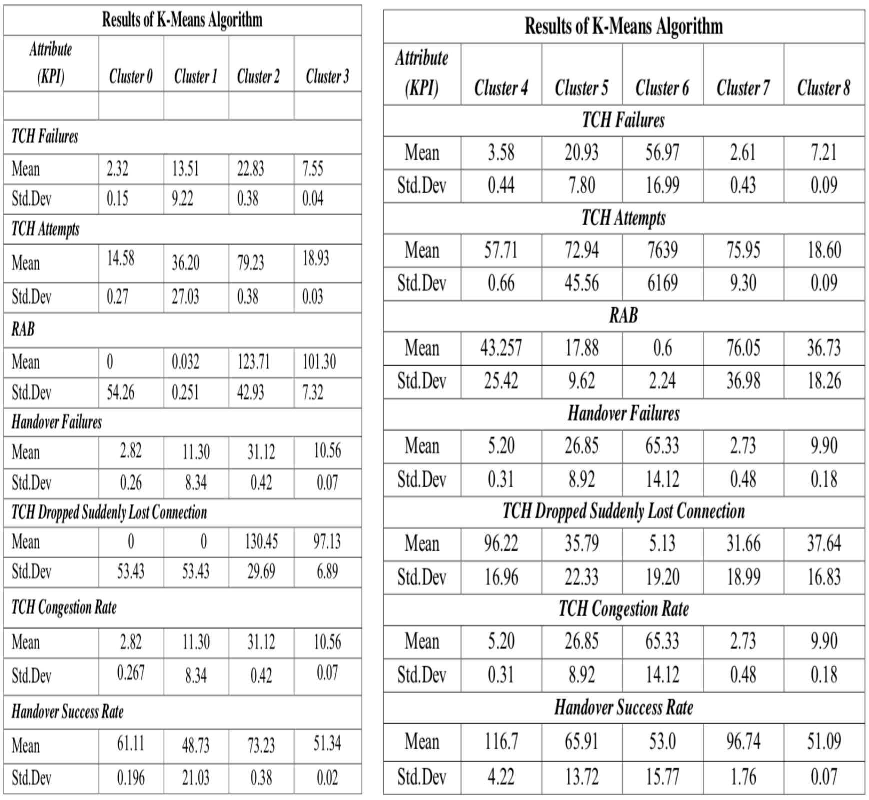
```
Rule-1: IF HandoverSussessSRate <= 71.23 and
        TCHCallDropRate <= 7.42 Handover failures <22.17
        THEN record is pre-classified as “Class A”
Rule-2: ELSE IF Handover failures >22.1 and RAB>3
        THEN symptom is pre-classified as “Class B”
        ELSE IF RAB< 3 THEN KPI is classified “Class A”
Rule-3: IF HandoverSussessSRate>57.99 and TCHCDR>1.18
        THEN record is classified as “Optimised”
        ELSE TCHCDR<=1.18 THEN record is classified as
       “Class C”
Rule-4: IF HandoverSussessSRate<= 57.99 and
        HandFailures>3.69 THEN record is “Class C”
Rule -5:IF HandFailures <3.69 and HandoverSussRate>25.15
       record is pre- classified as “Class A”
        ELSE HandoverSussessSRate<= 25.15
        THEN record is classified as “Class C”
Rule -6:IF RAB <=6 and TCHCallDropRate <= 7.65
       THEN record is “Class C”
        ELSE TCHCallDropRate >7.65
       THEN record is “Class A”
Rule-7: IF RAB >6 and TCHDSDLC>26
        THEN record is “Class B”
        ELSE TCHDSDLC<=26 and HandFailures <=2.87
        THEN record is “Class B”
Rule -8:IF HandFailures >2.87 and TCHCallDropR <= 3.18
        THEN record is “Class C”
        ELSE TCHCallDropR> 3.18
        THEN record is “Class C”
        END IF
```

Based on the results, an accuracy rate of 98.86% of instances were correctly classified using the decision tree model. Only 6 instances were incorrectly classified. The mean absolute error was only 0.0087%.

The results show that the weighted average precision of the algorithm was 98.64%, and that none of the classes had precision rates of less than 97%. Also 93 causes of class A were correctly predicted, 216 out of 220 causes of class B were also predicted correctly and 64 out 65 causes are classified as class C. Only 61 causes of class D were predicted correctly. The causes of the network faults are shown in figure 3.1 of the previous chapter,

The k-means clustering process that was based on the eight rules for identifying the causes of network faults involved introducing a small number of attributes. The reason for this is that with a smaller number of attributes, the results will be more precise and indicative for the technical area being examined. Moreover, adding more variables could provide the opportunity to examine the characteristics of 9 clusters that would represent the different technical areas of concern. If more variables were introduced into the model, the analysis would detail every area and provide a more general perspective of a possible optimisation solution. In such a situation, the combination of KPIs would not always be predictable (Rozaki, 2016).

TABLE 4.10 EXPERIMENTAL RESULTS OF K-MEANS



Results of K-Means Algorithm

Attribute (KPI)	*Cluster 0*	*Cluster 1*	*Cluster 2*	*Cluster 3*
TCH Failures				
Mean	2.32	13.51	22.83	7.55
Std.Dev	0.15	9.22	0.38	0.04
TCH Attempts				
Mean	14.58	36.20	79.23	18.93
Std.Dev	0.27	27.03	0.38	0.03
RAB				
Mean	0	0.032	123.71	101.30
Std.Dev	54.26	0.251	42.93	7.32
Handover Failures				
Mean	2.82	11.30	31.12	10.56
Std.Dev	0.26	8.34	0.42	0.07
TCH Dropped Suddenly Lost Connection				
Mean	0	0	130.45	97.13
Std.Dev	53.43	53.43	29.69	6.89
TCH Congestion Rate				
Mean	2.82	11.30	31.12	10.56
Std.Dev	0.267	8.34	0.42	0.07
Handover Success Rate				
Mean	61.11	48.73	73.23	51.34
Std.Dev	0.196	21.03	0.38	0.02

Results of K-Means Algorithm

Attribute (KPI)	*Cluster 4*	*Cluster 5*	*Cluster 6*	*Cluster 7*	*Cluster 8*
TCH Failures					
Mean	3.58	20.93	56.97	2.61	7.21
Std.Dev	0.44	7.80	16.99	0.43	0.09
TCH Attempts					
Mean	57.71	72.94	7639	75.95	18.60
Std.Dev	0.66	45.56	6169	9.30	0.09
RAB					
Mean	43.257	17.88	0.6	76.05	36.73
Std.Dev	25.42	9.62	2.24	36.98	18.26
Handover Failures					
Mean	5.20	26.85	65.33	2.73	9.90
Std.Dev	0.31	8.92	14.12	0.48	0.18
TCH Dropped Suddenly Lost Connection					
Mean	96.22	35.79	5.13	31.66	37.64
Std.Dev	16.96	22.33	19.20	18.99	16.83
TCH Congestion Rate					
Mean	5.20	26.85	65.33	2.73	9.90
Std.Dev	0.31	8.92	14.12	0.48	0.18
Handover Success Rate					
Mean	116.7	65.91	53.0	96.74	51.09
Std.Dev	4.22	13.72	15.77	1.76	0.07

Table 4.10 shows that in cluster 2, the results indicate unusual channel drops in which connections were suddenly lost. This occurred even though handover metrics and the traffic channel failures had reasonable and acceptable values. However, the cluster results show that the attempts might have been higher than normal, which could have affected the traffic of calls and data at that specific time. In that case, the next step would be to run the model including the traffic rate indicators to find out if the traffic equipment needs to be improved in that particular geographic area. If the results are also predictable, then other variables, such as peak times, would need to be checked that might affect the optimisation plan.

The causes of the problem could also be clustered to show additional evidence as part of the optimisation results. It should be noted that one limitation of the model might be that causes for some of the groups may not be shown within some of the clusters. In this situation, the categories and clusters would have to be extended and further analysed in the following chapters using meta-classification techniques and ensemble learning methods.

4.8 Summary

This chapter has presented an automated methodology that uses a machine learning approach for extracting an optimisation model from mobile data. This part of the thesis proposes a systematic approach for anomaly detection that is based on a KPI data analysis model using the machine learning techniques for those who are responsible for the monitoring and optimisation of GSM networks.

The final statistics show the accuracy of the model and visualise precision and recall (true positive rate vs false positive rate) (Ferri-Ramirez, Flach, & Hernandez-Orallo, 2002). TP rate also demonstrates the sensitivity of the model such as a scale of actual positive values and the FP rate shows the specificity such as the negative tuples that are incorrectly labelled (Gao & Wang 2011). In this regard, the slightly less accurate information of BayesNet and NaiveBayes classifiers across the entire system still results in a high level of ability to determine and localise the fault or faults that need to be corrected in order to return a network to optimisation (Shahzad, Asad, and Khan, 2013).

In the following chapter the aim of the research is to further examine the optimisation issues using cost sensitive classification algorithms to avoid overfitting, minimize optimisation costs and reducing misclassification costs. Then an optimisation framework will be proposed to evaluate cost and network performance.

PART 2

Part 2 includes Chapter 5 which improves the fault diagnosis outputs with the implementation of network cost data coupled with meta-learning classification techniques applied with neural network models. In this part of the thesis a network optimisation framework was proposed to measure cost effectiveness and network troubleshooting considerations using data mining models.

Chapter 5. Network Cost Modelling

Previously, network performance was considered and fault management purely in terms of their technical implications. It is imperative for network administrators and the companies that operate networks to also focus on the issue of cost. Such networks are already very expensive to operate, which is exacerbated if optimisation is not achieved. Network operators and administrators need a way to handle optimisation techniques in a way that reduces network costs and allows mobile data to be handled effectively and efficiently. As shown in the previous chapter, machine learning algorithms are able to detect faults from raw network data with high accuracy. This chapter Introduces cost considerations to identify the benefits arising from detecting optimisation faults and the penalty incurred if the outputs are incorrectly classified. The contribution of this chapter to the rest of the thesis is significant as it analyses the problems with imbalanced learning while suggesting the most appropriate classification models using meta-cost classification algorithms for an efficient optimisation framework. An optimisation framework is considered appropriate if it correctly identifies the class label for different and various instances such as numerical variables, categorical variables or labels. In our cost optimisation plan, the correct fault classification measure must balance the network cost of acquiring additional information against the penalties for incorrect fault classification.

Cost considerations benefit the analysis by calculating and adjusting a probability threshold for a set of KPIs in order to classify them into cost or benefit classes using a method for the comparison of classifier performance that is robust to imprecise class distributions and misclassification costs (Domingos, 1999). Decision trees & Bayesian models are first used to obtain reliable probability estimates of training examples based on the decision tree rules. The cost sensitive classifiers produced by decision tree and Bayesian methods result in robust network troubleshooting performance, including the optimisation of cost metrics. In the second part of this chapter the implementation of a feedforward artificial neural network model has been integrated using Multilayer Perceptron (MP) to search for hidden patterns between the variables of the cost classes. The main reason that the Multilayer Perceptron algorithm was used for the cost sensitive classification is that is a supervised model carrying out thorough propagation while it can also extract results from neutrons that use a non-linear activation function that will show hidden patterns between the optimisation costs and the KPIs.

This process can occur based on the threshold that selects the probability that minimises the misclassification cost rate thus all types of misclassification errors found on the training instances for network fault classes and cost classes. To model network faults, a set of design criteria are used to guide the development of the cost optimisation evaluation process The issues observed are based on the KPIs alarms, such as call drops due to interference, coverage limitations and missing adjacencies (TCH), congestion rate due to TRX hardware faults and handover success due to location issues and coverage limitations.

The cost optimisation in this thesis is concerned with the bandwidth cost of cells, that is the possibly non-monetised cost of providing the required level of service/traffic to a cell. Operators will typically have a target value of bandwidth cost, beyond which providing service becomes inviable. The purpose of the non-monetised expression of values is to include an optimisation fault valuation while also enclosing the values of multiple types of equipment needed for various network malfunctions such as *TRX for traffic issues or antenna replacement for infrastructure costs.

*TRX: Base Transceiver Station: transmits and receives according to the GSM standards, which specify eight TDMA timeslots per radio frequency. A TRX may lose some of this capacity as some information is required to be broadcast to handsets in the area that the BTS serves. BSC Base Station Controller (might control more base stations –sample size) (Base Station Subsystem, 2013).

5.1 Cost optimisation

Cost optimization is a continuous discipline used for cost reduction, while maximizing revenue for network performance and resource optimisation. Sensitivity to both cost and performance are important since customers are generally willing to pay more for the use of a network when performance is good. While there are a number of studies that specifically address cost optimisation within networks, these have not always been fully successful. For example, Xue (2003) studied the issue of minimum-cost QoS in communication networks and associated routing problems. Their multicast analysis optimally balanced minimum-cost and minimum-delay, but their unicast analysis did not result in an optimal trade-off. It should be noted that this analysis predated the data reliability concerns of Chu et al. (2011), since communications networks at that time did not need to take into account mixed voice and data. However, it does demonstrate the difficulty of balancing cost and QoS concerns. The cost optimisation that is the basis for this investigation relates bandwidth costs to the improvement of network performance and overall cost. In other words, the variance in cost with regards to different types of traffic on a network is vital to the larger cost structure of actual service and content delivery (Adler, Sitaraman & Venkataramani 2011).

5.1.1 Cost causation in telecommunication networks

Different types of costs such as hardware and ongoing costs need to be included in cost optimisation including those arising from the backbone, aggregation, equipment and maintenance (Salmelin & Metsälä, 2012). Transportation costs, including backhaul transmissions (or transmissions from peripheral to central nodes in the network) are usually avoidable costs associated with mobile transmissions. However, backhaul costs are relatively low because of the larger capacity and low number of links in the backbone network, while access costs (delivery to and receipt of data from the consumer unit) are much higher because of lower capacity and high numbers of nodes and links.

Thus, costs are highly dependent on the particular configuration of the network, and network planning needs to balance backhaul, aggregation, and access needs as well as QoS and cost needs. In this chapter, the aim is to address this problem using a meta-cost algorithm and identify the needs in additional equipment, software upgrades or number of nodes and base station controllers.

5.1.2 Cost analysis in multiple classes of KPIs

Common costs and directly attributable costs include fixed costs such as maintenance, hardware or end user software and variable costs such as capacity and coverage costs. In an optimisation problem, modifying the fixed costs only shifts the cost curve, that is the graph of the cost of production as a function of total quantity produced. As our purpose is to identify the expected costs and benefits into a higher or lower level, in this analysis variable costs only were considered (Seytnazarov 2010). These correspond to hardware installation and maintenance (such as TRX), improvement of network transmission limitations and increasing radio resources and power budget parameters to assist better cell handovers and improve QoS. Four categories of network cost were defined that are associate each with corresponding KPIs.

5.1.2.1 C1: Cost of infrastructure

Infrastructure costs arise from purchasing and maintaining antennas, towers, tower components, and equipment for data processing and data analysis (Um, Gille, Simon, & Rudelle, 2004). The KPI used in this investigation to evaluate infrastructure cost is the Handover Success Rate indicator (HOSR) explained in the previous chapter. The HOSR is a measure of the interference, either internal or external, which may affect call switching. HSR degradation can occur as a result of several issues such as problems in hardware infrastructure and coverage improvement related with the infrastructure costs of the network (Jagadesh et al. 2011).

The degradation can occur due to blocking channels, as well as due to hardware faults.

Hardware faults such as Base Transceiver Station and TRX transceiver faults and location area code boundaries wrongly planned can be also incorporated in a decreasing HSR, which is a part of Base Station System (BSS) failures. (Seytnazarov 2010, Jagadesh et al. 2011).

5.1.2.2 C2: Cost of traffic

The traffic model described in Chapter 3 was defined by the difference between the total and actually received transmitted calls divided by the total number of transmitted calls. Traffic channel congestion can occur due to hardware faults, as well as increases in the number of subscribers and the subsequent increase in the total traffic demand in the network.

Reducing the Traffic Channel Congestion Rate can be achieved by increasing the number of transceivers and transmission equipment in a network, as well as replacing or repairing faulty hardware. In addition, companies can install temporary equipment as a means of reducing congestion in areas where it is predicted that network traffic will increase, such as near a sports venue during a major sporting event (Haider, Zafrullah. & Islam, 2009). Additional sites are also required to provide increased levels of in-building coverage as mobile users expect to be offered service in all geographical locations (Michaelis, 2011).

5.1.2.3. C3: Cost of transmission

The major performance areas involved in this cost category are accessibility and retainability as defined in Chapter 3, whose costs include regular maintenance and replacement of equipment (Um, Gille, Simon, & Rudelle, 2004). Call Success Rate is impacted by radio interface congestion, a lack of radio resources, an increase in traffic, and faulty hardware. In addition, CSSR can increase when there are transmission limitations. In order to diagnosis an increase in CDR, radio uplink statistics may be monitored, along with path balance statistics, which are indicators and measures of uplink and downlink power on the network. Customer complaints regarding blocked calls should also be reviewed as an indicator of CDR problems. Furthermore, spectrum scanners and drive test reports can also be helpful in finding where CDR problems exist. Once these measures are obtained and examined, improving CDR typically involves replacing faulty hardware (Haider, Zafrullah. & Islam, 2009).

5.1.2.4. C4: Cost of maintenance

Both regular and periodic maintenance is required, such as replacement and upgrade of equipment (Walgampaya & Kantardzic 2006). The estimation of the cost of maintaining equipment and operations is based on the evaluation of KPI alarms described in Chapter 4. The cells that have been classified as “NORM” in Chapter 4 are already optimised but still have a cost regarding the salaries, operation and maintenance of equipment.

5.2 Design cost and performance optimisation algorithm

A cost optimisation framework has been developed as a learner to train our data model to predict probabilities and optimised KPI alarms values. Algorithmic rules have been developed using our input variables (show in tables 5.1& 5.2) to estimate each of the four types of optimisation cost. Each of these costs are associated with the KPI alarms and the causes of network faults previously performed for our technical optimisation and clustering analysis.

Our algorithm employs a tree structure to extract features for the KPIs inputs of CSSR, DCR, HOF and TR. The selected KPIs are the dependent variables that are included in our data model shown in table 5.1. The rules based on KPI values will support the classification of the cost classes. Our datasets contain raw data and derived variables from our previous experiments.

TABLE 5.1 NETWORK DATA MODEL

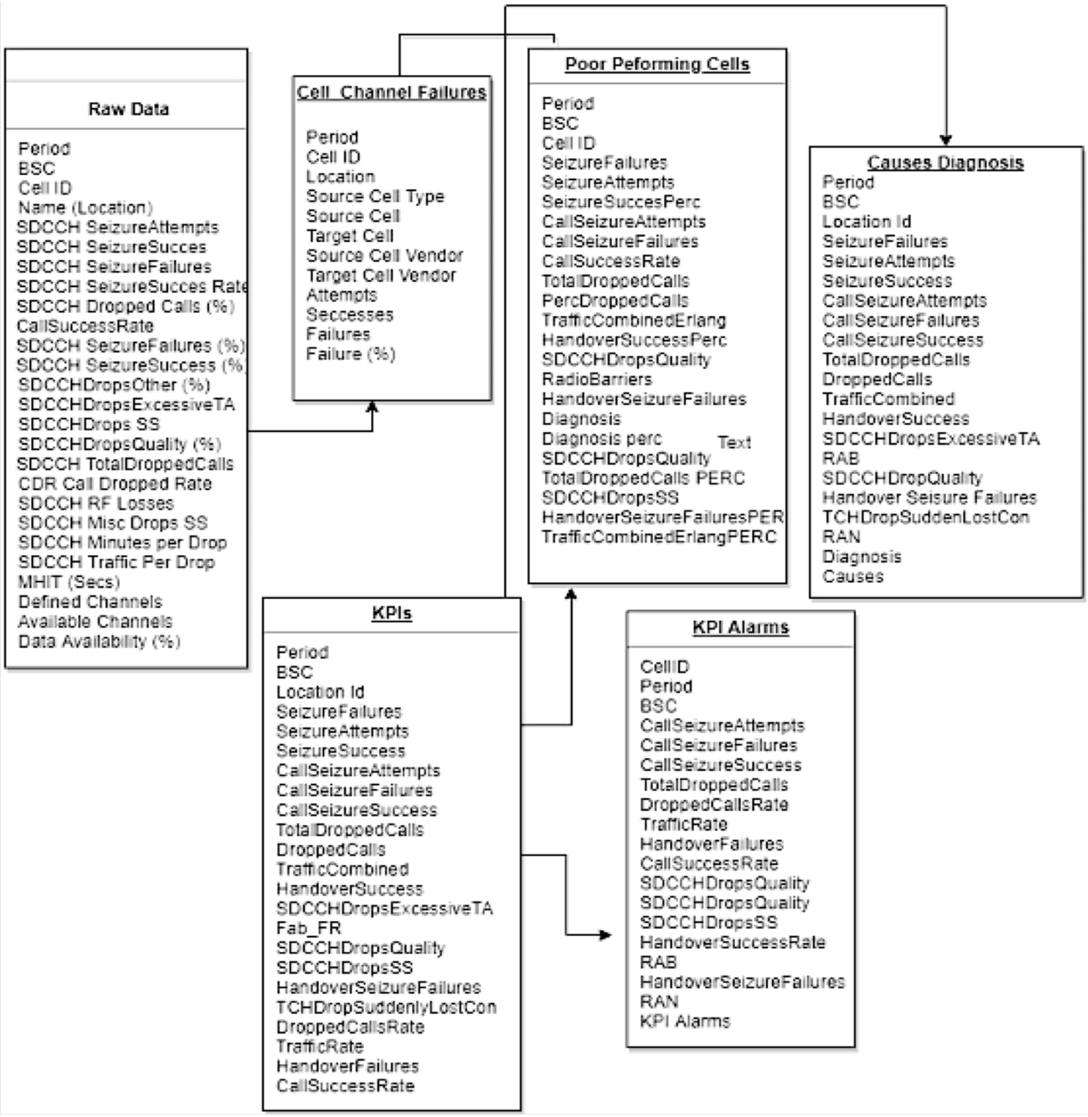


TABLE 5.2 INPUT TRAINING DATA

Input Variables (X)	**Key Performance Indicators (X)**	**KPI Alarms (X)**	**Class Labels (Y)**
Period (Data/Time)	CSSR (Call Success Rate)	NORM	Maintenance Cost
BSC Base Station Controller Location	TM (Traffic Model)	CRITICAL	Infrastructure Cost
Standalone dedicated control channel	HOFR (Handover Failure Rate)	WARN	Transmission Cost
SDCCHAvailabilityRate			
SDCCHCR			
SDCCHAccessRate			
SDCCHDropsExcessiveTA			
SDCCHDropSuddLostCon			
Traffic Channel	CDR (Call Dropped rate)		Traffic Cost
TCHTR			
TCH Seizure Attempts			
TCHCR			
TCHSS			
TCH Drop			
TCH Drop Rate			
TCH Availability			
Handover			
HOFR (Handover Failure Rate)			
HOSR (Handover Success Rate)			

Cost optimisation algorithm:

Aim: evaluate the network optimisation costs and predict the network areas that emerge to be optimised.

1. Purpose: Generate a cost classification based on cost evaluations of KPI metrics and KPI alarm values given by the optimisation algorithm in 4.5.1

   Input data: The input data used for this experiment are shown in table 5.1. Training data set with 2100 cell data referring to 150 locations in the area of Dublin in Ireland.

   Dependent variable: optimisation cost split into infrastructure costs, traffic cost, transmission costs and maintenance cost, a multi-class variable which is defined based on the KPI metrics and fault alarms values. Selected KPIs that used as an input of the cost optimisation algorithm referring to a certain amount of base stations (i=n) are Dropped Call Rate, Call Success Rate, Traffic Rate and Handover Failures {(DCR_i, $CSSR_i$, TR_i, $HOF_{i,}$)}.

Rule 1: Transmission costs

The cells of several base station locations that are chosen to be optimised accordingly to the transmission cost rules are the ones that they present a high Dropped Call Rate and simultaneously the Call Success Rate is lower than usual even if the traffic rate is very low. In conclusion, evidence that show that in a not so busy network (low traffic rate) the dropped call rate is unusually high and the call success rate is lower than expected, transmission equipment need to be integrated. Hence, the rules defining the transmission costs for the model are as follows:

If DCR>2 and CSSR<=95% while TR<70% then

classify the cells in the area location as → "Transmission Cost"

Rule 2: Traffic cost

When the Traffic network rate is relatively high, the call success rate is low and the Handover Failures are greater than usual then the traffic hardware equipment may be checked for further upgrades. The rules defining the transmission costs are as follows:

if TR>70% and CSS<90%
while HOF>=10% then
classify the cells in the area location (i) as → "Traffic Cost"

Rule 3: Infrastructure Cost

According to our previous experiment an obvious evidence of infrastructure issues are the KPI alarms that are resulting as "WARN". So the cell areas that present a high Handover Failures rate coupled with a "Warn" status given by the fault optimization algorithm have to revise the hardware configuration. The rules defining the Infrastructure costs are shown below:

```
else if KPI Alarm = "WARN" given by the fault optimisation algorithm
while HOF >=10%
then classify the cells in the area location () as → "Infrastructure Cost"
```

Rule 4: Maintenance Cost

Maintenance actions are required for all the cells that are categorsed as "NORM" from the fault optimisation algorithm. These costs include network upgrades, software configurations and personnel costs. The rules for estimating the maintenance costs are as follows:

```
do if KPI Alarm = "Norm" class given by the fault optimisation algorithm
then classify the cells in the area location () as → "Maintenance Cost"
```

Output 1:

Applying the C45 algorithm by removing the input attributes of your model to the cost classes to construct new decision tree rules using the subset of training cell data for the selected locations. The purpose of the decision tree algorithm is integrating cost and error rate in decision tree pruning to find hidden relations between the data and KPIs that are not part of the input variables.

Output 2:

A cost sensitive classification model based on Naïve Bayes network algorithm to show the cost probability of the attributes of data and KPIs of the training set. Naive Bayes classification has mainly examined how to reduce the misclassification costs by reviewing different classification risks based on our algorithm rules (Chai et al. 2004).

For classification, the Naive Bayes algorithm computes the posterior probability of the data and KPI alarms belonging to the cost class defined according to Bayes' rule. When missing values exist in a sample with the cells in different locations, the corresponding (output) attributes are simply left out in likelihood computation and the posterior cost probability is computed only based on the known (input) attributes (Chai et al. 2004). It is important to consider that both the output results are referring to the attributes of the data set that are not included in the selected input variables.

The algorithm employs a tree structure to extract particular features for particular inputs that refer to it as the Cost-Sensitive Tree of Classifiers (CSTC). Initially, the foundational concepts regarding the CSTC tree has been introduced that justify the derivation of a global cost term that extends to trees of classifiers. Next, the resulting loss function drive the model into a well-behaved optimisation problem (Xu, et al. 2014). The optimisation algorithm has also the ability to optimise the cost parameters based on the posterior probabilities that are determined by the neural network for the purpose of performance maximization.

5.3 Estimating alarm costs

A *cost matrix* (also known as loss matrix) expresses target values (class labels) that indicate the costs for correct and incorrect classifications. A cost matrix is a mechanism for influencing the decision making of a meta cost classification model. In cost-sensitive learning, it is usually assumed that the cost matrix automatically adjusts the target values in order to improve the accuracy of the classifiers. A cost matrix could bias the model to avoid this type of error. In this experiment it was used a meta-cost sensitive classification together with a Bayesian classifier based on the specified classification algorithm, training data, and a cost matrix.

The cost sensitive learning classification tries to avoid classification errors with a high error weight, in our case exceeded values in the KPI alarm metrics that may be costly for the network operators. To evaluate the cost approach, a study of the meta cost classification algorithm using a cost matrix was performed to estimate the cost of the false alarms. Given the cost matrix, the variables and attributes of the data set should be classified into classes with a minimum expected cost estimated for each type of network cost. The expected cost of classifying a cell into a class can be expressed as a 4-class problem with the associated cost matrix given in Table 5.3. The values of the matrix present approximately the costs of an individual cell. Thus, the number of errors increases while the cost of the errors decreases in comparison with the same classification without a cost matrix.

TABLE 5.3. NETWORK COST MATRIX

Cost Classification	*Predicted Class*	*Actual Class*			
Traffic Costs		-2.5	4	4.5	2
Transmission Costs		2.1	-3.5	5	0
Infrastructure Costs		1.2	1.3	-4	1
Maintenance Costs		0.9	1.1	2.1	-3

The matrix presents higher predicted values in terms of transmission costs in comparison to the matrix that shows that the infrastructure costs are higher. The diagonal of the network cost in the matrix show the benefits or alternatively the costs for correct classification, all of which are negative values. This is normal because an appropriate classification has benefits rather than costs.

5.4 Cost sensitive classification results using Bayes & J48

Meta cost classification models are available to run with the variety of classifiers. In our experiments Bayesian and J48 classifiers were applied to network performance data including calls, traffic channels, handovers etc. and KPIs, and a cost matrix. The main reason that Bayesian modeling was used in this part of the thesis is to measure prior distribution and the likelihood that network faults can occur based on KPI alarm values. Using Bayesian modeling will help to estimate the parameters of fault diagnosis in a particular network area while using J48 will extract additional rules based on the network cost matrix. The tool used for running meta –cost classification models with Bayesian modelling and J48 is Weka (Version 3.4.3)

The selected attributes of the KPIs of CSSR, DCR, TR and HOF are the dependent variables of our model. Our data were divided into training and testing sets. The training data were used for parameter estimation, while the test set was used for evaluation of the methodology. From 2100 instances in total, 1300 instances were used for training and 600 were used for testing.

In experiments, a low cell dropped rate of between 2 and 4 was chosen, along with a low call success rate of between 95% and 98% in order to achieve an optimised cost related to transmission. Traffic rate provides a clear measure of the conditions related to traffic costs on the network. The cells that contain KPIs faults with status "Warn" are classified as infrastructure costs. It should also be noted that for the cells in which KPI faults were defined as Norm, the algorithm would provide an estimation for the maintenance and operation costs. The purpose of the algorithm is to categorise an output based on 24 attributes of data and KPIs that present the independent variables of our model into the four different cost classes and evaluate the relations between the KPI alarms and the network costs considerations to assist financial predictions.

The outputs on the posteriori probability of the different classes are shown in the following tables:

TABLE 5.4 META COST SENSITIVE CLASSIFIERS NAÏVE BAYES RESULTS

Naïve Bayes Results		
Correctly Classified Instances	1999	95.19%
Incorrectly Classified Instances	101	4.80%
Kappa statistic	0.9307	
Total Cost	-5609	
Average Cost	-2.67	
Mean absolute error	0.27	
Root mean squared error	0.14	
Relative absolute error	7.90%	
Root relative squared error	35.00%	
Coverage of cases (0.95 level)	97.47%	
Mean rel. region size (0.95 level)	26.51%	
Total Number of Instances	2100	

TABLE 5.5 ACCURACY BY CLASS

	PT Rate	FT Rate	Precision	F-Measure	MCC	ROC Area	PRC Area	Class
	0.914	0.003	0.997	0.914	0.920	0.999	0.999	Infrastructure Cost
	0.996	0.003	0.996	0.996	0.987	1.000	1.000	Transmission Costs
	0.994	0.052	0.778	0.994	0.855	0.996	0.996	Maintenance Cost
	0.972	0.000	1.000	0.972	0.981	1.000	1.000	Traffic Cost
Weight Avg	0.952	0.010	0.961	0.954	0.934	1.000	0.999	

The Bayes net model considers the overall rating of each KPI as a weight of the probability to be included in a defined cost network class. The accuracy rate of the Bayes Network results was 95.2%, with a total cost of -5609 and an average cost per cell of -2.674. It can be seen that the Bayes net model show lower accuracy than the decision trees. Thus, there is a need to consider that Bayesian classifiers even if they evaluate the posterior probability distribution over a certain number of variables, the parameters given by the models are not highly precise.as they are based on approximately realistic estimates. The transmission and traffic costs show a higher rate of transmission and precision with regards to misclassified network cells.

TABLE 5.6 META COST SENSITIVE CLASSIFIERS J48 RESULTS

J48 Classifier Results		
Correctly Classified Instances	2100	100%
Incorrectly Classified Instances	0	0%
Kappa statistic	1	
Total Cost	-6278	
Average Cost	-2.9895	
Mean absolute error	0	
Root mean squared error	0	
Relative absolute error	0%	
Root relative squared error	0%	
Coverage of cases (0.95 level)	0%	
Mean rel. region size (0.95 level)	25%	
Total Number of Instances	2100	

The decision tree algorithm extracted the rules for cost classifications. The cost optimisation rules used as an input train the decision tree algorithm to extract the rules of the independent variables of our model. Table 5.6 shows the meta cost sensitive classifiers J48 results, which shown an accuracy rate of 100%, with a total cost of -6278 and an average optimisation cost of -2.9895 per cell.

Overall, what is shown is that the algorithm that was tested in this investigation reduces the misclassification cost.

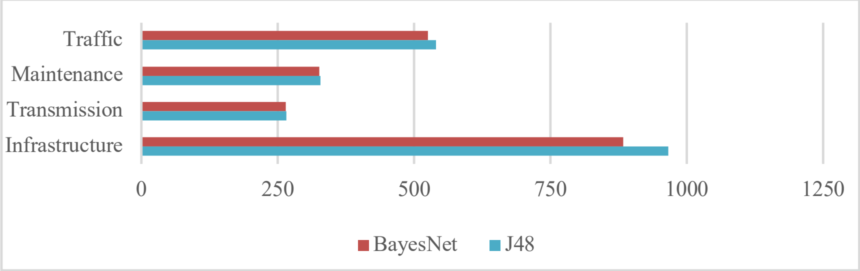


Figure 5.1 Bayes & J48 confusion metrics

Figure 5.1 shows the comparison of the confusion metrics of cost insensitive Bayes classification and J48 generated by Weka. The graph shows that the Traffic Costs and Infrastructure Costs in meta-cost classifier present a low misallocation cost of the cells in Bayesian networks. It can be clearly seen from the results that the cost sensitive model improves the accuracy of the results of our classification algorithms.

5.5 Artificial Neural Networks for Optimising Decisions

Artificial Neural Networks (ANN) are comprised of simple elements that operate in parallel to each other. The result is a multilayer perceptron (MLP) ANN that is fully connected with two hidden layers. The input layer was put into a high dimensional first hidden layer that allowed for effective features to be chosen. Then, for the output activation function, a nonlinear hyperbolic tangent function was used along with other activation functions in order to achieve the best results.

The overall result was a nonlinear transformation in the network (Iliya et al., 2015). Each neuron in an MLP takes the weighted sum of its input values. As a result, each input value is multiplied by a coefficient, all the outcomes are summed up. Each layer in a multi-layer neural network can be seen as a representation of the input obtained through a learned transformation.

The network architecture of neural networks can be identified as either *feed forward*, when the network has links that extend in only one direction, or *recurrent*, where connections between nodes form a directed cycle. A multilayer perception (MLP) network is one in which the output of one layer is connected to the input of the next layer through a synaptic weight. The recurrent type can have at least one feedback connection between the neurons that are in the same layer or in other layers with regards to the specific architecture. The training time of the feed forward has a lower comparison to the recurrent type. However, the recurrent type has better memory for recalling previous events (Iliya et al., 2015). In this experiment, the MLP classifier was used to identify the hidden relations and pattern recognition between data and KPIs in relation to optimisation costs. The main aim is to train the network for the four- class model. More specifically the input variables of the network data and KPIs will be classified into four network cost classes such as traffic, maintenance, infrastructure and transmission costs.

The network could be also monitored and modified during the training. The nodes in the network were sigmoid. The backpropagation neural network is a network comprised of simple processing elements that work together to produce some complex output.

The backpropagation algorithm that will be discussed in chapter 7 of this thesis performs the learning on a multilayer feed-forward neural network. The algorithm learns a set of weights for a prediction of the class label of costs. A multilayer feed-forward neural network is comprised of an input layer, at least one hidden layer, and an output layer (Arora and Suman 2012).

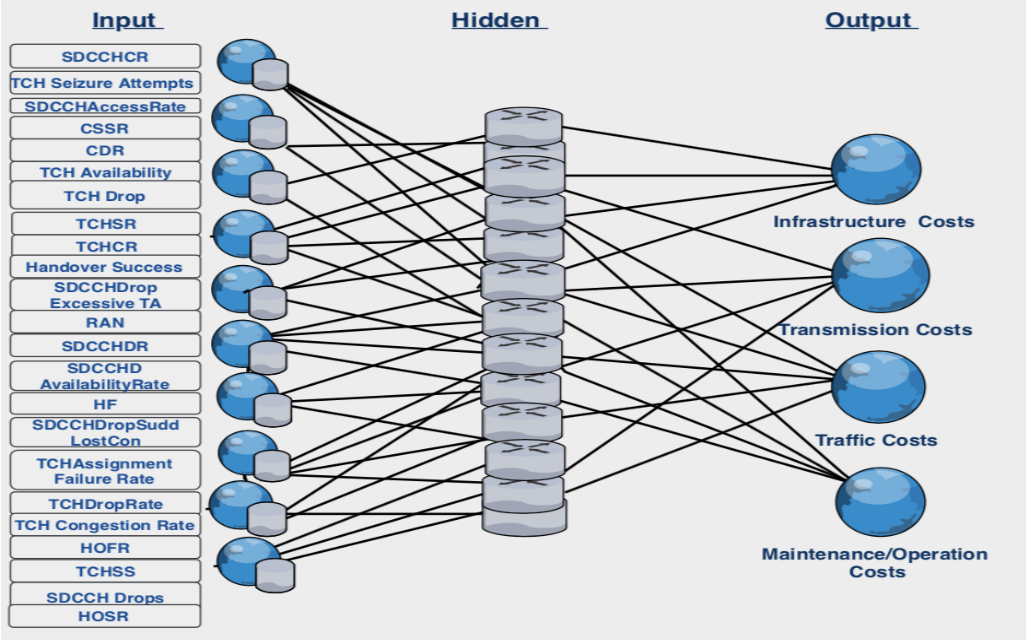


Figure 5.2 MLP Outputs for the final Cost Optimisation Model

Figure 5.2 shows the multilayer feed-forward neural network for the network costs that were created. The four cost classes of infrastructure, maintenance, transmission and traffic that present the neural networks nodes can be structured in different configurations. Weight values are determined by the iterative flow of training data through the network. For instance, in our experiment weight values are established during the training phase in which the network learns how to identify the predefined cost classes by their input of mobile data and KPIs characteristics

Initially, two different sets of KPIs such as Traffic channel rates, RAN and KPI alarms were selected in order to identify the hidden values of the technical areas. Then, the Standalone Dedicated Control Channel Congestion Rate (SDCCHCR) variations, traffic channels dropped rates (TCHDR), and Handover success (HOSR) variables were added to the model. The goal was to connect the categories of the network costs with the input layers that were related to the KPIs for network performance. Based on the results explained in the following paragraph, the four NN cost models were able to adapt to the variations of the optimisation costs (Iliya et al., 2015).

5.5.1 Output of ANN Sigmoid Nodes

The neural network model uses the 26 attributes of mobile data and KPIs to predict the weights for each variable associated with the cost classification. The sigmoid nodes are the nodes used in backpropagation and the associated cost optimisation data. The nodes in the hidden layers of the network are all sigmoid but the output nodes are linear units. The input layers were placed in the hidden layers in order to achieve appropriate selection of cost features. There are 13 hidden nodes labeled in 4 Sigmoid Nodes related to the network costs. The weight values, known also as biases, are given for each variable that feeds into each sigmoid node.

Weight values are determined by the iterative flow of training data through the network. For instance, in our experiment weight values are established during the training phase in which the network learns how to identify the predefined cost classes by their input of mobile data and KPIs characteristics

The 13 hidden nodes pass an output value to the sigmoid nodes shown in figure 5.4 which has a feed, or weight, from each of the 13 hidden neurons. According to our algorithm, the effect of nodes increases linearly from 7.36 to -4.6. A negative weight from input to hidden node means that this value is the one that our cost optimization algorithm found most suitable to satisfy the assigned cost categories. From the results, it is observed that maintenance and operation costs are most closely related with the handover rates, the KPI alarms defined as “NORM” and “WARN”, traffic channel availability and congestion rate, and finally with the call success rate.

In addition, sigmoid node 3 that presents the traffic cost is only related with the traffic channels congestion rate, the KPI alarms defined as “NORM” and the radio access network indicators. Infrastructure costs found to be more likely “Critical” KPI alarms connected with the radio access indicators. Finally, the Multilayer Perceptron Classifier results show that the costs of maintenance and operation are related to the SDCHAccessRate, TCHSeizureAttempts and the call success rate indicators. The proposed approach aimed at implementing the selection of relations given by the hidden neurons in Multilayer perceptron classification network for optimising the financial predictions of the given network areas.

5.6 Results, discussion and interpretation of findings

Results from the experiment with neural network classifiers show high accuracy with a relatively low error rate, as seen in tables 5.7 and 5.8. The results that MLP achieved 99% accuracy with only 2 incorrectly classified instances. Furthermore, the percentage measures that compare the true values and their estimates such as Mean Absolute Error, Relative Absolute Error and Root Relative Squared Error are very low. It can be seen in table 5.8 that the error estimates are low, hence that verifies the accuracy of the cost optimisation results. Moreover, kappa coefficient (κ) that measures the inter-rater reliability of the cost optimisation model is also only 0.9%. A low score for K statistic output shows that the cost optimisation model is reliable and there is a low possibility of the classification outputs occurring by chance. Table 5.9 shows that the outputs of recall, precision and F-measure performance measures are higher for the infrastructure and maintenance/operation classes.

The level of accuracy and generalization of ANN is primarily based on the initial weights and biases, learning rate, moment constant, training data and the network architecture.

TABLE 5.7 MPL CLASSIFIER RESULTS

Correctly Classified Instances	2098	99.90 %
Incorrectly Classified Instances	2	0.09%
Kappa statistic	0.9986	
Mean Absolute Error	0.0021	
Relative Absolute Error	0.6102%	
Root Relative Squared Error	5.2574%	
Total Number of Instances	2100	

TABLE 5.8 MLP RESULTS OF THE COST MATRIX

	PT Rate	FT Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Infrastructure Cost
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Transmission Costs
	1.000	0.001	0.996	1.000	0.998	0.998	1.000	1.000	Maintenance Cost
	0.992	0.000	1.000	0.992	0.996	0.996	1.000	1.000	Traffic Cost
Weight Avg	0.999	0.000	0.999	0.999	0.999	0.999	1.000	1.000	

In order to tackle the imbalance in misclassification errors and minimise the risk of the optimisation costs it was incorporated the cost sensitive classification method into various algorithms. In this experiment, the learning rate and the momentum were maintained at 100% accuracy for J48, 99.95% for MPL, and 95.2% for Bayesian network classifiers.

Three types of cost sensitive classification algorithms (decision trees, naive bayes networks and neural networks) have been trained using the class labels provided by the cost matrix in order to predict the causes of faults. The cost sensitive classification method assigns different weights given by a cost matrix for different network cost class instances. This effectively induces the classifiers to avoid high misallocation costs.

TABLE 5.9 EXPERIMENTAL RESULTS ON COST-SENSITIVE METHODS

Results	BAYES NET	J48	MPL
Correctly Classified Instances	95.2%	100%	99.9%
Kappa Statistics	0.93	1	0.99
Total Cost	-5609	-6278	-6274
Average Cost	-2.671	-2.989	-2.988
Mean absolute error	0.027	0	0.0019
Root mean squared error	0.1445	0	0.013
Relative absolute error	7.91%	0	0.55%
Root relative squared error	35%	0%	3.17%

Table 5.9 compares the results of all three types of classification models considered in this chapter. The experimental results show that the cost-sensitive Bayesian network classifiers have obvious advantages compared to the neural network and decision trees classifiers in terms of the total misclassification costs and the number of high cost errors. Even though the Bayesian classifiers showed the lower accuracy, it looks that the cells that they eventually classified present a lower risk for the cost optimisation plan. The classes of costs that present the parameters given by the operators show the higher misallocation errors are the infrastructure costs and the maintenance costs. It should be noted that the impact of the misallocation costs in infrastructure may be higher than the one of the maintenance costs. The Traffic Costs and the Infrastructure Costs present a lower level of misallocation cost in all the three classification models.

Figure 5.3 shows the output weight for the KPIs from the hidden units for the multilayer perceptron classifier model, as well as the variations in the output weights of the KPIs. Operators can update the output weights so that the hidden units can change cost classes based on network performance.

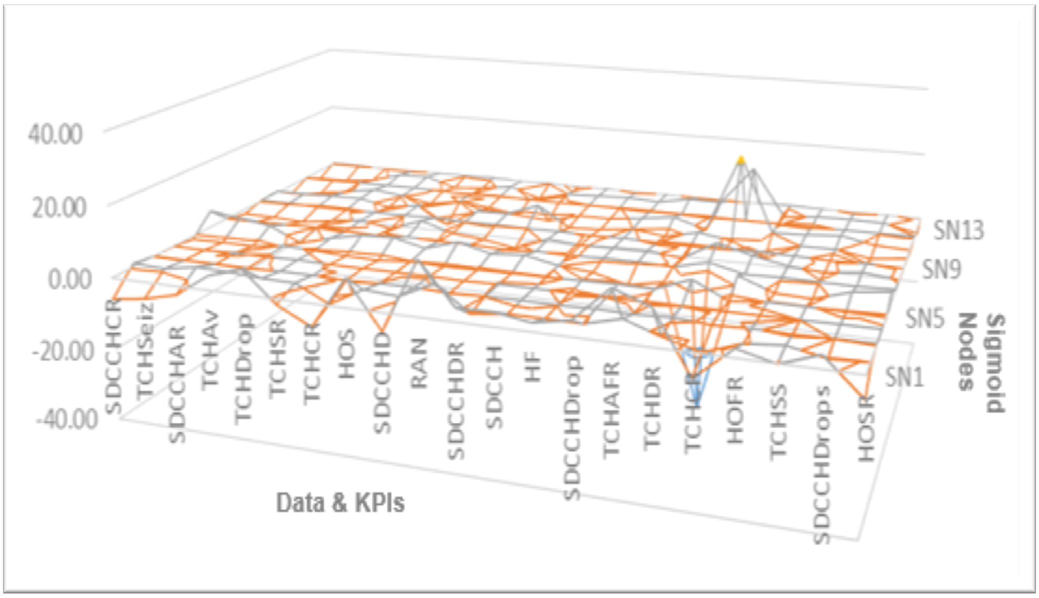


Figure 5.3 Multiple KPIs Output Weights

The hidden layer in the model acted as a feature detector so that during the training, the KPIs features were learned. The KPI values fluctuated based on the selected costs defined by the cost matrix while the operator would be able to localise the optimisation costs following the location of the cells. In situations in which the abnormal values were shown in the hidden values, they were checked. For example, the Handover Success Rate (HOSR) and the Traffic Channel Congestion Rate values had to be checked in order to ensure accurate cost optimisation.

The results also show that the cost-sensitive neural network classifier makes fewer cost errors than the original cost-sensitive classifiers.

The classes of costs that present the parameters given by the operators show the higher misallocation errors are the infrastructure costs and the maintenance costs. It should be noted that the impact of the misallocation costs in infrastructure may be higher than the one of the maintenance cost. The Traffic Costs and the Infrastructure Costs present a lower level of misallocation cost in all the three classification models

5.7 Summary

In the second part of the thesis the research experiments demonstrated the need of a cost optimisation framework to evaluate efficiency in network fault detection and budgeting and also ameliorate the overall quality of network services. Meta cost learning techniques contribute to a cost-effective optimisation framework that will anticipate higher accuracy in the following parts of this research.

This part of the thesis has been published by the International Journal of Advanced Engineering Forum in 2016. The paper was recently cited by a conference paper written by K. J. Lee. named as “Online Class Imbalance Learning for Quality Estimation in Manufacturing” for the cost-sensitive neural network model proposed. Publication can be found in the proceedings of IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA))2018.

In the following chapter the main aim will be focused on customer segmentation using data mining models. The scope of the third part of the thesis is to use the cost optimisation framework to evaluate mobile deals while estimating mobile customers' wiliness to pay using real customer data.

PART 3

In Part 3, chapter 6 provides an overview of customer segmentation in the telecommunication industry using data mining models. This part of the thesis proposes a framework for telecommunication companies to target deals and special programmes or incentives in order to prevent customers from churning and switching to another provider. Next, Chapter 7 proposes an end to end super-framework to identify the most important variables of our datasets and prioritise end improve optimisation decisions using ensemble learning algorithms.

Chapter 6. Customer Profile and Revenue Management

According to statistics presented by Zan M. et al, 2007, the annual customer churn rate is 30%, and over each month data contains 2.5% churn (real) customers and 97.5% online (virtual) customers. A customer segmentation framework proposed in this part of the thesis will facilitate the promotion of services and features that match with customer usage. It will also offer the appropriate details considering customer relationship strategies and improve technical and financial performance with enhanced targeted marketing. The use of machine learning techniques that was proposed for the mobile customer segmentation scheme allows mobile providers to dynamically distinguish their customers' preferences based on various criteria, including customer location, gender, spending behaviour, financial status and several telecommunication deals.

6.1 Customer segmentation and pricing strategy in telecommunication industry

The process of market segmentation can play an important role in the success of every business in identifying the demographic, socioeconomic, and geographic characteristics of customers. This is particularly the case in telecommunications, where many new digital services, hardware devices and technologies are being introduced making the market more competitive than before. To increase revenue in the telecommunication industry, it is important to have a clear understanding of the target customer segments. An efficient approach to segmenting and targeting should take account of comprehensive demographics and consumer expenditure data, associated with the actual buying behaviour, and customer's willingness to pay. (Noughabi, Albadvi and Far 2015). The competition among major cell phone operators raises several questions of revenue management, optimality of queuing systems and supply chain management. Wei, Xu, & Hu, 2014 discussed the concavities of the revenue function with demand and price respectively.

Zhongwen et al. 2007 focused their analysis on deals from a single service provider. According to the authors, the telecom industry includes a variety of bundled products, product mix, large user groups, and customer level branches of all their advantages.

All of these factors can seriously affect the income of network operators. Zhongwen et al. 2007 proposed a closed loop management system which included market prediction, calculation of yield model parameters, pricing model design, development of pricing strategies, actual receipts assessment and optimizing of yield model parameters.

Sladojevic, Culibrk, & Crnojevic, (2011) discussed transition of consumers in their research. The telecommunication industry is one of the industries in which a consumer can rapidly and easily move to competitors. Sladojevic, Culibrk, & Crnojevic, (2011) discussed churn and an effort was made to predict the customer's likely to move using data mining methods.

Esteves & Cerqueira (2017) discussed that successful pricing strategies should be based on an efficient customer profiling and billing system to be able to evaluate a wide range of telecommunication services and service bundles for new and existing customers, as well as provide consolidated customer data for price discrimination by purchase history. It also had to be adjustable enough to validate new offerings and billing models based on customer's needs and willingness to pay.

In this part of the thesis the aim is to apply machine learning methods to provide an automated scheme to address the two major business models in revenue management and churn prediction faced by telecommunication companies. In particular, a detailed analysis of the C.5 algorithm within Naive Bayesian modelling was provided for the task of profiling telecommunication customer behaviours according to their billing and socio-demographic features focusing on improving our ability to predict churn.

Due to the amount of data used for this part of the thesis C.5 algorithm was used as it takes care of many decisions automatically without overfitting the trees while using fairly reasonable defaults. It is also easy to interpret and identify the most relevant rules for the automated scheme in comparison to the rest of the decision tree algorithms. Therefore, the Naïve Bayesian model contributes to predict the probability of mobile deals to be purchased. The classifiers were run using Weka tool (version 3.8 & 3.9).

To demonstrate the approach, the results have been experimentally implemented using real data from the iD Mobile Ireland company, a start-up telecommunications provider in the Republic of Ireland. The company differentiates itself in the competitive Irish market by separating the mobile tariff from the purchase of the handset. This allows customers the flexibility to enter or leave a 12, 18 or 24-month contract without penalty, and purchase a new handset every three months, should they wish to do so, provided the previous handset cost is fully paid off. As the customer is not tied down into an extended contract where the cost of the phone is subsidised by the tariff price, customers may change their tariff call, text and data allowances every month, to suit their individual needs, allowing them more control over their account charges (Dullaghan & Rozaki 2017).

The customer profile framework provides important billing information about the customers transactions and their financial status which may be contributable to a fraud detection scheme that will monitor and track fraudulent activities while improving services for VIP customers.

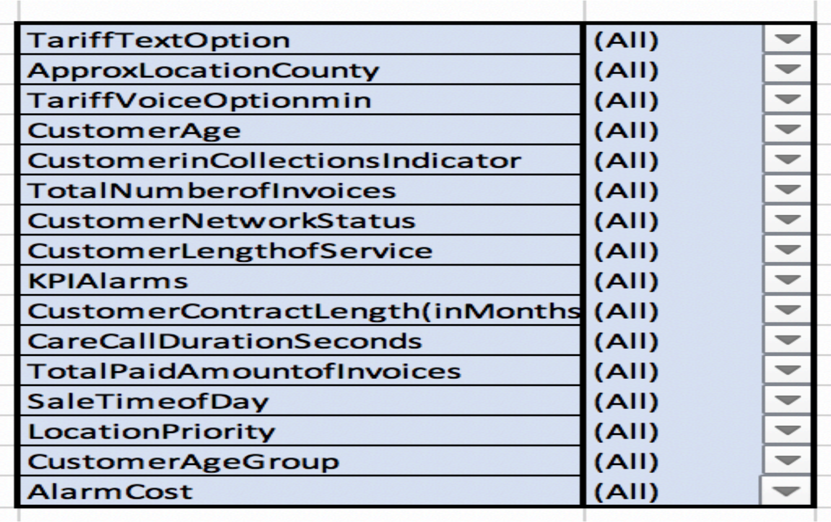
Consequently, the main aim is to create a framework to predict different customer preferences on deals and the probability of a customer switching to a different telecommunication provider.

6.2 Ground truth data for the segmentation model

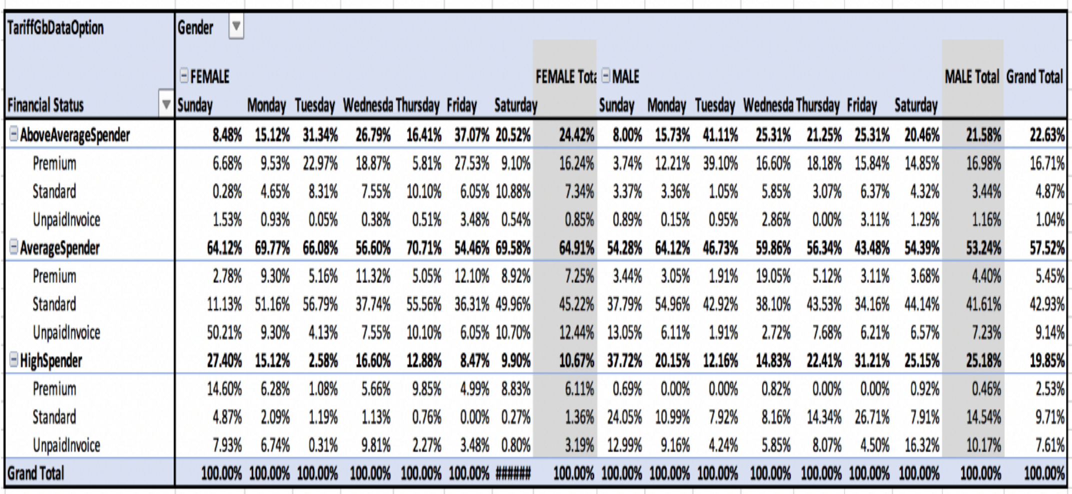
Understanding and carefully selecting the ground truth data to solve business problems is the second and most important challenge of a data mining process. Ground truthing in the segmentation model will recover the root causes of customer profiles classification. The selected ground truth data are composed of labeled features from the mobile operators based on demographics, socioeconomic and geographic characteristics of the telecommunication customers.

Results in this chapter are based on a large dataset that contained 26717 instances and 86 attributes with a further 11 columns comprised of formula derived values or classes to categorise the data. The derived attributes were labeled based on the processes applied by the operator. It is important to understand that derived attributes are new variables that are based on the original variables that were modified in order to enable an efficient analysis of the results. The data that the company can access are currently not being used to their full potential as a means of understanding the customers that are served, their sale patterns, the potential fraud risks, and churn patterns. Finally, the selection of variables in Table 6.1 illustrate the most relevant customer data.

TABLE 6.1 ID MOBILE IRELAND CUSTOMERS DATA



TariffTextOption	(All) ▼
ApproxLocationCounty	(All) ▼
TariffVoiceOptionmin	(All) ▼
CustomerAge	(All) ▼
CustomerinCollectionsIndicator	(All) ▼
TotalNumberofInvoices	(All) ▼
CustomerNetworkStatus	(All) ▼
CustomerLengthofService	(All) ▼
KPIAlarms	(All) ▼
CustomerContractLength(inMonths	(All) ▼
CareCallDurationSeconds	(All) ▼
TotalPaidAmountofInvoices	(All) ▼
SaleTimeofDay	(All) ▼
LocationPriority	(All) ▼
CustomerAgeGroup	(All) ▼
AlarmCost	(All) ▼



TariffGbDataOption	Gender ▼																
	⊟ FEMALE							FEMALE Tota	⊟ MALE							MALE Total	Grand Total
Financial Status ▼	Sunday	Monday	Tuesday	Wednesda	Thursday	Friday	Saturday		Sunday	Monday	Tuesday	Wednesda	Thursday	Friday	Saturday		
⊟ AboveAverageSpender	**8.48%**	**15.12%**	**31.34%**	**26.79%**	**16.41%**	**37.07%**	**20.52%**	**24.42%**	**8.00%**	**15.73%**	**41.11%**	**25.31%**	**21.25%**	**25.31%**	**20.46%**	**21.58%**	**22.63%**
Premium	6.68%	9.53%	22.97%	18.87%	5.81%	27.53%	9.10%	16.24%	3.74%	12.21%	39.10%	16.60%	18.18%	15.84%	14.85%	16.98%	16.71%
Standard	0.28%	4.65%	8.31%	7.55%	10.10%	6.05%	10.88%	7.34%	3.37%	3.36%	1.05%	5.85%	3.07%	6.37%	4.32%	3.44%	4.87%
UnpaidInvoice	1.53%	0.93%	0.05%	0.38%	0.51%	3.48%	0.54%	0.85%	0.89%	0.15%	0.95%	2.86%	0.00%	3.11%	1.29%	1.16%	1.04%
⊟ AverageSpender	**64.12%**	**69.77%**	**66.08%**	**56.60%**	**70.71%**	**54.46%**	**69.58%**	**64.91%**	**54.28%**	**64.12%**	**46.73%**	**59.86%**	**56.34%**	**43.48%**	**54.39%**	**53.24%**	**57.52%**
Premium	2.78%	9.30%	5.16%	11.32%	5.05%	12.10%	8.92%	7.25%	3.44%	3.05%	1.91%	19.05%	5.12%	3.11%	3.68%	4.40%	5.45%
Standard	11.13%	51.16%	56.79%	37.74%	55.56%	36.31%	49.96%	45.22%	37.79%	54.96%	42.92%	38.10%	43.53%	34.16%	44.14%	41.61%	42.93%
UnpaidInvoice	50.21%	9.30%	4.13%	7.55%	10.10%	6.05%	10.70%	12.44%	13.05%	6.11%	1.91%	2.72%	7.68%	6.21%	6.57%	7.23%	9.14%
⊟ HighSpender	**27.40%**	**15.12%**	**2.58%**	**16.60%**	**12.88%**	**8.47%**	**9.90%**	**10.67%**	**37.72%**	**20.15%**	**12.16%**	**14.83%**	**22.41%**	**31.21%**	**25.15%**	**25.18%**	**19.85%**
Premium	14.60%	6.28%	1.08%	5.66%	9.85%	4.99%	8.83%	6.11%	0.69%	0.00%	0.00%	0.82%	0.00%	0.00%	0.92%	0.46%	2.53%
Standard	4.87%	2.09%	1.19%	1.13%	0.76%	0.00%	0.27%	1.36%	24.05%	10.99%	7.92%	8.16%	14.34%	26.71%	7.91%	14.54%	9.71%
UnpaidInvoice	7.93%	6.74%	0.31%	9.81%	2.27%	3.48%	0.80%	3.19%	12.99%	9.16%	4.24%	5.85%	8.07%	4.50%	16.32%	10.17%	7.61%
Grand Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	######	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

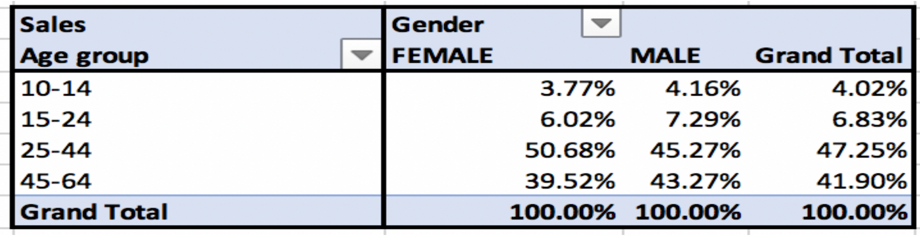
6.2.1 Customer Demographic Profiles

In this section the customer demographic profiles used to explain consumer behaviour.

6.2.1.1 Customer Age Groups

The extracted attribute of customer age groups categorises the customer's age into five groups, 10-14, 15-24, 25-44, 45-64 and 65+. These groups were chosen based on the Irish Census groupings to allow for future comparisons to be made

TABLE 6.2 MOBILE CUSTOMERS AGE GROUPS



Sales	Gender		
Age group	FEMALE	MALE	Grand Total
10-14	3.77%	4.16%	4.02%
15-24	6.02%	7.29%	6.83%
25-44	50.68%	45.27%	47.25%
45-64	39.52%	43.27%	41.90%
Grand Total	**100.00%**	**100.00%**	**100.00%**

6.2.1.2 Customer Location County

This variable is the approximate county location of the customers obtained from the county name in the Customer Bill Address column. Because the data are manually entered at point of sale, they are not consistent and may not always contain the county name. In that case the cell may contain null values However, Eircodes, which is Ireland's Post Code, has become mandatory as part of the sales process, which means that precise location categorisation can be found and manually replaced.

The customer details located in 30 different counties of Ireland were categorised with a higher demand in Dublin and Galway as well as in Donegal and Waterford. (Dullaghan & Rozaki 2017).

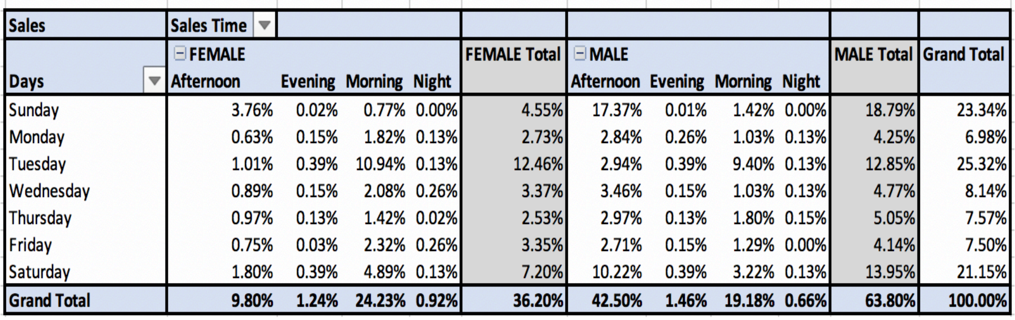
6.2.1.3 Customer Length of Service

This variable contains the total number of days a customer account has been in service. The variable was created by checking the Customer Network Status, and if inactive, subtracts the Subscriber inactive Date from the Customer Activation Date. If no inactive date was present, the variable was created by subtracting the Customer Activation Date from today's date to find the current length of service in days.

6.2.1.4 Service Sale Day

The service sales date attribute shown in table 6.3 was created from the value of the sale date column with the format of DD/MM/YYY and converted it to a day in the week. For example, 28/07/2015 was converted to Tuesday.

TABLE 6.3 DAY/TIME OF MOBILE SERVICE SALES



Sales	Sales Time ▼										
	⊟ FEMALE				FEMALE Total	⊟ MALE				MALE Total	Grand Total
Days ▼	Afternoon	Evening	Morning	Night		Afternoon	Evening	Morning	Night		
Sunday	3.76%	0.02%	0.77%	0.00%	4.55%	17.37%	0.01%	1.42%	0.00%	18.79%	23.34%
Monday	0.63%	0.15%	1.82%	0.13%	2.73%	2.84%	0.26%	1.03%	0.13%	4.25%	6.98%
Tuesday	1.01%	0.39%	10.94%	0.13%	12.46%	2.94%	0.39%	9.40%	0.13%	12.85%	25.32%
Wednesday	0.89%	0.15%	2.08%	0.26%	3.37%	3.46%	0.15%	1.03%	0.13%	4.77%	8.14%
Thursday	0.97%	0.13%	1.42%	0.02%	2.53%	2.97%	0.13%	1.80%	0.15%	5.05%	7.57%
Friday	0.75%	0.03%	2.32%	0.26%	3.35%	2.71%	0.15%	1.29%	0.00%	4.14%	7.50%
Saturday	1.80%	0.39%	4.89%	0.13%	7.20%	10.22%	0.39%	3.22%	0.13%	13.95%	21.15%
Grand Total	**9.80%**	**1.24%**	**24.23%**	**0.92%**	**36.20%**	**42.50%**	**1.46%**	**19.18%**	**0.66%**	**63.80%**	**100.00%**

6.2.1.5 Sale Time of Day

This variable contained the sales time in the 24 hour format HH:MM:SS and categorised into 4 types. If the hour was less than 6, it was defined as Night. If the hour was less than 12, it was defined as Morning. If the hour was less than 17, it was defined as Afternoon. Otherwise, the time of day was defined as Evening. For example, 10:17:55 was categorised as Morning.

6.2.2 Customer Account Information about Bills and Payments

This variable contains the billing information for each customer and service for a certain number of years. The last 16-months of Bill Pay customer account data are available and only used in the prediction system. However, different customers may have had different bill occurrences depending on when they joined the mobile network and their length of service. In this regard, the durations of customer bills might have been different (Huang & Kechadi 2012)

6.2.3 Total Invoice Amount Excluding Brought Forward

This variable contains the Total Invoice Amount Excluding Brought Forward in order to add in every month's invoice amount minus any amount unpaid and brought forward from the previous month. By performing this calculation, it was possible to find the total revenue generated by each customer's account.

6.2.4 Total Number of invoices

The attribute of Total Number of Invoices is to count the total number of invoices a customer has received. It required an additional row be created above the attribute names, with the string “Invoice” being placed above each column of data where the months’ invoice amount is. This is due to the data layout having the balance brought forward attributes between each month’s invoice amount and the Count function only being able to select a single range of columns.

6.2.5 Average Invoice Amount Excluding Brought Forward

The attribute of Average Invoice Amount Excluding Brought Forward checks to see if the value for Total Invoice Amount Excluding Brought Forward is zero, and if so returns zero, otherwise divides it by the value in Total Number of Invoices to find an average invoice amount.

6.2.6 Total Paid Amount of invoices

The Total Paid Amount of Invoices attribute takes a count of every paid invoice. As the attribute data is located in columns beside each other, a single range can be defined as paid (Pi,j) , unpaid (Ui,j) and decline status (DCi,j) of the invoices. Nji is a set of total invoices such as paid invoices, unpaid invoices and deferred company instalments. The total amount of invoices attributes defined in table 6.5: N = {Pji, DClji, Uji,} i represents the base station location and j the customer id.

6.3 Billing behaviour classification rules

The suggested proposal towards the pricing strategy is based on a set of rules that have been defined from a realistic pricing scheme of iD Mobile Ireland for checking customer willingness to pay based on average invoice amount. The set of rules shown in Table 6.1 were created based on the hypothesis of the first classification of their customer billing requirement and wiliness to pay.

TABLE 6.4 SPENDER STATUS

Status	Average invoice amount (X)
Low Spender (Class A)	X<= €15
Average Spender (Class B)	€15 <X< =€29
Above Average Spender (Class C)	€29 <X< =€50
High Spender Class D)	€50 <X< =€70
Very High Spender (Class E)	X> €70

The spender status classification facilitates the process of dynamic pricing based on behavior-based price discrimination (BBPD) or price discrimination by purchase history. (Esteves & Cerqueira 2017).

6.4 Ground truth

This algorithm provides the target attribute that categorises each customer account into four different classes based on their invoices being paid and spender status. If the Total Paid Amount of Invoices is greater than or equal to one less than the Total Number of Invoices, and if the Spender Status is either a Low Spender or an Average Spender, then they are defined as Standard. If the spender status is either a High Spender or an Above Average Spender, then they are defined as Premium. Alternatively, if the spender status is a Very High Spender and the duration of the customer contract is greater or equal than 24 moths then they are defined as VIP. If the Total Paid Amount of Invoices is less than one less of the Total Number of Invoices, then it is defined as Unpaid Invoice to highlight that a customer account is not paid up to date. The rules can be changed or modified by the mobile operator.

Consider a database S as follows:

When i= Location id and j= Customer id

$S_i = (N_{ji}, E_{ji}, T_{ji})$ where

N_{ji} is a set of the Total paid of invoice attributes: N = { P_{ji}, $DCI_{ji,}$ U_{ji}}

E_{ji} is a set of data related to Spender status: E = {$LowSpA_i$, Above $AverageSpB_i$, $AverageSpC_i$, $HighSpD_{i,}$ $VeryHighSpE_i$ }

T_{ji}= Length of time in months

```
For each location $L_i$ in S do
 If The value of Total paid invoices N>= (N-1)
          While Cust({$LowSpA_i$,}) OR Cust({Above $AverageSpB_i$)}
                  do Standard
Set the values of Total paid invoices N>=(N-1)
            If  Cust{($HighSpD_i$)} OR Cust{(Above $AverageSpB_i$)}
                  do Premium
 If Cust{($VeryHighSpD_i$)}) And Length of time >24 months
              then
     do VIP
Else Check for “Unpaid status”
```

6.5 Classification of customers segments

Customer profiling is one of the key factors in the successful operation of telecommunication networks (Lunn & Lyons 2018). In creating a suitable network design, it is essential to take into account customer profiles that have been constructed based on the analysis of their needs. Profile groups can then be used to understand how the network operation and limitations impact on different network resources and ultimately revenue.

Firstly, the suggested segmentation model aims to extract useful patterns and information about mobile customer behaviour. Using these, it will partition the customer base into groups of individuals that are similar in specific ways relevant to the company's sales deals, such as age, gender, location and spending habits. Finally, the output from the model will provide guidelines for actions to prevent customer churn based on the values associated with each mobile customer profile defined by the framework.

6.5.1 User identification based on classification rules

Customer profiling is an important part of the ability to successfully provide a variety of options in telecommunications services to satisfy customer needs and willingness to pay. The databases used in this study includes demographic information showing the location of the users and customers deals that demonstrate their wiliness to pay for mobile services. There are different groups of customers with different sets of needs behaviours towards telecommunication services. Market segmentation is an important part of the success of any business, and its importance is even greater significant in the telecommunications industry.

Companies offer a variety of new telecommunication services and technologies that increase the competitiveness of the industry. For example, companies that offered landline services experienced loss due to major shift of customers from fixed land lines to wireless communications. Telecommunication network providers generally serve different customer groups. Those different customer groups can be differentiated based on their quality of service (QoS) requirements. Some customers accept service delivery on a best-effort base, meaning services that are provided without guarantee that the job is executed. Other customers, however, require a guarantee that the job is executed according to the contracted service levels. The former customer group pays less for their service execution because of the lack of a guarantee, while the latter group only pays for service when service levels are met. Mobile providers favour customers with contracted service levels, known as gold clients, over best-effort customers who are processed only when resources are left. (Püschel et al. 2015).

6.5.1.1. Low income users (Silver Profile)

This profile included the consumers which are defined as “Low spenders” (ClassAji) or “Average Spenders” (ClassBji). For instance, students of having age group 18-24 belong to this group. These consumers are at an early stage of their professional career. Despite their limited financial resources this group shows strong desire to communicate with their peers using online applications and social media. This group can be divided further in to ‘heavy text users’ or ‘heavy data users’ depending on their preferred way of communicating. In the selected database Silver users could be either business or low-income customers depending on the pricing scheme (Yang et al. 2016, Yihua 2010).

6.5.1.2. Normal users (Gold Profile)

This group accesses their mobile phones on regular basis for calls and requires internet access at normal level. Normal profile users do not show any specific network demands in shape of capacity and services. They have price-oriented choices and make up the largest target group for mobile communication companies. They can be further divided into *business* and *prosumer* customers. Prosumer customers aged 30+ are generally familiar with new technologies and willing to pay more for services such as frequent internet use and video calls. Business users mostly use their mobile phone in context of their professional activity and are ‘heavy voice users’ with cell phones and mobile phone services an important part of their business. In our database the gold profile users have been classified either as “Above Average Spenders” ClassCji or “High Spenders”, (ClassDji).

6.5.1.3. Platinum users

Cell phone expenses of corporate users are financed by their employers or companies. Corporate packages typically involve a range of services such as unlimited nationwide calls that increases the demand for the capacity in network for such VIP users (Püschel et al. 2015). In this study, data are used to represent gold user such as VIP, Industrial and Business customers using the network in a specific time and data (Yihua 2010).The platinum users are presented in our data shown in Table 6.1 either with the status of “High Spender”, (ClassDji). or the “Very High Spender” (ClassEji).

6.6 Experiment 1: customer segmentation using decision tree classification rules

The attributes of an instance in the first experiment were used against the decision tree to determine the likelihood of the result by progressing through each step until the final decision. It should be noted that this is one of the most popular and widely used classification techniques.

First, classifying a customer’s VIP status is explained. Because of the large number of attributes in the dataset, the C.5 algorithm output was not clearly visualised. Hence, derived attributes of invoice information for several months were removed for this classification, the C.5 algorithm with 86 attributes was available for use.

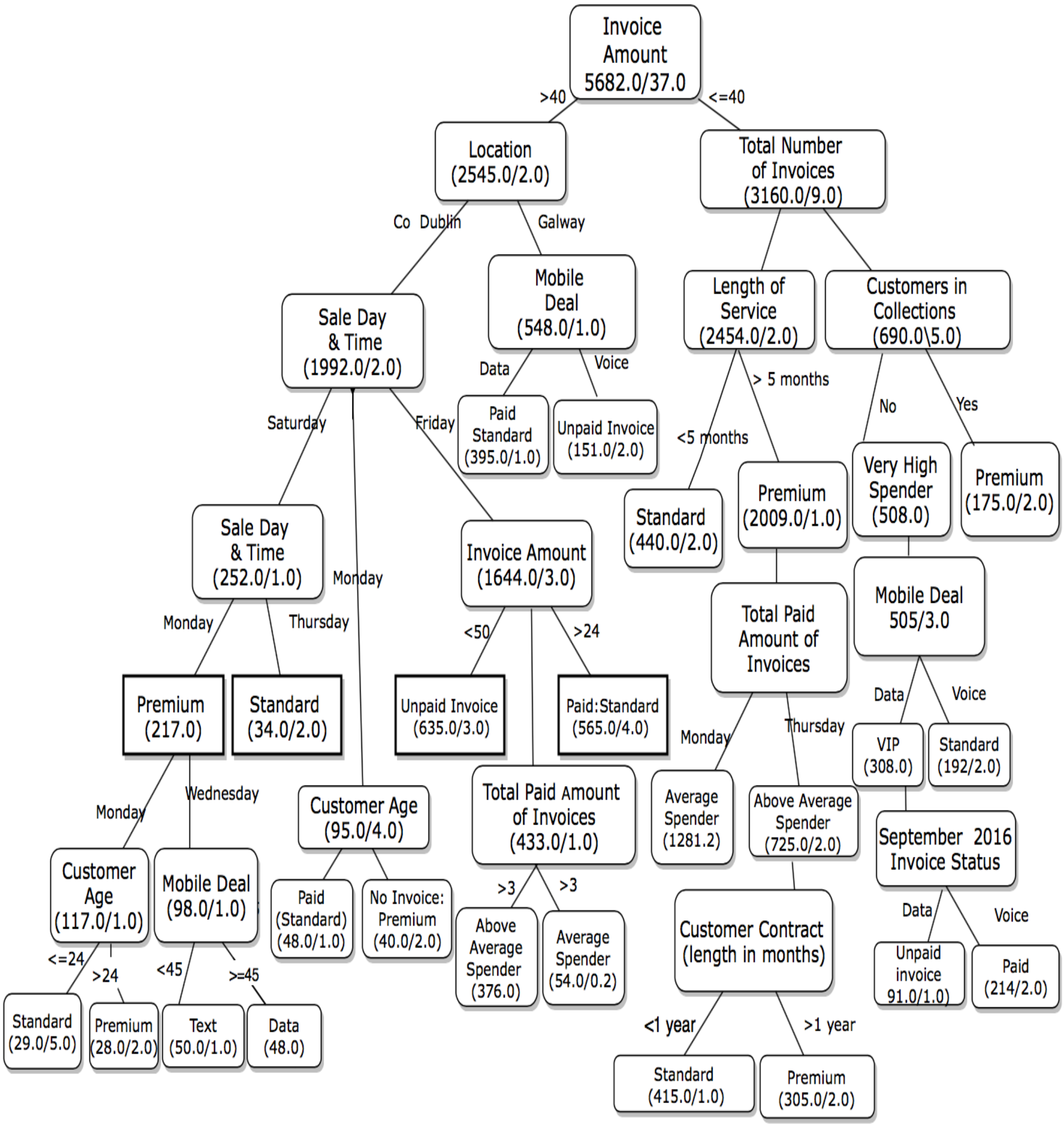

Invoice Amount 5682.0/37.0
>40
<=40
Location (2545.0/2.0)
Total Number of Invoices (3160.0/9.0)
Co Dublin
Galway
Mobile Deal (548.0/1.0)
Sale Day & Time (1992.0/2.0)
Length of Service (2454.0/2.0)
Customers in Collections (690.0\5.0)
Data
Voice
Paid Standard (395.0/1.0)
Unpaid Invoice (151.0/2.0)
> 5 months
<5 months
No
Yes
Saturday
Friday
Monday
Sale Day & Time (252.0/1.0)
Invoice Amount (1644.0/3.0)
Standard (440.0/2.0)
Premium (2009.0/1.0)
Very High Spender (508.0)
Premium (175.0/2.0)
Mobile Deal 505/3.0
Monday
Thursday
<50
>24
Total Paid Amount of Invoices
Premium (217.0)
Standard (34.0/2.0)
Unpaid Invoice (635.0/3.0)
Paid:Standard (565.0/4.0)
Data
Voice
Monday
Thursday
VIP (308.0)
Standard (192/2.0)
Monday
Wednesday
Customer Age (95.0/4.0)
Total Paid Amount of Invoices (433.0/1.0)
Average Spender (1281.2)
Above Average Spender (725.0/2.0)
September 2016 Invoice Status
Customer Age (117.0/1.0)
Mobile Deal (98.0/1.0)
Paid (Standard) (48.0/1.0)
No Invoice: Premium (40.0/2.0)
>3
>3
Customer Contract (length in months)
Data
Voice
<=24
>24
<45
>=45
Above Average Spender (376.0)
Average Spender (54.0/0.2)
Unpaid invoice 91.0/1.0)
Paid (214/2.0)
Standard (29.0/5.0)
Premium (28.0/2.0)
Text (50.0/1.0)
Data (48.0)
<1 year
>1 year
Standard (415.0/1.0)
Premium (305.0/2.0)

Figure 6.1 Customer Segmentation Decision Rules

The size of the resulting decision tree was 9124 and had 9062 leaves. The algorithm resulted in classification of the customer profiles into four categories of standard, unpaid invoice, premium and VIP status.

The rules extracted from the tree show the most reliable customers, their spending status, frequency of purchases during date and time, how to track unpaid invoices and block accounts and, whether they were VIP customers.

The VIP Status could be deemed to be a successful classifier because of the high accuracy of correctly classified instances. The decision tree algorithm showed additional rules about the VIP churners and their length of time, gender, location the services that have received, whether they were more likely to purchase a deal in regards of time and day, and the most popular addresses at which VIP customers were located.

After completing eleven tests, the Sample Mean of Correctly Classified Instances was 97.70315% for Percentage Split, while 10-fold cross-validation showed 97.6669% of instances were correctly classified.

6.7. Experiment 2: Bayesian modelling of customer profiles

In this experiment, Naïve Bayesian modelling was applied to the same data set. A machine learning technique based on Naïve Bayes model assumes the presence of a particular feature in a customer profile class that is unrelated to the presence of any other feature. As has been shown, the Bayesian network provided estimations about the utility of every possible attribute value in the spender status domain. In order to use these estimations to elicit customer preferences, a notion was needed regarding the relative importance of the attributes relative to each other. In the approach used in this study, the importance of an attribute depended on three factors: Customer demography, Customer Length of Service, Gender, Mobile deal Sale Day & Time of Day and financial considerations.

TABLE 6.5 NAÏVE BAYESIAN RESULTS

Naïve Bayesian Results		
Correctly Classified Instances	7964	87.71%
Incorrectly Classified Instances	1115	12.28%
Kappa statistic	0.7882	
Mean absolute error	0.0615	
Root mean squared error	0.2422	
Relative absolute error		20.24%
Root relative squared error		62.08%
Total Number of Instances	9079	

TABLE 6.6. ACCURACY BY CLASS

	PT Rate	FT Rate	Precision	F-Measure	MCC	ROC Area	PRC Area	Class
	0.980	0.210	0.844	0.907	0.793	0.960	0.941	Standard
	0.780	0.002	0.990	0.873	0.853	0.990	0.976	Unpaid Invoice
	0.771	0.029	0.893	0.827	0.781	0.964	0.910	Premium
	0.011	0.002	0.059	0.018	0.021	0.794	0.030	VIP
Weight Avg	0.877	0.120	0.120	0.871	0.794	0.966	0.942	

TABLE 6.7. CONFUSION MATRIX TABLE

Confusion Matrix				
A	B	C	D	Classified as
4778	9	65	4	a -Standard
323	1494	86	12	b-Unpaid Invoice
499	4	1491	0	c-Premium
59	2	32	1	d-VIP

Through the use of semi-structured data, it is proposed that a Bayesian approach be used to model the overall customer preferences in terms of the aspects identified from the consumption and the billing behaviour associated with a customer’s willingness to pay. This Bayesian model considered the overall rating of each deal as a weight and sum of the probability to select the individual offers.

This model allowed for the determination of an estimation of the tendencies for each deal aspect from each customer's perspective (Farhadloo et al. 2016). The use of the Naïve Bayes model provided a good, but still less accurate result than the C.5 model, of 87.7189% Correctly Classified Instances. The VIP type in the VIP Status class is where the precision is weakest and thus caused the reduced overall accuracy. This may occur for the reason that a significant number of VIP customers have unpaid invoices. It might be considered a point in the customer preferences where the model shows a high precision and accuracy related to the "premium offer" as evidence within the naïve Bayesian network leading to the posteriori probability distributions shown in Table 6.7. Evaluating the Influences as have been explained, the learning algorithm used in this study showed that the customer tends to choose a premium offer due to the additional need for cell services and mobile internet connectivity (Radde and Freitag 2010).

The Naïve Bayes results showed that the probability of selecting the premium offer was higher for the customers. However, there was also a very low independence among the preferences related to the "standard deal" offered by the mobile provider. So customer segmentation model encourage the mobile operators to reconsider the customer targets with low precision. Furthermore, churn management models should not only identify the customers that are most likely to leave the current service provider but also identity the customers that are most likely to respond positively to the right retention deal (Radde and Freitag 2010).

According to the results that were obtained, the “standard deal” needs to be reconsidered and reviewed in relations to customer needs and willingness to pay. Hence, the proposed optimisation framework benefits iD Ireland mobile company to efficiently segment its customer profiles and either remove or modify deals that are not demanding while providing deals for the customer with specific preferences using location, gender and payment behaviours data. In addition, a significant probability of fraud detection was found to be classified in the “unpaid Status” customer profile shown in the Bayesian results that need to be further investigated by the network providers.

6.8 Performance evaluation results

The customers’ billing behaviours and financial status were used to develop the two machine learning models that can be tailored according to the customer needs and customer’s wiliness to pay. In order to actually implement the plan, more retention efforts should be given to the potential churners who are most likely to react positively (Radde &Freitag 2010).

In this study, the prediction accuracy of each model for each data set in regards to the ROC curve, precision, recall and RPC area were examined. In addition, the predictions of models built and evaluated on more transactions in the training data, and models built and evaluated on customer’s deals were also examined. Furthermore, in addition to the True Churn and False Churn illustrated in the ROC Curve for decision tree models, the results of the Naïve Bayesian classification. PRC, and the overall accuracy were examined, and the results are shown in Figures 6.3.

These show the performance of the model achieved by each machine learning technique and the maximum accuracy that was archived by the decision tree model in all the deals (Lee at al 2004).

It should be noted that the main purpose of the Bayesian network was to derive utility estimations for attribute values from which customers' preferences were derived. In this regard, the churn probability in different deals shown in the Naïve Bayes results was based on the assumption that values that were more likely to satisfy the customer's needs were also more useful (Huang & Kechadi 2012).

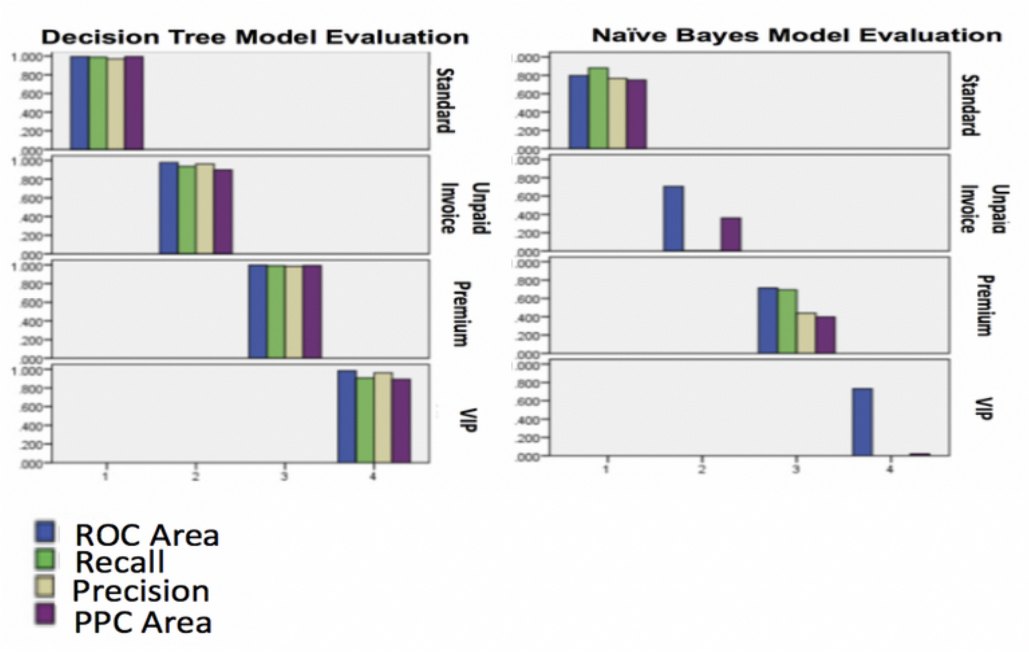


Figure 6.2 Models Evaluation

6.9 Summary

The use of machine learning techniques with a dataset built upon customer data generated from the telecommunications industry made it possible to test two different classifiers for categorising a Customer's Age Group, VIP Status, Spend Status and Customer Length of Service. All of the features used for the churn prediction in this study were either demographic billing, or usage features. The goal was to gain an understanding about the importance of these three types of deals for customers segmentation and churn prediction. The conclusion that was reached was that the billing and usage features had a very high importance for a customer segmentation scheme (Khan et al. 2010).

Finally, the two ways in which to evaluate customers' segmentation were equally important for predicting churn. The number assigned to a category of a spender showed the importance of a deal in customer segmentation. The result of the decision tree algorithm and preference given to the features are a bit different from the Bayesian model. However, usage and billing features were still of primary importance, while demographics may also affect the churn prediction (Khan et al. 2010). In the future, research is planned for mining further churning behaviours and developing retention strategies.

This part of the thesis has been published by the International Journal of Data Mining & Knowledge Management Process in 2017. The paper was recently cited by the book of *Marketing Concepts and Instruments in Supply Chain Management* written by T.Neukircheen, O. Gansser, and M. Klumpp and published by SpringerLink, with the purpose to show an example of machine learning techniques in customer churn predictions in the telecommunications industry (August 2019). An additional IEEE paper titled, "*A customer segmentation framework for targeted marketing in telecommunication*", (Namvar, Ghazanfari & Naderpour 2017) used our work to demonstrate the recent development in the 4G networks and the importance of the machine learning applications in terms of leading organisations to define appropriate marketing strategies. Finally, the last citation was received from a technical paper published by the Aalto University named "*Applications of Artificial Intelligence in Telecom Industry*" (V. Nguyen 2017). The variety of citations received for this work show that our purpose of demonstrating academically savvy work in machine learning with an applicable business approach in customer segmentation in the telecommunication industry was successful.

In the following chapter the findings of the customer segmentation framework will be integrated and amalgamated with the network optimisation and cost data and with the use of super learning techniques it will be provided a super-framework that will assist on an efficient end to end optimisation plan.

Chapter 7. Improving end to end Optimisation Decisions

The review of literature in Chapter 6 identified a clear need of a super-framework that can combine customer consumption and subscription data with insight into the network performance data, cell-site data, optimisation costs and customer preference data to trigger specific actions, which will help to enhance customer experience and reduce costs.

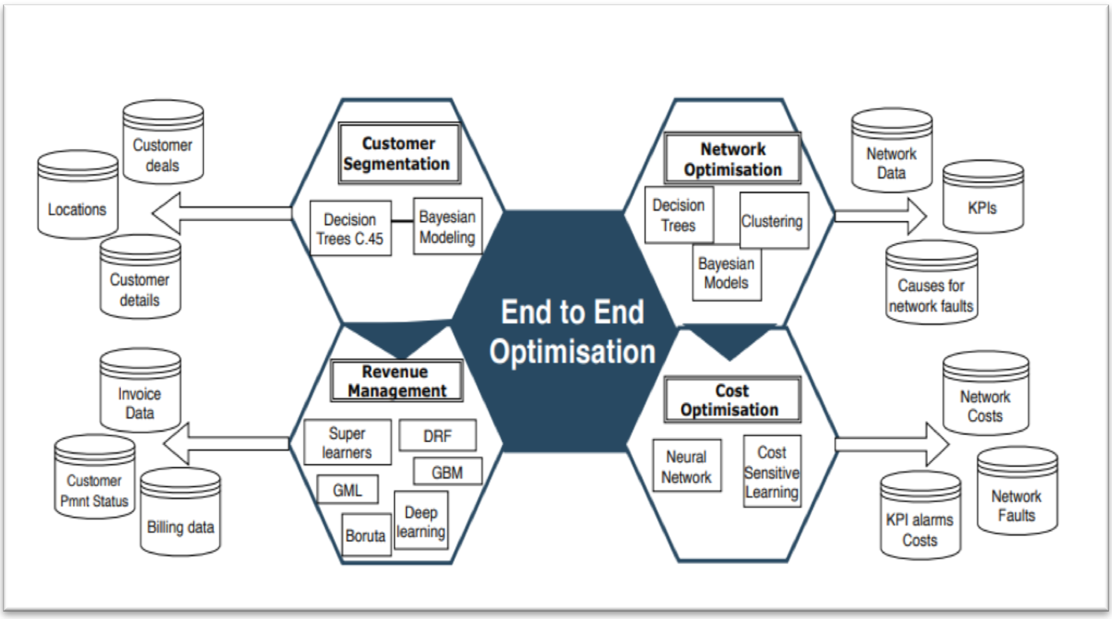


Figure 7.1 End to End Optimisation framework

Consequently, our final aim in this thesis is to propose an end to end optimisation super -framework capable of reducing network performance costs while improving the customer experience.

By connecting several data sets of customers' preferences and mobile network data in this chapter of the thesis it will be able to create a system that will distinguish potential issues (congestion, network parameters tuning, customer complaint etc.) and understand customer behavior.

From the data point of view the main challenge we face is to connect two distinct and diverse, yet important, data sources in the telecommunication industry. These are the databases containing technical optimisation data and customers' information data. As a matter of fact, many telecommunication companies recently are facing difficulties in running satisfactorily big data projects between IT, data analysts, and business. (Bughin 2016).

Connecting large data sets from different sources can be challenging. Wemegah & Zhu (2017) use Radio Frequency Identification (RFID) Data in order to allow traffic count volume (number of counts of vehicles in one hour) to be linked with associated locations and manage to control and facilitate traffic routes in Nanjing, China. Kalra & Lal (2016) search for intelligent methods based on information retrieval to connect heterogeneous data such as audio, video, online transaction, social media and E-commerce websites data from different domains including business, industries, and medical to allow users to access information according to their needs Furthermore, many companies such as NquiringMinds Ltd and SAS have efficiently addressed the challenges of data connectivity due to the needs of smart cities projects by using collaborative data platforms and suitable mobile applications for practical data sharing activities. (SAS, 2015)

To solve the connectivity difficulties, our datasets have been linked using the location IDs that define the customers; address connected with base station locations as shown in figure 7.2 using the latest Eircode system that facilitate the search of the location codes in Ireland and can be used to define the postcode identification.

The goal of revenue management is to enhance the revenues of a company by means of financial management decisions, such as predictive modeling and addressing network performance, as a means of improving end to end Quality of Services, managing different QoS (Quality of Service) requirements, providing sufficient network optimisation and planning in reasonable costs, and guaranteeing user satisfaction. The use of the revenue management process for a cellular system must also be based on the trade-off between network costs and the provider's profits (Raitoharju 2015).

Accurate network management is essential for appropriate scheduling and planning of the operations and functioning of telecommunication networks. (Pervaiz, 2010). The main objective of this part of the thesis is to present an end to end optimisation framework that will also provide a priority scheme for admission control. In order to provide service performance and user priority, a framework to prioritise network reconfiguration in network locations using data mining techniques was created to represent the user of different categories with the highest priority assigned to the most important customers

7.1. A model for end to end network optimisation

The objective of this model is to find patterns and trends across customers' needs and network utilisation and also help forecast and optimise network performance to minimize costs while improving QoS. The first driver of network efficiency for a revenue management predictive model should be datasets that include cell data and KPIs, network faults and alarms to find out the causes of network issues in order to support network optimisation decisions. Network queries should be modelled so that each query such as acceleration of traffic regulation and rate adaptation is represented by a dataset consisting of network optimisation techniques (Posnakides et al. 2015).

The information from the previous work on this area of research (Rozaki 2016) was focused on technical optimisation issues and cost evaluations of network faults to meet demand and provide high quality service. The extension of that research leads to the effort to minimise the loss of revenue of the network providers (Raitoharju et al. 2016, Zheng et al. 2017).

Our proposed end-to-end optimisation model identifies the most significant parameters of affected network areas and predicts the causes of network malfunctions while classifying the priorities of the optimisation functions. The objective of the model is to define criteria to prioritise network, costs and revenue optimisation actions using machine-learning algorithms to expose hidden relationships.

7.1.1 Input variables used for classification

The training data (70% of our sample) consists of a number of records, each specifying:

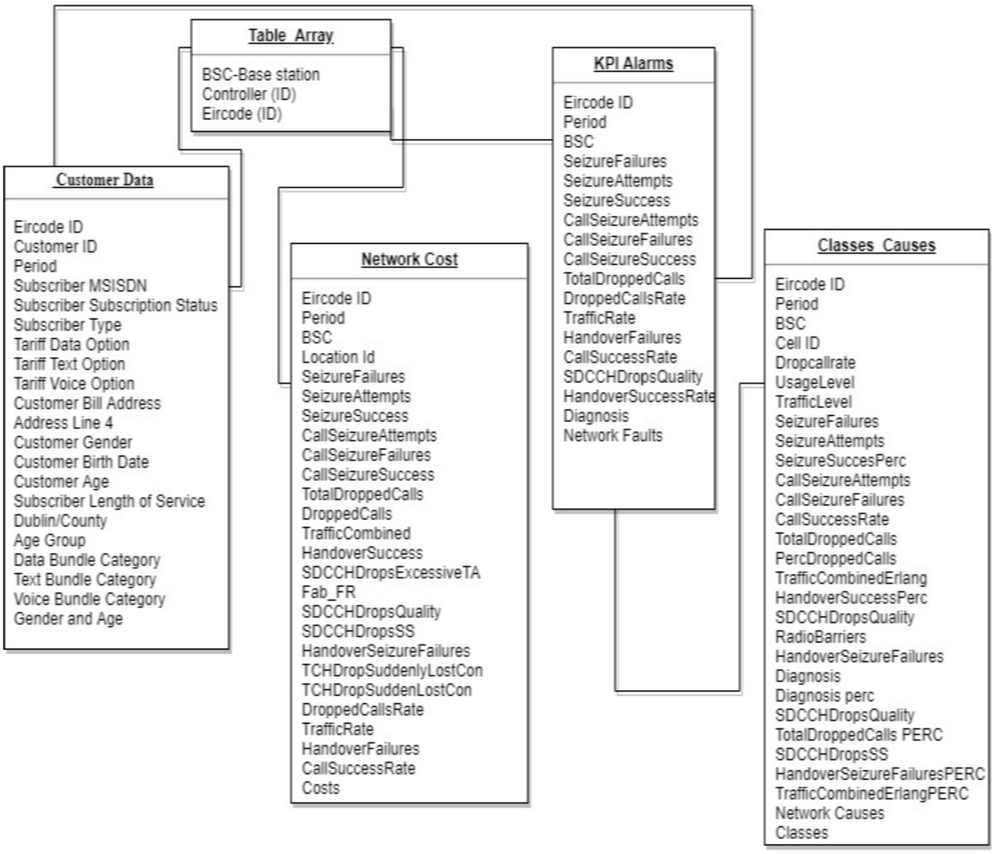
TABLE 7.1. INPUT VARIABLES

J	*The index of the location (cell) of this performance report*
N	*A KPI alarm status (generated by the technique in chapter 4) which takes the value NORM, WARN or CR*
E	*The cause of the network malfunction – a categorical value of either class A, B or C (see chapter 4).*
F	*The type of network cost – either MAIN (maintenance), TR (traffic), INF (infrastructure) or TRSM (transmission).*
C	*The customer willingness to pay profile associated with this report – either AbAvg (above average), Avg (average), HSpd (high spender) or VHSpd (very high spender) (Chapter 6)*
S	*Customer deal profile – VIP, Premium, Standard or Unpaid. (Chapter 6)*

The dataset used in this study has been provided by Id mobile Ireland and contains different types of variables such as network performance measurements, customer's data, network costs, KPIs and mobile deals. The dataset consists of 9,000 mobile customers and mobile locations.

The training set is used to learn and estimate the priority rules of our models, which are then evaluated on an unseen sample, the test set, to evaluate their performance. About 20% of these customers are VIP, whilst the others are business, prosumers or low budget customers using the network from various geographical locations. The database contains a table with the KPI alarms (Chapter 4), the classification of the causes of network malfunctions (Chapter 4) the network costs and a table array with the Eircode unique ids and the base station location.

Figure 7.2. Data model relations



7.2 Pre-processing Stage. The use of the learning algorithm to set up the priority rules.

The priority rules formula was used as a dependent variable in the model. The priority classes are the base of the categorisation rules based on the input variables. Thus, the input variables shown in Table 7.1 were used to calculate the priority classes. The purpose of the learning algorithm was to find the relation with the other variables in order to define the priority classes.

The ratio between the data is 40% test data to 60% training data, so it is imbalanced data. Classification models normally perform best when the class distribution is approximately even (Verbeke et al., 2012), so this may cause models to encounter difficulties in learning which customers should be prioritised, resulting in poor classification accuracy. For this reason, the data sets used have also been tested also using a ratio of 30% test data to 70% training data. In the following subsections, it will be described how features of the technical and customer data were generated.

Rule 1-Priority D:

```
Get values for fault diagnosis data sets $\{(N_{ij}, E_{xj}, F_{kj}, \mathrm{S}_{zj},)\}$
and the pattern of costs $F_{ki} =(MAIN_{ki}, TR_{ki}, INF_{ki},TRSM_{ki})$

Set $\mathrm{L}_i$, the starting location for attempted match, to $NORM_{ji}$ alarm

  For each location $_i$ in M do
    Set the value of causes $E_{xj}$ :{($ClassA_{ji}$, $ClassB_{ji}$, $CLassC_{ji}$)} to $ClassA_{ji}$,
      While $Alarm$ ($N_{ij}$ =Norm)
         AND $\mathrm{S}_{zj}$:{(VIP$_{zj}$, , Premium$_{zj,}$ Standard$_{zj}$)}
            OR $cost\ F_{ki}$ $:(\{(m= MAIN_{ki},)\})$
            do Priority D
```

Rule 2-Priority C:

```
Get values for M and N, the size and the pattern, respectively.
Get values for all data sets $\{(N_{ij}, E_{xj} C_{yj} F_{kj}, )\} n i=1,$
 and the pattern $N_{ij}\{ (NORM_{ji}, CR_{ji,} Warn_{ji})\}$
    Set the value of causes $\{(ClassA_{ji}, ClassB_{ji}, ClassC_{ji})\}$ to $ClassB_{ji}$,
        While alarm (j=CR$_{ji,,}$) AND $C_{yj}$:{($y$= Above averageSpB$_{yj}$ )}
            OR $C_{yj}$ :{($y$= AverageSpC$_i$)}
            Else cost $F_{kj}$, ({(k= $TR_i$ ) OR (k= $TRSM_i$)})
             do Priority C
```

Rule 3-Priority B:

```
Set up values for Key Performance Indicators data sets
 $\{( N_{ij}, E_{xj}, F_{kj}, S_{zj} C_{yj} )\} n i=1,$
     and the pattern $\{(NORM_{ji}, CR_{ji}, WARN_{ji})\} n i=1,$
      If alarm (j=CR$_{ji,,}$)then
        Set the value of causes $\{(ClassA_{ji}, ClassB_{ji}, CLassC_{ji})\}$ to $ClassB_{ji}$,
          while both CuSTM({(z= Premium$_{j,}$VIP$_j$)} OR CuSTM({($y$= HighSpD$_i$)}
            AND cost ({(k=$TR_{ki}$ i)} OR (k= $TRSM_i$)})
             do Priority B
```

Rule 4-Priority A:

```
Get values for M and N, the size of the alarms data set and the pattern, respectively.
  Get values for all data sets $\{( N_{ij},, F_{kj}, S_{zj} C_{yj} ,)\}$
    and the pattern $\{(NORM_{ji}, CR_{ji}, WARN_{ji})\}$
     If cost ({(k=$INF_{ki}$ i)} AND ($y$= VHighSpD$_i$)} OR $\{(z=VIP_j)\}$
       Else $\{(NORM_{ji}, CR_{ji}, WARN_{ji})\}$ to $WARN_{ji}$,
          then
           do Priority A
```

7.3 Identifying customers' requirements for prioritization

Here a k-mean clustering algorithm applied to predict relations between customers' consumption and willingness to pay. The aggregated data obtained from the five clusters ("5000 texts-5000 Minutes FEMALE 15-24", "100 texts-1000 Minutes-FEMALE 25-44", "100 texts-1000 Minutes-FEMALE 25-44" and ionization, "5000 texts-5000 Minutes-MALE 15-24", "100 texts-250 Minutes-MALE 45-64"). After labelling the five categories of mobile customers in order to supervise the algorithm, the data were passed to the clustering analysis. To perform a cross validation, five instances were divided into a training data set of 80% with 26717 rows and 86 attributes. All data were randomly mixed and passed to the classifiers.

Each test was repeated 10 times and the clustering predictions in classifications in customers' deals shown in Table 7.2.

TABLE 7.2. CLUSTERING RESULTS OF CUSTOMER PREFERENCES

Attribute	Cluster				
	Cluster 0: 5000 texts-5000 Minutes FEMALE 15-24 (0.2)	Cluster 1: 100 texts-1000 Minutes-FEMALE 25-44 (0.11)	Cluster 2: 5000 texts-5000 Minutes-MALE 15-24 (0.16)	Cluster 3: 5000 texts-5000 Minutes FEMALE 15-24 (0.28)	Cluster 4: 100 texts-250 Minutes-MALE 45-64 (0.25)
Customer in Collection Indicator					
Y	847.61	2905.54	4125.81	7308.94	6456.08
N	4476.77	72.80	150.32	151.37	217.71
Total	5324.38	2978.35	4276.14	7460.31	6673.80
Customer Gender					
Female	2035.70	2919.11	56.61	2863.25	2974.31
Male	3266.68	59.23	4219.52	4597.06	3699.49
Total	5324.38	2978.35	4276.14	7460.31	6673.80
Spender Status					
Average Spender	857.31	2667.10	3754.84	3282.72	1.009
Above Average Spender	1398.30	1.020	1.031	1.021	5468.62
High Spender	1198.77	1.012	1.040	1.014	1.00
Low Spender	554.08	311.19	521.20	4177.50	1.00
Very High Spender	1317.89	1.020	1.020	1.052	1.00
Total	5327.38	2981.35	4279.14	7479.31	6676.60
VIP Status					
Standard	1.52	2940.23	4130.93	7278.19	1.000
Unpaid Invoice	5062.25	38.01	145.14	182.11	56.45
Premium	1.650	1.000	1.000	1.000	6617.34
VIP	260.95	1.0155	1.0154	1.0099	1.0059
Total	5326.36	2980.35	4278.14	7462.31	6675.80
Customer Age Group					
45-64	1127.03	1016.63	1346.60	2566.27	2126.44
25-44	3594.79	1446.46	2153.43	3771.79	2554.46
15-24	535.19	249.23	585.73	593.40	433.38
65+	69.36	248.01	392.26	530.84	21.50
Total	5326	2980	4275	7462	6675
Customer Length of Service					
Mean	206.62	360.44	352.20	97.32	295.32
Std. Dev	101.00	102.125	100.19	67.14	129.79

Table 7.2 shows 5 clusters of the mean and standard deviation of the customer 16 profiles, gender, spending status, age group and the average of the length of service. Each cluster describes a group of customers. It is worth mentioning that the selection of these criteria is rational since most of these classes reflect the requirements of both customers deals and mobile network optimisation needs. (Dullaghan & Rozaki 2017.

TABLE 7.3: ATTRIBUTES SHOWING FIVE DIFFERENT CUSTOMER DEALS GIVEN BY ID MOBILE IRELAND

Clusters (k-mean)	**Attribute**	**Clusters Instances**
Unsupervised Clusters		
Cluster 0	20Gb Data -5000 Texts –5000 minutes Female 25-44	28% (7371)
Cluster 1	20Gb Data-100 Texts –1000 minutes Male 25-44	28% (7544)
Cluster 2	20Gb Data-100 Texts –1000 minutes Female 25-44	11% (3002)
Cluster 3	20Gb Data-100 Texts –5000 minutes Male 25-44	22% (5762)
Cluster 4	20Gb Data-100 Texts –250 minutes Male 45-64	11% (3024)

The attributes in Table 7.3 show five different customer deals given by iD Mobile Ireland. The clustering scheme shows an efficient system to manage revenue, costs and customers consumption issues in order to improve pricing planning and budget planning as well as to keep the customers satisfied. Customers between 25-44 found to be more demanding and they may require the higher amount of tariff options.

7.3.1 Priority scheme based on clustering techniques

Next database trained within the learning algorithm was applied to the base stations cell locations in order to cluster the problematic areas showing the total results of fault alarms, costs and customers' consumption. The resulting priority scheme will indicate the clusters that they will need to be prioritised based on the criteria set up by the learning algorithm.

The aim of the priority scheme using clustering techniques is to identify relations between customers' consumption and locations, technical issues and cost considerations. (Cheng et al 2016). In this study, a budget optimisation plan and a priority scheme based on customer needs, network potentiality and technical costs was evaluated. It is important to note that different clustering areas and distance metrics could be adopted for re-aggregation.

7.4 TABLE.PRIORITY SCHEME USING CLUSTERING TECHNIQUES

Attribute	Cluster 0: (0.16)	Cluster 1: (0.08)	Cluster 2: (0.06)	Cluster 3: (0.07)	Cluster 4 (0.1)	Cluster 5 (0.16)	Cluster 6 (0.17)	Cluster 7 (0.08)	Cluster 8 (0.12)
KPI Alarms									
WARN	110.61	152.76	52.09	7.86	3.039	326.12	53.69	23.82	244.74
CR	165.00	3.28	66.69	146.00	1	18.62	272.52	140.40	1.178
NORM	67.31	6.58	1	1	217.05	1.00	40.93	1	1.053
Total	343.13	162.64	120.05	154.87	221.14	345.75	367.16	165.23	246.99
Alarm Costs									
Infrastructure Costs	110.81	152.76	52.09	7.86	3.03	326.12	53.69	23.82	244.76
Transmission Costs	17.98	1.99	66.96	22.032	1	18.62	4.98	140.40	1.0017
Maintenance Costs	67.31	6.58	1	1	217.10	1.00	40.91	1	1.05
Traffic Costs	148.02	2.28	1	124.97	1	1	268.53	1	1.17
Total	344.13	163.64	121.05	155.87	222.14	346.75	368.16	166.23	247.99
Low Income Customers									
Mean	15822.0	12754.94	12333.85	2819.57	2705.88	1915.21	1911.20	2181.01	2013.92
Std. Dev	11051.26	10696.60	10429.48	2410.54	1958.84	582.71	624.47	1061.06	631.48
Prosumers									
Mean	16918.58	13190.77	12757.49	18882.51	2273.79	1576.05	1447.97	1826.51	1625.90
Std. Dev	10899.22	9804.01	11895.51	1015.16	2260.72	658.20	611.80	1404.51	772.74
Business Customers									
Mean	15539.55	13339.35	12466.31	3080.32	3088.24	1658.38	1638.63	2142.76	1767.89
Std. Dev	6675.43	6900.25	7747.02	2724.27	3142.45	812.09	686.54	1353.31	514.01
Industrial Customers									
Mean	16854.11	14287.52	14318.29	3192.10	3249.83	2099.12	2166.91	2461.04	2566.85
Std. Dev	8536.48	8865.13	8973.80	2313.71	2893.60	1398.24	1317.21	1766.35	1928.45
VIP Customers									
Mean	6463.74	5187.77	5409.79	1151.42	1313.04	1088.67	1034.78	1085.24	1127.59
Std. Dev	4475.53	4613.55	4508.29	763.91	1094.77	763.73	705.22	812.63	749.39
Priority Scheme									
Priority D	14.72	28.24	14.17	142.74	192.01	343.75	365.15	162.2	240.98
Priority C	44.54	16.57	16.68	5.77	9.10	1	1	1.31	1
Priority B	157.41	79.26	54.22	6.34	20.02	1	1.00	1.72	5.00
Priority A	127.45	39.56	35.98	1	1	1	1	1	1
Total	244.13	163.64	121.05	155.87	222.14	346.75	368.16	166.21	247.99

Table 7.4 shows 13 clusters of the mean and standard deviation of the KPI alarms, network costs and customers preferences within consumption.

Data similarity between the warn KPI alarms and infrastructure costs is shown as the number of instances classified as warn is the same number with the instances classified as infrastructure costs. The reason for the data similarity is because the conditions set up for the supervised clustering model in the network fault algorithm show that all selected cells are under warn alarm limits that need to be checked for infrastructure issues. The priority scheme shows that the cluster 0 cell ids and areas need to be clearly prioritised based on the revenue consideration scores. However, there is also a large number of VIP customers that are using the specific cells of this part of the network. In this regard, the revenue received from the usage of all the VIP and business customers should cause the operators to prioritise the clustered area number 0. As a result, the clustering scheme shows an efficient system to manage revenue, costs and technical issues in order to improve the network performance, as well as to keep the customers satisfied.

7.4 Priority parameters using neural networks

We apply a back-propagation algorithm to define additional variables that are related to the defined priority rules while treating learning as a revenue optimisation problem. The algorithm is used to help with the extraction of the priority criteria based on the hidden layers (Rozaki 2016). Firstly, combinations of the different scores of technical details, costs evaluations and the type of customers that might be affected at a specific location are extracted. The warn alarms serve to ensure that the VIP customers at any class cost are categorised in priority A.

The gold users that are defined as Business and industrial customers in the databases will be emerged by the priority B, which also includes the warn and critical alarms and the constructions and transmission costs issues. The priority C range includes the needs of silver users and the causes of network issues in a variety of costs and alarms in order to provide an efficient network optimisation.

Finally, priority D will be ordered by the revenue management system as maintenance issues and traffic updates for the low-income customers. However, it should be noted that the neural network learning process would extract the revenue optimisation rules without identifying the base station locations that serve the mobile customers. In this regard, the revenue scheme would be identified with the clustering methods that shows the network coverage areas where the users operate. This would need to be prioritised based on the revenue management and priority criteria given by the back-propagation algorithm.

The algorithm is appropriate to be implemented in distributed computing environments in which the entire revenue optimisation process is spread across multiple locations (Gandomi, & Haider, Beyond 2015).

In the neural network model, one hidden layer per interested output layer is enough for the learning of the network. Networks with more than one hidden layer are more complex and time consuming. The number of neurons in the hidden layer has the greatest effect on network performance. If there are not enough hidden neurons, the network will find difficulty in the learning, but over-dimensioning of the hidden layer above a certain threshold level does not result in a performance improvement, and in extreme cases can degrade performance. Additional attention should be given to number of iterations in order to avoid over-fitting (Beretka, & Varga, 2013)

Consequently, the neural network that was developed for this experiment consists of 23 input neurons related to the datasets, which included 13 hidden layer neuron showing as sigmoid nodes and 4 output layer neuron that represent the priority options related with the optimum combinations of the hidden layers of neurons in the hidden layer in order to define the sets of data that will emerge the optimisation plans.

Figure 7.3 visualises the neural network nodes that were created based on the variable importance. Changing the number of neurons (hidden layer) produces different results, thus all architectures have to be analysed and tested in order to find the optimal revenue management decisions. In this regard, the results of networks were tested with differing number of neurons (Beretka, & Varga, 2013).

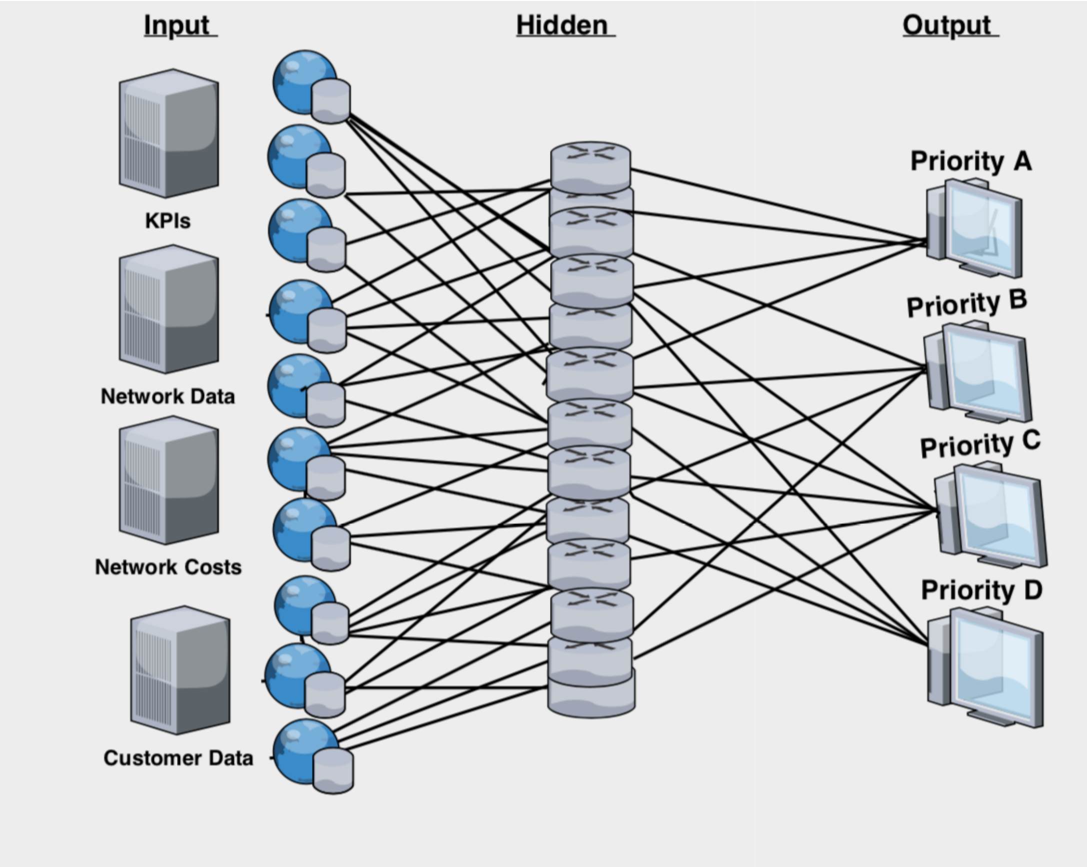


Figure 7.3. Threshold of priority parameters

Figure 7.3 shows the network of the end-to-end priority nodes throughput measurements for a test receiver with 23 datasets with 26717 rows loads. The results show that the best model consists of an input layer with 86 attributes, 13 hidden layers and an output layer with four output layers defining the priority rules layers.

The deployment problem considered here is specified by knowledge of locations for client infrastructure nodes, such as residences and businesses. The probability of subscription at each location is assumed to be known or estimated a priori and expressed as a function of price. This so called demand curve is dependent on the availability and price of competing services and economic circumstances, such as subscriber affluence at the deployment locations (Posnakides et al.2015).

7.5. Ensemble learning algorithms as support for end to end optimisation decisions

Ensemble learning algorithms are used in this final part of the thesis to combine the results of several base classifiers to achieve the final outcome. Our purpose is to assign the most appropriate priority class to customers by combining the results of several base learners using a powerful and efficient ensemble for multinomial classifiers. The way in which the algorithms combine the results impacts the performance of the ensemble classifiers. Stacked generalisation is an ensemble method that can be used to combine different predictive algorithms. Combining a library of different algorithms provides the ability to enhance performance, as well as to create the best combination of algorithms to increase accuracy (Zhai & Chen, 2018). Stacked learner uses K-fold cross-validation to create the optimal weighted combination of predictions from a library of algorithms. The benefit is that over-fitting is avoided with K-fold cross-validation.

It should be noted that optimality in this context is defined by a user-specified objective function. For example, optimality can be defined as minimizing mean squared error or maximizing the area under the receiver operating characteristic curve. In previous studies, researchers have successfully applied these techniques to improve predictive accuracy, avoid overfitting, and minimize parametric assumptions (Naimi and Balzer 2018). Common ensemble learning methods include a combination of base learners, as well as boosting, bagging and stacking.

As part of this thesis, the performance of these methods is evaluated and compared. The boosting process used involves initially training a model and incrementally creating new models with a focus on addressing the classifying errors made in the previous model (Young, Abdou & Bener 2018)., A gradient boosting decision tree algorithm uses trained trees to reconstruct the difference between the target function and an ensemble prediction (Sangani, Erickson & Hasan 2017).

The bagging process involves using training models on random subsamples in which each model votes with equal weight on the classification. Random forest uses a bagging process to enable the selection of a random set of features at each internal node. The training of the bagging ensemble occurs with the resulting training set comprised of the lowest margin instances. New margin values are calculated for each of the remaining training instances. This procedure, which is an iterative process, is repeated until the maximum training accuracy is reached and provides an optimal ensemble which has a reduced and more informative training set (Guo and Boukir 2017).

The overall result is that stacking takes the output of a set of models and feeds them into a separate algorithm that provides the final predictions. This entire process can involve any set of base learners and an ensemble learning algorithm. The method combines the predictions of the models and averages them with a simple or weighted measure.

The super learner process is a pooling method that is used to determine the optimal combination of variables for predicting an efficient network optimisation plan from technical, financial and customers' perspective (Young, Abdou & Bener, 2018).

7.6. Multi –Label super learners. Methodological approach and design

Ensemble learning algorithms are used in the last part of this thesis in order to combine the results of several base classifiers to achieve an optimal final outcome. It is the amalgamation of the results of the base learners that is important for the creation of a powerful and efficient ensemble of multinomial classifiers. The way in which the algorithms combine the results influences the performance of the ensemble classifiers.

The learning process begins with the initialisation of the cellular data population encoding of the randomly selected alarm faults and cost class examples for each individual set of customer data in order to set up the priority classes. The input data for the learning process includes training using clustering techniques the base learners used for the stacked ensemble model.

A super learner ensemble methodical approach is used to integrate the end to end optimisation super -framework for multi-class classification problems using H20 models. H20.ai provides an open source machine learning platform that provides an easy method to build smart applications. (Guo and Boukir 2017) H20 automates most of the steps of the super learning algorithm. In addition, it facilitates the process of building ensembles models. It should be noted that the h20.stack function is an alternative to the h20.ensemble function that allows the user to specify H20 models individually and then stack them together at a later time.

The analysis of the model represents the performance of the classifiers, the analysis of discretised priority groups, including identification of the important features of the prioritised locations, customer groups and additional characteristics of our variables. It has been also performed a comparison of the different base learners with other related models that are applied to the same dataset. Therefore, super learning enables all of the information contained in the final data models to be used in order to achieve a good performance in the pattern recognition of multidimensional data used for an end to end optimisation super-framework. Therefore, this method was used in the final part of the thesis as it has the potential to differentiate customer profiles and investigate network performance patterns and costs in the contrast of decision trees and Bayesian models that present lower accuracy with multidimensional databases.

An important part of this study is that it adds to the existing academic literature on stacked ensembles, but synthesis of the classifiers, GLM, GBM, RF and Deep Learning, is limited regarding the analysis of the multinomial classifiers using the H20 super learner platform. http://docs.h20.ai. The tools used for this part of the thesis is R version R 2.13.0 and H2O.

7.7 Base Learners Using H20 Libraries

For the selection of base classifiers in this study, the learner libraries were selected from the H20 package using R implementation tool. The classification depended on their efficiency and also the type of data being used. Combining classifiers and constructing ensemble classifier meant that weak classifiers were preferred as they were easy to understand. In addition, when they were amalgamated, they provided accurate results in comparison to the results obtained from individual classifiers. (Kumar, Siraj and Singh 2017).

7.7.1 Gradient Boost Machine (GBM)

Gradient boosting is an algorithm that operates based on the principle of ensemble learning in which an ensemble of decision trees is built, and individual trees are summed to calculate the overall prediction (Sangani, Erickson and Hasan 2017). A decision tree is a classifier that splits a space of features into regions by applying conditional splitting. For example, the metric *Customer Length of Service* might be used to split customers who gain the deal for more than 12 months and those who are with the company for less than 12 months. Gradient boosting merges a set of weak learners, which are individual trees and are considered to be poor predictors and provides an improved prediction accuracy by leveraging the entire ensemble. For any tree, the prediction outcome of the tree is weighted based on the outcome of the previous tree. In this regard, the more accurate an outcome, the more it is weighted. In addition, when an ensemble of trees is used to make a prediction, each tree that is part of the ensemble will have a different level of contribution based on the weighting of its outcome (Sangani, Erickson and Hasan 2017).

Furthermore, boosting is trained on weak learners with high bias and low variance, the error is the variance plus the square of bias (MSE). Boosting works to reduce error by reducing the bias. At the same time, the variance is also reduced because of aggregating the output from many models. Boosting assigns a weight to each sample, which serves to determine the importance of the samples in the modelling process. If a sample is classified correctly, its weight is decreased. In contrast, if the sample is classified incorrectly, its weight is increased (Arora, et al. 2015).

7.5 VARIABLE IMPORTANCE FOR GBM MODEL

	Variable	Relative importance
1	Address.Line.4	147.8953979
2	KPIAlarms	146.0876617
3	AlarmCost	121.9902115
4	Tariff.Gb.Data.Option	58.40407181
5	Tariff.Voice.Option.min	38.90882111
6	VIP.Status	24.70096827
7	Customer.Length.of.Service	20.33009052
8	Approx.Location.County	18.20639729
9	Customer.Age	0.137790591
10	Care.Call.Duration.Seconds	0.108223855
11	Customer.Network.Status	0.103469923
12	Tariff.Text.Option	0.039902668
13	Total.Paid.Amount.of.Invoices	0.023508856
14	Spender.Status	0.022195147
15	Customer.Gender	0.00621597
16	Sale.Date.Day	0.004331739
17	Customer.in.Collections.Indicator	2.05
18	Customer.Contract.Length..in.Months.	4.53
19	Customer.Age.Group	6.81
20	Total.Number.of.Invoices	5.55
21	Sale.Time.of.Day	0

The data set used for this experiment shown in Figure 7.2 The result of the importance of the variables show that the City (Line 4) that are located the mobile customers has a high importance for the model, following by the KPI alarms the Network costs and Data and voice tariff deals. The GBM model introduced two new variables of importance for the priority classes which were the tariff data deal and the call duration (in seconds) that were not shown in the other classifiers.

7.7.2 Boruta Algorithm

To explore features that would be relevant to train the model, the Boruta method was used. This comprises of a wrapper approach built around the Random Forest (RF), a tree-based ensemble classifier that consists of many decision trees. An important characteristic of Boruta is the creation of duplicate copies of all independent variables. At every iteration, Boruta calculates the maximum z score among the attributes of the database The algorithm stops calculating when all features are confirmed or rejected, or it reaches a specified limit of random forest runs (Kursa & Rudnicki 2010).

The Boruta algorithm performed 100 iterations in 3.07 mins using the dataset show in Figure 7.2. This analysis performed 100 trees for each Random Forest classifier. The results yielded 25 attributes, 23 were confirmed as being important, 1 attribute confirmed as being unimportant, and 1 tentative attribute. The results of the Boruta algorithm is shown in Figure 7.5.

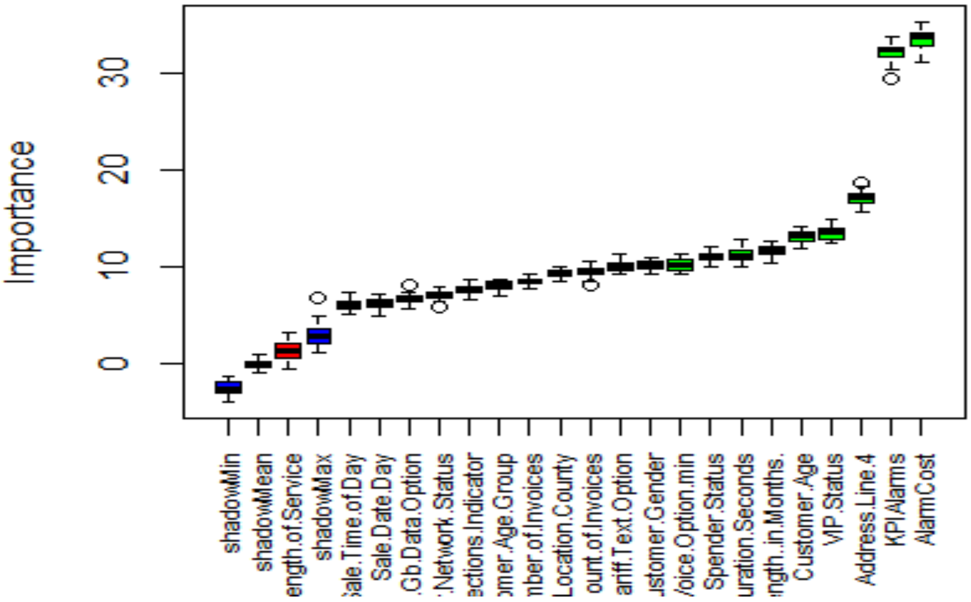


Figure 7.5. boruta model variables

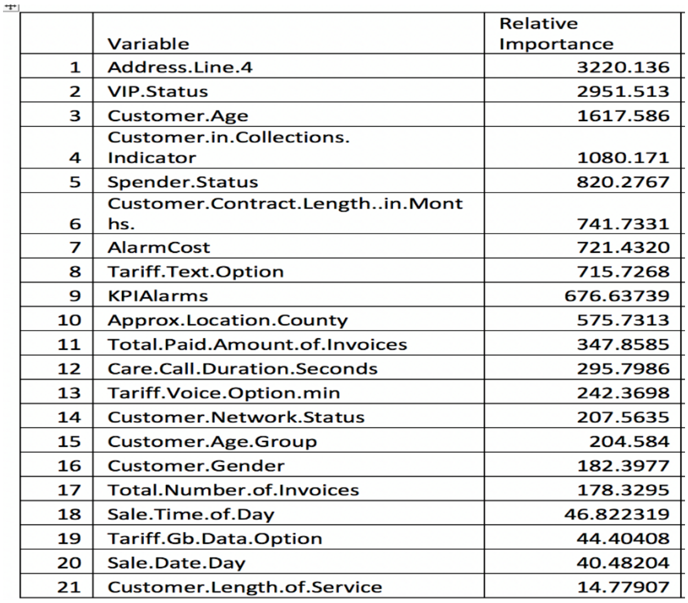
The attributes considered important are represented by the green boxplots, which include KPI alarms, Network costs, the address line 4 which is the city of the customer, the VIP status and the voice tariff option. The attributes that are not given a high importance were the sales time/day, collection indicator and the tariff text options. The only attribute that was confirmed to be unimportant was the length of service.

7.7.3 Distributed Random Forest (DRF)

Random Forest is known for its resilient embedded feature selection algorithm, which allows it to learn from high-dimensional data such as text data. Random Forest uses fully grown decision trees, which have low bias and high variance. It handles the error reduction task by reducing variance (Krauss, Do, & Huck 2017).

The model creates random uncorrelated decision trees which maximize the accuracy by reducing the variance, however the model is compromised by the bias/variance trade off, the algorithm cannot reduce the bias. The bias for the distributed random forest is slightly greater than the bias of an individual tree (Rai 2017). The multinomial model for distributed random forest resulted in 50 trees. In addition, the number of internal trees was 150, with a minimum number of 16 leaves. The maximum tree depth was 19, while the maximum number of leaves was 54.

TABLE 7.6 VARIABLE IMPORTANCE FOR RANDOM FOREST



	Variable	Relative Importance
1	Address.Line.4	3220.136
2	VIP.Status	2951.513
3	Customer.Age	1617.586
4	Customer.in.Collections. Indicator	1080.171
5	Spender.Status	820.2767
6	Customer.Contract.Length..in.Mont hs.	741.7331
7	AlarmCost	721.4320
8	Tariff.Text.Option	715.7268
9	KPIAlarms	676.63739
10	Approx.Location.County	575.7313
11	Total.Paid.Amount.of.Invoices	347.8585
12	Care.Call.Duration.Seconds	295.7986
13	Tariff.Voice.Option.min	242.3698
14	Customer.Network.Status	207.5635
15	Customer.Age.Group	204.584
16	Customer.Gender	182.3977
17	Total.Number.of.Invoices	178.3295
18	Sale.Time.of.Day	46.822319
19	Tariff.Gb.Data.Option	44.40408
20	Sale.Date.Day	40.48204
21	Customer.Length.of.Service	14.77907

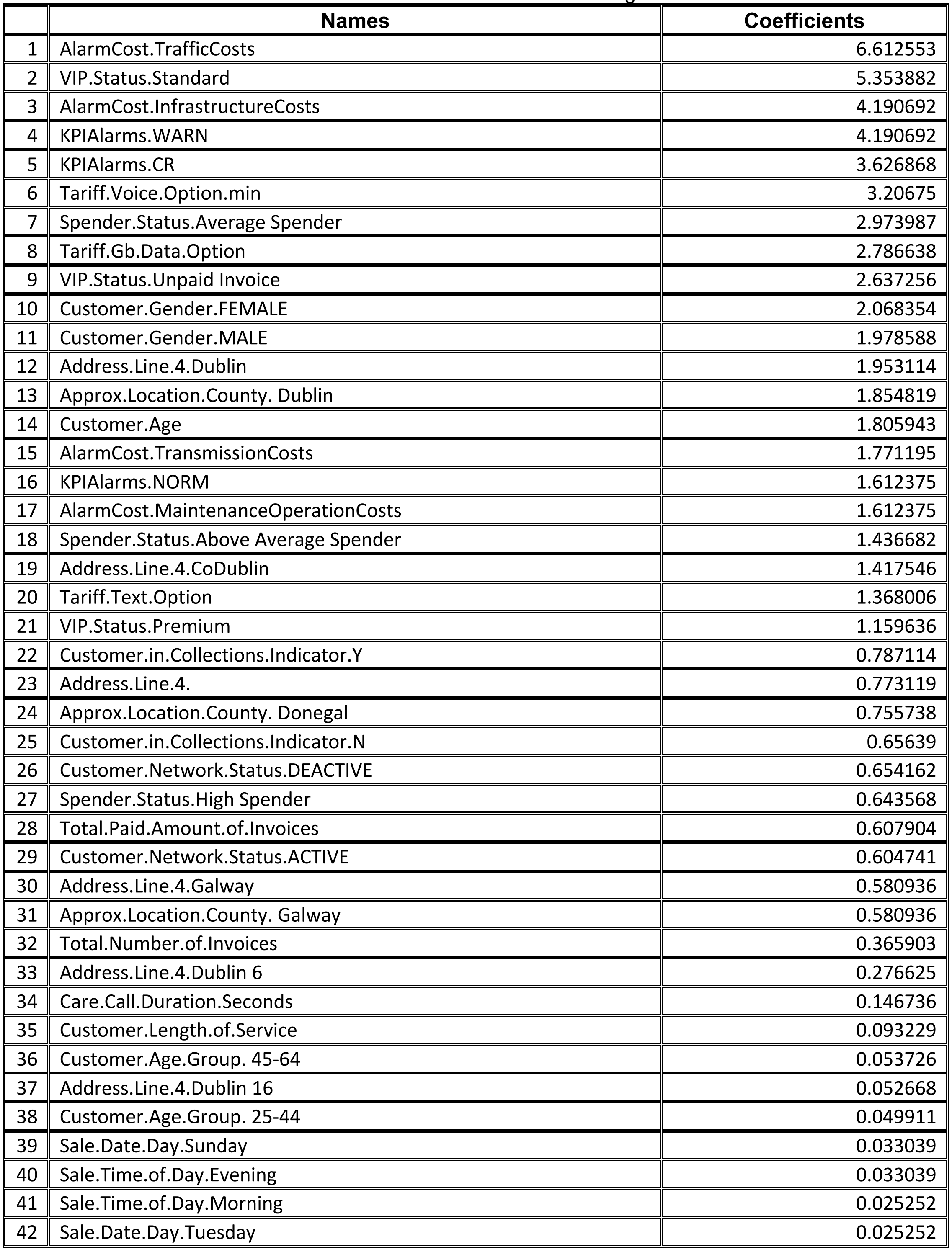
The variables considered significant in the model were location, VIP status, customer age and customers in collection indicator, these were determined by calculating their relative importance in relation to each other. This occurred, based on variable selection during the splitting of the nodes in the process of the tree construction, as well as the improvement in the squared error across all the trees.

Cross validation metrics were performed with a K-fold =5 where the squared error across all the trees was 0.7 and mean per class error was 0.20 and mean per class accuracy was 0.79.

7.7.4 General Linear model

In the generalized linear model (GLM) used in this study, each outcome of the dependent variable that was used to define the priority criteria was assumed to have been generated from a particular distribution of the exponential distribution's family (Kaur &. Batra 2017). In the experiment that was performed, a multinomial model GLM was presented with an alpha of -0.5 and a lambda of 0.0542 with 81 predictors and 23 active predictors.

TABLE 7.7 VARIABLE IMPORTANCE FOR glm model



	Names	Coefficients
1	AlarmCost.TrafficCosts	6.612553
2	VIP.Status.Standard	5.353882
3	AlarmCost.InfrastructureCosts	4.190692
4	KPIAlarms.WARN	4.190692
5	KPIAlarms.CR	3.626868
6	Tariff.Voice.Option.min	3.20675
7	Spender.Status.Average Spender	2.973987
8	Tariff.Gb.Data.Option	2.786638
9	VIP.Status.Unpaid Invoice	2.637256
10	Customer.Gender.FEMALE	2.068354
11	Customer.Gender.MALE	1.978588
12	Address.Line.4.Dublin	1.953114
13	Approx.Location.County. Dublin	1.854819
14	Customer.Age	1.805943
15	AlarmCost.TransmissionCosts	1.771195
16	KPIAlarms.NORM	1.612375
17	AlarmCost.MaintenanceOperationCosts	1.612375
18	Spender.Status.Above Average Spender	1.436682
19	Address.Line.4.CoDublin	1.417546
20	Tariff.Text.Option	1.368006
21	VIP.Status.Premium	1.159636
22	Customer.in.Collections.Indicator.Y	0.787114
23	Address.Line.4.	0.773119
24	Approx.Location.County. Donegal	0.755738
25	Customer.in.Collections.Indicator.N	0.65639
26	Customer.Network.Status.DEACTIVE	0.654162
27	Spender.Status.High Spender	0.643568
28	Total.Paid.Amount.of.Invoices	0.607904
29	Customer.Network.Status.ACTIVE	0.604741
30	Address.Line.4.Galway	0.580936
31	Approx.Location.County. Galway	0.580936
32	Total.Number.of.Invoices	0.365903
33	Address.Line.4.Dublin 6	0.276625
34	Care.Call.Duration.Seconds	0.146736
35	Customer.Length.of.Service	0.093229
36	Customer.Age.Group. 45-64	0.053726
37	Address.Line.4.Dublin 16	0.052668
38	Customer.Age.Group. 25-44	0.049911
39	Sale.Date.Day.Sunday	0.033039
40	Sale.Time.of.Day.Evening	0.033039
41	Sale.Time.of.Day.Morning	0.025252
42	Sale.Date.Day.Tuesday	0.025252

42 out of 44 variables were found to be related with the priority classes. The contract length (in months), whether the customers located in the county area of Dublin and purchases made in the Morning or on Tuesday were found to not be correlated with the priority classification. Based on the findings, those variables were considered to be unimportant in the model. It is important to note that GLM was the weaker model in this experiment because the results showed that there was lower accuracy and high levels of errors.

7.7.5 Deep Learning Model

Deep learning algorithms consist of layer by layer processing of features with a cascading hierarchy structure in which the information that is processed in one layer is fed to the next layer for further processing. The deep learning pre-processes the data, in order to standardize and normalise their range for compatibility with the activation functions.

Deep learning follows the model of multi-layer, feedforward neural networks for making predictions. The deep learning classifier that was used as a learner for the stacked ensemble model consisted of a purely supervised training protocol. The model was a 4-class classification model with 3,665 training samples and 46403 weights/biases with a mean bias of 0.47 and a bias rms of 0.022.

TABLE 7.8 VARIABLE IMPORTANCE FOR DEEP LEARNING MODEL

	Variable	Relative importance
1	Customer.in.Collections.Indicator.Y	1
2	Approx.Location.County. Dublin	0.997282803
3	AlarmCost.TrafficCosts	0.974452853
4	KPIAlarms.CR	0.972006798
5	Approx.Location.County. Donegal	0.970959723
6	Address.Line	0.969272912
7	Sale.Date.Day.Sunday	0.961897492
8	KPIAlarms.NORM	0.951708794
9	VIP.Status.Unpaid Invoice	0.950411737
10	AlarmCost.TransmissionCosts	0.948345006
11	Customer.Age.Group. 45-64	0.939581275
12	Customer.Gender.FEMALE	0.939478397
13	AlarmCost.InfrastructureCosts	0.936883628
14	Customer.Network.Status.ACTIVE	0.933715165
15	VIP.Status.Standard	0.93163079
16	Customer.Network.Status.DEACTIVE	0.929579794
17	Tariff.Text.Option	0.924869895
18	Spender.Status.High Spender	0.919214487
19	Address.Line.4.Dublin 16	0.918741107
20	Customer.Age	0.918459415
21	VIP.Status.Premium	0.918280005
22	KPIAlarms.WARN	0.914853036

23	Customer.Gender.MALE	0.911670506
24	Approx.Location.County. Galway	0.908834398
25	Tariff.Gb.Data.Option	0.9074471
26	Address.Line.4.CoDublin	0.903406203
27	Customer.in.Collections.Indicator.N	0.902760684
28	Address.Line.4.Dublin	0.886724174
29	Customer.Length.of.Service	0.886239588
30	Tariff.Voice.Option.min	0.88032943
31	Total.Paid.Amount.of.Invoices	0.879826069
32	Address.Line.4.Co Dublin	0.879175961
33	Customer.Age.Group. 25-44	0.878193915
34	Sale.Time.of.Day.Morning	0.877585948
35	Address.Line.4.Galway	0.873166621
36	Care.Call.Duration.Seconds	0.870609581
37	Address.Line.4.Dublin 6	0.866213679
38	Spender.Status.Average Spender	0.86337626
39	Spender.Status.Above Average Spender	0.860979497
40	AlarmCost.MaintenanceOperationCosts	0.859393954
41	Total.Number.of.Invoices	0.854149342
42	Sale.Date.Day.Tuesday	0.845640719
43	Customer.Contract.Length..in.Months.	0.825567245
44	Sale.Time.of.Day.Evening	0.815271974

The results in table 7.8 show that the main variables with a high level of influence on priority rules are customers in collections, located in Dublin or Donegal, the cost for traffic issues and the network faults that are defined as “Critical” based on our previous network optimisation scheme shown in chapter 4. This was followed by the customers that purchased company’s services or products on Sunday that according to the output are valuable customers and need to be prioritised. It can be seen that the deep learning results show that a significant number of VIP customers had missed payments and they were in collections. It is noted that the deep learner model searches the selected sample further while identifying the importance of a specific class in a categorical attribute. For example, a specific age group (45-64) of customers or a specific sales day or network cost, or an individual network fault.

7.8 Super Learner Performance

The super learner algorithm is used to identify the optimal combination of base learners and their variables for predicting an efficient end to end optimisation super-framework. The information presented below shows the tasks that were involved in training and testing the Super Learner ensemble.

Model 1: Super Learner Stacked ensemble using four base Classifiers

1. Set up the ensemble using the models

2. Specify the list L of the base algorithm <-list (glm,gbm,rf, dl)

3. Spefify the metalearner < - “gbm”

4. Select the family = ”multinomial”

5. Train the H stacked ensemble

```
stack <- h20.stackedEnsemble(x = x,
           y = "PriorityClasses",
          training_frame = train,
          base_models = glm, gbm, rf, dl,
          metalearner_algorithm = metalearner “gbm”)
```

The ensemble algorithm used in this study was comprised of four models: gradient boost machine, random forest, deep learning and general linear model. The metal learner that was used in the first experiment was the Gradient Boost Machine.

Model 2: Super learner stacked ensemble using three base classifiers:

The results signified that the classification accuracy for the GLM model was weaker than the other learners, to test the influence of this learner on the overall prediction accuracy, the next step in the experiment was to run the ensemble stacked algorithm again removing the GLM model.

Below is the stacked learner pseudocode for the gradient boost machine, random forest and deep learning models. The meta- leaner used for the second model was the GBM.

```
models <- list (gbm, rf, dl)

metalearner <- "gbm"

family="multinomial"

stack <- h20.stackedEnsemble(x = x,

            y = "PriorityClasses",

            training_frame = train,

            base_models = models,

            metalearner_algorithm = metalearner)

            Compute test set performance:

        perf <- h20.performance(stack, newdata = test)
```

The output of the second ensemble model using three classifiers and the GBM model as a meta-learner revealed that the deep learning algorithm was more efficient. Once the algorithms were run, cross-validation was used to generate out-of-sample predictions for the entire training set. It is important to note that the number of folds could have impacted the degree of under and over-fitting of the algorithm, as well as runtime because with a higher number of folds, each fold contains a larger portion of the training data to train on leading to a better fit. In this study, five folds were used to separate the data into five different learning sets, accordingly, research shows a good balance of fit and runtime can be achieved with three folds (Ibrahim, Siraj and Din. 2017).

7.8.1 Results of the stacked learners

In the end, deep learner showed higher accuracy and better overall results shown on Table 7.9, so it was chosen to run as a metalearner for the second ensemble model.

TABLE 7.9 BASE LEARNERS RESULTS

Cross Validation Metrics Summary for Base Learners				
CV-5 Folds	**DFR**	**GBM**	**Deep Learning**	**GLM**
	Mean	**Mean**	**Mean**	**Mean**
Accuracy	0.809	0.807	0.870	0.780
Error	0.197	0.182	0.129	0.178
Error Count	14	11.9	9.2	15.2
Logloss	0.415	0.399	0.394	0.418
Max per class error	0.257	0.249	0.296	0.241
Mean per class accuracy	0.793	0.829	0.849	0.790
Mean per class error	0.207	0.170	0.151	0.208
MSE	0.133	0.119	0.101	0.129
R^2	0.789	0.810	0.842	0.851
RMSE	0.364	0.345	0.316	0.381

The DRF and the GBM were both decision trees classifiers that had very similar results. The DRF performed with the Boruta algorithm and showed higher accuracy than GBM, which had been used for boosting and had a lower rate of error and logloss than the DRF classifier. GLM showed poor performance with regards to accuracy and errors. The decision was made to remove the GLM for the last stacked ensemble model in order to raise the accuracy of the selected classification model.

Because the GLM model has a lesser performance than the other models with an accuracy of 0.78, the experiment was performed on the three models of Gradient Boost, Random Forest and Deep learning.

TABLE 7.10 STACKED MODELS RESULTS

Stacked Models using H20 Mulitnomial on Cross Validation Data K-folds = 5	**Super Learner Stacked Ensemble with Four Base Learners (GLM, GBM, DRF, DL)**	**Super Learner Stacked Ensemble with three Base Learners (GBM, DRF, DL)**
MSE	0.117	0.109
RMSE	0.319	0.331
Logloss	0.388	0.401
Mean per Class Error	0.174	0.131
Accuracy	0.801	0.905

Overall Deep learning performed the best. In addition, removing the General linear from the stacked model resulted in increased performance of the ensemble model. There was a significant difference in the results of the GLM and Deep learning of 10%, which was significant. According to the base learners accuracy the most important variables that may need to be considered for an efficient end to end optimization framework are the following:

TABLE 7.11 STACKED MODELS IMPORTANCE VARIABLES

Variable importance GBM Model		
	Variable	**Relative Importance**
1	Address.Line.4	1478.953979
2	KPIAlarms	146.0876617
3	AlarmCost	121.9902115
4	Tariff.Gb.Data.Option	58.40407181
5	Tariff.Voice.Option.min	38.90882111
	Variable importance Deep Learning Model	
1	Customer.in.Collections.Indicator.Y	1
2	Approx.Location.County. Dublin	0.997282803
3	AlarmCost.TrafficCosts	0.974452853
4	KPIAlarms.CR	0.972006798
5	Approx.Location.County. Donegal	0.970959723
Variable importance DRF Model		
1	Address.Line.4	3220.136963
2	VIP.Status	2951.513916
3	Customer.Age	1617.586426
4	Customer.in.Collections.Indicator	1080.171021
5	Spender.Status	820.2767334
Variable importance GLM Model		
1	AlarmCost.TrafficCosts	6.612553
2	VIP.Status.Standard	5.353882
3	AlarmCost.InfrastructureCosts	4.190692
4	KPIAlarms.WARN	4.190692
5	KPIAlarms.CR	3.626868

7.9 Summary

The super-framework proposed in this study has been designed using machine learning in network analytics as a way of predicting the optimisation priority scheme through the process of selecting an appropriate range of troubleshooting management data for the customers. This scheme has been tested on different datasets using network analytics and data mining techniques for a smart revenue scheme. This general framework can be applied in a variety of industries. In addition, the scheme can be customised for specific applications and purposes (Gupta, & Pathak 2014).

In this part of the thesis, it was proposed a priority scheme using clustering methods and neural networks to perform the selected data while we are testing different base learners using ensemble learning methods. The use of the algorithms clarifies the rules for the proposed class specific approaches when different priority classes completely define a revenue management scheme (Cheng,2016). This contribution of this work performed in this study is based on the ability to conduct simulations for decision support in revenue management. It should be noted that additional learners can be used to perform the stacked model. However, the decision was made not to use the General Linear Model for a fair comparison, but will be used in a future study. In fact, the goal is to utilise additional data from external data sources, such as social media feeds, application downloads, customer reviews as a way of including data regarding customer perceptions and preferences with regards to their overall satisfaction.

This will allow for increased ability to determine optimal financial portfolios, financial derivatives pricing and budget plans in mobile services. (Raitoharju et al. 2015)

Chapter 8. Conclusions

This chapter summarises the research carried out and assesses the most important contributions of this thesis. In addition, the difficulties encountered are described and used to motivate the suggestion of future avenues of work. The initial scope of this research is applying different data mining techniques to network troubleshooting and proposing an efficient system to optimise the network. Following the successful demonstration of machine learning algorithms to identify technical faults, cost analysis and mobile customers' segments, an optimisation framework has been applied to predict the cost that may be saved, and improvements to revenue that drive an end to end optimisation super –framework. This connects all the data and information available and facilitate the network management decisions.

Nowadays data analytics is vital for financial returns linked to big data projects (Bughin 2016).Establishing an automated end-to-end optimisation super-framework which connects a variety of data sources available, will enable mobile operators to engage in more accurate and rapid network troubleshooting, revenue analysis and customers' segmentation. This will free resources, which can tackle upcoming challenges for the mobile companies brought by the continuing evolution of data analytics. The findings of this thesis clearly show that data mining algorithms can be used to diagnose and detect network optimisation faults while tracking the KPI alarms.

Amongst troubleshooting tasks, diagnosis of faults is the most complex, costly and time-consuming. Within the analysis of the causes of network fault detection as well as the cost evaluation of the KPI alarms a cost-effective optimisation framework proposed in the second part of this thesis which indicates strong evidence that cost-sensitive meta-learning techniques can provide efficient optimisation solutions to drastically reduce the network optimisation costs. With the support of a cost sensitive neural network classification model this research demonstrated that network providers can efficiently address budgeting considerations and network planning efforts, and also define pricing strategies and evaluate optimisation costs. The cost optimisation framework has been successfully integrated with demographic and revenue data in the third part of the thesis in order to generate recommendations for customer segmentation and customers willingness to pay for mobile services. However, very few references can be found on automatic diagnosis and network cost management using machine learning techniques. Hence, this thesis has drawn inspiration from other application domains where diagnosis is also required.

The first steps in automating troubleshooting in the radio access network of cellular networks has focused on network performance improvement and on fault detection. Hence, thanks to data mining algorithms such as decision trees, Bayes and neural network that show an accuracy of , efficient network performance was archived while fault detection and diagnosis are carried out more easily. Therefore, advanced machine learning techniques such as meta-learning and ensemble learning models have been proposed that will enable network specialists to raise the accuracy of network predictions up to 90% based on the findings of this thesis while running and amalgamating groups of classification and regression models in order to choose the best combination of algorithms to be run together and support an end to end optimisation super-framework.

The end to end optimisation super-framework that was proposed in the last part of the thesis contributes to Quality of Service processes by increasing operational efficiency and network performance and facilitate the customers' segmentation process. The final results show an accuracy of 90% as an overall score including all the variables involved in the end to end optimisation super-framework. Thus, it is possible to obtain better performance from the existing network at a lower cost. The foundation of a well-performing network is the basic radio platform, as determined by the findings of this research as well as input from the literature. Issues such as the location of the sites as well as cost optimisation and the customer's chain have a major impact in telecommunication sector.

8.1 Recommendations

Wireless networks have become more complex with increasing numbers of users and network devices. (M.Elkhord et al.2016). This only adds to the burden of finding automatic methods of fault diagnosis that are efficient, both in terms of accuracy of output and actual usage. The proposed network optimisation system based on data mining techniques, represents a systematic approach for anomaly detection that is based on a KPI data analysis model for those who are responsible for the monitoring and optimisation of mobile networks.

The variety of data mining models used in this study provide a strong means by which to perform network diagnosis efficiently and quickly. The rules and decision trees set up that are extracted by the data mining process simplify the optimisation process.

A weakness of the system as this point is that the decision trees do not always show the causes of the symptoms that are creating inefficiency in a network. However, even with this weakness, decision trees provide the rules of the alarms, and have a high rate of accuracy in the information that is provided.

Data mining algorithms have been demonstrated to be capable of showing where causes of network congestion occur, as well as the symptoms that can create the lack of network optimisation so that engineers and operators can work to regain optimisation for end users (Rozaki 2016). In addition, the results of this study have shown that the use of meta learners such as meta cost and ensemble learning algorithms algorithms provide a higher level of accuracy in determining network, revenue and cost optimisation decisions.

Furthermore, meta learning methods applied in end to end optimization models retains the ability to identify patterns and . In the end, the results presented in this thesis allow for the conclusion that the proposed method has produced very promising results in the classification of multi-class optimisation faults.

Data mining for cost analysis in this thesis is based upon meta-cost classification models that provide the ability to determine the bandwidth cost that can be achieved through network troubleshooting management. In this way, network performance and cost are optimised at the same time. Meta-cost classification techniques are effective and accepted tools for pattern recognition and categorisation but require careful application using the appropriate learner to perform the models. The cost sensitive classifiers evaluate the cost values and they are able to obtain a high level of accuracy while supplying an interpretable model for the end cost evaluation of optimisation faults.

Minimisation of the cost function makes it possible for meta learning algorithms to both forecast and learn the optimal cost to facilitate revenue decisions. Then the output of optimisation process becomes more realistic because specific KPIs are selected based on the selected costs and customers' wiliness to pay in question. Finally the aspects of finding the optimal network service levels according to the mobile customers preferences and willingness to pay become part of the ensemble learning classification methods. This study evaluated a number of machine learning methods to classify telecommunication data and compare them using a stacked generalization approach. Super learning applications using H20 platform used in this thesis to perform a multinomial classifications and improve technical and financial decisions.

8.2 Future Work

The results of this study are not only important in terms of what can be concluded from them about an applicable method of network fault diagnosis that was presented, but also the broader context of network management. The analysis of network fault diagnosis in this study, and most other studies, requires some knowledge of statistical analysis. The question that arises, however, is whether the people who are responsible for network troubleshooting in practice have the statistical analysis and the data analysis tools that are needed to efficiently perform the same procedures and analysis conducted.

Another area for future research might be to examine the business decisions that underlie network optimisation efforts on the part of network providers. While the focus of so much of this area of research is on the techniques and methods, the reality is that there are important business decisions that dictate which techniques and methods are used, as well as how they are used (Bresfelean, 2009).It would be useful to understand the decision-making process and the business realities that determine the specific methods that companies that operate wireless networks use to perform fault diagnosis.

An additional direction for future research relates to the development of quality of service management that can be utilised in an end-to-end approach in overcoming the problem of parameter sensitivity for solutions that may be less integrated. This approach also suffers from the possibility that the model will be over-fitted to the training data. Thus, it has been suggested that future research be conducted into the possibility of utilising association rules for an end-to-end approach to QoS management.

It has been also suggested that further research should be aimed at exploring text mining in categorising customers' preferences and behavioural segmentation. Hence, the contextual features for different behaviour using the approach and find that high precision predictive models can be built for most mobile users (Allen et al. 2008, Boratto et al, 2016). Clustering and regression modelling remain the subject of active research in many fields such as pattern recognition, machine learning and statistics.

Text mining adds to data mining complications of large computation requirements on the relevant clustering algorithms. There algorithms available that meet such requirements. They can be applied in real-life data mining problems. These algorithms are subject of new research.

References

Abellan, J., Mantas C.J., Gasteliano G. J., Moral-Garcian S., 2018. Increasing diversity in random forest learning algorithm via imprecise probabilities. Expert systems with applications, Science Direct, 97, pp. 228-243.

Adler, M., Sitaraman R., K., & Venkataramani H. 2011. Algorithms for optimizing the bandwidth cost of content delivery. *Computer Networks*, 55(18), pp. 4007-4020.

Allen, S., M., Whitaker, R., M., & Hurley, S., 2008. Personalised subscription pricing for optimised wireless mesh network deployment. *Computer Networks,* 52(11), pp. 2172-2188.

Ali, M., Shehzad, A., & Akram, M., A., 2010. Radio access network audit & optimization in gsm (Radio access network quality improvement techniques). *International Journal of Engineering & Technology,* Vol. 10, pp. 55-58.

Altman, Z., Skehill, R., Barco, R., Moltsen, L., Brennan, R., Samhat, A. 2006. The Celtic Gandalf framework. In *Proceedings of the IEEE Mediterranean Electrotechnical Conference MELECON'06*. Benalmadena, Spain.

André, R., Denis, A., Pérez, C., Priol, T. 2003. Component-Based Software Infrastructure for the Grid Computing**.** In *17th International Parallel and Distributed Processing Symposium (IPDPS 2003)*, Nice, France, April 2003. IEEE Computer Society.

Ashwinkumar.U.M & Anandakumar.K.R 2013. Data Preparation by CFS an Essential approach for decision making using C 4.5 for Medical data miningfindings. In ACCT' 13 *Proceedings of the 2013 Third International Conference on Advanced Computing & Communication Technologies.* pp 77-85.

Arora S., Candel, A., Lanford, J., LeDell E., and Parmer V., 2016. Deep Learning with H2O, *Mountain View CA H20ai 2015*

Arnold, R.A., 2008. *Economics*. London: Cengage.

Arumugam S., Bagga J., Lowell W. Beineke, Panda B.S. 2016. Theoretical Computer Science and Discrete Mathematics. *First International Conference, ICTCSDM 2016, Krishnankoil, India, Revised Selected Papers.*

Asche V.C., Kim M., Brown A., Golden A., Laack A.T.,, Rosario J., Strother C., Totten Y.V., and Okuda Y., 2017. Communicating Value in Simulation: Cost–Benefit Analysis and Return on Investment *Consensus Conference.*

Banerjee A., 2013. Big Data & Advanced Analytics in Telecom: A multi-Billion Dollar Revenue Opportunity*. Huawei.*

Bangladesh Telecommunication Regulatory Commission 2014
Directives on Quality of Service (QoS) for Cellular Mobile Telecom Operators. [Online] Available at:
http://www.btrc.gov.bd/sites/default/files/news_files/Directive%20on%20QoS%20fo%20Cellular%20Mobile%20Telecom%20Operators.pdf

Barco, R., Díez, L., Wille, V., and Lázaro, P., 2009. Automatic diagnosis of mobile communications networks under imprecise parameters. In *Expert Systems with Applications: An International Journal,* 36 (1). Pergamon Press: NY.

Barco, R., Wille, V., and Díez, L., 2005. System for automated diagnosis in cellular networks based on performance indicators. *European Transactions on Telecommunications,* 16 (5), pp. 399–409.

Barco, R., Wille, V., Díez, L., and Lázaro, P., 2006. Comparison of probabilistic models used for diagnosis in cellular networks. In *Proceedings of the IEEE vehicular technology conference (VTC'06)*. Melbourne, Australia.

Barreto, G. A., Mota, J. C. M., Souza, L. G. M., Frota, R. A., Aguayo, L., Yamamoto, J. S., & Macedo, P. E. 2004. Competitive neural networks for fault detection and diagnosis in 3G cellular systems. In *Telecommunications and Networking-ICT 2004* (pp.207-213). Springer Berlin Heidelberg.

Bayram, S., & Gezici, S., 2012. Stochastic resonance in binary composite hypothesis-testing problems I the Neyman-Pearson framework. *Digital Signal Processing* 22, 391-406.

Beyer, J., & Mao, L., 2012. Analysis of the multi-cell correlation of the slow fading from UMTS measurements and its impact on radio network planning. In *Vehicular Technology Conference (VTC)*. New York, 2012. IEEE.

Beretka, S., F., & Varga, E., D., 2013. Efficient encoding of customer class load profiles. *In 2013 Africon*, 22, pp. 1-5.

Bhowmik, B., Roy, S., Guha Thakurta, P. K., & Sarkar, A. 2011. Priority Based Hard Handoff Management Scheme for Minimizing Congestion Control in Single Traffic Wireless Mobile Networks. *International Journal of Advancements in Technology*, *2*(1), pp.90-99.

Breidert, C., 2006. *Estimation of willingness-to-pay*. Berlin: Springer.

Brett, L.,2016. Machine Learning with R. *Packt Publishing Ltd*. Second edition

Bresfelean, V., 2009. Analysis and Prediction of Student's Behavior Using Decision Tress in WEKA Environment. *Proceedings of the ITI 2007 29th International conference on Information Technology Interfaces,* pp.51-56.

Brefelean V., P., 2009. Data mining in continuing education. *INTED, the International Technology, Development and Education Development Valencia, Spain*.

Bughin J.2016. Reaping the benefits of big data in telecom *Journal of Big Data,* DOI 10.1186/s40537-016-0048-1

Cao P., Zhao D., and Zaiane, O. 2013. Measure optimized cost-sensitive neural network ensemble for multiclass imbalance data learning. 2013 13th *International Conference on Hybrid Intelligent Systems* pp. 35-40.

Chai X., Deng L., Yang Q., and Ling, C., X., 2004. Test-Cost Sensitive Naive Bayes Classification. *Fourth IEEE International Conference on Data Mining (ICDM'04)* pp. 51-58.

Chandrasekhar, V., & Andrews, J. 2009. Uplink capacity and interference avoidance for two-tier femtocell networks. *Wireless Communications, IEEE Transactions on*, *8*(7), pp.3498-3509.Charnay, C., Lanchiche, N., & Braud, A., 2013. Pairwise optimization of Bayesian classifiers for multi-class cost-sensitive learning. *Wireless Communications, 2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, pp. 499-505.

Chu, K., Lin, F.,Y., & Wang, C., 2011. Cost optimisation of integrated network planning based on adaptive sectorization in hybrid F/CDMA telecommunications system via Lagrangian relaxation. *Expert Systems with Applications*, 38, pp.14819-31.

Churchill, S., 2012. *South Korea completes nationwide LTE coverage*. [Online] Available at: http://www.dailywireless.org/2012/06/21/south-korea-completes-nationwide-lte-coverage.

Cisco, 2010. Radio Resource Management Concepts *White paper*, Chapter 2. Available at: https://www.cisco.com/c/en/us/td/docs/wireless/controller/technotes/8-3/b_RRM_White_Paper/b_RRM_White_Paper_chapter_01.html

Cormio, C., & Chowdhury, K., R., 2010. Common control channel design for cognitive radio wireless ad hoc networks using adaptive frequency hopping. *Ad Hoc Networks*, *8*(4), pp.430-438.

Cui H., Wang J., Sun F., Liu Y. & Chen K-C., 2014. Streaming media traffic characteristics analysis in mobile internet, 17 *International Symposium on Wireless Personal Multimedia Communications*, pp. 1-5.

Cusani, R., Inzerilli, T., & Valentini, L., 2007. Network monitoring and performance evaluation in a 3.5G network. *Computer Networks*, 51, pp. 4412-4420.

Domingos, P., 1999. MetaCost: A general method for making classifiers cost-sensitive. In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining *ACM* pp.155-164.

Deljac, Z., Mostak, R., & Stjepanovic, T., 2010. The use of Bayesian networks in recognition of faults causes in the BB networks *Computer Networks, MIPRO, 2010 Proceedings of the 33rd International Convention* pp.771-775.

Deloitte MCS Limited, 2016. Digital transformation for telecom operators. Adapting to a customer centric, mobile first world*, The Creative Studio at Deloitte J5613.*

Demars, M., Fourestié, D., Mourlon, J., Picard, J., & Renou, S., 2005. 3G Network QoS Estimation in a Multi Service Context. 2005 *IEEE 61st Vehicular Technology Conference, 2005. VTC 2005-Spring*. , 3, 182-1-1824.

Deng, L., Kawamura, T., Taoka, H., & Sawahashi, M. 2011. Combined effect of transmit diversity and frequency hopping for DFT-precoded OFDMA in uplink frequency-selective fading channels. In *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd* (pp.1-5). IEEE.

ECD, 2018. *Broadband statistics update*. [Online] Available at: http://www.oecd.org/sti/broadband/broadband-statistics-update.htm.

Elkhodr M., Shahrestani S.,Cheung H., 2016. The Internet of Things. New Interoperability, Management and Security Challenges. *International journal of Network Security and Its Applications (IJNSA)*. Vol.8 No.2

Esteves, R. B., Cerqueira, A., 2017. Behavior-based pricing under imperfectly informed customers. *Information Economics and Policy*, vol. 40, pp. 60-70.

Farhadloo, R., Patterson, A., and Rolland, E., 2015. Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, vol. 90, pp. 1–11.

Ferri-Ramirez, C., Flach, P., & Hernandez-Orallo, J., 2002. Multi-dimensional roc analysis with decision trees. *Technical Report* pp. 1-36.

Flood, J. 1997. Telecommunication Networks. *IEE Telecommunications Series, No 36*, Institution of Engineering and Technology.

Gabel, D., 1991. An application of stand-along costs to the telecommunications industry. *Telecommunications Policy*, 15(1), pp. 75-84.

Gallego, G., & Van Ryzin, G 1994. Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons*. Management Science,* 40(8), p.p. 999-1020.

Gandomi, A., & Haider, M. Beyond 2015. The hype: big data concepts, methods, and analytics, *International Journal of Information Management.* pp. 35(2), 137-144.

Gangyi, J., Yun, Z., Mei., K., & Yu, T., 2009. A Low-complexity quantization for the H.265/AVC", *Journal of Real-Time Image Processing,* Springer, 4 (1) 3-12.

Gao Y. & Wang J. 2011. Active learning method of Bayesian networks classifier based on cost-sensitive sampling, *2011 IEEE International Conference on Computer Science and Automation Engineering.* pp. 233-236.

Genuer R., Poggi J. M., Tuleau - Malot C., 2017. Random Forests for Big Data, *2Big Data Research Science Direct.* 9 pp. 28-46.

Ghosh, S.C., Whitaker, M. & Hurley, S.M.A.S., 2009. Service coverage bounds through efficient load approximation in UMTS network planning. In *First International Conference on Networks and Communications (NETCOM '09)*. New York, 2009. IEEE.

Guo L., and Boukir., S., 2017. Building an ensemble classifier using ensemble margin. Application to image classification, *IEEE International Conference on Image Processing*, pp. 4492-4496

Gupta, L., Salman T., Zolanvari M., Erbad A., Jain R., 2019. Fault And Performance Management In Multi-Cloud Virtual Network Services Using AI: A Tutorial And A Case study,*Computer Networks*, V. 165,106950 .

Gupta, R., Pathak, C., 2014. A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing, *Procedia Computer Science*, 36, pp. 599-605.

Ramzi A. H., Dimishkieh, M.,and Masud M. 2015. An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare. *International Journal of Distributed Sensor Networks.* Volume 2015, Article ID 615740, 11 pages.

Haider, B., Zafrullah M., & Islam, M., K., 2009. Radio frequency optimization & QoS evaluation in operational GSM network. *Proceedings of the World Congress on Engineering and Computer Science,* pp. 393-398.

Hammed Lasis and Aderinkola, B. Fatimah 2018. Comparative analysis of traffic congestion of mobile communication networks in Osogbo, Nigeria. *Journal of engineering and Manufacturing Technology* JEMT 6 25-32.

Han S. S., Azad D.T., Suarez A.P., Ratliff K.J., 2019. A machine learning approach for predictive models of adverse events following spine surgery. *The spine Journal*. Vol 19, pp. 1772−1781.

Haizan, W., Nor, Mohamed, W., Salleh, M., N., M., & Omarv, A., H., 2012 A comparative study of reduced error pruning method in decision tree algorithms, *IEEE International Conference on Control System, Computing and Engineering*, pp. 392-397.

He, Z., Deng, S., Xu, X., & Huang, J. Z. 2008. Mining class outlier: Algorithms, Concepts and applications in CRM. *The Expert Systems with the application*, 27, 681–697.

Hu, Q., Wei, Y. and Xia, Y. 2010. Revenue management for a supply chain with two streams of customers. *1st ed. ELSEVIER*, pp.582-598.

Huang, K. Kechadi,T ,Buckley B., 2013 Customer churn prediction in telecommunications*, Expert Systems with Applications,* vol. 39, no. 1, pp. 1414–1425.

Huang, K. and Liang, Y. 2011. A dynamic programming algorithm based on expected revenue approximation for the network revenue management problem. *ELSEVIER*, pp.333-341.

Ibrahim, A., J., M. Siraj M., and Din., M., M. 2017. Ensemble classifiers for spam review detection. *IEEE Conference on Application, Information and Network Security,* pp. 130-134.

ICAP, 2009. *Customer Profiling and 3G Optimization: Updating Marketing Strategies in the Telecommunication Industry*. London: ICAP.

Iliya S., Goodyer E., Gow J., Shell J. and Gongora M. (2015). Applications of artificial neural network and support vector regression in cognitive radio networks for RF power prediction using compact differential evolution algorithm. *Proceedings of the Federated Conference on Computer Science and Information Systems,* Vol.5, pp. 55-66.

Jagadesh, B.L., Kullayamma, I. Vivek, Naresh, 2011. Handover analysis. *International Journal of Engineering Research and Applications (IJERA)* V.1, Issue 2, pp.287-291.

Jain, R. K., Katiyar, S., & Agrawal, N. K. 2011. Hierarchical Cellular Structures in High-Capacity Cellular Communication Systems. *arXiv preprint arXiv:1110.2627*.

Johnson, J.L. (2003). *Probability and statistics for computer science*. Wiley, New Jersey, USA.

Jaudet, M., Iqbal, N., Hussain, A., & Sharif, C., 2005. Temporal classification for fault-prediction in a real-world telecommunications network. *2005 International Conference on Emerging Technologies,* pp.209-214.

Ji, R., Gao, J., & Xie, G., Flowers, G.T. & Chen C., 2014. A fault diagnosis method of communications connectors in wireless receiver front-end circuits. *60th Holm Conference on Electrical Contacts,* pp.1-6.

Johnson, J.L. 2003. *Probability and statistics for computer science*. Wiley, New Jersey, USA. Katzela, I. and Schwartz, M. 1995. Schemes for fault identification in communication networks. *IEEE/ACM Transactions on Networking*, 3 (6), pp. 753–764.

Kalra M., Lal M., 2016. Data Mining of Heterogeneous Data with Research Challenges, *Symposium on Colossal Data Analysis and Networking (CDAN), pp. 1-6.*

Keramati A, Jafari-Marandi R., Aliannejadi M., Ahmadian I., Mozaffari M., and. Abbasi U, 2014. Improved churn prediction in telecommunication industry using data mining techniques, *Applied Soft Computing, vol. 24, pp. 994–1012.*

Khan A, Jamwal, S. , and Sepehri, M. 2010 Applying Data Mining to Customer Churn Prediction in an Internet Service Provider, *International Journal of Computer Applications, vol. 9, no. 7 pp. 8-14.*

Khanafer, R., Moltsen, L., Dubreil, H., Altman, Z., & Barco, R. 2006. A Bayesian approach for automated troubleshooting for UMTS networks. In *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on* pp.1-5. IEEE.

Kihara, T., Tateishi N., & Seto, S., 2011. Evaluation of network fault-detection method based on anomaly detection with matrix eigenvector. *2011 13th Asia-Pacific Network Operations and Management Symposium,* pp. 1-7.

Kim, J., Choi, K., Kim, G., & Suh, Y., 2012. Classification cost: An empirical comparison among traditional classifier cost-sensitive classifier, and meta-cost. *Expert Systems with Applications*, 39, pp. 4013-4019.

King N.J., & Forder J., 2016. Data analytics and consumer profiling: Finding appropriate privacy principles for discovered data. *Computer Law and Security Review.* V. 32 (5), pp 696-714

Krakovsky R., & Forgac R., 2011. Neural network approach to multidimensional data classification via clustering, *IEEE 9th International Symposium on Intelligent Systems and Informatics,* pp. 169-174.

Krauss C, Do X.A., Huck N, 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500*, European Journal of Operational Research,* pp. 689-702.

Kryftis Y, Mastorakis G, Mavromoustakis C, Batalla J. M., Kormentzas G., 2016. Efficient Entertainment Services Provision over a Novel Network Architecture*, IEEE Communications Society* pp. 14-21.

Kumar, S., Siraj M. M. and Singh., R. 2018. Comparative analysis of ensemble classifiers for sentiment analysis and opinion mining. *3rd International Conference on Advances in Computing, Communication & Automation, (ICACCA), IEEE publisher*

Kursa Y., Rudnicki R., W., 2010. Feature selection with the Boruta Package *, Journal of Statistical Software,* vol 36, pp. 1-13.

Kwok TY., & Yeung DY, 1996. Bayesian regularization in constructive neural networks. In: von der Malsburg C., von Seelen W., Vorbrüggen J.C., Sendhoff B. (eds) Artificial Neural Networks, *ICANN 96. ICANN 1996. Lecture Notes in Computer Science, vol 1112. Springer, Berlin, Heidelbergon*

Lariviere M., A., and E., L., Porteus, E., L., 2001. Selling to the Newsvendor: An Analysis of Price-Only Contracts. *Manufacturing & Service Operations Management,3* (4), 293-305

Lee, H., Lee Y., Cho, K. Im K., and Kim Y.S. 2004. Mining Churning behaviors and developing retention strategies based on a partial least squares (PLS) model. *Decision Support Systems,* Vol 52 no 1 pp. 207-216.

Lescuyer, P., 2004. UMTS: Origins, architecture and the standard. *Berlin: Springer.*

Lehtimäki, P., & Raivio, K. 2005. A knowledge-based model for analyzing GSM network performance. In *Advances in Intelligent Data Analysis VI* (pp.204-215). Springer Berlin Heidelberg.

Li, J., & Manikopoulos, C., 2002. Network fault detection: Classifier training method for anomaly fault detection in a production network using test network information. *27th Annual IEEE Conference on Local Computer Networks*. pp. 473-482.

Lopez, V., Del Rio, S., Benitez, J., M., & Herrera f., 2015. Cost-sensitive linguistic fuzzy role based classification systems under the Map Reduce framework for imbalanced big data, *Fuzzy Sets and Systems.* Vol 258C., pp. 5-38.

Loukeris, N., 2008. Radial basis functions networks to hybrid neuro-genetic RBFNs in financial evaluation of corporations. *12 WSEAS International Conference on Computers* pp.812-819.

Liu P. & Li Z. 2009, Study on a new clustering optimisation algorithm and its application in network faults analysis, *International Conference on Information Technology and Computer Science*, pp. 134-137.

Lunn P. & Lyons S. 2018, Consumer switching intentions for telecom services: evidence from Ireland, *Heliyon 4, Elsevier Ltd.* Article no e00618

Krishnamoorthy, G., Ashok, P., and Tesar, D., 2014. A Simultaneous sensor and process fault detection and isolation in multiple-input-multiple-output systems. *IEEE Systems Journal Data Analysis.* pp. 1-15.

Malhotra P., and Dureja A., 2013. A survey of weight-based clustering algorithms in MANET, *Journal of Computer Engineering*. Vol. 9, pp. 34-40.

Mason, L., G., Juda, M., & Dziong, Z., 1997. A framework for bandwidth management in ATM networks-aggregate equivalent bandwidth estimation approach. Journal of Networking, IEEE/ACM Transactions, 5 (1). [Online] Available at: http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?tp=&arnumber=554728&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D554728.

Masmoudie, A., Zaghlache, D., & Tabbane, S., 2005. Resource and scheduling optimisation in HSDPA based UMTS networks. In *Proceedings of the World Wireless Congress (WWC 2005)*. New York, 2005. IEEE.

Michaelis, S., 2011. Balancing high-load scenarios with next cell predictions and mobility pattern recognition. In *MOBILITY 2011, The First International Conference on Mobile Services, Resources, and Users* pp.164-169.

Monacelli, L., and Francescangeli, R., 2011. Fault management for VoIP applications over wireless and wired NGN networks: an operational prospective. *IEEE* 36th *conference Local Computer Networks* pp.711-718*.*

Morales, D. R., and Wang, J., 2010. Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research: an operational prospective, 202(2)* pp.554-562*.*

Naimi A. I., and Balzer, L. B., 2018. Stacked generalization: An introduction to super learning. *European Journal of Epidemiology, Vol. 33, pp.* 459-464*.*

Nasreddine J and Hassan S. E. H. 2016. Interference Mitigation and Traffic Adaptation Using Cell Clustering For LTE-TDD Systems. *IEEE International Multidisciplinary Conference on Engineering Technology (IMCET).*

Musa S.,A.,2016, Performance Appraisal of Mobile Telecommunication Network in Dutse, Jigawa State, *Journal of Science, Technology & Education* (JOSTE); Vol. 4 (1),

Ng, G. and Ong, K. 2000. Using a qualitative probabilistic network to explain diagnostic reasoning in an expert system for chest pain diagnosis. *Computers in Cardiology conference.*

Nor Haizan W, Mohamed W, Salleh M. N. M., and Omar A. H. 2012. A comparative study of reduced error pruning method in decision tree algorithms, *IEEE International Conference on Control System, Computing and Engineering,* pp. 392-397.

Noughabi E. A. Z., Albadvi A, and Far B. H. 2015. How Can We Explore Patterns of Customer Segments' Structural Changes? A Sequential Rule Mining, *IEEE International Conference on Information Reuse and Integration, pp.273-280*

Okamoto, K., Higashi, M., Yokota, M., Nishikiori, R., Osaki, M., Yasunaga, T., & Takagi, T. 2006. New computer-intensive procedures for testing null hypotheses comparing two parameters approximately. *Chemometrics and Intelligent Laboratory Systems* 82, pp.66-74.

Oladimeji, M., O., Ghavami M., & Dudley, S., 2015. A new approach for event detection using k-means clustering and neural networks, *IEEE International Joint Conference on Neural Networks*, pp. 1-5.

Oseni M., O., & Pollitt G., M., 2017. The prospects for smart energy prices: Observations from 50 years of residential pricing for fixed line telecoms and electricity, *Renewable and Sustainable Energy Reviews*, Science Direct, *70* pp. 150-160.

Ouyang, Y., & Hosein Fallah, M., 2010. A performance analysis for UMS packet switched network based on multivariate KPIs. *Wireless Telecommunications Symposium (WTS),* pp.1-10.

Panda, M., & Padhy, S., P., 2009. Traffic analysis and optimization of GSM network. *IJSCI International Journal of Computer Science Issues,* Vol. 1 pp. 28-31.

Patil, B., Toshniwal D., & Joshi, R., 2009. Predicting burn patient survivability using decision tree in weka environment. *IEEE International Advance Computing Conference*, pp. 1353-1356.

Park, B., Won, Y., J., H., Yu, J., W-K., Hong, H-S Noh, & J., J., Lee, 2009. Fault detection in ip-based process control networks using data mining. *International Symposium on Integrated Network Management,* pp. 211-217.

Perez-Costa, X., Banchs, A., Noguera, J., & Sallent-Ribes, S., 2004. Optimal Radio Access Bearer Configuration for Voice over IP in 3G UMTS networks. *Available at* http://www.it.uc3m.es/banchs/papers/ew04.pdf.

Popoola S.I., Atayeroi, A.A., Faruk, N., 2018a. Received signal strength and local terrain profile data for radio network planning and optimization at GSM frequency bands. Data in Brief 16 (2018) 972–981.

Popoola S.I. Aderemi A. Atayero, Faruk N., Badejo J.A. 2018b Data on the key performance indicators for quality of service of GSM networks in Nigeria. *Data in Brief* 16 (2018) 914–928

Posnakides D., Mavromoustakis, C., Skourletopoulos, G., Mastorakis G, Pallis Mongay B.. J., 2016. Performance Analysis of a Rate-Adaptive Bandwidth Allocation Scheme in 5G Mobile Networks. *20th IEEE Symposium on Computers and Communications (ISCC 2015)*.

Püschel, T., Schryen, G., Hristova, D., Neumann, D., 2015. Revenue management for Cloud computing providers: Decision models for service admission control under non-probabilistic uncertainty. *European Journal of Operational Research,* 244(2), pp. 637-647.

Quinlan, R. J., 1993. Programs for machine learning. *Morgan Kaufmann Publishers, Inc*. Unites States of America.

Qureshi, A., Rehman, A., S., Qamar, A. M. and Kamal, A., 2013. Telecommunication subscribers' churn prediction model using machine learning. *In Proceedings of the 8th International Conference on Digital Information Management (ICDIM '13)* Vol 27, pp. 2458-2471.

Rahnema, M., 2008. UMTS network planning, optimisation, and inter-operation with GSM. *London: John Wiley and Sons.*

Rai, B. 2017. Feature selection and predictive modelling of housing data using random forest. *International Journal of Industrial and Systems Engineering, Vol.11,* pp. 940-944.

Raitoharju,J. Kiranyaz, S. Gabbouj, M., 2016. Training Radial Basis Function Neural Networks for Classification via Class-Specific Clustering. *IEEE Transactions on Neural Networks and Learning Systems .* pp 131-136.

Rodrigues, E.B., Cavalcanti, F.R.P. & Wänstedt, S., 2009. Congestion control for wireless cellular systems with applications to UMTS. In *Optimizing Wireless Communication Systems*. Berlin: Springer. Chapter 4 pp. 141-57.

Rohit A. and Suman 2012. Comparative analysis of classification algorithms on different datasets using WEKA. *International Journal of Computer Applications* Vol54(13) pp. 21-25.

Rosich, A., Voos, H. & Pan, L., 2014. Network design for distributed model-based to fault detection and isolation. *IEEE conference on Control Applications* pp. 1226-1231.

Rowe, M., 2013. Network Mining User Lifecycles from Online Community Platforms and their Application to Churn Prediction. *Data Mining (ICDM), 2013 IEEE 13th International Conference* pp. 637-646.

SAS 2015, The Connected Vehicle: Big Data, Big Opportunities, White paper.

Sánchez-González J., Sallent O., Perez-Romero J., 2008. A new methodology for RF failure detection in UMTS networks. In *Network Operations and Management Symposium 2008*. New York, 2008. IEEE.

Santoyo-González, A. Cervello-Pastor C., 2018. Latency-aware cost optimisation of the sevice infrastructure placement in 5G networks. *Journal of Networks and computer applications. V114 pp.29-37.*

Salmelin, J. & Metsälä, E., 2012. *Mobile backhaul*. London: John Wiley and Sons.

Sangani D., . Erickson K. and Al Hasan, M., 2017. Predicting zillow estimation error using linear regression and gradient boosting. IEEE 14 International Conference on Mobile Ad Hoc and Sensor Systems*, pp.530-534.*

Sawson, K., Aiken, L.S., Cohen, J., Cohen, P., & West, S.G, 2007. Applied multiple correlation/regression analysis for behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sevtnazarov, Sh.O., 2010. RF optimisation issues in GSM networks. In *4th International Conference on Application of Information and Communication Technologies*. New York, 2010. IEEE.

Shahzad W, S. Asad, Khan M. A., 2013, S.O., 2010. Feature subset selection using association rule mining and JRip classifier. *International Journal in Physical Sciences*. Vol. 1, pp. 28-31.

Sharma T., C., & M. Jain, M., 2013 WEKA approach for comparative study of classification algorithm. *International Journal of Advanced Research in Computer and Communication Engineering.* Vol. 2, pp. 1925-1931..

Skianis, C., (2013), Introducing automated procedures in 3g network planning and optimization, *The Journal of Systems and Software*, Vol. 86, pp. 1596-1602.

Srimani P.K. & Balaji K., A comparative study of different classifiers on search engine based educational data, *International Journal of Conceptions on Computing and Information Technology*, Vol 2, ISSN: 2345 – 9808, 2014.

Siomina, I., Varbrand, P. & Di, Y., 2006. Automated optimisation of service coverage and base station antenna configuration in UMTS networks. *IEEE Wireless Communications*, 13(6), pp.16-25.

Sladojevic, S., Culibrk, D. and Crnojevic, V. (2011). Predicting the Churn of Telecommunication Service Users using Open Source Data Mining Tools. *TELSIKS*.

Srimani, P., K., & Pikulík, T., 2014. GDPR principles in Data protection encourage pseudonymization through most popular and full-personalized devices - mobile phones. *International Journal of Conceptions on computing and Information Technology.* Vol. 2, pp. 6-11.

Štarchoň, P. and Sethi, A., 2019. A survey of fault localization techniques in computer networks. *The 10th International Conference on Ambient Systems, Networks and Technologies (ANT), Procedia Computer Science* Vol 151, pp. 303–312

Stevens, R.E., 2012. *Market opportunity analysis*. London: Routledge.

Steinder, M. and Sethi, A., 2004. A survey of fault localization techniques in computer networks. *Science of Computer Programming*, 53 (2).

Su, X., & Khoshgoftaar, T., 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence* 1-19. doi:10.1155/2009/421425.

Tang, A., Y., C., Azami, N., H., & Osman, N., 2011. Application of Data Mining Techniques in Customer Relationship Management for an Automobile Company. *ICIMU 2011, 14-16 November, Malaysia.*

Temprado, Y., Garcia, C. Molinero, F., J., and Gomez, J., 2008. Knowledge discovery from trouble ticketing reports in a large telecommunications company. *International Conference on Computational Intelligence of Modelling Control and Automation*, pp. 37-42.

Tsoumakas, G., Katakis, I., 2009. Multi-Label Classification: An overview. *International Journal of Data Warehouse and Mining*, 3 (3) pp. 1-13.

Tung, L., 2013. *EE lights up nine more UK towns and cities with 4G*. [Online] Available at: http://www.zdnet.com/uk/ee-lights-up-nine-more-uk-towns-and-cities-with-4g-7000011952/.

Um, P., N., Gille, L., Simon, L., & Rudelle, C., 2004. A Model for Calculating Interconnection costs in Telecommunications. *Washington, D.C.: The World Bank*.

Wachter S., 2018. Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR Center of foreign languages and communications, *Computer Law & Security Review*. Vol.34, pp. 436–449

Walgampaya, C., K., & Kantardzic, M., 2006. Cost-sensitive analysis in multiple time series prediction. *Conference on Data Mining Research on the 3G mobile network optimisation* on pp. 17-23.

Wang, D., Yu, Y., Zong, K., & Yan, F. 2011. Research on the 3G mobile network optimisation. In *Communication Technology and Application (ICCTA 2011), IET International Conference on* pp.423-427. IET.

Wang, Y-Y., Wu, K-W., Huang, C-M., & Chan, C-C. 2014. Quality management and network fault diagnosis for iptv service, *16th Asia-Pacific Network Operations and Management Symposium,* pp.1-4.

Wang, L., Li, J., Diang, L. and Li, P. 2010. E- Learning Evaluation System Based on Data Mining. S*haanxi: Center of foreign languages and communications,* Shaanxi University.

Wei, L., Yanyan, C. Dapu, Z. and Yu L. 2014. Research on reliability cost-benefit analysis and optimization for distribution network planning based on multi-measures decomposition. *Chine international conference on electricity distribution* (CICED 2014) IEEE pp.57-60.

Wei, Y., Xu, C. and Hu, Q. 2014. Transformation of optimization problems in revenue management, queueing system and supply chain management. *Shanghai: Shanghai University*, pp.588-597.

Wemerah, T., D, and Zhu, S. 2017. Big Data Challenges in Transportation: A case study of Traffic Volume Count from Massive Radio Frequency Identification *International Conference on the Frontiers and Advances in Data Science (FADS) IEEE*, pp.58-63.

Widanapathirana, C., Şekercioğlu, Y. A., Ivanovich, M.V., Fitzpatrick, P. G., Lee J.C., 2012. Automated inference system for end-to-end diagnosis of network performance issues in client-terminal devices. In *International Journal of Computer Networks & Communications (IJCNC),* Vol.4, No.3.

Wietgrefe, H. Tuchs, K-D., Jobman, K. Carls, G., Fronhlich, P., Nejdl, W. Steinfield, S., 1997. Using neural networks for alarm correlation in cellular phone networks. In *International Workshops on Applications of Neural Networks to Telecommunications*.

Wikimedia Commons, 2012. *Structure of an UMTS Network*. [Online] Available at: http://commons.wikimedia.org/wiki/File:UMTS_structures.svg.

Wireless Telegraphy 2012. Liberalised use and preparatory licenses in the 800MHz 900 MHz and 1800 MHz. Regulations 2012 Ireland.

Witten I. A., and Frank E., 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. *San Francisco, CA: Morgan-Kaufmann.*

World Economic Forum Patterson G., Mittal S.B., Weinelt B. 2017. Digital Transformation Initiative. Telecommunications Industry. *World Economic Forum in collaboration with Accenture. White Paper.*

Wu C-W., Chiang T-C, and Fu L-C, 2014. An ant colony optimisation algorithm for multi-objective clustering in mobile ad hoc networks. *IEEE Congress on Evolutionary Computation.* pp. 2963-2968.

Wu M., Kim, C, 2010. A cost matrix agent for shortest path routing in ad hoc networks. *Journal of network and computer applications*. 33 (2010) 646-852.

Wu Y., Wang H, Zang B., and Du K.-L. 2011. Using Radial Basic Function Approximation and classification. *ISRN Applied Mathematics.* Article id 324194,pp. 34.

Xue, G., 2003. Minimum-cost QoS multicast and unicast routing in communication networks. *IEEE Transactions on Communications*, 51(5), pp.817-24.

Yamashita, A., Kawamura, H. & Suzuki, K., 2010. Similarity computation method for collaborative filtering based on optimisation. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 14(6), pp.654-55.

Yang, Z., C., & Kuang, H., Xu, J., S., Sun, H., 2016. Credit evaluation using eigenface method for mobile telephone customers. *Applied Soft Computing*, 40, pp.10-16.

Ye, O. & Fallah, M.H., 2010. A performance analysis for UMTS packet switched network based on multivariate KPIs. *International Journal of Next-Generation Networks*, 2(1), pp.80-94.

Ye, L. & Qiu-ru, X., 2012. Telecom Customer Segmentation with K-means Clustering. *The 7th International Conference on Computer Science & Education (ICCSE 2012)*.

Young, S., Abdou, T., & Bener A., 2018. Deep super learner: A deep ensemble for classification problems. *Advances in Artificial Intelligence: 7: 31st Canadian Conference on Artificial Intelligence.* pp.84-95.

Zachariadis, G., & Barria, J., 2007. Load distribution for telecommunications networks using revenue management. *15th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Publisher: IEEE COMPUTER SOC*, pp. 325-330.

Zan M., Shan Z., Li L., Ai-jun L., 2007. A Predictive Model of Churn in Telecommunications Based on Data Mining. *Control and Automation, 2007 IEEE International Conference pp. 809–813.*

Zeng, L., Veeravalli, B., & Li, X., 2015. SABA: A security-aware and budget-aware workflow scheduling strategy in clouds. *Journal of Parallel and Distributed Computing,* 75, pp. 141-151.

Zhai B., & Chen, J., 2018. Development of stacked ensemble model for forecasting and analyzing daily average $PM_{2,5}$ concentration in Beijing China. *Science of the total Environment*, vol. 635 pp.644-658.

Zhang, Q., Zhu, W., & Zhang, Y., 2005. End-to-end QoS for video delivery over wireless Internet. *Proceedings of the IEEE*, 93(1), pp.123-34.

Zhang, Q., Meng Z., & Zhuo, H., 2009. Network fault diagnosis using hierarchical svms based on kernel method. *2nd International Workshop on Knowledge Discovery and Data Mining*, pp.753-756.

Zhou, L., Chen, L., Pung, H.K. & Ngoh, L.H., 2008. End-to-end diagnosis of QoS violations with neural network. In *33rd IEEE Conference on Local Computer Networks, 2008*. New York, 2008. IEEE.

Zhongwen, Z., Xu, C. and Du, Z. (2007). *Research on the Pricing Strategy of Telecom Products Based on RM*. Chengdu: IEEE, pp.1-4.

Yihua, Z., 2010. VIP customer segmentation based on data mining in mobile-communications industry. *5th International Conference on Computer Science and Education (ICCSE),* pp. 156-159.

Author's Publications

Conferences:

Rozaki, E., 2015 *Dubai*. Network fault diagnosis using data mining classifiers. *Computer Science & Information Technology (CS & IT)*. *Third International Conference on Database and Data Mining* Vol.5, pp. 29-40.

Rozaki, E., 2015. Automated Network Optimization Using Data Mining Techniques *International Conference on Signal Processing & Data Mining (ICSPDM),* Rome, Italy.

Rozaki, E., 2015 Rome, Italy. Data mining modeling and cost sensitive analysis for network faults *International Conference on Signal Processing & Data Mining (ICSPDM)*.

Journals:

Rozaki, E., 2015. Design and implementation for automated network troubleshooting using data mining. *International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.5 pp. 9-27. doi:10.5121/ijdkp.2015.5302

Rozaki, E., 2016. Clustering Optimisation Techniques in Mobile Networks. *International Journal on Recent and Innovation Trends in Computing and Communication. (IJRITCC)* Vol.4 (2) pp. 22-29

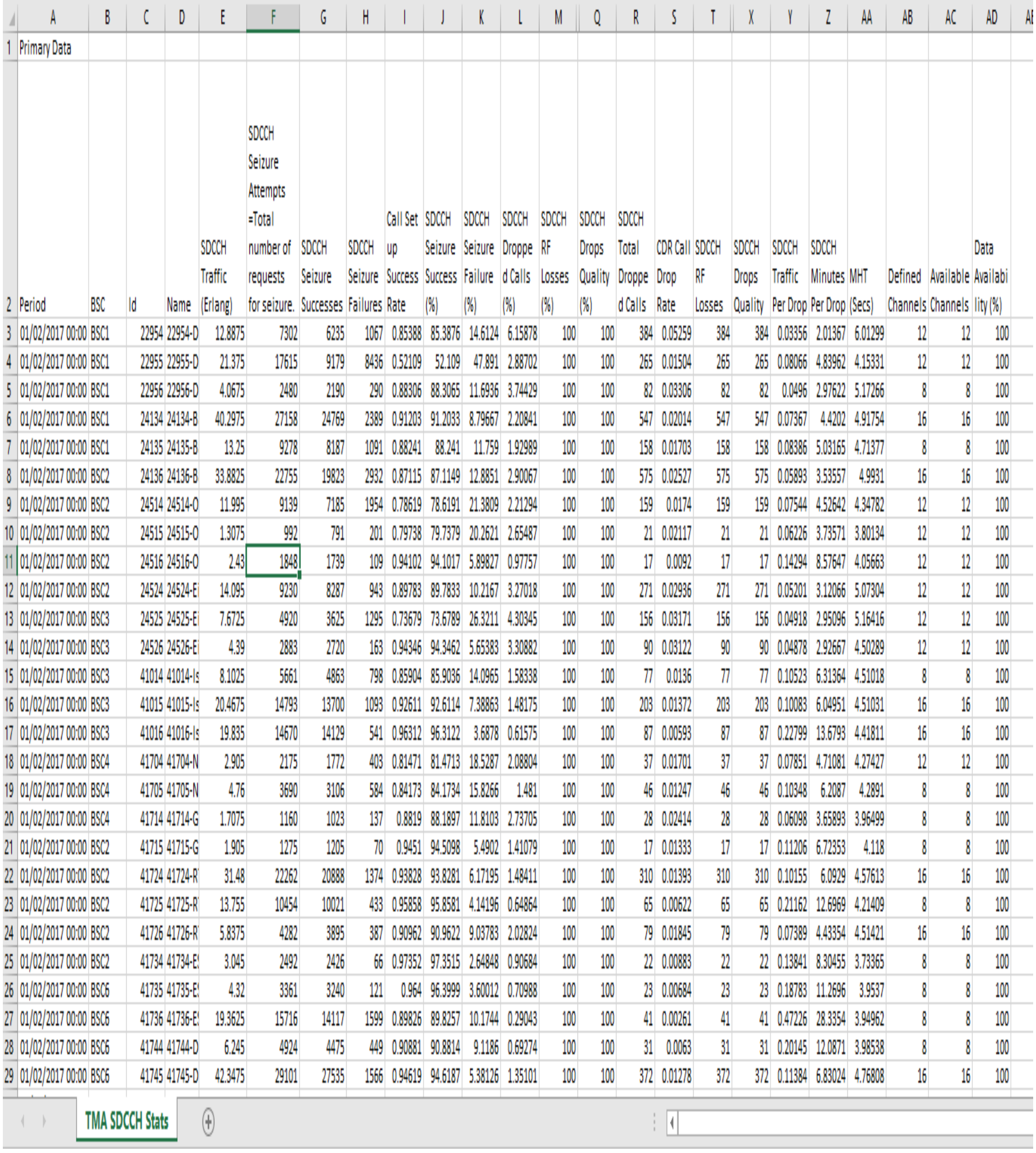
Rozaki, E., 2016. Financial Predictions Using Cost Sensitive Neural Networks for Multi-Class Learning, *Advanced Engineering Forum*, Vol. 16, pp. 104-116. doi: 10.4028/www.scientific.net/AEF.16

Dullaghan C. Rozaki, E., 2017. Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers. *International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.7 pp. 13-24. doi: 10.5121/ijdkp.2017.7102

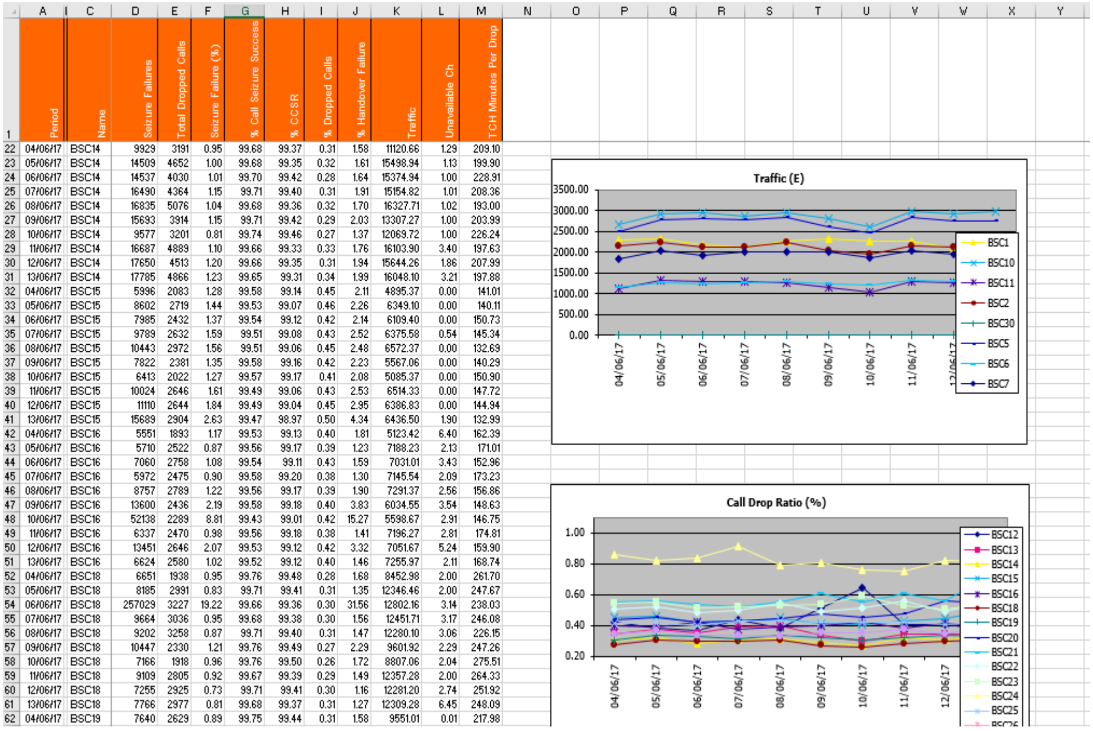
Data sets provided for this thesis:

- **iD Mobile Ireland** (licensed data set by the legal department of the iD Mobile company in Ireland)
- **TIM big data challenge 2015 competition** (licensed data set due to the competition of the TIM big data challenge)

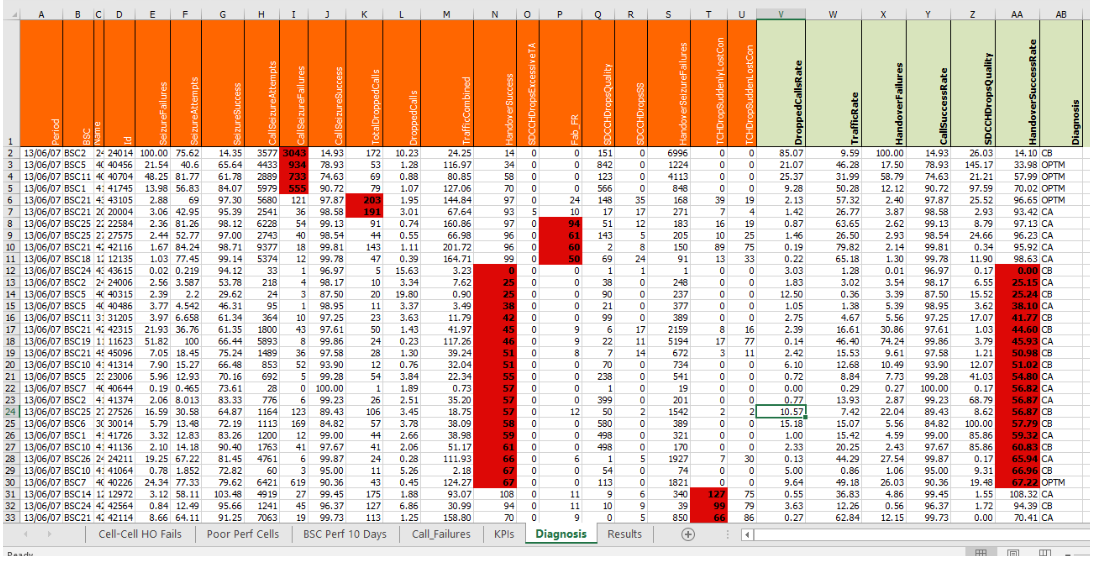
Appendix A Data and KPIs



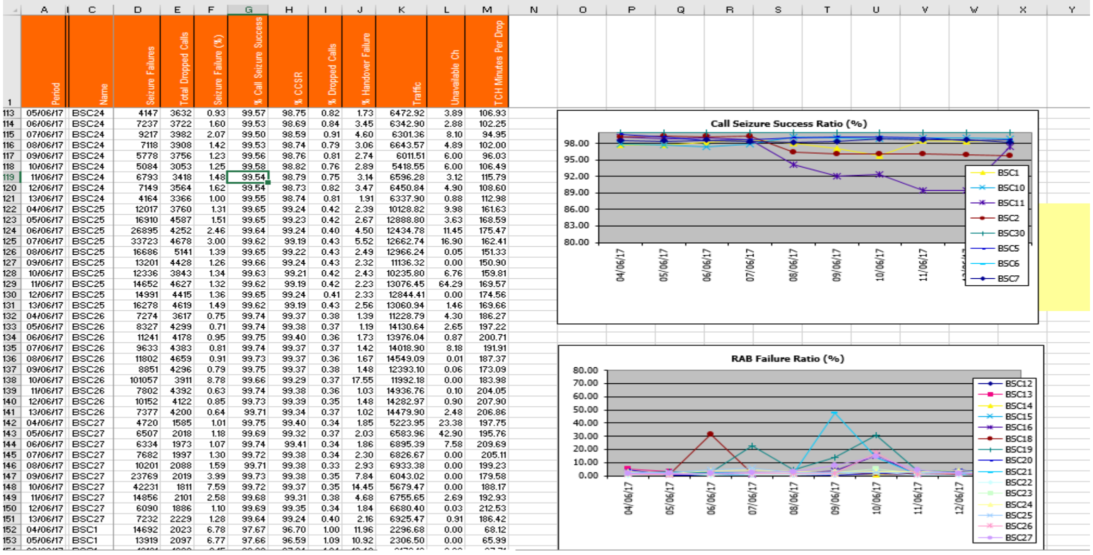
	A	B	C	D	E	F	G	H	I	J	K	L	M	Q	R	S	T	X	Y	Z	AA	AB	AC	AD
1	Primary Data																							
2	Period	BSC	Id	Name	SDCCH Traffic (Erlang)	SDCCH Seizure Attempts =Total number of requests for seizure.	SDCCH Seizure Successes	SDCCH Seizure Failures	Call Set up Success Rate	SDCCH Seizure Success (%)	SDCCH Seizure Failure (%)	SDCCH Dropped Calls (%)	SDCCH RF Losses (%)	SDCCH Drops Quality (%)	SDCCH Total Dropped Calls	CDR Call Drop Rate	SDCCH RF Losses	SDCCH Drops Quality	SDCCH Traffic Per Drop	SDCCH Minutes Per Drop	MHT (Secs)	Defined Channels	Available Channels	Data Availability (%)
3	01/02/2017 00:00	BSC1	22954	22954-D	12.8875	7302	6235	1067	0.85388	85.3876	14.6124	6.15878	100	100	384	0.05259	384	384	0.03356	2.01367	6.01299	12	12	100
4	01/02/2017 00:00	BSC1	22955	22955-D	21.375	17615	9179	8436	0.52109	52.109	47.891	2.88702	100	100	265	0.01504	265	265	0.08066	4.83962	4.15331	12	12	100
5	01/02/2017 00:00	BSC1	22956	22956-D	4.0675	2480	2190	290	0.88306	88.3065	11.6936	3.74429	100	100	82	0.03306	82	82	0.0496	2.97622	5.17266	8	8	100
6	01/02/2017 00:00	BSC1	24134	24134-B	40.2975	27158	24769	2389	0.91203	91.2033	8.79667	2.20841	100	100	547	0.02014	547	547	0.07367	4.4202	4.91754	16	16	100
7	01/02/2017 00:00	BSC1	24135	24135-B	13.25	9278	8187	1091	0.88241	88.241	11.759	1.92989	100	100	158	0.01703	158	158	0.08386	5.03165	4.71377	8	8	100
8	01/02/2017 00:00	BSC2	24136	24136-B	33.8825	22755	19823	2932	0.87115	87.1149	12.8851	2.90067	100	100	575	0.02527	575	575	0.05893	3.53557	4.9931	16	16	100
9	01/02/2017 00:00	BSC2	24514	24514-O	11.995	9139	7185	1954	0.78619	78.6191	21.3809	2.21294	100	100	159	0.0174	159	159	0.07544	4.52642	4.34782	12	12	100
10	01/02/2017 00:00	BSC2	24515	24515-O	1.3075	992	791	201	0.79738	79.7379	20.2621	2.65487	100	100	21	0.02117	21	21	0.06226	3.73571	3.80134	12	12	100
11	01/02/2017 00:00	BSC2	24516	24516-O	2.43	1848	1739	109	0.94102	94.1017	5.89827	0.97757	100	100	17	0.0092	17	17	0.14294	8.57647	4.05663	12	12	100
12	01/02/2017 00:00	BSC2	24524	24524-E	14.095	9230	8287	943	0.89783	89.7833	10.2167	3.27018	100	100	271	0.02936	271	271	0.05201	3.12066	5.07304	12	12	100
13	01/02/2017 00:00	BSC3	24525	24525-E	7.6725	4920	3625	1295	0.73679	73.6789	26.3211	4.30345	100	100	156	0.03171	156	156	0.04918	2.95096	5.16416	12	12	100
14	01/02/2017 00:00	BSC3	24526	24526-E	4.39	2883	2720	163	0.94346	94.3462	5.65383	3.30882	100	100	90	0.03122	90	90	0.04878	2.92667	4.50289	12	12	100
15	01/02/2017 00:00	BSC3	41014	41014-Is	8.1025	5661	4863	798	0.85904	85.9036	14.0965	1.58338	100	100	77	0.0136	77	77	0.10523	6.31364	4.51018	8	8	100
16	01/02/2017 00:00	BSC3	41015	41015-Is	20.4675	14793	13700	1093	0.92611	92.6114	7.38863	1.48175	100	100	203	0.01372	203	203	0.10083	6.04951	4.51031	16	16	100
17	01/02/2017 00:00	BSC3	41016	41016-Is	19.835	14670	14129	541	0.96312	96.3122	3.6878	0.61575	100	100	87	0.00593	87	87	0.22799	13.6793	4.41811	16	16	100
18	01/02/2017 00:00	BSC4	41704	41704-N	2.905	2175	1772	403	0.81471	81.4713	18.5287	2.08804	100	100	37	0.01701	37	37	0.07851	4.71081	4.27427	12	12	100
19	01/02/2017 00:00	BSC4	41705	41705-N	4.76	3690	3106	584	0.84173	84.1734	15.8266	1.481	100	100	46	0.01247	46	46	0.10348	6.2087	4.2891	8	8	100
20	01/02/2017 00:00	BSC4	41714	41714-G	1.7075	1160	1023	137	0.8819	88.1897	11.8103	2.73705	100	100	28	0.02414	28	28	0.06098	3.65893	3.96499	8	8	100
21	01/02/2017 00:00	BSC2	41715	41715-G	1.905	1275	1205	70	0.9451	94.5098	5.4902	1.41079	100	100	17	0.01333	17	17	0.11206	6.72353	4.118	8	8	100
22	01/02/2017 00:00	BSC2	41724	41724-R	31.48	22262	20888	1374	0.93828	93.8281	6.17195	1.48411	100	100	310	0.01393	310	310	0.10155	6.0929	4.57613	16	16	100
23	01/02/2017 00:00	BSC2	41725	41725-R	13.755	10454	10021	433	0.95858	95.8581	4.14196	0.64864	100	100	65	0.00622	65	65	0.21162	12.6969	4.21409	8	8	100
24	01/02/2017 00:00	BSC2	41726	41726-R	5.8375	4282	3895	387	0.90962	90.9622	9.03783	2.02824	100	100	79	0.01845	79	79	0.07389	4.43354	4.51421	16	16	100
25	01/02/2017 00:00	BSC2	41734	41734-E	3.045	2492	2426	66	0.97352	97.3515	2.64848	0.90684	100	100	22	0.00883	22	22	0.13841	8.30455	3.73365	8	8	100
26	01/02/2017 00:00	BSC6	41735	41735-E	4.32	3361	3240	121	0.964	96.3999	3.60012	0.70988	100	100	23	0.00684	23	23	0.18783	11.2696	3.9537	8	8	100
27	01/02/2017 00:00	BSC6	41736	41736-E	19.3625	15716	14117	1599	0.89826	89.8257	10.1744	0.29043	100	100	41	0.00261	41	41	0.47226	28.3354	3.94962	8	8	100
28	01/02/2017 00:00	BSC6	41744	41744-D	6.245	4924	4475	449	0.90881	90.8814	9.1186	0.69274	100	100	31	0.0063	31	31	0.20145	12.0871	3.98538	8	8	100
29	01/02/2017 00:00	BSC6	41745	41745-D	42.3475	29101	27535	1566	0.94619	94.6187	5.38126	1.35101	100	100	372	0.01278	372	372	0.11384	6.83024	4.76808	16	16	100



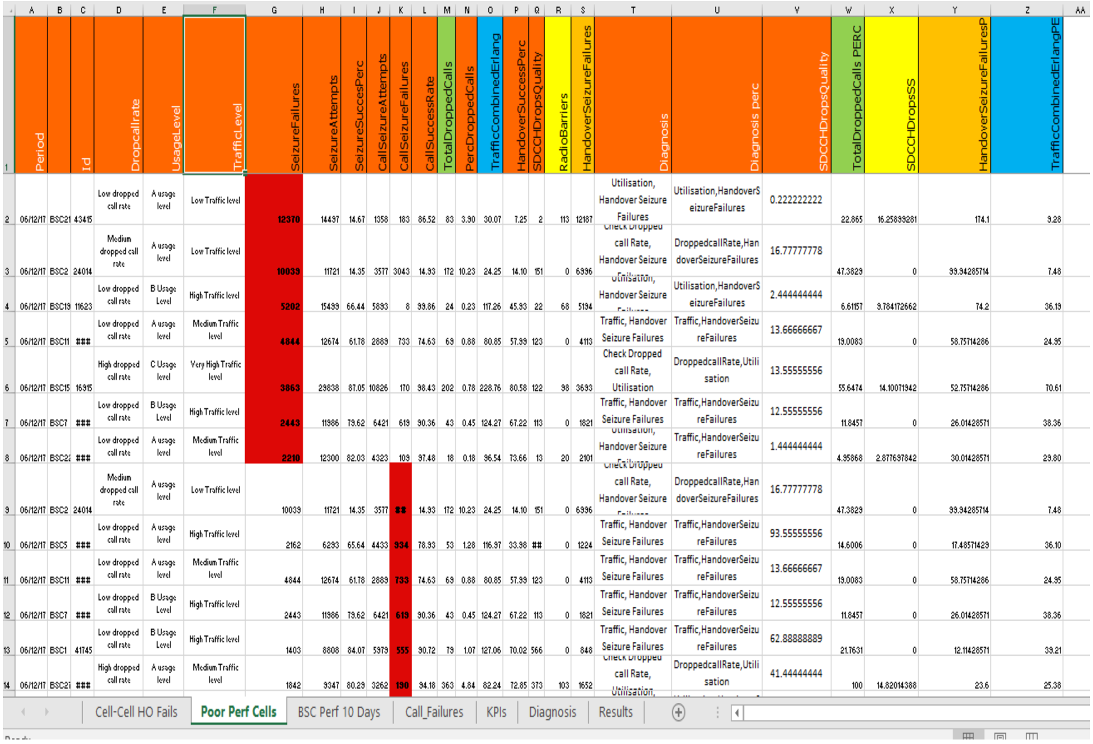
	Period	Name	Seizure Failures	Total Dropped Calls	Seizure Failure (%)	% Call Seizure Success	% CCSR	% Dropped Calls	% Handover Failure	Traffic	Unavailable Ch	TCH Minutes Per Drop
22	04/06/17	BSC14	9929	3191	0.95	99.68	99.37	0.31	1.58	11120.66	1.29	209.10
23	05/06/17	BSC14	14509	4652	1.00	99.68	99.35	0.32	1.61	15498.94	1.13	199.90
24	06/06/17	BSC14	14537	4030	1.01	99.70	99.42	0.28	1.64	15374.94	1.00	228.91
25	07/06/17	BSC14	16490	4364	1.15	99.71	99.40	0.31	1.91	15154.82	1.01	208.36
26	08/06/17	BSC14	16835	5076	1.04	99.68	99.36	0.32	1.70	16327.71	1.02	193.00
27	09/06/17	BSC14	15693	3914	1.15	99.71	99.42	0.29	2.03	13307.27	1.00	203.99
28	10/06/17	BSC14	9577	3201	0.81	99.74	99.46	0.27	1.37	12069.72	1.00	226.24
29	11/06/17	BSC14	16687	4889	1.10	99.66	99.33	0.33	1.76	16103.90	3.40	197.63
30	12/06/17	BSC14	17650	4513	1.20	99.66	99.35	0.31	1.94	15644.26	1.86	207.99
31	13/06/17	BSC14	17785	4866	1.23	99.65	99.31	0.34	1.99	16048.10	3.21	197.88
32	04/06/17	BSC15	5996	2083	1.28	99.58	99.14	0.45	2.11	4895.37	0.00	141.01
33	05/06/17	BSC15	8602	2719	1.44	99.53	99.07	0.46	2.26	6349.10	0.00	140.11
34	06/06/17	BSC15	7985	2432	1.37	99.54	99.12	0.42	2.14	6109.40	0.00	150.73
35	07/06/17	BSC15	9789	2632	1.59	99.51	99.08	0.43	2.52	6375.58	0.54	145.34
36	08/06/17	BSC15	10443	2972	1.56	99.51	99.06	0.45	2.48	6572.37	0.00	132.69
37	09/06/17	BSC15	7822	2381	1.35	99.58	99.16	0.42	2.23	5567.06	0.00	140.29
38	10/06/17	BSC15	6413	2022	1.27	99.57	99.17	0.41	2.08	5085.37	0.00	150.90
39	11/06/17	BSC15	10024	2646	1.61	99.49	99.06	0.43	2.53	6514.33	0.00	147.72
40	12/06/17	BSC15	11110	2644	1.84	99.49	99.04	0.45	2.95	6386.83	0.00	144.94
41	13/06/17	BSC15	15689	2904	2.63	99.47	98.97	0.50	4.34	6436.50	1.90	132.99
42	04/06/17	BSC16	5551	1893	1.17	99.53	99.13	0.40	1.81	5123.42	6.40	162.39
43	05/06/17	BSC16	5710	2522	0.87	99.56	99.17	0.39	1.23	7188.23	2.13	171.01
44	06/06/17	BSC16	7060	2758	1.08	99.54	99.11	0.43	1.59	7031.01	3.43	152.96
45	07/06/17	BSC16	5972	2475	0.90	99.58	99.20	0.38	1.30	7145.54	2.09	173.23
46	08/06/17	BSC16	8757	2789	1.22	99.56	99.17	0.39	1.90	7291.37	2.56	156.86
47	09/06/17	BSC16	13600	2436	2.19	99.58	99.18	0.40	3.83	6034.55	3.54	148.63
48	10/06/17	BSC16	52138	2289	8.81	99.43	99.01	0.42	15.27	5598.67	2.91	146.75
49	11/06/17	BSC16	6337	2470	0.98	99.56	99.18	0.38	1.41	7196.27	2.81	174.81
50	12/06/17	BSC16	13451	2646	2.07	99.53	99.12	0.42	3.32	7051.67	5.24	159.90
51	13/06/17	BSC16	6624	2580	1.02	99.52	99.12	0.40	1.46	7255.97	2.11	168.74
52	04/06/17	BSC18	6651	1938	0.95	99.76	99.48	0.28	1.68	8452.98	2.00	261.70
53	05/06/17	BSC18	8185	2991	0.83	99.71	99.41	0.31	1.35	12346.46	2.00	247.67
54	06/06/17	BSC18	257029	3227	19.22	99.66	99.36	0.30	31.56	12802.16	3.14	238.03
55	07/06/17	BSC18	9664	3036	0.95	99.68	99.38	0.30	1.56	12451.71	3.17	246.08
56	08/06/17	BSC18	9202	3258	0.87	99.71	99.40	0.31	1.47	12280.10	3.06	226.15
57	09/06/17	BSC18	10447	2330	1.21	99.76	99.49	0.27	2.29	9601.92	2.29	247.26
58	10/06/17	BSC18	7166	1918	0.96	99.76	99.50	0.26	1.72	8807.06	2.04	275.51
59	11/06/17	BSC18	9109	2805	0.92	99.67	99.39	0.29	1.49	12357.28	2.00	264.33
60	12/06/17	BSC18	7255	2925	0.73	99.71	99.41	0.30	1.16	12281.20	2.74	251.92
61	13/06/17	BSC18	7766	2977	0.81	99.68	99.37	0.31	1.27	12309.28	6.45	248.09
62	04/06/17	BSC19	7640	2629	0.89	99.75	99.44	0.31	1.58	9551.01	0.01	217.98



	Period	BSC	Name	Id	SeizureFailures	SeizureAttempts	SeizureSuccess	CallSeizureAttempts	CallSeizureFailures	CallSeizureSuccess	TotalDroppedCalls	DroppedCalls	TrafficCombined	HandoverSuccess	SDCCHDropsExcessiveTA	Fab_FR	SDCCHDropsQuality	SDCCHDropsSS	HandoverSeizureFailures	TCHDropSuddenlyLostCon	TCHDropSuddenLostCon	DroppedCallsRate	TrafficRate	HandoverFailures	CallSuccessRate	SDCCHDropsQuality	HandoverSuccessRate	Diagnosis
2	13/06/07	BSC2	24	24014	100.00	75.62	14.35	3577	3043	14.93	172	10.23	24.25	14	0	0	151	0	6996	0	0	85.07	9.59	100.00	14.93	26.03	14.10	CB
3	13/06/07	BSC5	40	40456	21.54	40.6	65.64	4433	934	78.93	53	1.28	116.97	34	0	0	842	0	1224	0	0	21.07	46.28	17.50	78.93	145.17	33.98	OPTM
4	13/06/07	BSC11	40	40704	48.25	81.77	61.78	2889	733	74.63	69	0.88	80.85	58	0	0	123	0	4113	0	0	25.37	31.99	58.79	74.63	21.21	57.99	OPTM
5	13/06/07	BSC1	41	41745	13.98	56.83	84.07	5979	555	90.72	79	1.07	127.06	70	0	0	566	0	848	0	0	9.28	50.28	12.12	90.72	97.59	70.02	OPTM
6	13/06/07	BSC21	43	43105	2.88	69	97.30	5680	121	97.87	203	1.95	144.84	97	0	24	148	35	168	39	19	2.13	57.32	2.40	97.87	25.52	96.65	OPTM
7	13/06/07	BSC21	20	20004	3.06	42.95	95.39	2541	36	98.58	191	3.01	67.64	93	5	10	17	17	271	7	4	1.42	26.77	3.87	98.58	2.93	93.42	CA
8	13/06/07	BSC25	22	22584	2.36	81.26	98.12	6228	54	99.13	91	0.74	160.86	97	0	94	51	12	183	16	19	0.87	63.65	2.62	99.13	8.79	97.13	CA
9	13/06/07	BSC25	27	27575	2.44	52.77	97.00	2743	40	98.54	44	0.55	66.98	96	0	61	143	5	205	10	25	1.46	26.50	2.93	98.54	24.66	96.23	CA
10	13/06/07	BSC21	42	42116	1.67	84.24	98.71	9377	18	99.81	143	1.11	201.72	96	0	60	2	8	150	89	75	0.19	79.82	2.14	99.81	0.34	95.92	CA
11	13/06/07	BSC18	12	12135	1.03	77.45	99.14	5374	12	99.78	47	0.39	164.71	99	0	50	69	24	91	13	33	0.22	65.18	1.30	99.78	11.90	98.63	CA
12	13/06/07	BSC24	43	43615	0.02	0.219	94.12	33	1	96.97	5	15.63	3.23	0	0	0	1	1	1	0	0	3.03	1.28	0.01	96.97	0.17	0.00	CB
13	13/06/07	BSC2	24	24006	2.56	3.587	53.78	218	4	98.17	10	3.34	7.62	25	0	0	38	0	248	0	0	1.83	3.02	3.54	98.17	6.55	25.15	CA
14	13/06/07	BSC5	40	40315	2.39	2.2	29.62	24	3	87.50	20	19.80	0.90	25	0	0	90	0	237	0	0	12.50	0.36	3.39	87.50	15.52	25.24	CB
15	13/06/07	BSC5	40	40486	3.77	4.542	46.31	95	1	98.95	11	3.37	3.49	38	0	0	21	0	377	0	0	1.05	1.38	5.39	98.95	3.62	38.10	CA
16	13/06/07	BSC11	31	31205	3.97	6.658	61.34	364	10	97.25	23	3.63	11.79	42	0	0	99	0	389	0	0	2.75	4.67	5.56	97.25	17.07	41.77	CB
17	13/06/07	BSC21	42	42315	21.93	36.76	61.35	1800	43	97.61	50	1.43	41.97	45	0	9	6	17	2159	8	16	2.39	16.61	30.86	97.61	1.03	44.60	CB
18	13/06/07	BSC19	11	11623	51.82	100	66.44	5893	8	99.86	24	0.23	117.26	46	0	9	22	11	5194	17	77	0.14	46.40	74.24	99.86	3.79	45.93	CA
19	13/06/07	BSC21	45	45096	7.05	18.45	75.24	1489	36	97.58	28	1.30	39.24	51	0	8	7	14	672	3	11	2.42	15.53	9.61	97.58	1.21	50.98	CB
20	13/06/07	BSC10	41	41314	7.90	15.27	66.48	853	52	93.90	12	0.76	32.04	51	0	0	70	0	734	0	0	6.10	12.68	10.49	93.90	12.07	51.02	CB
21	13/06/07	BSC5	23	23006	5.96	12.93	70.16	692	5	99.28	54	3.84	22.34	55	0	0	238	0	541	0	0	0.72	8.84	7.73	99.28	41.03	54.80	CA
22	13/06/07	BSC7	40	40644	0.19	0.465	73.61	28	0	100.00	1	1.89	0.73	57	0	0	1	0	19	0	0	0.00	0.29	0.27	100.00	0.17	56.82	CA
23	13/06/07	BSC2	41	41374	2.06	8.013	83.33	776	6	99.23	26	2.51	35.20	57	0	0	399	0	201	0	0	0.77	13.93	2.87	99.23	68.79	56.87	CA
24	13/06/07	BSC25	27	27526	16.59	30.58	64.87	1164	123	89.43	106	3.45	18.75	57	0	12	50	2	1542	2	2	10.57	7.42	22.04	89.43	8.62	56.87	CB
25	13/06/07	BSC6	30	30014	5.79	13.48	72.19	1113	169	84.82	57	3.78	38.09	58	0	0	580	0	389	0	0	15.18	15.07	5.56	84.82	100.00	57.79	CB
26	13/06/07	BSC1	41	41726	3.32	12.83	83.26	1200	12	99.00	44	2.66	38.98	59	0	0	498	0	321	0	0	1.00	15.42	4.59	99.00	85.86	59.32	CA
27	13/06/07	BSC10	41	41136	2.10	14.18	90.40	1763	41	97.67	41	2.06	51.17	61	0	0	498	0	170	0	0	2.33	20.25	2.43	97.67	85.86	60.83	CB
28	13/06/07	BSC26	24	24211	19.25	67.22	81.45	4761	6	99.87	24	0.28	111.93	66	0	6	1	5	1927	7	30	0.13	44.29	27.54	99.87	0.17	65.94	CA
29	13/06/07	BSC10	41	41064	0.78	1.852	72.82	60	3	95.00	11	5.26	2.18	67	0	0	54	0	74	0	0	5.00	0.86	1.06	95.00	9.31	66.96	CB
30	13/06/07	BSC7	40	40226	24.34	77.33	79.62	6421	619	90.36	43	0.45	124.27	67	0	0	113	0	1821	0	0	9.64	49.18	26.03	90.36	19.48	67.22	OPTM
31	13/06/07	BSC14	12	12972	3.12	58.11	103.48	4919	27	99.45	175	1.88	93.07	108	0	11	9	6	340	127	75	0.55	36.83	4.86	99.45	1.55	108.32	CA
32	13/06/07	BSC24	42	42564	0.84	12.49	95.66	1241	45	96.37	127	6.86	30.99	94	0	11	10	9	39	99	79	3.63	12.26	0.56	96.37	1.72	94.39	CB
33	13/06/07	BSC21	42	42114	8.66	64.11	91.25	7063	19	99.73	113	1.25	158.80	70	0	9	0	5	850	66	86	0.27	62.84	12.15	99.73	0.00	70.41	CA



	A	C	D	E	F	G	H	I	J	K	L	M
1	Period	Name	Seizure Failures	Total Dropped Calls	Seizure Failure (%)	% Call Seizure Success	% CCSR	% Dropped Calls	% Handover Failure	Traffic	Unavailable Ch	TCH Minutes Per Drop
113	05/06/17	BSC24	4147	3632	0.93	99.57	98.75	0.82	1.73	6472.92	3.89	106.93
114	06/06/17	BSC24	7237	3722	1.60	99.53	98.69	0.84	3.45	6342.90	2.88	102.25
115	07/06/17	BSC24	9217	3982	2.07	99.50	98.59	0.91	4.60	6301.36	8.10	94.95
116	08/06/17	BSC24	7118	3908	1.42	99.53	98.74	0.79	3.06	6643.57	4.89	102.00
117	09/06/17	BSC24	5778	3756	1.23	99.56	98.76	0.81	2.74	6011.51	6.00	96.03
118	10/06/17	BSC24	5084	3053	1.25	99.58	98.82	0.76	2.89	5418.55	6.00	106.49
119	11/06/17	BSC24	6793	3418	1.48	99.54	98.79	0.75	3.14	6596.28	3.12	115.79
120	12/06/17	BSC24	7149	3564	1.62	99.54	98.73	0.82	3.47	6450.84	4.90	108.60
121	13/06/17	BSC24	4164	3366	1.00	99.55	98.74	0.81	1.91	6337.90	0.88	112.98
122	04/06/17	BSC25	12017	3760	1.31	99.65	99.24	0.42	2.39	10128.82	9.98	161.63
123	05/06/17	BSC25	16910	4587	1.51	99.65	99.23	0.42	2.67	12888.80	3.63	168.59
124	06/06/17	BSC25	26895	4252	2.46	99.64	99.24	0.40	4.50	12434.78	11.45	175.47
125	07/06/17	BSC25	33723	4678	3.00	99.62	99.19	0.43	5.52	12662.74	16.90	162.41
126	08/06/17	BSC25	16686	5141	1.39	99.65	99.22	0.43	2.49	12966.24	0.05	151.33
127	09/06/17	BSC25	13201	4428	1.26	99.66	99.24	0.43	2.32	11136.32	0.00	150.90
128	10/06/17	BSC25	12336	3843	1.34	99.63	99.21	0.42	2.43	10235.80	6.76	159.81
129	11/06/17	BSC25	14652	4627	1.32	99.62	99.19	0.42	2.23	13076.45	64.29	169.57
130	12/06/17	BSC25	14991	4415	1.36	99.65	99.24	0.41	2.33	12844.41	0.00	174.56
131	13/06/17	BSC25	16278	4619	1.49	99.62	99.19	0.43	2.56	13060.94	1.46	169.66
132	04/06/17	BSC26	7274	3617	0.75	99.74	99.37	0.38	1.39	11228.79	4.30	186.27
133	05/06/17	BSC26	8327	4299	0.71	99.74	99.38	0.37	1.19	14130.64	2.65	197.22
134	06/06/17	BSC26	11241	4178	0.95	99.75	99.40	0.36	1.73	13976.04	0.87	200.71
135	07/06/17	BSC26	9633	4383	0.81	99.74	99.37	0.37	1.42	14018.90	8.18	191.91
136	08/06/17	BSC26	11802	4659	0.91	99.73	99.37	0.36	1.67	14549.09	0.01	187.37
137	09/06/17	BSC26	8851	4296	0.79	99.75	99.37	0.38	1.48	12393.10	0.06	173.09
138	10/06/17	BSC26	101057	3911	8.78	99.66	99.29	0.37	17.55	11992.18	0.00	183.98
139	11/06/17	BSC26	7802	4392	0.63	99.74	99.38	0.36	1.03	14936.76	0.10	204.05
140	12/06/17	BSC26	10152	4122	0.85	99.73	99.39	0.35	1.48	14282.97	0.90	207.90
141	13/06/17	BSC26	7377	4200	0.64	99.71	99.34	0.37	1.02	14479.90	2.48	206.86
142	04/06/17	BSC27	4720	1585	1.01	99.75	99.40	0.34	1.85	5223.95	23.38	197.75
143	05/06/17	BSC27	6507	2018	1.18	99.69	99.32	0.37	2.03	6583.96	42.90	195.76
144	06/06/17	BSC27	6334	1973	1.07	99.74	99.41	0.34	1.86	6895.39	7.58	209.69
145	07/06/17	BSC27	7682	1997	1.30	99.72	99.38	0.34	2.30	6826.67	0.00	205.11
146	08/06/17	BSC27	10201	2088	1.59	99.71	99.38	0.33	2.93	6933.38	0.00	199.23
147	09/06/17	BSC27	23769	2019	3.99	99.73	99.38	0.35	7.84	6043.02	0.00	179.58
148	10/06/17	BSC27	42231	1811	7.59	99.72	99.37	0.35	14.45	5679.47	0.00	188.17
149	11/06/17	BSC27	14856	2101	2.58	99.68	99.31	0.38	4.68	6755.65	2.69	192.93
150	12/06/17	BSC27	6090	1886	1.10	99.69	99.35	0.34	1.84	6680.40	0.03	212.53
151	13/06/17	BSC27	7232	2229	1.28	99.64	99.24	0.40	2.16	6925.47	0.91	186.42
152	04/06/17	BSC1	14692	2023	6.78	97.67	96.70	1.00	11.96	2296.68	0.00	68.12
153	05/06/17	BSC1	13919	2097	6.77	97.66	96.59	1.09	10.92	2306.50	0.00	65.99



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Period		Id	Dropcallrate	UsageLevel	TrafficLevel	SeizureFailures	SeizureAttempts	SeizureSuccesPerc	CallSeizureAttempts	CallSeizureFailures	CallSuccessRate	TotalDroppedCalls	PercDroppedCalls	TrafficCombinedErlang	HandoverSuccessPerc	SDCCHDropsQuality	RadioBarriers	HandoverSeizureFailures	Diagnosis	Diagnosis perc	SDCCHDropsQuality	TotalDroppedCalls PERC	SDCCHDropsSS	HandoverSeizureFailuresP	TrafficCombinedErlangPE
2	06/12/17	BSC21	43415	Low dropped call rate	A usage level	Low Traffic level	12370	14497	14.67	1358	183	86.52	83	3.90	30.07	7.25	2	113	12187	Utilisation, Handover Seizure Failures	Utilisation,HandoverSeizureFailures	0.222222222	22.865	16.25899281	174.1	9.28
3	06/12/17	BSC2	24014	Medium dropped call rate	A usage level	Low Traffic level	10039	11721	14.35	3577	3043	14.93	172	10.23	24.25	14.10	151	0	6996	Check Dropped call Rate, Handover Seizure Failures	DroppedcallRate,HandoverSeizureFailures	16.77777778	47.3829	0	99.94285714	7.48
4	06/12/17	BSC19	11623	Low dropped call rate	B Usage Level	High Traffic level	5202	15499	66.44	5893	8	99.86	24	0.23	117.26	45.93	22	68	5194	Utilisation, Handover Seizure Failures	Utilisation,HandoverSeizureFailures	2.444444444	6.61157	9.784172662	74.2	36.19
5	06/12/17	BSC11	###	Low dropped call rate	A usage level	Medium Traffic level	4844	12674	61.78	2889	733	74.63	69	0.88	80.85	57.99	123	0	4113	Traffic, Handover Seizure Failures	Traffic,HandoverSeizureFailures	13.66666667	19.0083	0	58.75714286	24.95
6	06/12/17	BSC15	16915	High dropped call rate	C Usage level	Very High Traffic level	3863	29838	87.05	10826	170	98.43	202	0.78	228.76	80.58	122	98	3693	Check Dropped call Rate, Utilisation	DroppedcallRate,Utilisation	13.55555556	55.6474	14.10071942	52.75714286	70.61
7	06/12/17	BSC7	###	Low dropped call rate	B Usage Level	High Traffic level	2443	11986	79.62	6421	619	90.36	43	0.45	124.27	67.22	113	0	1821	Traffic, Handover Seizure Failures	Traffic,HandoverSeizureFailures	12.55555556	11.8457	0	26.01428571	38.36
8	06/12/17	BSC22	###	Low dropped call rate	A usage level	Medium Traffic level	2210	12300	82.03	4323	109	97.48	18	0.18	96.54	73.66	13	20	2101	Utilisation, Handover Seizure Failures	Traffic,HandoverSeizureFailures	1.444444444	4.95868	2.877697842	30.01428571	29.80
9	06/12/17	BSC2	24014	Medium dropped call rate	A usage level	Low Traffic level	10039	11721	14.35	3577	##	14.93	172	10.23	24.25	14.10	151	0	6996	Check Dropped call Rate, Handover Seizure Failures	DroppedcallRate,HandoverSeizureFailures	16.77777778	47.3829	0	99.94285714	7.48
10	06/12/17	BSC5	###	Low dropped call rate	A usage level	High Traffic level	2162	6293	65.64	4433	934	78.93	53	1.28	116.97	33.98	##	0	1224	Traffic, Handover Seizure Failures	Traffic,HandoverSeizureFailures	93.55555556	14.6006	0	17.48571429	36.10
11	06/12/17	BSC11	###	Low dropped call rate	A usage level	Medium Traffic level	4844	12674	61.78	2889	733	74.63	69	0.88	80.85	57.99	123	0	4113	Traffic, Handover Seizure Failures	Traffic,HandoverSeizureFailures	13.66666667	19.0083	0	58.75714286	24.95
12	06/12/17	BSC7	###	Low dropped call rate	B Usage Level	High Traffic level	2443	11986	79.62	6421	619	90.36	43	0.45	124.27	67.22	113	0	1821	Traffic, Handover Seizure Failures	Traffic,HandoverSeizureFailures	12.55555556	11.8457	0	26.01428571	38.36
13	06/12/17	BSC1	41745	Low dropped call rate	B Usage Level	High Traffic level	1403	8808	84.07	5979	555	90.72	79	1.07	127.06	70.02	566	0	848	Traffic, Handover Seizure Failures	Traffic,HandoverSeizureFailures	62.88888889	21.7631	0	12.11428571	39.21
14	06/12/17	BSC21	###	High dropped call rate	A usage level	Medium Traffic level	1842	9347	80.29	3262	190	94.18	363	4.84	82.24	72.85	373	103	1652	Check Dropped call Rate, Utilisation	DroppedcallRate,Utilisation	41.44444444	100	14.82014388	23.6	25.38

Appendix 2 - Run information in Weka

A summary of various result outputs of tests performed:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2

Relation:     iD DataMining Weka CSV_Seperate Date-

weka.filters.unsupervised.attribute.Remove-R29-39

Instances:    26703

Attributes:   86

              Customer

              Customer Network Status

              Customer Activation Date

              Customer Activation Time

              Subscriber Type

              Tariff Data Option

              Tariff Text Option

              Tariff Voice Option

              Customer Bill Address

              Address Line 4
```

Customer in Collections Indicator

Customer Gender

Customer Birth Date

Customer Age

Customer Contract Length (in Months)

Customer Last MO Usage Date

Customer Last MT Usage Date

Care Call Count

Care Call Duration Seconds

Subscriber Subscription Status

Subscriber Active Date

Subscriber Deactive Date

Subscriber Port-In Flag

Subscriber Port-In Date

Subscriber Port-In Operator

Subscriber Port-Out Date

Subscriber Port-Out Operator

Agent

July 2015 Invoice Balance Brought Forward

July 2015 Invoice Amount

August 2015 Invoice Balance Brought Forward

August 2015 Invoice Amount

September 2015 Invoice Balance Brought Forward

September 2015 Invoice Amount

October 2015 Invoice Balance Brought Forward

October 2015 Invoice Amount

November 2015 Invoice Balance Brought Forward

November 2015 Invoice Amount

December 2015 Invoice Balance Brought Forward

December 2015 Invoice Amount

January 2016 Invoice Balance Brought Forward

January 2016 Invoice Amount

February 2016 Invoice Balance Brought Forward

February 2016 Invoice Amount

March 2016 Invoice Balance Brought Forward

March 2016 Invoice Amount

April 2016 Invoice Balance Brought Forward

April 2016 Invoice Amount

May 2016 Invoice Balance Brought Forward

May 2016 Invoice Amount

June 2016 Invoice Balance Brought Forward

June 2016 Invoice Amount

July 2016 Invoice Balance Brought Forward

July 2016 Invoice Amount

August 2016 Invoice Balance Brought Forward

August 2016 Invoice Amount

September 2016 Invoice Balance Brought Forward

September 2016 Invoice Amount

October 2016 Invoice Balance Brought Forward

October 2016 Invoice Amount

July 2015 Invoice Status

August 2015 Invoice Status

September 2015 Invoice Status

October 2015 Invoice Status

November 2015 Invoice Status

December 2015 Invoice Status

January 2016 Invoice Status

February 2016 Invoice Status

March 2016 Invoice Status

April 2016 Invoice Status

May 2016 Invoice Status

June 2016 Invoice Status

July 2016 Invoice Status

August 2016 Invoice Status

September 2016 Invoice Status

Total Invoice Amount Excluding Brought Forward

Total Number of Invoices

Average Invoice Amount Excluding Brought Forward

Total Paid Amount of Invoices

Spender Status

VIP Status

Sale Date Day

```
              Sale Time of Day

              Approx Location County

              Customer Age Group

              Customer Length of Service

Test mode:    10-fold cross-validation



=== Classifier model (full training set) ===



J48 pruned tree

------------------



Customer in Collections Indicator = N

|   Spender Status = Average Spender

|   |   September 2016 Invoice Status = Paid

|   |   |   August 2016 Invoice Status = Paid

|   |   |   |   March 2016 Invoice Status = Paid: Standard (5072.0/5.0)

|   |   |   |   March 2016 Invoice Status = Unpaid: Unpaid Invoice (3.0)

|   |   |   |   March 2016 Invoice Status = No Invoice: Standard (2410.0)
|  |  |  August 2016 Invoice Status = Unpaid
|  |  |  |  July 2016 Invoice Status = Paid: Standard (2.0)
|  |  |  |  July 2016 Invoice Status = Unpaid: Unpaid Invoice (7.0)
|  |  |  |  July 2016 Invoice Status = No Invoice: Unpaid Invoice (0.0)
|  |  |  August 2016 Invoice Status = No Invoice: Standard (376.0)
|  |  September 2016 Invoice Status = No Invoice
|  |  |  February 2016 Invoice Status = Paid
|  |  |  |  January 2016 Invoice Status = Paid: Standard (535.0/34.0)
|  |  |  |  January 2016 Invoice Status = Unpaid
...
|  |  |  |  |  June 2016 Invoice Balance Brought Forward = -€ 3.16 : Unpaid Invoice (0.0)
|  |  |  |  Total Paid Amount of Invoices > 6
|  |  |  |  |  Spender Status = Average Spender: Standard (8.0)
|  |  |  |  |  Spender Status = Above Average Spender: Premium (4.0)
|  |  |  |  |  Spender Status = High Spender: Standard (0.0)
|  |  |  |  |  Spender Status = Low Spender
|  |  |  |  |  |  Care Call Count <= 0: Standard (4.0)
|  |  |  |  |  |  Care Call Count > 0: Unpaid Invoice (2.0)
|   |   |   |   |   Spender Status = Very High Spender: VIP (1.0)
|   |   April 2016 Invoice Status = Unpaid
|   |   |   Total Paid Amount of Invoices <= 3: Unpaid Invoice (2664.0)
|   |   |   Total Paid Amount of Invoices > 3
|   |   |   |   Total Number of Invoices <= 6
|   |   |   |   |   November 2015 Invoice Status = Paid: Standard (8.0)
|   |   |   |   |   November 2015 Invoice Status = Unpaid: Standard (0.0)
|   |   |   |   |   November 2015 Invoice Status = No Invoice: Unpaid Invoice (3.0/1.0)
|   |   |   |   Total Number of Invoices > 6: Unpaid Invoice (324.0/2.0)
|   |   April 2016 Invoice Status = No Invoice
|   |   |   July 2016 Invoice Status = Paid: Standard (54.0/23.0)
|   |   |   July 2016 Invoice Status = Unpaid: Unpaid Invoice (112.0/12.0)
|   |   |   July 2016 Invoice Status = No Invoice: Unpaid Invoice (76.0/30.0)
|   Customer Network Status = BLOCKED
|   |   September 2016 Invoice Status = Paid
|   |   |   Customer Contract Length (in Months) <= -1: Standard (4.0/1.0)
|   |   |   Customer Contract Length (in Months) > -1
|   |   |   |   June 2016 Invoice Status = Paid: Premium (5.0)
|   |   |   |   June 2016 Invoice Status = Unpaid: Unpaid Invoice (2.0)

|   |   |   |   June 2016 Invoice Status = No Invoice: Premium (0.0)

|   |   |   |   June 2016 Invoice Status = 2: Premium (0.0)

|   |   September 2016 Invoice Status = No Invoice: Unpaid Invoice (0.0)

|   |   September 2016 Invoice Status = Unpaid: Unpaid Invoice (630.0/3.0)

|   Customer Network Status = SUSPENDED: Unpaid Invoice (1.0)


Number of Leaves  :     9062


Size of the tree :      9124



Time taken to build model: 3.54 seconds


=== Stratified cross-validation ===

=== Summary ===


Correctly Classified Instances       26080               97.6669 %
```

```
Incorrectly Classified Instances      623              2.3331 %

Kappa statistic                     0.9614

Mean absolute error                   0.019

Root mean squared error                 0.1022

Relative absolute error               6.2573 %

Root relative squared error            26.2166 %

Total Number of Instances            26703


=== Detailed Accuracy By Class ===


            TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area

Class

            0.993    0.030    0.974      0.993    0.983      0.964    0.993     0.992     Standard

            0.915    0.007    0.972      0.915    0.943      0.929    0.983     0.960     Unpaid

Invoice

            0.994    0.004    0.987      0.994    0.990      0.987    0.998     0.991     Premium

            0.954    0.001    0.936      0.954    0.945      0.944    0.988     0.945     VIP

Weighted Avg.    0.977    0.019    0.977      0.977    0.976      0.962    0.992     0.985
```

```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as

 14244    98     5     0 |     a = Standard

   366  5014    82    17 |     b = Unpaid Invoice

     6    37  6574     0 |     c = Premium

     4     8     0   248 |     d = VIP
```

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iD DataMining Weka CSV_Seperate Date-
weka.filters.unsupervised.attribute.Remove-R29-39
Test mode:    split 70.0% train, remainder test


Customer in Collections Indicator = N
|   Spender Status = Average Spender
|   |   September 2016 Invoice Status = Paid
|   |   |   August 2016 Invoice Status = Paid
|   |   |   |   March 2016 Invoice Status = Paid: Standard (5072.0/5.0)
|   |   |   |   March 2016 Invoice Status = Unpaid: Unpaid Invoice (3.0)
|   |   |   |   March 2016 Invoice Status = No Invoice: Standard (2410.0)
|   |   |   August 2016 Invoice Status = Unpaid
|   |   |   |   July 2016 Invoice Status = Paid: Standard (2.0)
|   |   |   |   July 2016 Invoice Status = Unpaid: Unpaid Invoice (7.0)
|   |   |   |   July 2016 Invoice Status = No Invoice: Unpaid Invoice (0.0)
|   |   |   August 2016 Invoice Status = No Invoice: Standard (376.0)
...
|   |   |   |   |   June 2016 Invoice Balance Brought Forward = -€ 3.16 : Unpaid Invoice (0.0)
|   |   |   |   Total Paid Amount of Invoices > 6
|   |   |   |   |   Spender Status = Average Spender: Standard (8.0)
|   |   |   |   |   Spender Status = Above Average Spender: Premium (4.0)
|   |   |   |   |   Spender Status = High Spender: Standard (0.0)
|   |   |   |   |   Spender Status = Low Spender
|   |   |   |   |   |   Care Call Count <= 0: Standard (4.0)
|   |   |   |   |   |   Care Call Count > 0: Unpaid Invoice (2.0)
|   |   |   |   |   Spender Status = Very High Spender: VIP (1.0)
|   |   April 2016 Invoice Status = Unpaid
|   |   |   Total Paid Amount of Invoices <= 3: Unpaid Invoice (2664.0)
|   |   |   Total Paid Amount of Invoices > 3
|   |   |   |   Total Number of Invoices <= 6
|   |   |   |   |   November 2015 Invoice Status = Paid: Standard (8.0)
|   |   |   |   |   November 2015 Invoice Status = Unpaid: Standard (0.0)
|   |   |   |   |   November 2015 Invoice Status = No Invoice: Unpaid Invoice (3.0/1.0)
|   |   |   |   Total Number of Invoices > 6: Unpaid Invoice (324.0/2.0)
|   |   April 2016 Invoice Status = No Invoice
|   |   |   July 2016 Invoice Status = Paid: Standard (54.0/23.0)
|   |   |   July 2016 Invoice Status = Unpaid: Unpaid Invoice (112.0/12.0)
|   |   |   July 2016 Invoice Status = No Invoice: Unpaid Invoice (76.0/30.0)
|   Customer Network Status = BLOCKED
|   |   September 2016 Invoice Status = Paid
|   |   |   Customer Contract Length (in Months) <= 1: Standard (4.0/1.0)
|   |   |   Customer Contract Length (in Months) > 1
|   |   |   |   June 2016 Invoice Status = Paid: Premium (5.0)
|   |   |   |   June 2016 Invoice Status = Unpaid: Unpaid Invoice (2.0)
|   |   |   |   June 2016 Invoice Status = No Invoice: Premium (0.0)
|   |   |   |   June 2016 Invoice Status = 2: Premium (0.0)
|   |   September 2016 Invoice Status = No Invoice: Unpaid Invoice (0.0)
|   |   September 2016 Invoice Status = Unpaid: Unpaid Invoice (630.0/3.0)
|   Customer Network Status = SUSPENDED: Unpaid Invoice (1.0)


Number of Leaves  :   9062
```

```
Size of the tree :        9124


Time taken to build model: 3.28 seconds


=== Evaluation on test split ===


Time taken to test model on training split: 0.04 seconds


=== Summary ===


Correctly Classified Instances        7778             97.0915 %

Incorrectly Classified Instances        233              2.9085 %

Kappa statistic                    0.9518

Mean absolute error                   0.0219

Root mean squared error                 0.1141

Relative absolute error               7.222  %
```

```
Root relative squared error            29.2778 %

Total Number of Instances              8011


=== Detailed Accuracy By Class ===


                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area

Class

                 0.991    0.038    0.968      0.991    0.980      0.955    0.992     0.988     Standard

                 0.899    0.009    0.962      0.899    0.930      0.912    0.978     0.935     Unpaid

Invoice

                 0.992    0.005    0.985      0.992    0.988      0.985    0.997     0.990     Premium

                 0.892    0.000    0.961      0.892    0.925      0.925    0.983     0.908     VIP

Weighted Avg.    0.971    0.024    0.971      0.971    0.971      0.953    0.990     0.977


=== Confusion Matrix ===


    a    b    c    d   <-- classified as

 4278   39    0    0 |    a = Standard

  137 1517   30    3 |    b = Unpaid Invoice
    3   12 1909    0 |    c = Premium
    0    9    0   74 |    d = VIP
```

```
=== Run information ===


Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2

Relation:     iD DataMining Weka CSV_Seperate Date-

weka.filters.unsupervised.attribute.Remove-R29-39

Instances:    26703

=== Classifier model (full training set) ===


J48 pruned tree

------------------


Customer Age <= 44

|   Customer Age <= 24:  15-24  (2192.0)

|   Customer Age > 24:  25-44  (14816.0)

Customer Age > 44

|   Customer Age <= 64:  45-64  (8178.0)

|   Customer Age > 64:  65+  (1517.0)


Number of Leaves  :   4
```

```
Size of the tree :       7




Time taken to build model: 0.47 seconds



=== Evaluation on test split ===

Time taken to test model on training split: 0.01 seconds





=== Confusion Matrix ===


    a    b    c    d   <-- classified as
 2401    0    0    0 |    a =  45-64
    0 4491    0    0 |    b =  25-44
    0    0  634    0 |    c =  15-24
```

```
   0    0    0  485 |    d =  65+



J48 pruned tree

------------------

Customer Age <= 44

|   Customer Age <= 24:  15-24  (2192.0)

|   Customer Age > 24:  25-44  (14816.0)

Customer Age > 44

|   Customer Age <= 64:  45-64  (8178.0)

|   Customer Age > 64:  65+  (1517.0)



Number of Leaves  :    4



Size of the tree :     7



Time taken to build model: 0.49 seconds
```

=== Confusion Matrix ===

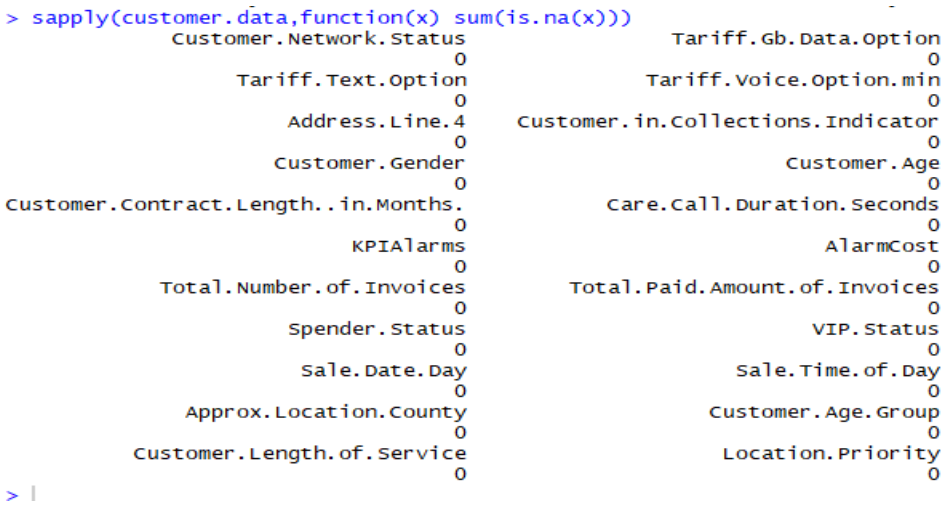
```
    a     b     c     d   <-- classified as
 8178     0     0     0 |     a =  45-64
    0 14816     0     0 |     b =  25-44
    0     0  2192     0 |     c =  15-24
    0     0     0  1517 |     d =  65+
```

Appendix 3

Stacked Learners output in H20

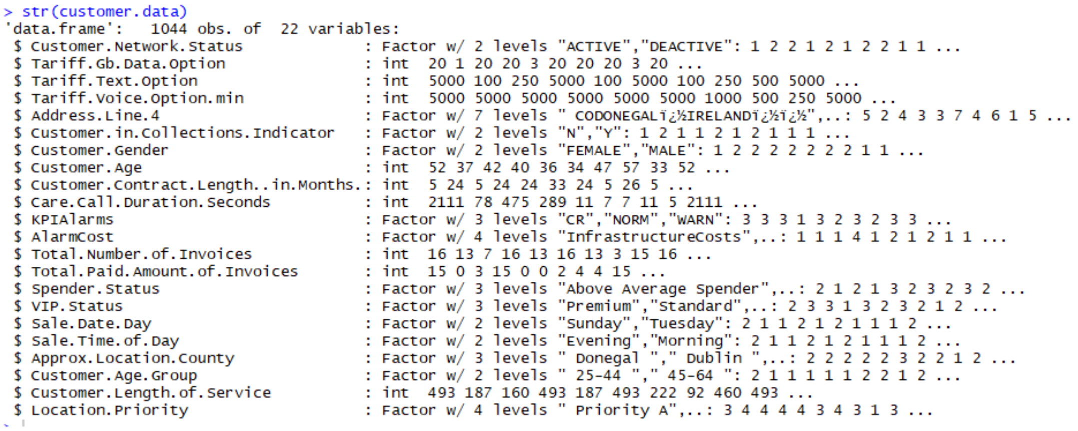
1. BORUTA ALGORITHM

Showing any null values – there is no null values in our data

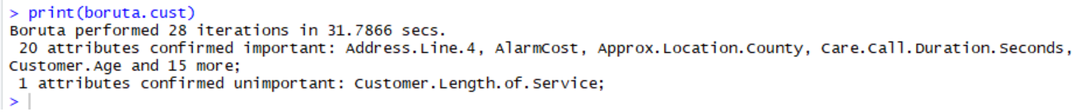


```
> sapply(customer.data,function(x) sum(is.na(x)))
                Customer.Network.Status                 Tariff.Gb.Data.Option 
                                      0                                     0 
                     Tariff.Text.Option               Tariff.Voice.Option.min 
                                      0                                     0 
                         Address.Line.4      Customer.in.Collections.Indicator 
                                      0                                     0 
                        Customer.Gender                          Customer.Age 
                                      0                                     0 
Customer.Contract.Length..in.Months.            Care.Call.Duration.Seconds 
                                      0                                     0 
                              KPIAlarms                             AlarmCost 
                                      0                                     0 
              Total.Number.of.Invoices         Total.Paid.Amount.of.Invoices 
                                      0                                     0 
                         Spender.Status                            VIP.Status 
                                      0                                     0 
                          Sale.Date.Day                       Sale.Time.of.Day 
                                      0                                     0 
                 Approx.Location.County                    Customer.Age.Group 
                                      0                                     0 
             Customer.Length.of.Service                     Location.Priority 
                                      0                                     0 
> 
```

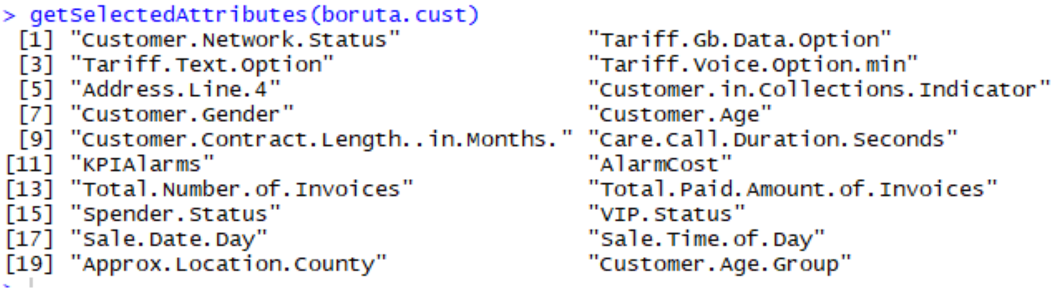
Variables within file



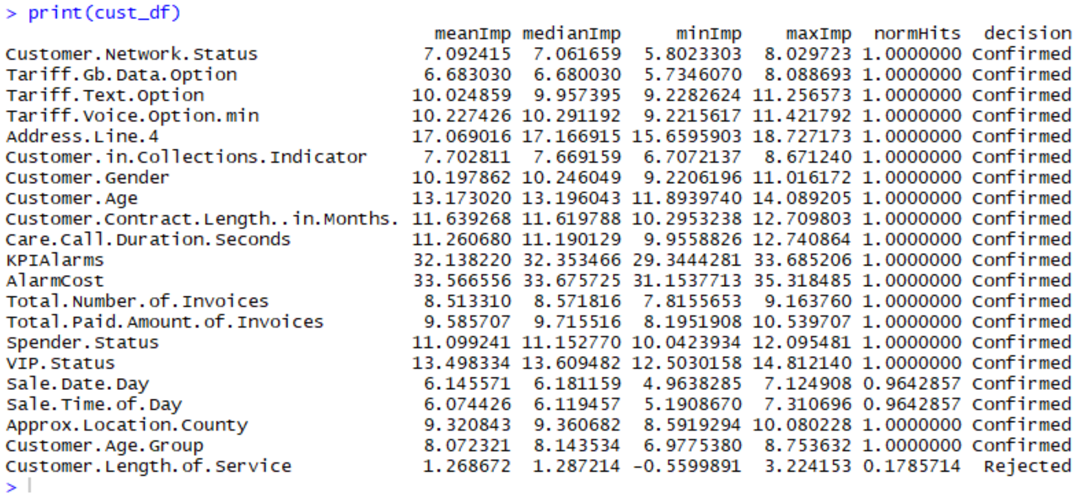
```
> str(customer.data)
'data.frame':	1044 obs. of  22 variables:
 $ Customer.Network.Status             : Factor w/ 2 levels "ACTIVE","DEACTIVE": 1 2 2 1 2 1 2 2 1 1 ...
 $ Tariff.Gb.Data.Option               : int  20 1 20 20 3 20 20 20 3 20 ...
 $ Tariff.Text.Option                  : int  5000 100 250 5000 100 5000 100 250 500 5000 ...
 $ Tariff.Voice.Option.min             : int  5000 5000 5000 5000 5000 5000 1000 500 250 5000 ...
 $ Address.Line.4                      : Factor w/ 7 levels " CODONEGALï¿½IRELANDï¿½ï¿½",..: 5 2 4 3 3 7 4 6 1 5 ...
 $ Customer.in.Collections.Indicator   : Factor w/ 2 levels "N","Y": 1 2 1 1 2 1 2 1 1 1 ...
 $ Customer.Gender                     : Factor w/ 2 levels "FEMALE","MALE": 1 2 2 2 2 2 2 2 1 1 ...
 $ Customer.Age                        : int  52 37 42 40 36 34 47 57 33 52 ...
 $ Customer.Contract.Length..in.Months.: int  5 24 5 24 24 33 24 5 26 5 ...
 $ Care.Call.Duration.Seconds          : int  2111 78 475 289 11 7 7 11 5 2111 ...
 $ KPIAlarms                           : Factor w/ 3 levels "CR","NORM","WARN": 3 3 3 1 3 2 3 2 3 3 ...
 $ AlarmCost                           : Factor w/ 4 levels "InfrastructureCosts",..: 1 1 1 4 1 2 1 2 1 1 ...
 $ Total.Number.of.Invoices            : int  16 13 7 16 13 16 13 3 15 16 ...
 $ Total.Paid.Amount.of.Invoices       : int  15 0 3 15 0 0 2 4 4 15 ...
 $ Spender.Status                      : Factor w/ 3 levels "Above Average Spender",..: 2 1 2 1 3 2 3 2 3 2 ...
 $ VIP.Status                          : Factor w/ 3 levels "Premium","Standard",..: 2 3 3 1 3 2 3 2 1 2 ...
 $ Sale.Date.Day                       : Factor w/ 2 levels "Sunday","Tuesday": 2 1 1 2 1 2 1 1 1 2 ...
 $ Sale.Time.of.Day                    : Factor w/ 2 levels "Evening","Morning": 2 1 1 2 1 2 1 1 1 2 ...
 $ Approx.Location.County              : Factor w/ 3 levels " Donegal "," Dublin ",..: 2 2 2 2 2 3 2 2 1 2 ...
 $ Customer.Age.Group                  : Factor w/ 2 levels " 25-44 "," 45-64 ": 2 1 1 1 1 1 2 2 1 2 ...
 $ Customer.Length.of.Service          : int  493 187 160 493 187 493 222 92 460 493 ...
 $ Location.Priority                   : Factor w/ 4 levels " Priority A",..: 3 4 4 4 4 3 4 3 1 3 ...
```



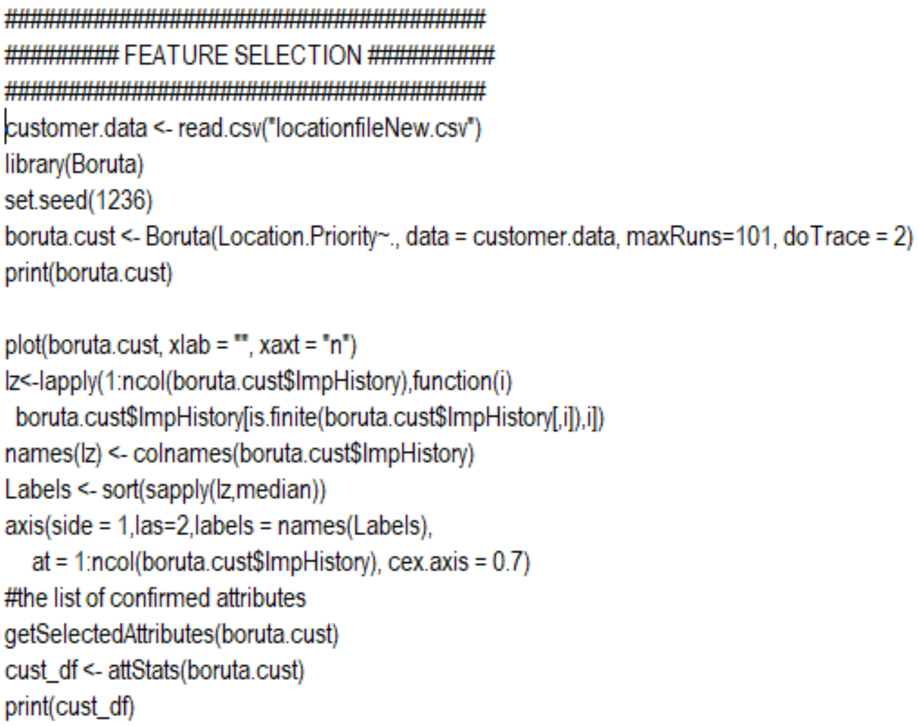
```
> print(boruta.cust)
Boruta performed 28 iterations in 31.7866 secs.
 20 attributes confirmed important: Address.Line.4, AlarmCost, Approx.Location.County, Care.Call.Duration.Seconds,
Customer.Age and 15 more;
 1 attributes confirmed unimportant: Customer.Length.of.Service;
>
```



```
> getSelectedAttributes(boruta.cust)
 [1] "Customer.Network.Status"                  "Tariff.Gb.Data.Option"
 [3] "Tariff.Text.Option"                       "Tariff.Voice.Option.min"
 [5] "Address.Line.4"                           "Customer.in.Collections.Indicator"
 [7] "Customer.Gender"                          "Customer.Age"
 [9] "Customer.Contract.Length..in.Months."     "Care.Call.Duration.Seconds"
[11] "KPIAlarms"                                "AlarmCost"
[13] "Total.Number.of.Invoices"                 "Total.Paid.Amount.of.Invoices"
[15] "Spender.Status"                           "VIP.Status"
[17] "Sale.Date.Day"                            "Sale.Time.of.Day"
[19] "Approx.Location.County"                   "Customer.Age.Group"
>
```



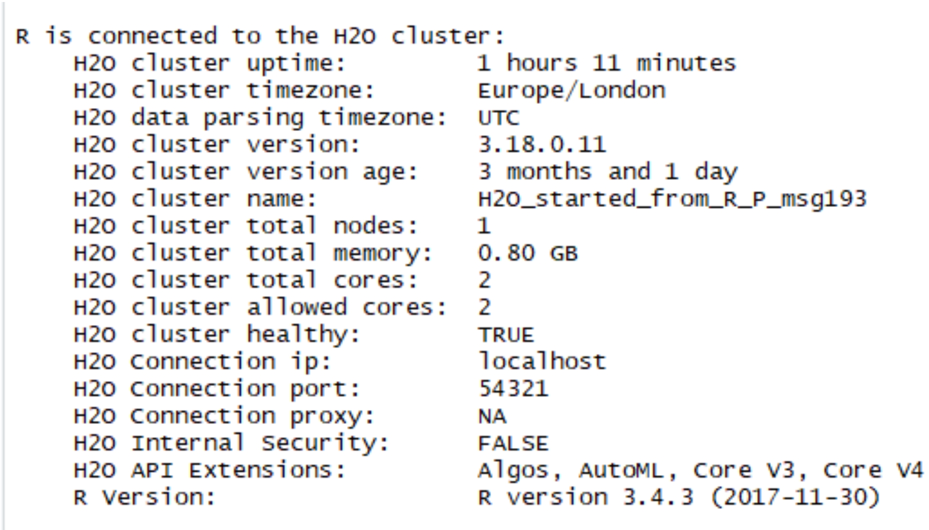
```
> print(cust_df)
                                      meanImp medianImp     minImp    maxImp  normHits  decision
Customer.Network.Status              7.092415  7.061659  5.8023303  8.029723 1.0000000 Confirmed
Tariff.Gb.Data.Option                6.683030  6.680030  5.7346070  8.088693 1.0000000 Confirmed
Tariff.Text.Option                  10.024859  9.957395  9.2282624 11.256573 1.0000000 Confirmed
Tariff.Voice.Option.min             10.227426 10.291192  9.2215617 11.421792 1.0000000 Confirmed
Address.Line.4                      17.069016 17.166915 15.6595903 18.727173 1.0000000 Confirmed
Customer.in.Collections.Indicator    7.702811  7.669159  6.7072137  8.671240 1.0000000 Confirmed
Customer.Gender                     10.197862 10.246049  9.2206196 11.016172 1.0000000 Confirmed
Customer.Age                        13.173020 13.196043 11.8939740 14.089205 1.0000000 Confirmed
Customer.Contract.Length..in.Months. 11.639268 11.619788 10.2953238 12.709803 1.0000000 Confirmed
Care.Call.Duration.Seconds          11.260680 11.190129  9.9558826 12.740864 1.0000000 Confirmed
KPIAlarms                           32.138220 32.353466 29.3444281 33.685206 1.0000000 Confirmed
AlarmCost                           33.566556 33.675725 31.1537713 35.318485 1.0000000 Confirmed
Total.Number.of.Invoices             8.513310  8.571816  7.8155653  9.163760 1.0000000 Confirmed
Total.Paid.Amount.of.Invoices        9.585707  9.715516  8.1951908 10.539707 1.0000000 Confirmed
Spender.Status                      11.099241 11.152770 10.0423934 12.095481 1.0000000 Confirmed
VIP.Status                          13.498334 13.609482 12.5030158 14.812140 1.0000000 Confirmed
Sale.Date.Day                        6.145571  6.181159  4.9638285  7.124908 0.9642857 Confirmed
Sale.Time.of.Day                     6.074426  6.119457  5.1908670  7.310696 0.9642857 Confirmed
Approx.Location.County               9.320843  9.360682  8.5919294 10.080228 1.0000000 Confirmed
Customer.Age.Group                   8.072321  8.143534  6.9775380  8.753632 1.0000000 Confirmed
Customer.Length.of.Service           1.268672  1.287214 -0.5599891  3.224153 0.1785714  Rejected
>
```



```
##########################################
########## FEATURE SELECTION ###########
##########################################
customer.data <- read.csv("locationfileNew.csv")
library(Boruta)
set.seed(1236)
boruta.cust <- Boruta(Location.Priority~., data = customer.data, maxRuns=101, doTrace = 2)
print(boruta.cust)

plot(boruta.cust, xlab = "", xaxt = "n")
lz<-lapply(1:ncol(boruta.cust$ImpHistory),function(i)
  boruta.cust$ImpHistory[is.finite(boruta.cust$ImpHistory[,i]),i])
names(lz) <- colnames(boruta.cust$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),
     at = 1:ncol(boruta.cust$ImpHistory), cex.axis = 0.7)
#the list of confirmed attributes
getSelectedAttributes(boruta.cust)
cust_df <- attStats(boruta.cust)
print(cust_df)
```

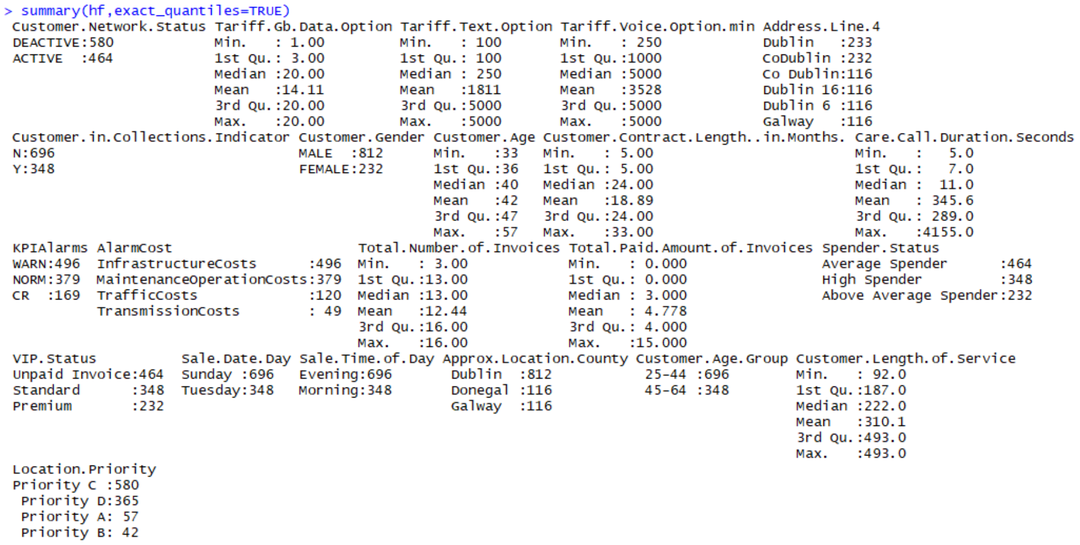
2. H20 Super Learner

Cluster

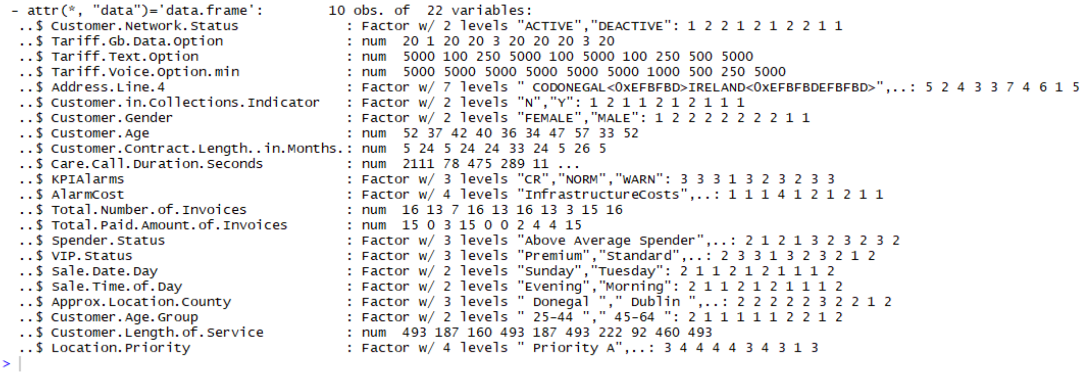


```
R is connected to the H2O cluster:
    H2O cluster uptime:         1 hours 11 minutes
    H2O cluster timezone:       Europe/London
    H2O data parsing timezone:  UTC
    H2O cluster version:        3.18.0.11
    H2O cluster version age:    3 months and 1 day
    H2O cluster name:           H2O_started_from_R_P_msg193
    H2O cluster total nodes:    1
    H2O cluster total memory:   0.80 GB
    H2O cluster total cores:    2
    H2O cluster allowed cores:  2
    H2O cluster healthy:        TRUE
    H2O Connection ip:          localhost
    H2O Connection port:        54321
    H2O Connection proxy:       NA
    H2O Internal Security:      FALSE
    H2O API Extensions:         Algos, AutoML, Core V3, Core V4
    R Version:                  R version 3.4.3 (2017-11-30)
```

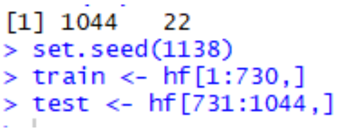
Distribution - quantiles



```
> summary(hf,exact_quantiles=TRUE)
 Customer.Network.Status Tariff.Gb.Data.Option Tariff.Text.Option Tariff.Voice.Option.min Address.Line.4
 DEACTIVE:580            Min.   : 1.00         Min.   : 100       Min.   : 250            Dublin   :233
 ACTIVE  :464            1st Qu.: 3.00         1st Qu.: 100       1st Qu.:1000            CoDublin :232
                         Median :20.00         Median : 250       Median :5000            Co Dublin:116
                         Mean   :14.11         Mean   :1811       Mean   :3528            Dublin 16:116
                         3rd Qu.:20.00         3rd Qu.:5000       3rd Qu.:5000            Dublin 6 :116
                         Max.   :20.00         Max.   :5000       Max.   :5000            Galway   :116
 Customer.in.Collections.Indicator Customer.Gender Customer.Age Customer.Contract.Length..in.Months. Care.Call.Duration.Seconds
 N:696                             MALE  :812      Min.   :33   Min.   : 5.00                        Min.   :   5.0
 Y:348                             FEMALE:232      1st Qu.:36   1st Qu.: 5.00                        1st Qu.:   7.0
                                                   Median :40   Median :24.00                        Median :  11.0
                                                   Mean   :42   Mean   :18.89                        Mean   : 345.6
                                                   3rd Qu.:47   3rd Qu.:24.00                        3rd Qu.: 289.0
                                                   Max.   :57   Max.   :33.00                        Max.   :4155.0
 KPIAlarms AlarmCost                     Total.Number.of.Invoices Total.Paid.Amount.of.Invoices Spender.Status
 WARN:496  InfrastructureCosts      :496 Min.   : 3.00            Min.   : 0.000                Average Spender      :464
 NORM:379  MaintenanceOperationCosts:379 1st Qu.:13.00            1st Qu.: 0.000                High Spender         :348
 CR  :169  TrafficCosts             :120 Median :13.00            Median : 3.000                Above Average Spender:232
           TransmissionCosts        : 49 Mean   :12.44            Mean   : 4.778
                                         3rd Qu.:16.00            3rd Qu.: 4.000
                                         Max.   :16.00            Max.   :15.000
 VIP.Status          Sale.Date.Day Sale.Time.of.Day Approx.Location.County Customer.Age.Group Customer.Length.of.Service
 Unpaid Invoice:464  Sunday :696   Evening:696        Dublin  :812           25-44 :696       Min.   : 92.0
 Standard      :348  Tuesday:348   Morning:348        Donegal :116           45-64 :348       1st Qu.:187.0
 Premium       :232                                   Galway  :116                            Median :222.0
                                                                                              Mean   :310.1
                                                                                              3rd Qu.:493.0
                                                                                              Max.   :493.0

 Location.Priority
 Priority C :580
  Priority D:365
  Priority A: 57
  Priority B: 42
```

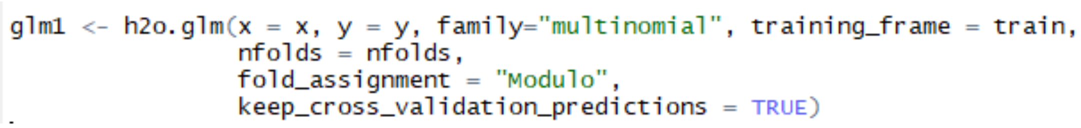


```
 - attr(*, "data")='data.frame':        10 obs. of  22 variables:
  ..$ Customer.Network.Status            : Factor w/ 2 levels "ACTIVE","DEACTIVE": 1 2 2 1 2 1 2 2 1 1
  ..$ Tariff.Gb.Data.Option              : num  20 1 20 20 3 20 20 20 3 20
  ..$ Tariff.Text.Option                 : num  5000 100 250 5000 100 5000 100 250 500 5000
  ..$ Tariff.Voice.Option.min            : num  5000 5000 5000 5000 5000 5000 1000 500 250 5000
  ..$ Address.Line.4                     : Factor w/ 7 levels " CODONEGAL<0xEFBFBD>IRELAND<0xEFBFBDEFBFBD>",..: 5 2 4 3 3 7 4 6 1 5
  ..$ Customer.in.Collections.Indicator  : Factor w/ 2 levels "N","Y": 1 2 1 1 2 1 2 1 1 1
  ..$ Customer.Gender                    : Factor w/ 2 levels "FEMALE","MALE": 1 2 2 2 2 2 2 2 1 1
  ..$ Customer.Age                       : num  52 37 42 40 36 34 47 57 33 52
  ..$ Customer.Contract.Length..in.Months.: num  5 24 5 24 24 33 24 5 26 5
  ..$ Care.Call.Duration.Seconds         : num  2111 78 475 289 11 ...
  ..$ KPIAlarms                          : Factor w/ 3 levels "CR","NORM","WARN": 3 3 3 1 3 2 3 2 3 3
  ..$ AlarmCost                          : Factor w/ 4 levels "InfrastructureCosts",..: 1 1 1 4 1 2 1 2 1 1
  ..$ Total.Number.of.Invoices           : num  16 13 7 16 13 16 13 3 15 16
  ..$ Total.Paid.Amount.of.Invoices      : num  15 0 3 15 0 0 2 4 4 15
  ..$ Spender.Status                     : Factor w/ 3 levels "Above Average Spender",..: 2 1 2 1 3 2 3 2 3 2
  ..$ VIP.Status                         : Factor w/ 3 levels "Premium","Standard",..: 2 3 3 1 3 2 3 2 1 2
  ..$ Sale.Date.Day                      : Factor w/ 2 levels "Sunday","Tuesday": 2 1 1 2 1 2 1 1 1 2
  ..$ Sale.Time.of.Day                   : Factor w/ 2 levels "Evening","Morning": 2 1 1 2 1 2 1 1 1 2
  ..$ Approx.Location.County             : Factor w/ 3 levels " Donegal "," Dublin ",..: 2 2 2 2 2 3 2 2 1 2
  ..$ Customer.Age.Group                 : Factor w/ 2 levels " 25-44 "," 45-64 ": 2 1 1 1 1 1 2 2 1 2
  ..$ Customer.Length.of.Service         : num  493 187 160 493 187 493 222 92 460 493
  ..$ Location.Priority                  : Factor w/ 4 levels " Priority A",..: 3 4 4 4 4 3 4 3 1 3
>
```



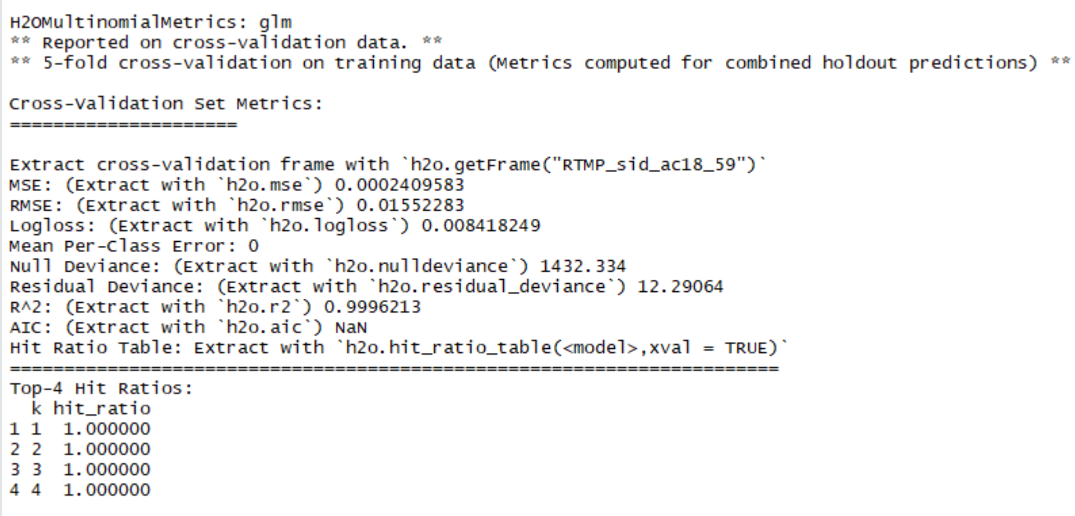
```
[1] 1044   22
> set.seed(1138)
> train <- hf[1:730,]
> test <- hf[731:1044,]
```

3. Run the first base learner model in the H20 cluster – Generalised Linear Model Using cross validation 5 folds

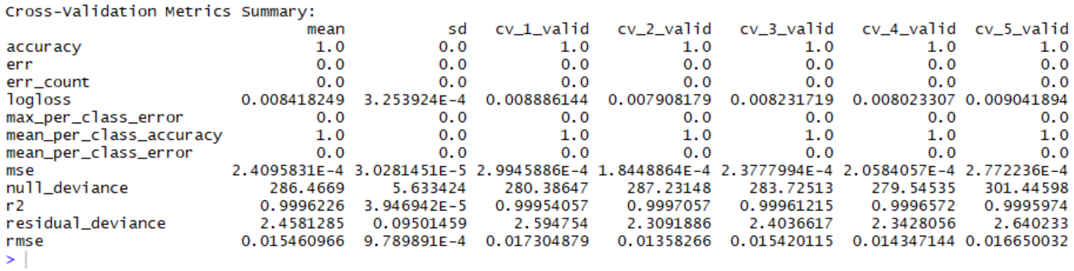


```
glm1 <- h2o.glm(x = x, y = y, family="multinomial", training_frame = train,
                nfolds = nfolds,
                fold_assignment = "Modulo",
                keep_cross_validation_predictions = TRUE)
```

Cross Validation

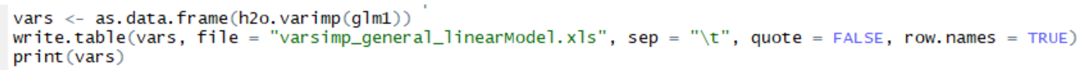


```
H2OMultinomialMetrics: glm
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **

Cross-Validation Set Metrics:
=====================

Extract cross-validation frame with `h2o.getFrame("RTMP_sid_ac18_59")`
MSE: (Extract with `h2o.mse`) 0.0002409583
RMSE: (Extract with `h2o.rmse`) 0.01552283
Logloss: (Extract with `h2o.logloss`) 0.008418249
Mean Per-Class Error: 0
Null Deviance: (Extract with `h2o.nulldeviance`) 1432.334
Residual Deviance: (Extract with `h2o.residual_deviance`) 12.29064
R^2: (Extract with `h2o.r2`) 0.9996213
AIC: (Extract with `h2o.aic`) NaN
Hit Ratio Table: Extract with `h2o.hit_ratio_table(<model>,xval = TRUE)`
=======================================================================
Top-4 Hit Ratios:
  k hit_ratio
1 1  1.000000
2 2  1.000000
3 3  1.000000
4 4  1.000000
```

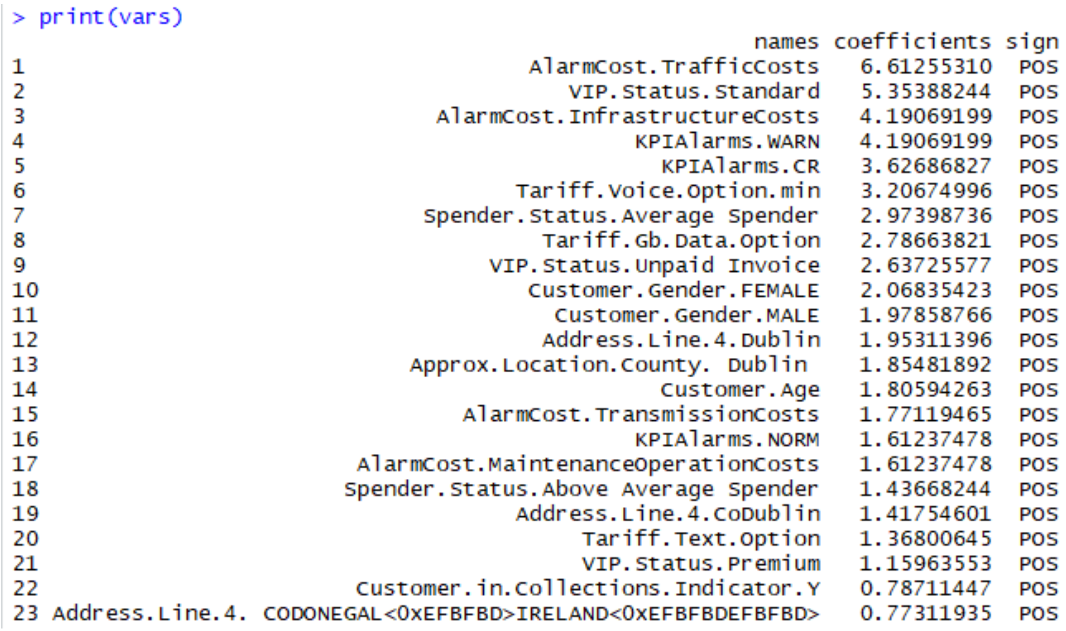


```
Cross-Validation Metrics Summary:
                                mean           sd   cv_1_valid   cv_2_valid   cv_3_valid   cv_4_valid  cv_5_valid
accuracy                         1.0          0.0          1.0          1.0          1.0          1.0         1.0
err                              0.0          0.0          0.0          0.0          0.0          0.0         0.0
err_count                        0.0          0.0          0.0          0.0          0.0          0.0         0.0
logloss                  0.008418249  3.253924E-4  0.008886144  0.007908179  0.008231719  0.008023307 0.009041894
max_per_class_error              0.0          0.0          0.0          0.0          0.0          0.0         0.0
mean_per_class_accuracy          1.0          0.0          1.0          1.0          1.0          1.0         1.0
mean_per_class_error             0.0          0.0          0.0          0.0          0.0          0.0         0.0
mse                     2.4095831E-4 3.0281451E-5 2.9945886E-4 1.8448864E-4 2.3777994E-4 2.0584057E-4 2.772236E-4
null_deviance               286.4669     5.633424    280.38647    287.23148    283.72513    279.54535   301.44598
r2                         0.9996226  3.946942E-5   0.99954057    0.9997057   0.99961215    0.9996572   0.9995974
residual_deviance          2.4581285   0.09501459     2.594754    2.3091886    2.4036617    2.3428056    2.640233
rmse                     0.015460966  9.789891E-4  0.017304879   0.01358266  0.015420115  0.014347144 0.016650032
> |
```

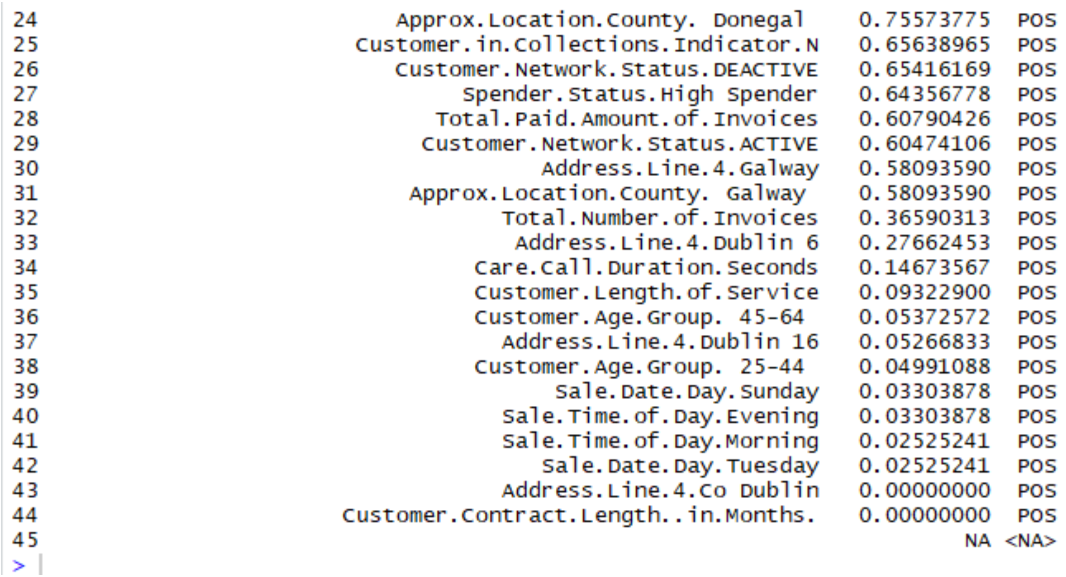
Variable importance for GLM



```
vars <- as.data.frame(h2o.varimp(glm1))
write.table(vars, file = "varsimp_general_linearModel.xls", sep = "\t", quote = FALSE, row.names = TRUE)
print(vars)
```



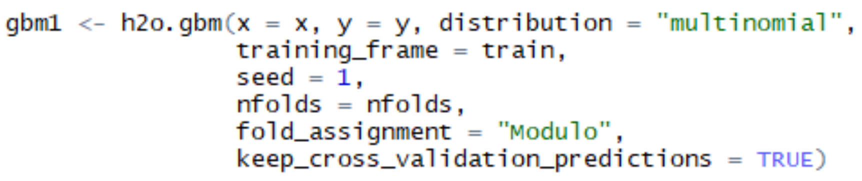
```
> print(vars)
                                                        names coefficients sign
1                                      AlarmCost.TrafficCosts   6.61255310  POS
2                                         VIP.Status.Standard   5.35388244  POS
3                               AlarmCost.InfrastructureCosts   4.19069199  POS
4                                              KPIAlarms.WARN   4.19069199  POS
5                                                KPIAlarms.CR   3.62686827  POS
6                                     Tariff.Voice.Option.min   3.20674996  POS
7                              Spender.Status.Average Spender   2.97398736  POS
8                                       Tariff.Gb.Data.Option   2.78663821  POS
9                                   VIP.Status.Unpaid Invoice   2.63725577  POS
10                                     Customer.Gender.FEMALE   2.06835423  POS
11                                       Customer.Gender.MALE   1.97858766  POS
12                                      Address.Line.4.Dublin   1.95311396  POS
13                            Approx.Location.County. Dublin    1.85481892  POS
14                                               Customer.Age   1.80594263  POS
15                                AlarmCost.TransmissionCosts   1.77119465  POS
16                                             KPIAlarms.NORM   1.61237478  POS
17                        AlarmCost.MaintenanceOperationCosts   1.61237478  POS
18                       Spender.Status.Above Average Spender   1.43668244  POS
19                                    Address.Line.4.CoDublin   1.41754601  POS
20                                         Tariff.Text.Option   1.36800645  POS
21                                         VIP.Status.Premium   1.15963553  POS
22                        Customer.in.Collections.Indicator.Y   0.78711447  POS
23 Address.Line.4. CODONEGAL<0xEFBFBD>IRELAND<0xEFBFBDEFBFBD>   0.77311935  POS
```



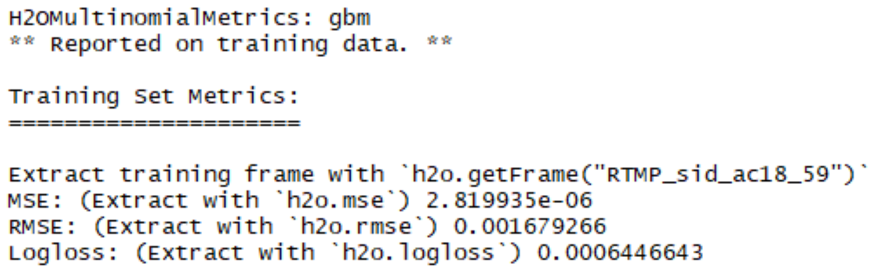
```
24                           Approx.Location.County. Donegal    0.75573775  POS
25                        Customer.in.Collections.Indicator.N   0.65638965  POS
26                           Customer.Network.Status.DEACTIVE   0.65416169  POS
27                                Spender.Status.High Spender   0.64356778  POS
28                              Total.Paid.Amount.of.Invoices   0.60790426  POS
29                             Customer.Network.Status.ACTIVE   0.60474106  POS
30                                      Address.Line.4.Galway   0.58093590  POS
31                            Approx.Location.County. Galway    0.58093590  POS
32                                   Total.Number.of.Invoices   0.36590313  POS
33                                    Address.Line.4.Dublin 6   0.27662453  POS
34                                 Care.Call.Duration.Seconds   0.14673567  POS
35                                 Customer.Length.of.Service   0.09322900  POS
36                                 Customer.Age.Group. 45-64    0.05372572  POS
37                                   Address.Line.4.Dublin 16   0.05266833  POS
38                                 Customer.Age.Group. 25-44    0.04991088  POS
39                                       Sale.Date.Day.Sunday   0.03303878  POS
40                                   Sale.Time.of.Day.Evening   0.03303878  POS
41                                   Sale.Time.of.Day.Morning   0.02525241  POS
42                                      Sale.Date.Day.Tuesday   0.02525241  POS
43                                   Address.Line.4.Co Dublin   0.00000000  POS
44                       Customer.Contract.Length..in.Months.   0.00000000  POS
45                                                                      NA <NA>
> 
```

4. Run the first base learner model in the H20 cluster – Gradient Boost Machine

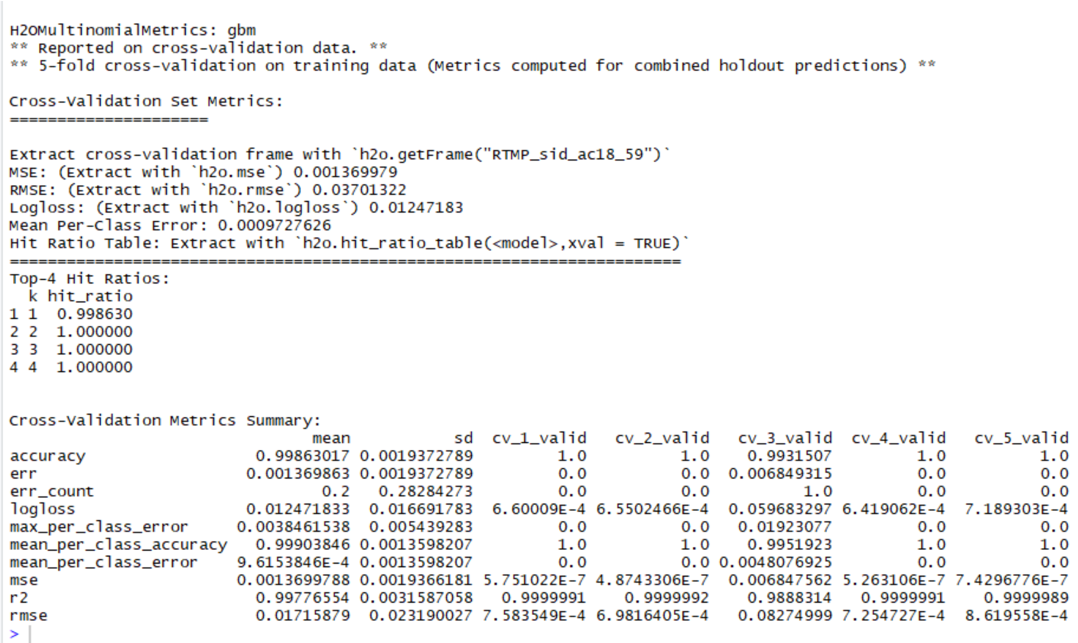
Using cross validation 5 folds



```
gbm1 <- h2o.gbm(x = x, y = y, distribution = "multinomial",
                training_frame = train,
                seed = 1,
                nfolds = nfolds,
                fold_assignment = "Modulo",
                keep_cross_validation_predictions = TRUE)
```

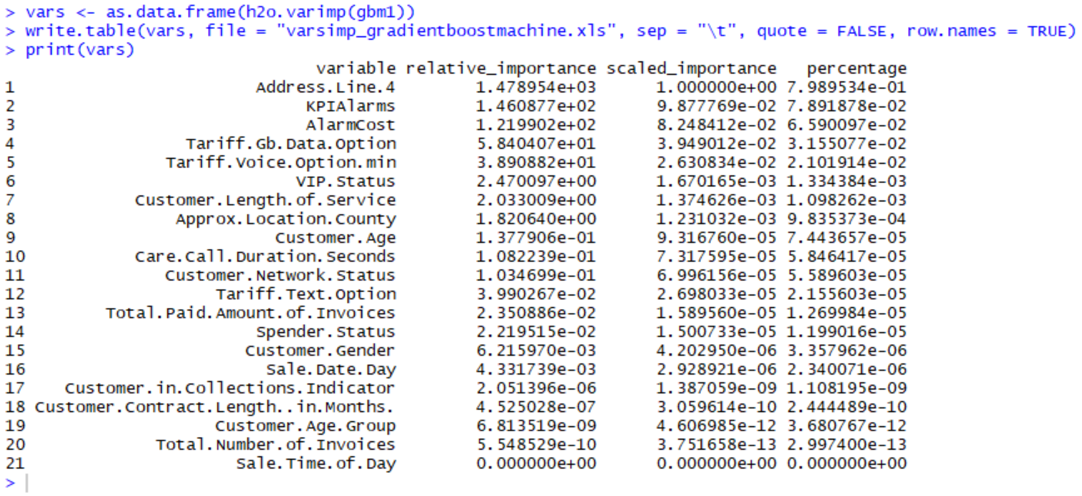


```
H2OMultinomialMetrics: gbm
** Reported on training data. **

Training Set Metrics:
=====================

Extract training frame with `h2o.getFrame("RTMP_sid_ac18_59")`
MSE: (Extract with `h2o.mse`) 2.819935e-06
RMSE: (Extract with `h2o.rmse`) 0.001679266
Logloss: (Extract with `h2o.logloss`) 0.0006446643
```



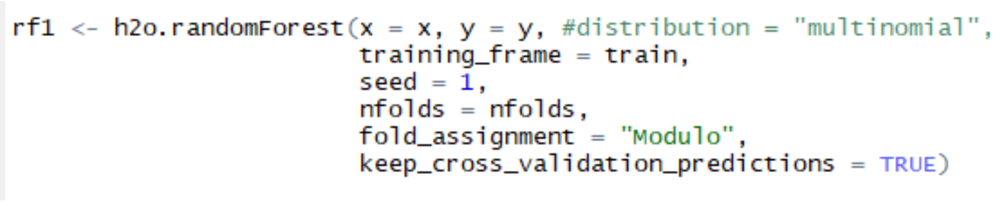
```
H2OMultinomialMetrics: gbm
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **

Cross-Validation Set Metrics:
=====================

Extract cross-validation frame with `h2o.getFrame("RTMP_sid_ac18_59")`
MSE: (Extract with `h2o.mse`) 0.001369979
RMSE: (Extract with `h2o.rmse`) 0.03701322
Logloss: (Extract with `h2o.logloss`) 0.01247183
Mean Per-Class Error: 0.0009727626
Hit Ratio Table: Extract with `h2o.hit_ratio_table(<model>,xval = TRUE)`
=======================================================================
Top-4 Hit Ratios:
  k hit_ratio
1 1  0.998630
2 2  1.000000
3 3  1.000000
4 4  1.000000


Cross-Validation Metrics Summary:
                                mean           sd  cv_1_valid   cv_2_valid   cv_3_valid  cv_4_valid   cv_5_valid
accuracy                  0.99863017 0.0019372789         1.0          1.0    0.9931507         1.0          1.0
err                      0.001369863 0.0019372789         0.0          0.0  0.006849315         0.0          0.0
err_count                        0.2   0.28284273         0.0          0.0          1.0         0.0          0.0
logloss                  0.012471833  0.016691783  6.60009E-4 6.5502466E-4  0.059683297 6.419062E-4  7.189303E-4
max_per_class_error     0.0038461538  0.005439283         0.0          0.0   0.01923077         0.0          0.0
mean_per_class_accuracy   0.99903846 0.0013598207         1.0          1.0    0.9951923         1.0          1.0
mean_per_class_error    9.6153846E-4 0.0013598207         0.0          0.0 0.0048076925         0.0          0.0
mse                     0.0013699788 0.0019366181 5.751022E-7 4.8743306E-7  0.006847562 5.263106E-7 7.4296776E-7
r2                        0.99776554 0.0031587058   0.9999991    0.9999992    0.9888314   0.9999991    0.9999989
rmse                      0.01715879  0.023190027 7.583549E-4 6.9816405E-4   0.08274999 7.254727E-4  8.619558E-4
> 
```

VARIABLE IMPORTANCE FOR GBM

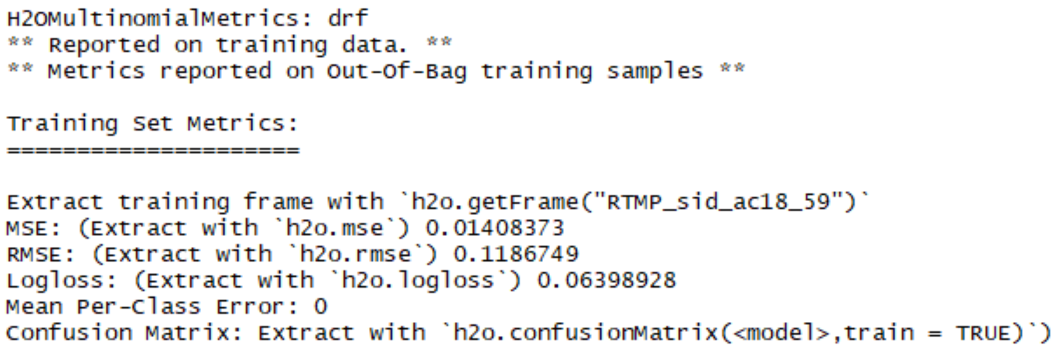


```
> vars <- as.data.frame(h2o.varimp(gbm1))
> write.table(vars, file = "varsimp_gradientboostmachine.xls", sep = "\t", quote = FALSE, row.names = TRUE)
> print(vars)
                               variable relative_importance scaled_importance   percentage
1                        Address.Line.4        1.478954e+03      1.000000e+00 7.989534e-01
2                             KPIAlarms        1.460877e+02      9.877769e-02 7.891878e-02
3                             AlarmCost        1.219902e+02      8.248412e-02 6.590097e-02
4                 Tariff.Gb.Data.Option        5.840407e+01      3.949012e-02 3.155077e-02
5               Tariff.Voice.Option.min        3.890882e+01      2.630834e-02 2.101914e-02
6                            VIP.Status        2.470097e+00      1.670165e-03 1.334384e-03
7            Customer.Length.of.Service        2.033009e+00      1.374626e-03 1.098262e-03
8                Approx.Location.County        1.820640e+00      1.231032e-03 9.835373e-04
9                          Customer.Age        1.377906e-01      9.316760e-05 7.443657e-05
10           Care.Call.Duration.Seconds        1.082239e-01      7.317595e-05 5.846417e-05
11              Customer.Network.Status        1.034699e-01      6.996156e-05 5.589603e-05
12                   Tariff.Text.Option        3.990267e-02      2.698033e-05 2.155603e-05
13        Total.Paid.Amount.of.Invoices        2.350886e-02      1.589560e-05 1.269984e-05
14                       Spender.Status        2.219515e-02      1.500733e-05 1.199016e-05
15                      Customer.Gender        6.215970e-03      4.202950e-06 3.357962e-06
16                        Sale.Date.Day        4.331739e-03      2.928921e-06 2.340071e-06
17    Customer.in.Collections.Indicator        2.051396e-06      1.387059e-09 1.108195e-09
18 Customer.Contract.Length..in.Months.        4.525028e-07      3.059614e-10 2.444489e-10
19                   Customer.Age.Group        6.813519e-09      4.606985e-12 3.680767e-12
20             Total.Number.of.Invoices        5.548529e-10      3.751658e-13 2.997400e-13
21                     Sale.Time.of.Day        0.000000e+00      0.000000e+00 0.000000e+00
> 
```

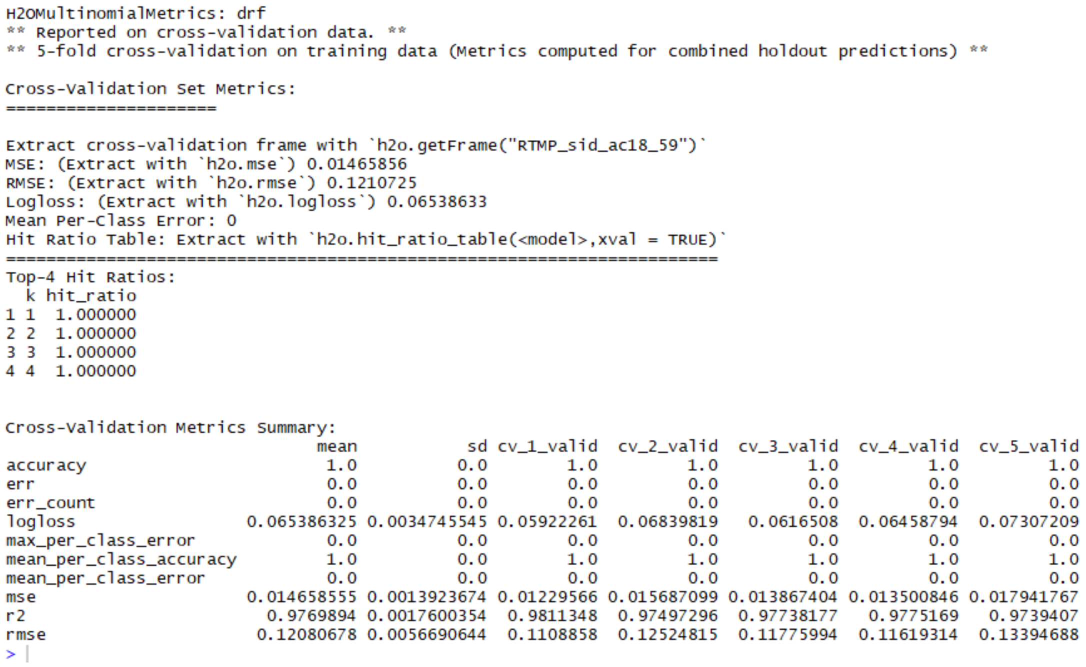
5. Run the first base learner model in the H20 cluster – Distributed Random Forest Using cross validation 5 folds



```
rf1 <- h2o.randomForest(x = x, y = y, #distribution = "multinomial",
                         training_frame = train,
                         seed = 1,
                         nfolds = nfolds,
                         fold_assignment = "Modulo",
                         keep_cross_validation_predictions = TRUE)
```

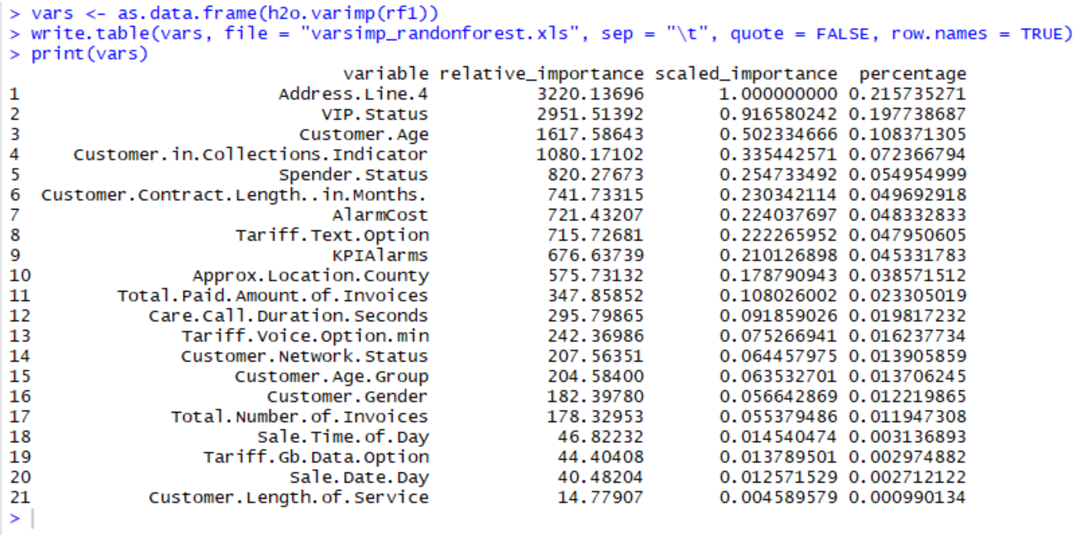


```
H2OMultinomialMetrics: drf
** Reported on training data. **
** Metrics reported on Out-Of-Bag training samples **

Training Set Metrics: 
======================

Extract training frame with `h2o.getFrame("RTMP_sid_ac18_59")`
MSE: (Extract with `h2o.mse`) 0.01408373
RMSE: (Extract with `h2o.rmse`) 0.1186749
Logloss: (Extract with `h2o.logloss`) 0.06398928
Mean Per-Class Error: 0
Confusion Matrix: Extract with `h2o.confusionMatrix(<model>,train = TRUE)`)
```



```
H2OMultinomialMetrics: drf
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **

Cross-Validation Set Metrics: 
=====================

Extract cross-validation frame with `h2o.getFrame("RTMP_sid_ac18_59")`
MSE: (Extract with `h2o.mse`) 0.01465856
RMSE: (Extract with `h2o.rmse`) 0.1210725
Logloss: (Extract with `h2o.logloss`) 0.06538633
Mean Per-Class Error: 0
Hit Ratio Table: Extract with `h2o.hit_ratio_table(<model>,xval = TRUE)`
=======================================================================
Top-4 Hit Ratios: 
  k hit_ratio
1 1  1.000000
2 2  1.000000
3 3  1.000000
4 4  1.000000


Cross-Validation Metrics Summary: 
                               mean           sd cv_1_valid  cv_2_valid  cv_3_valid  cv_4_valid  cv_5_valid
accuracy                        1.0          0.0        1.0         1.0         1.0         1.0         1.0
err                             0.0          0.0        0.0         0.0         0.0         0.0         0.0
err_count                       0.0          0.0        0.0         0.0         0.0         0.0         0.0
logloss                 0.065386325 0.0034745545 0.05922261  0.06839819   0.0616508  0.06458794  0.07307209
max_per_class_error             0.0          0.0        0.0         0.0         0.0         0.0         0.0
mean_per_class_accuracy         1.0          0.0        1.0         1.0         1.0         1.0         1.0
mean_per_class_error            0.0          0.0        0.0         0.0         0.0         0.0         0.0
mse                     0.014658555 0.0013923674 0.01229566 0.015687099 0.013867404 0.013500846 0.017941767
r2                        0.9769894 0.0017600354  0.9811348  0.97497296  0.97738177   0.9775169   0.9739407
rmse                     0.12080678 0.0056690644  0.1108858  0.12524815  0.11775994  0.11619314  0.13394688
> 
```

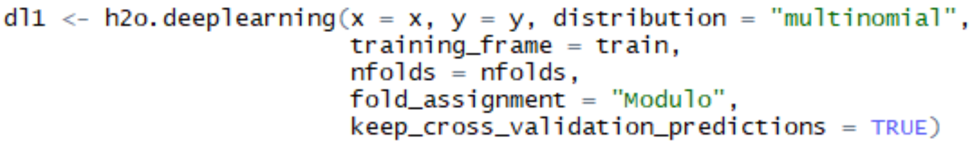
VARIABLE FOR DRF



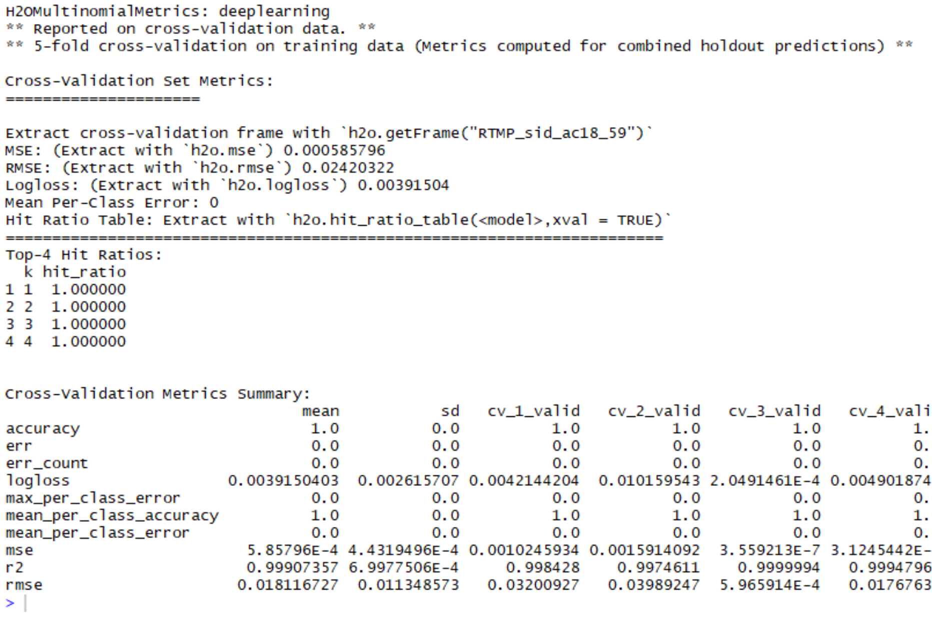
```
> vars <- as.data.frame(h2o.varimp(rf1))
> write.table(vars, file = "varsimp_randonforest.xls", sep = "\t", quote = FALSE, row.names = TRUE)
> print(vars)
                               variable relative_importance scaled_importance  percentage
1                        Address.Line.4          3220.13696       1.000000000 0.215735271
2                            VIP.Status          2951.51392       0.916580242 0.197738687
3                          Customer.Age          1617.58643       0.502334666 0.108371305
4     Customer.in.Collections.Indicator          1080.17102       0.335442571 0.072366794
5                        Spender.Status           820.27673       0.254733492 0.054954999
6  Customer.Contract.Length..in.Months.           741.73315       0.230342114 0.049692918
7                             AlarmCost           721.43207       0.224037697 0.048332833
8                    Tariff.Text.Option           715.72681       0.222265952 0.047950605
9                             KPIAlarms           676.63739       0.210126898 0.045331783
10               Approx.Location.County           575.73132       0.178790943 0.038571512
11        Total.Paid.Amount.of.Invoices           347.85852       0.108026002 0.023305019
12           Care.Call.Duration.Seconds           295.79865       0.091859026 0.019817232
13              Tariff.Voice.Option.min           242.36986       0.075266941 0.016237734
14              Customer.Network.Status           207.56351       0.064457975 0.013905859
15                   Customer.Age.Group           204.58400       0.063532701 0.013706245
16                      Customer.Gender           182.39780       0.056642869 0.012219865
17             Total.Number.of.Invoices           178.32953       0.055379486 0.011947308
18                     Sale.Time.of.Day            46.82232       0.014540474 0.003136893
19                Tariff.Gb.Data.Option            44.40408       0.013789501 0.002974882
20                        Sale.Date.Day            40.48204       0.012571529 0.002712122
21           Customer.Length.of.Service            14.77907       0.004589579 0.000990134
> 
```

6. Run the first base learner model in the H20 cluster – Deep Learning

Using cross validation 5 folds

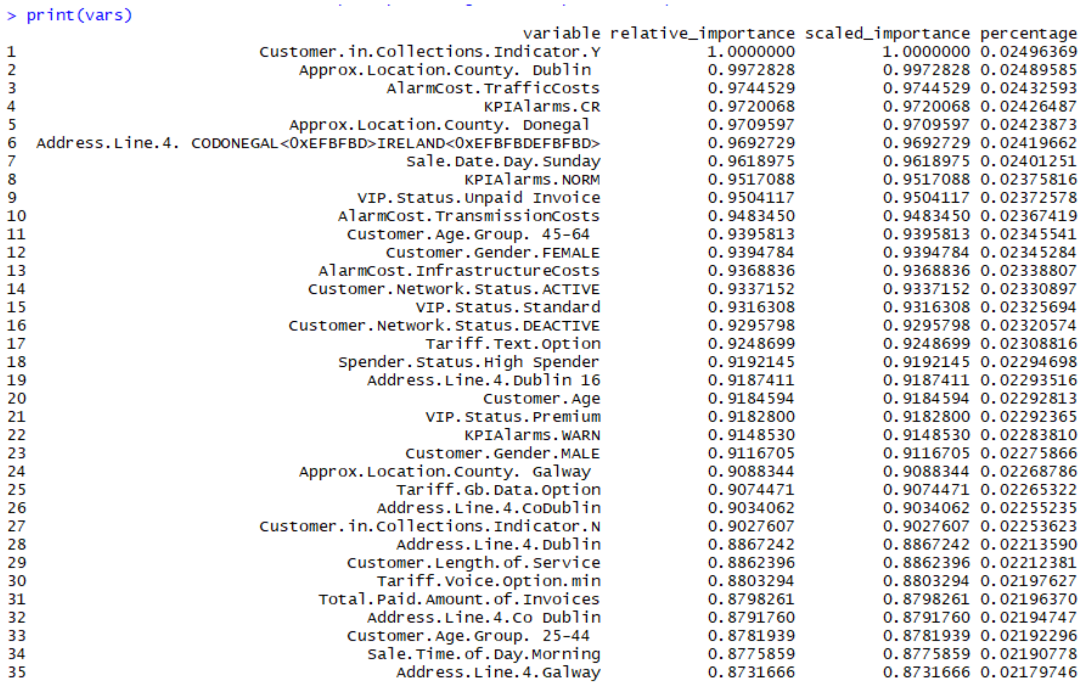


```
dl1 <- h2o.deeplearning(x = x, y = y, distribution = "multinomial",
                        training_frame = train,
                        nfolds = nfolds,
                        fold_assignment = "Modulo",
                        keep_cross_validation_predictions = TRUE)
```

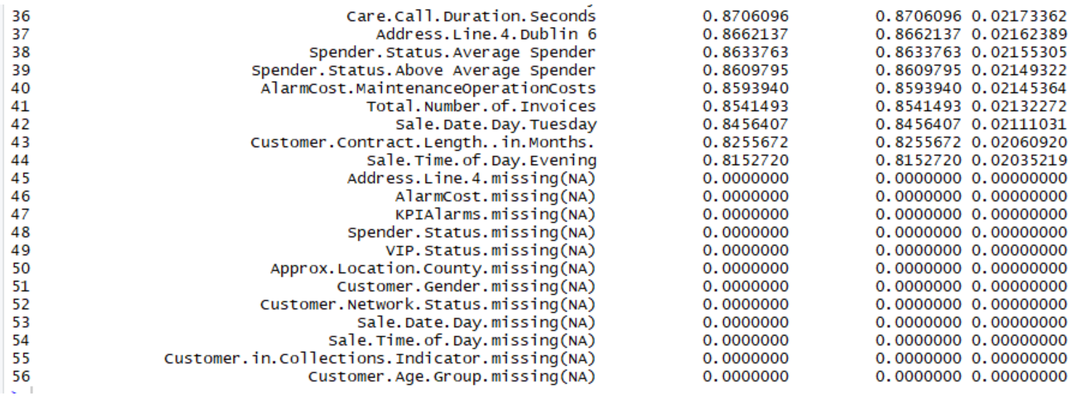


```
H2OMultinomialMetrics: deeplearning
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **

Cross-Validation Set Metrics:
=====================

Extract cross-validation frame with `h2o.getFrame("RTMP_sid_ac18_59")`
MSE: (Extract with `h2o.mse`) 0.000585796
RMSE: (Extract with `h2o.rmse`) 0.02420322
Logloss: (Extract with `h2o.logloss`) 0.00391504
Mean Per-Class Error: 0
Hit Ratio Table: Extract with `h2o.hit_ratio_table(<model>,xval = TRUE)`
=======================================================================
Top-4 Hit Ratios:
  k hit_ratio
1 1  1.000000
2 2  1.000000
3 3  1.000000
4 4  1.000000


Cross-Validation Metrics Summary:
                                mean           sd   cv_1_valid   cv_2_valid   cv_3_valid   cv_4_valid   cv_5_valid
accuracy                         1.0          0.0          1.0          1.0          1.0          1.0          1.0
err                              0.0          0.0          0.0          0.0          0.0          0.0          0.0
err_count                        0.0          0.0          0.0          0.0          0.0          0.0          0.0
logloss                 0.0039150403  0.002615707 0.0042144204  0.010159543 2.0491461E-4 0.0049018743  9.444945E-5
max_per_class_error              0.0          0.0          0.0          0.0          0.0          0.0          0.0
mean_per_class_accuracy          1.0          0.0          1.0          1.0          1.0          1.0          1.0
mean_per_class_error             0.0          0.0          0.0          0.0          0.0          0.0          0.0
mse                       5.85796E-4 4.4319496E-4 0.0010245934 0.0015914092  3.559213E-7 3.1245442E-4 1.6722045E-7
r2                        0.99907357 6.9977506E-4     0.998428    0.9974611    0.9999994   0.99947965   0.99999976
rmse                     0.018116727  0.011348573   0.03200927   0.03989247  5.965914E-4   0.01767638   4.08926E-4
>
```

VARIABLE IMPORTANCE for DEEP LEARNING

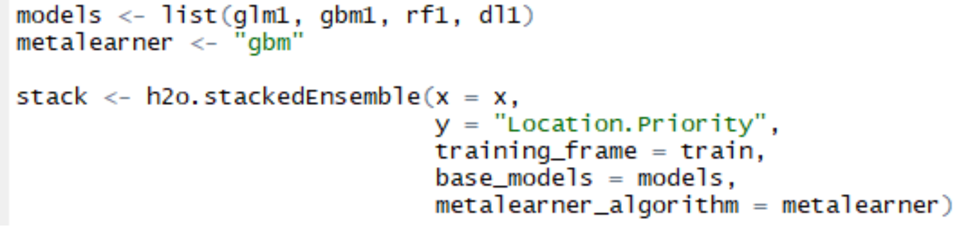


```
> print(vars)
                                                     variable relative_importance scaled_importance percentage
1                         Customer.in.Collections.Indicator.Y           1.0000000         1.0000000 0.02496369
2                             Approx.Location.County. Dublin            0.9972828         0.9972828 0.02489585
3                                      AlarmCost.TrafficCosts           0.9744529         0.9744529 0.02432593
4                                                KPIAlarms.CR           0.9720068         0.9720068 0.02426487
5                            Approx.Location.County. Donegal            0.9709597         0.9709597 0.02423873
6  Address.Line.4. CODONEGAL<0xEFBFBD>IRELAND<0xEFBFBDEFBFBD>           0.9692729         0.9692729 0.02419662
7                                        Sale.Date.Day.Sunday           0.9618975         0.9618975 0.02401251
8                                              KPIAlarms.NORM           0.9517088         0.9517088 0.02375816
9                                   VIP.Status.Unpaid Invoice           0.9504117         0.9504117 0.02372578
10                                AlarmCost.TransmissionCosts           0.9483450         0.9483450 0.02367419
11                                 Customer.Age.Group. 45-64            0.9395813         0.9395813 0.02345541
12                                     Customer.Gender.FEMALE           0.9394784         0.9394784 0.02345284
13                              AlarmCost.InfrastructureCosts           0.9368836         0.9368836 0.02338807
14                             Customer.Network.Status.ACTIVE           0.9337152         0.9337152 0.02330897
15                                        VIP.Status.Standard           0.9316308         0.9316308 0.02325694
16                           Customer.Network.Status.DEACTIVE           0.9295798         0.9295798 0.02320574
17                                         Tariff.Text.Option           0.9248699         0.9248699 0.02308816
18                                Spender.Status.High Spender           0.9192145         0.9192145 0.02294698
19                                   Address.Line.4.Dublin 16           0.9187411         0.9187411 0.02293516
20                                               Customer.Age           0.9184594         0.9184594 0.02292813
21                                         VIP.Status.Premium           0.9182800         0.9182800 0.02292365
22                                             KPIAlarms.WARN           0.9148530         0.9148530 0.02283810
23                                       Customer.Gender.MALE           0.9116705         0.9116705 0.02275866
24                            Approx.Location.County. Galway            0.9088344         0.9088344 0.02268786
25                                      Tariff.Gb.Data.Option           0.9074471         0.9074471 0.02265322
26                                    Address.Line.4.CoDublin           0.9034062         0.9034062 0.02255235
27                        Customer.in.Collections.Indicator.N           0.9027607         0.9027607 0.02253623
28                                      Address.Line.4.Dublin           0.8867242         0.8867242 0.02213590
29                                 Customer.Length.of.Service           0.8862396         0.8862396 0.02212381
30                                    Tariff.Voice.Option.min           0.8803294         0.8803294 0.02197627
31                              Total.Paid.Amount.of.Invoices           0.8798261         0.8798261 0.02196370
32                                   Address.Line.4.Co Dublin           0.8791760         0.8791760 0.02194747
33                                 Customer.Age.Group. 25-44            0.8781939         0.8781939 0.02192296
34                                   Sale.Time.of.Day.Morning           0.8775859         0.8775859 0.02190778
35                                      Address.Line.4.Galway           0.8731666         0.8731666 0.02179746
```

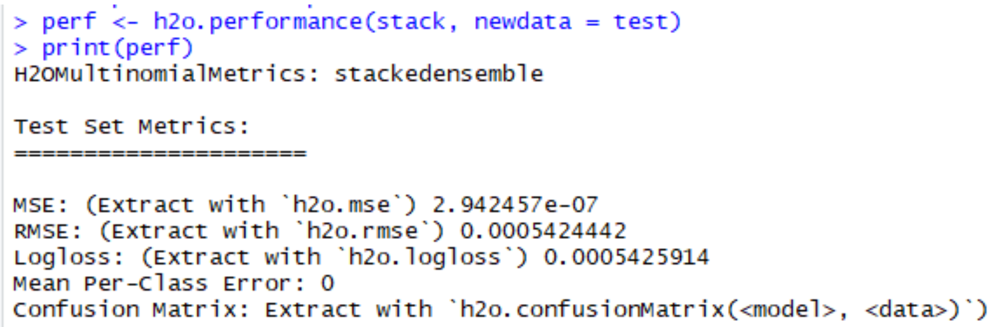


```
36                 Care.Call.Duration.Seconds      0.8706096      0.8706096 0.02173362
37                   Address.Line.4.Dublin 6       0.8662137      0.8662137 0.02162389
38              Spender.Status.Average Spender      0.8633763      0.8633763 0.02155305
39        Spender.Status.Above Average Spender      0.8609795      0.8609795 0.02149322
40         AlarmCost.MaintenanceOperationCosts      0.8593940      0.8593940 0.02145364
41                   Total.Number.of.Invoices      0.8541493      0.8541493 0.02132272
42                    Sale.Date.Day.Tuesday        0.8456407      0.8456407 0.02111031
43          Customer.Contract.Length..in.Months.    0.8255672      0.8255672 0.02060920
44                   Sale.Time.of.Day.Evening      0.8152720      0.8152720 0.02035219
45                 Address.Line.4.missing(NA)      0.0000000      0.0000000 0.00000000
46                      AlarmCost.missing(NA)      0.0000000      0.0000000 0.00000000
47                      KPIAlarms.missing(NA)      0.0000000      0.0000000 0.00000000
48                 Spender.Status.missing(NA)      0.0000000      0.0000000 0.00000000
49                     VIP.Status.missing(NA)      0.0000000      0.0000000 0.00000000
50         Approx.Location.County.missing(NA)      0.0000000      0.0000000 0.00000000
51                Customer.Gender.missing(NA)      0.0000000      0.0000000 0.00000000
52        Customer.Network.Status.missing(NA)      0.0000000      0.0000000 0.00000000
53                  Sale.Date.Day.missing(NA)      0.0000000      0.0000000 0.00000000
54               Sale.Time.of.Day.missing(NA)      0.0000000      0.0000000 0.00000000
55 Customer.in.Collections.Indicator.missing(NA)  0.0000000      0.0000000 0.00000000
56            Customer.Age.Group.missing(NA)      0.0000000      0.0000000 0.00000000
```

7. SUPER LEARNER – STACKED ENSEMBLE

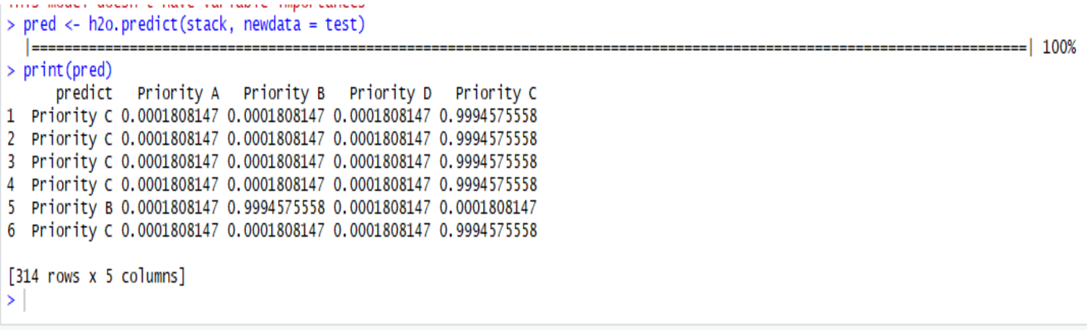


```
models <- list(glm1, gbm1, rf1, dl1)
metalearner <- "gbm"

stack <- h2o.stackedEnsemble(x = x,
                             y = "Location.Priority",
                             training_frame = train,
                             base_models = models,
                             metalearner_algorithm = metalearner)
```



```
> perf <- h2o.performance(stack, newdata = test)
> print(perf)
H2OMultinomialMetrics: stackedensemble

Test Set Metrics:
=====================

MSE: (Extract with `h2o.mse`) 2.942457e-07
RMSE: (Extract with `h2o.rmse`) 0.0005424442
Logloss: (Extract with `h2o.logloss`) 0.0005425914
Mean Per-Class Error: 0
Confusion Matrix: Extract with `h2o.confusionMatrix(<model>, <data>)`)
```

Prediction:



```
This model doesn't have variable importances
> pred <- h2o.predict(stack, newdata = test)
  |=====================================================================================================================| 100%
> print(pred)
     predict   Priority A   Priority B   Priority D   Priority C
1  Priority C 0.0001808147 0.0001808147 0.0001808147 0.9994575558
2  Priority C 0.0001808147 0.0001808147 0.0001808147 0.9994575558
3  Priority C 0.0001808147 0.0001808147 0.0001808147 0.9994575558
4  Priority C 0.0001808147 0.0001808147 0.0001808147 0.9994575558
5  Priority B 0.0001808147 0.9994575558 0.0001808147 0.0001808147
6  Priority C 0.0001808147 0.0001808147 0.0001808147 0.9994575558

[314 rows x 5 columns]
> 
```

And even more text