DATA MINING AND PREDICTIVE ANALYTICS APPLICATION ON CELLULAR NETWORKS TO MONITOR AND OPTIMIZE QUALITY OF SERVICE AND CUSTOMER EXPERIENCE

by:

MUWAWA JEAN NESTOR DAHJ

submitted in accordance with the requirements
for the degree of

MAGISTER TECHNOLOGIAE

In the Subject

ELECTRICAL ENGINEERING

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: DR KINGSLEY A. OGUDO

NOVEMBER 2018

SUMMARY

This research study focuses on the application models of Data Mining and Machine Learning covering cellular network traffic, in the objective to arm Mobile Network Operators with full view of performance branches (Services, Device, Subscribers). The purpose is to optimize and minimize the time to detect service and subscriber patterns behaviour. Different data mining techniques and predictive algorithms have been applied on real cellular network datasets to uncover different data usage patterns using specific Key Performance Indicators (KPIs) and Key Quality Indicators (KQI). The following tools will be used to develop the concept: R-Studio for Machine Learning and process visualization, Apache Spark, SparkSQL for data and big data processing and clicData for service Visualization. Two use cases have been studied during this research. In the first study, the process of Data and predictive Analytics are fully applied in the field of Telecommunications to efficiently address users' experience, in the goal of increasing customer loyalty and decreasing churn or customer attrition. Using real cellular network transactions, prediction analytics are used to predict customers who are likely to churn, which can result in revenue loss. Prediction algorithms and models including Classification Tree, Random Forest, Neural Networks and Gradient boosting have been used with an exploratory Data Analysis, determining relationship between predicting variables. The data is segmented in to two, a training set to train the model and a testing set to test the model. The evaluation of the best performing model is based on the prediction accuracy, sensitivity, specificity and the Confusion Matrix on the test set. The second use case analyses Service Quality Management using modern data mining techniques and the advantages of in-memory big data processing with Apache Spark and SparkSQL to save cost on tool investment; thus, a low-cost Service Quality Management model is proposed and analyzed. With increase in Smart phone adoption, access to mobile internet services, applications such as streaming, interactive chats require a certain service level to ensure customer satisfaction. As a result, an SQM framework is developed with Service Quality Index (SQI) and Key Performance Index (KPI). The research concludes with recommendations and future studies around modern technology applications in Telecommunications including Internet of Things (IoT), Cloud and recommender systems.

DECLARATION

I declare that **DATA MINING AND PREDICTIVE ANALYTICS APPLICATION ON CELLULAR NETWORKS TO MONITOR AND OPTIMIZE QUALITY OF SERVICE AND CUSTOMER EXPERIENCE** is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I submitted the thesis/dissertation to originality checking software and that it falls within the accepted requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at Unisa for another qualification or at any other higher education institution.

_____ November 10, 2018

SIGNATURE DATE

COPYRIGHT CLASSIFICATION

DEDICATION

I dedicate this dissertation to God Almighty who gave me the inspiration, strength and intelligence to tackle this hot topic of today's century. This dissertation is also dedicated to my Parents Tjoppen Muwawa and Isamanga Therese for their belief in me.

ACKNOWLEGDEMENTS

Firstly, I would like to thank my Supervisor Dr. Kingsley of the Electrical Engineering Department for his unlimited support and direction throughout this research; His door was always opened whenever I needed assistance. He has consistently allowed this dissertation to become my own and individual work. I thank all the academic figures and experts who have been involved in the realization and validation of these research projects.

I would also like to thank the following people:

1. My brother Muwawa Salam Smart and Sister Muwawa Mireille for their mental and spiritual support through my studies and research, without who this could not have been achieved.

2. Mss. Divine Mutombo and Karice Beghela for their encouragement and moral support through the draft of this dissertation, for always pushing me to complete my task on time.

3. UNISA for granting me the opportunity to study and all support regarding the research and publication during my study period.

4. Mr. Minerve Mampaka, my colleague and business partner with whom we have shared many research ideas around new technologies. Without whom, it would have been quite difficult to be where I am with my dissertation.

5. Mr. Bagula Patrick for his facilitation of tools and resources to help me complete this course. And everyone who in a way or another contributed to this research study, including my colleagues, friends and extended family.

ABSTRACT

Cellular networks have evolved and are still evolving, from traditional GSM (Global System for Mobile Communication) Circuit switched which only supported voice services and extremely low data rate, to LTE all Packet networks accommodating high speed data used for various service applications such as video streaming, video conferencing, heavy torrent download; and for say in a near future the roll-out of the Fifth generation (5G) cellular networks, intended to support complex technologies such as IoT (Internet of Things), High Definition video streaming and projected to cater massive amount of data. With high demand on network services and easy access to mobile phones, billions of transactions are performed by subscribers. The transactions appear in the form of SMSs, Handovers, voice calls, web browsing activities, video and audio streaming, heavy downloads and uploads. Nevertheless, the stormy growth in data traffic and the high requirements of new services introduce bigger challenges to Mobile Network Operators (NMOs) in analysing the big data traffic flowing in the network. Therefore, Quality of Service (QoS) and Quality of Experience (QoE) turn in to a challenge. Inefficiency in mining, analysing data and applying predictive intelligence on network traffic can produce high rate of unhappy customers or subscribers, loss on revenue and negative services' perspective. Researchers and Service Providers are investing in Data mining, Machine Learning and AI (Artificial Intelligence) methods to manage services and experience. This research study focuses on the application models of Data Mining and Machine Learning covering network traffic, in the objective to arm Mobile Network Operators with full view of performance branches (Services, Device, Subscribers). The purpose is to optimize and minimize the time to detect service and subscriber patterns behaviour. Different data mining techniques and predictive algorithms will be applied on cellular network datasets to uncover different data usage patterns using specific Key Performance Indicators (KPIs) and Key Quality Indicators (KQI). The following tools will be used to develop the concept: R-Studio for Machine Learning, Apache Spark, SparkSQL for data processing and clicData for Visualization.

KEY TERMS

Data Mining; Predictive Analytics; Big Data; Quality of Service (QoS); Customer Experience; Business Intelligence (BI); Network Churn; Key Quality Index (KQI); Key Performance Index (KPI); Service Quality Management (SQM); Neural Network (NN); Deep Learning (DL); Random Forest (RF); Classification Tree; Regression; In-memory Data processing; Data Science.

LIST OF PUBLICATIONS

1. D. M. J. Nestor and K. A. Ogudo, "Practical Implementation of Machine Learning and Predictive Analytics in Cellular Network Transactions in Real Time," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Durban, 2018, pp. 1-10.
   https://ieeexplore.ieee.org/document/8465476

2. K. A. Ogudo and D. M. J. Nestor, "Modeling of an Efficient Low Cost, Tree Based Data Service Quality Management for Mobile Operators Using in-Memory Big Data Processing and Business Intelligence use Cases," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Durban, 2018, pp. 1-8.
   https://ieeexplore.ieee.org/document/8465410

3. D.M.J Nestor and K.A. Ogudo, "Geo and Graph Analytics for Dynamic Cellular Transactions Insights, Improving Quality of Service and Business Decisions: Quality X Map". Submitted to *International Conference on Intelligent and Innovative Computing Applications* (ICONIC 2018), Plaine Magnien, Mauritius, December 6-7. 2018

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BI	Business Intelligence
BPM	Business Performance Management
BSC	Base Station Controller
BTS	Base Transceiver Station
CDR	Call Data Records
CEM	Customer Experience Management
CRISP-DM	Cross-industry standard process for data mining
CRM	Customer Relationship Management
CS	Circuit-Switch
CSP	Communication Service Provider
DNS	Domain Name Server
ETSI	European Telecommunications Standards Institute
FTP	File Transfer Protocol
GGSN	Gateway GPRS Support Node
GSM	Global System for Mobile Communication
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
ITU	International Telecommunication Union
KPI	Key Performance Indicator
KPI	Key Performance Indicator
KQI	Key Quality Indicator
MNO	Mobile Network Operator

MSC	Mobile Switching Center
PS	Packet-Switch
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RDBMS	Relational Database Management System
RNC	Radio Network Controller
ROC	Receiving Operating Characteristics
RTT	Round Trip Time
SGSN	Serving GPRS Support Node
SQI	Service Quality Index/Indicator
SQL	Structured Query Language
SQM	Service Quality Management
TCP	Transmission Control Protocol
UE	User Equipment
UP	User Plane

TABLE OF CONTENTS

TABLE OF FIGURES

TABLES

CHAPTER 1. INTRODUCTION AND BACKGROUND

To protect and keep a stable business revenue stream, Network Mobile Operators need to strive for perfection on network quality with reduced or no customer affecting subjects, using effective techniques. Thus, ensuring a very good reference of customer satisfaction by putting a big accent on Quality, be it of Service, Experience, Device or Network (QoX). With Big transactions' Data generated by users on daily basis, and with the projection on number of devices to be connected soon, Network Operators need transformation in their business models. Communication Service Providers (CSP) focus on different behavioural patterns in network traffic to pinpoint opportunities of service improvement and predict the likelihood of customers to terminate their contracts or/and move to a competitor [1]. CSPs have indeed managed to shape robust IT platforms which efficiently store subscriber transactions and any other traffic originated from devices and customers. Illustrations of such platforms include the Customer Relationship Management (CRM), Billing systems, Intelligent Network systems and so on. With the current rate of data application usage such as WhatsApp, Skype, Instagram and other Over-The-Top (OTT) applications, on top of the traditional voice services, CSPs clearly highlight the need for business model adjustment. Coupled with the increasing adoption of Smartphones and other smart devices, more network resources need to be added and undoubtedly more human technical expertise, to manipulate the humongous transactional data that will be generated; All in the road to maintaining good quality of service and experience.

Also, the more CSPs focus on QoS, the complex the entire business process becomes due to the exponential growth in connected devices. The coming of 5G, which will bring a more flexible architecture to the current Cellular Network architecture, aims to address challenges to come, IoT, Machine Type Communications, Virtualization. A study by Ericsson shows a projection of approximately 50 Billion devices to be connected by 2020 [2]; The question now is not to provide technological IoT platforms only, but to also develop intelligent mechanisms to draw values out the Big Data generated by the connected devices. Hence, the need to bring forth Data Mining and Predictive Analytics in Telecommunications.

Positioned at the crossroad of Statistics, Machine Learning and Computer Science, Data mining brings more value to the Analytics process by exploring the different traffic patterns that can guide in improving business decisions. Another essential facet of telecommunications data analytics is the visual portrayal of the outcome, which opens another hole of research on efficient data visualization and representation. For instance, after exploring traffic patterns on

analysed transactions, the result needs to be portrayed in such a way that everyone in the Organization needs to understand the significance of each report or dashboard to simplify decision making; So, choosing the best way to present the analysis result to the end users is as important as mining the data itself. Organizations need to move in the direction of Business Intelligence (BI) systems to explore and represent in a flexible manner, the insight of processed large datasets. The dynamic depiction of data generated by humans and machines, involves an analytics methodology and mathematical models for information and knowledge retrieval processes to support easy and extremely complex business decisions [3]. This research study presents a practical, simplistic and efficient system models to support various decisions in the Telecommunications Environment.

1.1. Context of the Study

Millions of people today own Mobile devices for calling or intensively using internet services and they generate huge amount of transactions in Mobile Networks every day if not every hour. With the expansion of IoT (Internet of Things) and Machine to Machine Communication (MTC), even more resource intensive transactions are passed through the Mobile Networks. Those transactions are all stored in complex storage systems in the form of structured data systems such as Customer Relationship Management (CRM), storing customer related information and unstructured data system storing pictures, videos and pure voice traffic [4]. Despite the complex IT infrastructures put in place, one of the pain points of Communication Service Providers (CSP) is to ensure very good quality of service for all the users. By good Service Quality, we call high number of satisfied customers and low rate of attrition or contract terminations. The scope of this research study involves two very important aspects of science and technology, highly on demand today.

In one aspect, the notion of QoS and QoE improvement in the Mobile Network environment is tackled through an SQM (service Quality Management) model, an area in which international organizations such as the Institute of Electrical and Electronics Engineers (IEEE), the International Telecommunication Union (ITU), the European Telecommunications Standard Institute (ETSI) and others including the 3rd Generation Partnership Project (3GPP), the Multimedia Communication Forum are spending comprehensive research time to provide the required basic Standards to govern QoS and QoE in Network environment.

The second aspect of the study brings the concept of applied Data Science and Predictive measures to efficiently improve the first concept. The hikes of Big Data Analytics, Artificial Intelligence (AI) and Predictive Machine Learning do not exempt the Telecommunications environment, instead, become critical components of the environment to improve customer experience and service experience. In this research, we explore a basic hierarchical tree model, leaning on Big Data Analytics, Predictive Analytics using some of the popular algorithms such as Regression, Random Forest, Boosting Tree, Neural Networks and Deep Learning (NN) and Business Intelligent (BI) methodology to methodically output the result of the model, in the aim to give Communication Service Providers (CSPs) a skeletal framework to apply all the mentioned technologies in their business environments. Two practical case studies are explored in this dissertation to go along with the theoretical background:

- A practical approach in the application of Machine Learning in Telecommunications Network transactions, using different algorithms. In this case study, transaction information from the CRM system is collected and analysed to find useful patterns and presented for decision support process. The use case addresses one of the hottest topics of any CSP: the prediction of customers that could churn (terminate their contract with one CSP). This will assist CSPs to protect customers [1].
- A low-cost model is explored for an efficient Service Quality Management system based on internet Data traffic. Using a tree model in this case, a navigational Analysis is performed on Network data of popular application services such as Interactive services, Streaming Services and Background Services [5].

In the context of this research study, we leverage on existing systems and platforms, providing a more practical methodology, facilitating the research reproducibility. We involve Big Data Analytics with in-memory data processing to increase the speed of data retrieval and discuss the concept for real time application analysis.

1.2. Conceptual Background

In comparison to the Second Generation and the third Generation of Cellular Mobile Network (2G and 3G respectively), the Fourth Generation has given the users (customers) the freedom to maximize data usage by intensively using resource demanding applications such as video and audio streaming, video-conferencing and other real time applications. Nevertheless, with more high demanding traffic increasing, Quality of Service (QoS) and Quality of Experience (QoE) are affected. The high use of data related services by Mobile users is an indication for CSPs to shift the business towards Packet data. With high usage of Over-The-Top (OTT) services, Data quality becomes a key parameter for customer satisfaction. In association to the high usage of services, is the current adoption of Smart devices. More devices, more services used, more data generated, more **technical expertise** required to manage Network Quality. For the reason elaborated above, Communication Service Providers (CSPs) appear to be investing heavily on expensive Customer Experience and Service Quality platforms to remain competitive in the market. The lofty investment is backed-up by the integration and deployment of new technologies including but not limited to Big Data, Real Time Data Processing Solutions, Machine Learning, Predictive Analytics, Business Intelligence and Smart visualization platforms. Be that as it may, the Return on Investment (ROI) of such expensive and explosive QoS/QoE systems is not as evident as the functionalities and integrated business rules.

In this research study, two important conceptual studies are addressed based on the concept illustrated in the above paragraph, to give CSPs a basic framework to transformation of CSPs business.

- Provide a Practical Approach of Predictive Analytics in the Telecommunication environment.
- Model of low-cost Service Quality Management System, tackling cost reduction.

1.2.1. Why Quality of Service (QoS)?

The use of Cellular Mobile devices has become an important factor in human life as they are becoming more and more human companions. Massive use of mobile internet on social media, video streaming and other protocol applications is attracting everyone to have a mobile phone. Mobile phones use SIM cards which are the main chip, provided by Mobile Network Operators

(MNOs) to access different network services. Ensuring and anticipating user's Quality of Service needs, is the factor that draw the line between service providers and their competitions. The framework proposed by the ITU for QoS is based on a 3x3 Matrix, using speed, accuracy, and dependability as parameters. These parameters are then used to evaluate the quality with which the basic user functions of connection set-up, user information transfer, and connection release are executed. By following closely and efficiently the requirements of the framework for telecommunications services, CSPs should have the user quality of experience facilitated [6].

The quality of a service is a great separator in the Mobile business market. Its parameters and measures are very critical in providing indication of how well a specific service is or behaves, and therefore, it becomes an important selector of offered services by different service providers. In the case of equal service billing and features, quality becomes the differentiator for network users, as well as, service providers can make use of quality to have an image of a "respected" provider [7]. Today mobile users are switching from one operator to the other due to poor services complaining about poor internet speed, not being able to watch YouTube, unable to live stream soccer game on the smart phones. And this is because Quality of Service is not looked at efficiently.

1.2.2. Why Data Science and Predictive Analytics?

Data mining and prediction Analytics fields are attracting increasing interest from scientists and technologists who want to solve real-life problems. Data mining refers to the science of finding useful information or patterns in a dataset stored in different ways, structured or unstructured. In processes of Data Mining, the point of interest is not very specified at priori and often the data is searched by Analytics. The concept of "useful information" in data mining depends on the context of the domain in which the science is applied and different objectives that are set. The aspects of mining in different protocols layers are distinguished in two ways as shown by Azzalini et al. [4] **:**

- ***Global behaviour*** of the phenomenon examined in the data. The objective in this case is to construct a global model, taken from the available data.
- Characterization of details or the ***pattern structure*** of the data. In this case the interest is outside the standard behaviour. It means the point of interest when analysing pattern

structure in data is exterior to the standard behaviour of data. The objective is to identify various variances in the structure instead of looking at the standard known behaviours.

Cellular Network traffic is collected, examined and analysed in different layers to identify trends that allows the Communication Service Provider (CSP) to forecast customers' behaviour or experience according to their data usage, voice calls, geographical positions and other elements, reducing the gap between the customers and the Service provider. The subject on importance of Data Mining and Predictive Analytics is clarified in the statements of D.J Hand [8] in which Data mining is presented as an applied discipline, requiring a wide span of knowledge to understand both statistical and computational issues. Predictive Analytics is practically the new way to improve customer experience by studying their behaviours.

One of the challenges of Service Providers have always been to know what customers exactly want, which give competitive advantage and increase customer loyalty. Different customers have different needs and different perception of the same services. Data Mining and Predictive Analytics allows customer's customization of needs, moving the business towards a customer-centric services as described by J. Betser [9], a system in which customer preference plays the crucial role. Several methods are used from simple classification and categorization algorithms to more complex forecasting and deep learning algorithms through which customers are clustered in line with their service interest, service usage leaning on their past service usage patterns.

Associating Data Mining and Predictive Analytics to the Telecommunications Environment provides to the CSPs (Customer Service Providers) a glimpse of what to invest on in the future. For example, by analysing **YouTube, ShowMax, Netflix, Black** application usage across the country in the past, a CSP can predict the group of customers or regions which is likely to adopt a new video-streaming or content provider service. This will help the CSP to target the correct market, tailor the service to suit the targeted group of customers.

1.2.3. Big Data in the Picture

Activities or transactions done by the users in a Cellular Network can be structured or unstructured, they can have **different natures** as some users may focus on browsing only, others on video-streaming, others on audio-streaming and pictures. With the high adoption of Smartphones in the last decade and a projected increase in number in the future, **billions** of

transactions are being generated by the users in the Network in terms of voice calls and internet usage. The figure below shows the adoption of Smartphones in South Africa from 2014 to 2022 as published by [10]. Depending on the applications and the service type of the CSPs, data **retrieval speed** is very capital in building business decision models.

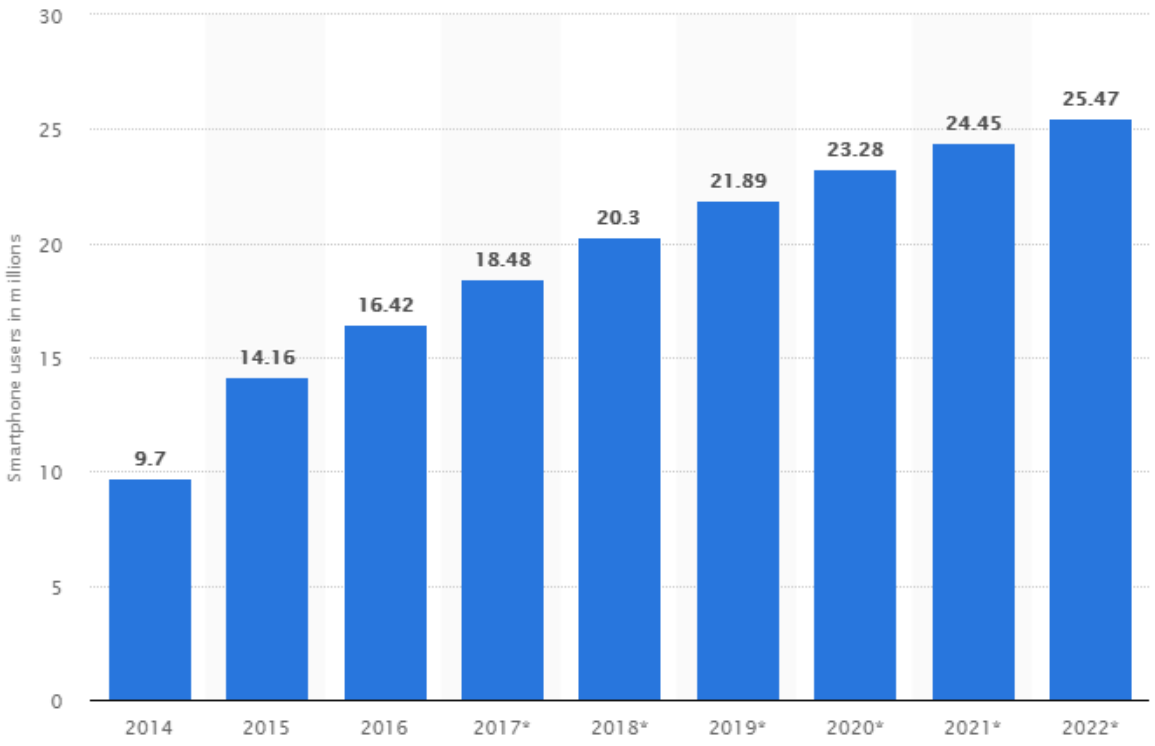


Figure 1 Smartphone Adoption in South-Africa 2014-2022 by Statista

Imagining a query which takes more than 5 minutes to run in the system database to retrieve information could be a big nightmare for a CSP because it can delay the business process; the heterogeneousity, the amount of transactions generated by the customers and the fast requirement to retrieve information already characterized Big Data's three Vs (Variety, Volume and Velocity). In this study, different Big Data Solutions are looked at and one of them is applied along with some traditional data manipulation algorithms.

Volume	Velocity	Variety
• Amount of Data • In the order of 100s of Gigabytes.	• Retrieval speed • Real time applications	• Data Diversity • Video, pictures, texts, binaries. • Data stored in different file format

Figure 2 The 3 Vs of Big Data Characteristics

1.2.4. Why Business Intelligence?

Data Analytics facilitates Data manipulation and pattern analysis. But one of the biggest concerns has been how to represent the results of Analysis in an efficient and understanding way to ease the decision-making process. Through Business Intelligence, the presentation of relevant dashboards and reports as described by [11] **is** used to convert or transform data into actionable information. Through BI systems, productivity is improved, saving time to take decisions. In the Telecommunications arena, the concept "performance" (Quality) is defined by several Key Performance Indicators (KPI) which defines the metric used to measure the performance. Going beyond the scope of KPIs, BI systems also introduce the concept of smart visualizations such as Scorecards and complex dashboards. Figure 3 illustrates the basic concept of BI which is also used in this study.

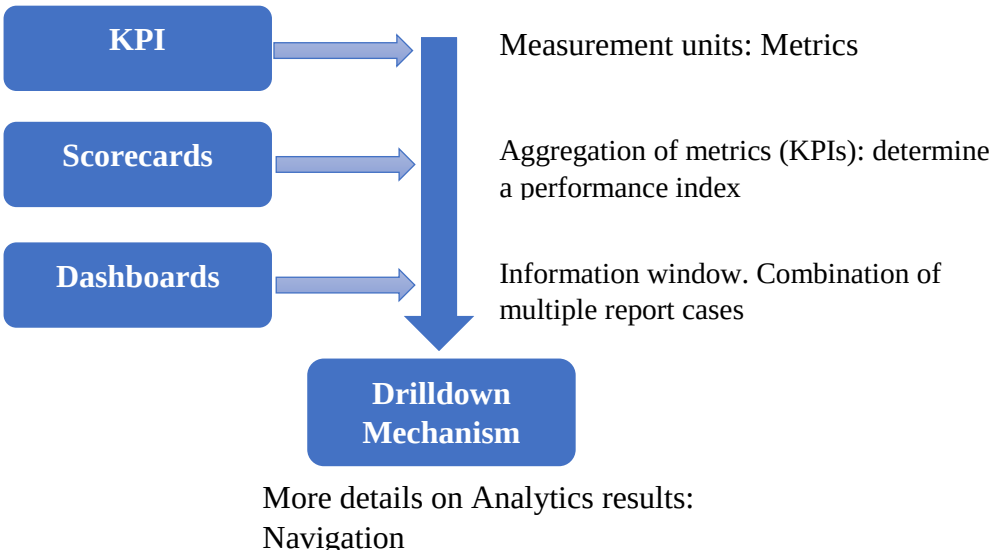


Figure 3 Business Intelligence Basic Concept

The Business Intelligence also introduces the concept of drilldown or navigation from one dashboard or report to another in the objective to get more details which can support the decision-making process. The Cellular Network is made of different interfaces and protocols. Defining the Quality of Service is not as straight forward as it seems because customer transactions can have different identities in different interfaces of the Network. Therefore, when doing Analytics on Cellular transactions it is important to define the interface at which the Analytics is done, facilitating the representation of data. In this Study, we have also introduced the concept of SQI (Service Quality Index) along with the rest of the BI components.

1.3. Global Requirements

1.3.1. Hardware Requirements

To conduct the research study in the area proposed, a list of requirements on the Hardware and operating system software has been put in place. All the use cases built in the scope of the research requires high computational power. The Hardware used for all the use cases is at minimum as follow (high performing PC):

- Processor Intel® Core i7-5500U 2.4GHz boost up to 3.0 GHz
- NVIDIA 920M with 2 GB Dedicated VRAM.
- RAM: 16GB DDR3 Memory
- 1TB Storage Disk.
- Operating System: Windows10.

The processing power of the Hardware limits the amount of data that can be processed in memory (Data < 16 GB) and stored in the disk (Data < 1TB).

1.3.2. Software Requirements

In order to build an efficient low-cost Data Analysis and predictive models, a good knowledge on Data Mining platforms and programming languages are required. In the same line as section

1.3.1, a list of Software is also stipulated in this section to work on the experiment. The following list of software is used during the research for Data manipulation and processing:

- MySQL: For data manipulation based on the popular traditional Sequence Query Language.
- R & R-Studio: a modern platform for Data Analysis based on the R-Project team. The R platform is a solid Open-Source platform for Data Analysis and statistics. It is based on the R language [12].
- Apache-Spark: with the growth of data in the Network and high usage of Mobile phones, the ideal way to process or mine data is by exploring the advantages of distributed computing architecture such as Apache Spark and explore the in-memory capability of the system [13].
- ClicData BI tool: Data representation is one of the key aspect of Data Analysis, including Telecommunications Data Analysis. This is because the end result is what is used for business decisions. For this research, a low-cost BI tool is used to represent data graphically and allows navigation from one layer to another.

CHAPTER 2. LITERATURE REVIEW

QoS as a concept, in Telecommunications is reviewed and defined by many researchers and boards of standards such as the ITU, IEEE… And researchers have also applied many models and algorithms to address QoS in different network domains. QoS is described as the degree of satisfaction of a service usage during a given communication session. Consistently monitoring and improving the QoS is the point that distinguishes successful communications service and network operators from their competitors [6]. When talking about QoS, the parameter traffic Class needs to be taken into consideration. The QoS mechanisms provided in the Cellular Network must be efficient with reasonable resolution. Depending on the way latency is handled, four QoS classes are described: Streaming, Conversational, Background and Interactive classes. The provision of QoS for data applications in a cellular network imposes a series of challenges because of the unreliable wireless channels and the mobility of mobile devices [14]. The measurement of QoS is based on parameters like delay, jitter, packet loss, throughput and many others, depending on the application and management scheme [7].

2.1. QoS Concept Overview

2.1.1. Quality of Service Overview

End user experience satisfaction is the main objective of QoS from a Communication Service Provider perspective. Service requirements can differ from one service to another, depending on the type, the nature and the amount of Network resources utilized by the service in question. A typical illustration is listening to music online and sending a mail to a friend. In terms of basic requirements, it can be stated that listening to music on the internet is very sensitive to delay then sending a mail to a friend. However, sending an email requires a high reliability then internet audio streaming. This shows that different services yield different quality requirements, which a certain category tolerating delays and others not. Another illustration would be tolerance to error. Certain services do not tolerate error in transmission (packet loss for example); a service application such as File transfer could have no meaning to the destination if there are errors in the transmission or loss of packets. In this case, there could be a need to retransmit packets to ensure that integrity of the file is maintained. R. Rodriguez et al. identify the above overview of QoS as the starting point for several developments and

researches, leading to the setting up of mechanisms and protocols to differentiate service requirements [15].

R. Rodriguez et al. in their book section classify the QoS in two, quantitative QoS and qualitative QoS depending on the methodology followed by the CSP.

- **Qualitative QoS**: focuses on priority of service in the Network. for instance, based on their requirements and user experience, traffic priority could be given to YouTube streaming then e-mail, providing faster traffic flow for YouTube.
- **Quantitative QoS**: focuses on measuring network capabilities. It guarantees certain performance level. Metrics are specified such as latency, packet loss, and throughput as data speed. While the study of R. Rodriguez and his associates have gone deep into both Quantitative and Qualitative QoS, the scope of our study is limited to measurable metrics, thus only developing on quantitative QoS.

2.1.1.1. QoS Services Class Categorization

When addressing QoS matter in a Network, it is crucial to group services in relation to their requirements. The 3GPP defines QoS classes as shown on the below table, considering limitation of the air interface. The Network treats or should at least treat services of the same QoS requirements accordingly.

Table 1 QoS Service Class

Service Class	**Characteristics**	**Applications**
Conversational	• Low delay. • Preserve time variation	• Voice Call • VoIP
Streaming	• Preserve time variation	Audio & Video Streaming
Interactive	• Request response pattern • Preserve payload content	• Web browsing • Instant Messaging
Background	• Not delay sensitive • Preserve payload content	• Email • Files download

2.1.1.2. QoS Levels

Users being able to effectively and efficiently stream a basketball match on the phone or tablet when connecting to a 3G or 4G network is a very important link between the service user and the Network Operator who provides that service. This means, as concluded by J.D. Power, that the operator with a strong focus on QoS is likely to attract a high number of customers than others [16]. To establish a robust QoS mechanism, the needs to situate the QoS in the general model is mandatory. The QoS in the general model can be situated in three levels as detailed in the work done by W.C. Hardy on QoS measurement and evaluation [17]:

- **Intrinsic QoS**: related to performance. In Intrinsic QoS, the focus is on the ability of the Network to provide basic services to the end-users. Network centricity of intrinsic QoS opens the flow to more NOC (Network Operation Centre) related applications and tools.
- **Perceived QoS**: related to customer perception of the quality of a specific used service. Many factors influence the perceived QoS including real customer experience. This QoS is measured in a non-technical way. The Service Level Agreement is the main assessor of this kind of QoS.
- **Assessed QoS**: build as an umbrella of the first two QoS levels, Assessed QoS relates to the probability of users to continue using specific services. This QoS makes the main driver of QoE (Quality of Experience). The requirements are then based on tangible network metrics such as delay, packet loss, jitter, and throughput.

2.1.2. QoS Monitoring Literature Review

This section is based on the work done by David Soldani et al. on QoE and QoS monitoring [18]. In their study, the perception of subscribers on provided services is analysed in terms of 3 parameters which are:

- **Integrity**: related to quantitative QoS metrics such as packet loss, delay and throughput.
- **Accessibility**: related to the availability of the service itself, the time it takes to set up connection to the Network for that service.
- **Retainability**: related to loss of connections in the signal flow across the Network.

The conceptual architecture of the QoS monitoring system as developed by their study, is shown in the figure 4. It is to be noted that although the entire architecture is shown, our research study focuses on one aspect of the architecture which is the **Service Quality Management**.

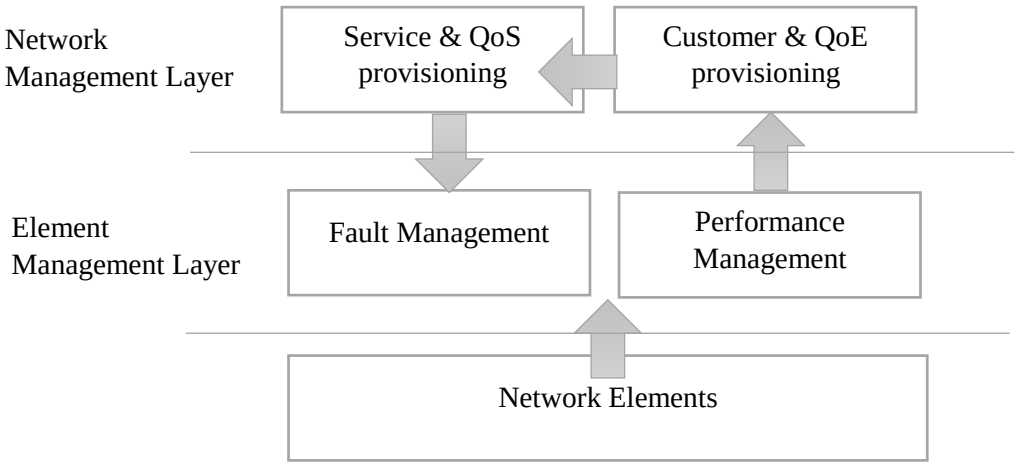


Figure 4 Conceptual Architecture of QoS as studied by David Soldani et al.

The study goes further with adding layers for data post-processing, visualization and drilldown mechanism on performance metrics used. The conceptual architecture as presented by D. Soldani, enhanced with the data processing is shown in the below Figure:

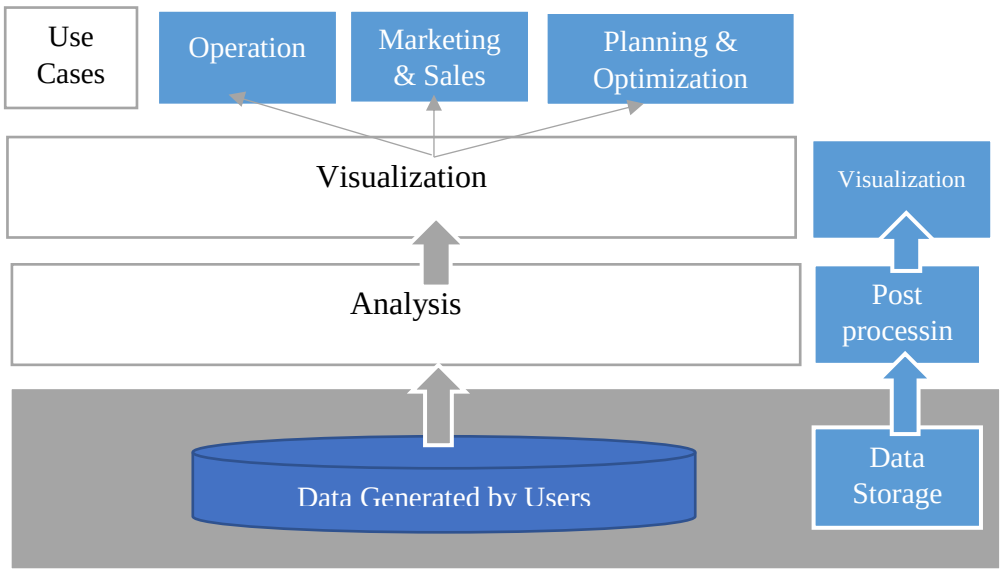


Figure 5 Illustration of Conceptual Architecture with Data post-processing

- The Data Storage is responsible for data collection and data storage.
- The Post-processing layer is the layer with possessing the intelligence for transaction data analysis.
- The visualization layer is responsible for the displaying of the data analysis results, with drilldown capabilities. This facilitates the investigation of network issues and help identify the unsatisfactory criteria by showing metrics separately.

Network management layer addresses two big questions: SQM (Service Quality Management) which takes performance and fault data as inputs and the Customer Experience management which takes as inputs the SQM data and the user data.

- The study also proposes a framework for QoS monitoring by formulating metrics that can be used to measure services performance across different domains of the network, carrying less on the content carried by upper layers' protocols.
- **Integrity monitoring in BSS**: the researchers introduce the metrics used to assess integrity of services belonging to a certain class.
    - **BLER (Block Error Rate)** monitoring: to determine how successful transmission is over the physical layer, which is nothing else than the ratio between the blocks of data with error over the total number of blocks sent over a transmission line. This is given by:

$$BLER_p = 1 - \frac{\sum_{i=1}^{MCS-9} N^P_{CorrectRxRLCblocks,i}}{\sum_{i=1}^{MCS-9} N^P_{TotalTxRLCblocks,\ i}} \qquad (2.1.2.1)$$

      Where $N_{CorrectRxRLCblocks}$ is the blocks of data correctly received over the transmission line and the $N_{TotalTxRLCblocks}$ is the total number of blocks of data sent over the transmission line for the used encoding techniques (MCS 1 to MCS 9) and p is related to the QoS Class.

    - **Throughput**: popularly known as number of bits delivered per seconds. This given by:

$$b^p = \sum_{k=CS-1}^{MCS-9} \sum_{i=1}^{N} r_k B_i^{p,k} \qquad (2.1.2.2)$$

Where b^p represents the total bits delivered correctly over the transmission line which is the parameter of interest in the computation of the throughput. And the time taken to deliver b^p , $D^p$is given by:

$$D^p = \sum_{i=1}^{N} d_i^p \qquad (2.1.2.3)$$

The average throughput per user, is then given by the ratio (2.1.2.2) over (2.1.2.3):

$$t^p = \frac{b^p}{D^p} \qquad (2.1.2.4)$$

- **Integrity monitoring in RAN** [18]: introduces the metrics used to assess integrity of services belonging to a certain class now in a more advanced network, the 3G.
    - **BLER monitoring in the downlink**:

$$BLER^m = \frac{\sum_{i=1}^{N} \bar{B}_i^{\,m}}{\sum_{i=1}^{N} (B_i^m + \bar{B}_i^m)} \qquad (2.1.2.5)$$

Based on the formula definition of Block Error Rate, we can deduce that $\bar{B}_i^m$ is related to the block of data unsuccessfully delivered over the transmission line and B_i^m is related to the block of data successfully delivered to the destination over the same transmission line.

- **Throughput**: the throughput in the RAN (Radio Access Network) takes into consideration the correctly transmitted bits over a certain period. This is an important (the throughput) QoS metric to assess the quality of experience (QoE) of customers [19]. The throughput t^m is given by, as unit of bits over time:

$$t^m = \frac{\sum_{i=1}^{N} \sum_{k=1}^{C} r_k B_i^{m,k}}{\sum_{i=1}^{N} d_i^m} \qquad (2.1.2.6)$$

- **Integrity monitoring in Core**: the study done by Soldani et al. also introduce the metrics used for integrity in the Packet Core domain, which will be the main area of interest of our research study herein.
    - **Throughput:** The throughput is given on the downlink and uplink as follow:

$$AveUL.APN = \frac{GTPBytesSent.APN}{K.PDPcontextActive.APN} \quad (2.1.2.7)$$

$$AveDL.APN = \frac{GTPBytesReceived.APN}{K.PDPcontextActive.APN} \quad (2.1.2.8)$$

Where $GTPbytesSent\ and\ GTPbytesReceived$ represent the number of User plane Bytes sent and received respectively at the APN level. K is a constant and the $PDPcontextActive.APN$ represents the number of active PDP(Packet Data Protocol) contexts at APN (Access Point Name) level.

- The study of Soldani et al. [18] also provides different examples of service assurance solution for Network Management Systems such as the:
    - Centralized Performance Management: in this system the data is collected from the Network element layer, pre-processed in the databases and sent for visualization. It mentions also that the storage capacity and retention period are CSP dependent.
    - Active, Service Management tools: this includes probing systems. Easily manage alarming systems.
    - Service Quality Manager: consisting of KPIs (metrics), Real time data processing, Alarming systems, and more functionalities.

2.1.3. Scope of the above QoS Review

The work done presented by D. Soldani et al. as summarized above is one of the most detailed and inspirational research studies in Telecommunications and Quality of Service (QoS). The work presents the conceptual architectures of QoS and QoE taking into consideration integrity and accessibility. The topic on Retainability has not been covered in the study. And, their

approach is based on a layered architecture focusing on Network Element Management layer (Performance Management and Fault Management) and Network Management layer which is the baseline of SQM and CEM. QoS metrics computation are presented in a clear way. It also presents different SQM systems and tools for QoS improvement.

From the previous review, our research follows some of the methodologies described in the work of the researchers such as following a layered architecture for our Network Data analysis. However, the work presented by D. Soldani et al. does not tackle the methodology or technology used for data pro-processing phase and does not detail much on the SQM use cases that could be useful for different departments although the use case layer was added on the conceptual architecture. The administration database system presented is not detailed as for how to proceed with the management of the database.

In our research we include the steps used for data processing and pre-processing, including SQL and Big Data techniques to manipulate network elements data. And also, deduce some basic SQM use cases from which CSPs can build upon.

2.2. Data Mining and Predictive Analytics Review

Finding information pattern in a bunker of data and predicting the trend of performance have become the new song of today Data Science techniques and it is being applied to almost every domain in life (Medicine, Telecommunications, Finances, …). Data mining is simply related to applying a thorough explanatory Data Analysis on the datasets. The review in this section is mostly based on the work by Daniel T. Larose and Chantal D. Larose in Data Mining and Predictive Analytics [20]. Their study describes six tasks that can be accomplished through data mining techniques as summarized in the below table:

Table 2 Illustration of Data Mining Tasks

Tasks	Summary.
Description	• Description of pattern information inside the data. • Use Exploratory Data Analysis and any graphical model to uncover data pattern.
Prediction	• Relying on future trend, or future results. • Predicting a certain service adoption 2 to 3 months from now (Marketing campaign related strategy). • Can use Regression methods
Classification	• Focus on categorical predictors. • Classifying the customers likely to "*churn*" from a Communication Service Provider.
Estimation	• Approximation model of numerical variables based on categorical predictors. • Estimating the number of customers that could be affected by poor YouTube streaming quality based on other predictors. • Can also use Regression models and other statistical methods.
Clustering	• Categorizing attributes of the same nature together. • No target predictor required for clustering. • Groups of customer contracts (Prepaid, Post-paid, …)
Association	• Prediction of Degradation of services in a Communication Service Provider. • Creating relationship between attributes or predictors. • Applying the rule of "If … then …" with support and confidence rules.

The methodology used by Daniel T. Larose and Chantal D. Larose is based on the Cross-Industry Standard process for Data Mining (CRISP-DM) developed by analysts and researchers [21]. This methodology will be adopted by our research, but with a bit of customization on the process.

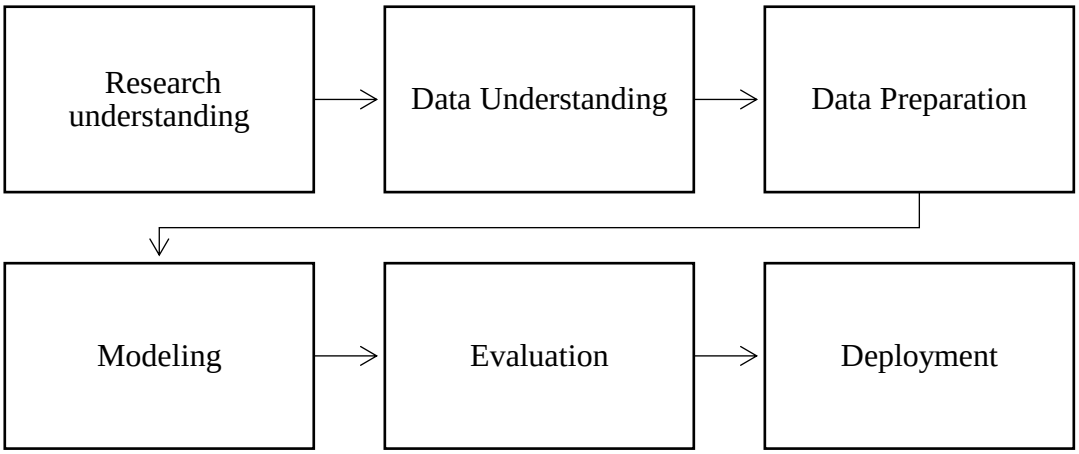


Figure 6 The Cross-Industry Standard Process for Data mining: CRISP-DM (**Chapman et al., 2000**)

The CRISP-DM methodology defines a process model that provides a framework for carrying data mining projects which is independent of the industry sector and the technology used. This gives the model a certain flexibility in terms of area of application. Adapting the methodology to fit the Mobile Cellular Networks is the main objective of the research.

- **Research Understanding Phase**: Understanding what needs to be done and requirements from a research perspective. The knowledge is then converted in to Data Mining problem definition as defined in the problem statement. Finally prepare a preliminary strategy to meet the requirements.
- **Data understanding Phase**: Initial data collection, get familiar with the data, identify quality problem in the data using exploratory data analysis, discover first Insight in to the data. Detect interesting subsets to form a hypothesis for hidden information (subsets that may contain actionable patterns).
- **Data preparation phase**: Construct the final data set. Data feeding to the model. The tasks in this process includes record, tables, attribute selection, data cleaning, construction of new attributes, transformation of data for the modelling tool.
- **Modeling Phase**: Apply and calibrate the model(s) to optimal values. In this case focus will be on the Regression trees.

- **Evaluation Phase**: Evaluate and review the steps that used to build the model to ensure that the model has met the objectives set. A key point in the phase is to evaluate if there is any research area that has not been considered sufficiently. At the end, the model should be ready to be deployed.
- **Deployment Phase**: Be able to reuse the research output. The output of the research model will be a reporting graphical representation of Modelled data.

2.2.1. Big Data Review

The amount of Data that are generated in a Mobile Cellular Network by subscribers, grow in an exponential way with ease adoption of Smartphones, affordable subscriber plans, and wide span of attractive applications. Even in the Telecommunications, data generated is of higher magnitude comparing to the size of Telecommunications data of years ago. Because of the number of Open source Big Data projects being developed, several start-up companies move towards the big data direction. Since Big Data approach is one of the methods that we are using in this research study, we browse through some of the big data technologies available and at the end, we use one technology to approach our study.

2.2.1.1. Hadoop MapReduce Overview

Hadoop is an open-source Big Data technology, which has been used for years now for processing large scale data. And it run on a cluster of servers. Using simple programming model, Hadoop is based on distributing processing computing where tasks are distributed among many computers. Google back in 2014 analysed the fastest methodology to process Terabytes of information by developing a distributed computing mechanism, the MapReduce [22]. MapReduce can be integrated with many big data platforms, as an example here, the open source Hadoop.

Advantages of the Hadoop Big Data:

- Open source framework and can run on a cluster of commodity servers. Cost reduction. No need to purchase extravagant Hardware.
- Fault tolerance and high availability are provided with the Hadoop big data platform.
- Hadoop is very good for Batch processing of large amount of data.

- Distributed computing facilitating the movement of codes around the cluster than moving around a large dataset. The framework facilitates the writing of codes for distributed applications.

Components of the Hadoop Big Data:

- **HDFS**: the Hadoop Distributed File System is the file system used to store data and it is used to have faster access to big amount of data. Data is partitioned into blocks and stored in the cluster's individual machines. The HDFS consists of two types of nodes: Data Node and Name Node.
    - **Name Node:** Manages the systems files, also stores the Metadata.
    - **Data Node:** Stores the blocks of Data files (content).

The below figure shows a typical Architecture of an HDFS with four Data Nodes and one Name Node.

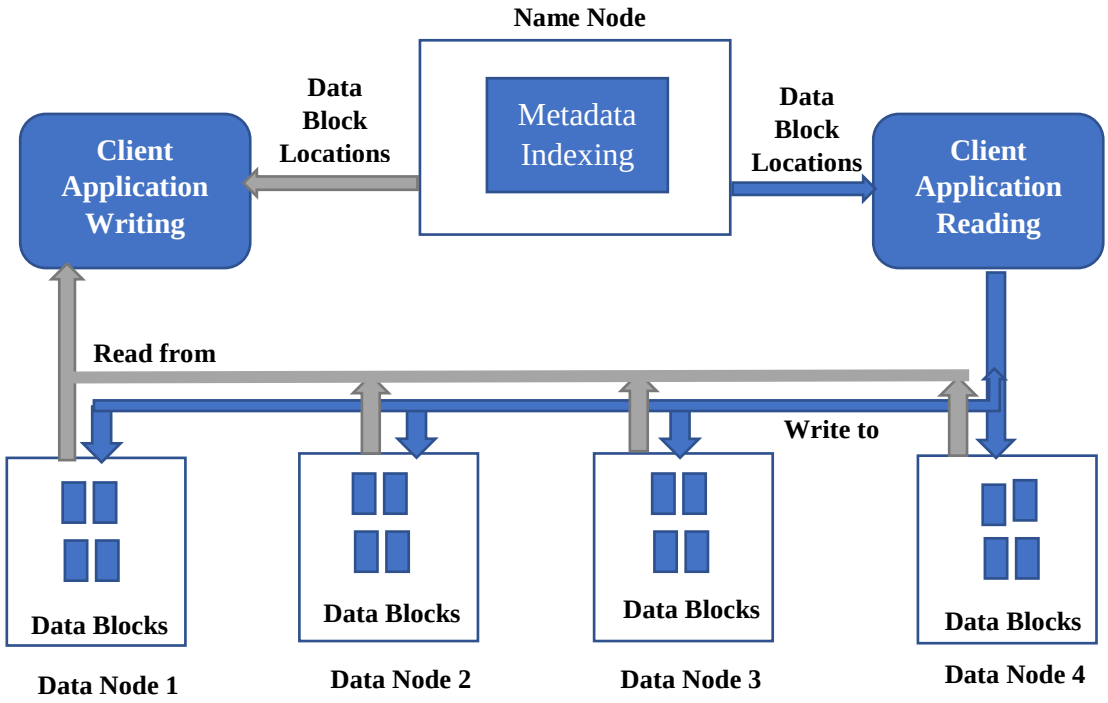


Figure 7 Hadoop Physical Architecture

- To ensure that the Data Nodes are all functional, they send "heartbeats" messages to the Name node.
- To provide the list and status of data blocks, the Data Nodes send "Blockreport" messages to the Name node.

- Every process of Data Read or Write goes through the Name Node. A client application sends the request to “read” or “write” data from or to respectively to the Name Node. The Name Node responds with the locations of the specific data blocks that constitute the file.
- For the write process, the Name Node creates new entries in the HDFS namespace every time a write request is initiated by the client application.
- However, the Name Node does not take part into the actual retrieval of data from the cluster to the client.

- **MapReduce**: The processing engine for large datasets. It uses parallel processing capability and provides a higher-level environment for drafting distributed application codes that can run on clusters of machines. The MapReduce has two main functions:
  - **Map:** The **Map** splits and distributes the data partitions across the data nodes in the cluster of machines. It uses key-value pairs as inputs and output intermediates keys. It arranges the output and put together the values which belong to the same intermediate keys. Then this is passed to the **Reduce** function.
  - **Reduce:** The Reduce function of the MapReduce aggregates the intermediate value keys received from the Map function, and outputs one aggregated value. Presented by Khadija et al., the tasks of the MapReduce are shown in the figure below [23]:

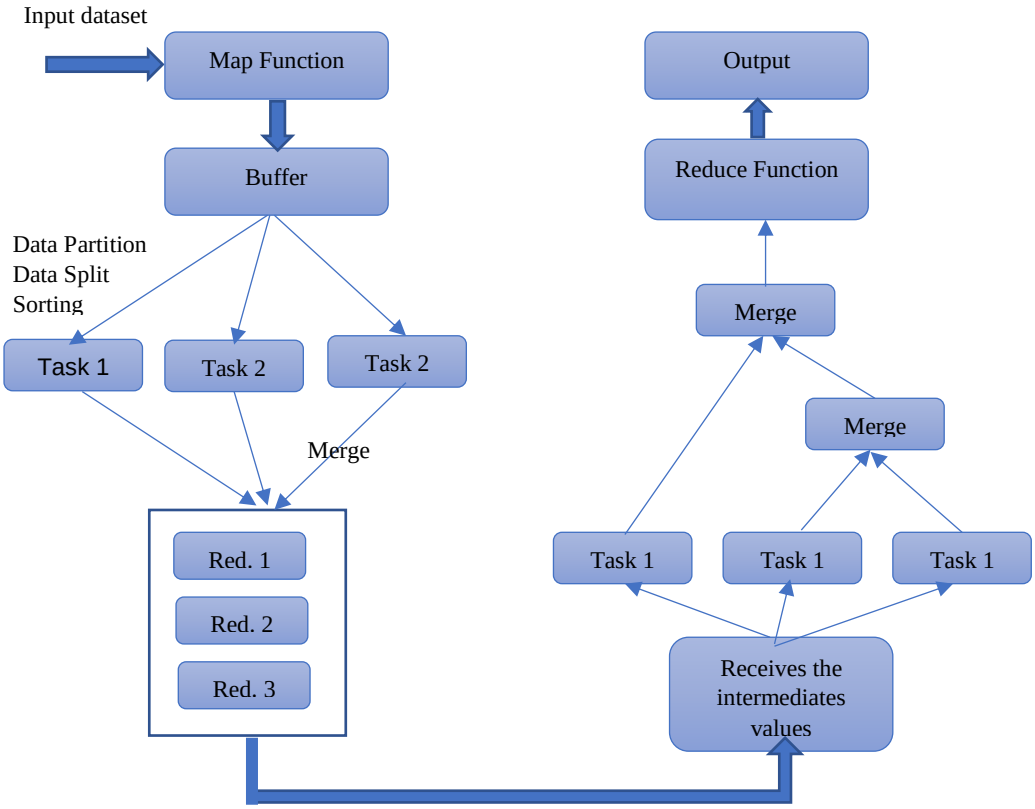


Figure 8 MapReduce Task Overview

- **Limitations of Hadoop Technology**:

  Although considered as one of the biggest big data technologies by many researchers and scientists, the Hadoop technology has shown some limitations which are listed below:

  - **Small Files problems**: The design capacity of Hadoop is very high. Therefore, it doesn't process efficiently small files, with small size defined to be less than the HDFS size of a block, which by default is 128 Mb.

  - **Slow processing speed**: in processing large datasets, the two functions of Hadoop, described above (Map and Reduce) need to be performed in a parallel way. This increases the latency since the data is also stored in the disk.

- **Good for Batch Processing Only**: Hadoop does not work well with streamed data, which also affects the performance speed. Datasets for Hadoop are processed in Batches.

- **Real time capability not supported**: For huge amount of data processing, batch processing could be the ideal solution, but also requires high computational power. Hadoop takes high volume data and produces output. However, the solution is not suitable for real time processing.

- **Not user friendly**: For every operation that needs to be done on the cluster, codes need to be handed over, this makes the system complex to use. MapReduce on its own downs provide interactivity.

- **No Caching capability**: MapReduce doesn't provide the capability to cache data in memory. The intermediate data cannot be cached for further processing, which slows down Hadoop performance.

2.2.1.2. In-Memory Processing with Apache Spark

As mentioned in section 2.2.1, Hadoop technology has shown limitations in terms of data caching, in-memory processing. Not suitable for data streaming and real time processing. In the same line, many data processing technologies have been relying on data manipulation in disk, with a good illustration being the popular Structure Query Language (SQL). Many services provided today such as video streaming, social media streaming (twitter stream) and audio-streaming applications require low latency, fast data processing and quicker data retrieval speed. Data stored in the memory is accessed faster than data stored in disk; for that reason, for real time applications and streaming, in-memory data processing happens to be the best methodology to process data [24]. In contrary to the Big Data Hadoop system, in-memory data processing provides a lower hardware footprint relying on the single computer processing power. To meet the business needs, computational elements such as CPU, Memory size, disk storage are all extremely important for fast data processing. The objective is to process data at rest and data on the move to answer different business needs.

Apache Spark, built on Java, is a big data framework, based on distributed processing; however, Apache Spark is faster than many existing big data processing frameworks. The speed

of Apache Spark is deducted and demonstrated with the fact that data processing is completely done in memory of the computer(s) and remove the needs for using I/O processes on the storage disks.

- **Some Advantages of Apache Spark**:
  - Powerful processing framework with a uniform programming model
  - Support multiple programming languages including Python, Java, Scala, R.
  - Faster data processing than other big data processing because of the in-memory capability.
  - Supports both stand-alone (single machine) and cluster configuration.
  - Comes with a diverse stack of libraries and functionalities, used for different data processing needs.

- **The Spark Libraries**:

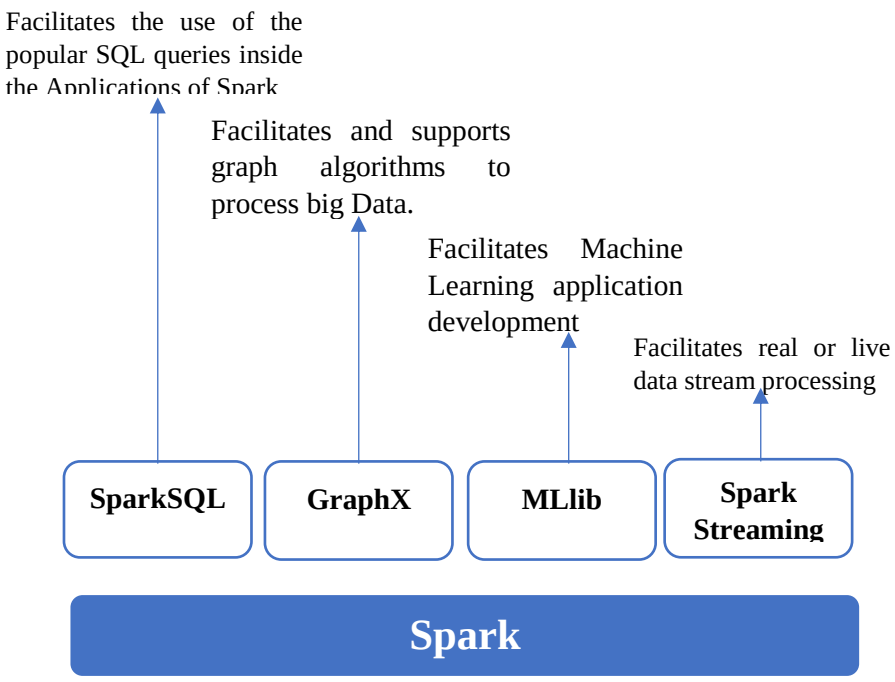


Figure 9 Apache Spark Supported Libraries

- **SparkSQL**:

  The technological enhancement introduces many challenges including but not limited to development of skills. Organizations for very long time has been using Relational Database Management Systems (RDBMS) and have invested a lot in such technologies.

Re-investing totally in a new technology could be a considerable obstacle towards adopting new technologies. Hence, the need for many platforms to enable interoperability between existing systems, technologies and the latest ones. SparkSQL is an Apache Spark module and library, developed to process Relational Data Structure with Spark [25]. It enables the application of SQL commands to retrieve information. Hence, SQL Users can take advantage of the Big Data processing. SparkSQL presents three main capabilities as presented by [26]:

- Data frame abstraction in different programming languages including Java, Python and Scala to efficiently work with structured data sets.
- Read and write many popular structured data formats such as JASON, Parquet, and Tables).
- Using SQL, query data from inside Spark program and from external tools that use database connectors to communicate with Spark (JDBC/ODBC). This scenario is used in this paper since the connection to the dataset is via JDBC connector.

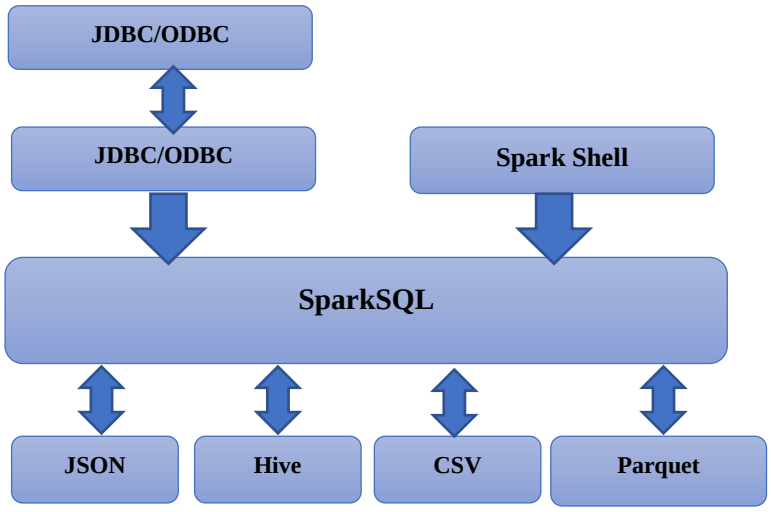


Figure 10 Scope of Application of SparkSQL Library

- **In-memory Caching and Data Access for SparkSQL**:

  Spark SQL takes the advantage of the in-memory capability of Apache Spark to provide to cache data in the memory in a structured way. It reduces the memory footprint by applying columnar compression techniques on the structured data dictionary encoding

scheme. Data cached in memory can be retrieved for different applications including visualization and Machine Learning at a later stage.

2.2.1.3. GraphX

GraphX is a Spark Library, as shown in Figure 9, which provides a compelling way of processing Big Data, by representing and exploring the connections between data points in a large dataset. Data points are represented as vertices and all the connections are shown as edges, in between vertices pairs. With the accent put on finding meaning from relationships, Big Data Graph processing can be applicable to domains such Mobile Cellular systems, Social Networks, Web Content Analysis and any intense data manipulation use cases. Graph processing has become a strong pillar of online advertisement and product recommendation methodology.

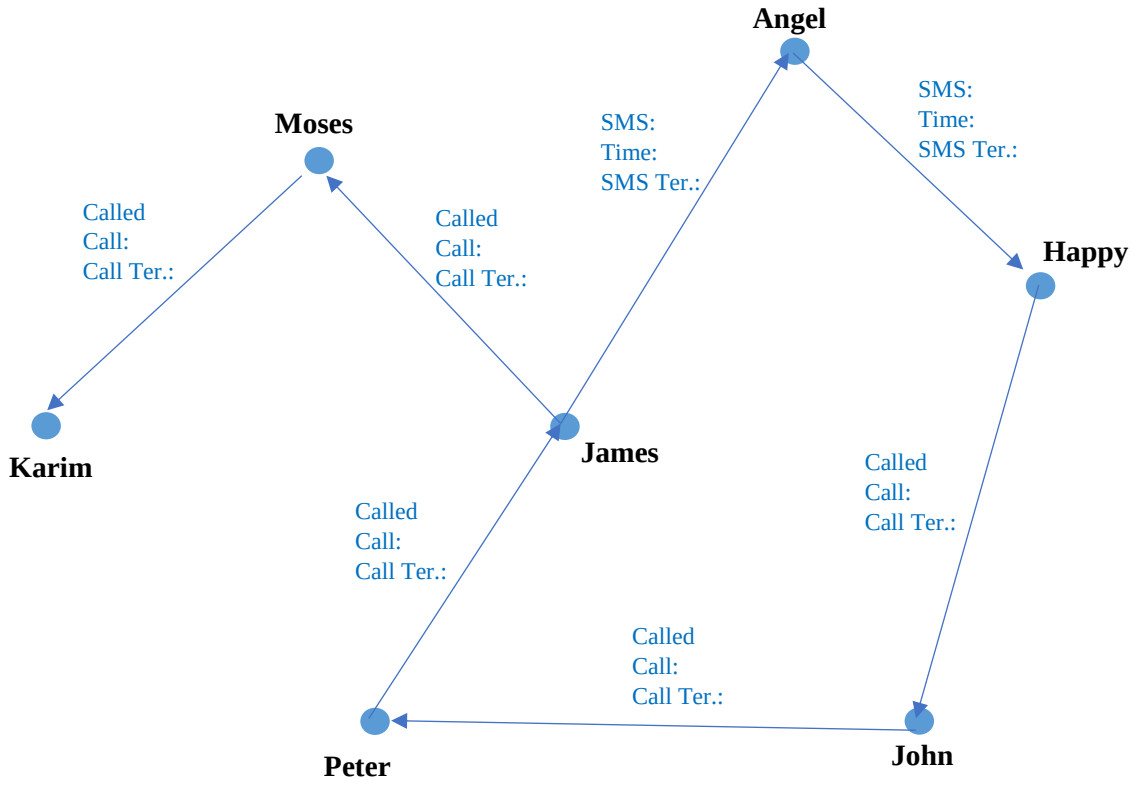


Figure 11 Customer calls Graph processing Illustration

Figure 11 illustrates a simple graph processing of customers calls, which could be very complex to represent using a tabular format or standard graph. The figure shows the activities done by users, referencing customer James. James, Karim, Moses and all the name of the customers, are referred to as "**vertices**" and activities such as calls, SMSs represent the "**Edges**". The vertices are also referred to as Nodes and are used as entities; on the other hand, the Edges also referred to as labels illustrates the role of vertices in the selected, specified domain. Several Edges can be assigned to vertices, adding a certain level of constraints to the vertices.

On the research study, Jakob Smedegaard and Olaf Zunkunft evaluate the scaling of Graph for big data GraphX [27]. The paper runs experiments to evaluate the scalability of different algorithms and GraphX framework. They emphasize on the graph and its properties as follow; the study introduces the preliminary of graph processing models, with the property-graph approach based on the study by R. Diestel [28] and the Pregel methodology proposed by G. Malewics et al. [29] which provides an iterative programming model which is mostly vertex-oriented. A super-step is used to name an iteration, and three phases are applied as below:

$G = (V, E)$ where V is the set of Vertices and E is the set of Edges. Element $e \in E$ represents the link between vertices.

- **Gather Phase**: using a combiner to aggregate all messages destined to v, in which the operation $\otimes$ has some associative and commutative characteristics.

$$M_{\sum} \leftarrow \otimes\, m\left(M_{u,v}\right) \quad \text{With } u \in N_{in}(v) \qquad (2.2.1.3.1)$$

$M_{u,v}$ is the message being directed to receiving node v from the sending node u.

- **Apply phase**: from the gather phase, current properties of v and message $M_{\sum}$ are accessed to apply a transformation to compute new properties of v. The new properties are given by:

$$P_V^{new}(v) \leftarrow a(P_V(v), M_{\sum}) \qquad (2.2.1.3.2)$$

- **Scatter Phase**: using new properties of the vertex, the edges and the end node are all used to send messages to the surrounding nodes of v.

(2.2.1.3.3)

$$\forall w \in N_{out}(v):$$

$$M_{v,w} \leftarrow s(P_V^{new}(v), P_E((v,w)), P_V(w))$$

The methodology on what the Pregel approach is based on, follows the Bulk Synchronous Parallel processing, introduced by L.G Valiant [30]. The study goes further, using GraphX to analyze social graph using semi-clustering, still based on the Pregel Model. The analysis is based on SCALA programing language. Using the same approach, we can use the same concept to analyze customers behaviors in a Network and cluster them according to their services or any other criteria.

Jakob Smedegaard and Olaf Zunkunft in their experiment, analyze Social Networks using Graph based algorithms, applying several semi-clustering methods.

Using Graph Model, the study uses the semi-clustering method introduced by [29], to determine the number of clusters, bundles of individuals who have strong connections with the individuals in the same cluster and have weaker connections with people outside the cluster. Vertices represent people and edges represent relationships between individuals. The same approach has been researched thoroughly also by J. Scott [31]. The result was the scaling of the number of people, studying the behavior of the graph structure. The below figure shows the scaling number of users using semi-clustering graph algorithm, as run by Jakob Smedegaard and Olaf Zunkunft [27]. Other method such as collaborative filtering, is also used to predict sales performance based on other purchased items.

The same Semi-clustering graph algorithm method can be used in Telecommunications to tailor marketing campaigns based on subscribers belonging to a certain cluster group with some specific characteristics and service categorization to predict the group of subscribers who could easily adopt a new service.

Number of Workers	Observed runtime in s	Expected runtime with linear scaling	Observed scaling factor
1	215,84	215,84	1
2	132,41	107,92	1,63
3	91.65	71,95	2,355
4	79,98	53,96	2,699

Figure 12 Scale of the Number of users using Semi Clustering *[27]*

2.2.2. Prediction Algorithms Review

Machine Learning, Predictive Analytics and Artificial Intelligence have become a popular song for every business and sectors in today's era. Prediction algorithms are being built constantly to improve different aspects of the industry. Figure 13 summarizes some of the most popular Machine Learning algorithms in use currently. However, we are not going to detail all of them rather review a couple of them and researches that have taken the advantages of such algorithms.

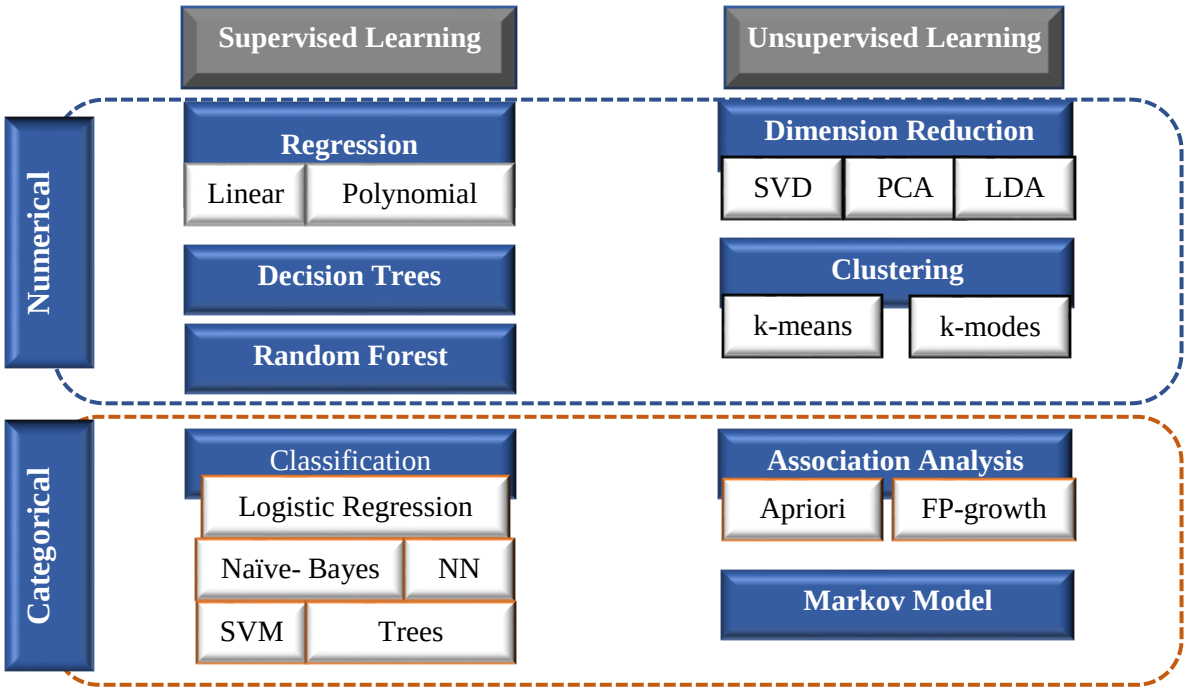


Figure 13 Popular Predictive Algorithms' Summary

SVM: Support Vector Machine; NN: Neural Network; SVD: Singular Value Decomposition; PCA: Principal Component Analysis; LDA: Latent Dirichlet Analysis.

It has been highlighted that some of the algorithms such as Random Forest, Neural Networks, Decision Tree and Gradient boosting tree (not listed above) can be used for both categorical and numerical predictions.

2.2.2.1. Regression Models

Regression Models have been for a long time, the work horse of data science and predictive analytics as they have produced many individual models. Because of their simplicity, ease of parsing and ability to interpret the results, regression models act as the gateways (points of entry) in Data Analytics and as powerful tools to solve practical statistic problems. In this section we have a look at the low-cost introduction to regression models as studied by Brian Caffo [32], using the Galton Francis regression model to predict children heights from parents' height. To evaluate the relationship between two variables, the initial assumption would be to suggest a simple linear regression line of the below caliber:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (2.2.2.1.1)$$

Where Y_i represents the Height of Child i, in other way, a numeric variable of the dataset, X_i is the height of the parent i, β is the regression parameter and ϵ_i is the error. The equation is then generalized as follow, considering ϵ_i as independent and identically distributed:

$$y = f(x, \beta) + \epsilon \qquad (2.2.2.1.2)$$

$$E[Y_i | X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i \qquad (2.2.2.1.3)$$

and

$$Var(Y_i | X_i = x_i) = \sigma^2 \qquad (2.2.2.1.4)$$

The least squares estimate β_0 and β_1 are given by:

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)} \text{ and } \hat{\beta}_0 = \hat{Y} - \hat{\beta}_1 \hat{X} \qquad (2.2.2.1.5)$$

The study also highlight regression as a very powerful tool for prediction; to guess the outcome or result at a specific value X of the predictor, the Regression model estimates:

$$\hat{\beta}_0 + \hat{\beta}_1 X \qquad (2.2.2.1.6)$$

In an easy way, predicting with regression consists of finding the value of Y on the line with the corresponding X value. In terms of predictions, regression, especially linear regression does not have a high accuracy; however, it provides parsable and interpretable results. The result of the research from Francis Galton is still as relevant today as it was.

- Some results of the study: Using R-Package, Caffo shows the inside of the dataset used by Francis Galton. The data is also available in R as part of the predictive analytics packages.

Loading and plotting Galton's data.

```
library(UsingR); data(galton); library(reshape); long <- melt(galton)
g <- ggplot(long, aes(x = value, fill = variable))
g <- g + geom_histogram(colour = "black", binwidth=1)
g <- g + facet_grid(. ~ variable)
g
```

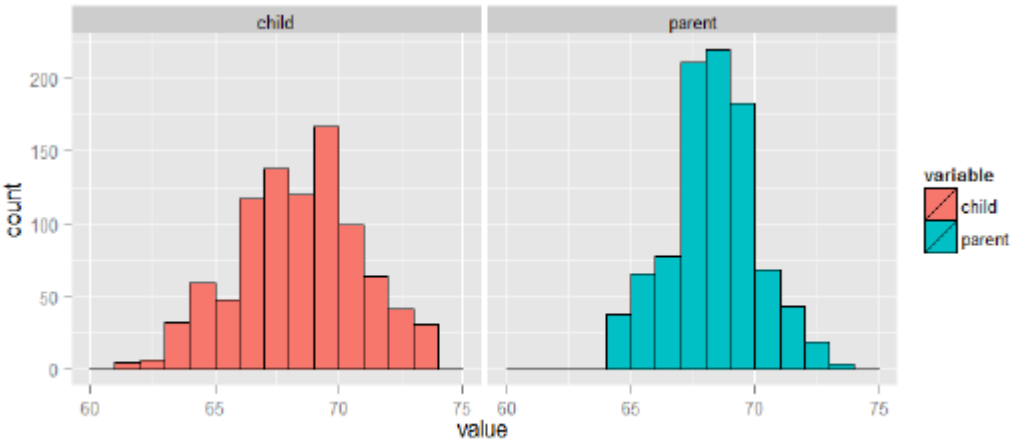


Figure 15 Plot of the Galton dataset as by Caffo *[32]*

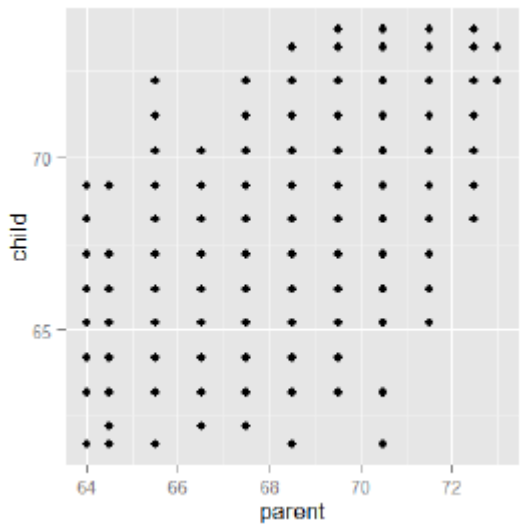


Figure 16 Plot of Children vs. Parents Heights as illustrated by Caffo *[32]*

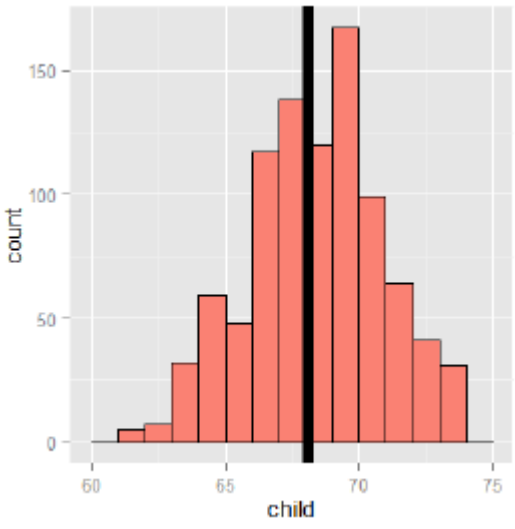


Figure 14 The plotted best mean of the Galton dataset *[32]*

The research of Galton, as studied by Caffo stipulates the interest of Galton to the fact that tall parents have children who tend to be tall, but a little bit shorter than their parents. And short parents tend to have short children but who are not necessarily as short as their parents. Thus, the concept of "Regression to mediocrity". The below table illustrates the original dataset table used by Galton for his regression problem [33]:

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Figure 17 Francis Galton Genetic Dataset for Regression *[33]*

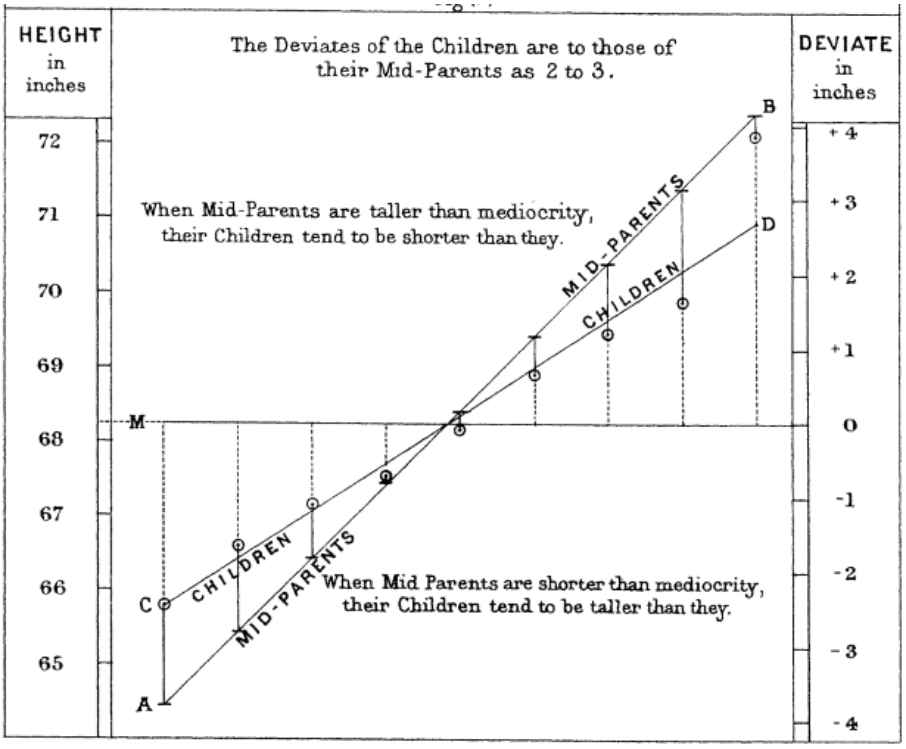


Figure 18 Deviation, Coherence & Precision of Galton's Regression Experiment *[33]*

The Analysis provides a numerical value of Regression for the experiment in the case of human as from 1 to 2/3 with some unexpected coherence and precision as illustrated in his graph, represented in Figure 18.

The application of Regression models since then has been broaden in many analysis and research paper in the objective to predict future events and outcome. Xuan Feng et al. used Multiple Linear Regression model to predict contact temperature of High Voltage Switchgear [34]. The study used a Map Reduce Model and illustrated how the longitudinal regression in multivariate linear regression could fit to predict the temperature of a High Voltage Switch gear. The study also compared the results of the two regression models used (longitudinal and Lateral regressions) with the results shown in the below figure table:

Table 3 Comparison between Longitudinal & Lateral Regression models *[34]*

Performance Index	Transient Lateral Response	Transient Longitudinal Response	Long term Lateral Response	Long term Longitudinal Response
MSE	0.394	0.409	0.144	0.102
MAPE %	2.91	3.20	1.34	0.94

During our research, regression models will be used to train Network Data in the objectives to predict outcomes and improve network and user experiences. Other regression models such as regression trees, logistic regression will be exploited practically as the experiment goes through.

2.2.2.2. Neural Networks and Deep Learning

Inspired by determinants such as neuroscience, mimicking the fundamental architecture of human nervous system, Neural Networks and Deep Learning are transforming the Technology industry. Ranging from Computer vision, Search Engines, Recommender systems, Advanced robotics, the field, Neural Networks and Deep Learning are becoming the major drivers of AI industry. Deep learning indicates the training of large Neural Networks or multi-layers Neural Networks. Figure 19 shows a simple, 1-layer Neural Network where $x = x_1, x_2, x_3, \dots x_n$ is the input vector, W, b are the parameters of the Neural Network, also called Weight and bias vectors respectively and y the output of the Neural Network. The complexity of a Neural Network can grow with the increase in terms of number of layers.

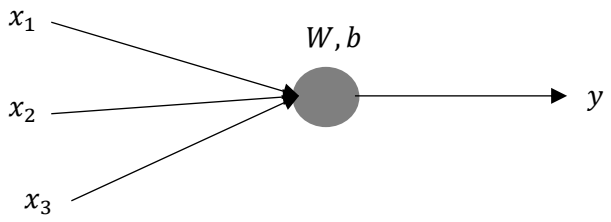


Figure 19 Illustration of a simple, 1-layer Neural Network

$$y = \sigma\,(Wx + b) \qquad (2.2.2.1)$$

which defines the sigmoid function of $Wx + b$. Let $Wx + b = z$, the output $y = \sigma(z)$. The sigmoid function of z is given by:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad (2.2.2.2)$$

Given many inputs parameters (training samples) x, a prediction algorithm can be used to predict a binary value 1 or 0. The objective of programming such a Network is to learn parameters W and b such that the output y is an understanding estimate of what needs to be predicted [35]. A more complex Neural Network, or Deep Learning Network, is illustrated in the below figure.

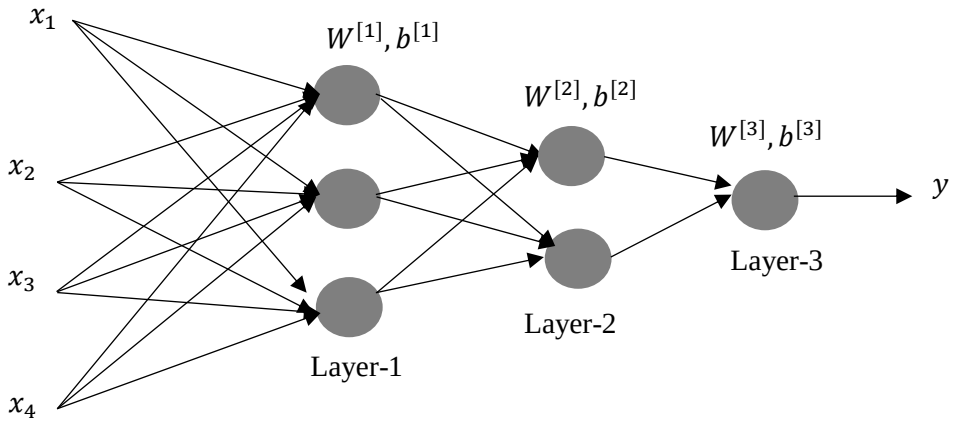


Figure 20 Illustration of a 3-Layer Neural Networks

Yuming Hua et al. in their research paper, Deep Belief Networks and Deep Learning [35], introduced how to process a Deep Belief Network, using Restricted Boltzmann Machines. In their paper, they also detailed Neural Networks and Deep Learning concepts. Their study also uses the SoftMax classifier to recognize handwritten numbers. The paper uses the below steps to tackle the problem:

- Finding the Energy function of the Restricted Boltzmann Machine:

$$E(v,h) = -\sum_{i\in v} a_i v_i - \sum_{j\in h} b_j h_j - \sum_{i,j} v_i h_j W_{ij} \qquad (2.2.2.3)$$

  Where h is the hidden layer, v is the visible layer, W is the Weight vector (Matrix), b the bias vector of the Restricted Boltzmann Machine (RBM).

- Finding the Probability distribution and the marginal distribution:

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)}, \quad P(v) = \frac{1}{Z} \sum_h e^{-E(v,h)} \qquad (2.2.2.4)$$

- Creating a Deep Network by using the output of the Restricted Boltzmann Machines:

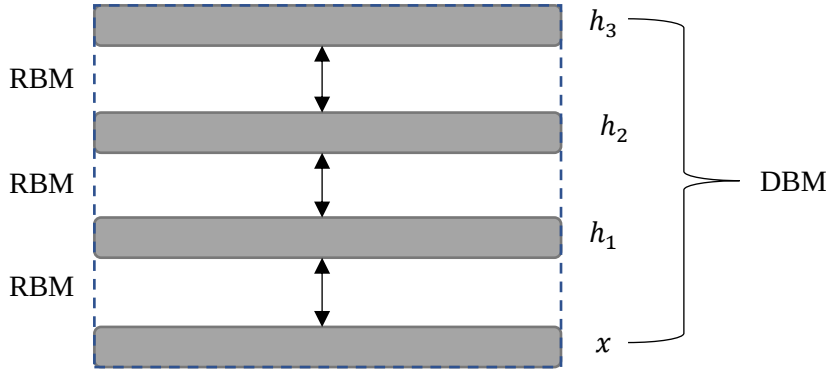


Figure 21 Creating Deep Belief Network by Using 3 RBMs *[35]*

- Training the Deep Network: Uses a supervised learning with cost function equals to,

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\}\log\frac{e^{\theta_j^T x(i)}}{\sum_{l=1}^{k} e^{\theta_l^T x(i)}}\right] \qquad (2.2.2.5)$$

- Attaching a Softmax classifier and fine-tuning the whole Deep Network.

The study achieved a 93.12 % accuracy on a simulation over 70000 pieces of input pictures. The sample of the simulation of the study is shown below:

Figure 22 Illustration of Simulation for Handwriting Recognition Study using Deep Network *[35]*

Pink Kuang, Wei-Na Cao and Qiao Wu illustrates the structures and benefits to shallow learning of deep learning and did some analysis of the currently most used algorithm in detail [36]. The study describes the Neural Networks and Deep Learning methods as follow:

- Convolutional Neural Network (CNN): used in many visual recognitions, it is a multilayer sensor Neural Network, in which each layer consists several 2-dimmensional plane. Planes have more neurons. More literature on the CNN can be found in the work studied by Zhijun Sun [37] and Jianwei Liu [38].
- Deep Belief Network (DBN): which can also connect several Restricted Boltzmann Machines to build a DBN. Hidden and Visible layers are interconnected.
- Deep Auto Encoder: Deep Learning structure and multilayer network of high dimensional data through unsupervised layer, introducing pre-parameter optimization training and system.

In our study, in this research-based dissertation, Neural Network is one of the prediction algorithms used to predict potential network churners, providing input parameters from Network transactions. Alongside other algorithms, the accuracy of prediction is tested towards the best algorithm for that use case.

2.2.3. Business Intelligence

Data representation is a very important factor when it comes to Data Analysis as it is the fundamentals indicator for decision making processes. In this section we look at some works or researches conducted in the areas of BI (Business Intelligence).

Yan Shi and Xiangjun Lu analyse processes, methodologies and technologies which underline Business Performance Management (BPM) and establish relationship between BPM and BI [39]. Yan Shi and X. Lu proposed in their research a framework to integrate Business Intelligence and Performance Management in a comprehensive method of controlling business performance. The study ties performance indicators to business strategy. The BPM as introduced by Yan Shi, provides three important outputs:

- Information in an understandable way, especially for Managers and executives.
- Performance Insight to facilitate business management.
- Performance effectiveness, which gives business improvement points.

The BPM framework presented by Yan Shi and X. Lu is shown in the figure below. The framework integrates Business strategies, IT processes and BI.

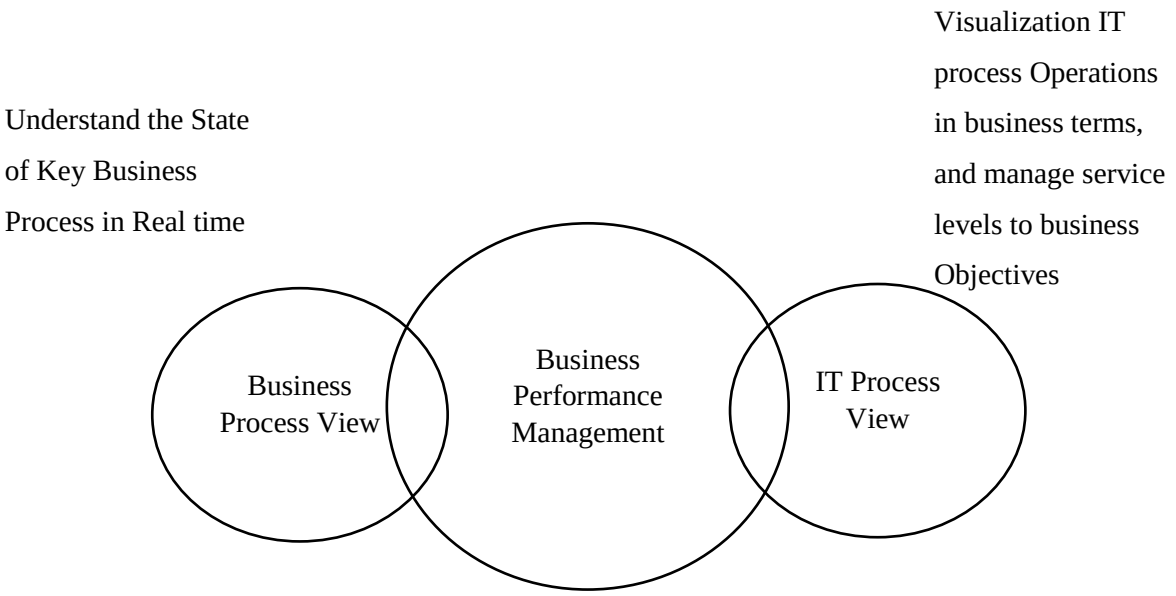


Figure 23 Business Performance Management Framework as presented by Yu Shin and X. Lu *[39]*

Yah Shin and X. Lu list new trends of BI systems, including:

- Ensuring a link between business process and operations of activities to enable a complete view of the Network.
- Defining business rules and KPIs which are consistent with the management process.
- Implementing an alerting mechanism for problem avoidance. Adopting a proactive rather than reactive problem handling.
- Define a flow of data to allow efficient management of business data or process.

According to Yah Sin and X. Lu, a BI, to enable proactive management of an Enterprise, should include the above trends.

Tong Gang, Cui Kai and Song Bei set out an overview of BI, defining the key technology of BI, the establishment and the application of the technology [40]. The study presents the characteristics of BI and the key technology of BI as shown in the below figures.

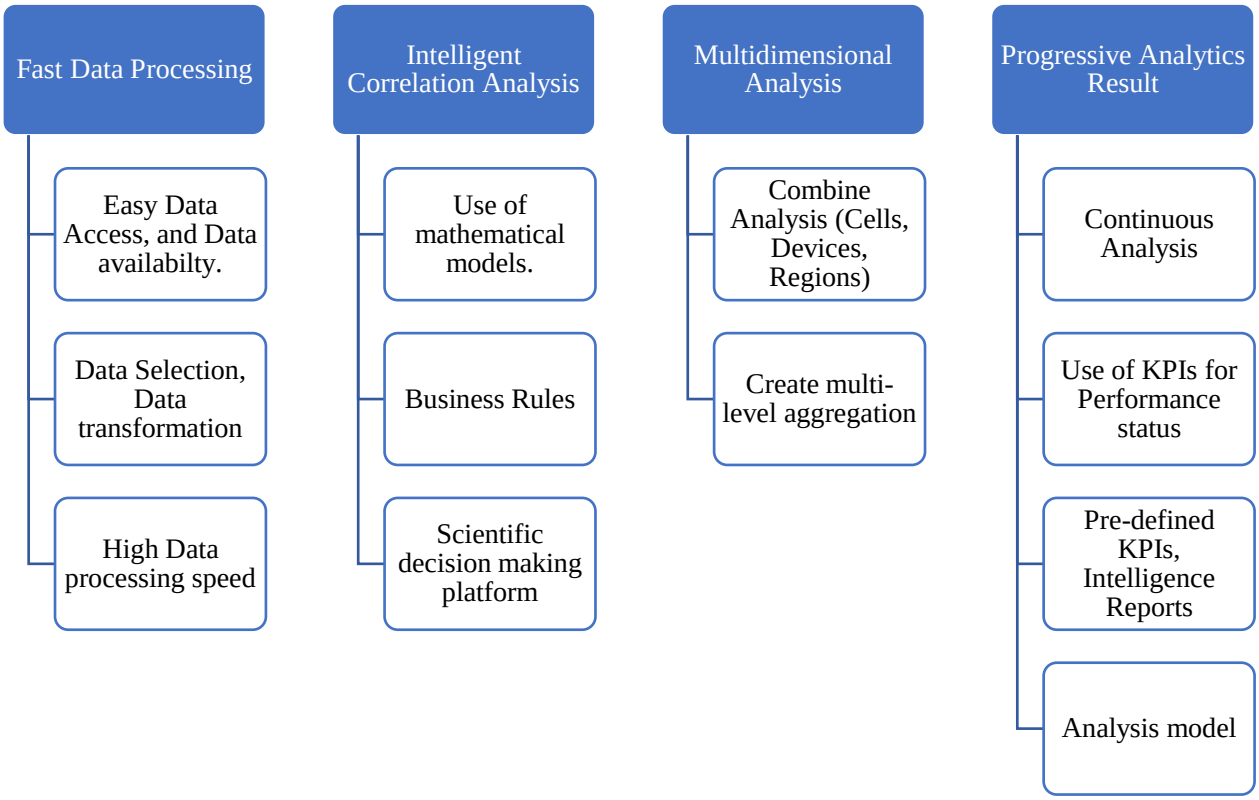


Figure 24 Characteristics of Business Intelligence as studied by Tong Gang et al. *[40]*.

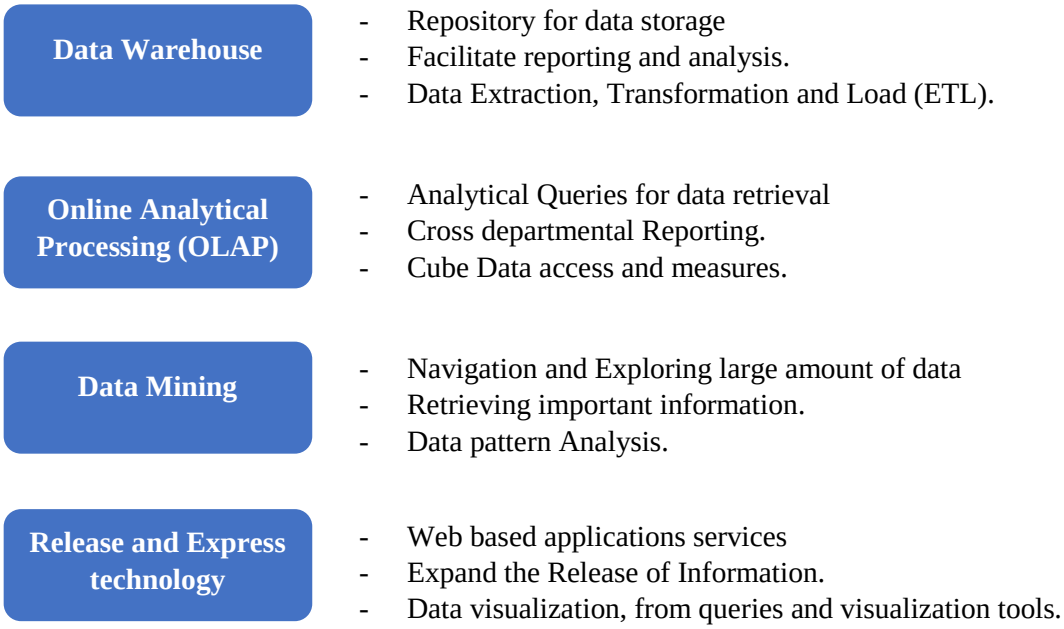


Figure 25 Key technology of Business Intelligence as shown by Tong Gang et al. *[40]*

The concept developed by Tong Gang et al. will be very helpful in this dissertation to facilitate the visualization and data manipulation.

2.3. Related Works

M. Torres Vega et al. proposed a machine learning algorithm to address the QoE issue for Video streaming services in the objectives for Communication providers to initiate proactive control on customer experience [41]. Their study compares various prediction algorithms, detailing supervised and unsupervised learning predictive analysis, based on their performance and QoE elements that they need to address. The results of the study are detailed and summarized in the below table:

Table 4 Machine Learning Research Study for video QoE *[41]*

Predictive Approach	**Explanation**
Client-based video quality modeling	Classification method based on metrics. Not very suitable for real time applications. Benefits from training and benchmarking.
Control loop	Suitable for real time applications.
Predictive monitoring	The solution of choice from M. Torres Vega et al. it is the base of video traffic pattern recognition. It classifies video traffic according to its derived characteristics.

An amazing analysis by the group of researchers. The study however only covers QoE on Video application. While the focus is more on the predictive side of the analysis, the methodology, the video metrics, the technologies behind the data processing are not discussed in the study.

Malekmohamadi et al. studied the classification of video in accordance to the temporal and spatial data, using clustering method based on K-means and linear discriminant analysis. The study focused more on video analysis and there was a good performance result on 3D Video,

with a limited dataset [42]. Khan et al. used the video content variables to predict video performance by applying linear regression algorithm.

In the same angle of research, acknowledging the explosion and prevalence of multimedia applications, Hongli Luo and Mei-Ling Shyu address the Quality of Service (QoS) provision in mobile multimedia through a survey study [43]. In their study, they enumerate some of the real-time multimedia QoS requirements including bandwidth, Packet Loss ratio, delay, jitter and other complex QoS protocols. The study attributes the provision of QoS in Multimedia environment to the imposition of challenges related to the wireless channels and mobility of devices, shown in the below figure:

Table 5 Challenges Related to Wireless Channels and Devices *[43]*

Level	**Challenges**
Physical Channel	Wireless channels are not very reliable. Factors such as fading and multipath effects make the physical channel unreliable and high packet loss rate and bit error rate. The study also highlights the challenges to provide guaranteed end-to-end delays for applications.
Mobility	Ensure continuous service while in movement. Multimedia applications QoS should provide smooth handovers and continuous playback of video for example without compromising the QoS.
Routing	QoS topology should match the network routes. The changes in topology should be accommodated by the routing mechanism of the Network
Resource constraints	The Device constraints affects the application services' QoS. Factors such as Power consumption, Battery type, screen size and resolution, codecs. Those limitations pose a challenge to QoS.
Heterogeneity	Adapting the services to the capabilities of the Mobile devices, network access type and available infrastructure also pose a challenge to QoS.
Evaluation Metrics	Associating the QoE (Quality of Experience) of users to QoS by using metrics (KPIs).

Hongli Luo and Mei-Ling Shyu however focuses their research only on the provision of QoS at the MAC layer of the IEEE 802.11e WLAN and the cross-layer approach of QoS provision.

The study only focuses on the Network aspect of QoS and doesn't describe the mathematical model or techniques behind every QoS metric.

However, the study of Hongli and Shyu opens a good hole of challenges and research directions. With the growth of social applications such as Facebook, LinkedIn and all, a new area of research is exposed, social multimedia computing as discussed by [44].

2.4. Problem Statement

Communication Service Providers in general, and particularly Cellular Operators are experiencing challenges to monitor QoS/QoE on user's, application's and/or service's perspectives for data usage comparing to traditional voice quality. With millions of transactions generated by users and devices, it appears easier to find the root causes of high call failure rate than finding the cause of a low video or audio streaming speed in a network. Different services yield different quality requirements. Many researches, technologies and algorithms have been put it in place to address the QoS in the network environment; each of them presents or tackles specific applications and services or is network architecture depends. The reason would be different protocol applications respond differently to certain network conditions. For example, **WhatsApp, internet browsing,** and **other social media** will certainly give no big problem in the case of low throughput than **YouTube** and **torrent-download** which require high throughput for a good user experience. Factors such as Delay, Jitter, Packet Loss or Retransmission, Throughput are parameters that contribute to the efficient building of a solid QoS system for data application [7]. The above-mentioned parameters have a big impact on the QoS/QoE of users when using data applications. Let us imagine an on-line movie streaming with high latency and poor throughput, that insinuates a video that will be stopping constantly. And envisage the frustration of the user watching that video. With transactions downloaded from Internet or/and provided by Operators for research, Decision tree models, using regression methods will be examined and applied into the dataset to classify leaf of the tree contributing to the poor QoS/QoE of the user, service or network. The same applies to traditional voice services. A Customer who experiences poor voice quality when on the phone, or who consistently experiences call failures while on the move is likely to switch to a better Network to preserve the quality. To efficiently proceed with the research, the following question can be analysed: How can we efficiently identify and predict the areas of poor QoS in a

telecommunications network to improve network resources, customer retention and satisfaction?

During this research, we provide the basic framework for QoS and QoE using modern Data Analytics methods. The research is a cross-road of two important data fields, which are Telecommunications and Computer Science. This shows that the implication on Network Data Analysis requires more than just Telecommunications expertise.

CHAPTER 3. STUDY FRAMEWORK

While lots of researches in the area of Data Mining and Predictive Analytics in general have been focusing on developing and constructing important algorithms and data discovery patterns scripts to address specific use cases, across various domains, some researches have always focusing on bettering the processes of data pattern analysis, analysing the questions related to the interface, addressing the topic related to databases and data storage systems and discussing the best ways to visualize data mining output to support decision making processes [45]. The scope of the problem involving data pattern discovery is at first practical, meaning the need to find useful information in the large-scale data produced by users and devices. Although many researches currently focus on the development and formulation of unified data mining framework, the framework approach in this study is moving towards data mining as a single-step approach, addressing use cases related to individual tasks such as classification, regression and prediction. One of the important aspects of our study is to discover useful information in Cellular Network transactions to support decision making process. The Interest independently of the way data is presented (in a Service Quality Management Model, a Customer Experience Model, a Device Management Model or a recursive reporting habit), is finding patterns in the data that increases applicability [46]. We follow the approach introduced by [47] in the microeconomic view of data mining in which:

$$x \rightarrow f(x) \tag{3.1}$$

$$f(x) = \sum f_i(x) \tag{3.2}$$

Where x is the decision parameter which leads to maximum utility or value of the decision $f(x)$. We will be finding directly or indirectly the decision parameter that will help us get the optimal utility. The theory also shows that the maximum utility as shown in equation (2) is the sum of utilities for each case. $f_i(x)$ is a function of complex function. Let y_i be the data generated by customer i, x the decision parameter, the utility function is represented by:

$$f_i(x) = g(x, y_i) \tag{3.3}$$

The decision x belongs to a specific domain D which is the domain of all the decision that can be taken in to the decision (Customer Relationship, Service Quality Management, and Device Management). D is also a complex function of the same nature as the utility function. Nonetheless, J. Kleinberg establishes a difference between the domain of decisions and the value of decision assuming that D is internal to the Enterprise, in the scope of our study, the Communication Service Provider (CSP) and $f(x)$ reflects the communication process between the CSP and the other agents of the markets including subscribers, device manufacturers, services, etc. taking into consideration the above statement, the equation (1) is rewritten as:

$$f(x) = \sum_{i \in C} f_i(x) \qquad (3.4)$$

Where C is the set of additional factors that have a certain level of influence on the Enterprise operations. So, the technique is always to find the decision that would maximize the sum of the values of the decision (utility). However, an important observation by J. Kleinberg on the microeconomic approach of data mining is that data mining technique is useful when $g(x, y_i)$ is linear. This approach is used during the experimental use cases to determine the value of each use case. It is the base framework approach that is used during the study, in association with the Data Mining methodology or model selected and suitable for this research (discussed in chapter 5).

In his study of Conceptual framework for data mining, Yiyu Yao et al. describe three existing framework views used for Data pattern discovery, summarized as below [48].

Function-Oriented view	Theory-Oriented view	Process-view
•Targets the performance of the system. •Discover knowledge in Data •Considerable energy put in mining, searching and using patterns embedded in data storage systems	•Considerable enery on studies of Data mining techniques and how they relate to other external domains. •Establishes a relationship between scientific researches and data mining processes.	•Focuses on methodologies and Algorithms •Methods to improve existing algorithms and methodologies.

Figure 26 Currently existing Data Mining Framework views

3.1. Research Objectives and Aims

The aim of the study reflects the meaning that could be extracted from the title of the topic. Studying the possibility to get in-depth knowledge on user generated Network transactions and study different user and network data patterns which can convey important business decision rules. From simple categorization and classification to prediction, Communication Service Providers can detect in a quick way, with no need to implement expensive systems, services performances, Network performances and Device performances in the network. We divide the research objectives into two main sections: The Main objective which constitutes the ground of the research study and the specific objectives which provide the points that are addressed in this research.

3.1.1. Main Objective

The research study deals with large data generated by users, devices and services as captured from the Network. The main objective to facilitate and improve Quality of Service in the Telecommunications environment using Data Mining and predictive analytics techniques. The adaptation of modern techniques to Telecommunications user experience is a motivation to business strategy change. We move with the assumption that every piece of information in the network can help in decision making and business growth. Many open source data mining platforms are being developed by scientists. Exploiting those techniques can save a lot on Enterprise OPEX and CAPEX. The models will then be used or applied in to a data usage transaction set to identify based on pattern classifications, areas or domains of poor QoS in a Telecommunications network. The result of the models is a movement towards customer satisfaction and network improvement, giving a competitive edge to Communication Service Providers (CSPs). The study will leverage on the existing researches and models used to customize its scope to the Telecoms industry. Data mining is a field that is also being developed and researched upon by many universities and group of researchers.

3.1.2. Specific objectives

To enrich the main objective presented in section 3.1.1, other specific objectives of the research are stipulated. They are a projection of the benefits of the study detailed in section 3.3. The objectives are listed below:

- Conduct Explanatory Data Analysis (EDA) on Network and customer data as follow:
    - On CRM (Customer Relationship Management) data to uncover data patterns to determine the customers or subscribers that may terminate their contracts due to bad experience. Several algorithms are used to conduct the analysis. However, the aftereffect of the analysis is predicted by using algorithms with the best accuracy.
    - On User plane data to uncover data pattern related to service and Device behaviours.
- Develop a low-cost Service Quality Management, based on Data Mining processes to leverage on the current demand of Network operators and other CSPs. An efficient SQM system will help us attend the below objective:
    - Reduce the time it takes to identify the problematic leaf or spot in terms of QoS in a network, therefore, reducing the mean time to repair (time for Network engineers to attend to problems).
- Deduce effective decision rules based on specific predictors to be used on the decision trees.
- Efficiently analyse transactional dataset to classify services, network elements or subscribers at risk of poor QoS using Regression models with trees.

The objectives mentioned above will be addressed in the coming sections in the form of Data Mining methodology and use cases developed during the research study.

3.2. Research Core Questions

QoS can be improved and optimized if underperforming leaves or areas of the network can be easily identified, being able to pin point the cells, the devices or the services which have poor QoS is the ideal way to render optimization processes. A good structure of different aggregation domains (service, customer, nodes) of the network, the selection of variable indicators (predictor sets), the choice of the right Key Performance Indicators (KPI) and an efficient

mining technique can influence the QoS/QoE perception in a network. This is possible by establishing some linear relationship between different predictors such as throughput, latency, retransmission, domain name server (DNS) performance, which are explained in detail in chapter 5. Poor throughput can impact data QoS, High latency can impact data QoS, High Packet loss or Packet Retransmission can impact data QoS. If the above-mentioned parameters can be aggregated properly to define a KQI (Key Quality Indicator) on different levels of aggregation, the QoS can really be improved and optimized. The following questions constitute the Core of the research study:

- How can a CSP determine customers that are likely to terminate their contracts due to service or network performance using low cost methodologies?
- How can a CSP implement an efficient Service Quality monitoring system taking into consideration cost?
- What is the impact of IoT services on Device performance in the Network?

The core research questions however are motivated by a set of sub questions that act as pathways to the objectives. Few of the sub-questions are illustrated below:

- What are the parameters which are considered crucial in determining the retention index of customers (loyalty to the CSP)?
- How do you select the best algorithm to be used for Network data mining and predictive analytics problems?
- What are the top services performance metric in the Network?
- What are the problematic cells of the Network, having poor YouTube experience?
- What are the top devices that perform well on WhatsApp, based on the amount of data used and the number of customers who have that phone model?

3.3. Benefits of the Study

The benefits of the study are legion as it benefits both Communication Service Providers (CSPs), Data Analysis organizations and many other public sectors dealing with data in daily basis. In the scope of Telecommunications, the below benefits can be highlighted from the research study:

- Improve QoS monitoring in Network environments with a different view of the entire network comparing to traditional ways of conducting QoS improvement.
- With predictive capabilities, address potential customer attrition and ensure customer retention. Competitive advantage for Operators.
- The research will facilitate decision-making based on the reported QoS tree, enabling the operator to take a competitive edge by having a very clear Insight of the network, service and subscribers' performance and perceived QoS.
- Bring the concept of data mining and Machine Learning in front of the house meaning allowing organizations in various domains to use the benefits of the technologies to increase efficiency in business decision and competition.
- Despite the scientific and technological benefits of the study, there is individual benefit in terms of personal growth. There is a need in data analysts in the current market. According to the McKinsey Global Institutes reports there will be a shortage of talent necessary for organizations to take advantage of big data and machine learning. Therefore, the projection that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140 000 to 190 000 positions [49].

3.4. Delimitation of the Study

It is practically impossible to cover all the aspects of Quality of Service; Quality of Experience in the Entire Cellular Network and it is not painless to also cover the whole scope of Data Mining and Predictive Analytics techniques in this research study.

3.4.1. In Scope Topics

This research will be efficiently conducted to get the expected results but will be limited to:

- **User Plane Data experience**: which means focusing only on the real user data no control plane and no circuit switched voice. This can be another area of research also. The Cellular Network is made of different domains. In this area we will focus more on Data (Internet) related transactions.

- **CRM Data**: churn data analysis is done based on CRM data only. In the CRM system, contractual service data for customers and aggregated statistical usage of network services for all the subscribers are stored. Although the CRM appears to be rich in customer data warehouse, it does not provide a good insight on Quality of Service (QoS) and Quality of Experience (QoE) of each subscriber. To better optimize on user experience, more data management systems, as data sources would be of great importance including but not limited to probes, Performance Management systems, Fault management systems, and mostly CDRs (Call Data Records). However, it is still necessary to run Predictive Analytics on CRM data to study the patterns of customer data and build an optimized marketing strategy.
- **Decision trees models**: the scope covers regression with trees and classification with trees. Other prediction models such as Boosting, Random Forest will also be touched. However only two best algorithms will be used for predictions.
- **Quality of Service & Experience**: Service Quality Management and Device Quality management are in scope for the Data Analysis. While in this research experience on customer level is not addressed (an area of research on its own).
- Computer algorithms and models for the validation of the research. Although it is difficult to find a practical environment to validate the research, the result will be usable in a live environment.
- For solving the research question, data mining and predictive techniques will be used. This will include data analysis language such as data query languages, python or R, Apaches software.

R-Studio will be used as the Machine Learning environment and different visualization techniques can be adopted to show different output graphs.

3.4.2. Out of Scope Topics

- The techniques used in the process of data mining do not cover unsupervised data mining, which is also referred to as indirect Data mining. We are going to use fields of the dataset to run analysis and prediction in contrary to unsupervised where the system learns everything by itself [50].

- The theoretical aspect of Data Mining frameworks and models is not in the scope of the study. The practical aspect only is considered, the one-step process focusing on algorithms and methodologies that will be used for Data Mining.
- For QoS, Radio part is out of scope and constitutes an area of research on its own with the big question of Geo-Localisation in Communication Environment.

CHAPTER 4. THEORITICAL AND MATHEMATICAL BACKGROUND

To develop the theoretical and mathematical background of the research, it is important to understand the transactional datasets that will be dealt with along the experiments and development of our theories. The transactional data is divided and collected into two: Circuit-Switch transactions, which are related to traditional voice services and Packet-Switched User Plane (UP) transactions which are related to customers' data transactions. Although the research does not give the entire insight on the Cellular Network Architecture, the illustration of Network Interfaces data capturing is shown in the below figure. The figure also shows the types of data that are captured.

The data used is coming from a Mobile Network Operators. For privacy purposes, all sensitive information is hashed or encrypted.

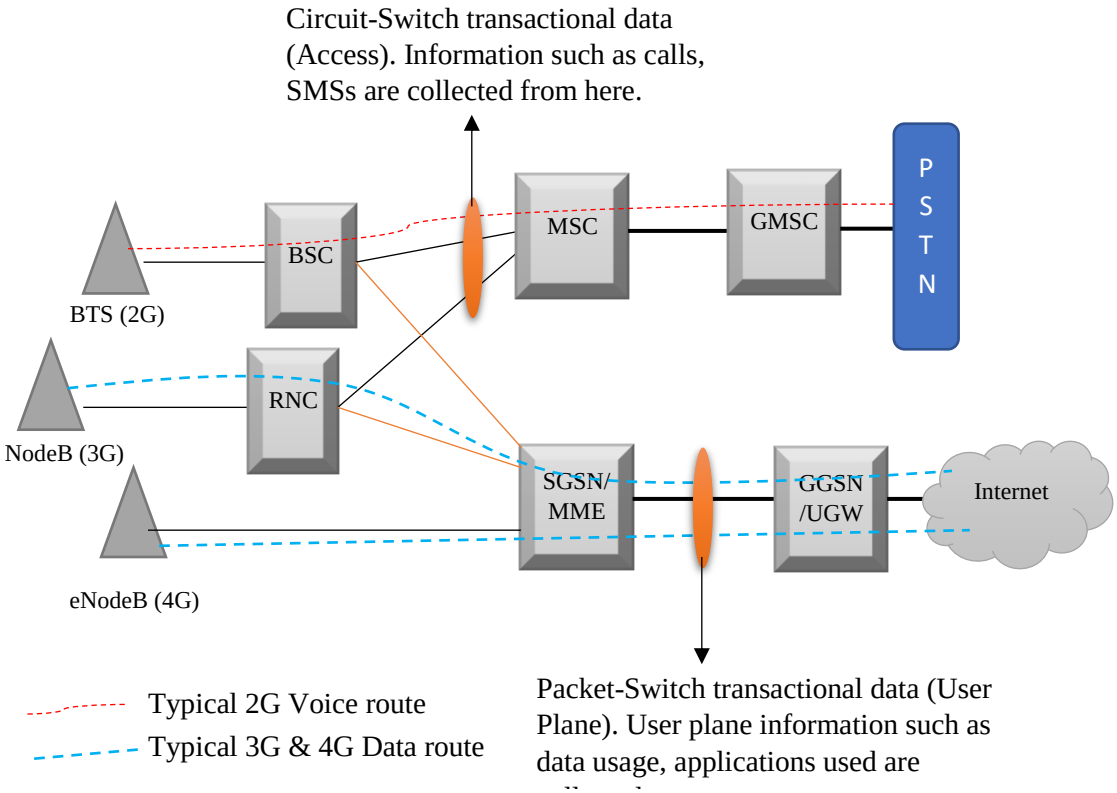


Figure 27 Simplified Cellular Network Architecture and point of transaction Data Collection

- BTS (Base Transceiver Station), NodeB, eNodeB are network equipments for wireless connection between the User Equipment (Telephone, Smart phones) and the Network.
- BSC, Base Station Controller (2G), RNC, Radio Network Controller (3G), equipments that control Radio resources and are the connection point to the Core Network.
- MSC (Mobile Switching Center) for Circuit-Switch, SGSN (Serving GPRS Support Node) for Packet-Switch: Core network equipments used for switching of Circuit-Switched services such as call release, call set-up, call routing and Packet-Switched services such as mobility management and user authentication respectively. The MSC and the SGSN perform the same task in different domains (Circuit and Packet respectively).
- GMSC (Gateway MSC) for Circuit-Switch: network equipment used to route calls outside the network. And when a call comes from outside the operator's network, the call is routed through the GMSC.
- GGSN (Gateway GPRS Support Node) for Packet-Switch: network equipment used to connect to the Internet or external packet-switch networks.
- MME (Mobility Management Entity): 4G network equipment and the main signalling equipment for LTE.
- UGW (Unified Gateway): 4G network equipment similar to the GGSN.

4.1. Understanding the Model Approach

The Packet-Switch dataset application is subdivided into 3 main categories: Streaming applications, Interactive Applications and Background Applications. The defined model allows data drilldown from High Level to lower level of the information, providing an efficient platform for business decision making. The model follows a tree approach to facilitate expansion of Information and providing a wide ground for managing and optimizing Network performance. The model is shown in the below figure. The model describes the SQM (Service Quality Management) tree model that is used in this research to build a low-cost Network Analytics system.

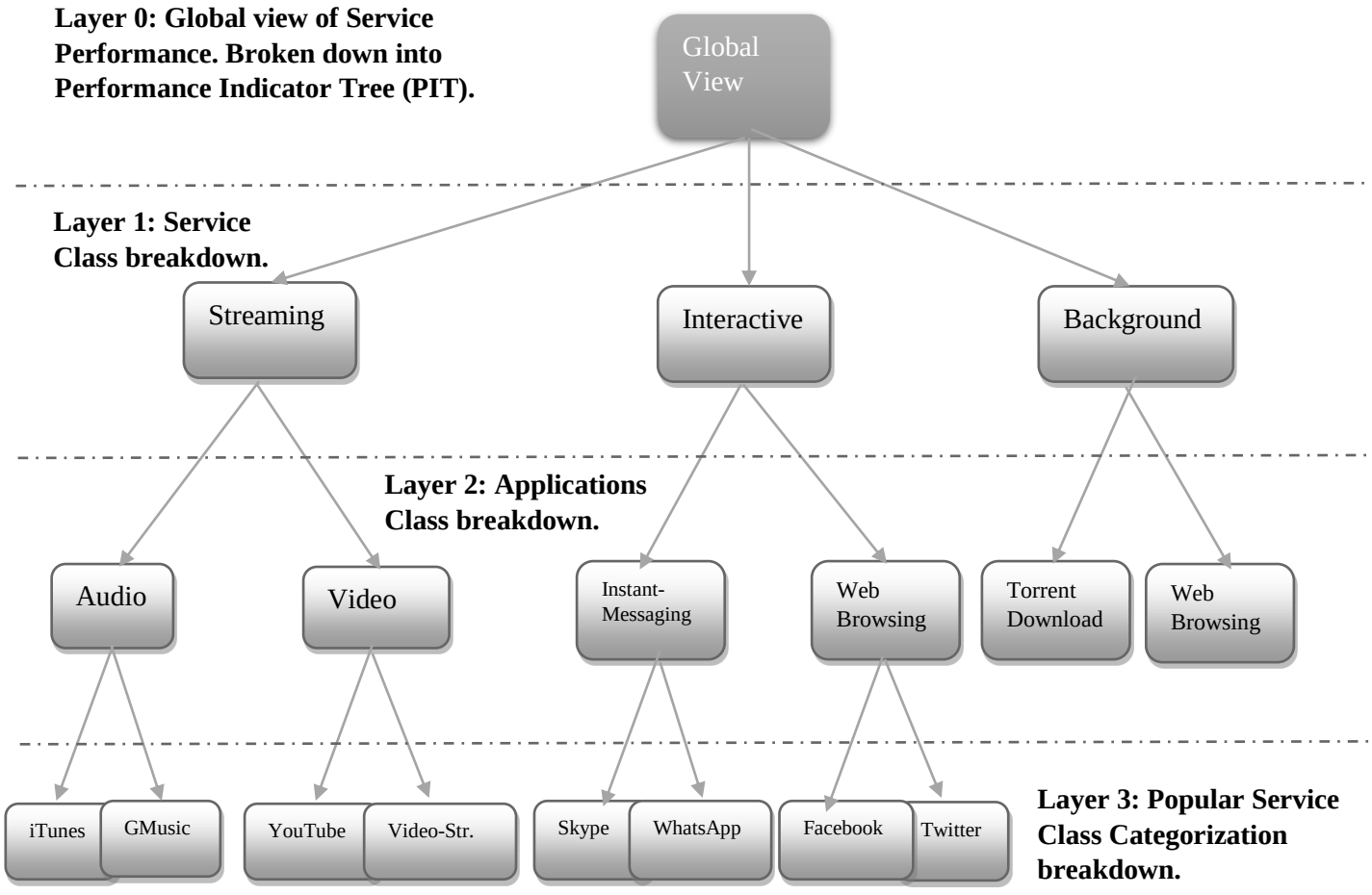
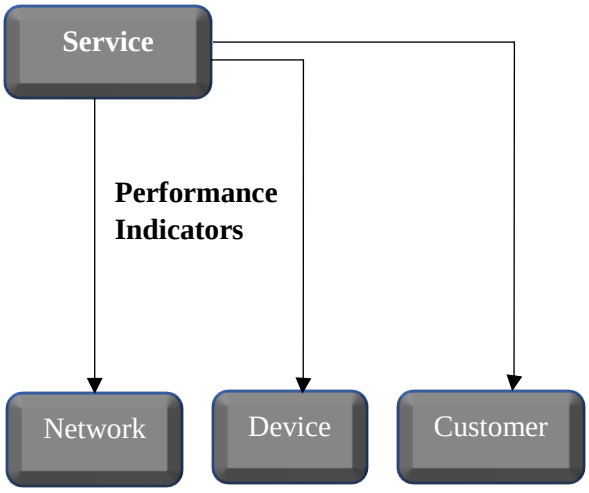


Figure 28 SQM Service Application Tree Model

The tree approach allows drilldown from upper layer to lower layer, extending the details as the drilldown is executed. To allow multiple aggregation dimension, it is important to oversee quality from different perspectives: perspective “User”, perspective “Network”, perspective “Device”. While focus is on “Service”, the model allows to oversee the SQM system from different several stand-points as shown in the below Figure.



- **Service**: The service represents the application that is being analyzed, as point of reference. For example, YouTube, Netflix, iTunes… or traditional Voice (Originating Calls).
- **Network**: Impact of Network Quality on a specific Service. For example, worst areas in South Africa in terms of YouTube streaming.
- **Device**: Impact of Device Quality on a specific Service. For example, device manufacturers with worst Netflix experience.
- **Customer**: Customer Experience based on specific service, which is a very important factor of QoE.

Figure 29 SQM Aggregation Model

4.2. Understanding the SQI and KPIs Approach

4.2.1. The KPI Approach

- Let us assume a customer with an active data session S, of t seconds in which the user is sending N Bytes of Data for a service application p. The estimated **throughput** $T(p)$ at which data is sent is given by:

$$T(p) = \frac{N(p)}{S(t)} \quad (4.2.1)$$

  $T(p)$ is expressed in Bytes over time. The required throughput depends on the application being used by the customer. Thus, throughput is one of the crucial Key Performance Indicators to consider for user, network and service performance. We consider both throughput in the Downlink and Uplink.

- As packets are transmitted across the network, they queue and experience delay. Let us assume latency L_{DL} the time it takes a packet to move from the Network to the User Equipment (End device) and the latency L_{UL} the time it takes to move from the User Equipment to the Network as illustrated in the below figure. The round-**trip time (RTT)** r_t is given by:

$$r_t = L_{DL} + L_{UL} \quad (4.2.2)$$

$$R_t(p) = \frac{1}{m}\sum r_t(p) \quad (4.2.3)$$

  Where $R_t(p)$ is the average round-trip-time (RTT) of a specific application p; and m is the number of network transactions or data sessions related to the application p. Therefore, the latency and the RTT are two important performance indicators for data performance, which will be considered in our Analysis.

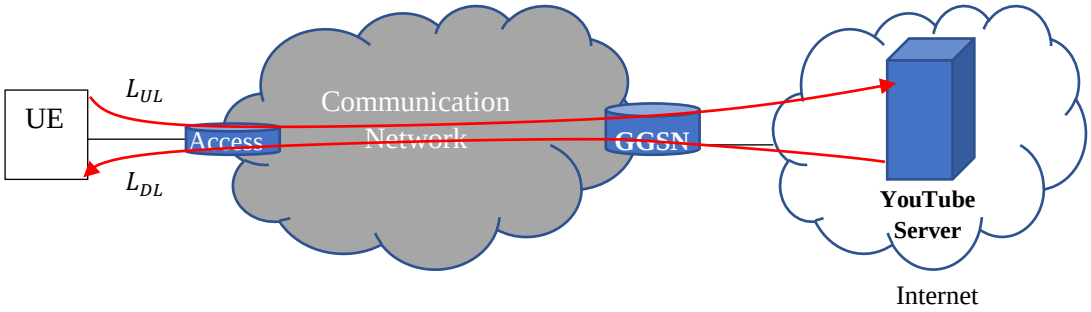


Figure 30 Illustration of Network Latency

- Another important performance indicator is the packet retransmission, which provides retainability in packet transmission.

  Let N_R be the retransmitted packets, the retransmission rate is given by:

$$r_{tx} = 100\frac{N_R}{N} \tag{4.2.4}$$

  So, the retransmission related to a specific application is:

$$r_{tx}(p) = 100\frac{1}{m}\sum_p \frac{N_R(p)}{N(p)} \tag{4.2.5}$$

  Where $N_R(p)$ and $N(p)$ are the Bytes Retransmitted and Bytes Used, for a specific service p respectively. For example, let us assume data for 1000 Bytes of data sent over the network, 500 Bytes are retransmitted or lost along the transmission line; that provides a 50% retransmission rate, which poorly affects a service performance. Thus, the **retransmission rate** (or packet loss in some case) is a very crucial performance indicator for data service performance.

- Let I be the number of address translation requests or DNS (Domain Name Server) queries, DNS failure rate D_r is given by:

$$D_r = \frac{I_f}{I_t} \rightarrow I_t = I_f + I_s \tag{4.2.6}$$

  Where I_f is the number of failed DNS requests, I_t is the total number of requests and I_s is the number of successful requests.

The filtering of service p is done in the algorithms. The same way the service is filtered, network and customer components will be filtered also as n and c respectively. Thus, the KPIs become as shown in the below table.

Table 6 KPI Aggregation Model

KPI	Service	Network	Customer
Throughput	$T(p)$	$T(n)$	$T(c)$
Data Usage	$N(p)$	$N(n)$	$N(c)$
Round Trip Time	$R_t(p)$	$R_t(n)$	$R_t(c)$
Latency (Uplink and Downlink)	$L_{DL}(p), L_{UL}(p)$	$L_{DL}(n), L_{UL}(n)$	$L_{DL}(c), L_{UL}(c)$
Retransmission	$r_{tx}(p)$	$r_{tx}(n)$	$r_{tx}(c)$
DNS Failure rate	$D_r(p)$	$D_r(n)$	$D_r(c)$

4.2.2. The SQI Model Approach

The Service Quality Index Model is used to provide the overall and combined Performance Indicator. It defines a non-standard algorithmic method to give the entire network quality index, helping to support business and operations decisions in High-Level or Layer-1 of our SQM Model. SQI is a weighted combination of KPIs. Depending on the Operator's needs, the SQI is modifiable and the weights can be changed as a bias factor.

Let β be the weight coefficient of Key Performance Indicator K. Since many KPIs can be defined as shown previously, the **Key Quality Index** (KQI) of a service class $Q(p)$ for n KPIs, is given by:

$$Q(p) = \sum_{i=1}^{n} \beta_i K_i \qquad (4.2.2.1)$$

If all the KPIs have the same weights value, then the formula becomes:

$$Q(p) = \beta \sum_{i=1}^{n} K_i \qquad (4.2.2.2)$$

Let α be the weight coefficient of Key Quality Index Q. The overall SQI for m KQIs is given by:

$$SQI = \sum_{j=1}^{m} \alpha_j Q_j \qquad (4.2.2.3)$$

(4.2.3.1) in (4.2.3.3) provides the general SQI formula:

$$SQI = \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_j (\beta_i K_i)_j \qquad (4.2.2.4)$$

If the coefficients are equal for all KPIs and KQIs, the formula becomes:

$$SQI = \alpha\beta \sum_{j=1}^{m} \sum_{i=1}^{n} K_{ij} \qquad (4.2.2.5)$$

4.3. Prediction Algorithms Used

In this section, we look at the various mathematical models for the Prediction and Machine Learning algorithms that are used for deep analysis on network transactions. Several models are used to train our different datasets, however, only the two best performing algorithms are then used on the test data set. Five algorithms are used to train the data. All the algorithms used are supervised machine learning algorithms. In this research, we will be predicting binary value true or false, 0 or 1. For example predict the possibility of a customer to churn or not. The algorithms and models will be used to predict binary values.

4.3.1. Classification and Decision Tree

Fundamentally used in Regressions and Classifications problems, we are going to use the tree model, precisely the classification tree, to train some of the Network and Customer outcomes. The algorithm has nodes and links. The nodes are attributes and the links are decisions that help build the tree. A node can have two or more links or branches. The concept root is also introduced as the top most node, corresponding directly to the best predictor. The example of a decision tree is shown in the below figure. Let us have a quick look at decision tree models. The response value in the decision tree is "*categorical*". Let us define two classes 0 and 1. The

probability that a customer with attributes x belongs to class 1 by $p(x) = \mathbb{P}\{Y = 1|x\}$. We use a step function and approximate the value $p(x)$.

$$\hat{p}(x) = \sum_{j=1}^{j} P_j I\ (x \in R_j) \qquad (4.3.1.1)$$

Where P_j represents the probability, in the R_j domain of $Y = 1$. To estimate the value of P_j, arithmetic mean is used:

$$\hat{p}_j = M(y_i : x_i \in R_j)$$

$$\hat{P}_j = \frac{1}{n_j} \sum_{i \in R_j} I(y_i = 1) \qquad (4.3.1.2)$$

the $\hat{p}_j$ is an indication of 1 in R_j. y is a binary and it is important that we calculate the deviance function. The deviance function D is given by:

$$D = 2n \sum_{j} \frac{n_j}{n} Q(\hat{p}_j) \qquad (4.3.1.3)$$

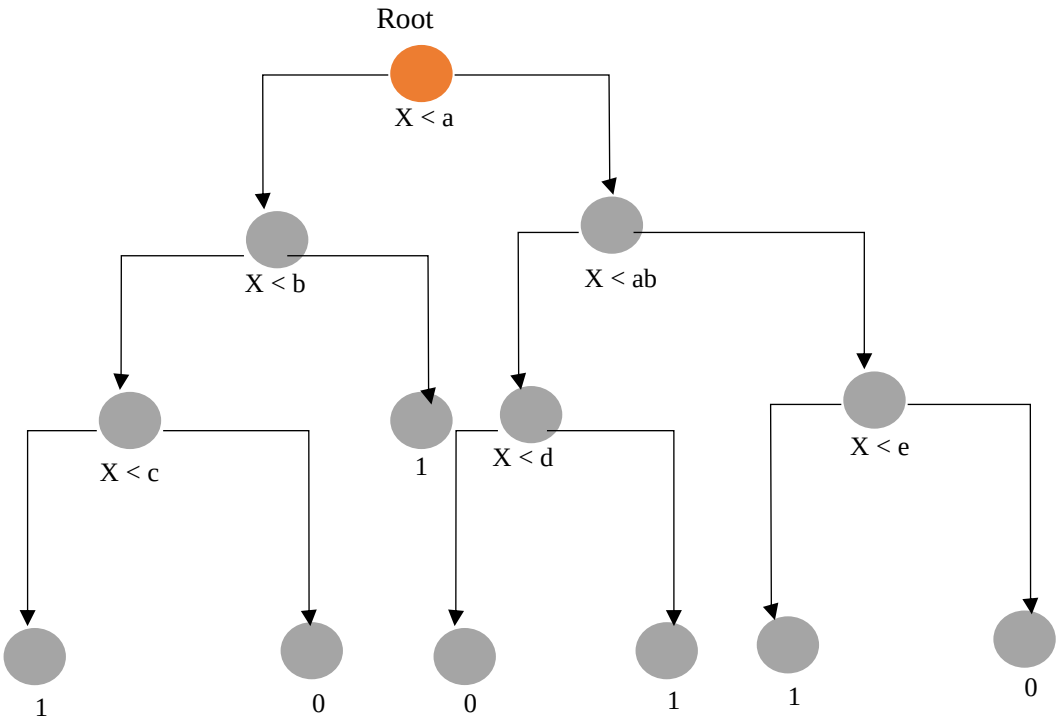


Figure 31 Example of Decision Tree Algorithm

The tree algorithm uses Entropy and Information Gain to build the decision tree. In the decision tree, all predictors will be included. From the root node, data is partitioned to subsets containing homogenous values; and the entropy helps compute a data sample's homogeneity. The entropy $E(S)$ is given by:

$$E(S) = -\sum_{i=1}^{K} P_{ji} \log_2 P_{ji} \qquad (4.3.1.4)$$

Where P is the probability of classification and S is the sample data. When we construct a decision tree for a dataset, the objective is to determine the attributes that yield the highest gain. The information gain of an attribute A is given by:

$$G(S, A) = E(S) - \sum_{v \in Values(A)} rac|S_v||S| \,.\, E(S_v) \qquad (4.3.1.5)$$

in which S_v is the subset of the data sample giving the attribute A a certain value v; $Values(A)$ is the set of values that the Attribute A can have.

In the experiment in this research, decision tree is used for classification processes, to train our dataset for customer information.

4.3.2. Gradient Boosting

Gradient Boosting is based on regression and classification trees. The model improves and optimizes accuracy using Gradient descent techniques. An important aspect of the model is that selecting variables is executed during the model fitting process, with no dependency on heuristic function methods such as stepwise selection [51].

Let y be the outcome of the prediction and $x_1, x_2, \ldots x_p$ as predicting variables or simply predictors. The objective is to define a relationship between y and $x = (x_1, x_2, \ldots x_p)^T$ such that an accurate prediction of y is obtained given x. The above is obtained by reducing the loss function $\rho(y, f) \in \mathbb{R}$ over the function f.

Basing on Logistic Regression, the algorithm predicts or estimates $\hat{y} \rightarrow \mathbb{P}(y = 1 \; or \; 0)|x$. In order to do that, the objective is to find parameters w and b to fulfil that condition.

$$\hat{y} = w^T x + b$$

Applying a step function (activation function) σ, $\hat{y} = \sigma(w^T x + b)$. For us to train a model using Regression model, the cost function needs to be calculated. Given the training dataset with m samples, $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$. The loss (error) function is given by the below equation and is a measure of the performance of the algorithm:

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$$

This is related to finding the square error. And the cost function used to find parameters w and b is given by:

$$J(w, b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i)$$

Gradient descent consists in minimizing the cost function to improve the performance of the model, using derivative functions in which the derivatives of the parameters are given by:

$$w := w - \alpha \frac{dJ(w, b)}{dw}$$

$$b := b - \alpha \frac{dJ(w, b)}{db}$$

In which w: and b: are updated parameters; and α is the Learning rate.

4.3.3. Random Forest

Developed by L. Breiman, Random Forest is a more recent Supervised Machine Learning algorithm, which is based on Regression and Classification tree [52]. The algorithm improves and optimizes the accuracy of prediction without overfitting the datasets. Contrary to other regression models, the predictor variables, in Random Forest are ranked in an unbiased way, which indeed provides the best accuracy [53]. The mathematical theory approach in this section is based on the white paper developed by Gerard Biau [54].

Let assume that we have a training dataset, $D_n = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ which has been randomly selected from a dataset distribution $D = \{(X_i, Y_i)\}$. The objective is to construct a classifier which from x be able to predict y. Several randomized regression trees $\{r_n(x, \theta_n, D_n)\}$ in which θ_n represents randomized variables of θ, are collected to form the Random Forest algorithm. All the random trees are aggregated:

$$\bar{r_n}(X, D_n) = E_\theta[r_n(X, \theta, D_n)] \tag{4.3.3.1}$$

E_θ is an expectation in relation to the three parameters, randomized variable, input parameter and selected dataset. E_θ can be implemented using Monte Carlo to generate random trees, and then select the average of each tree. The randomized variable θ shows the way cuts are performed when building each tree, it determines parameters such as selection of split and position. Each randomized tree is calculated by:

$$r_n(X, \theta) = \frac{\sum_{i=1}^{n} Y_i 1_{[X_i \in A_n(X,\theta)]}}{\sum_{i=1}^{n} 1_{[X_i \in A_n(X,\theta)]}} 1_{E_n(X,\theta)} \tag{4.3.3.2}$$

With:

$$E_n(X, \theta) = \left[\sum_{i=1}^{n} 1_{[X_i \in A_n(X,\theta)]} \neq 0\right] \tag{4.3.3.3}$$

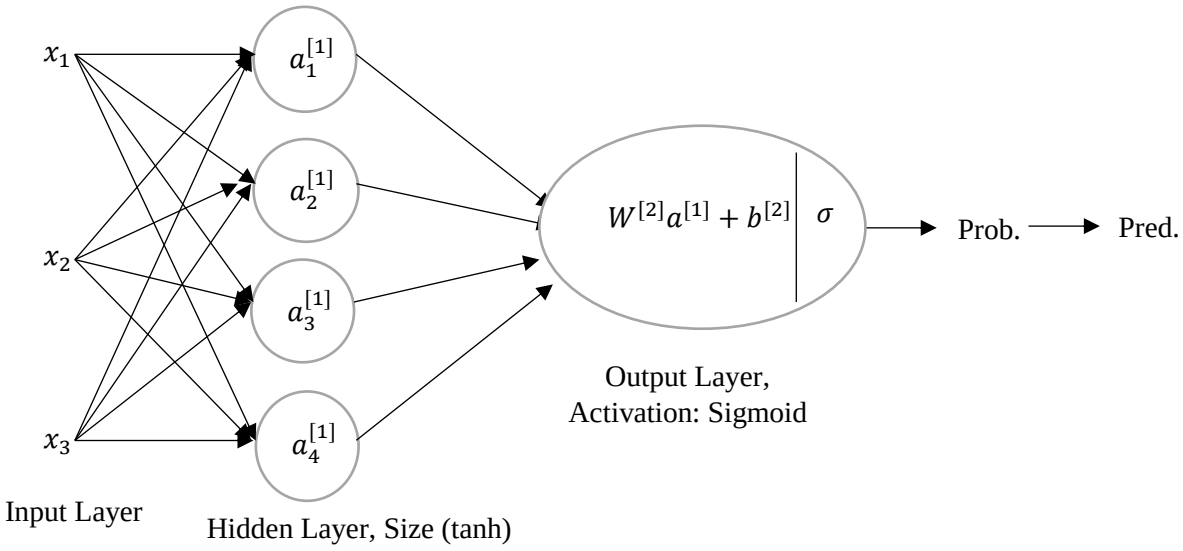
The average of all trees over all Y_i are computed from each randomized tree, for which X_i is in the cell of random partition, as X. $A_n(X, \theta)$ is the cell of the random data partition containing X. When a cell is empty, the model sets to 0 all the estimate.

The random forest regression estimate is then given by:

$$\bar{r_n}(X) = E_\theta[r_n(X, \theta)] = E_\theta\left[\frac{\sum_{i=1}^{n} Y_i 1_{[X_i \in A_n(X,\theta)]}}{\sum_{i=1}^{n} 1_{[X_i \in A_n(X,\theta)]}} 1_{E_n(X,\theta)}\right] \tag{4.3.3.4}$$

4.3.4. Neural Networks

An introduction to Neural Networks has been detailed in section 2.2.2.2. In this section, we go through the mathematical theories that we use in our Data Analysis. We build a Neural Network of one hidden layer as shown in the below figure. The objective is to calculate the Output of the Neural Network.



$$a^{[1]} = \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \\ a_3^{[1]} \\ a_4^{[1]} \end{bmatrix} = \tanh(z^{[1]}), \quad z^{[1]} = W^{[1]}X + b^{[1]}, \tag{4.3.1.6}$$

If we consider one training example or sample (input) $x^{(i)}$, the following equations apply:

- The output vector of the hidden layer related to the selected training example, $z^{[1](i)}$ is given by:

$$z^{[1](i)} = W^{[1]}x^i + b^{[1]} \tag{4.3.1.7}$$

  In which W is the Weight Vector and b is the bias vector.

- The activated Output for the hidden layer related to the selected training example, $a^{[1](i)}$ is given by:

$$a^{[1](i)} = \tanh(z^{[1](i)}) \tag{4.3.1.8}$$

Considering that the activation function for the hidden layer is a hyperbolic tangent function.

- The output vector of the Output Layer, related to the selected training example, $z^{[2](i)}$ is given by:

$$z^{[2](i)} = W^{[2]}a^{[1](i)} + b^{[2]} \tag{4.3.1.9}$$

In which $a^{[1](i)}$ has become the input vector to the Output layer.

- The activated Output of the second layer, which in this case is also the predicted value is given by the below equation. The second activation function is a sigmoid function.

$$\hat{y}^{(i)} = \sigma(z^{[2](i)}) \tag{4.3.1.10}$$

- The predicted value is given by:

$$\hat{y}^{(i)}_{pred.} = \begin{cases} 1 \; if \; a^{[2](i)} > value \\ \quad . \\ 0 \quad otherwise \end{cases} \tag{4.3.1.11}$$

- The cost function J can be compute as below:

$$J = -\frac{1}{m}\sum_{i=0}^{m}\left(y^{(i)}\log\left(a^{[2](i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - a^{[2](i)}\right)\right) \tag{4.3.1.12}$$

Until now we have implemented above the Forward path propagation. However, to train a Neural Network, it is also important to calculate or compute the Backward propagation, by deriving the functions of the forward propagation; the computation of backward propagation is also referred to as Gradient descent. The backpropagation mathematical computations are shown below:

$$dz^{[2]} = a^{[2]} - y$$

$$dW^{[2]} = dz^{[2]}a^{[1]^T}$$

$$db^{[2]} = dz^{[2]}$$

$$dz^{[1]} = W^{[2]T}dz^{[2]} * g^{[1]'}\left(z^{[1]}\right)$$

$$dW^{[1]} = dz^{[1]}x^T$$

$$db^{[1]} = dz^{[1]}$$

The methodology to build and train the Neural Network used in this research is shown and detailed in the next chapter.

CHAPTER 5. METHODOLOGY

Tackling Analytics problems such as Machine Learning and Predictive Analytics require a certain workflow and processes. Our methodology is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining), introduced in Chapter 2. However, to fit our problem, different customizations are applied to the model, based on the problem being solved. In this section, we detail the different methodologies used for the practical use case studies conducted. The actual execution of each process in the methodologies are followed and detailed in the next Chapter, Result and Discussion.

5.1. Predicting Churn, Practical Machine Learning for Telecoms CS Transactions, Use Case 1.

As defined in the context of the research study, the objective of the use case is to predict customers churn, based on the CRM (Customer Relationship Management) system in a simple but efficient way. The use case allows CSPs and Data Scientists to overlook the need to apply Machine Learning and Predictive Analytics to improve customer retention. The methodology and design used for this use case is shown in the figure 32. The primary question is to determine what needs to be predicted and which data is being used to predict the outcome.

- **Problem Definition**: Define the problem that needs to be addressed and the goal that needs to be achieved. A churn prediction problem, with a classification objective.
- **Data Collection and pre-processing**: Data fetching through File Transfer Protocol (FTP). Understanding of the dataset. And for this research, sensitive information has been hashed.
- **Exploratory Data Analysis**: In this section, the relationship between predictors and the main predictor, "CHURN_FLAG" is elaborated. Considered as a continuity of Data processing, it consists to know the dataset in detail, establish the interrelationship between different variables, and observe the behaviour of the variables towards the target variable.
- **Machine Learning and Data training**: Partition of the dataset and application of Machine Learning algorithms on the training set. Several algorithms are used to train and predict. The choice of Machine Learning algorithms is not the critical section of a Predictive Analytics process. Pre-processing, processing and the understanding of the

data to work on is. As the target variable “CHURN_FLAG” has been specified in the list of variables, the selected models must learn the interrelationship between the target variable and the rest of predictor variables, hence supervised Machine Learning algorithms are used. We focus more on tree-based models, including the Classification tree, Random Forest, Boosting models and gradient boosting models.

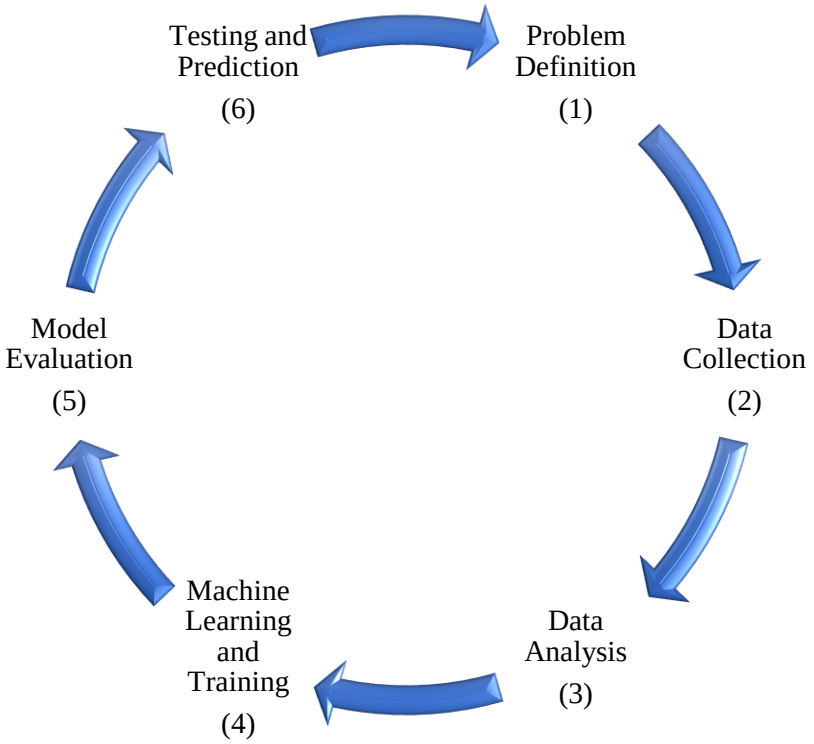


Figure 32 Methodology and Design for Churn Prediction

- **Model Evaluation**: Evaluate the performance of the used models to determine the best performing prediction algorithm to be used for the new dataset. Performance evaluation components are used to compare each model. The best performing algorithm is not just selected based on accuracy.
- **Testing and Prediction**: use the best performing model or algorithm to predict on the testing data and show the customers who are likely to churn. This is very important for the end-user.

5.2. Low Cost Data Service Quality Management: Methodology

The methodology used for the use case number two, which is the low-cost SQM is no different to the previous one. Still based on the CRISP-DM, Data Analytics method is customized. A database connection is established between the original data source and the processing server (here in represented by the High-performance PC used; refer to the Global Requirements section). The processing server caches the data frame in memory for rapid processing of information using Spark SQL or R. This means that SQL is used to query information from the dataset and display on the dashboard. The SQM system is separated into two parts: The Back-End and the Front-End as shown on the below figure, taking an in-house approach, contrary to a Cloud approach.

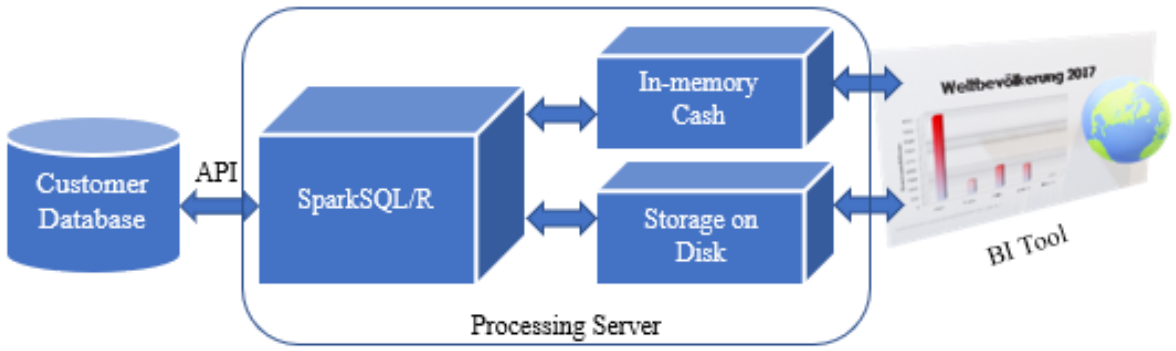


Figure 33 Low Cost In-House SQM System Architecture

- Database: contains Network Information from User Plane Gn Interface, with all data usage and used applications.
- Processing server: fetches information from the customer database, using an API collection algorithm. Based on an In-memory capable framework (Apache SparkSQL, R), data is loaded in memory and stored in the disk for batch processing or historical reference.
- Information is then visualized by the Front-End tool, in forms of dashboards.

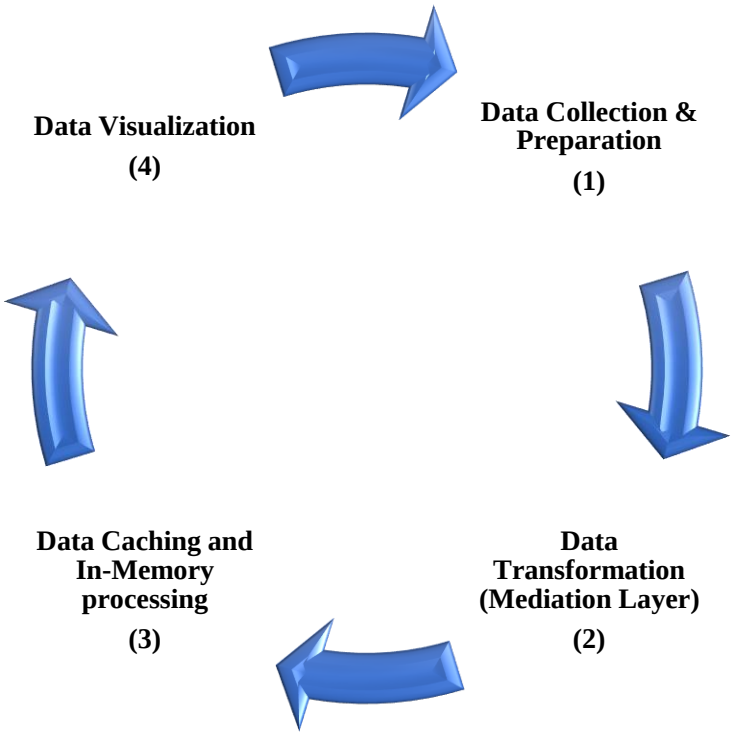


Figure 34 Low-Cost SQM Design Methodology

In its simplified scheme, relating the process to a Data Mining problem, data is collected from the customer database through API (Access Programming Interface) and pre-processed, then the same data is transformed to a format understandable by the SQM Back-End, the mediation layer. The data is cached in memory to allow rapid processing of data. The processed and re-structured data is smartly visualized to improve business decisions, through reports and dashboards. The methodology used for the SQM data analysis is shown in figure 34.

CHAPTER 6. RESULTS AND DISCUSSION

After discussing about the background of the research, the different works done around Data and Predictive Analytics and methodologies followed in this research, this chapter focuses on the experimentation and practical processes to support the methodologies detailed in the previous chapter and provides the ultimate answer to the research questions, in the goal to reach the objective of the research. The chapter is structured based on the use cases tackled. The codes and algorithms are shown where necessary.

6.1. Churn Prediction Analytics Using CRM Data in Real Time

We use R platform to run the experiments on the dataset, as it represents a solid platform for Data Analytics, Machine Learning and Predictive Analytics. The study takes the advantage of R's use of In-memory processing of information to speed the data computation.

6.1.1. Problem Definition and Objectives

Is it possible to depict subscribers that are likely to terminate their contract with a certain CSP or directly move to the competitor? Such is the general question to define the scope of the Analytics in this use case. Based on quantitative and qualitative data from CRM, we want to be able to predict if a certain customer will terminate his contract or not. The objectives of the Analytics are:

- Do an Exploratory Data Analysis on CRM information to examine relationships between different variables or predictors.
- Predict the outcome of subscribers, using the Algorithm with the best accuracy but training the data with more than one algorithm.
- Find based on tree models, the weight of predictors that have decided on the outcome of the model to support the use case.

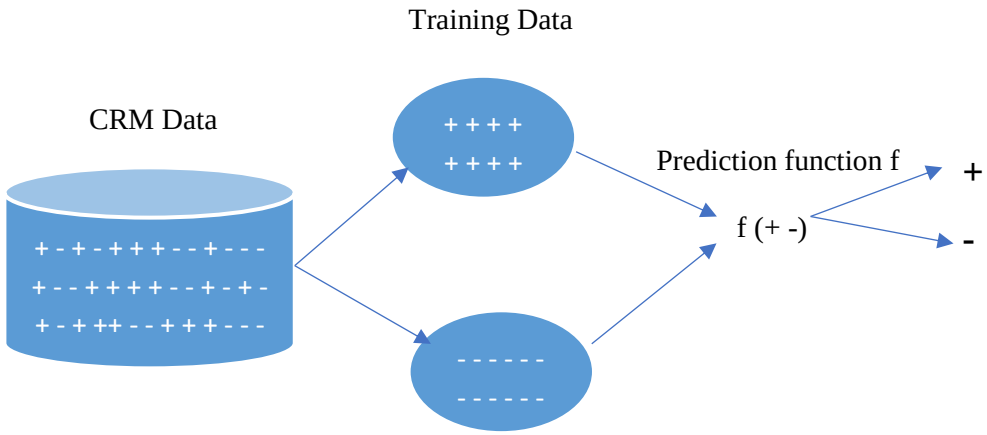


Figure 35 CRM Analytics Creed Illustration

The CRM information is partitioned, following the prediction analytics creed as shown in the figure 35

The CRM Data is partitioned and train using a predictive function such that the algorithm can predict Customers likely to churn. The data based on probability is divided to obtain the training set which contains different characteristics of the data from which the model learn the behaviour. A new set of CRM data is then fed to the prediction function to predict if a customer will churn or not.

6.1.2. Data Collection and Pre-processing

Information is stored in the CRM database in a structured way and collected in a systematic way either through API (Access Programming Interface) or through File management FTP (File Transfer Protocol) client. In this case, the data has been collected through FTP, fetching directly from a specific repository. We can't nevertheless, presume that CSP's data are stored in a single location for easy access, they are stored in different IT infrastructure depending on the transactions to be stored. The raw data used in this research paper are captured through FTP, coming from an Operator's CRM system.

6.1.2.1. Pre-processing and Hashing of Sensitive variables

The raw data collected from the CRM can have Null values for some variables and parameters that could negatively affect the prediction processes, models and results: these are outliers, missing values and incorrect fields. The objective in this phase is to reduce the amount of Garbage in the models and understanding the predictors. The variables are as follow, after pre-processing. The description allows the understanding of each predictor, even for those who have little or no knowledge of Telecommunications.

Table 7 Variable Predictors for Churn Prediction Analytics (CRM data)

Predictor Name	**Data Type of the Predictor**	**Description**
SERVICE_NUMBER*	Character	The Mobile Station International Subscriber Directory Number. This is the Mobile number
SERVICE_LINE	Character	Types of contracts a subscriber can be on.
SERVICE_PLAN	Character	Customer Service Plan
SUBSCRIBER NUMBER*	Character	Unique ID for Subscriber Identification in the Network.
SUBSCRIBER STATUS	Character	Status of Subscriber in the Network.
ACTIVATION_MONTH	Character	The Activation date and month of the subscriber contract.
REGION	Character	the region where the subscriber is based. The region of the contract initiated
GENDER	Character	Gender of the Subscriber.
CUSTOMER_ID*	Character	Unique number that identifies the customer in the CRM database
INT_CALLS	Integer	Number of International calls made by a customer
INT_CALLS_REVENUE	Decimal	Revenue generated on the International transactions by a customer.
CALLS_TOTAL_NUMBERS	Integer	Total Number of Calls by a user, including both national and international
CALL_REVENUE	Decimal	Revenue generated on all calls for a user.
CHARGEABLE_DURATION	Decimal	The amount of time billed by the CSP on a user service

TOTAL_OUTGOING_SMS	Integer	Total number of SMSs sent by a user
SMS_REVENUE	Decimal	Revenue generated on SMS.
CHARGEABLE_UNIT	Integer	Amount of Billable SMS Units of a user.
TOTAL_DATA_VOLUME	Decimal	Amount of Data used for Internet and other related Packet services, by a user.
DATA_REVENUE	Decimal	Revenue generated by a user on Data
CHARGEABLE_VOLUME	Decimal	Amount of Data Volume billed on by a CSP to a user.
INTERNATIONAL_PLAN	Character	Indicates if a user has got international plan in place.
CHURN_FLAG	Character	Indicates if a user has churned already or not.

The training dataset has got 24 predictors and variables over information from 5000 customers, in which the predictor variable is the field “CHURN_FLAG”. From the table, all the variables with an * mark indicates sensitive information and is hashed or encrypted.

- Connecting directly to the database and Loading the data in Memory:

```
library(dbConnect)
lapply(dbListConnections(MySQL()),dbDisconnect)
dbconnection=dbConnect (MySQL (), user="***", password="***",
db="churn_prediction_db", host="localhost")
db_frame=data.frame(dbGetQuery(dbconnection, "select * from crm_churn_dataset;"))
Mydata=data.frame(dbGetQuery (dbconnection, "select * from churn_proc_dataset;"))
Mydata$CHURN_FLAG=as.factor(Mydata$CHURN_FLAG)
Mytest=data.frame(dbGetQuery (dbconnection, "select * from churn_proc_testset;"))
Mytest$CHURN_FLAG=as.factor(Mytest$CHURN_FLAG)
```

- Viewing the internal data structure after being loaded in memory

```
str (Mydata)
```

Figure 36 Internal Structure of the Training Set

6.1.2.2. Computing the Min-Max Normalization, Mean, Median and Spread

In the normalization process, the distance between fields' values are adjusted to avoid the tendency of great values to influence the results of the model: scale standardization. The Minimum and Maximum values are found for every numerical predictor fields. This is to show how far the maximum value is from the minimum. The min-max normalization of a column Y, denoted Y^*_{mm} is given by:

$$Y^*_{mm} = \frac{Y - \min(Y)}{\max(Y) - \min(Y)} \qquad (6.1.2.2)$$

The mean is used to determine the average value for every numerical field. The mean is given by:

$$\bar{y} = \frac{\sum y}{n} \qquad (6.1.2.3)$$

Where y is the values of each filed and n is the sample size of the data, in other way the number of records in the data. The median is also calculated and used because the mean is easily affected by the outliers and noise; the median is the centre of the field with the latest changed to ascending order.

```
summary(db_frame)
```

```
GROUP_SERVICE_LINE SUBSCRIBER_STATUS_201708    REGION            INTL_CALLS     INTL_CALLS_REVENUE
Length:5899        Length:5899              Length:5899        Min.   :  0.00   Min.   :      0
Class :character   Class :character         Class :character   1st Qu.:  7.00   1st Qu.:   2812
Mode  :character   Mode  :character         Mode  :character   Median : 10.00   Median :   7749
                                                              Mean   : 15.85   Mean   :  15583
                                                              3rd Qu.: 18.00   3rd Qu.:  18067
                                                              Max.   :523.00   Max.   :1059879
CALLS_TOTAL_NUMBER TOTAL_VOICE_DURATION  CALL_REVENUE     CHARGEABLE_DURATION TOTAL_OUTGOING_SMS
Min.   :   0.0     Min.   :      0       Min.   :      0   Min.   :     0      Min.   :   0.00
1st Qu.:  76.0     1st Qu.:   5235       1st Qu.:   5296   1st Qu.:  4914      1st Qu.:   0.00
Median : 187.0     Median :  14788       Median :  14925   Median : 14268      Median :  13.00
Mean   : 280.7     Mean   :  24846       Mean   :  29399   Mean   : 24120      Mean   :  39.53
3rd Qu.: 367.0     3rd Qu.:  33986       3rd Qu.:  34778   3rd Qu.: 32857      3rd Qu.:  43.00
Max.   :5016.0     Max.   :363294        Max.   :1155175   Max.   :330741      Max.   :6403.00
 SMS_REVENUE       CHARGEABLE_UNITS  TOTAL_DATA_VOLUME    DATA_REVENUE     CHARGEABLE_VOLUME   INTERNATIONAL_PLAN
Min.   :     0.0   Min.   :   0.00   Min.   :0.000e+00   Min.   :      0   Min.   :0.000e+00   Length:5899
1st Qu.:     0.0   1st Qu.:   0.00   1st Qu.:1.257e+05   1st Qu.:      0   1st Qu.:5.202e+03   Class :character
Median :     0.0   Median :  12.00   Median :5.706e+08   Median :      0   Median :5.568e+08   Mode  :character
Mean   :   818.7   Mean   :  36.81   Mean   :2.009e+09   Mean   :   8261   Mean   :1.991e+09
3rd Qu.:   456.1   3rd Qu.:  39.00   3rd Qu.:2.197e+09   3rd Qu.:   1697   3rd Qu.:2.172e+09
Max.   :213290.0   Max.   :6403.00   Max.   :5.830e+10   Max.   :3125711   Max.   :5.822e+10
CHURN_FLAG
No :4999
Yes: 900
```

Figure 37 Computation of Data Characteristics

6.1.2.3. Identification of Data Outliers in Numerical Predictors

It is necessary to check and double check our CRM dataset, to highlight values of numeric fields that are not following the trend of the rest of the data. These are known as outliers and can negatively impact the performance and accuracy of the model. Certain statistical methods thus, are sensitive to outliers [20]. The below figures show the graphs of numerical predictors. shows the Histogram plot of International calls against their count. The graph shows the presence of no outliers, or negative numbers of international calls. For example, there is no outlier on “negative” number of calls or international calls. number of calls can also be equal to zero. Thus, all the values are falling in between the range they should be.

```
require(gridExtra)
 par(mfrow=c(1,2))
 hist (db_frame$INTL_CALLS, breaks = 200, xlim = c(0,600),col = "blue", border = "blue", ylim = c(0,600), xlab = "# International Calls", ylab = "Count of International Calls", main = "Count of Internation Calls")
 #par(mfrow=c(1,1))
```

```
 hist (db_frame$CALLS_TOTAL_NUMBER, breaks = 200, xlim = c(0,6000),col = "blue", border = "blue", ylim = c(0,400), xlab = "Total Calls", ylab = "Count of Calls", main = "Count of Total Calls Distribution")
  #par(mfrow=c(1,2))
  hist (db_frame$TOTAL_OUTGOING_SMS, breaks = 200, xlim = c(0,6500), col = "blue", border = "blue", ylim = c(0,500), xlab = "Total SMS sent", ylab = "Count of Outgoing SMS", main = "Count of Tot.Outgoing SMS Distribution")
```

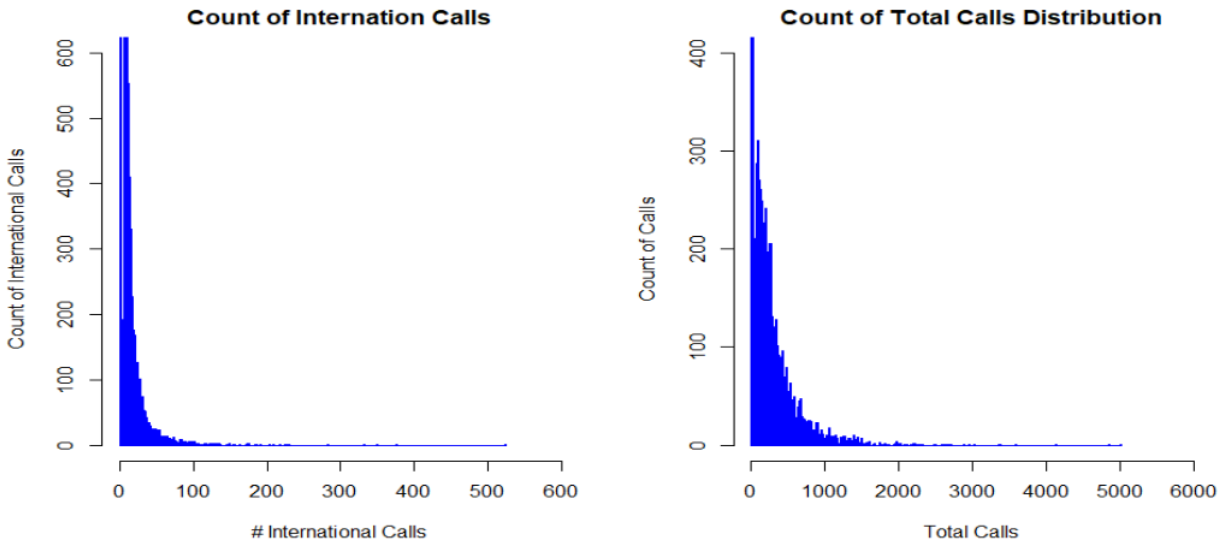


Figure 39 Histogram of Calls to Check Outliers

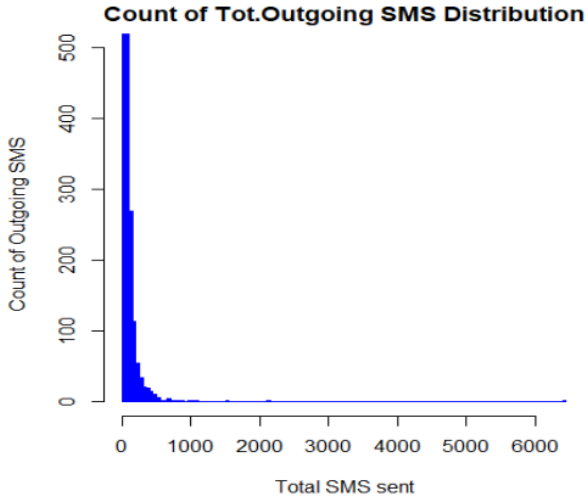


Figure 38 Histogram of SMS counts to check outliers

The graph shows that Subscribers in the send between 0 to 1000 SMSs, with few subscribers having more than 1000 SMSs, which is a normal behaviour in the network. Illustrated in the left figure. no abnormal phenomena or outliers are observed such as negative number of SMSs.

The same we have verified outliers for calls, data and SMS count, we verify the outliers for Revenue data to ensure that no abnormal data is present in the dataset. The below figure shows the Revenue information against customer services including Data, Voice and SMS.

```
require(gridExtra)
par(mfrow=c(2,2))
plot (db_frame$TOTAL_DATA_VOLUME, db_frame$DATA_REVENUE, xlim =
c(0,70000000000), ylim = c(0,350000), xlab = "Total Data Volume", ylab = "Data Revenue", main =
"Data_V vs. Revenue", type = "p", col = "blue")
plot (db_frame$TOTAL_OUTGOING_SMS, db_frame$SMS_REVENUE, xlim = c(0,500), ylim =
c(0,27000), xlab = "Total Outgoing SMS", ylab = "SMS Revenue", main = "Outgoing SMS vs.
SMS_Revenue", type = "p", col = "blue")
plot (db_frame$CALLS_TOTAL_NUMBER, db_frame$CALL_REVENUE, xlim = c(0,5000), ylim
= c(0,240000), xlab = "Total Number of Calls", ylab = "Revenue on Call", main = "Data_V vs.
Charg.Unit", type = "p", col = "blue")
plot (db_frame$INTL_CALLS, db_frame$INTL_CALLS_REVENUE, xlim = c(0,50), ylim =
c(0,30000), xlab = "Total Number of International Calls", ylab = "Revenue on International Calls",
main = "# Int Calls vs. Revenue", type = "p", col = "blue")
```

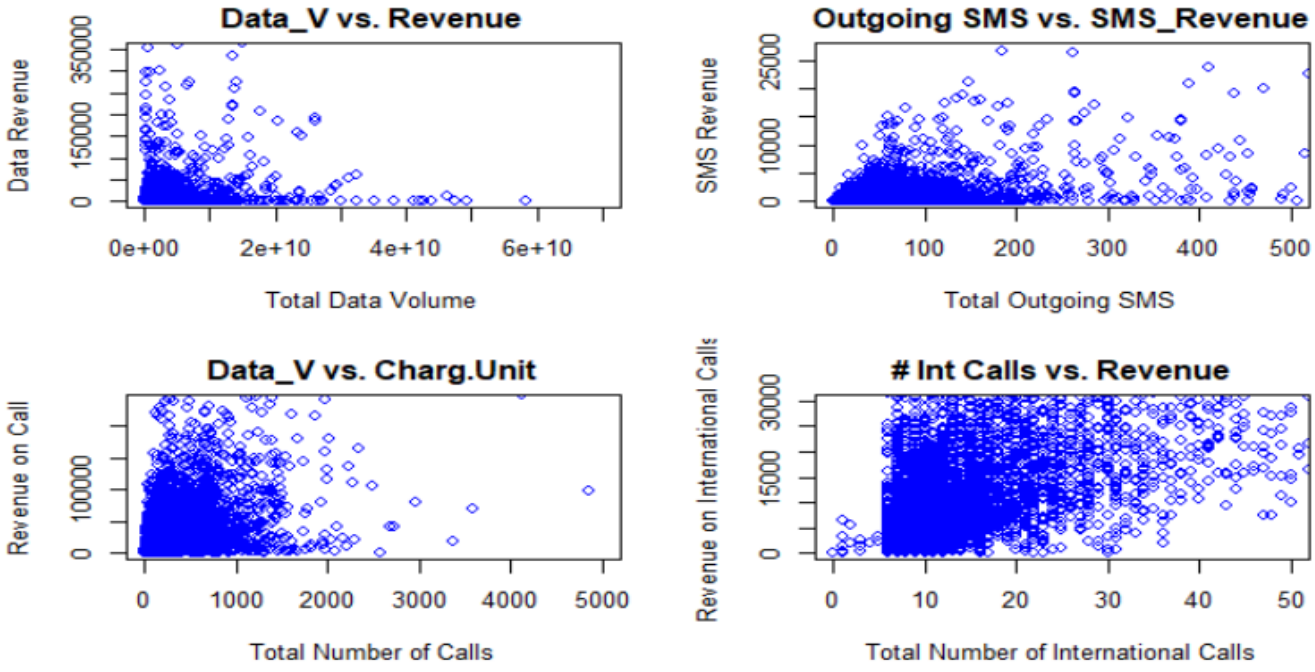


Figure 40 Data Plot for Outlier verification Revenue Data

Another important numerical variable or predictor is the chargeable unit and duration which also need to be checked for outliers. The graph is shown below. The same codes as above are modified to display different predictors. We can therefore, see that there are no outliers in our datasets. No need to apply processes such as Imputation of Data to solve null values and abnormal values.

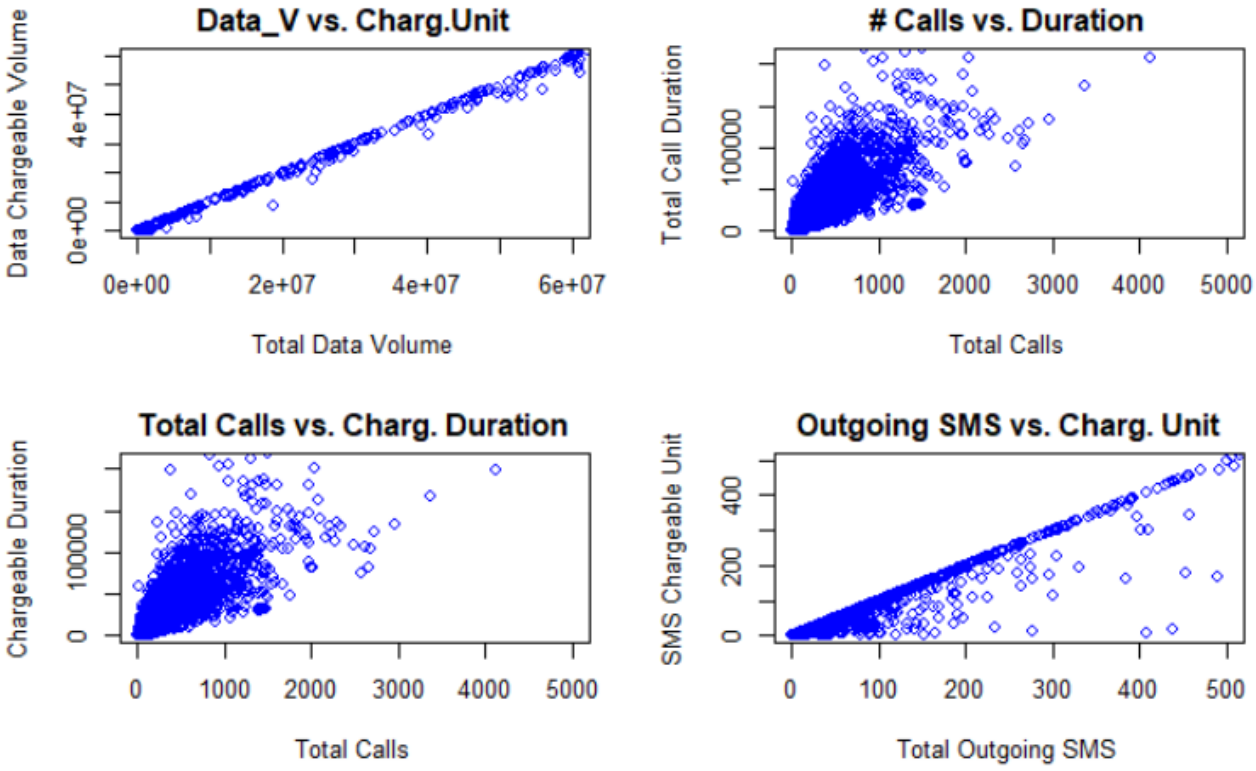


Figure 41 Data Plot for Outlier Check on Durations and Chargeable Units

6.1.3. Exploratory Data Analysis

6.1.3.1. Categorical Variable Analysis

The categorical variables on the dataset contains the characters data type. The objective in this section is to determine in advance from the CRM dataset, the patterns that will assist scaling down the proportion of churners. The below table illustrates the proportion of Subscribers who have churned vs. those who have not churned. The percentage of churned customer is 15.25%.

```
rate_churn=100*sum(db_frame$CHURN_FLAG=="Yes")/length(db_frame$CHURN_FLAG)
 rate_churn
 sum_flag=summary(db_frame$CHURN_FLAG)
 sum_flag
 barplot (sum_flag, ylim = c(0,5000), main = "Churn Proportion in the Dataset", col = "blue")
```

```
[1] 15.25682
  No  Yes
4999  900
```

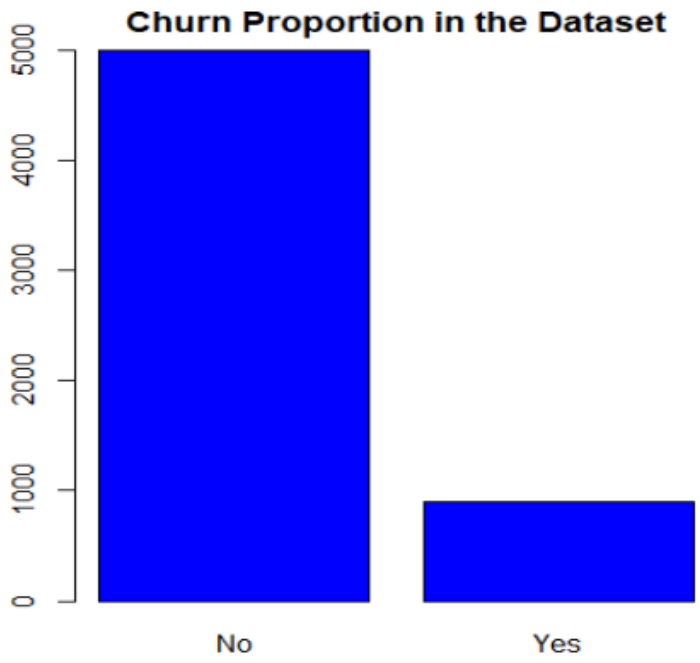


Figure 42 Subscriber Churn by International Plan Subscription

The categorical variables relationship or proportion to the CHURN_FLAG is shown in the below figure.

- Churn proportion per service line: provides the number of churners per customer service line
- Churn proportion per Region: displays the distribution of churners per geographical locations.
- Churn proportion per International Call Subscription: displays the churners based on the fact that they have international subscription or not.

```
require(gridExtra)
par (mfrow=c (2,2))
counts=table (db_frame$CHURN_FLAG, db_frame$GROUP_SERVICE_LINE, dnn = c("Churn_Flag","Group_Service_Line"))
counts
serviceLinechurn=table (db_frame$CHURN_FLAG, db_frame$GROUP_SERVICE_LINE)
  barplot (serviceLinechurn, legend=rownames(serviceLinechurn), col = c("blue", "black"),
       ylab="Count of Churned Subs", xlab = "Service Line Type",
       main="Churn Proportion per Service Line")

counts2=table (db_frame$CHURN_FLAG, db_frame$REGION, dnn = c("Churn_Flag","Region"))
counts2
```

```
regionchurn=table (db_frame$CHURN_FLAG, db_frame$REGION)
  barplot(regionchurn, legend=rownames(regionchurn), col = c("blue", "black"),
        ylab="Count of Churned Subs", xlab = "Region",
        main="Churn Proportion per Region")

counts3=table (db_frame$CHURN_FLAG, db_frame$INTERNATIONAL_PLAN, dnn = c("Churn_Flag","Internat. Plan"))
counts3
intplanchurn=table (db_frame$CHURN_FLAG, db_frame$INTERNATIONAL_PLAN)
  barplot (intplanchurn, legend=rownames(intplanchurn), col = c("blue", "black"),
        ylab="Count of Churned Subs", xlab = "Int. Plan Subscription",
        main="Churn Prop. per Int. Plan Subscription")
```

```
Group_Service_Line
Churn_Flag Hybrid HybridBroadband Postpaid PostpaidBroadband PostpaidFTTH
       No   2327             178     2463                31            0
       Yes   298             192      302               105            3
          Region
Churn_Flag Central EasternCape Gauteng  KZN NorthRegion Unknown WesternCape
       No      246         112    2806  469         638       0          728
       Yes      49          49     296  168         181      61           96
          Internat. Plan
Churn_Flag    N    Y
       No     0 4999
       Yes  880   20
```

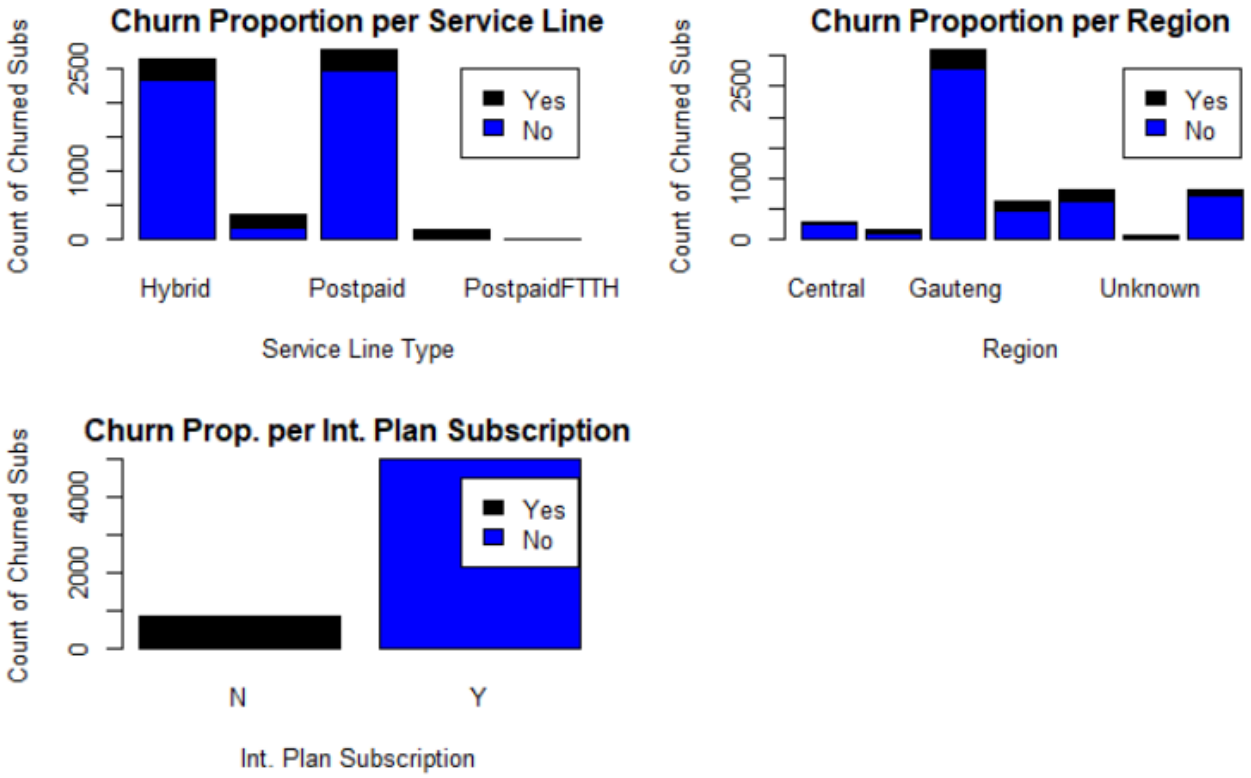


Figure 43 Data Relationship between Categorical Variables and the predictor variable

6.1.3.2. Numerical Variables Analysis

The goal to analyse numerical variables is to navigate in detail into the data, uncovering interrelationship between different numerical variables and between variables and the target predictor. Using R-plotting and the result in the below figure, the observations below are pin-pointed:

- If we compare call revenue to international call revenue, we notice that churners are in low usage, showing a trend-like or linear behaviour.
- If we compare the Call revenue to SMS revenue, we observer a similar observation, with churners showing a low revenue stream.
- If we compare Data revenue to call revenue we also see that churners show a low revenue stream on both data and calls. The observations will help in building the prediction models in the sense that churners seem to be low income generating users.

```
require(gridExtra)
 plot1=qplot (INTL_CALLS_REVENUE, CALL_REVENUE, colour=CHURN_FLAG, data=db_frame)
 plot2=qplot (CALL_REVENUE, SMS_REVENUE, colour=CHURN_FLAG, data=db_frame)
 plot3=qplot (CALL_REVENUE, DATA_REVENUE, colour=CHURN_FLAG, data=db_frame)
 plot4=qplot (CHARGEABLE_VOLUME, CHARGEABLE_DURATION, colour=CHURN_FLAG,
data=db_frame)
 grid.arrange(plot1, plot2, plot3, plot4, ncol=2)
```

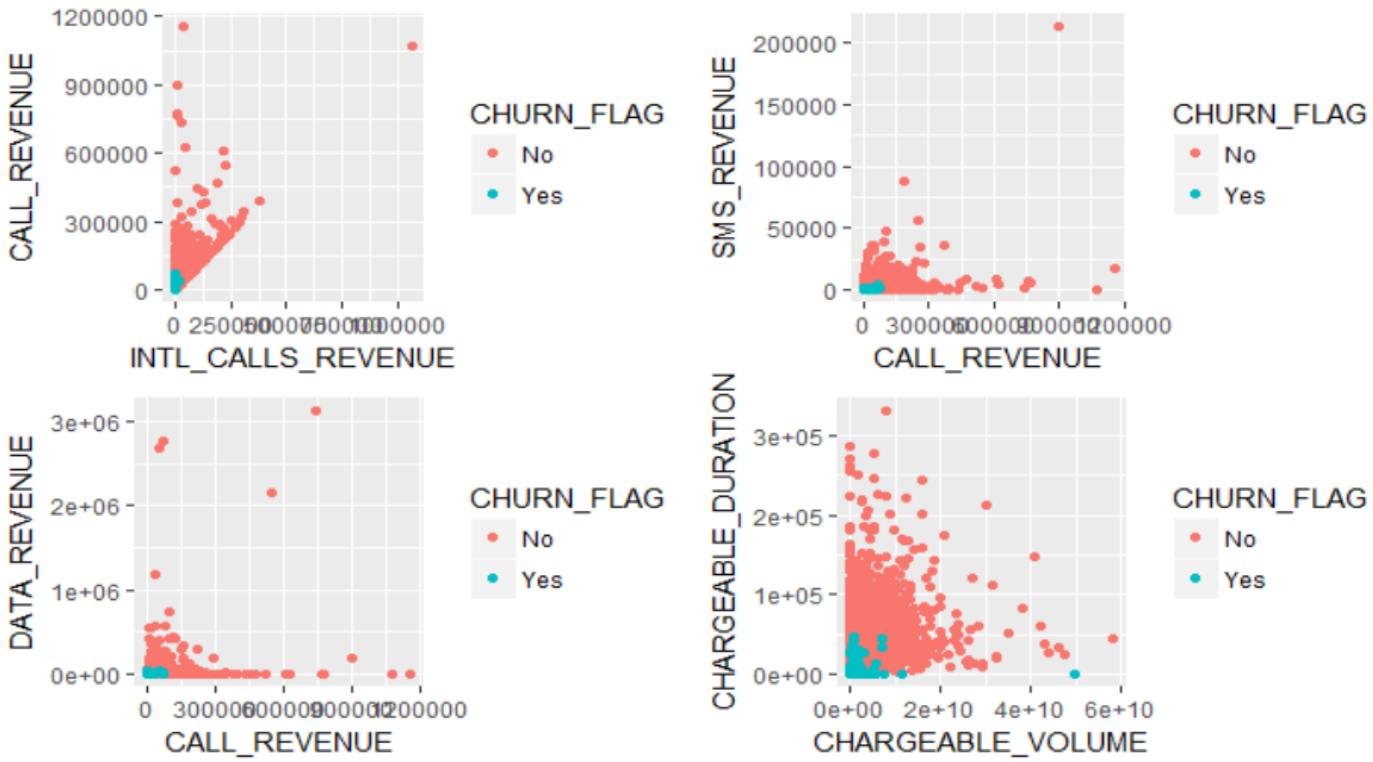


Figure 44 Exploratory Data Analysis for Numerical Variables

6.1.4. Machine Learning and Training

6.1.4.1. Data Partitioning and Model Fitting

The processed or cleaned dataset is split into two sub datasets, training and testing sets. The partition is such that 70% of the dataset is used to train the model and 30% is used to test the model. The below table shows the dimension of the split datasets.

Table 8 Dimension of training and testing datasets

	Number of Records	Number of Predictors
training dataset:	4130	17
Testing dataset:	1769	17

```
set.seed(332335)
db_frame=db_frame[,-c(1,5,7,9,10)]
inTrain=createDataPartition (y=db_frame$CHURN_FLAG, p=0.7, list = FALSE)
db_training=db_frame[inTrain,]
db_testing=db_frame[-inTrain,]
```

inTrain is the partitioned dataset variable, db_frame is the main processed dataset, CHURN_FLAG is the target variable, db_training is the training dataset and db_testing is the testing dataset.

6.1.4.2. Model Fitting

6.1.4.2.1. Classification Trees

```
library(caret)
model_lm=train(CHURN_FLAG~., method="rpart", data=db_training)
model_lm
print(model_lm$finalModel)
```

```
CART

4130 samples
  16 predictor
   2 classes: 'No', 'Yes'
```

```
No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 4130, 4130, 4130, 4130, 4130, 4130, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.0000000   0.9994186  0.9977340
  0.4992063   0.9994186  0.9977340
  0.9984127   0.9448148  0.6382457

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.4992063.


n= 4130

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 4130 630 No (0.8474576271 0.1525423729)
  2) INTL_CALLS>=5.5 3501    1 No (0.9997143673 0.0002856327) *
  3) INTL_CALLS< 5.5 629    0 Yes (0.0000000000 1.0000000000) *
```

Based on the model fit's output result, the classification and Regression tree model provides an accuracy of 99.94%. the below figure shows the decision tree used for our Classification model. The model is based on the International plan and calls. Customers who have churned have made less international call.

```
library(rpart.plot)
fancyRpartPlot(model_lm_n$finalModel)
```

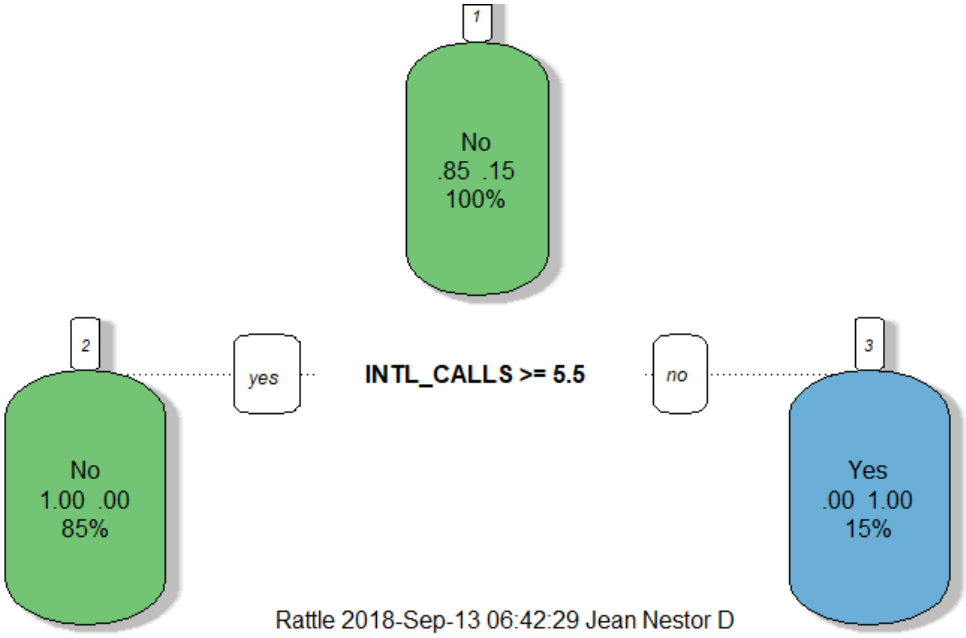


Figure 45 Related Decision Tree Model for the Classification Learning

- **Variable (Predictor) Importance**:

```
rpart variable importance

  only 20 most important variables shown (out of 25)

                                 Overall
INTL_CALLS                          1066
SUBSCRIBER_STATUS_201708Suspended   1064
INTL_CALLS_REVENUE                  1056
INTERNATIONAL_PLANY                 1054
CALL_REVENUE                        1030
CHARGEABLE_DURATION                    0
REGIONEasternCape                      0
GROUP_SERVICE_LINEPostpaid             0
TOTAL_OUTGOING_SMS                     0
GROUP_SERVICE_LINEPostpaidFTTH         0
CHARGEABLE_VOLUME                      0
TOTAL_VOICE_DURATION                   0
GROUP_SERVICE_LINEHybridBroadband      0
CHARGEABLE_UNITS                       0
REGIONWesternCape                      0
```

REGIONUnknown	0
REGIONGauteng	0
CALLS_TOTAL_NUMBER	0
TOTAL_DATA_VOLUME	0
GROUP_SERVICE_LINEPostpaidBroadband	0

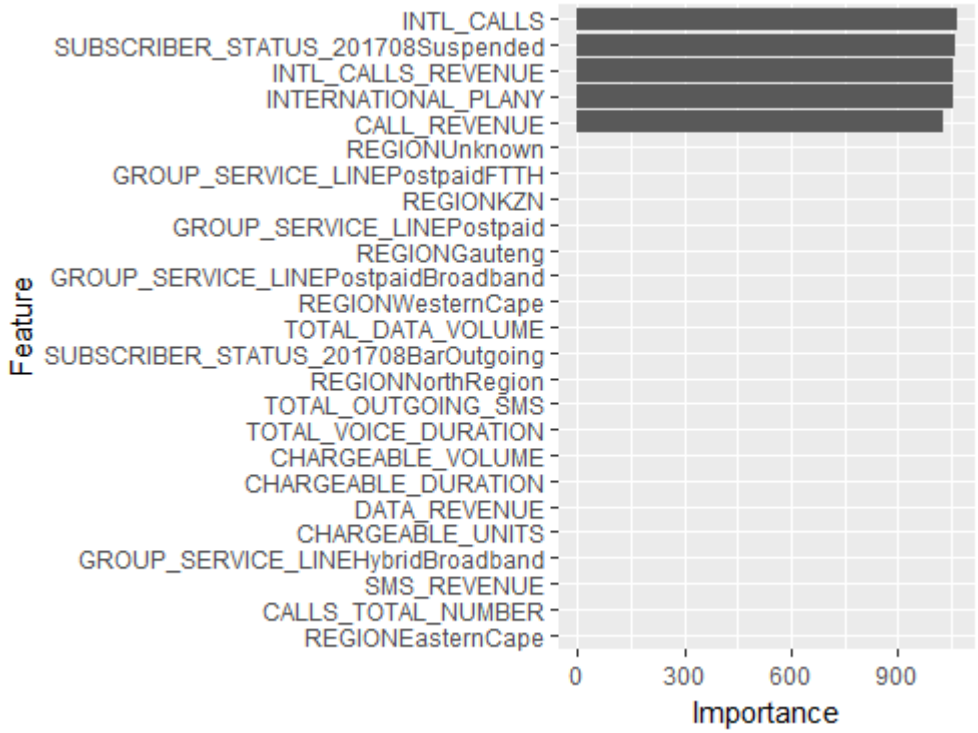


Figure 46 Classification Model Predictors Importance

6.1.4.2.2.Boosting Trees

Boosting Trees model is used to reduce bias and variance in most supervised Learning [55]. As one of the gradient boosting algorithms, it is based on weak learners, predictor variables which highly biased and lowly variant. Several open-source libraries are developed by researchers on a practical stand-point to tune and predict using boosting models [56]. The model is trained by fitting the training dataset, db_training into it using the "*gbm*" package. The result of the model fit is shown in below.

```
library(plyr)
model_bt=train (CHURN_FLAG~., method="gbm", data = db_training, verbose=FALSE)
model_bt
print(model_bt$finalModel)
```

```
Stochastic Gradient Boosting

4130 samples
  16 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 4130, 4130, 4130, 4130, 4130, 4130, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.9992092  0.9969574
  1                  100      0.9992092  0.9969574
  1                  150      0.9992092  0.9969571
  2                   50      0.9992092  0.9969574
  2                  100      0.9992882  0.9972590
  2                  150      0.9992614  0.9971592
  3                   50      0.9992619  0.9971586
  3                  100      0.9992882  0.9972587
  3                  150      0.9992614  0.9971592

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode'
 was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 100, interaction.depth = 2, shr
inkage = 0.1
 and n.minobsinnode = 10.

A gradient boosted model with bernoulli loss function.
100 iterations were performed.
There were 25 predictors of which 12 had non-zero influence.
```

The best achieved accuracy with boosting tree algorithm is 99.93%.

6.1.4.2.3. Random Forest

Using the training set db_training, the Random Forest algorithm is used to train and fit the model as below. The result of the model fit is shown below:

```
model_rf=train (CHURN_FLAG~., method="rf", data = db_training, prox=TRUE)
model_rf
print(model_rf$finalModel)
```

```
Random Forest

4130 samples
  16 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 4130, 4130, 4130, 4130, 4130, 4130, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9994166  0.9977154
  13    0.9996294  0.9985636
  25    0.9994972  0.9980459

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 13.


Call:
 randomForest(x = x, y = y, mtry = param$mtry, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 13

        OOB estimate of error rate: 0.02%
Confusion matrix:
      No Yes class.error
No  3500   0 0.000000000
Yes    1 629 0.001587302
```

The best achieved accuracy with Random Forest algorithm is 99.96%. The predictor importance is also shown below.

```
fitrf=varImp(model_rf_n, scale = FALSE)
fitrf
p=ggplot (data=fitrf, aes (x=weight, y=Predictor))
p+coord_flip ()
```

```
rf variable importance

  only 20 most important variables shown (out of 25)

                                    Overall
INTL_CALLS                        4.422e+02
SUBSCRIBER_STATUS_201708Suspended 3.903e+02
INTL_CALLS_REVENUE                1.383e+02
INTERNATIONAL_PLANY               6.577e+01
CALL_REVENUE                      2.090e+01
CHARGEABLE_DURATION               8.079e+00
CALLS_TOTAL_NUMBER                1.974e+00
DATA_REVENUE                      1.487e-01
TOTAL_VOICE_DURATION              1.228e-01
CHARGEABLE_UNITS                  1.009e-01
TOTAL_OUTGOING_SMS                9.172e-02
CHARGEABLE_VOLUME                 6.288e-02
TOTAL_DATA_VOLUME                 4.606e-02
REGIONWesternCape                 2.707e-02
SMS_REVENUE                       1.589e-02
GROUP_SERVICE_LINEPostpaid        5.767e-03
REGIONEasternCape                 5.000e-03
REGIONUnknown                     0.000e+00
REGIONNorthRegion                 0.000e+00
SUBSCRIBER_STATUS_201708BarOutgoing 0.000e+00
```

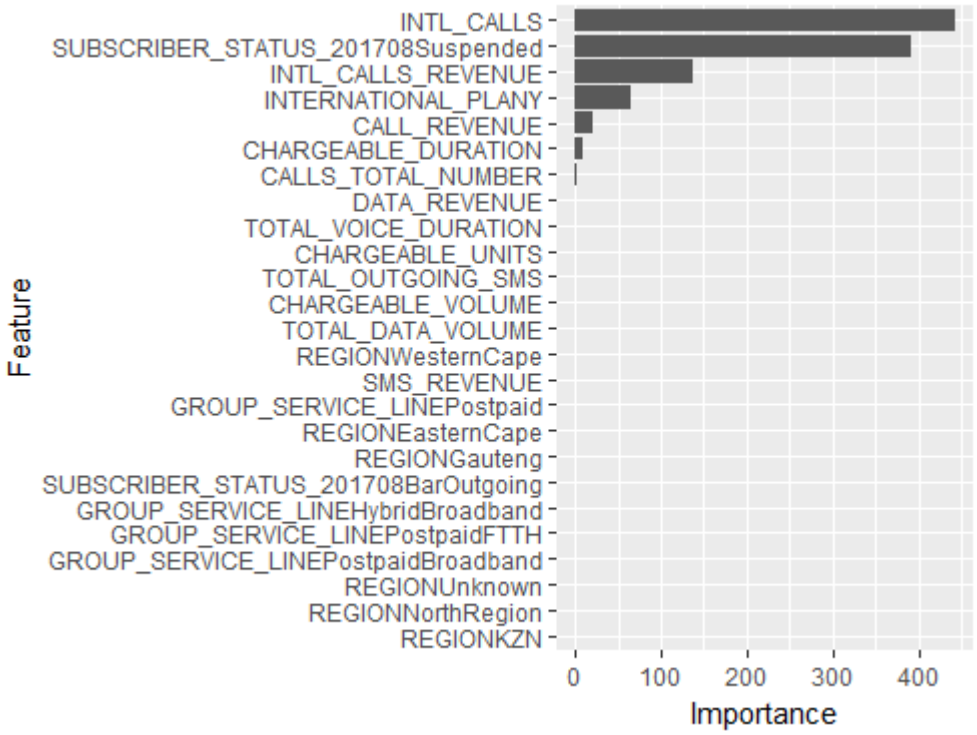


Figure 47 Random Forest Model's Predictor Importance

6.1.4.2.4. Neural Network

Neural Network is also used to train the CRM data. The Neuralnet package is used in R, to train a Neural Network. To avoid predictor influences, scaling is used for Neural Network. Neural Network is a black box supervised Machine Learning model, meaning that what happens in each Neuron is not controlled by the Data Scientist or Telecoms Engineer. As shown under data pre-processing, we use min-max normalization method.

The step to follow to train a Neural Network for a classification problem is not the same as training using regression-based methods, used above. The following processes are considered to design and train our model using Neural Network:

- Dataset Scaling: the CRM churn dataset is scaled to minimize the influence of large predictors in the design of the Neural Network. If unscaled data is used, this can lead to meaningless output. The min-max normalization is used to scale the churn dataset which also retain the distribution of predictors.
- Data Sampling and Partition: Using 70% of the scaled data as training set and 30% as testing set, using random sampling of data.

- Model Fitting: design and train the model, plot the designed Neural Network with 3 hidden layers.
- Prediction on the test data set and evaluation of the model

---

```
library(neuralnet)
dataframe_n_nn = db_frame_n[,-c(1,2,3)]
dataframe_n_nn$CHURN_FLAG= as.numeric(as.factor(dataframe_n_nn$CHURN_FLAG))
dataframe_n_nn$INTERNATIONAL_PLAN =
as.numeric(as.factor(dataframe_n_nn$INTERNATIONAL_PLAN))
maxim=apply (dataframe_n_nn, 2, max)
minim=apply (dataframe_n_nn, 2, min)
scaled_data = as.data.frame (scale (dataframe_n_nn, center = minim, scale = maxim - minim))
samplesize = 0.70 * nrow(dataframe_n_nn)
index = sample (seq_len (nrow (dataframe_n_nn) ), size = samplesize )
trainNN = scaled_data[index,]
testNN = scaled_data[-index,]
set.seed(332335)
model_nn_n =
neuralnet(CHURN_FLAG~INTL_CALLS+INTL_CALLS_REVENUE+CALLS_TOTAL_NUMBE
R+TOTAL_VOICE_DURATION+
CALL_REVENUE+CHARGEABLE_DURATION+TOTAL_OUTGOING_SMS+SMS_REVENUE
+

CHARGEABLE_UNITS+TOTAL_DATA_VOLUME+DATA_REVENUE+CHARGEABLE_VOL
UME, trainNN, hidden = 3, linear.output = T)
plot(model_nn_n)
```

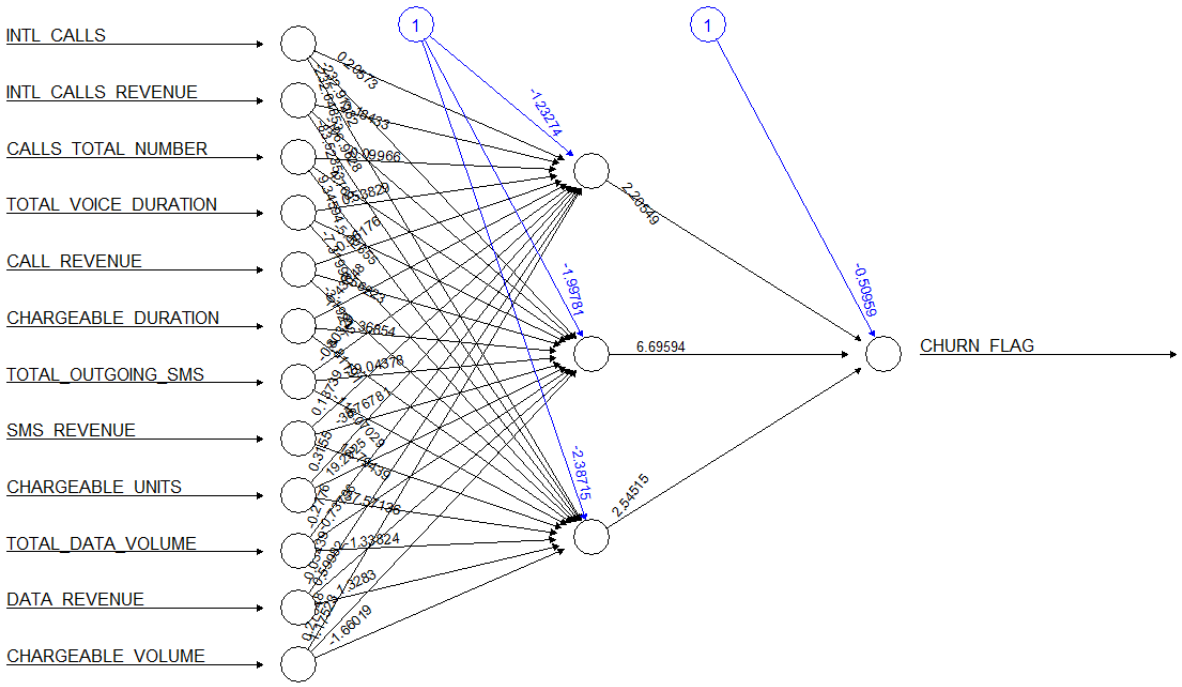


Figure 48 Neural Network Mode Fit with 3 Hidden Layer

model_nn_n2$result.matrix

error	3.18828523734
reached.threshold	0.00953340671
steps	2324.00000000000
Intercept.to.1layhid1	-1.23274015878
INTL_CALLS.to.1layhid1	0.20573026445
INTL_CALLS_REVENUE.to.1layhid1	0.18433486531
CALLS_TOTAL_NUMBER.to.1layhid1	-0.09966402207
TOTAL_VOICE_DURATION.to.1layhid1	0.53828838274
CALL_REVENUE.to.1layhid1	-0.06175948537
CHARGEABLE_DURATION.to.1layhid1	-0.43047895318
TOTAL_OUTGOING_SMS.to.1layhid1	-0.30328853095
SMS_REVENUE.to.1layhid1	0.13738919756
CHARGEABLE_UNITS.to.1layhid1	0.31549792634
TOTAL_DATA_VOLUME.to.1layhid1	-0.27759801202
DATA_REVENUE.to.1layhid1	-0.03439034885
CHARGEABLE_VOLUME.to.1layhid1	0.27348481265
Intercept.to.1layhid2	-1.99781462534
INTL_CALLS.to.1layhid2	-232.91231673990
INTL_CALLS_REVENUE.to.1layhid2	-106.95279700198
CALLS_TOTAL_NUMBER.to.1layhid2	4.76899607944
TOTAL_VOICE_DURATION.to.1layhid2	-5.12655006927
CALL_REVENUE.to.1layhid2	6.56223360337
CHARGEABLE_DURATION.to.1layhid2	2.36853982835
TOTAL_OUTGOING_SMS.to.1layhid2	-19.04378342432
SMS_REVENUE.to.1layhid2	-38.76780829932
CHARGEABLE_UNITS.to.1layhid2	19.29249638301
TOTAL_DATA_VOLUME.to.1layhid2	-0.73796133144
DATA_REVENUE.to.1layhid2	0.59981684830
CHARGEABLE_VOLUME.to.1layhid2	1.17522845090
Intercept.to.1layhid3	-2.38715365140
INTL_CALLS.to.1layhid3	-232.64853485130
INTL_CALLS_REVENUE.to.1layhid3	-83.52352692968
CALLS_TOTAL_NUMBER.to.1layhid3	9.34593716733
TOTAL_VOICE_DURATION.to.1layhid3	-7.31998098430
CALL_REVENUE.to.1layhid3	-3.19211681453
CHARGEABLE_DURATION.to.1layhid3	4.81190559382
TOTAL_OUTGOING_SMS.to.1layhid3	-116.07029398965
SMS_REVENUE.to.1layhid3	12.79439006307
CHARGEABLE_UNITS.to.1layhid3	37.57136005669
TOTAL_DATA_VOLUME.to.1layhid3	-1.33823986780
DATA_REVENUE.to.1layhid3	1.32829884508
CHARGEABLE_VOLUME.to.1layhid3	-1.66018594108
Intercept.to.CHURN_FLAG	-0.50958838347
1layhid.1.to.CHURN_FLAG	2.20549146538
1layhid.2.to.CHURN_FLAG	6.69593593680
1layhid.3.to.CHURN_FLAG	2.54515147748

To get the accuracy of the Neural Network trained above, we need to predict on the test data or the forecasting Churn network data.

6.1.5. Models' Evaluation

Four Machine Learning models have been used to train the dataset. In this section we review the performance based on the training dataset and compare the different model components.

Performance comparison:

By looking at the comparison table below, Random Forest achieved the best accuracy and Kappa for the regression-based algorithms. The second-best performing model is the Classification tree. Boosting Tree has a good accuracy over 99% but is the least of the 3. Random Forest along with Neural Network will be used to predict the test set.

```
results_model = resamples(list(Classification_tree=model_lm_n, Boosting_Trees=model_bt_n,
Random_Forest=model_rf_n))
summary(results_model)
dotplot(results_model)
```

```
Call:
summary.resamples(object = results_model)

Models: Classification_tree, Boosting_Trees, Random_Forest
Number of resamples: 25

Accuracy
                          Min.       1st Qu.       Median         Mean      3rd Qu. Max. NA's
Classification_tree 0.9986657772 0.9993324433 0.999343832 0.9994186027 0.9993552547    1    0
Boosting_Trees      0.9986495611 0.9986936643 0.999339498 0.9992882117 1.0000000000    1    0
Random_Forest       0.9986657772 0.9993390615 1.000000000 0.9996293871 1.0000000000    1    0

Kappa
                          Min.       1st Qu.       Median         Mean      3rd Qu. Max. NA's
Classification_tree 0.9945489125 0.9973798708 0.9974171403 0.9977340006 0.9975737822    1    0
Boosting_Trees      0.9948035925 0.9951671186 0.9973272053 0.9972589897 1.0000000000    1    0
Random_Forest       0.9945489125 0.9974299635 1.0000000000 0.9985636398 1.0000000000    1    0
```

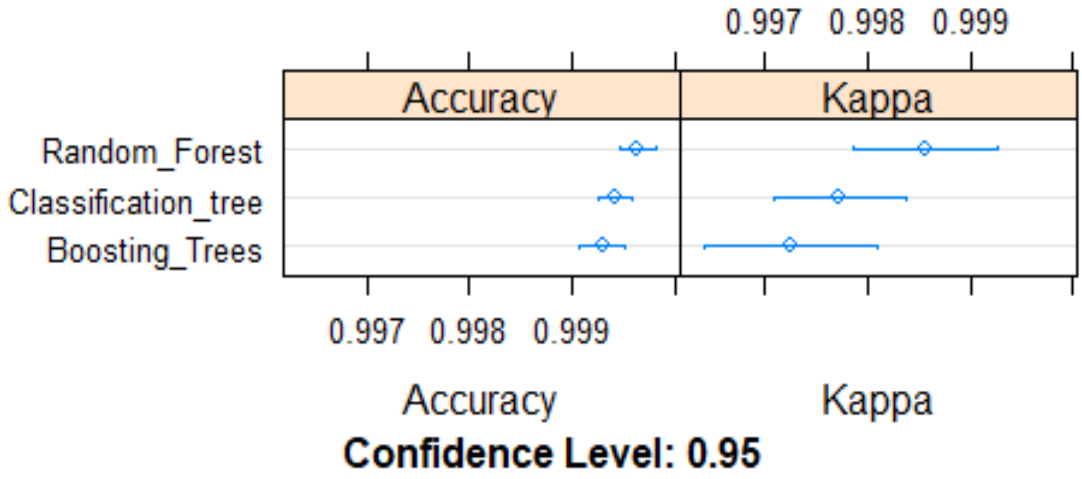


Figure 49 Model Performance Comparison

6.1.6. Predicting on the new Dataset and ROC Curves

Now that predictive models have been trained and accuracy on the training dataset has been evaluated, in the case of Neural Network, the Neural Network has been designed, the next step is used a completely new dataset to test the models and use the Confusion Matrix and ROC curve to check how far the predictive values scatter on the training and test datasets.

- **The Confusion Matrix**:

  The **confusion matrix** is an important indication of how the model is performing on a new dataset, it is a table which groups predictions based on whether the predicted value and the real value match or not. One table component provides the possible categories of values that have been predicted, and the other component provides the same for actual or real values. If the value predicted matches the actual or real value, then the prediction model gives an exact classification. The matrix is made of Positive and negative groups of values predicted. The exact predictions are named “True”. And the True can be either positive or Negative. The confusion Matrix is shown in the below table:

Table 9 Confusion Matrix of The Study

	Predicted Values		
Actual Values		No	Yes
	No	True Negative (Correctly predicted as Non-churners)	False Positive (Incorrectly predicted as Churners)
	Yes	False Negative (Incorrectly predicted as Non-churners)	True Positive (Correctly predicted as Churners)

The Accuracy and the sample error rate are computed using the Confusion Matrix by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.1.6.2)$$

The Sample Error is given by: $Err = 1 - A$

In which TP = True Positive, TN=True Negative, FP=False Positive and FN=False Negative.

- **Precision and Recall**:

  Computing precision and recall provides an indication of how the models are performing with emphasis on the relevance of the output results of the model. The impact of noise in the prediction function is determined by the two evaluators, Precision and Recall. The precision is the proportion of positive predicted values, and the Recall is the measure of precision of the result.

$$Pr = \frac{TP}{TP + FP} \quad (6.1.6.3)$$

$$Rec = \frac{TP}{TP + FN} \quad (6.1.6.4)$$

6.1.6.1. Predicting on the New Dataset

6.1.6.1.1. Prediction based on Classification Tree Model

The testing dataset is used by the classification tree model for the prediction of Churners. As seen in the output below, the Classification model has achieved a much better performance on the new dataset with an accuracy of 99.887 %, sensitivity of 100% and Specificity of 99.259%.

```
pred_lm=predict (model_lm_n, newdata = db_testing_n)
confusionMatrix (pred_lm, db_testing_n$CHURN_FLAG)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1499    2
       Yes    0  268

               Accuracy : 0.9988694
                 95% CI : (0.995922, 0.9998631)
    No Information Rate : 0.8473714
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9956159
 Mcnemar's Test P-Value : 0.4795001

            Sensitivity : 1.0000000
            Specificity : 0.9925926
         Pos Pred Value : 0.9986676
         Neg Pred Value : 1.0000000
             Prevalence : 0.8473714
         Detection Rate : 0.8473714
   Detection Prevalence : 0.8485020
      Balanced Accuracy : 0.9962963

       'Positive' Class : No
```

6.1.6.1.2. Prediction based on Gradient Boosting Trees Model

The testing dataset is used by the boosting trees model for the prediction of Churners. As seen in the output below, the boosting trees model has achieved a much better performance on the new dataset with an accuracy of 100 %, sensitivity of 100% and Specificity of 100%.

```
pred_gbm = predict (model_bt_n,newdata = db_testing_n)
confusionMatrix(pred_gbm, db_testing_n$CHURN_FLAG)
```

```
Reference
Prediction   No  Yes
       No  1499    0
       Yes    0  270

               Accuracy : 1
                 95% CI : (0.9979169, 1)
    No Information Rate : 0.8473714
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000000
            Specificity : 1.0000000
         Pos Pred Value : 1.0000000
         Neg Pred Value : 1.0000000
             Prevalence : 0.8473714
         Detection Rate : 0.8473714
   Detection Prevalence : 0.8473714
      Balanced Accuracy : 1.0000000

       'Positive' Class : No
```

6.1.6.1.3. Prediction based on Random Forest Model

The testing dataset is used by the Random Forest Model for the prediction of Churners. As seen in the output below, Random Forest model has achieved a much better performance on the new dataset with an accuracy of 100 %, sensitivity of 100% and Specificity of 100%; and the model has a higher 95% CI of 0.9979.

```
  Reference
Prediction   No  Yes
       No  1499    0
       Yes    0  270

               Accuracy : 1
                 95% CI : (0.9979169, 1)
    No Information Rate : 0.8473714
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000000
            Specificity : 1.0000000
         Pos Pred Value : 1.0000000
         Neg Pred Value : 1.0000000
             Prevalence : 0.8473714
         Detection Rate : 0.8473714
   Detection Prevalence : 0.8473714
      Balanced Accuracy : 1.0000000

       'Positive' Class : No
```

6.1.6.1.4. Prediction based on Neural Network

The testing dataset is used by the Neural Network constructed in the previous sections for the prediction of Churners. While confusion Matrix is used for classification models, Neural Network relies on the Mean Square Error to compute the performance based on the predicted values. The closer to zero the Mean Square Error is, the better the performance.

```
pred_nn = compute (model_nn_n2, testNN [, c(1:12)])
pred_NN2 = (pred_nn$net.result * (max(dataframe_n_nn$CHURN_FLAG) -
min(dataframe_n_nn$CHURN_FLAG))) + min(dataframe_n_nn$CHURN_FLAG)
realvalues = (testNN$CHURN_FLAG) *(max(dataframe_n_nn$CHURN_FLAG)-
min(dataframe_n_nn$CHURN_FLAG))+min(dataframe_n_nn$CHURN_FLAG)
plot(realvalues, pred_NN2, col='blue',main='Real values   vs  predicted values',pch=18,cex=0.7)
abline(0,1,lwd=2)
legend ('bottomright', legend='NN', pch=18,col='red',  bty='n')
MSE = sum ((realvalues - pred_NN2) ^2)/nrow(testNN)
MSE
```

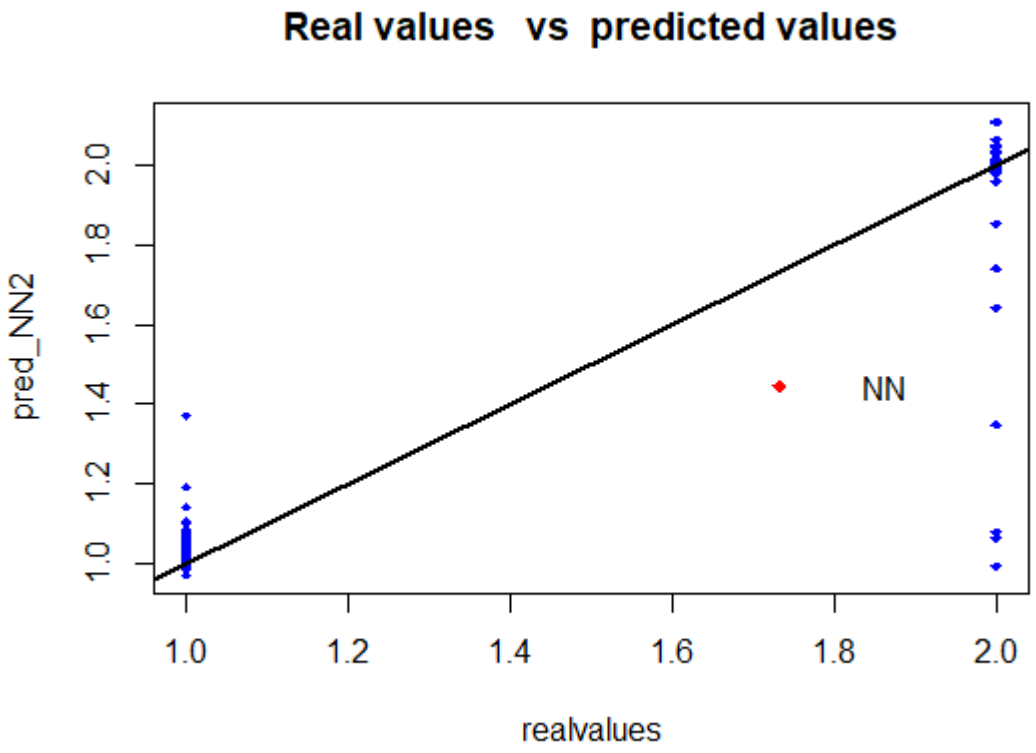


Figure 50 Neural Network Model Real values vs. Pred. Values

```
> MSE
[1] 0.002247233494
```

Checking the number of True Positives and negatives generated by the Neural Network Prediction. And this can be used to calculate the Accuracy of the model.

```
results = data.frame (actual = testNN$CHURN_FLAG, predicted = pred_nn$net.result)
results
result_round = sapply (results, round, digits=0)
result_round_df=data.frame (result_round)
attach(result_round_df)
table(result_round_df)
```

```
predicted
actual    0    1
     0 1498    0
     1    4  268
```

$$Acc = \frac{1498 + 268}{1498 + 268 + 4} = 99.78\ \%$$

We can see that the Neural Network predicts with an Accuracy of 99.78%. For this classification problem, Neural Network is the least performing algorithm. Random Forest is the highest performing algorithm with Boosting Trees, with 100% accuracy, then comes the Classification Tree with an accuracy of 99.88%.

6.1.6.2. The ROC Curve for the models

The ROC (Receiving Operating Characteristics) provides sensitivity against specificity. In the curve, we show the relationship between True Positive and False positive at different threshold levels. As seen in the below graph, the curve makes a rectangular form closer to the ideal values. The further the graph is from the 45 degrees line, the better the performance of the model. We are predicting binary value, TRUE or FALSE, 1 or 0. In order to plot the ROC curve, the predictor is converted from factor to numeric.

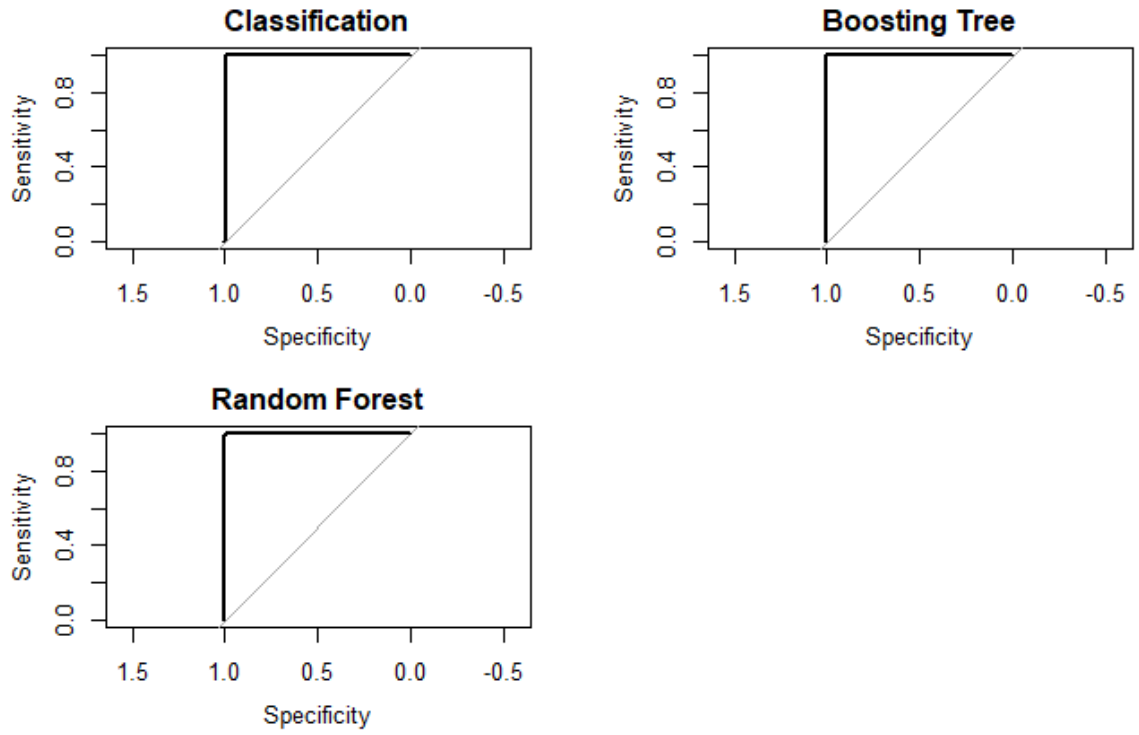


Figure 51 ROC Curve for the 3 Regression-based model

6.1.6.3. Predicting Churners using the best Performing Algorithm

Random Forest model has performed better than the other models, even though the performance gap is not huge. The model has achieved a precision and Recall of 1, an excellent result of the model. In summary, the model has had the best performance sampling error, Good accuracy on predicting new dataset, with the precision and recall of 1. In this section we now predict the customer who are likely to churn.

```
churners=data.frame(db_testing_n,pred_rf)
churners_n = sqldf("select GROUP_SERVICE_LINE, REGION,INTERNATIONAL_PLAN, pred_rf
                   from churners where pred_rf = 'Yes';")
summary (churners_n)
```

```
GROUP_SERVICE_LINE     REGION           INTERNATIONAL_PLAN pred_rf
 Length:270          Length:270          Length:270         No :  0
 Class :character    Class :character    Class :character   Yes:270
 Mode  :character    Mode  :character    Mode  :character
```

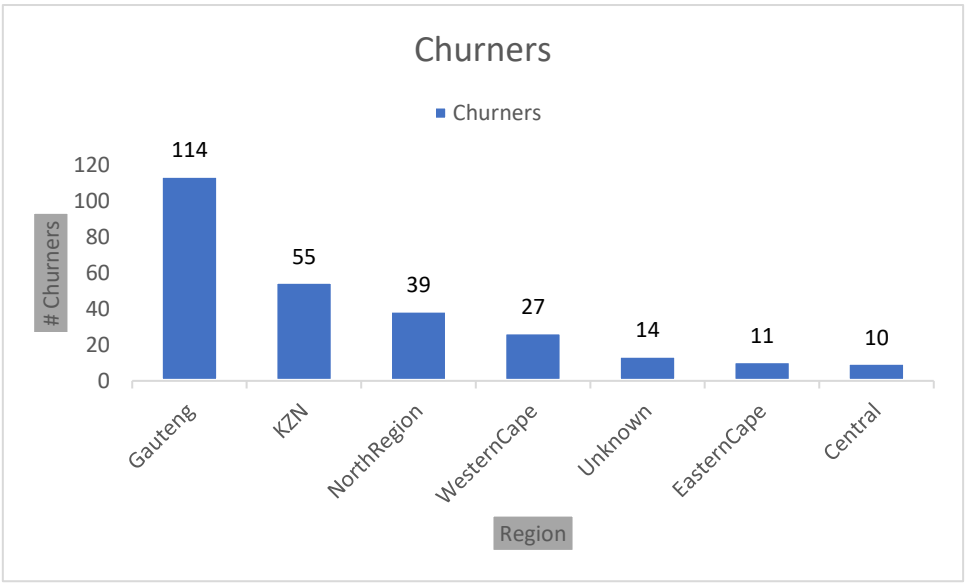


Figure 52 Number of Churners per Region

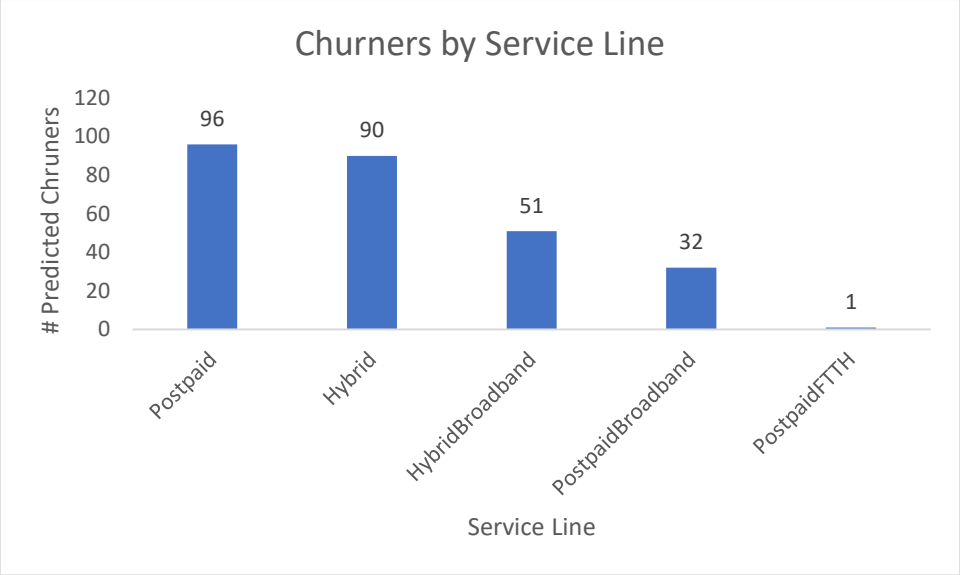


Figure 53 Number of Predicted Churners by Service Line

From the prediction above, we can see that 270 customers are likely to terminate their contracts and leave the network. The Analytics also shows the repartition of churners by provinces and by Service line.

6.2. Low Cost SQM Tree Model Implementation

6.2.1. Data Collection, Preparation and Pre-processing

In this section, data connection to the database, data understanding, and data pre-processing are done. The requirements here are that: Apache Spark Stand-Alone is installed in the processing server and or R framework. The SQL queries in this section on the ContextSQL from Spark for SQL users.

Data Connection and Data Understanding

The information exchange from the Customer database to the processing server is done through transport and application protocols which carry packets. TCP (Transmission Control Protocol) is used as the transport protocol for connection-oriented packets. TCP applications are very sensitive to the conditions of the communication pipes (bandwidth requirements). The selected performance metrics used for the SQM are as follow:

Table 10 Performance Metrics Selected for the Data SQM

Metric	Description
TCP_DL	Downlink Delay
TCP_UL	Uplink Delay
ActiveSession_DL	Active session time in the Downlink
ActiveSession_UL	Active session time in the Uplink
Bytes_DL	Volume of data used in the Downlink
Bytes_UL	Volume of data used in the Uplink
Bytes_Retransmitted_DL	Volume of data retransmitted in the Downlink
Bytes_Retransmitted_UL	Volume of Data retransmitted in the Uplink
DNS_Query_OK	Amount of Domain Name Resolution Queries Success.
DNS_Query_NOK	Amount of Domain Name Resolution queries failure.

Apart from the metrics collected from the customer database, the SQM Model uses a set of categorical parameters to aggregate the information. Data is aggregated using the following parameters which also form part of the collected dataset:

Table 11 Categorical Aggregation Metrics for the SQM system

Aggregation Level	Description
Service	Protocol application, and the main layer of the SQM
Network	Network Element (cell) used.
Device	IMEI converted to Device Manufacturer, model and capability.
IMSI	Single Customer Information.

6.2.2. Data Caching, Processing and Adaptation

Using the theorical and mathematical background detailed in section 4.2, the below KPIs and SQIs are computed to optimize the tree model. The data structure is adapted in a format which is understandable to the processing server or Mediation layer of the processing server. All the data is queried from the Memory of the server.

```
round (100*sum (DNS_OK)/ (sum (DNS_OK) + sum (DNS_ERROR)),2) as DNS_Performance,
round (sum (BYTESUL)+sum (BYTESDL)/1024/1024/1024,2) as DataVolume,
round (100*sum (BYTE_RTX_DL)/sum (BYTESDL),2) as DL_RTX,
round (0.001*sum (BYTESDL)/sum (ACTSEC_DL),2) as Throughput,
round (2*(avg (TCP_NW) +avg (TCP_MS))) as Latency
```

Table 12 SQM Model Computed KPIs

KPI	Description
DNS_Performance	Domain Name Resolution Success Rate.
DL_RTX	Downlink Retransmission Rate/Packet Loss
Data Speed /Throughput	Download Data speed
Latency	Packet or traffic latency.

SQIs are computed by aggregating and summing the KPIs together. The global SQI is the sum of KQIs or application level SQIs. The below tables show the algorithm for both the Global and local SQIs.

- Global SQI:

```
SELECT (((SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.video_str_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.audio_streaming_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.binary_download_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.facebook_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.https_secured_browsing_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.http_layer_1_index) + (SELECT 100-
(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM  soc_development.instant_messaging_layer_1_index)
+
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.itunes_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.other_services_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.p2p_torrents_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.skype_layer_1_index) +
(SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM soc_development.twitter_layer_1_index))/12) AS
GLOBAL_SQI
```

- Local SQI

```
SELECT 100-(0.25*(RTX_I+DNS_I+RTT_I+THPUT_I)) AS SQI FROM
"Application_Protocol"_layer_1_index
```

From the algorithm, we can see the algorithm used as an aggregator for the SQI approach. Note that in this case, variables α and β are identical and equal to $\alpha = \beta = 0.25$

6.2.3. Visualization of the Output Results

Data Visualization is an important part of Business Intelligence, as it is explored by different industrial domains. It is important that the visualization be built on the idea to display simultaneous multiple views of the Network Service insight data in the form of graphs and charts, allowing interactions and analysis between the different graphs, dashboards [57]. Several elements of reporting are put together to give a deep dive in data performance in a unified view to support decision making. Mathematical models or algorithms are applied to the BI system to minimize calculation of inherited files such as percentages and totals, that can be graphically represented by Charts, Gauges, and Maps. Taking into consideration the objective to keep the system cost low, in other word the cost effectiveness, the visualization of the result is done in a simple and concise way. From a high level, Layer 0 which is the Global view, the Communication Service Provider can drilldown to Layer 1, and from Layer 1 to Layer 2, Layer

2 to Layer 3 for more insight on the information. The generated SQM Dashboard is shown in the below sections.

6.2.3.1. Layer 1 Use Case: Overall Network Service Quality

To engage to Optimization, troubleshooting and improvement of the Network, it is crucial to have a global analysis of the network performance, comparing to an overall threshold set by the Operator. The advantage of the use case Layer 1 is to pinpoint underperforming services or services with poor SQI (Service Quality Index) as illustrated in the below figure. With a gauge chart chosen as the representation graph, services with SQI less than a certain threshold can be identified at a glance. The following services are analysed: Facebook, http_web browsing, audio-streaming, P2P applications, iTunes, Binary downloads, Instant messaging, skype, twitter and video-streaming. The graph also shows the Data Volume usage per protocol application in Gigabytes.

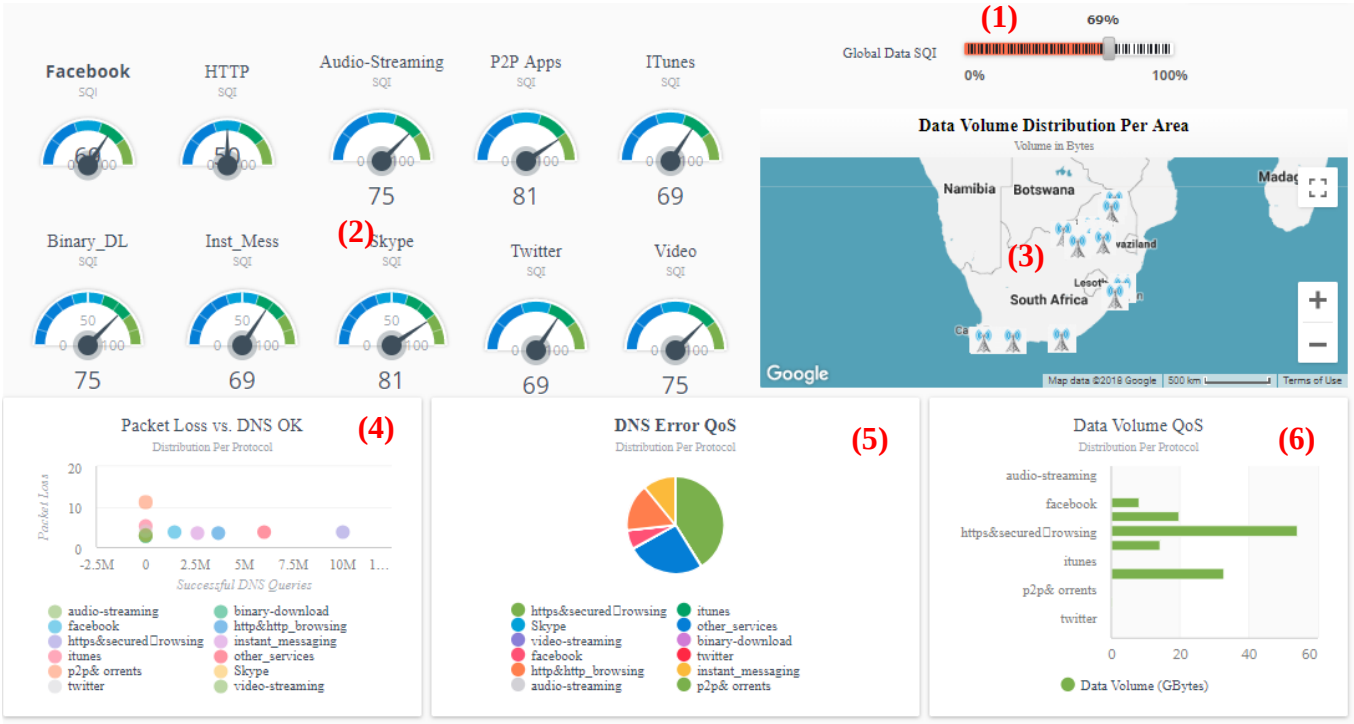


Figure 54 SQM Model Layer 1 Use Case Illustration

(1): Global Data SQI: for the internal Network, the Service Index is 69%

(2) Individual SQI or KQI for each service: It can be seen from the graph that HTTP protocol applications have an SQI of 50% which is very low. That specific service needs to be looked upon.

(3) GIS view of Performance Data: one of the most efficient way to represent data, GIS shows a lot on performance based on Geographical aggregation.

(4) Packet Loss vs. DNS OK: Packet loss (rate) vs. the number of Domain Name Server successful queries.

(5) DNS Error QoS: Failure on the DNS performance. The DNS server's can be easily analyzed to see if the issue is global or affect only a specific service.

(6) Data Volume: Usage of real data in Giga Bytes, per service.

6.2.3.2. Layer 2 Use Case: Specific Application Service Performance

Layer 2 Use case provides a specific service performance as illustrated on the below figure. A selected service application's, in this case Audio-Streaming quality is analysed on a single dashboard. The path to Layer 2 is a click from Layer-1 service gauge. The dashboard displays:

(1): The overall throughput or data speed in Kilobits per second: Average data speed of 677.35 Kbps.

(2), (3) The latency on the Downlink and Uplink

(4) The DNS Statistics Information: how many DNS queries failed or were successful.

(5) Packet loss of the audio-streaming application, in this case below 5%, ~ 3%.

(6) DNS Failure rate: 1%.

(7) A summary table of Performance aggregated based on Cells, Handsets and Subscribers. This allows to track which dimension of the Network affects the Performance.

Based on the data analysis, the worst Key Performance Indicator (KPI) can be evaluated quickly to take optimization decision. The Operator can benchmark the QoS value against the target QoS.

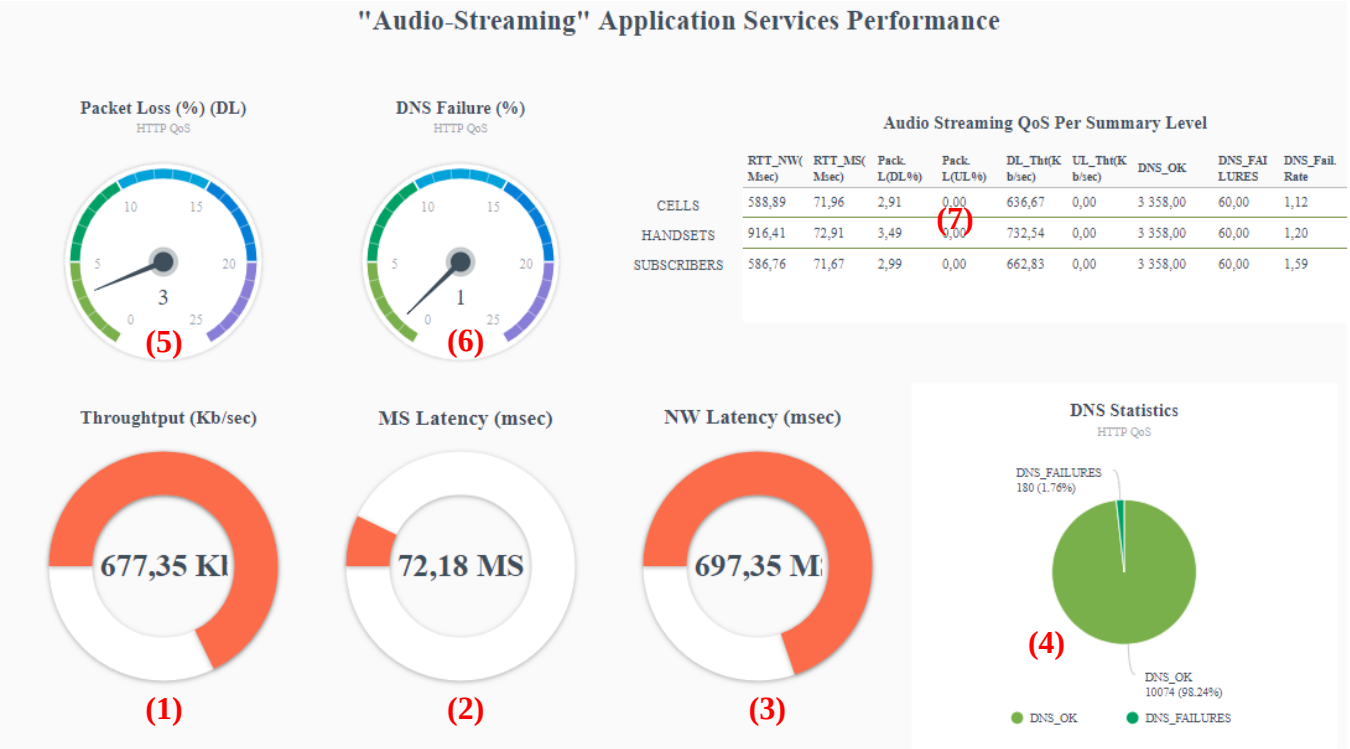


Figure 55 SQM Model Layer 2 Use case Illustration

6.2.3.3. Layer 3 Use Case: Specific Aggregation Performance Impact of a specified Service

The SQM Model Layer 3 use case illustrates the performance of the selected service application's (Audio streaming application) impact on the Network, showing performance per Network cells. It must be noted that the choice of the aggregation from the previous layer could be on Device/Handsets or Customer. Thus, handsets or customer aggregation levels should be selectable to have a multi-layer aggregation view. The below figure illustrates the visualization of such a use case. Using this layer, the worst and top network elements in terms of performance can be identified, evaluated against the number of impacted customers for priority in optimization processes.

(1): Packet Loss per cell: Top cells with higher Packet loss rate.

(2): QoS Data Activities: Cells with highest Retransmission Bytes comparing to Data Volume

(3): QoS DNS vs. Packet Delay: Domain Name Server performance against the Network delay, latency per cell.

(4): Map Distribution of Cells: GIS view of cells coordinates across the country.

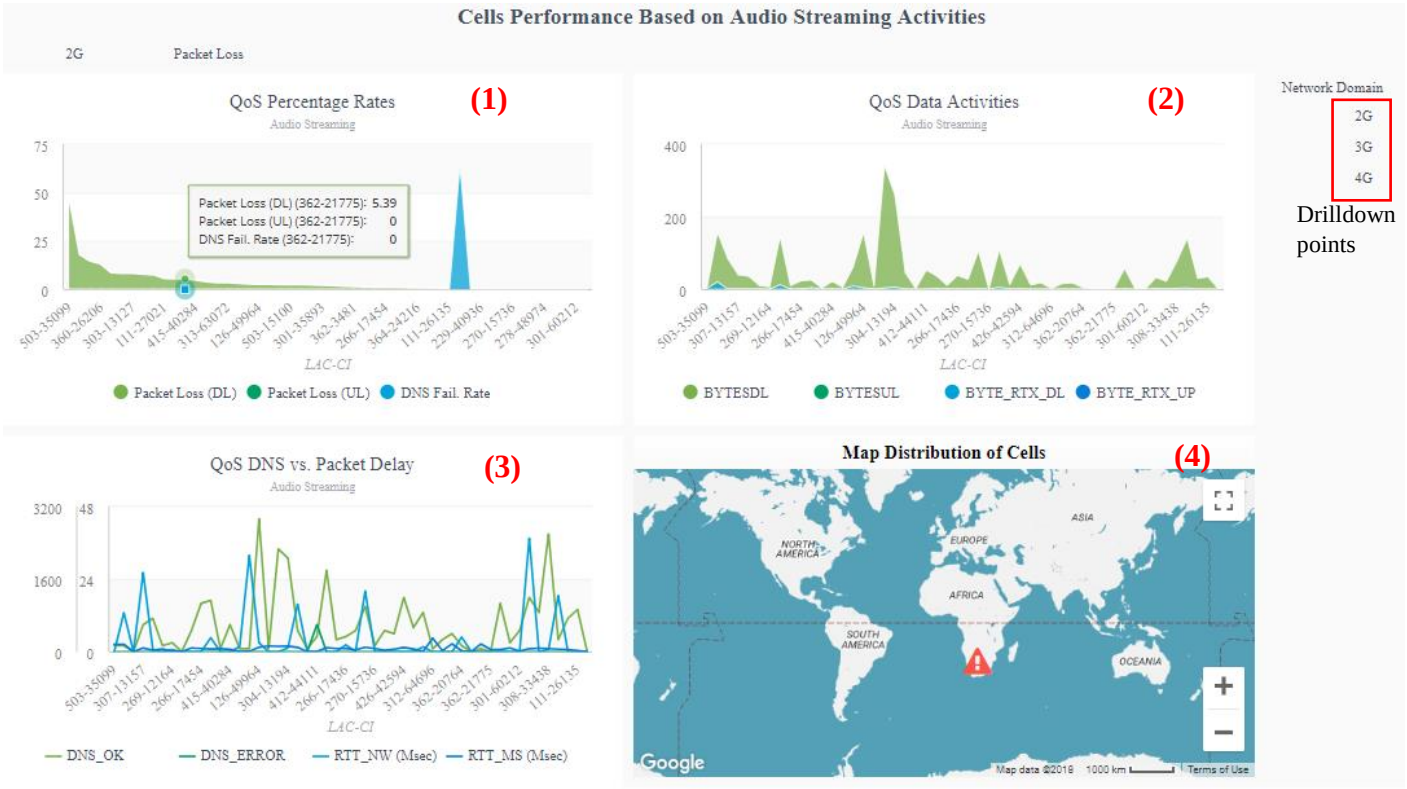
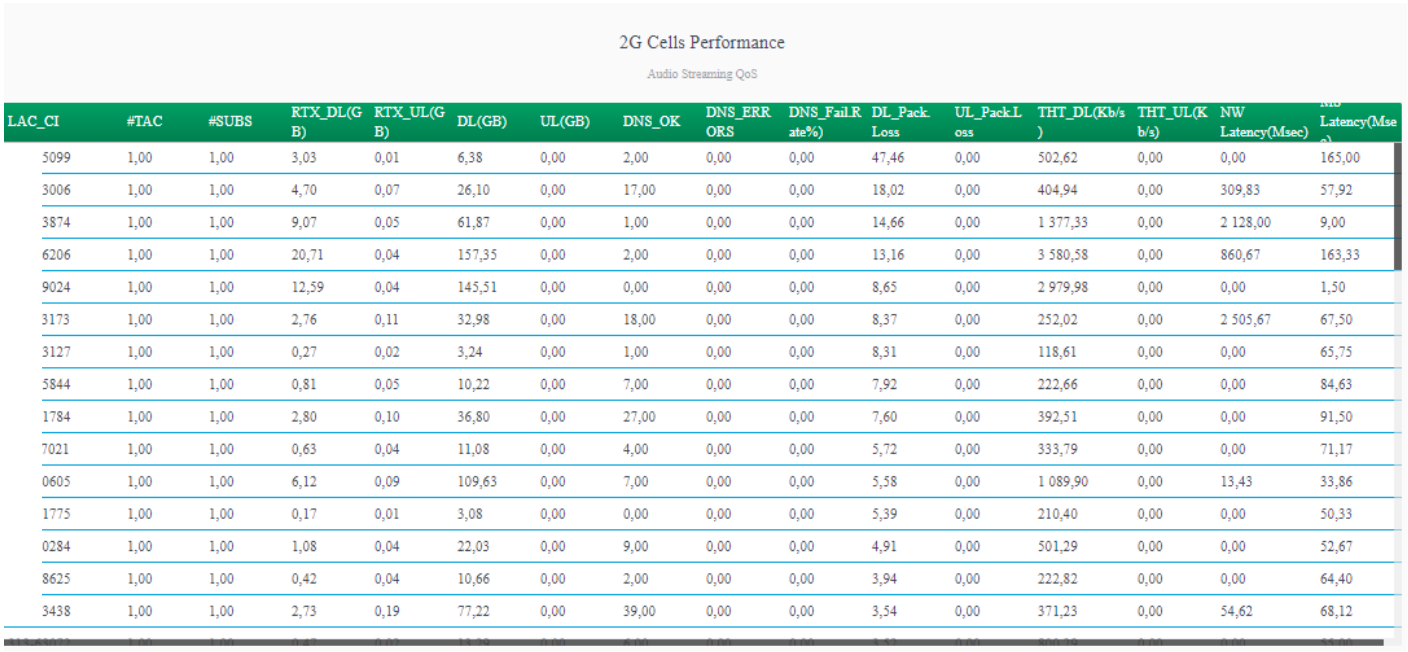


Figure 56 SQM Model Layer 3 Use Case Illustration

6.2.3.4. Layer 4: Technology based Performance based on a specific Service

The SQM Model Layer 4, considered the lowest element of the model is analysed on the selected service application, Audio streaming service application. The illustration is shown in a tabular format representation, displaying for each network cell, representing a geographical region, the number of subscribers seen in the network using Audio Streaming, the number of devices used for audio streaming, and the relevant KPIs. The representation of such a layer use case is illustration in the below figure. The dashboard can be exported to excel or a pdf format for further actions.



2G Cells Performance

Audio Streaming QoS

LAC_CI	#TAC	#SUBS	RTX_DL(GB)	RTX_UL(GB)	DL(GB)	UL(GB)	DNS_OK	DNS_ERRORS	DNS_Fail.Rate%)	DL_Pack. Loss	UL_Pack.Loss	THT_DL(Kb/s)	THT_UL(Kb/s)	NW Latency(Msec)	MS Latency(Msec)
5099	1,00	1,00	3,03	0,01	6,38	0,00	2,00	0,00	0,00	47,46	0,00	502,62	0,00	0,00	165,00
3006	1,00	1,00	4,70	0,07	26,10	0,00	17,00	0,00	0,00	18,02	0,00	404,94	0,00	309,83	57,92
3874	1,00	1,00	9,07	0,05	61,87	0,00	1,00	0,00	0,00	14,66	0,00	1 377,33	0,00	2 128,00	9,00
6206	1,00	1,00	20,71	0,04	157,35	0,00	2,00	0,00	0,00	13,16	0,00	3 580,58	0,00	860,67	163,33
9024	1,00	1,00	12,59	0,04	145,51	0,00	0,00	0,00	0,00	8,65	0,00	2 979,98	0,00	0,00	1,50
3173	1,00	1,00	2,76	0,11	32,98	0,00	18,00	0,00	0,00	8,37	0,00	252,02	0,00	2 505,67	67,50
3127	1,00	1,00	0,27	0,02	3,24	0,00	1,00	0,00	0,00	8,31	0,00	118,61	0,00	0,00	65,75
5844	1,00	1,00	0,81	0,05	10,22	0,00	7,00	0,00	0,00	7,92	0,00	222,66	0,00	0,00	84,63
1784	1,00	1,00	2,80	0,10	36,80	0,00	27,00	0,00	0,00	7,60	0,00	392,51	0,00	0,00	91,50
7021	1,00	1,00	0,63	0,04	11,08	0,00	4,00	0,00	0,00	5,72	0,00	333,79	0,00	0,00	71,17
0605	1,00	1,00	6,12	0,09	109,63	0,00	7,00	0,00	0,00	5,58	0,00	1 089,90	0,00	13,43	33,86
1775	1,00	1,00	0,17	0,01	3,08	0,00	0,00	0,00	0,00	5,39	0,00	210,40	0,00	0,00	50,33
0284	1,00	1,00	1,08	0,04	22,03	0,00	9,00	0,00	0,00	4,91	0,00	501,29	0,00	0,00	52,67
8625	1,00	1,00	0,42	0,04	10,66	0,00	2,00	0,00	0,00	3,94	0,00	222,82	0,00	0,00	64,40
3438	1,00	1,00	2,73	0,19	77,22	0,00	39,00	0,00	0,00	3,54	0,00	371,23	0,00	54,62	68,12

Figure 57 SQM Model Layer 4: Use Case Illustration

KPIs are show on the above Figure for each cell. However, one of the important aspects is the addition of the impacted number of subscribers and number of devices.

CHAPTER 7. CONCLUSION

7.1. Conclusion

We have shown in this study that Telecommunications Environment is not spared by the virus of Data Analysis, Predictive Analytics and Machine Learning. Building the next generation of Customer Experience Management (CEM), Service Quality Management (SQM) and Network Performance Management (NPM) requires more than just Telecommunications knowledge. With the high adoption of Smart Phones, the arrival of 5G and the prospect of connected devices and Internet of Things (IoT), Communication Service Providers (CSPs) must adopt flexible strategies in managing billions of generated transactions and providing capabilities such as:

- Customer profiling and individual recommendations.
- Prediction of potential Network Failures and break-ups.
- Prediction of Service adoption, revenue loss, fraud detection and abnormal traffic pattern Analysis.
- Advanced Roaming Analysis.

Data and Predictive Analytics, when applied to Telcos environment, should support decision making, notably at strategic level, using various and diverse indicators, both qualitative and quantitative to improve business performance and reach the defined objectives.

Extended flow of Data from various sources, in various formats, structured and unstructured in today's era drives the global economy [58]; hence, Data Analysis and Predictive Analytics take a high position on the spot. A fundamental point in Data Analysis and Machine Learning is the development and choice of algorithms for building models, selecting the right variables and training different models [59]. However, executing predictive studies is more than just selecting an algorithm for prediction. A detailed understanding of data to be used (expertise in the area) and variables is mandatory, mostly when dealing with supervised Machine Learning. The choice, performance and evaluation of Machine Learning algorithm depends on several components where in our case, the accuracy, the sampling error, the precision and the recall have been considered.

The Out-of-sample error, on the new data set is an indicator that counts most. On the new dataset, Random Forest provides an out-of-sample error of "0". The study shows that Telecommunications Operators can benefit from the advancement in technology.

The study has also proposed a simple but efficient, and low cost in implementation Service Quality Management (SQM) system for Communication Service Providers using new technologies (Fast processing and Business Intelligence). The path towards this model is motivated by the perpetual high investment from CSPs on plug-in tools. The approach in this study to leverage on the CSP's skills of adopting new technologies, transferring a certain level of management and control to the Operator. With the emerging of services such as Over-The-Top (OTT) applications, Online Streaming, applications download, combined with the boost in Mobile devices usage, create a driving force for CSP business models. To survive the competitive market, CSPs need transformation of customer and Network data to information that can be used to make intelligent, cost effective business decisions.

Different Network Operators will take different values from the network data depending on two factors:

- Business transformation, cost reduction by increasing efficiency in processes.
- Faster decision making based on real-time Network data information.

The SQM model studied in this research is a pure combination of two disciplines: Telecommunications and Data Science. The model is an introduction to an in-house SQM development that can help CSPs in reducing the cost on Network tools.

The proposals to Telecommunications Business transformation consist in leveraging on the expertise where necessary, on the technology in a competitive and cost effective way, making data a priority and accessible to the entire Organization.

7.2. Recommendations and Future Studies

7.2.1. QoS Data Prediction and Unification

Data Analytics, Big Data and Machine Learning are growing rapidly. Many efforts are being put in to them. Telecommunications area with the deployment, launching of 5G and the ascent of the Internet of Things (IoT), is also investing in Advance Analytics. CSPs will have to adapt their current infrastructure to embrace new technologies. The process involves integrating Predictive Analytics and Machine Learning in to their business strategy to stay ahead of competition.

To provide robust Machine Learning and Predictive Analytics, more data and data sources will be required. The CRM database accommodates only customer related information. The scope needs to be expanded to contain multiple cases. In the future, Telecommunications predictive Studies must integrate all different sources of CSP data including QoS (Quality of Service) data. And advanced Machine Learning Algorithms such as Neural Networks, Support Vector Machines (SVM) which provide a black box kind of analysis, an improved performance on accuracy and learning, comparing to other traditional models need to be considered also.

7.2.2. Recommender Systems, Service and Customer Auto-Profiling

Integrating Predictive Analytics and Advance Analytics in Telecommunications will also help CSPs in building Intelligent systems, capable of adapting to the type of services, customers, and network elements. Once the system can understand the internal processes (through Artificial Intelligence), functionalities such as SQM (Service Quality Management), CEM (Customer Experience Management), NFD (Network Fraud Detection) become automated.

- A typical scenario would be a system that can learn in advance and recommend a certain service package to a specific customer.
- Another scenario would be a system or Network which learn in advance and recommend a change of cells configuration due to a predicted increase in capacity or usage.

Many use cases will be set up, should CSPs integrate Artificial Intelligence in to the business strategy.

7.2.3. IoT and Device Performance Analytics

With the rise of Data Analytics, the increase on devices, the ascent in generated data and the need to deploy intelligently connected network of things to improve the human ecosystem, open new windows to change and flexibility of the actual Network Infrastructures. Areas such as IoT, Machine to Machine Communication need to also take advantage of Artificial Intelligence to draw efficient business values. Discuss IoT and MTC (Machine Type Communication) should also be in the business strategy of CSPs to fight competition. Hence, the importance of Data Analytics for Device and IoT transactions.

7.2.4. Telecommunications Cloud Solution

On top of SQM and CEM, the big data superpower on data processing, Business Intelligence and Predictive capabilities will take Customer Service Providers to another level. The Operators are in search of effective SQM and CEM systems which takes into consideration not only the Analytics and Prediction, but also the needs for speed in data retrieval, which plays a huge role on market competition. An area of future researches in the design and implementation of SQM, CEM and related systems, is the adaptation of data computing environment to tackle both speed, security and infrastructure management, area such as [60]:

- Cloud Computing
- Grid Computing
- In-Memory computing for which some features have been used in this paper

REFERENCES

[1] D. M. J. Nestor and K. A. Ogudo, "Practical Implementation of Machine Learning and Predictive Analytics in Cellular Network Transactions in Real Time," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD),* pp. 1-10, 2018.

[2] Ericsson, "More than 50 Billion Connected Devices," February 2011. [Online]. Available: http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf.

[3] R. Agrawal and J. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering,* no. 8, p. 962–969, 1998.

[4] A. Azzalini and B. Scarpa, *Data Analysis & Data Mining, and Introduction,* London: Oxford University Press, 2013, pp. 2-200.

[5] K. A. Ogudo and D. M. J. Nestor, "Modeling of an Efficient Low Cost, Tree Based Data Service Quality Management for Mobile Operators Using in-Memory Big Data Processing and Business Intelligence use Cases," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD),* pp. 1-8, 2018.

[6] P. Staffan Fredricsson and C. Perey, "Quality of Service for Multimedia Communication," Technology Futures, inc, 1995.

[7] F. C. d. Gouveia and M. Thomas, "QUALITY OF SERVICE IN TELECOMMUNICATION," in *Telecommunication Systems and Technologgies,* P. Bellavista, Ed., Oxford, United Kingdom, Encyclopedia of Life Support Systems (EoLSS) Publisher, 2009, pp. 77-79.

[8] H. D.J., H. Mannila and P. Smyth, "Principles of Data Mining," *Cambrige, Mass.: MIT Press,* 2001.

[9] J. Betser and D. Belanger, "Architecting the Entreprise via Big Data Analytics," in *Big Data and Business Analytics,* J. Liebowitz, Ed., London, CRC Press, Taylor & Francis Group, 2013, pp. 1-20.

[10] Statista, "Smartphone users in South Africa 2014-2022," 2018. [Online]. Available: https://www.statista.com/statistics/488376/forecast-of-smartphone-users-in-south-africa/. [Accessed 25 May 2018].

[11] A. Aspin, "SQL Server Reporting Services as a Business Intelligence Platform," in *Business Intelligence with SQL Server Reporting Services,* Apress, Ed., New York, Springer Science, 2015, pp. 21- 41.

[12] R Development Core team, "R: A language and environment for statistical computing," 2008. [Online]. Available: http://www.R-project.org. [Accessed 15 October 2017].

[13] M. Zaharia, R. S. Xin, P. Wendell and al, "Apache Spark: a unified engine," *Communications of the Acm,* vol. 59(11), pp. 56-65, 2016.

[14] H. Luo and M.-L. Shyu, “Quality of service provision in mobile multimedia,” Human-centric Computing and Information Sciences, Fort Wayne, 2011.

[15] R. Rodríguez, D. Fernández, H. Montes, S. Hierrezuelo and G. Gómez, “Quality of Service Mechansms,” in *End-to-End Quality of Service over Cellular Networks*, R. G. Gomez, Ed., John Wiley & Sons,, 2005, pp. 103-137.

[16] A. (. J.D. Power, “Wireless network quality assessment study™,” 11 November 2004. [Online]. Available: http://www.jdpower.com/studies/pressrelease.asp?StudyDD=891.

[17] W.C. Hardy, ““QoS” Measurement and Evaluation of Telecommunications Quality of Service,” John Wiley & Sons. Ltd, 2001.

[18] D. Soldani, D. Chiavelli, J. Laiho, M. Li, N. Muhammad, G. Giambiasi and C. Rodriquez, “QoE and QoS Monitoring,” in *QoS and QoE Management in UMTS Cellular Systems*, M. L. a. R. C. David Soldani, Ed., John Wiley & Sons, Ltd, 2006, pp. 315-383.

[19] D. Soldani, “QoS management in UMTS terrestrial radio access FDD networks, dissertation for the degree of Doctor of Science in Technology (Doctor of Philosophy),” Helsinki, 2005.

[20] C. D. L. Daniel T. Larose, Data Mining and Predictive Analytics, First, ed., New Jersey: Inc. John Wiley & Sons, 2015.

[21] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinart, C. Shearer and R. Wirth, “CRISP-DM Step-by-Step Data Mining Guide,” 2000.

[22] J. Dean and S. Ghemawat, “MapReduce: a flexible data Processing Tool,” *Communications of the ACM*, vol. 53, no. 1, pp. 72-77, 2010.

[23] K. AZIZ, D. ZAIDOUNI and M. BELLAFKIH, “Real-Time Data Analysis Using Spark and Hadoop,” *IEEE*, pp. 1-6, 2018.

[24] P. Hasso and Z. Alexander, “In-Memory Data Management: Technology and Applications,” *Springer Science & Business Media*, 2012.

[25] H. Zhang, G. Chen, B. C. Ooi, K.-L. Tan and M. Zhang, “In-Memory Big Data Management and Processing: A Survey,” *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 27, no. 7, 2015.

[26] H. Karan, A. Konwinski, P. Wendell and M. Zaharia, “Spark SQL,” in *Learning Spark Lightining Fast Data Analysis*, Ed. By O'Reilly, 2015, pp. 161-183.

[27] O. Z. S.A. Jacob, “Evaluating the Scaling of Graph–Algorithms for Big Data using GraphX,” *IEEE 2nd International Conference on Open and Big Data*, pp. 1-18, 2016.

[28] R. Diestel, Graph Theory, 4th ed., Heidelberg: Springer, 2010.

[29] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser and G. Czajkowski, “Pregel: A system for large-scale graph processing,” in *Proc. of the 2010 ACM SIGMOD Intl. Conf. on Management of Data*, SIGMOD’10., Ed., New York, NY, USA: ACM, 2010, pp. 135-146.

[30] L.G. Valiant, “A bridging model for parallel computation,” *Commun. ACM,* vol. 33, no. 8, pp. 103-111, 1990.

[31] J. Scott, “Social network analysis,” *Sage,* 2012.

[32] Brian Caffo, “Regression Models for Data Science in R,” 2015.

[33] F. Galton, “Regression towards mediocrity in hereditary stature,” Anthropological Miscellanea, 19th, Century.

[34] X. Feng, Y. Zhou, T. Hua, Y. Zou and J. Xiao, “Contact Temperature Prediction of High Voltage Switchgear Based on Multiple Linear Regression Model,” *IEEE International,* pp. 277-282, 2017.

[35] Y. Hua, J. Guo and H. Zhao, “Deep Belief Networks and Deep Learning,” *International Conference on Intelligent Computing and Internet of Things (ICTT),* pp. 1-4, 2015.

[36] K. Ping, C. Wei-Na and W. Qiao, “PREVIEW ON STRUCTURES AND ALGORITHMS OF DEEP LEARNING,” *2014 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing(ICCWAMTIP),* pp. 176-179, 2014.

[37] Zhijun Sun, “Review on the Study of deep Learning,” 2012.

[38] Jianwei Liu, “Research Progress of Learning depth,” 2014.

[39] Y. Shi and X. Lu, “THE ROLE OF BUSINESS INTELLIGENCE IN BUSINESS PERFORMANCE MANAGEMENT,” *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering,* vol. 4, pp. 184-186, 2010.

[40] T. Gang, C. Kai and S. Bei, “The Research & Appliction of Business Intelligence System in Retail Industry,” *Proceedings of the IEEE International Conference on Automation and Logistics,* pp. 87-91, 2008.

[41] M. Vega, C. Perra, F. D. Turck and A. Liotta, “A Review of Predictive Quality of Experience Management in Video Streaming Services,” *IEEE TRANSACTIONS ON BROADCASTING,* vol. 64, no. 2, pp. 432-444, 2018.

[42] H. Malekmohamadi, W. Fernando and A. Kondoz., “Automatic QoE Prediction in Stereoscopic videos,” *Proc. IEEE Int. Conf. Multimedia Expo Workshops,* pp. 581-586, Jul 2012.

[43] H. Luo and M.L. Shyu, “Quality of Service Provision in Mobile Multimedia-a Survey,” *SpringerOpen,* vol. 1, no. 5, pp. 1-15, 2011.

[44] Y. Tian, J. Srivastava, T. Huang and N. Contractor, Social Multimedia Computing, vol. 43, Computer, 2010, pp. 27-36.

[45] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery: An Overview,” in *Advances in Knowledge Discovery and Data Mining,* U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, Eds., Menlo Park, AAAI Press, 1996, pp. 1-34.

[46] Mannila Heikki, “Theoretical Frameworks for Data Mining,” *SIGKDD Explorations,* vol. 1, no. 2, pp. 30-32, 2000.

[47] J. Kleinberg, C. Papadimitriou and P. Raghavan, “A Microeconomic View of Data Mining,” in *Data Mining and Knowledge Discovery 2*, Kluwer Academic Publishers, 1998, pp. 311-324.

[48] Y. Yao, N. Zhong and Y. Zhao, “A Conceptual Framework of Data Mining,” *Springer-Verlag,* vol. 1, no. 118, pp. 501-515, 2008.

[49] James Manyika et al., “Big data: The next frontier for innovation, competition, and productivity,” 2011. [Online]. Available: www.mckinsey.com. [Accessed 16 March 2014].

[50] Berry M.J, “Data Mining Techniques: For Marketing, Sales and Customer Relationship Management,” John Wiley & Sons Incorporated, Hoboken, NJ, USA, 2004.

[51] P. Buhlmann and B. Yu, “Boosting with the¨ L2 loss: Regression and classification,” *Journal of the American Statistical Association,* vol. 98, p. 324–338, 2003.

[52] L. Breiman, “Random forests. Machine Learning,” vol. 45, pp. 5-32, 2001.

[53] C. Gonzalez, J. Mira-McWilliams and I. Juárez, “Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests, Statistical Laboratory,” *Escuela Técnica Superior de Ingenieros Industriales, Technical University of Madrid,* 2014.

[54] G. Biau, “Analysis of a Random Forests Model,” *Journal of Machine Learning Research,* vol. 13, pp. 1063-1095, 2012.

[55] L. Brieman, “BIAS, VARIANCE, AND ARCING CLASSIFIERS",” *Arcing [Boosting] is more successful than bagging in variance reduction,* 1996.

[56] P. Hothorn and T. Buhlmann, “Boosting algorithms: Regularization, prediction and model fitting (with¨ discussion),” *Statistical Science,* p. 22:477–522, 2007.

[57] M. O. Ward, G. Grinstein and D. Keim, “Interactive Data Visualization: Foundations, Techniques, and Applications,” *Second Edition. A. K. Peters, Ltd,* 2015.

[58] M. James Manyika et al., “Big data: The next frontier for innovation, competition, and productivity,” Global Institute, May, 2011. [Online]. Available: www.mckinsey.com. [Accessed 16 March 2014].

[59] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., New York: Springer, 2009.

[60] P. Kent, R. Kulkarni and U. Sglavo, “Finding Big Value in Big Data: Unlocking the Power of High-Performance Analytics,” in *Big Data and Business Analytics*, J. Liebowitz, Ed., CRC Press: Taylor & Francis Group, 2013, pp. 87-102.