University of Cape Town

Masters Thesis

Prediction of Mobile Network Subscriber Satisfaction by using Network Probing Experience Measures and Machine Learning

Author:
Martin J. Kruger

Supervisor:
Dr. Etienne A.D. Pienaar

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Data Science

Department of Statistical Sciences

June 11, 2025

Declaration of Authorship

I, Martin J. Kruger, declare that this thesis titled, *"Prediction of Mobile Network Subscriber Satisfaction by using Network Probing Experience Measures and Machine Learning"* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Signed by candidate

Date: June 11, 2025

“*The world is one big data problem.*”

Andrew McAfee

UNIVERSITY OF CAPE TOWN

Abstract

Faculty of Science

Department of Statistical Sciences

Master of Science

Prediction of Mobile Network Subscriber Satisfaction by using Network Probing Experience Measures and Machine Learning

by Martin J. Kruger

The mobile telecommunications market is a highly competitive and mature market and mobile network operators (MNOs) increasingly rely on the quality and reliability of the core services they offer to distinguish themselves from other market players. Customer satisfaction plays a crucial role in such a landscape where negative word of mouth could severely damage the reputation of a business. Customer satisfaction has therefore become a key differentiator for many companies. A popular metric to track customers' experience with a business is the *Net Promoter Score*® (NPS). NPS is measured by customer surveys, prompting them to answer a simple question: *"How likely are you to recommend company X to a friend or colleague?"* The response ranges between zero, representing not likely, to ten, representing very likely. The score value is obtained by grouping responses into three categories: *Promoters*, *Neutrals* or *Detractors*, and calculating the percentage difference between promoters and detractors. The more positive the value, the better overall customer perception is likely to be. A key shortcoming of NPS is that it does not provide tangible and directly interpretable reasons for customer responses.

This thesis aims to establish whether machine learning models, combined with network experience data collected by passive probing of mobile network interfaces, can accurately predict whether a subscriber will likely be a detractor. In addition, we would like to understand which network experience metrics are the best indicators of poor performance and negatively influence subscriber perception. We make use of survey and network data sourced from a large mobile network operator in South Africa over six months to create modelling features for cross validation of classification models with varying complexity to predict the NPS class of subscribers.

We find that mobile network data provided by present Customer Experience Management (CEM) systems may not be ideal for use in machine learning applications. The standard library of metrics and data structures used to perform classical CEM requires much effort to clean and prepare it as viable input to machine learning models. In addition, we find that all tested machine learning models, whether linear or non-linear, are poor predictors of NPS. This suggests that NPS may instead be driven by other factors, such as pricing or the interaction of customers with other processes that are more important and not represented within the present data.

Acknowledgements

I want to express my sincere gratitude to my loving family, which has supported me throughout this journey. Especially to my wife, Marlene. Thank you for your patience and enduring support. I would not have been able to complete this project without all your love and understanding.

To my supervisor, Dr Etienne Pienaar, thank you for your patience and guidance through the last hurdles.

Last but not least, I would like to thank my Heavenly Father for gracing me with his Holy Spirit and giving me the insight and perseverance to endure tough times.

Contents

List of Figures

List of Tables

List of Abbreviations

3GPP	3rd Generation Partnership Project
AUC	Area Under the Curve
CEI	Customer Experience Indicator
CS	Circuit-Switched
CX	Customer Experience
E-UTRAN	Evolved Universal Mobile Telecommunications System Terrestrial Radio Access Network
EDA	Exploratory Data Analysis
ETL	Extract Transform Load
ETSI	European Telecommunications Standards Institute
GDPR	General Data Protection Regulation
GERAN	GSM/Edge Radio Access Network
GLM	Generalized Linear Model
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
ICASA	Independent Communications Authority of South Africa
ICT	Information Communication Technology
LTE	Long Term Evolution
LTR	Likelihood To Recommend
MNO	Mobile Network Operator
MO	Mobile Originating
MSISDN	Mobile Station International Subscriber Directory Number
MT	Mobile Terminating
NOC	Network Operations Centre
NNPS	Network Net Promoter Score
NPS	Net Promoter Score
OSI	Open Systems Interconnection
PoPI	Protection of Personal Information
PS	Packet-Switched
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
SMS	Short Message Service
SQL	Structured Query Language
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
VoLTE	Voice over LTE

In memory of my Loving Father…

Chapter 1

Introduction

The mobile telecommunications (telco) market is a well saturated and mature market world-wide. Mobile network operators are in strong competition to attract new, and to retain existing subscribers in a cut-throat market characterised by rapid technological development, while at the same time being forced to continue lowering service margins. Additionally, the modern telco market landscape is subject to major issues and challenges related to fast maturing privacy policies and regulations, as outlined in the report "*Safety, Privacy and Security Across the Mobile Ecosystem*" by The GSM Association (2017).

Within this landscape, the viability of telco business relies increasingly on the quality and reliability of core services. Indeed, although there are only a handful of major market players in the South African context, the sector exhibits a highly liquid customer base and operators are having to compete to grow their share of the prepaid and lower-income markets, from which the majority of future growth is expected to come. In conjunction, our local market is under immense pressure from various other factors, including the lack of state awarded radio frequency spectrum, environmental concerns, unreliable electricity supply, and a general slowdown in economic activity; as outlined in the "*Electronic Communications Act: National Integrated ICT Policy White Paper*" by the Department of Telecommunications and Postal Services (2016).

In consequence, a principal metric in the day-to-day running of such a business is that of the health of the interface between the services provided, and the customer. This is not merely an engineering problem, i.e. probing the health of systems, but also the perceived quality of service on the customer's end. *Customer Experience* (CX) is the umbrella term that is commonly used by practitioners and scholars to refer to this concept. It is a complex topic debated and researched by many authors, and at present no common understanding exists regarding what customer experience entails (Becker and Jaakkola, 2020). Recent business practice has broadly defined customer experience as "encompassing every aspect of a company's offering – the quality of customer care, of course, but also advertising, packaging, product and service features, ease of use, and reliability. It is the internal and subjective response customers have to any direct or indirect contact with a company" (Meyer and Schwager, 2007) (as cited in Lemon and Verhoef, 2016). In the context of the present research project we interpret CX as the perceived quality that subscribers have while using the telco services that is provided by a mobile network operator (MNO).

There is a common understanding amongst business leaders that there's merit in continiously improving on CX – in the long run it not only builds positive relationships with customers, but it also leads to higher return on investment and creates sustainable competitive advantage. Work by Johnston and Michel (2008) has demonstrated the impact that improving the customer experience has on customer loyalty (as cited in Anaman and Lycett, 2010). However, to improve on the current experience, service providers must first be able to effectively *measure* and *model* customer experience. This means that service providers must also understand what matters most to their customers (Spiess et al., 2014). CX *measurement*, so to speak, plays a critical role in making insights actionable to business (Lemon and Verhoef, 2016). Although any respectable market player would have key quality metrics in place, these are often rudimentary in nature, and in the present context in need of improved precision. Granted, it cannot be claimed *a priori* that improvement is guaranteed to be possible – that is a function of the nature of the data – however, it stands to reason that it should be amendable to innovation. "In practice the majority of initiatives oriented at understanding CX are reactive and based on gathering explicit data related to experience (most commonly gathered through customer surveys)" (Anaman and Lycett, 2010).

A popular metric, used by many companies for more than a decade to guage and improve their overall customer satisfaction and loyalty, is the *Net Promoter Score®* (NPS). Reichheld proposed this single metric in his 2003 *Harvard Business Review* article *"The One Number You Need to Grow"*, as an alternative to other complex methods that attempt to measure customers' perceptions of their experience. NPS is calculated based on a customer's response to a single survey question: "*How likely, on a scale from 0 to 10, are you to recommend company X to a friend or colleague?*" The feedback classifies a subscriber either as a *promoter* (score 9-10); as a *passive* or *neutral* (score 7-8); or as a *detractor* (score 0-6). The NPS score is calculated as the percentage of promoters, less the percentage of detractors – producing a score between -100 and 100 (Reichheld, 2003). In summary, NPS assesses the likelihood of a subscriber to recommend the service to another, and can therefore be seen as a discriminator of customer experience (Hamilton, Lane, Gaston, Patton, Macdonald, Simpson, and Howie, 2014). The successful adoption of NPS by many large companies as their main measure of subscriber satisfaction and loyalty has been driven by a few factors. Firstly, because the NPS survey is short and quick to respond to, subscriber's response rate to the survey is higher than other lengthy questionnaires that require more effort to complete. Further, Lemon and Verhoef (2016) suggests that businesses tend to use simple, usually single-item measures like NPS, because it is intuitive in nature and easily understood by top management. Zeithaml, Bolton, Deighton, Keiningham, Lemon, and Petersen (2006) have also suggested that NPS is more of a forward-looking metric focusing on the present status, whereas satisfaction is more of a backward-looking metric (as cited in Lemon and Verhoef, 2016).

Once an understanding of CX is obtained, it is important for service providers to translate the gathered information into actionable insights and to have a strategy of pro-actively addressing the issues that result to bad experience (Anaman and Lycett, 2010). Service providers need to make the necessary targeted investments and implement actions to optimize customer experience (Spiess et al., 2014). *Customer Experience Management* (CEM) is the hypernym that is generally used to refer to this process, of extracting usefull information from customer interactions to gain insights to what business strategy needs to be applied to optimise CX.

A notable drawback of the NPS measurement however, is the lack of ability to identify and act upon the driving factors behind customers' responses based on solely the NPS feedback. To improve the bad experience of a detractor, business first needs to establish the root cause that resulted to an unhappy customer. Secondly, even though the response rate to NPS is greater than other customer surveys, the sample size is often still too small to make significant statistical inference from the data. In particular, when for example a MNO wants to focus on subsets of the data – like smaller geographic managed regions of its network, or on specific segments of the subscriber base. The present research project aims to address both of the last mentioned drawbacks of NPS.

Project Positioning

Current research is situated within a leading MNO in South Africa (referred to as Telco hereafter), and more specifically the focus is from a network performance management perspective. The function of network performance within Telco, is to look after the long term "health" of the cellular network, and to strategically drive the day-to-day operational activities within the core-network as well as the radio-network operated by Telco. Customer Experience Management (CEM) forms part of this network-centric function, and predominantly makes use of probing data from network interfaces (see Section 2.1) to measure the quality of service provided to mobile subscribers – using metrics labeled as *Customer Experience Indicators* (CEIs). The key benefits of using CEIs to measure service quality – in contrast to using network element based data – is that it is vendor agnostic[1]; and more importantly, the quality of mobile services can be monitored down to the individual subscriber level. It is important to note the network-centric setting of CEM within Telco, which is quite different to the general understanding of CEM within marketing and business domains.

Concurrently to the above mentioned network performance management activities, Telco is also collecting NPS data from subscribers in order to track CX from a holistic marketing and business perspective. Within Telco, the particular NPS survey is instead referred to as *Network NPS* (or NNPS), in order to distinguish it from another NPS survey, which is used to measure the official South African Customer Satisfaction Index *(SA-csi)* of mobile telecommunications services. The SA-csi is an independent national benchmark of customer satisfaction of the quality of products and services available to household consumers in Africa (Consulta, n.d.), and is conducted by an external third party. We shall therefore refer to the Telco managed survey as NNPS hereafter in order to avoid ambiguity.

NNPS data is collected in-house by Telco through an electronic self-managed process, with the key objective of gathering a large number of customer experience measurements on a regular and ongoing basis as it pertains to the use of services on Telco's network. The results from the NNPS survey is then used for closed-loop feedback to guide strategic and targeted improvement of Telco's infrastructure. In particular, where there are areas containing higher concentrations of detractors, investment into network improvement initiatives will be prioritised to improve the network experience of customers within those areas. The before mentioned issues with NPS are however impeding the success of the process, and potential value from the NNPS data is possibly not fully harnessed.

[1] "Vendor agnostic" in the current setting refers to having a common interpretation of network quality, instead of having multiple manufacturer specific definitions; more specifically when network equipment from multiple manufacturers are used by a MNO. Interoperability standards defined by organisations like *3rd Generation Partnership Project* (3GPP) does define the network functions and communication protocols used on the interfaces between network nodes, but internally to a manufacturer's equipment the provided functionality and metric definitions can be unique.

Aim and Objectives

It is the purpose of the present research project to determine whether network interface probing data can be used effectively to address the shortcomings of NPS-based customer experience management in the context of cellular mobile telephone networks.

The aforementioned shortcomings of NPS are:

- The inability to determine driving factors behind NPS detraction and low subscriber satisfaction;
- Inadequate NPS sample size to represent the population at large for generalised statistical inference of customer experience.

In order to meet the aforementioned goals we fit a variety of classification machine learning models that predict the NPS class (*detractor* or *promoter*) of subscribers based on CEI metrics collected from a probing based CEM solution used by Telco. Training data containing the known NPS responses, comes from the electronic NNPS customer surveys collected by SMS.

A best performing classification model is identified based on evaluation of the highest prediction *accuracy* and *area under the curve* (AUC) of all models being assessed. In conjuction to model performance, we are also interested in gaining insight into the key *network related factors* that drive negative subscriber satisfaction – the predicted likelihood of detraction being a proxy for dissatisfaction or negative sentiment. We also investigate the impact of various data transformations and the removal of outliers on model performance.

The specific objectives of the project include:

- Prepare a consolidated dataset that merges subscriber NNPS results with customer experience metrics from multiple cellular network interfaces, into a format that is suitable for modelling;
- Determine the level of accuracy that can be obtained with prediction of NPS detraction based on CEI network metrics;
- Identify which type of machine learning model has the highest classification accuracy and AUC for prediction of NPS detraction from network probing data;
- Identify the most important network experience factors and conditions that drive NPS detraction.

Thesis Outline

The rest of this paper is structured as follows:

In Chapter 2 we provide an overview of the mobile telecommunications network as well as the measurement of customer experience within this setting. We also explore existing research that is in some way related to the present project. In Chapter 3 we discuss the data used in the project, as well as its preparation in order to allow machine learning to be applied. In Chapter 4 we visually examine the data using exploratory data analysis methods. In Chapter 5 we describe the application of the methods to analyse the data and to select the best performing models. Finally in Chapter 6 we give a summary of the key results and provide concluding remarks.

Chapter 2

Contextual Overview and Literature Review

The key objective of the current project is to predict the NPS class of mobile network subscribers by utilising machine learning models. The predictions will be based on data collected during user interactions with core services provided by cellular networks. This is done through a process known as *network interface probing*, which will be discussed in the following sections. Another objective of the project is to identify the primary *network-related* factors that deteriorate *quality of experience* and negatively impact customer satisfaction.

The outcome of having such capability has far-reaching implications. Using this novel approach, any mobile network operator (MNO) may eventually be able to determine which network-coupled quality factors contribute to low overall satisfaction with their service, without necessarily receiving customer complaints. Business strategy can therefore be aligned accordingly, to achieve optimal customer experience improvement, and bring with it the associated positive by-products like customer loyalty and positive brand perception. The concept can even be taken one step further. Network operators can *pre-emptively* make subscribers aware of improvements to infrastructure – for example, by distributing SMS messages to all subscribers that are likely to be detractors; to inform them of present changes to improve network quality. The hypothesis is that sensitising subscribers beforehand, may trigger them to be consciously aware of changes to their experience. Subsequently, even the slightest improvement to the network may lead to positive effects on their overall perception (a concept known as *confirmation bias*), and ultimately be reflected by higher NPS scores.

We start this section with a short introduction to cellular communication networks and the basic principles needed to understand how the data used in the project is collected.

2.1 The Cellular Network

Cellular networks are complex wireless infrastructures that mobile network operators (MNOs) provide as a communications platform and service to end users. Services in the context of mobile telecommunications generally refer to any communication function, such as *voice calls* or *data transfers*, utilising a wireless cellular network infrastructure to carry user information between source and destination, regardless of whether the user is a natural person or an application server. MNOs usually have their own network infrastructure comprising various interconnected communication systems. The systems consist of multiple pieces of equipment, referred to as *network elements*, each providing a specific function needed for the combined system to function as a whole. Network elements are colloquially also referred to as *network nodes*.

Mobile Service Types

The services which are provided by cellular networks can be divided into mainly two categories or types of services, namely *circuit-switched* services and *packet-switched* services.

Circuit-switched Services

Circuit-switching (CS) is a technique used in telecommunications networks, where two network nodes establish a dedicated communications channel or circuit through the network before the nodes can communicate. The technique emulates the early analog telephone network, where switches within the telephone exchange created a continuous wire circuit between two telephones for as long as a call lasted. Circuit switching is most commonly used for connecting voice circuits, presenting the benefit of reducing routing overhead and the disadvantage of inefficient uncontended use of the total available network capacity.

Packet-switched Services

Packet-switching (PS), as alternative to CS, is a method of grouping data transmitted over a digital network into packets, consisting of a header section used for routing and a payload section containing the user data. The packetisation of data allows for variable bitrate transmission and more efficient resource contention using statistical multiplexing or dynamic bandwidth allocation techniques. This contention results in variable latency and throughput, and as such, are not well suited to time- and bandwidth-sensitive services such as voice.

In billable services, such as cellular communication, CS-based services is characterised by a fee per unit of connection time, even when no data is transferred. At the same time, PS-based services may be represented by a cost per unit of information transmitted; such as bytes, characters, packets, or messages.

Network Architecture

Network architecture refers to the physical as well as logical design of network services and the supporting network appliances, to serve the connectivity requirements of end users. The architecture of cellular networks is forever changing, primarily driven by technological advancement in wireless technologies and mobile devices, which are guided by end users' adoption of data-hungry services. We refer to these different phases as *network generations*, where the *genera* are colloquially referred to as 1G, 2G, 3G, 4G, and 5G at present. Each is associated with varying radio capabilities and a progressive increase in data bitrates.

The first introduction of wireless cellular telecommunication networks, 1G, took place during the 1980s and was originally based on *analog* technology. During the early 1990s the evolution to *digital* technology was termed the "second generation", 2G, and was formalised and defined as a standard termed the *Global System for Mobile Communication* (GSM) by the *European Telecommunications Standards Institute* (ETSI). Today, even though 3G through 5G evolutions came about, the common term "GSM" is still often used to refer to cellular telecommunication.

The present project covers 2G, 3G and 4G technologies, included in the network architecture used by Telco, and for which experience measurement data is readily available.

Network Components

Cellular networks consist of the interconnection of largely five main components:

1. User devices called **Mobile Stations** (MS) or **User Equipment** (UE), which refers to the combination of all user equipment and software needed for communication with a mobile network.
2. **Radio Access Network** (RAN), which can be a combination of 2G, 3G, and 4G generation networks.
3. **Core Network** (CN) which provides centralised functions and connects the RAN to other external networks.
4. **Operations and Support Subsystem** (OSS) is a combination of various functional entities that provide a network operator with the ability to monitor and control the cellular network.
5. External public networks like the wider **Public Switched Telephone Network** (PSTN) and **Public Data Network** (PDN) or Internet.

Internal to each of the main network components, multiple network nodes are responsible for the end-to-end communication services provided to end users. The connections between the nodes are termed *interfaces*, which we discuss in more detail in the following section. Figure 2.1 provides a high-level architectural overview of a typical cellular network, containing all functionalities covered within the current project. The point of reference regarding experience measurements collection for the project, is predominantly on the interfaces that form the border between the *radio access network* (RAN) and the *core network* (CN).

Besides the main function, i.e. transport of user-information between source and destination, network nodes are also responsible to store configuration data, and to collect *quality of service* (QoS) measurements. This information is used by network operators – within the *Operations Support Subsystem* (OSS) component (shown at the bottom of Figure 2.1) – to control the network, and to track the level of service provided to end users.

Network Interfaces

From the standpoint of a cellular network, a *network interface* can be described as a component; that contains all the hardware and software required, to connect a network element to the network physically, and to allow it to communicate with other devices in the network. Similarly, it is the interconnection point between a network node and its surrounding network. The concept of network *interface* is used within the industry to define standards around the interconnection of network entities and to enable interoperability of equipment from different manufacturers.

The *"3rd Generation Partnership Project"* ("3rd Generation Partnership Project", n.d.) is a global initiative, specific to cellular telecommunications, that unite seven standard development organisations known as "organisational partners" – providing their members with a stable environment to produce the reports and specifications that define 3GPP technologies. Current 3GPP specifications cover cellular telecommunications technologies, including radio access, core network and service capabilities – providing

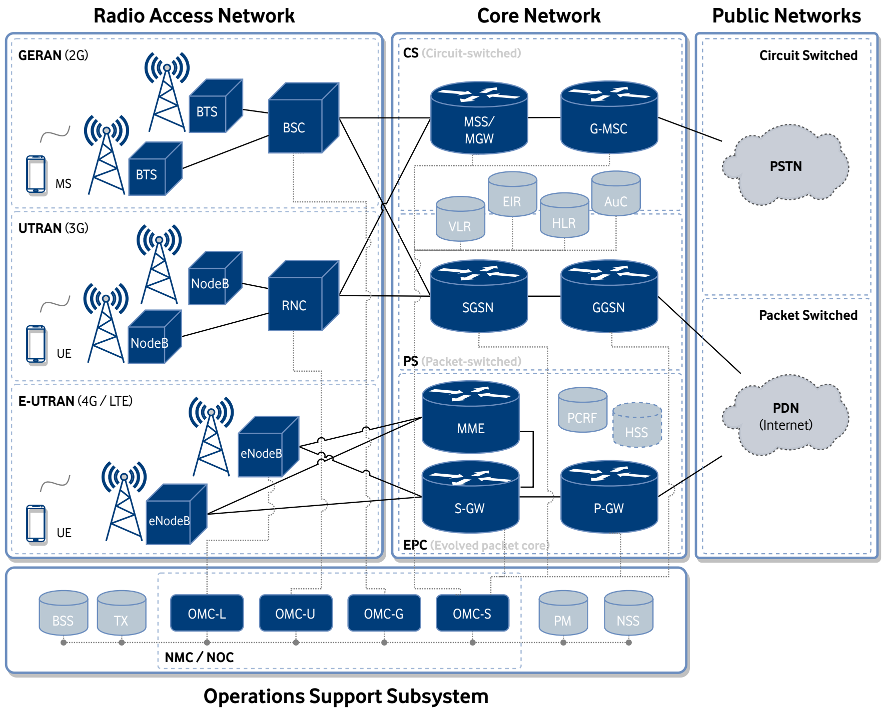


FIGURE 2.1: High level architectural overview of a cellular network containing elements of the 2G, 3G and 4G network generations.

a complete system description for mobile telecommunications. Naming of interfaces in cellular networks follow a standard convention defined by 3GPP. The interfaces are named using a combination of letters and numbers, that represent the function and location of the interface within the network. Nomenclature is based on the *International Telecommunication Union* (ITU) standard, which provides a standardised set of abbreviations for telecommunications terms. Overall, the naming convention is intended to provide a consistent way to refer to the various cellular network interfaces, and it helps to ensure interoperability between different network elements from different vendors. Below are some examples of interface names and their meanings. The **bold** labeled interfaces are relevant to the present project:

- *Um*: Air interface between the mobile device (MS or UE) and the base transceiver station (BTS).
- *Abis*: Interface between the BTS and the base station controller (BSC).
- ***A***: Interface between the BSC and the mobile switching center (MSC).
- ***Gb***: Interface between the BSC and the serving GPRS support node (SGSN).
- ***IuCS***: Interface between the radio network controller (RNC) and the MSC.
- ***IuPS***: Interface between the RNC and the SGSN.
- ***Gn***: Interface between the SGSNs and the GGSNs.
- ***S1-U***: Interface between the eNodeB (eNB) and the serving gateway (SGW).
- ***S1-MME***: Interface between the eNB and the mobility management entity (MME).
- ***S11***: Interface between the SGW and the MME.
- *Gi*: Interface between the GGSN and an external packet data network (PDN), like the Internet.

The common function of all network interfaces, is to transport payloads of application data, from one node to other supporting nodes; according to a structured and conformed set of rules known as *communication protocol.* The data as such, is related to the specific functions provided by each node, and is a combination of user data and information controlling the end-to-end flow between source and target.

Communication Protocols

Conceptually, network communication protocols are described by the *Open Systems Interconnection* (OSI) model, shown in Figure 2.2. The model describes seven logic protocol layers that computer- and telco-systems use, to communicate over a network. It was the first standard model for network communications, adopted by all major telecommunication and computer companies during the 1980s.

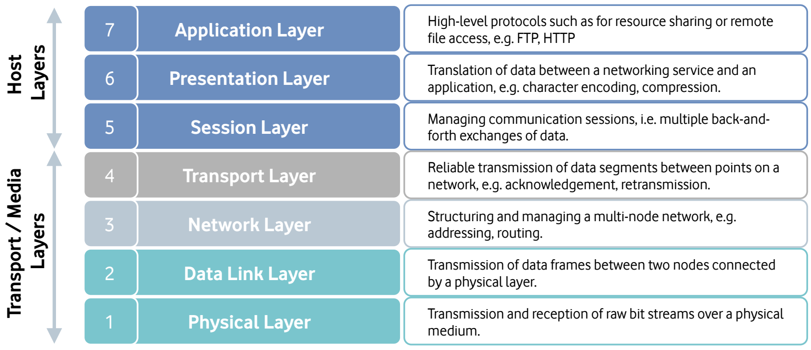


FIGURE 2.2: The *Open Systems Interconnection* (OSI) model describes seven conceptual layers that computer and telecommunication systems use to communicate over a network. The model was formalised in 1984 by the *International Organisation for Standardisation* (ISO) as standard ISO/IEC 7498.

By definition, *communication protocols* enable an entity in one host, to interact with a corresponding entity (at the same layer) in another host. The OSI model abstractly describes the functions provided to a layer N, by a layer N – 1; where N is one of the seven protocol layers operating in the local host. At each level N, two entities at the communicating devices – layer N *peers* – exchange *protocol data units* (PDUs) by means of a layer N protocol. Each PDU contains a payload, called the *service data unit* (SDU), along with protocol-related headers or footers. In consequence, data transmission by two communicating OSI-compatible devices proceeds as follows:

1. The data to be transmitted is composed at the topmost layer N of the transmitting device into a PDU, and passed to layer N – 1, where it is known as a SDU.
2. At layer N – 1 the SDU is concatenated with a header, a footer, or both – producing a layer N – 1 PDU. It is then passed to layer N – 2.
3. The process continues until reaching the lowermost level, from which the data is transmitted to the receiving device through some physical transmission medium.
4. At the receiving device the data is passed from the lowest to the highest layer as a series of SDUs while being successively stripped from each layer's header or footer until reaching the topmost layer, where the last of the data is consumed.

Practical implementations of network protocols are generally more straightforward and do not combine all of the functions described by the seven-layer model as a single entity. Protocols are usually focused on specific networking functions, and the conceptual tasks defined within the OSI model are either combined to span multiple layers; or segmented further to represent only a portion of a layer. Protocols also don't exist in isolation and are usually packaged as modular *"protocol stacks"* that can be re-purposed by different interfaces. As such, protocol stacks are commonly abstracted as *host* or *transport* layer protocols – to allow independent operation of the upper layers from the lower layers, as shown in Figure 2.2.

This abstraction allows applications higher up in the protocol stack, to focus on achieving more complex tasks, and to entrust lower level tasks to the transport layers. Figure 2.3 explains the concept – it shows the different protocols used to transport *voice* in a GSM network. At the highest level, the *Connection Management* (CM) protocol is responsible for *Call Control* and *Short Message Service*. Below CM, other protocols are responsible to manage tasks related to cellular communications, like *Mobility Management* (MM) allowing seamless hand-over of calls between towers as subscribers move around in the RAN; resulting to changing radio conditions.

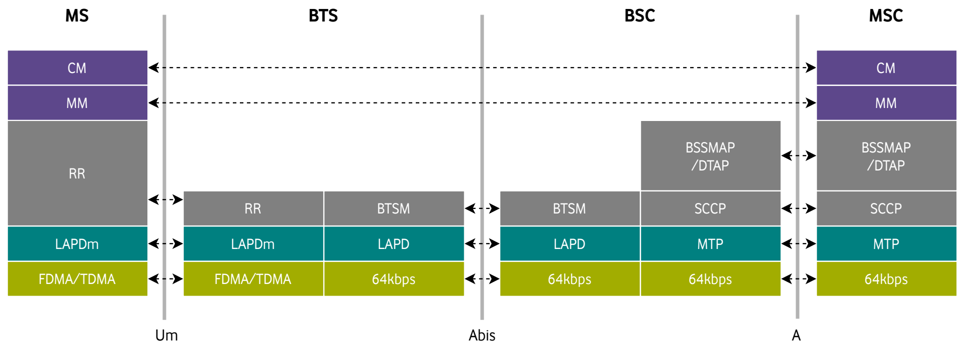


FIGURE 2.3: The GSM protocol stack used to transport voice in a 2G cellular network.

Dissimilar services, such as voice calls in contrast to WhatsApp messaging, generally make use of discrete interfaces to transmit its data. More specifically, CS services are managed distinct from PS services. This however was not the case in older network generations. Before the transition from 2G, to more recent generations, the same interfaces were used to provide multiple services; like CS voice, SMS and dial-up data modem connections – all via the *A* interface which is still around today. Table 2.1 shows the subscriber services and the supporting interfaces for each service used as data sources in the present project.

Network Probing

In Section 2.1, Network Components, we pointed out that network nodes collect "health" measurements during communications. By analysing the recorded QoS data, network operations personnel can gain valuable insight into the quality of service provided to end users. This is achieved through mechanisms such as alarming and presentation of specialised reporting via graphical user interfaces. Generally, MNOs deploy alarms to continuously monitor QoS levels and to trigger root-cause investigations by human intervention when service levels exceed predefined thresholds. Reporting on the other hand, is used for *post-mortem* analysis by visually presenting network personnel with statistical graphs that have slice-and-dice capability, allowing them to scrutinise detail from multiple perspectives. The reporting process also aids in identifying long-term quality trends and determining the required changes to the network that will improve customer experience in the future.

Interface	Radio Technology	Service Type	Subscriber Services	Description
A	2G	CS	Voice / SMS	Circuit-switched voice and SMS services provided by the GERAN radio access network.
Gb	2G	PS	GPRS Data	Packet-switched data control plane provided by the GERAN radio access network.
IuCS	3G	CS	Voice / SMS	Circuit-switched voice and SMS services provided by the UTRAN radio access network.
IuPS	3G	PS	3G Data	Packet-switched data control plane provided by the UTRAN radio access network.
Gn	2G, 3G	PS-CP / PS-UP	Data Session *(CP)* / Data *(UP)*	Carries data session control plane signalling and user plane packet switched services provided by the GERAN or UTRAN radio access networks.
S1-MME	4G	PS	LTE/4G Data	Packet-switched data control plane provided by the E-UTRAN radio access network.
S1-U	4G	PS-UP	Data *(UP)* / PS Voice (VoLTE)	User plane data services provided by the E-UTRAN radio access network.
S11	4G	PS-CP	Data Session *(CP)*	Data service control plane provided by the EPC core network.

CS: Circuit-switched, **PS**: Packet-switched, **CP**: Control-plane, **UP**: User-plane, **VoLTE**: Voice over LTE

TABLE 2.1: Cellular network interfaces and the subscriber services that are transported via each interface.

Motivation for Probing

MNOs can provide a good quality of service by merely using the health indicators and rudimentary tools mentioned above. However, beyond that, a more specialised *modus operandi* is required if MNOs want to deliver a level of service, that meets and exceeds individual subscribers' expectations. The aim should be to gradually close the gap between what service levels are as measured from the network's perspective, versus what the experience is like from a customer's point of view – as shown in Figure 2.4.

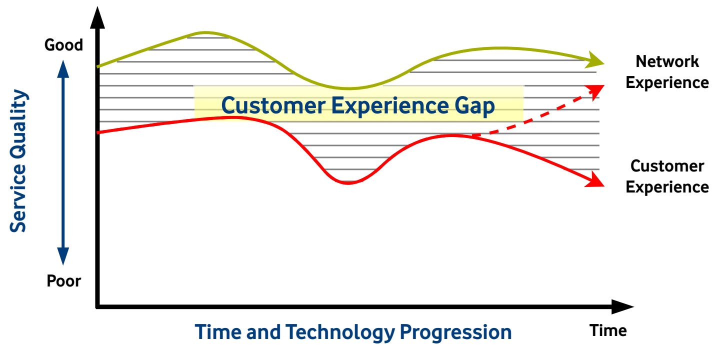


FIGURE 2.4: The customer experience gap is the difference between, service quality as measured from a network perspective by using *quality of service* (QoS) measurement data, versus the actual subscriber *quality of experience* (QoE).

At this point, it is worth outlining a few limitations of node-based QoS measurements:

- Measurements are generally aggregated and can only be examined to a level of detail that corresponds to some characteristic of a node itself or to a level that corresponds to the logical hierarchical position of the node within the network; for example, a BTS-node will collect measurements that logically represent down to a cell level, or internally to a physical radio module. The network performance of a

MNO can therefore only be assessed and fine-tuned to some average value, which is an aggregate of all subscriber activity that passes through a given network point.

- Due to the localised recording of QoS metrics, MNOs have to manually correlate statistics from multiple network elements and locations within the network to get a complete end-to-end view of network performance.
- Definitions of key performance indicators and the interpretation of fault conditions may differ between equipment manufacturers. Subsequently, when MNOs use equipment from multiple equipment vendors within the same network, no common understanding of a particular QoS metric, or the underlying fault conditions affecting the metric, exists through the entire service footprint of the service provider.

Probing Methodology

To this end, in as far as measuring customer experience, the telecommunication industry (amongst others) deploy network probes – a concept also known as *passive network monitoring*. These (non-invasive) nodes collect communication protocol information from network interfaces, and provide operators with subscriber level information about the interaction between customer and network. Figure 2.5 shows some of the complexity behind basic services, like voice calls in a GSM network. The example shown covers the end-to-end process of protocol signalling and message exchange, between a subscriber's mobile device (MS) and the supporting network nodes, for the scenario when a subscriber initiates a voice call in a 2G network setting.

The call flow shown will be more complex and look entirely different in the case of later network generations – not even considering error scenarios. Probes are consequently essential in the modern day-to-day operation of any MNO. Not only does it enable operators to intercept the plethora of protocol messages without affecting the live network traffic, but it further enables more sophisticated interpretation of service quality, by feeding the data into specialised external applications.

Customer Experience Management Systems

As a result of the emergence and evolution of a few key technologies in the telecom industry in recent decades, operators are now closer to understanding each subscriber's experience each time they interact with operator services. Traditional probing solutions have improved to now be able to collect detailed call and data-session records – from multiple network interfaces, for all subscribers, all the time – as well as providing consolidated exports of the information in real-time to downstream systems via technologies such as *Apache Kafka*[1].

On the downstream side, we have seen the development of industry-specific *Customer Experience Management* (CEM) systems, which incorporate *big data* technologies to ingest and process the large volumes of transactional information associated with customer-level probing within a telecommunications environment. Figure 2.6 illustrates how probe data is extracted from a live network interface – using a method known as

[1] Apache Kafka® is a real-time data streaming technology capable of handling trillions of events per day. Initially conceived as a messaging queue, Kafka is based on an abstraction of a distributed commit log. Since being created and open sourced in 2011, Kafka has since become the industry standard for working with streaming data.

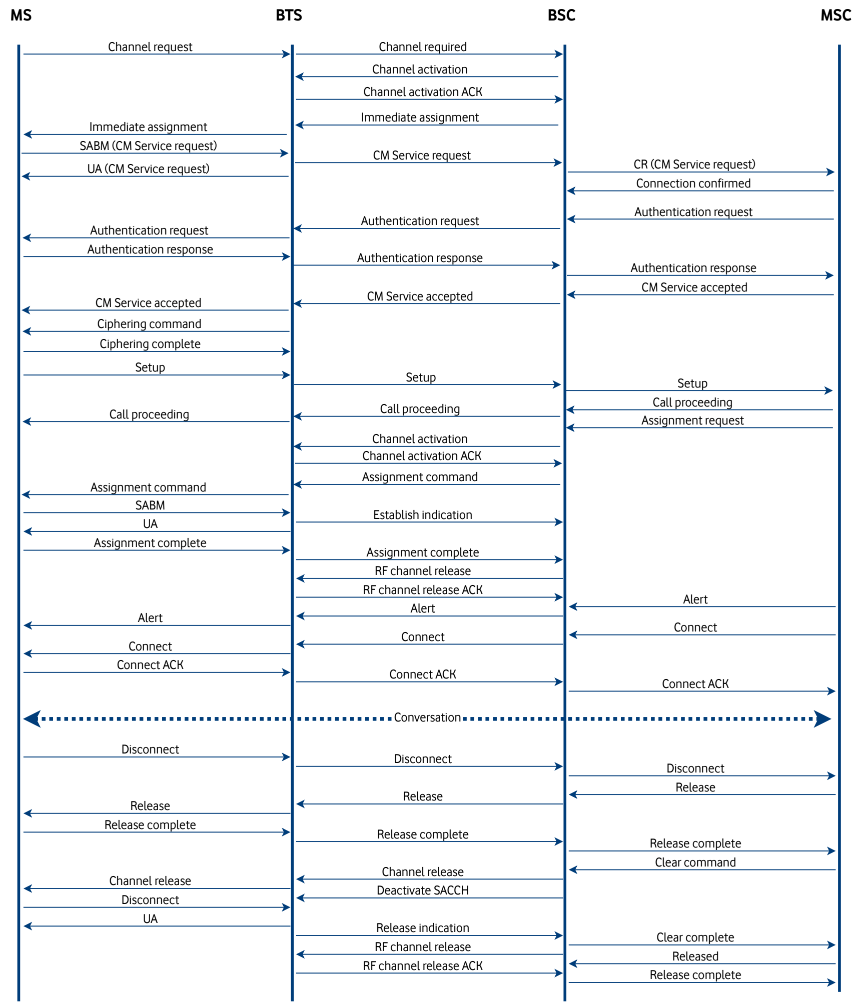


FIGURE 2.5: A signalling ladder diagram describing the typical flow of protocol messages between network nodes and a subscriber handset (MS), when a subscriber is making a mobile originating voice call in a 2G cellular network.

"tapping" – and fed into an external CEM system for correlation, processing, and aggregation, before being saved in a data store.

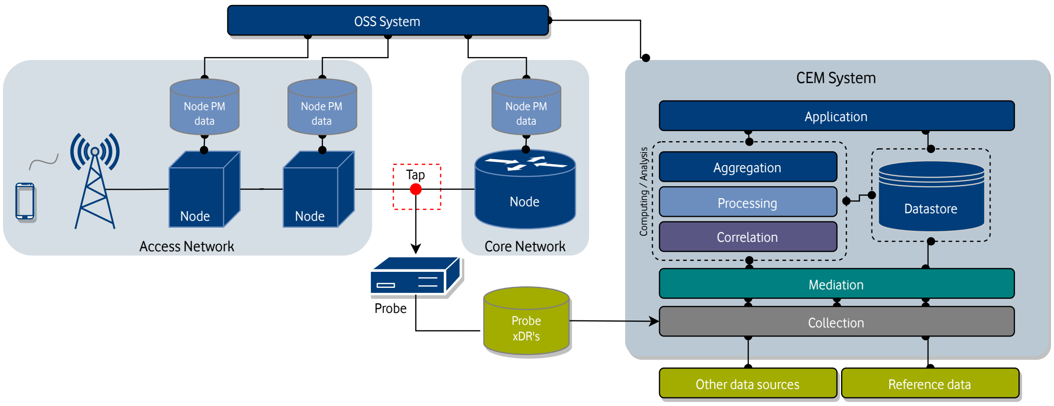


FIGURE 2.6: The process of collecting CEM-data by "tapping" live network interfaces with probes. The image also shows the conceptual difference between probing data in comparison to node performance data.

CEM systems allow network operators to capture and process raw network events from probes, turning it into realistic "*quality of experience*" metrics – generally referred to as *Customer Experience Indicators* (CEIs).

Customer Experience Indicators

CEI metrics are typically calculated by complex event processing models integrated into the CEM platforms mentioned earlier. The objective of the models is to quantify the *quality of experience* (QoE) of each network interaction – from the *customer's* perspective. QoE assessment is performed by analysing consecutive flows of protocol messages collected from probe sources and providing summary metrics summarising the entire end-to-end transaction. An example of one such end-to-end scenario is shown in Figure 2.5, which entails signalling associated with an originating voice call in a 2G network.

Typical CEIs collected from such an end-to-end voice session may include but are not limited to metrics like the ones listed below – focussing on a single 2G subscriber during a given time period:

1. ***Mobile Originating 2G Call Attempts*** – The total number of *mobile originating* (MO) voice calls attempted by the subscriber on 2G. This is a count of *CM Service request* messages seen on the *A* interface originating from the subscriber's MS.
2. ***Mobile Originating 2G Setup Fail Network*** – The total number of MO voice calls attempted by the subscriber on 2G, that failed due to *network* related error conditions. This is a count of the *CM Service request* messages without successful *Setup* responses, and having *network* related release cause values in the *Clear command* response messages.
3. ***Mobile Originating 2G Setup Duration*** – The average time in milliseconds (ms) that the network takes to setup a MO voice call on 2G. This is the average time between *Connection confirmed* and the *Call proceeding* messages on the *A* interface.
4. ***Mobile Originating 2G Call Answered*** – The total number of MO voice calls answered by the subscriber on 2G. This is the count of *Connect ACK* messages having *conversation* time greater than zero.
5. ***Mobile Originating 2G Call Dropped Network*** – The total number of MO voice calls that were unintentionally disconnected due to *network* related error conditions.

This is the count of *Connect ACK* messages having conversation time greater than zero, without the associated *Disconnect* messages initiated by the MS, where the *Clear command* contains *network* related release cause values.

6. ***Mobile Originating 2G Call Duration*** – The average time in seconds (s) that the subscriber spends on active 2G MO voice calls. This is the average time between *Connect ACK* and *Disconnect* messages for all calls initiated by the subscriber MS.

Similar metrics are available for the *terminating* (MT) side of calls and across all network generations (2G, 3G and 4G). The metrics for other end-user services, such as *SMS* and *PS data*, are different altogether, and are more appropriate for quantifying the user experience as it relates to each communication protocol used by that particular service. We provide a complete summary of all aggregate CEI metrics that forms part of the present project in Chapter 3, and detail descriptions of the underlying discrete metrics used to quantify each service's quality in Chapter B. The below list contains references to the relevant tables for each subscriber service:

- **Voice** Chapter 3 Tables B.6 and B.7, and Appendix B Table B.2.
- **SMS** Chapter 3 Table B.9 and Appendix B Table B.3.
- **PS data control-plane** Chapter 3 Tables B.10 and B.11 and Appendix B Table B.4.
- **PS data user-plane** applications Chapter 3 Tables B.13 and B.14 and Appendix B Table B.5.

2.2 Customer Centric Service Management

The competitive nature of the telco landscape is one of the primary motivations for proper network management. Consequently, the need to constantly evaluate service quality is central to maintaining market position. Laghari and Connelly (2012) explain how network management concepts are evolving and that autonomic network management paradigms aspire to bring human-like intelligence to telecommunication management tasks. Thanks to these technical advancements, the fulfilment of customer demands and user experience requirements, are becoming the main differentiators for the effectiveness of MNOs. In this era of competition, poor customer experience leads to a chain reaction of negative word of mouth, pushing customers into the arms of waiting competitors. Humans are inherently quality meters today and their expectations, perceptions, and needs concerning a particular product or service carry significant value.

Recently the concept of *quality of experience* (QoE) has also become a major research theme within the telecommunications community. QoE can be described as an assessment of human experience when interacting with technology and business entities in a particular context. QoE is a fast emerging field based on multidisciplinary fields of study involving engineering science, cognitive science, economics, marketing and social psychology – all focused on trying to establish a better understanding of overall human quality requirements (Möller and Raake, 2014). In consequence, the services rendered by MNOs has also shown a gradual shift in focus from being *quality of service* (QoS) driven, to becoming *quality of experience* (QoE) driven. We believe that the present project can assist MNOs to bridge the gap by transitioning from existing methods to be more aligned with customer experience.

From Quality of Service to Quality of Experience

Traditional network performance management methods were largely focussed around the use of QoS measures or *key performance indicators* (KPIs) coming from network nodes to manage networks and to guide operators on where to focus their investment and improvement initiatives. The "old school" ways of purely QoS driven focus has however changed over the last decades as the ICT industry has been influenced by ideas coming from other domains like Customer Relations Management (CRM), Human Computer Interface (HCI) design and marketing fields. Quality management in these fields are more aimed at boosting of sales and improvement of customer retention and brand loyalty. Concepts coming from these fields, like the *perceived level of quality* by customers or the *quality of experience* (QoE), motivated the need for a radical shift from purely technical quality requirements to a more customer focussed approach of experience guarantees.

The use of CEM systems in the telecommunications industry has brought about a significant change in how Mobile Network Operators (MNOs) manage service quality. With the ability to focus on the quality of service provided to individual customers and customize the network according to their experience, Telcos have become more *customer-centric* in their approach to service management. However, measuring the quality of experience directly is challenging. Therefore, in this study, we aim to bridge the gap between the network-centric view of experience and the customer-centric view by using machine learning to search for patterns in data. Our goal is to relate the subjective perceived quality to the existing objective quality of service measures that operators use to improve customer experience. We will be utilising the Net Promoter Score (NPS) as the subjective measurement for this analysis, which represents the customer's likelihood of recommending the service to others.

2.3 Net Promoter Score®

The Net Promoter Score (NPS) is generally regarded to be a good measure of the overall willingness of customers to recommend a company's products or services to others. It is a measure of customer loyalty to a brand and used as a proxy for gauging customer satisfaction with a company's products or services. It serves as an alternative to traditional customer satisfaction research and claims to be correlated with revenue and business growth. NPS can be summarised as being a management tool or loyalty metric that can be used to gauge the overall status of a company's customer satisfaction. It has been widely adopted by Fortune 500 companies and other organizations as a benchmark of customer satisfaction (Reichheld, n.d.; CG, 2023).

The American business strategist Frederick Reichheld introduced the measure of customer loyalty in his 2003 Harvard Business Review article *"The one number you need to grow"*. The measure that he introduced is called the Net Promoter Score® (NPS®) and was developed by, and is a registered trademark of Reichheld in partnership with Bain & Company and Satmetrix (Reichheld, 2003).

Reichheld stated in the article that one of the most effective ways to gauge an organizations imprint in an industry, is to measure the customer satisfaction. Customer satisfaction leads to loyalty, but it is very hard to measure customer satisfaction. What exactly does "loyalty" mean? Loyalty cannot be defined by simply measuring the rate of store visits or the rate at which services are being used. A customer that visits the only supermarket in his home town or neighbourhood on a regular basis does not have to be loyal to that brand of stores, he merely visits because it is the only supermarket at his disposal. A customer that always buys his cars from the same company may be very loyal, but he does not visit that company very often because he doesn't regularly buy a new car.

Reichheld therefore reasoned that customer loyalty should rather be defined as the *willingness of a customer to recommend* a company to his family, friends or colleagues. With this in mind loyalty could therefore be measured by one simple question, instead of using long and complex customer surveys. The method he proposed involves asking participants to choose a point on an eleven point scale to rate their overall satisfaction with a company. The single question that he proposed as best measure of customer loyalty is:

"On a scale from zero to ten, how likely is it that you would recommend [company X] to a friend or a colleague?"

Reichheld derived this question through research by linking customer responses to their customer experience and to levels of loyalty or advocacy towards companies. He claimed to observe a clear correlation between the Net Promoter Score® and company growth.

Customer Loyalty Groups

The results from the survey is used to split customers into three groups that Reichheld calls *detractors*, *passives* (or *neutrals*) and *promoters*. The classification is based on their response to the survey question:

- **Detractors**: Gave a score less than or equal to 6.
- **Passives** (or **neutrals**): Gave a score of 7 or 8.
- **Promoters**: Gave a score of 9 or 10.

General behaviours associated with the aforementioned customer groups are:

Detractors are the group of customers that will most likely have a negative influence on a company's image. They are not satisfied with the products or services and in all likelihood will not purchase from the company again, and could potentially damage the company's reputation through negative word of mouth.

Promoters are those customers that will most likely have a positive effect on a company. They are satisfied with the company's products and services and are likely repeat buyers, or the enthusiastic evangelists who recommend the company's products and services to other buyers.

Passives or *Neutrals* are those customers that are somewhat satisfied, but could easily switch to a competitor's offering if they are given the opportunity. They probably won't spread any negative word of mouth, but are not enthusiastic enough about the company's products or service to actually promote them.

NPS® Calculation

The Net Promoter Score® is a dimensionless quantity determined by subtracting the percentage of customers who are detractors, from the percentage of customers who are promoters as summarised in Figure 2.7 below. What is generated is a score between -100 and 100 called the NPS®. To put this in context, if all of the customers were answering the question with a 9 or 10, then the NPS would be 100. On the other side of the spectrum, if all of the customers responded with a score of 0 to 6, the NPS would be -100. In industry, a positive score is well regarded, and scores over 50 are thought to highlight good performance.

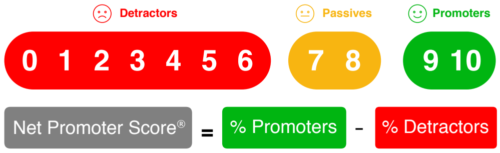


FIGURE 2.7: Net Promoter Score® calculation.

Critique Against the Use of NPS

While many companies around the world use Net Promoter Score (NPS), there are some who are not convinced of its usefulness. Grisaffe, in his paper *"Questions about the Ultimate Question: Conceptual Considerations in Evaluating Reichheld's Net Promoter Score (NPS)"*, agrees that NPS can be an indication of a company's overall state. However, he also points out that NPS has its weaknesses. He questions Reichheld's definition of loyalty and criticizes the idea that NPS is 'the one' number that companies should increase. According to Grisaffe, NPS is *"just a number among others"* (Grisaffe, 2007).

While we concur with Grisaffe's perspective that NPS can signal dissatisfaction among mobile network service provider customers, we recognize that it alone cannot offer a complete picture. Rather, it must be supplemented with actionable insights. The crux of the matter is as follows: "What is the source of subscriber discontent with *MNO Company X's* offerings?" This query cannot be resolved solely by relying on NPS. Our belief is that NPS is as frequently utilised because it is presently the most effective alternative, owing to its superior response rate and, as a result, the greater sample size that may be gleaned from it.

2.4 Literature Review

Reviews covered in this section include published research that either focused on determining the driving factors behind *Net Promoter Score* (NPS) or aimed to predict NPS – using machine learning techniques and input data from a telecommunications environment. Although the NPS is reportedly a key measure of customer loyalty for thousands of companies, only a few published papers specifically examine NPS in this particular setting. In addition, a significant criticism of the NPS metric is its failure to account for the factors driving detraction. This knowledge gap suggests a need for further research in this area.

Tong, Wang, Wen, and Li (2017) highlighted the potential benefits of using data mining techniques to analyse NPS data to gain insights that can help improve customer loyalty in the telco industry. The study was conducted in a Chinese telecommunications company, and the authors used NPS data to identify customer satisfaction and loyalty levels. Data mining techniques, such as *eXtreme Gradient Boosting* (XGBoost) and *Information Gain* (IG) association rule mining, were applied to analyse data and to identify the critical factors related to customer consumption that influence customer loyalty, such as the call duration, volume of data used and Average Revenue Per User (ARPU). Data sources used for the research included five distinct categories: *NPS* *feedback*, *customer attributes* combined with *tariff*-, *usage*- and *behaviour*-related data. The study also found that data mining techniques like k-Means clustering could help telcos to identify *customer segments* with different loyalty levels, and that targeted strategies can be developed to improve loyalty for each group.

Overall the key objectives of the research by Tong et al. (2017) has very little overlap with the current project, with the key differences being that network performance data was not included as a source, and low-complexity linear modelling techniques like logistic regression were not considered. In addition, the analysis of the key driving factors behind NPS was positioned from a different perspective, i.e. to identify the features that impact the relationship between *good NPS* and *financial performance*, instead of the driving factors behind *negative NPS* related to *network factors*, which is our present focus.

The research paper by Ickin, Ahmed, Johnsson, and Gustafsson (2019) discussed the necessity for MNOs to estimate user perceived quality of service by using Key Performance Indicators (KPIs) measured in different components of mobile telecommunication networks. The authors suggest that machine learning can potentially be used to first develop a good estimator of poor Quality of Experience (QoE), and subsequently to identify key factors contributing to poor performance.

Prior research titled *"Bridging the Gap between QoE and User Engagement in HTTP Video Streaming"* (by Moldovan and Metzger, 2016), which showed correlation between service usage related to network KPIs and QoE; in combination with NPS as measure of customer loyalty (by Reichheld, 2003), were used as basis for the research by Ickin et al. (2019). The authors recognise that there is a clear need for a model that is able to map objective network KPIs to user perceived experience scores, and that an understanding of the internal logic of a model is paramount in the area of QoE. Reference is made to prior studies that have shown *decision tree* machine learning models to be a viable option for addressing this requirement (Casas, D'Alconzo, Wamser, Seufert, Gardlo, Seufert, Tran-Gia, and Schatz, 2017). The authors also recognised that it is often challenging to obtain both accurate and human-interpretable machine learning models, meaning that a simple decision tree model might not model QoE as accurate as a more complex "black-box" model, like a neural network.

In consequence, Ickin et al. (2019) followed an approach that addresses the trade-off between the accuracy and interpretability of models, based on a comprehensive prior survey of various model distillation techniques presented by Cheng, Wang, Zhou, and Zhang (2017). They utilised a two-step *teacher/student* model, as introduced by Hinton, Vinyals, and Dean (2015), to extract the internal logic of a complex machine learning model. The authors first developed a complex *teacher* machine learning model to map subjective NPS feedback to KPI metrics – and then trained a more interpretable *student* model supervised by the teacher model. The second step involves extracting the rules and essential features of the "distilled" student model. The authors claim that this method improves accuracy by at least 10% compared to conventional direct training of a decision tree model. Four types of models were trained and compared with a fifth baseline model, representing random classification, using a 70% and 30% training- and test-set, respectively. Model performance was evaluated using *precision*, *recall* and *F1-score*. The four types of models compared included: (1) *Decision Tree*, (2) *XGBoost*, (3) *Random Forest*, and (4) *Neural Network* (NN); by using *Python scikit-learn* and *TensorFlow/Keras* libraries.

The study by Ickin et al. (2019) involved a large-scale analysis of network KPIs and NPS data from a mobile telecommunications provider in Sweden, which mainly offered 4G packet-switched data services at the time. The research utilised input data that consisted of 30 000 *NPS responses* collected via SMS over three months. Additionally, the study used daily aggregations of network KPIs, which were categorised into five groups including (1) *utilisation*, indicating physical radio resource load in a cell; (2) *availability*, indicating the percentage of cell resources being in service; (3) *down-link throughput* indicating data throughput to a network cell; (4) *up-link throughput* indicating data throughput from a network cell; and (5) *total data volume* consumed through a network cell. KPIs were collected over 30 days for each network cell, and the data across multiple cells were aggregated into clusters – each cluster potentially including tens of network cells. Seven aggregation methods were used to summarise the data, including the *mean, minimum, maximum, standard deviation, 1st percentile (1p), median* and *99th percentile (99p)* of each KPI in each cell cluster. The final dataset consisted of 480 observations of 1 050 features (5 KPI/aggregation × 7 aggregations/day × 30 days), and excluded 225 samples due to missing values.

The detailed model distillation procedure used by Ickin et al. (2019) to obtain the student model, was as follows: (1) A robust teacher model was trained with high accuracy as objective – without imposing any constraints on model complexity. The final teacher model consisted of an ensemble of the XGBoost and Neural Network models, with weights of 0.3 and 0.7, respectively – the authors claim that this combination had slightly reduced accuracy, but with reduced variance as benefit. (2) An inference task was executed using the teacher model and the same set of 70% training data used to create the model. (3) A student decision tree regression model, with a depth of 8 layers, was trained using the 70% training dataset as input, and the *class probability* of the "stacked" ensemble as target variable. The authors state that the difference between training a decision tree with actual class labels, versus training it with the estimated probabilities of the stacked model, is that in the latter case the probabilities capture the *relative difference* between similar samples. The student model accuracy was evaluated on the 30% test set after converting the predicted probabilities back to class labels. (4) Finally, the authors extracted the rules and important features from the student decision tree by using the frequency of indicative features extracted throughout cross validation evaluations. To determine which days of the month were having the greatest impact on low NPS scores, the researchers extracted the day number of the features that were deemed significant.

In summary, the study by Ickin et al. (2019) involved training of several machine learning models using a telco dataset, including XGBoost, Random Forest, Neural Network, and a Decision Tree. The more complex models outperformed the simpler decision tree, and all models performed better than a random baseline model. A stacked ensemble, combining XGBoost and Neural Network models, performed the best and had lower variance. The student decision tree, trained on estimated probabilities from the stacked model, outperformed the conventional decision tree by at least 10%. The student decision tree identified essential features, with median uplink throughput values within the last few days of the month being the most significant. The top features showed an increasing correlation with NPS towards the end of the month, indicating a recency effect. Additionally, cell clusters predicted to have poor NPS scores, had lower data volume network traffic per site, which aligned well with the previous research on user engagement and QoE by Moldovan and Metzger (2016) (as cited by Ickin et al., 2019). Hence, the approach taken by Ickin et al. varies significantly from the present project on a number of aspects:

1. NPS observations were associated with- and aggregated to cell level – through a mechanism that is not clear from the published paper. In our opinion it is very difficult to associate mobile subscribers to specific cells, which are usually fixed stationary geographic tower locations within a cellular network. The setting, in which Ickin et al. (2019) performed their research, may have allowed associating subscribers to fixed cell locations, based on the type of service predominantly offered by the MNO, which in the authors' case was packet-switched data services. In the present project we utilise subscriber level data from probing, and we don't associate subscribers to any particular cell.

2. Network services included by the research of Ickin et al. covered packet-switched services and metrics – measured per cell and aggregated to cell cluster level. In the present project we utilise subscriber level CEI metrics covering multiple services, including *voice*, *SMS* and *packet-switched signalling*, over and above *packet-switched user-plane* metrics as perused by Ickin et al.

3. Although Ickin et al. tested a range of models including both complex as well as basic models, they did not cover any of the classic modelling approaches like logistic regression. We find the *teacher/student* approach they take as an innovative approach to understanding model output that we might consider employing in future research. In the current project we utilise stand-alone modelling techniques, ranging from basic models, like the classic logistic regression model, through complex models, like Support Vector machines and XGBoost, and we instead use modern explanatory modelling techniques like *Partial-dependence Profiles* to interpret model output.

Chapter 3

Data Sources and Preparation

This chapter discusses the different data sources and the process of collecting, cleaning, and preparing them for modelling and analysis using machine learning techniques. In the context of the current project, our primary goal is to address a supervised learning problem of classification. We aim to achieve the following objectives:

1. Determine if any common data patterns connect subscribers' subjective responses from NPS satisfaction surveys to their objective customer experience measurements obtained through probing cellular network interfaces. Specifically, we want to predict the likelihood of a subscriber being an *'NPS detractor'* based on their experience metrics collected from the cellular network interfaces.
2. If there are identifiable patterns that accurately identify NPS detractors, we want to identify the key factors that contribute to negative sentiment.

The data for this project was obtained from a prominent Mobile Network Operator (referred to as Telco from here on) in South Africa. It consists of 627 112 individual NPS survey responses gathered from subscribers over six months, specifically from October 2018 to March 2019. The survey data was merged with network experience measurements obtained from network probes during the same timeframe. To ensure the protection of subscribers' personal information, the data were de-identified by removing mobile telephone numbers as soon as datasets from respective sources were merged.

3.1 Main Data Sources

As input to the project, three primary data sources from two business areas within Telco were identified. The data sources included: *subscriber attributes*, survey *responses* from an internally managed satisfaction improvement campaign, and *experience measurements* collected by probing of Telco's cellular network interfaces. A summary of the data sources is briefly outlined below:

Data from the **customer analytics** environment:

1. Survey data:
   - Network NPS (NNPS) campaign responses – target variable.
2. Subscriber information:
   - Market segment attributes.
   - Demographic attributes.

Data from the **network management** environment:

3. Customer experience measurements from probed network interfaces related to:
    - Circuit-switched control-plane services like *voice calls*.
    - Packet-switched control-plane services related to *data session establishment*.
    - *Application* use, e.g. web browsing or video streaming – mobile data services transporting information via the packet-switched user-plane.

3.2 Customer Analytics Data

The customer analytics dataset includes Network NPS (NNPS) survey responses enriched with subscriber attributes, providing valuable segmentation and demographic information about subscribers. Telco obtains the NNPS data through an ongoing internal process, described next.

NNPS Surveys

NNPS surveys are sent daily to a randomly selected sample of customers representative of the entire subscriber base. The size of the daily sample is dimensioned in such a way that the whole subscriber base can be surveyed once a year. The survey is also governed by rules around spamming and adherence to the Protection of Personal Information Act 4 of 2013 (South African Government, 2013) to protect customers.

The survey is conducted through text message and consists of two steps. During the first step, an introduction message is sent to inform target subscribers about the study and to request participation. A subscriber has the option to opt-out or to participate by replying with a free SMS. The introduction message typically has the following wording:

"Hello, this is a message from Telco. We would like to hear about your recent experience using our network. Reply '1' to continue and share your feedback with us free of charge, or reply with '0' to opt-out."

Non-responders are automatically excluded from the survey, and only if a subscriber responds with a *'1'* will they receive the survey question. The next part of the survey will then proceed with a typical NPS question:

"Based on your recent experience using our mobile network, how likely are you to recommend Telco to friends/family? SMS a score from 0 (not at all likely) to 10 (extremely likely)."

The NNPS score value provided by a customer, a number between 0 and 10, is recorded in the customer analytics database under the field name `NNPS_LTR` or *Likelihood to Recommend* (LTR). During the modelling process, the LTR score will be interpreted in the same fashion as common NPS; scores from 0 to 6 are labelled *'Detractor'*, 7 or 8 as *'Neutral'*, and 9 or 10 as *'Promoter'*.

Approximately two million customers are surveyed by Telco through this process every month. In determining the number of target subscribers, we consider the number of subscribers who have opted out and are excluded in adherence to the PoPI Act (South African Government, 2013), as well as those who were excluded due to spamming rules limiting surveys to one per 365 days. The number of completed survey responses is usually in the order of 100 000 samples per month, representing a relatively low response rate of only 5%. However, the sample obtained from the in-house survey is still much larger than the official NPS survey – mainly due to the high cost of externally managed surveys, which limits the number of samples that can be sourced using third parties. A self-managed electronic process is much more cost-effective and, therefore, even with the low response rate, still a much more efficient option to collect subjective feedback from subscribers about their experience.

Subscriber Information

As mentioned previously, the analytics dataset includes NNPS LTR scores, subscriber segmentation, and demographic information. This enriched dataset enhances the survey data by incorporating Telco organisation-specific attributes, providing additional details about subscribers. The added features enable analysts to group survey results of similar subscribers together during reporting and analysis, giving more context than the uni-dimensional LTR scores alone. For instance, analysts can present stakeholders with an aggregated regional perspective of survey results to identify geographical areas with lower subscriber sentiment than others. Our prediction model will utilise these segmentation attributes as additional input variables. A comprehensive list of all available subscriber attributes is listed in Table B.1, and a summarised list of the most relevant features used during the present analysis is listed in Table 3.1.

Attribute	Database field name	Description
NNPS Response	`IF_NNPS_STATUS`	The NNPS survey response category of the subscriber. Values: *{'Promoter','Neutral','Detractor'}*
Payment method	`IF_PMT_METH_MED`	The payment method of the subscriber. Values: *{'Postpaid','Prepaid'}*
Network region	`IF_NETWK_REGN_NM`	The geographic network region of the subscriber. Values: *{'Central', 'Eastern', 'Kwazulu Natal', 'Limpopo', 'Mpumalanga', 'Northern Gauteng', 'Southern Gauteng', 'Western'}*
Device type	`IF_DVC_TYPE_GRP_NM`	The type of device used by the subscriber. Values: *{'Smart Phone', 'Feature Phone', 'Tablet', 'Basic Phone'}*

Table 3.1: A list of the subscriber attributes from the customer analytics dataset that are relevant to the present project.

Analytics Data Flow

The complete process of collecting and preparing the analytics dataset is illustrated in Figure 3.1. An automated system manages daily survey distribution and subscriber response gathering. Before storing the data in the analytics database, the survey data is enriched with subscriber information obtained from the customer data warehouse. Once stored, end-users like analysts or data scientists can retrieve the data from the analytics database for reporting and analysis purposes using *structured query language* (SQL)[1].

The mobile subscriber identifier number (MSISDN) field is used in anonymised format to join the analytics dataset to the experience measurements from probing sources before it gets dropped from the data to protect personal information.

[1] *Structured query language* (SQL) is a programming language for storing and processing information in a relational database.

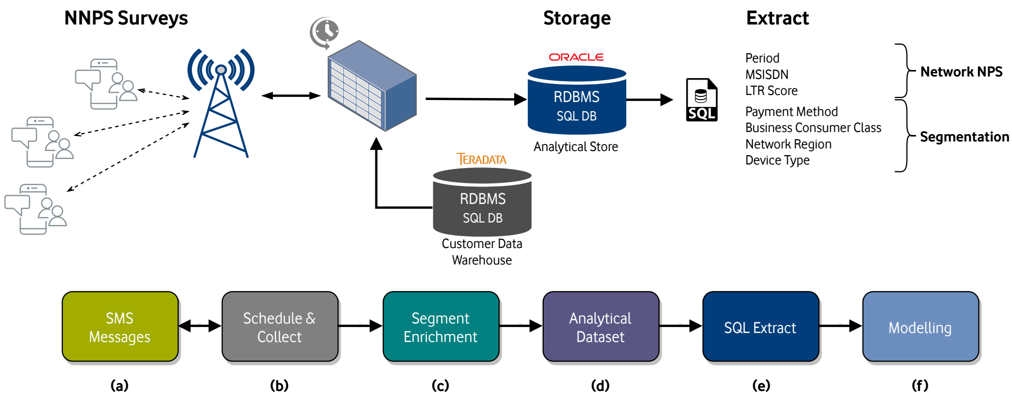


FIGURE 3.1: Process flow of the analytics data. From left to right: (a) Distribution of NNPS surveys to subscribers via SMS. (b) Daily automated scheduling for distribution and collection of surveys. (c) Enrichment of survey responses with subscriber information from the customer data warehouse. (d) The enriched survey data records is stored within the analytical database. (e) *Structured query language* (SQL) is used to extract the analytical dataset for use within modelling. (f) The analytics dataset is combined with probing data and prepared for modelling.

3.3 Customer Experience Data

The capture, interpretation, aggregation and storing of *Customer Experience Indicators* (CEI) from network interfaces involves network probes and a Customer Experience Management (CEM) application, as discussed in Section 2.1. The CEM application backend utilises a large-scale database system where various aggregated groupings of probing metrics are stored for analysis and reporting. Telco employed two different CEM applications at the time of this project, which resulted in some challenges and compromises that needed to be made in terms of how subscriber-level experience could be collected and summarised in a format suitable as input to machine learning algorithms. One of the applications covered only the *control-plane* – including services like voice, sms and data-bearer control, whereas the second application only covered the *packet-switched user-plane* – including services like web browsing, social media and video streaming. Each service entails a distinct set of experience measurements and varying aggregation rules to be considered when summarising CEIs. Database technologies were also vastly different between the two systems, ranging from a structured relational Oracle™ database for control-plane data to a Cloudera™ Hadoop *big data* implementation for the user-plane data. Consequently, 11 datasets from 8 distinct interfaces – described in Section 2.1, had to be collected from 2 different platforms and combined through complex preprocessing to provide a thorough set of customer level metrics that are comprehensive enough for modelling of customer satisfaction as a function of mobile cellular service quality.

Attribute Level Aggregation

The aggregation of experience metrics is done for each cellular service, for example, *voice*, and for each network generation. The network generation is determined by the type of network interface, for example, the *A* interface that carries *2G* voice or the *IuCS* interface in the case of *3G*. Experience metrics are further grouped into distinct logical levels representing the entities involved during the communications. The level at which aggregation is done depends on the type of analysis or reporting the data will be used for. For example, metrics can be compiled at an *MSISDN* level, which allows individual subscriber experiences to be reported. Similarly, metrics can be aggregated at the cellular level. This enables the reporting and analysis of metrics for a specific geographical coverage area provided by single or multiple network cells. We focus on aggregating metrics at the subscriber (MSISDN) level, as we need to merge this data with the analytics dataset (described in Section 3.2) one-to-one as input for the current project. Figure 3.2 provides a summary architectural diagram of the mobile telephone network and it indicates the eight probed *Core Network* (CN) interfaces relevant to this project – already discussed in detail in Section 2.1.

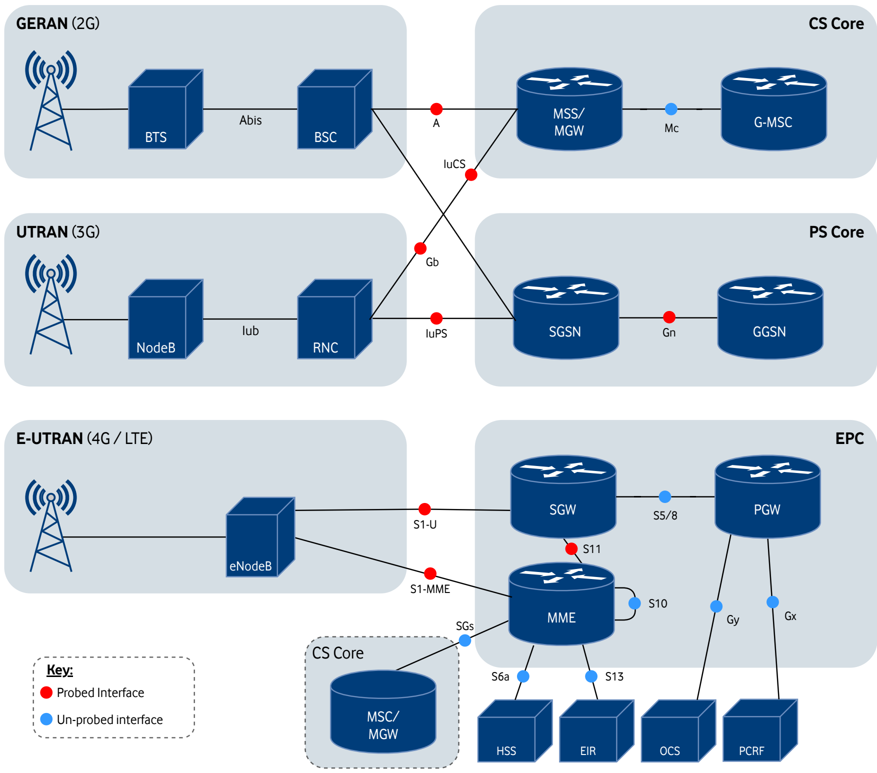


FIGURE 3.2: A summarised architectural diagram of a mobile telephone network indicating the core network (CN) interfaces between nodes we are probing for collection of customer experience indicators used in the present project. See Table 2.1 for more detail about the function of each interface.

Temporal Aggregation

Metrics are also aggregated *temporally* into various discrete time intervals that allow granular reporting over short periods – with detail down to a 5- or 15-minute level, or over extended periods – with daily or weekly granularity. After considering limitations of system storage capacity and different historical data retention policies between the two CEM platforms, this project's only common long-term aggregations of subscriber-level data available for analysis were at the *weekly* aggregation level. Weekly summarised experience metrics follow the *"ISO8601-1:2019 Date and time – Representations for information interchange – Part 1: Basic rules"* standard (International Organization for Standardisation, 2019) and range from *Monday* through *Sunday*. The approach to aligning the weekly aggregated probing experience metrics most closely with the daily survey data is described next.

Extraction Windows

Subscriber surveys are continuously collected daily, through the process outlined earlier in Section 3.2. Based on the assumption that a subscriber's response to the NNPS survey may be influenced by the time at which they are surveyed – as recent network experiences may negatively or positively impact a subscriber's view – a data alignment mechanism was implemented in order to ensure that the *weekly* probing measurements were as close to the actual date of the survey as possible.

The approach used was to collect historical data of a *four-week* period prior to the survey response date for each subscriber. However, the starting Monday of the four weeks was shifted by one week later for surveys responded to on Thursday through Sunday. The motivation behind the chosen approach was to reduce missed experience measurements in close proximity prior to a survey response. For example, Figure 3.3 shows the extraction windows for subscribers that responded to NNPS surveys during the week from 2018/10/01 to 2018/10/07.

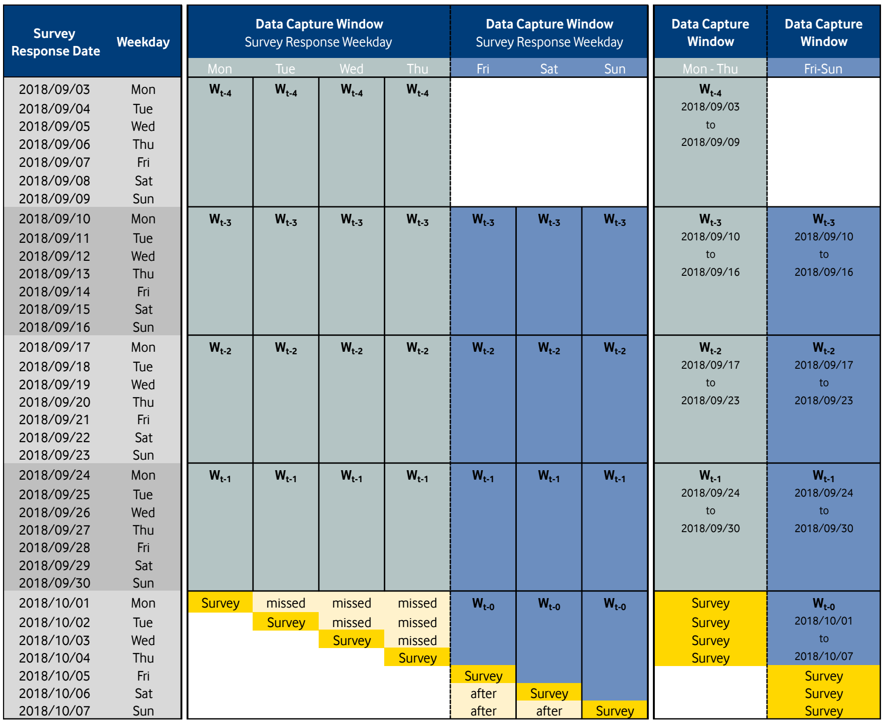


FIGURE 3.3: Probing data extraction temporal windows are driven by the survey response date of each subscriber.

Metric Aggregation

Data collected for each subscriber based on the specified time window approach could encompass any number of weeks, from one week to a maximum of four weeks' data, depending on actual subscriber activity. The weekly granularity data was aggregated into single records per subscriber, service and network radio type by summating each numerical CEI counter. CEI metrics used for modelling are weighted average values based on formulas combining the counters, for example, the *Call Setup Fail Rate* consisting of *Call Setup Failure Count* divided by *Call Attempt Count*. In the case of metrics indicating total consumed volumes, for example, the *Total Call Duration*, we normalised the values by dividing it by the number of weeks that data existed for each subscriber, resulting in the *Average Weekly Total Call Duration*. Scaling the metrics prevents excessive bias towards subscribers with significantly higher call duration due to more consistent regular use.

Customer Experience Data Flow

The end-to-end flow of customer experience data processing taking place behind this project – from the initial raw data collection originating at probes through interpretation, aggregation, correlation, storage and extraction for further post-processing and modelling – is illustrated in Figure 3.4. As already mentioned, two distinct CEM systems were used as sources of network experience data. The data flow for control-plane interfaces managed by the one application is shown in the top half of the figure, while the user-plane flow of data from interfaces covered by the second application is shown in the bottom half.

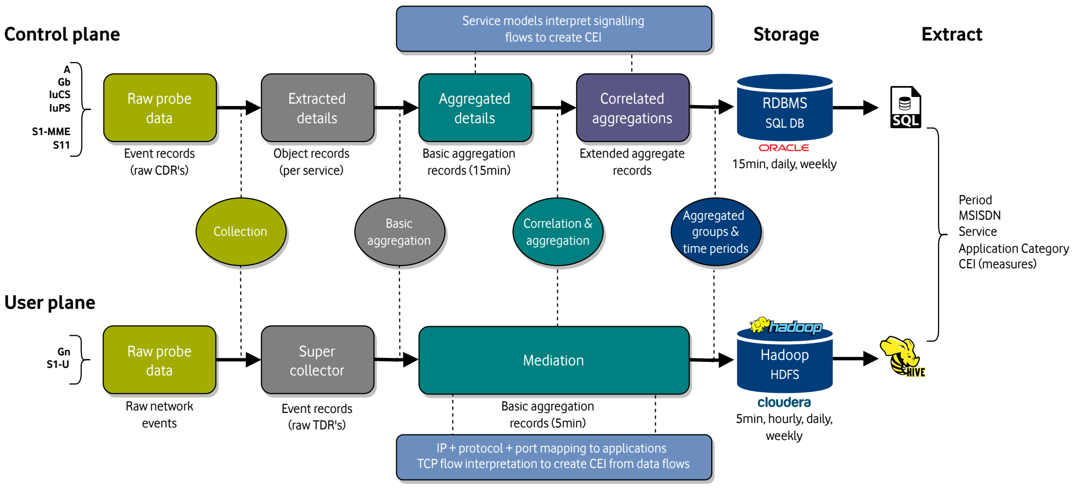


FIGURE 3.4: Process flow of the customer experience data. Raw data is collected from network interfaces by passive probing. A Customer Experience Management system manages the collection, aggregation, correlation and interpretation of raw events from probing into interpretable Customer Experience Indicators that are stored in a database for use by analytical applications or statistical modelling down-stream. The present project utilised SQL scripts to extract metrics for a specified list of MSISDN that were part of the NNPS survey respondent cohort.

Among the significant differences between the data flows of the two applications are:

- The type and structure of data collected through probing.
- Processing of raw data into interpretable metrics describing the user experience of each service.
- The discrete periods and rules employed to aggregate data.
- Database technology utilised by each application to store the data.
- Both applications utilise SQL as the extraction method, although Hive and Oracle SQL differ slightly in their dialects.

SQL scripts were used at the end of each calendar month to extract the CEIs of subscribers that formed part of the recent NNPS respondent cohort during the corresponding period. The process was repeated each month for the six months of data collection for the project and extraction windows following Section 3.3 logic were used to determine the activity period to collect data of each subscriber.

3.4 Feature Engineering and Combining Datasets

The individual data sources we have discussed must be combined into a single uniform structure to be suitable as input to machine learning or statistical models. The required format is one row of data per subscriber, containing the associated independent predictors and the dependent outcome variable. Regarding the present project, we need to create a single dataset that combines the profiling and survey information described in Section 3.2 with the mobile network customer experience (CE) metrics of Section 3.3. A common anonymised identifier variable, the *MSISDN*, which uniquely identifies a subscriber in any mobile cellular network, was included in all sources during data collection and used exclusively for the purpose of joining datasets. The end result is a modelling dataset with structure, as shown in Figure 3.5. For ease of discussion, we will refer to a single row of data from this dataset as a *Subscriber Modelling Record* from this point forward.

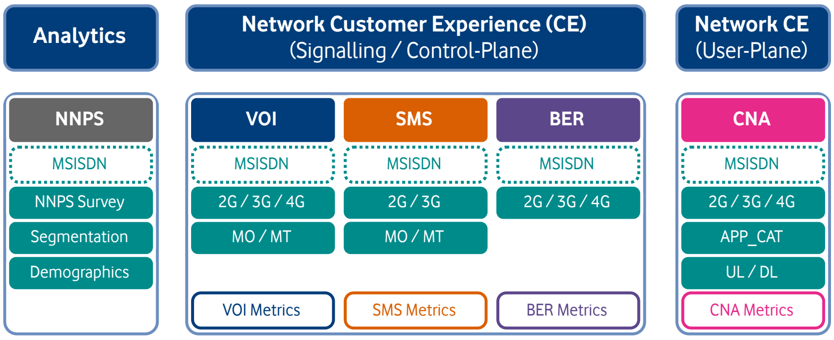


FIGURE 3.5: High-level structure of the modelling data used in the present project. The data consists of a single row of data per subscriber, containing the independent variables from the analytics dataset combined with the experience metrics from each mobile network service and the dependent outcome variable collected through the NNPS survey. The customer experience metrics were sourced from two independent CEM applications – one covering the control-plane services and the other covering the user-plane services. All datasets have a common anonymised identifier, the `MSISDN`, which serves the purpose of joining individual sources.

Sparse Mobile Network Data

The basic joining of each dataset on a common field is trivial. However, during the Exploratory Data Analysis (EDA) covered in Chapter 4, we found gaps in our data after doing the joins using data in its original raw format. The gaps will result in issues during modelling if the models used are not robust against incomplete data. Within the spectrum of classification models, some algorithms can function with incomplete data, for example, Classification Trees. Still, for others, like Logistic Regression, the presence of missing values in the data will impact the robustness of the model. Therefore, we must explicitly handle missing values during the data preprocessing steps. Incomplete records are generally discarded during modelling or circumvented by using imputation methods. In our scenario, the gaps are too significant for effective modelling using either of these methods. To address this issue, we use methods like aggregation and filtering or grouping of sparse categorical variable levels to restructure the data into a more forgiving format that can be used across multiple model types. Two primary reasons for incomplete modelling records have been identified – these are linked to differences in users' communication needs and usage patterns. In this regard, let's discuss the common scenarios observed using Figure 3.5 as a reference.

Service-Level Gaps

The first scenario relates to which specific mobile services are used by each subscriber. Differences in service usage can significantly impact the completeness of modelling records at the *entire service level.* For instance, gaps can be caused by including data of subscribers that only make voice calls and never use data during the observation period. Such subscribers' modelling records will have no metrics reflected within the SMS, BER or CNA services. Similarly, subscribers who solely use mobile data will have metrics captured within the BER and CNA services but no measurements within VOI and SMS.

Within-Service Gaps

The second scenario relates more to how and where subscribers use a particular service. There are subtle differences between the various service types, and the levels available within each service's attributes may be slightly different, but underlying the common scenario is that incomplete data will be *isolated to a particular service.* For example, in the case of VOI and SMS services, there is a *directional* component associated with the metrics that indicate whether a subscriber initiated or received the signalling for the communication. Directionality is reflected by the mobile originating (MO) and mobile terminating (MT) attributes. When a subscriber only ever makes calls during the measurement period, the MT attribute will have incomplete experience measurements. Regarding packet-switched services, we have similar uplink (UL) and down-link (DL) attributes to indicate directionality. Fortunately, gaps are less likely to occur in data services due to the bi-directional nature of a UL user request towards an application server, usually accompanied by a DL response back to the user.

At a *technology level*, it is possible to define attributes that characterise the radio types used for each service, such as 2G, 3G, or 4G. It is worth noting that during a measurement period, certain subscribers may not be active on all radio technologies due to network coverage or handset limitations. Consequently, there may be gaps in the data pertaining to radio-type attributes within some subscriber records. In the case of the bearer (BER) service, additional metrics are available that describe user experience; however, their scope is limited to the $4G$ technology. For subscribers who only use older-generation technologies, by default, these technology-specific metrics will be missing.

Finally, within the CNA user-plane dataset, an *application category* attribute serves to indicate the nature of the applications that a given subscriber employs. While some subscribers utilise a limited number of applications, such as only using messaging apps, others may use a more extensive list of application types, encompassing web browsing, security and software updates, file transfer, email and streaming applications. Notably, subscribers who use fewer application categories are likely to present gaps in their experience measurements compared to those who use a more extensive list of application types.

We make use of one commonly used service, voice, as a general example to demonstrate the data gaps that typically result when data from all underlying sources are combined into a single wide-format modelling dataset.

Voice Service Example: According to Table 2.1, when it comes to representing voice as a service, it is necessary to collect probing data from three interfaces: *A*, *IuCS* and *S1-U*. This data includes a subscriber identifier (`MSISDN`), a radio-type attribute, and voice metrics. Fortunately, voice experience metrics obtained from the CEM application used by Telco are similar for all radio technologies, and we can easily combine the individual sources into a comprehensive overview of voice performance across all radio technologies by performing a simple union of the data. As a result, the raw data for voice is consolidated into a *"long"* format, as demonstrated in Table 3.2. There can be anywhere from one to three records per subscriber, based on the number of sources where the subscriber appears.

Subscriber$_{ID}$	Radio_Type	VOI.CEI$_1$	VOI.CEI$_2$	...	VOI.CEI$_p$
S_1	2G	S_1.2G.CEI$_1$	S_1.2G.CEI$_2$	...	S_1.2G.CEI$_p$
S_1	3G	S_1.3G.CEI$_1$	S_1.3G.CEI$_2$	...	S_1.3G.CEI$_p$
S_2	2G	S_2.2G.CEI$_1$	S_2.2G.CEI$_2$	...	S_2.2G.CEI$_p$
S_3	3G	S_3.3G.CEI$_1$	S_3.3G.CEI$_2$	...	S_3.3G.CEI$_p$
S_4	2G	S_4.2G.CEI$_1$	S_4.2G.CEI$_2$	...	S_4.2G.CEI$_p$
S_4	3G	S_4.3G.CEI$_1$	S_4.3G.CEI$_2$	...	S_4.3G.CEI$_p$
S_4	4G	S_4.4G.CEI$_1$	S_4.4G.CEI$_2$	...	S_4.4G.CEI$_p$
S_5	3G	S_5.3G.CEI$_1$	S_5.3G.CEI$_2$	...	S_5.3G.CEI$_p$
S_5	4G	S_5.4G.CEI$_1$	S_5.4G.CEI$_2$	...	S_5.4G.CEI$_p$
...	...	...	...	...	...
S_N	2G	S_N.2G.CEI$_1$	S_N.2G.CEI$_2$	...	S_N.2G.CEI$_p$
S_N	3G	S_N.3G.CEI$_1$	S_N.3G.CEI$_2$	...	S_N.3G.CEI$_p$
S_N	4G	S_N.4G.CEI$_1$	S_N.4G.CEI$_2$	...	S_N.4G.CEI$_p$

TABLE 3.2: Example of the *"long"* data structure for voice customer experience indicators (CEI) obtained from the CEM application. The dataset has multiple records per `Subscriber-Id`, one for each radio technology where the subscriber was active, and each column represents a voice customer experience metric.

The next step in data preparation involves converting the extended layout depicted by Table 3.2 into a *"wide"* format to obtain a voice modelling dataset. The *"wide"* format is obtained by transposing the voice metrics into three *radio-type-metric* blocks, which are positioned next to the subscriber identifier so that we end up with a single record per `Subscriber-Id`. The radio-type attribute is used to differentiate between separate metrics per technology. As a result, we obtain a *"wide"* data structure, as depicted in Table 3.3.

Subscriber$_{ID}$	VOI.2G.CEI$_1$	VOI.2G.CEI$_2$	...	VOI.2G.CEI$_p$	VOI.3G.CEI$_1$	VOI.3G.CEI$_2$	...	VOI.3G.CEI$_p$	VOI.4G.CEI$_1$	VOI.4G.CEI$_2$	...	VOI.4G.CEI$_p$
S_1	S_1.2G.CEI$_1$	S_1.2G.CEI$_2$	...	S_1.2G.CEI$_p$	S_1.3G.CEI$_1$	S_1.3G.CEI$_2$	...	S_1.3G.CEI$_p$			...	
S_2	S_2.2G.CEI$_1$	S_2.2G.CEI$_2$	...	S_2.2G.CEI$_p$			...				...	
S_3			...		S_3.3G.CEI$_1$	S_3.3G.CEI$_2$	...	S_3.3G.CEI$_p$			...	
S_4	S_4.2G.CEI$_1$	S_4.2G.CEI$_2$	...	S_4.2G.CEI$_p$	S_4.3G.CEI$_1$	S_4.3G.CEI$_2$	...	S_4.3G.CEI$_p$	S_4.4G.CEI$_1$	S_4.4G.CEI$_2$	...	S_4.4G.CEI$_p$
S_5			...		S_5.3G.CEI$_1$	S_5.3G.CEI$_2$	...	S_5.3G.CEI$_p$	S_5.4G.CEI$_1$	S_5.4G.CEI$_2$	...	S_5.4G.CEI$_p$
...	...	...	...	...	...	...	...	...	...	...	...	...
S_N	S_N.2G.CEI$_1$	S_N.2G.CEI$_2$	...	S_N.2G.CEI$_p$	S_N.3G.CEI$_1$	S_N.3G.CEI$_2$	...	S_N.3G.CEI$_p$	S_N.4G.CEI$_1$	S_N.4G.CEI$_2$	...	S_N.4G.CEI$_p$

TABLE 3.3: Example of the *"wide"* data structure for voice customer experience indicators after the data shown in Table 3.2 was transposed into a modelling format. The empty yellow blocks represent scenarios where a subscriber did not have measurements on the associated radio technology.

It is important to note the gaps that emerge – marked in yellow – where the metrics for a particular technology are missing. This situation arises again when we partition the data further into MO and MT detail. In mobile telecom networks, incomplete coverage across all levels of the available attributes is quite common. Given the sparsity that occurs when radio technology and call directionality are used for modeling, dropping incomplete records or using imputation methods are not feasible options to address this issue. Instead, we employ the methods described next when transforming the data into a modelling format, to avoid gaps in our data based on the known patterns we discovered during EDA; discussed at lenght in Chapter 4.

Methods for Managing Sparse Mobile Network Data

Aggregation of Metrics

To avoid *within-service* gaps, we use metric aggregation by summating each counter variable across the categorical attribute levels of *radio technology* and *directionality*. All CEI metrics were extracted in their raw form during data collection, meaning that the underlying counters were used instead of already calculated weighted average values. For example, `DROPNETWORKRATE` consists of `DROPNETWORK` and `CONNECTSUCCESS`. We have kept all the base counters in our data, per technology and direction, and having them in raw form allows us to calculate each CEI metric *value* as a last step – once the data is aggregated up to the point that we are confident that no more gaps will result once we move to the modelling format. Mathematically the aggregation of `DROPNETWORKRATE` can be expressed as:

$$\begin{aligned}\texttt{DROPNETWORKRATE} &= \frac{\texttt{DROPNETWORK}}{\texttt{CONNECTSUCCESS}} \times 100 \\ &= \frac{\sum_{i\in\{\texttt{MO},\texttt{MT}\}}\sum_{j\in\{\texttt{2G},\texttt{3G},\texttt{4G}\}}(\texttt{DROPNETWORK})_{i,j}}{\sum_{i\in\{\texttt{MO},\texttt{MT}\}}\sum_{j\in\{\texttt{2G},\texttt{3G},\texttt{4G}\}}(\texttt{CONNECTSUCCESS})_{i,j}} \times 100,\end{aligned} \tag{3.1}$$

where i iterates the levels within the *directionality* attribute, and j iterates the levels within the *technology* attribute. Similar calculations are used across all services. In Section 3.5 we provide a detailed summary of all CEI metric formulas used.

To allow models to detect the possible effect of radio technology on the outcome variable, we feature-engineer new factor variables that capture the most used radio type for each subscriber, based on the technology where a subscriber has the most service access attempts (`RA_MOST_USED_ATT`) – and we do this separately for each service. We also add a variable that captures the *relative magnitude* of activity against the most used radio type (`RATIO_RA_MOST_USED_ATT`) – to allow models to possibly detect patterns related to subscribers having *'solid'* versus *'interrupted'* radio coverage. For example, a subscriber having `RA_MOST_USED_ATT` = 4G and `RATIO_RA_MOST_USED_ATT` > 0.7 on average has pretty stable 4G coverage during the entire measurement period, whereas a similar 4G subscriber with a ratio of less than 0.6 probably experience a lot of switching between technologies – which can potentially negatively impact their experience. Bear in mind that only *weekly* data was available for the present project, as explained in Section 3.3, and therefore, it was impossible to feature-engineer any variables relying on more granular time measurements – for example, to gauge the *frequency of radio technology changes*, which might have been more descriptive of the user experience.

Similarly, the effect of MO and MT directionality on voice experience was catered to by introducing the `RATIO_MODUR` metric that measures the ratio of MO call to total call duration. The reason behind introducing this metric was to consider that subscribers who make more calls than receive them are also likely to spend more money with their service provider. Such subscribers may have higher expectations than those who primarily receive calls and pay less for their subscriptions and may, therefore, be more prone to respond negatively to their NPS survey than their counterparts spending less, with the slightest deterioration of network experience.

The same aggregation method was applied to the CNA service at an *application-category* level to model the effect of application type on the outcome variable. We identify the most used category based on the number of data packets transmitted for each type of application, giving us the new factor variable `APP_MOST_USED_PACKETS` to use in our models. Similar variables to the ones created for voice, namely `RATIO_MURAT` and `RATIO_MUAPP` was introduced to consider the relative magnitude of the most used technology and application within the CNA service.

Separate Modelling Data

In scenarios with *service-level* gaps, it is necessary to adopt a different approach as the aggregation method is not suitable. This is because the aggregation method requires identical counter definitions across the levels to be combined. Hence, for example, aggregating VOI and SMS counters is infeasible as each service type employs unique mechanisms and metric definitions to evaluate the user experience. Moreover, as explained in Section 3.4, *"Within-Service Gaps"*, we face an unusual situation with the BER service in which metrics differ even between radio technologies. Consequently, we cannot entirely aggregate data across radio types to address the within-service gaps for that particular service. However, by treating *2G-3G* versus *4G* BER as distinct services, denoted as BER_a and BER_b respectively, we can also circumvent the sparseness of modelling data across the 2G-3G and 4G levels by leveraging the following method.

We use *discrete models*, each trained with their individual modelling dataset, to circumvent the problem of incomplete service-level data. The grouping of records for each modelling dataset is based on known and explainable patterns we identified during the EDA process. We group sections of the data where the features are complete at the subscriber modelling record level. In other words, only the records of those individuals who use the same combination of services will be included in a particular modelling dataset, and we subsequently create separate models for each *subscriber usage class*. Figure 3.6 summarises the grouping we performed to obtain the individual modelling datasets.

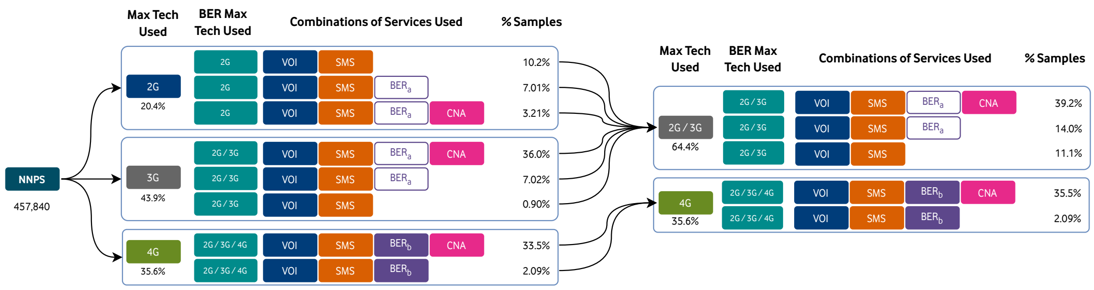


FIGURE 3.6: The grouping of similar subscriber modelling records is based on the combination of services and the maximum radio technologies individuals use. On the left-hand side, datasets are shown separately for each *max technology used* category of subscriber usage class. On the right-hand side, we illustrate the final combined levels of subscriber modelling records after grouping data by similar service usage patterns.

As part of our current analysis, we aim to assess the impact of incorporating features from each service on the accuracy of our predictive model. To accomplish this, we have meticulously selected ten distinct datasets for training models of varying complexity to predict the NPS classification of subscribers. Each dataset is centered on a particular group of subscribers who have utilised a specific set of services.

Figure 3.7 shows the composition of the ten datasets we isolated for the analysis, indicating which services are included in each. The least complex modelling data, containing only a single service's metrics as features, are represented by the first five datasets at the top of the list – up until dataset `MOD_M5_CNA`. The following five datasets, from `MOD_M6_VOI-SMS` through `MOD_M10_VOI-SMS-BERb-CNA`, each contain incremental additions of features from the remaining services, gradually widening the explored feature space for models to train from. The first multi-service model, `MOD_M6_VOI-SMS`, combines VOI and SMS datasets, representing the most common services used by the most significant proportion of our survey responder group. Moving down the list, we incrementally add more services until we have a dataset containing all available services.

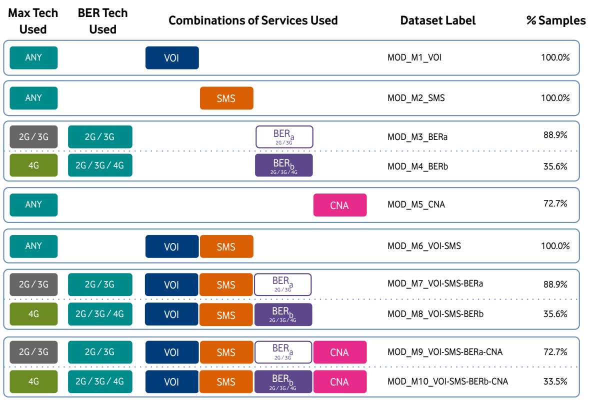


FIGURE 3.7: The ten datasets that we used for modelling in our analysis. Each dataset contains different combinations of services from our probing data. The purpose is to create a list of modelling datasets that represent the typical patterns we encounter. This will enable us to assess the significance of including data from a specific mobile service, both in isolation and when combined with other services.

The implication of gradually adding network data, one service at a time, allows for thorough testing of the value of each dataset. Consequently, we can establish which services are most valuable to include in a machine learning pipeline to predict NPS with a certain amount of accuracy, still to be determined. The benefit to MNOs will be that capital expenditure on expensive probing solutions can be scaled according to the level of accuracy they aim to achieve from this type of use case. For example, an MNO with no probing infrastructure can start by investing in probing only the most valuable network interfaces and then gradually expand on their investment, adding more network interfaces to increase the accuracy.

Filtering of Low-Represented Groups

The final method we use to circumvent sparse modelling data entails filtering small subsets of records based on our findings from the initial exploratory analysis. We cover the identification and discard of the small sample groups at length in Section A.1. In summary, this method – which is also our last resort for avoiding incomplete data, is to discard samples where the number of records representing particular subsets of subscribers is minimal. In a full-production machine learning pipeline, we would ideally have to cater for these scenarios, but to simplify the present project, we removed the data. Overall, we retained 91% of our original dataset by carefully considering each scenario before discarding any samples.

3.5 Predictor Variables

The present study aims to investigate the viability of predicting the customer experience of mobile network users by using objective measurements derived from the mobile network. It is hypothesised that the overall cellular experience of any end-user can be described by a combination of Customer Experience Indicators (CEIs) from multiple network interfaces, depending on the mobile services and applications they use during a particular period. To this end, machine learning techniques will be employed to search across this multidimensional space for common patterns in the data that can be utilised to predict the Net Promoter Score (NPS) for any new non-surveyed subscriber based on their network experience data alone.

At the outset of the project, the key CEIs that are essential to indicate the actual perceived experience of subscribers and, therefore, may play a role in the prediction of NPS were unknown. It is plausible that some metrics may be more indicative of user experience and how subscribers perceive or rate the overall service provided by a particular Mobile Network Operator (MNO) compared to others. As such, machine learning models will be utilised, in combination with techniques like regularisation, to process an exhaustive list of plausible CEIs from probing, to attempt to identify those that are most important in describing the relationship between CEIs and NPS. Figure 3.8 provides an overview of the complete list of variables deemed as essential to include as input for the analysis. The selected list of variables will describe any subscriber's overall mobile network experience and provide the necessary data as input into machine learning models that will be trained to predict NPS.

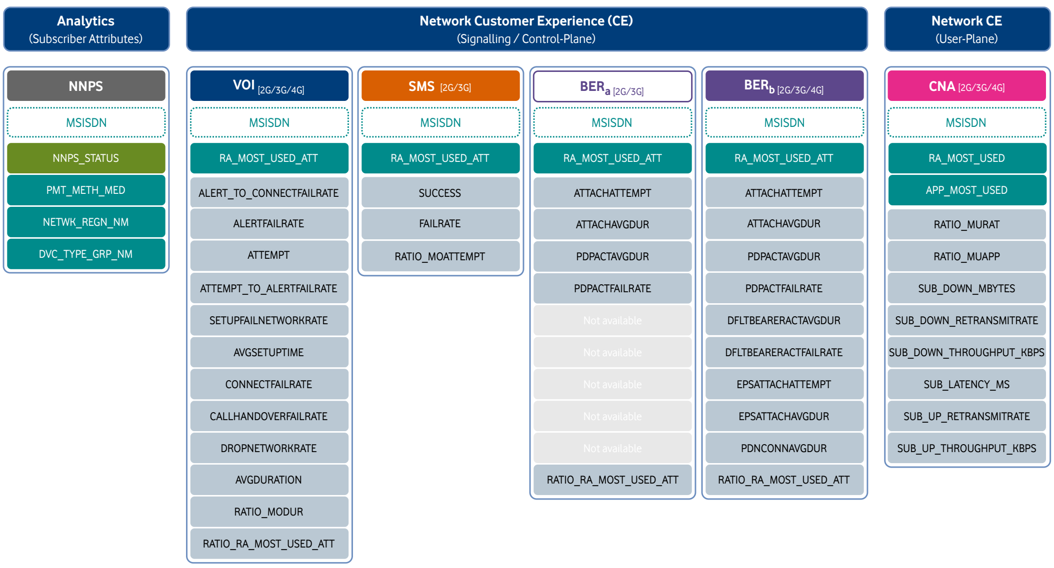


FIGURE 3.8: High-level summary of the complete dataset used for modelling. Each data source is represented by variables consisting of categorical and numerical variables. The analytics source on the left contains the target `NNPS_STATUS` variable, and the remainder of sources represent the various services provided by a mobile cellular network.

The following tables each summarises the calculated CEIs used as predictor variables per network service and describes how they are derived from the underlying basic probing counters. A more comprehensive and detailed description of all data fields is provided in Chapter B.

Table 3.4 shows the Short Message Service experience predictor variables and calculations.

CEI	Type	Description and Formula
SMS_RA_MOST_USED_ATT	Factor	Most used radio access type, based on the radio access type with maximum attempt count of all SMS radio technologies including 2G and 3G. $= \max\left[\mathtt{ATTEMPT}_{\mathtt{2G}}, \mathtt{ATTEMPT}_{\mathtt{3G}}\right]$
SMS_SUCCESS	Count	Total number of successful SMS message send or receive attempts on the 2G and 3G radio layers. $= \sum_{i \in \{\mathtt{MO},\mathtt{MT}\}} \sum_{j \in \{\mathtt{2G},\mathtt{3G}\}} (\mathtt{SMS\_SUCCESS})_{i,j}$
SMS_FAILRATE	Rate	SMS fail rate calculated as the ratio of SMS that failed, versus the total number of SMS attempts. $= \left(1 - \frac{\mathtt{SUCCESS}}{\mathtt{ATTEMPT}}\right) \times 100$
SMS_RATIO_MOATTEMPT	Rate	SMS ratio MO, calculated as the ratio of MO attempts, versus the total number of MO and MT attempts. $= \frac{\mathtt{MOATTEMPT}}{\mathtt{MOATTEMPT} + \mathtt{MTATTEMPT}} \times 100$

MO: Mobile-originating, MT Mobile-terminating, SMS: Short Message Service

Table 3.4: Customer experience SMS service predictor variable descriptions. Counter definitions follows in Section B.2.1.

Table 3.5 shows the voice experience predictor variables and calculations.

CEI	Type	Description and Formula
VOI_RA_MOST_USED_ATT	Factor	Most used radio access type based on the radio access type with maximum VOI attempt count of all radio technologies including 2G, 3G and 4G. $= \max\left[\texttt{ATTEMPT}_{\texttt{2G}}, \texttt{ATTEMPT}_{\texttt{3G}}, \texttt{ATTEMPT}_{\texttt{4G}}\right]$
VOI_ATTEMPT	Count	Total number of MO and MT voice calls that where attempted on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MO},\texttt{MT}\}} \sum_{j\in\{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{ATTEMPT})_{i,j}$
VOI_ATTEMPT_TO_ALERTFAILRATE	Rate	Attempt to alert fail rate calculated as the ratio of calls that are failing between the attempt and alerting phases, versus the total number of call attempts. $= \frac{\texttt{ATTEMPT\_TO\_ALERTFAIL}}{\texttt{ATTEMPT}} \times 100$
VOI_ALERTFAILRATE	Rate	Alert fail rate calculated as the ratio of calls that are failing during the alerting phase, versus the total number of alert attempts. $= \left(1 - \frac{\texttt{ALERTSUCCESS}}{\texttt{ALERTATTEMPT}}\right) \times 100$
VOI_ALERT_TO_CONNECTFAILRATE	Rate	Alert to connect fail rate calculated as the ratio of calls that are failing between the alerting and connection phases, versus the total number of successful alert attempts. $= \frac{\texttt{ALERT\_TO\_CONNECTFAIL}}{\texttt{ALERTSUCCESS}} \times 100$
VOI_SETUPFAILNETWORKRATE	Rate	Setup failure network rate calculated as the ratio of calls that failed to setup due to network related error causes, versus the total number of call attempts. $= \frac{\texttt{SETUPFAILNETWORK}}{\texttt{ATTEMPT}} \times 100$
VOI_CONNECTFAILRATE	Rate	Connect fail rate calculated as the ratio of calls that failed to connect, versus the total number of connect attempts. $= \left(1 - \frac{\texttt{CONNECTSUCCESS}}{\texttt{CONNECTATTEMPT}}\right) \times 100$
VOI_DROPNETWORKRATE	Rate	Call drop network rate calculated as the ratio of calls that dropped due to network related error causes, versus the total number of calls that successfully connected. $= \frac{\texttt{DROPNETWORK}}{\texttt{CONNECTSUCCESS}} \times 100$
VOI_CALLHANDOVERFAILRATE	Rate	Call handover fail rate calculated as the ratio of calls that failed to handover, versus the total number of calls that attempted to handover. $= \left(1 - \frac{\texttt{CALLHANDOVERSUCCESS}}{\texttt{CALLHANDOVERATTEMPT}}\right)$
VOI_RATIO_RA_MOST_USED_ATT	Rate	Ratio RA most used calculated as the ratio of call attempts on the most used radio access type, versus the total number of call attempts. $= \frac{\texttt{ATTEMPT}_{\texttt{RA\_MOST\_USED\_ATTEMPT}}}{\texttt{TOTAL\_ATTEMPT}} \times 100$
VOI_AVGSETUPTIME	Time *(ms)*	Average call setup time in milliseconds calculated as the total time calls are in the setup phase divided by the total number of call attempts. $= \frac{\texttt{AVGSETUPTIME\_N}}{\texttt{AVGSETUPTIME\_D}}$
VOI_AVGDURATION	Time *(s)*	Average voice call duration in seconds calculated as the total time calls were active divided by the total number of connected calls. $= \frac{\texttt{AVGDURATION\_N}}{\texttt{AVGDURATION\_D}}$
VOI_RATIO_MODUR	Rate	Ratio of mobile originating call duration versus total call duration across 2G, 3G and 4G radio layers. $= \frac{\texttt{MOAVGDURATION\_N}}{\texttt{MOAVGDURATION\_N} + \texttt{MTAVGDURATION\_N}} \times 100$

MO: Mobile-originating, MT Mobile-terminating

TABLE 3.5: Customer experience voice service predictor variable descriptions. Counter definitions follows in Section B.2.1.

Table 3.6 shows the data access bearer service predictor variable calculations for cases where the maximum radio technology is 2G or 3G.

CEI	Type	Description and Formula
`BER_RA_MOST_USED_ATT`	Factor	Most used radio access type, based on the radio access type with maximum attempt count of all BER radio technologies including 2G and 3G. $= \max\left[\mathtt{ATTEMPT}_{\mathtt{2G}}, \mathtt{ATTEMPT}_{\mathtt{3G}}\right]$
`BER_ATTACHATTEMPT`	Count	Total number of Attach Requests attempted on the 2G or 3G radio layers. $= \sum_{j \in \{\mathtt{2G},\mathtt{3G}\}} (\mathtt{ATTACHATTEMPT})_j$
`BER_ATTACHAVGDURATION`	Time *(ms)*	Average Attach Request time in milliseconds calculated as the total Attach Request time divided by the total number of Attach Request messages sent. $= \frac{\mathtt{ATTACHAVGDURATION\_N}}{\mathtt{ATTACHAVGDURATION\_D}}$
`BER_PDPACTFAILRATE`	Rate	PDP Activation fail rate calculated as the ratio of unsuccessful PDP Activation Request messages, versus the total number of PDP Activation Request messages. $= \frac{\mathtt{PDPACTFAIL}}{\mathtt{PDPACTATTEMPT}} \times 100$
`BER_PDPACTAVGDURATION`	Time *(ms)*	Average PDP Activation Request time in milliseconds calculated as the total PDP Activation Request time divided by the total number of PDP Activation Request messages sent. $= \frac{\mathtt{PDPACTAVGDURATION\_N}}{\mathtt{PDPACTAVGDURATION\_D}}$
`BER_RATIO_RA_MOST_USED_ATT`	Rate	Ratio RA most used calculated as the ratio of bearer attempts on the most used radio access type, versus the total number of bearer attempts. $= \frac{\mathtt{ATTEMPT}_{\mathtt{RA\_MOST\_USED\_ATTEMPT}}}{\mathtt{TOTAL\_ATTEMPT}} \times 100$

`ATTACH`: Attach Request, `PDPACT`: PDPActivation Request

TABLE 3.6: Customer experience data access bearer predictor variables in cases where the radio technology used was 2G or 3G generation at most.

Table 3.7 shows the data access bearer predictor variables and calculations for cases where the maximum radio technology is 4G.

CEI	Type	Description and Formula
BER_RA_MOST_USED_ATT	Factor	Most used radio access type, based on the radio access type with maximum attempt count of all BER radio technologies including 2G, 3G and 4G. $$= \max\left[\texttt{ATTEMPT}_{\texttt{2G}}, \texttt{ATTEMPT}_{\texttt{3G}}, \texttt{ATTEMPT}_{\texttt{4G}}\right]$$
BER_EPSATTACHATTEMPT	Count	Total number of EPS Attach Requests attempted on the 4G radio layer. $$= \sum_{j \in \{\texttt{4G}\}} (\texttt{EPSATTACHATTEMPT})_j$$
BER_EPSATTACHAVGDUR	Time *(ms)*	Average EPSAttach Request time in milliseconds calculated as the total EPSAttach Request time divided by the total number of EPSAttach Request messages sent. $$= \frac{\texttt{EPSATTACHAVGDURATION\_N}}{\texttt{EPSATTACHAVGDURATION\_D}}$$
BER_DFLTBEARERACTFAILRATE	Rate	Default Bearer Activation fail rate calculated as the ratio of unsuccessful Default Bearer Activation Request messages, versus the total number of Default Bearer Activation Request messages. $$= \frac{\texttt{DFLTBEARERACTFAIL}}{\texttt{DFLTBEARERACTATTEMPT}} \times 100$$
BER_DFLTBEARERACTAVGDUR	Time *(ms)*	Average duration for Default Bearer Activation Request message types in milliseconds calculated as the total Default Bearer Activation Request time divided by the total number of Default Bearer Activation Request messages sent. $$= \frac{\texttt{DFLTBEARERACTAVGDUR\_N}}{\texttt{DFLTBEARERACTAVGDUR\_D}}$$
BER_PDNCONNAVGDUR	Time *(ms)*	Average duration for PDN Connectivity Request message types in milliseconds calculated as the total PDN Connectivity Request time divided by the total number of PDN Connectivity Request messages sent. $$= \frac{\texttt{PDNCONNAVGDUR\_N}}{\texttt{PDNCONNAVGDUR\_D}}$$
BER_RATIO_RA_MOST_USED_ATT	Rate	Ratio RA most used calculated as the ratio of bearer attempts on the most used radio access type, versus the total number of bearer attempts. $$= \frac{\texttt{ATTEMPT}_{\texttt{RA\_MOST\_USED\_ATTEMPT}}}{\texttt{TOTAL\_ATTEMPT}} \times 100$$

EPSATTACH: EPS Attach Request, DFLTBEARERACT: Default Bearer Activation Request, PDNCONN: PDN Connectivity Request

TABLE 3.7: Customer experience data access bearer predictor variables in cases where the radio technology used was up to the 4G generation. Counter definitions follows in Section B.2.2.

Table 3.8 shows the mobile application experience predictor variables and calculations.

CEI	Type	Description and Formula
CNA_RA_MOST_USED_PACKETS	Factor	The most used radio access type. The most used RA type is based on the radio access type with the maximum packet count of all radio technologies including 2G, 3G and 4G. $= \max\left[\mathtt{PACKETS}_{\mathtt{2G}}, \mathtt{PACKETS}_{\mathtt{3G}}, \mathtt{PACKETS}_{\mathtt{4G}}\right]$
CNA_APP_MOST_USED_PACKETS	Factor	The most used application category. The most used application category is based on the application category with the maximum packet count of all application categories. $= \max\left[\mathtt{PACKETS}_{\mathtt{Cat1}}, \mathtt{PACKETS}_{\mathtt{Cat2}}, \ldots, \mathtt{PACKETS}_{\mathtt{CatN}}\right]$
CNA_DOWN_MBYTES	Count	Total volume in Megabytes of data received on the down-link for all radio technologies. $= \sum_{j\in\{\mathtt{2G},\mathtt{3G},\mathtt{4G}\}} \mathtt{DOWN\_BYTES}_j / 1 \times 10^6$
CNA_UP_RETRANSMITRATE	Rate	Ratio of retransmitted uplink packets to total uplink packets. $= \frac{\mathtt{UP\_RETRANSMIT\_PACKETS}}{\mathtt{UP\_PACKETS}} \times 100$
CNA_DOWN_RETRANSMITRATE	Rate	Ratio of retransmitted down-link packets to total down-link packets. $= \frac{\mathtt{DOWN\_RETRANSMIT\_PACKETS}}{\mathtt{DOWN\_PACKETS}} \times 100$
CNA_UP_THROUGHPUT_KBPS	Rate *(kbps)*	Average uplink throughput for applications measuring network usage calculated as the total upload volume in Kilobits divided by the total active upload time in seconds. $= \frac{\mathtt{UP\_THROUGHPUT\_MBIT}}{\mathtt{UP\_THROUGHPUT\_SECONDS}} \times 1 \times 10^3$
CNA_DOWN_THROUGHPUT_KBPS	Rate *(kbps)*	Average down-link throughput for applications measuring network usage calculated as the total download volume in Kilobits divided by the total active download time in seconds. $= \frac{\mathtt{DOWN\_THROUGHPUT\_MBIT}}{\mathtt{DOWN\_THROUGHPUT\_SECONDS}} \times 1 \times 10^3$
CNA_LATENCY_MS	Time *(ms)*	Average Round Trip Time in milliseconds calculated as the total TCP Round Trip Time for all TCP message pairs divided by the total number of TCP Round Trip events. $= \frac{\mathtt{LATENCY\_MS\_N}}{\mathtt{LATENCY\_MS\_D}}$
CNA_RATIO_MURAT	Rate	Ratio RA most used calculated as the ratio of transmitted data packets on the most used radio access type, versus the total number of data packets transmitted on all radio types. $= \frac{\mathtt{PACKETS}_{\mathtt{RA\_MOST\_USED\_PACKETS}}}{\mathtt{TOTAL\_PACKETS}} \times 100$
CNA_RATIO_MUAPP	Rate	Ratio RA most used calculated as the ratio of transmitted data packets on the most used radio access type, versus the total number of data packets transmitted on all radio types. $= \frac{\mathtt{PACKETS}_{\mathtt{RA\_MOST\_USED\_PACKETS}}}{\mathtt{TOTAL\_PACKETS}} \times 100$

MBYTE: Megabyte, MBIT: Megabit, KBPS: Kilobit per Second

Table 3.8: Mobile application experience predictor variables and calculations. Counter definitions follows in Section B.2.3.

Chapter 4

Exploratory Data Analysis

4.1 Network NPS Surveys

Survey Sample Size

Telco's original unaltered *Network NPS* (NNPS) dataset consists of 627 095 survey responses collected over six months from October 2018 to March 2019. Table 4.1 shows the number of survey response samples collected during each month.

Period	Count	%
2018-10-01	107 105	17.1
2018-11-01	113 134	18.0
2018-12-01	68 558	10.9
2019-01-01	126 706	20.2
2019-02-01	106 829	17.0
2019-03-01	104 763	16.7

TABLE 4.1: The number of NNPS survey responses received per month during the analysis period from October 2018 to March 2019.

We see that, on average, approximately 100 000 samples were collected each month, dipping in December when the number of samples dropped below 70 000, and peaking in January when the count was greater than 126 000. This change might be due to subscribers that were less responsive to the December survey as a result of the holiday period, and some of the December survey group only providing their responses in January.

The approach we used for collecting network experience data, discussed in Section 3.3, takes each subscriber's response date into consideration, not the date when a survey was sent out. This method aims to ensure that a customer's most recent network experience will be reflected by the *Likelihood to Recommend* (LTR) score given during the survey.

Network NPS Score

The primary objective of the present project is to predict whether a subscriber is likely to be a NPS *Detractor* and, in turn, gain valuable insights into the network conditions that warrant attention to improving the Quality of Experience (QoE) for customers. We obtain the NPS status of each subscriber, by mapping their LTR score values provided during the NNPS surveys, which are integer values between zero and ten, to three discrete NPS classes *{Promoter, Neutral, Detractor}* based on the definitions given in Section 2.3.

We have provided a summary of the survey results for the entire six-month period in Table 4.2. The table displays the total number and percentage of samples for each LTR score value and the counts and percentages per NPS status group.

LTR Score	LTR Count	% LTR	Status	Status Count	% Status
10	252 036	40.2	Promoter	303 018	48.3
9	50 982	8.1			
8	59 644	9.5	Neutral	92 688	14.8
7	33 044	5.3			
6	20 821	3.3	Detractor	231 389	36.9
5	44 076	7.0			
4	14 952	2.4			
3	23 491	3.7			
2	19 328	3.1			
1	53 934	8.6			
0	54 787	8.7			

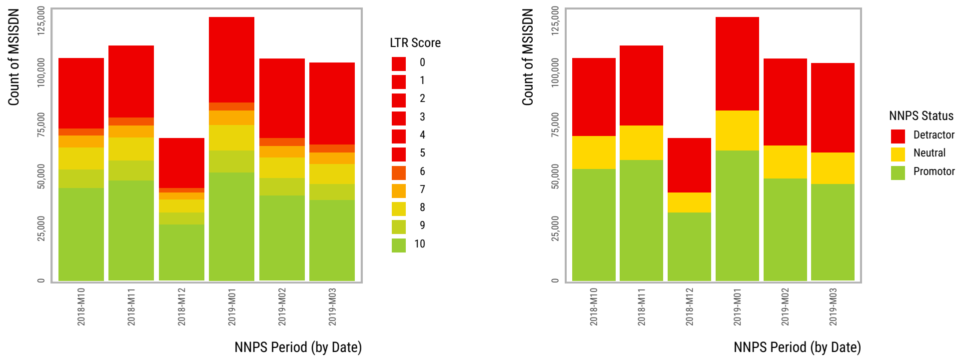
Table 4.2: Summary of the NNPS survey data from Telco for the six months from October 2018 to March 2019.

We notice that the LTR score of ten has the highest frequency, followed by score values of zero, one, eight and nine, with similar counts. Interestingly, the score value five has quite a high frequency compared to values in its immediate vicinity. The higher frequency might be due to customers' indecisive position on whether they want to promote Telco. Therefore, they pick the middle value because they don't feel very *likely to recommend*, nor do they feel very *likely not to recommend*. If this is the case, it might be related to a common phenomenon often seen with Likert-type scales, with an odd number of response options to allow a neutral response (Kulas and Stachowski, 2009). In the present case, though, the suspected *'five-neutral'* responses will translate to definite detractors on the NPS scale, which includes all scores between zero and six. At this point, we take note of the observation but don't aim to address the issue in the present project – it may be a beneficial investigation for future research to determine whether it impacts models produced from the data. Finally, we see from Table 4.2 that the majority, or 48.3% of all responders, fall in the *Promoter* category, followed by 36.9% in the *Detractor* and 14.8% in the *Neutral* category.

Based on the NPS® calculation in Section 2.3, Telco has a positive average Net Promoter Score of *11.4%* (48.3% − 36.9%), indicating that on average 11.4% more subscribers are willing to promote products from Telco than not. The problem, however, is that the margin separating the MNO from the negative side of the NPS scale is relatively small. With 14.8% of customers being *neutral*, there is a risk that sudden changes in customer perception can easily cause the scale to tip towards the negative side, which is a concern.

Network NPS Trend

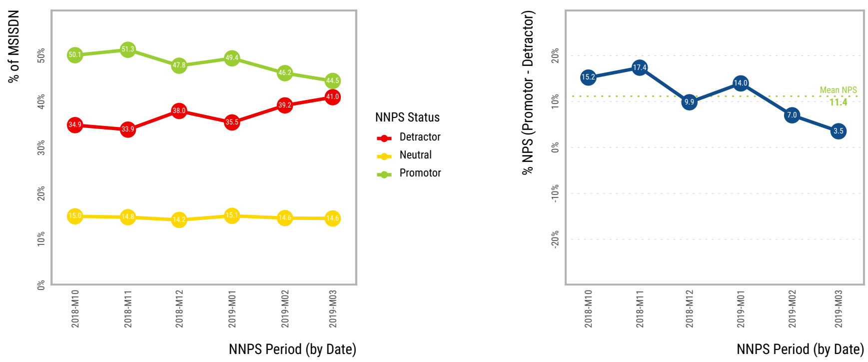
We investigate the stability of the NPS score by plotting a trend of the survey results over time to see if our data has a sufficient time range to overcome any transient effects and whether it's moving in any particular direction. In Figure 4.1, we show the monthly trend of sample counts for each LTR score and the counts per NPS status group based on the class definitions given in Section 2.3.



(a) Samples by LTR score. (b) Samples grouped by status.

FIGURE 4.1: Trend of the number of monthly NNPS survey responses broken down by LTR score and by status category based on the NPS® definitions in Section 2.3 (Reichheld, 2003).

The changing sample size every month makes it difficult to notice any underlying changes in NPS. Therefore, we plot the trends of sample proportions in each category and the overall network NPS score in Figure 4.2.



(a) Ratio per status group (b) Network NPS score

FIGURE 4.2: Trends of the monthly sample ratio per NPS status group and the overall NPS score.

Figure 4.2a shows that the *Neutral* group remains steady at approximately 14.7%, but noticeable upward and downward trends exist for the *Detractor* and *Promoter* groups, respectively. This graph suggests that subscribers may be moving between the two extreme groups. However, it is also possible that there are similar gradual movements between the *Promoter* $\rightarrow$ *Neutral* and the *Neutral* $\rightarrow$ *Detractor* groups, which may result in the net change we see. Such behaviour could be modelled using a simple Markov chain, but this would require keeping track of score changes at the customer level, which is beyond the scope of the current project. Despite individual subscriber sentiment volatility, the overall Network NPS score in Figure 4.2b suggests that the net result is more significant towards the *Detractor* group due to the negative trend.

Network NPS Distribution

In Figure 4.3, we compare the LTR score distributions between survey months to gain insight into the negative trend we saw in Figure 4.2b. The change in opacity of the histogram bars from translucent to opaque illustrates the progressive change in the LTR score distributions over time, with opaque indicating more recent and translucent earlier months.

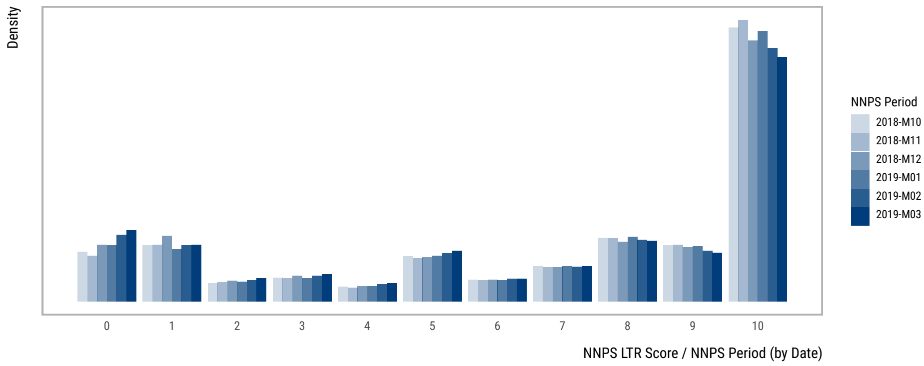


Figure 4.3: The NNPS LTR score distribution of each month. We show the gradual change in distribution of LTR scores over time by a change in the opacity of the graph as months progress, from translucent to opaque – with opaque depicting the most recent month.

Figure 4.3 shows that the monthly distributions have similar multimodal shapes with pronounced modes around LTR score values of zero, one, five, eight, nine and ten. There is a noticeable downward shift in the distribution of scores eight to ten for recent months and an opposite shift for score values zero to seven. The score-level trends correspond to the negative slope of the net NNPS graph we saw earlier in Figure 4.2b, indicating an increase in *Detractors* over time, causing a decline in *Promoters*. The distribution graphs show that most of the movement on the detractor side of the scale is in the LTR score range between zero and one, indicating an increase in *deep detraction*.

Network NPS Summary

We saw from the analysis of the survey results that Telco's Network Net Promoter Score has an average positive value of 11.4% over the entire six-month period, indicating overall positive subscriber sentiment. There is, however, a clear indication of gradual deterioration of sentiment due to a negative trend in the month-to-month NPS, from 15.2% in October down to 3.5% in March. Naturally, this should raise concerns with the network operator, as it means some underlying factors are making subscribers unhappy.

The focus should now be on determining the *key drivers behind the detraction* to comprehend the problem. This, however, points to a negative aspect of NPS. We cannot deduce the critical drivers behind detraction from the NPS survey results alone. The goal of our project is to address precisely that. We aim to establish, through predictive modelling, what the critical network experience metrics and factors are that drive negative NPS; in other words, which network-related QoE features have the most significant influence on customer dissatisfaction of subscribers with a mobile network operator?

4.2 Subscriber Segmentation Attributes

We investigate the distribution and stability of the NPS outcome variable against each segmentation attribute introduced in Section 3.2 by plotting trends of the sample distributions across each attribute level and the distribution of NPS classes within each attribute level.

The segmentation attributes that we deemed relevant features to be included in our prediction model and the possible values of each attribute are listed in Table 4.3 for reference.

Attribute	**Values**
Payment method	`Postpaid, Prepaid, Top Up`
Network region	`Central, Eastern, Kwazulu Natal, Limpopo, Mpumalanga, Northern Gauteng, Southern Gauteng, Western`
Device type	`Smart Phone, Feature Phone, Tablet, Basic Phone`

TABLE 4.3: Subscriber segmentation attributes that were deemed as relevant features to be be included into our prediction model, with the possible values for each attribute.

Network NPS by Payment Method

Table 4.4 shows the total number and percentage of samples in each level of the *Payment method* attribute for the entire six-month period. We see that the largest group of 84.3% subscribers are from the *Prepaid* class, the second largest group is the *Top Up* class with 11.3%, and the smallest group is *Postpaid* with a 4.4% sample representation.

Payment Method	**Count**	**%**
Prepaid	528 736	84.3
Top Up	70 593	11.3
Postpaid	27 783	4.4

TABLE 4.4: The count and percentages of total samples per *Payment method* for the six months.

In Figure 4.4, we show trends of the monthly sample ratio for each *Payment method*. From the figure, we see that the sample ratio of each group remained relatively constant, with only a slight amount of drift between the groups.

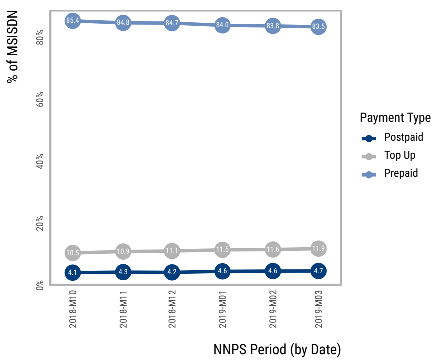


FIGURE 4.4: Trends of the monthly NNPS sample ratios per *Payment method*.

Over the six months, the *Prepaid* group reduced by 1.9% from 85.4% to 83.5%, the *Top Up* group gained 1.4% from 10.5% to 11.9%, and the *Postpaid* group gained 0.6%, from 4.1% to 4.7%.

Figure 4.5 shows trends of the monthly NPS class ratios within each *Payment method* group. The figure indicates that sentiments between the groups are notably different. The *Postpaid* group has the most significant ratio of *Detractors*, between 46% and 53%, versus values between 34% and 45% for the *Prepaid* and *Top Up* subscribers. The higher detractor ratio suggests that, on average, the postpaid subscriber group is more likely to provide low scores during the NNPS survey, which might indicate some underlying factor that negatively affects the satisfaction of contract subscribers compared to the other payment types. One possible explanation might be that contract subscribers are unhappy about their network experience but cannot turn to alternative MNOs due to contractual obligations. Consequently, they are even more likely to be disappointed with Telco. Another reason may be that contract subscribers have higher QoE expectations due to their higher premiums than prepaid subscribers.

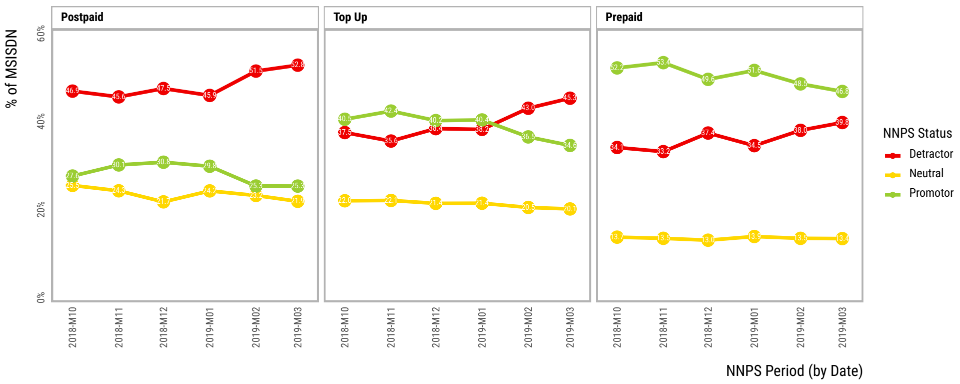


FIGURE 4.5: Trend of the monthly NPS class ratios broken down by *Payment method*.

Figure 4.5 further indicates that *Top Up* and *Prepaid* groups have similar *Detractor* ratios but a notable difference in their *Neutral* proportions. A difference in Neutral proportions implies a difference in Promotors. Consequently, there will be an impact on the net NPS between the two groups because neutral samples are excluded during the NPS calculation.

Figure 4.6 shows the monthly trends of net NPS broken down by *Payment method*. We notice similar negative trends for each NPS class, with a clear distinction in relative score values between classes. The negative trends align with the overall trend in Figure 4.2b. We note a dip in net NPS across all groups in December, but it is more pronounced for the *Prepaid* group. The more excessive dip in NPS might indicate some underlying factor that was more annoying to prepaid subscribers – for example, recharge platforms that could have failed when the load on the network increased, resulting in prepaid subscribers being unable to use mobile services at peak times. Additionally, from January to March, there is an evident change in the trend of net NPS across all three groups, from almost flat to negative. The negative shift seems related to a common underlying factor for all subscriber types, which might indicate overall network degradation.

Key takeaways from the *Payment method* attribute analysis is that there are apparent absolute level differences in the subscriber satisfaction between the different payment

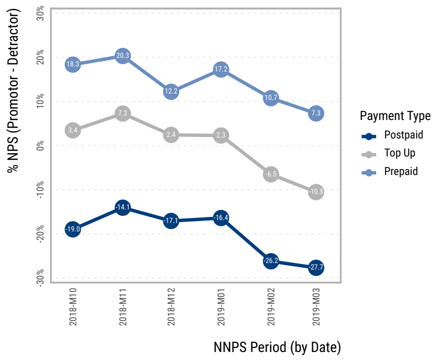


FIGURE 4.6: Trends of the monthly net NPS score broken down by *Payment method*.

type categories. All *Payment method* categories have similar negative trends for NPS and their sample ratios remain constant during the analysis period.

Network NPS by Region

The *Network region* attribute facilitates the geographical separation of our data into spatial clusters that correspond to the administrative layout of the Telco network. Each network region, in turn, is managed by a dedicated team responsible for maintaining local infrastructure and funding network operations and improvement initiatives in their respective areas. Our analysis, conducted over a six-month period, revealed that the *Southern Gauteng* and *Kwazulu Natal* regions have the largest cohorts, each representing more than 18% of the base. The *Northern Gauteng*, *Limpopo* and *Mpumalanga* regions each represent between 14% and 16%. The remaining three regions, namely *Western*, *Eastern*, and *Central*, combined have a lower representation than the top two. These figures align with the overall network subscriber-base distribution of Telco. Table 4.5 illustrates the count and percentage of NNPS samples within each network region over the six months of our analysis.

Network Region	**Count**	**%**
Southern Gauteng	118 837	18.9
Kwazulu Natal	115 012	18.3
Northern Gauteng	101 362	16.2
Limpopo	95 209	15.2
Mpumalanga	89 328	14.2
Western	44 214	7.1
Central	31 792	5.1
Eastern	31 358	5.0

TABLE 4.5: Count and ratio of total samples per *Network region* over the six-month period.

The graph in Figure 4.7 displays the monthly sample ratio for each *Network region*. We note that the sample ratios in all areas remained consistent over the six months of the present analysis, except for March, the last survey month. During this month, there was a slight increase in the two Gauteng areas and the Western region, particularly in the Southern Gauteng. On the other hand, there was a corresponding decrease in the Kwazulu Natal and Limpopo regions, which seems to match the growth seen in the mentioned areas.

After consulting with the data owners, it was discovered that the change in sample ratios for these regions was due to a correction made by Telco to ensure that the

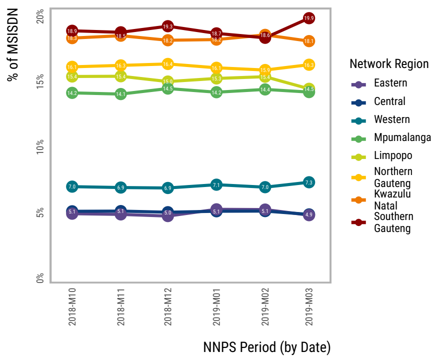


FIGURE 4.7: Trends of the monthly NNPS sample ratios per *Network region*.

sampling strategy of the NNPS surveys is more representative of the actual subscriber base composition of each area, from the beginning of the next financial cycle. We acknowledge this change. However, we don't anticipate any significant impact on our analysis because the change is small and data will be merged across months when we perform modelling.

Figure 4.8 shows the trends of the monthly NPS class ratios broken down by *Network region*.

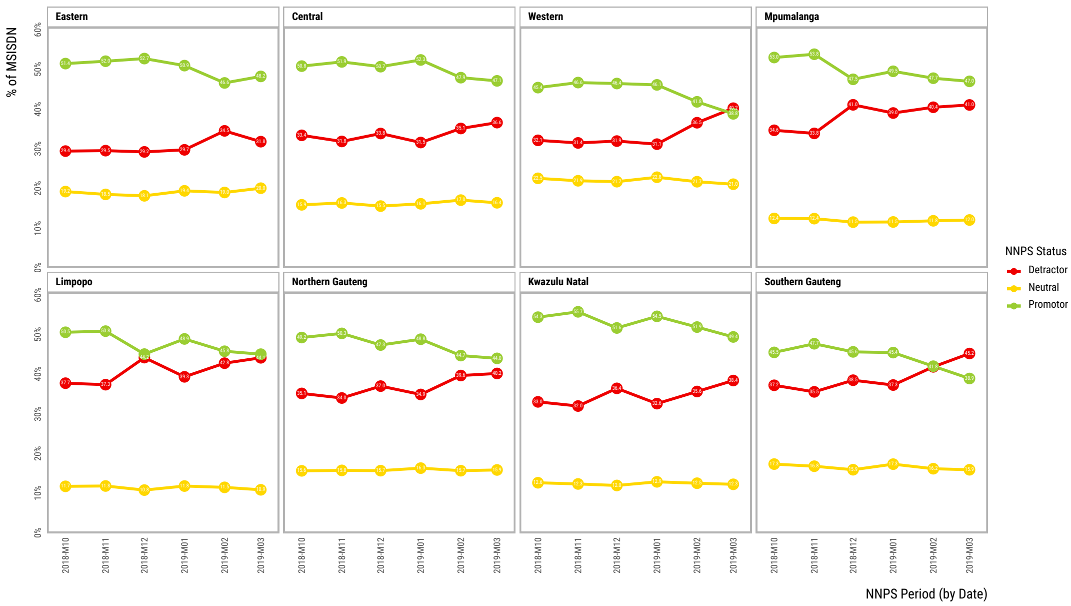


FIGURE 4.8: Trends of the monthly NPS class ratios within each *Network region*.

The graphs reveal a few noteworthy observations, which are outlined below:

- The ratios of *Detractors* across all regions are quite similar, with only minor differences due to individual transient effects. All regions have comparable ratios of detractors, ranging from 29% to 45%, with only subtle variations. The overall trend indicates that detractor ratios tend to increase over time, reflecting a similar message as previously observed. Furthermore, paying closer attention to the transient effects, we observe spikes at different periods for specific regions. For instance, the *Mpumalanga*, *Limpopo*, and *Kwazulu Natal* regions exhibited spikes in December, whereas the *Eastern* region experienced a spike in February.

All areas except Eastern displayed increasing detractor ratios over the last two months. Conversely, the Eastern region showed an increase in February but a decline in March.

- The *Neutral* groups remain stable in all regions, with only slight drift in some instances. Despite the stability, a more noticeable difference between regions is the absolute levels of neutral groups. The *Western* region has the most significant proportion of neutral subscribers, averaging above 21%, followed closely by the *Eastern* region. The *Limpopo* and *Mpumalanga* regions, on the other hand, have the smallest ratios of neutral subscribers, averaging between 11% and 12%.
- Owing to the stability of the Neutral groups in each region, the *Promoter* groups track opposite directions to the Detractor groups in each case.

Figure 4.9 shows the monthly net NPS for each *Network region* over the six months we covered in our analysis.

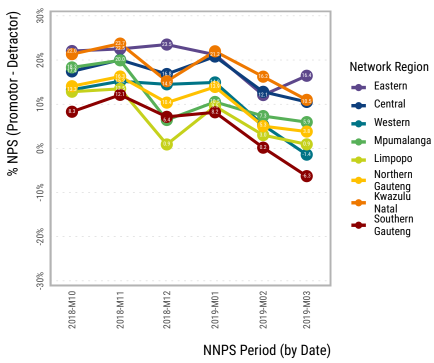


FIGURE 4.9: Trends of the monthly net NPS score broken down by *Network region*.

The observations from this graph are similar to those we made in Figure 4.8, but the net effects on NPS, which are related to the trends and transient effects we noticed in the trends of Figure 4.8, becomes more apparent from Figure 4.9. Excluding December, a known period of drastic traffic changes in the network, we see that the trend of net NPS remains relatively constant during the first four months of observation but drastically becomes more negative after January. Overall, the average net NPS trends are negative across all regions. The *Eastern* region is the only exception, with an upward trend in the last month. We also notice that some areas are tracking each other closely, such as *Kwazulu Natal*, *Northern Gauteng* and *Southern Gauteng*, or *Mpumalanga* and *Limpopo*. The *Western*, *Eastern* and *Central* regions seem to be tracking together, except during March, when the *Eastern* region shows different behaviour from all other areas. Trends of *Mpumalanga* and *Limpopo* indicate a substantial step down from December onwards, with only a slight recovery in January, after which trends follow similar negative trends like the other regions.

Another observation, informed by experience and intimate knowledge of the Telco network and its subscriber base, is that the dip in NPS we saw in Figure 4.6 for prepaid subscribers during December is limited primarily to a subset of the network regions, such as *Limpopo*, *Mpumalanga* and *Kwazulu Natal*, as shown in Figure 4.9. These regions also happen to have the largest concentrations of prepaid subscribers. Therefore, it's a chicken and egg situation where we don't know whether the cause of the dip could be related to some issue associated with the area, or the problem could be related to the subscriber type, as discussed earlier - the effects we see in Figure 4.6 and Figure 4.9 could be driven from either side.

The key takeaways from the *Network region* attribute analysis are that there are apparent absolute level differences in subscriber sentiment between the different geographic areas of the network, and, in general, all network areas follow a similar negative trend in NPS, similar to Figure 4.2b, which indicates that the *Detractor* likelihood is increasing over time at similar rates across all network regions.

Network NPS by Device Type

As for the last subscriber attribute we analyse, it is the *Device type*. We consider it essential to include this attribute as a possible factor influencing subscribers' perceived network quality because mobile devices play a critical role in their interaction with network services. There are various ways to categorise devices, such as by manufacturer brand and device model. However, this categorisation would create a feature space too complex to explore in the present project. To simplify, we will use the *Device type* attribute, which Telco uses internally to classify devices into four categories based on standard device capabilities: *Smart Phone*, *Feature Phone*, *Tablet*, and *Basic Phone*.

Table 4.6 provides the sample counts and ratios for each *Device type* category – based on the grouping of devices according to the definitions used by Telco, as mentioned. We note that the most significant proportion of devices belong to the *Smart Phone* category, with a sample ratio of 71% over the entire six-month period. *Feature Phones* comprise the most considerable remainder of devices at 25%, and only a tiny proportion of less than 5% combined belong to *Tablets* and *Basic Phones*. During modelling data preparation, as discussed in Chapter A, we excluded the two smaller groups from the data due to their limited sample size and to simplify the analysis.

Device Type	**Count**	**%**
Smart Phone	446 854	71.3
Feature Phone	155 300	24.8
Tablet	20 218	3.2
Basic Phone	4 725	0.8

Table 4.6: The count and percentages of total samples per *Device type* over the six-month period.

Figure 4.10 shows the monthly trends of sample ratios per *Device type*. We note the ratio of *Smart Phone* subscribers shows a gradual total increase of 1.0% over the first three months of the analysis period and remains constant over the following three months. The ratio of *Feature Phone* devices shows an opposite gradual decline during the first three months, suggesting that subscribers may be switching from older devices to newer 'smart' devices – possibly due to targeted upliftment campaigns performed by Telco – to free up older generation radio spectrum for rolling out newer technologies like 4G. Due to the very slight change over the six months of our analysis, we don't anticipate any significant impact. We won't be investigating this any further to simplify the present analysis.

Figure 4.11 shows the monthly trends of NNPS class ratios broken down by *Device type*. We notice slight differences in the absolute ratios of *Neutral* NPS classes between the device types – similar to behaviour we've seen with the *Payment method* and *Network region* attributes. *Detractor* ratios are very similar between device categories and show

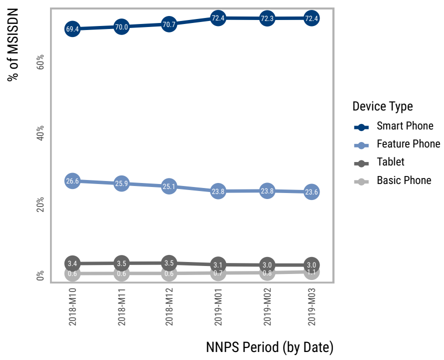


FIGURE 4.10: Trends of the monthly NNPS sample ratios per *Device type*.

a gradual increase over time. Once again *Promotor* ratios follow an inverse trend to detractors due to the constant ratio of the neutral group.

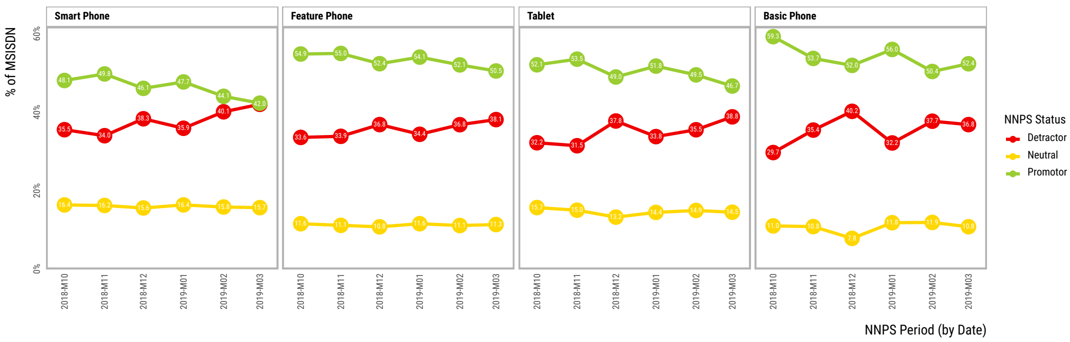


FIGURE 4.11: Trends of the monthly NNPS class ratio within each *Device type* level.

Figure 4.12 shows the monthly trends of net NPS broken down by *Device type*. We notice similar negative slopes for all device types, with the net NPS of the smaller two groups being slightly eratic due to their small sample size. All device types show a similar dip during December as noted with the analyses of other subscriber attributes.

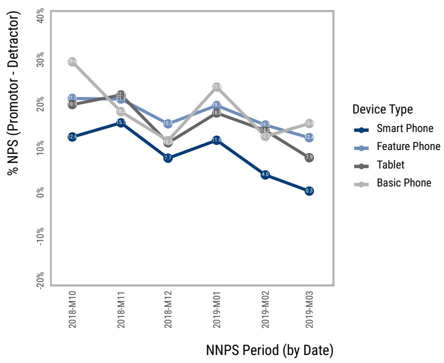


FIGURE 4.12: Trends of the monthly NPS score broken down by *Device type* category.

Key take aways from the *Device type* category analysis are similar to the *Payment method* and *Network region* attributes, where we note absolute level differences between the levels in each category, and similar negative trends of net NPS over time.

4.3 Network Probing Experience Metrics

We explore the data obtained from the Customer Experience (CE) network probing sources. This dataset, in particular, proved to be quite complex and challenging to prepare for modelling, primarily due to incomplete data and differences in the granularity of measurements within each source. The probing data, therefore, went through an extensive cleanup process, which is explained at length in Chapter A, Probing Data Preparation, before we reach this point of the analysis.

Samples per Service

After cleanup of samples with low representation we are left with a probing dataset with the composition as shown in Table 4.7.

Service	Count	%
NNPS	502 884	100.0
VOI	502 884	100.0
SMS	502 884	100.0
BER	450 741	89.6
CNA	382 233	76.0

TABLE 4.7: The number of observations per service for the final cleaned probing dataset.

Samples per Radio Type

After cleanup of samples with low representation we are left with a probing dataset with the radio type activity proportions as shown in Table 4.8.

Radio Type	Count	%
2G	502 884	100.0
3G	402 722	80.1
4G	183 952	36.6

TABLE 4.8: The number of observations per radio technology of the final cleaned probing dataset.

Samples per Combination of Services

By grouping probing data based on the common combinations of services used, we could isolate three main clusters of subscriber types based on service use patterns. Table 4.9 shows each service group's probing data sample number and percentage. The clustering of data based on the predominant patterns allowed us to discard a lot of complexity while retaining a representative sample of our original dataset.

4.4 Detailed Analysis of Voice

We present the exploratory data analysis (EDA) results from a single service, VOI, as each service is a replication of the same process. Special cases are presented in a dedicated Section 4.5 Special Service Cases.

Voice	SMS	Data (CP)	Data (UP)	Count	%
Y	Y	Y	Y	382 233	76.0
Y	Y	Y	N	68 508	13.6
Y	Y	N	N	52 143	10.4

CP: Control-plane, **UP**: User-plane

TABLE 4.9: The number of observations for groups of services that were used together for the final cleaned probing dataset.

Complete Cases and Missing Values

Table 4.10 shows summary statistics of the numerical features in the VOI dataset. We note the number of complete and missing values for each variable, as well as variables that have large proportions of zero-values. Values highlighted were discarded based on correlation analysis which is performed in Section 4.4.

Feature	Complete	%	NA	%	Zero	%	GT-Zero	%
VOI_ATTEMPT	502 882	100.00	0	0.00	0	0.00	502 882	100.00
VOI_CONNECT	502 882	100.00	0	0.00	2 688	0.54	500 194	99.47
VOI_AVGDURATION	500 194	99.47	2 688	0.54	0	0.00	500 194	99.47
VOI_AVGSETUPTIME	500 983	99.62	1 899	0.38	0	0.00	500 983	99.62
VOI_TOTALDURATION	500 194	99.47	2 688	0.54	0	0.00	500 194	99.47
VOI_ALERT_TO_CONNECTFAILRATE	502 882	100.00	0	0.00	6 886	1.37	495 996	98.63
VOI_ALERTCONNECTRATE	500 979	99.62	1 903	0.38	1 204	0.24	499 775	99.38
VOI_ALERTFAILRATE	502 882	100.00	0	0.00	109 862	21.85	393 020	78.15
VOI_ATTEMPT_TO_ALERTFAILRATE	502 882	100.00	0	0.00	4 111	0.82	498 771	99.18
VOI_CALLHANDOVERFAILRATE	502 882	100.00	0	0.00	224 022	44.55	278 860	55.45
VOI_CONNECTFAILRATE	502 882	100.00	0	0.00	42 156	8.38	460 726	91.62
VOI_DROPNETWORKRATE	502 882	100.00	0	0.00	220 402	43.83	282 480	56.17
VOI_SETUPFAILNETWORKRATE	502 882	100.00	0	0.00	93 437	18.58	409 445	81.42
VOI_RATIO_2G_DUR	502 882	100.00	0	0.00	44 398	8.83	458 484	91.17
VOI_RATIO_3G_DUR	502 882	100.00	0	0.00	125 129	24.88	377 753	75.12
VOI_RATIO_4G_DUR	502 882	100.00	0	0.00	485 606	96.56	17 276	3.44
VOI_RATIO_MODUR	502 882	100.00	0	0.00	21 663	4.31	481 219	95.69
VOI_RATIO_RA_MOST_USED_ATT	502 882	100.00	0	0.00	0	0.00	502 882	100.00
VOI_RATIO_RA_MOST_USED_DUR	502 882	100.00	0	0.00	2 573	0.51	500 309	99.49

Complete: Complete cases, **NA**: Missing values (Not Available), **Zero**: Zero values, **GT-Zero**: Greater than zero values

TABLE 4.10: Summary statistics indicating the number of observations and missing values for each of the numerical VOI features. The variables highlighted in blue were identified as highly collinear variables as shown in Figure 4.18 and was removed from our final modelling dataset.

Unprocessed Variables Visualised

Distributions

In Figure 4.13, we can see the kernel density estimate plots of the continuous voice service features, categorized by NPS class, *before* undergoing any transformations. Upon examination, we observe that some variables naturally follow a near-normal Gaussian distribution. However, the majority of variables are significantly right-skewed, and some have value ranges of much larger magnitude. Therefore, we will need to transform the predictors to standardise them to similar scales before using them as model input. Additionally, we observe that certain variables, such as `ATTEMPTS`, `AVGSETUPTIME` and `AVGDURATION`, show apparent outliers as they exhibit large ranges. To gain a clearer understanding of this, we will use box plots.

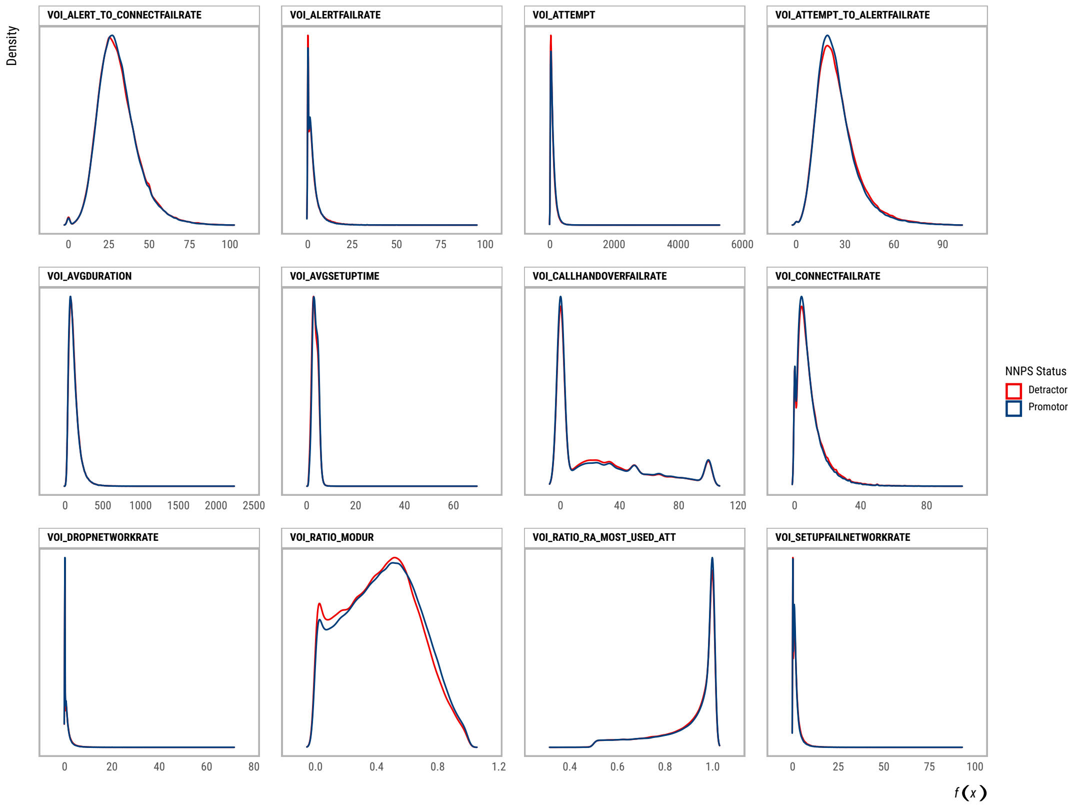


FIGURE 4.13: Kernel density estimate plots of the original VOI continuous variables, broken down by NPS class. The plots show variable distributions over the complete analysis period, prior to the application of normalisation transformations.

Box-Plot Trends

Figure 4.14 shows monthly box plots of the numerical voice service features broken down by NPS class *before* applying normalisation or standardisation transformations. Distributions for all, except `CALLHANDOVERFAILRATE`, seem to be stable during the analysis period. We also confirm the existence of extreme outliers in the cases of `ALERTFAILRATE`, `ATTEMPT`, `DROPNETWORKRATE` and `SETUPFAILNETWORKRATE`, as well as milder cases of outliers for `AVGSETUPTIME` and `AVGDURATION`.

We have noticed that the distribution of the `CALLHANDOVERFAILRATE` variable was significantly different during October compared to the following periods. This difference was caused by a problem with the handover-correlation process on the CEM platform. The issue resulted in a falsely high handover failure rate due to uncorrelated handover transactions.

One possible workaround for this problem is to use a package like `MICE` that uses *Multivariate Imputation* to allocate feasible values at a sample level based on the values of other variables (van Buuren and Groothuis-Oudshoorn, 2011). However, for the sake of simplicity, we have decided to keep the current values and revisit the issue during modeling to determine whether it has a significant impact on our model. Since a large percentage of zero values are present for this variable, we do not anticipate any significant impact.

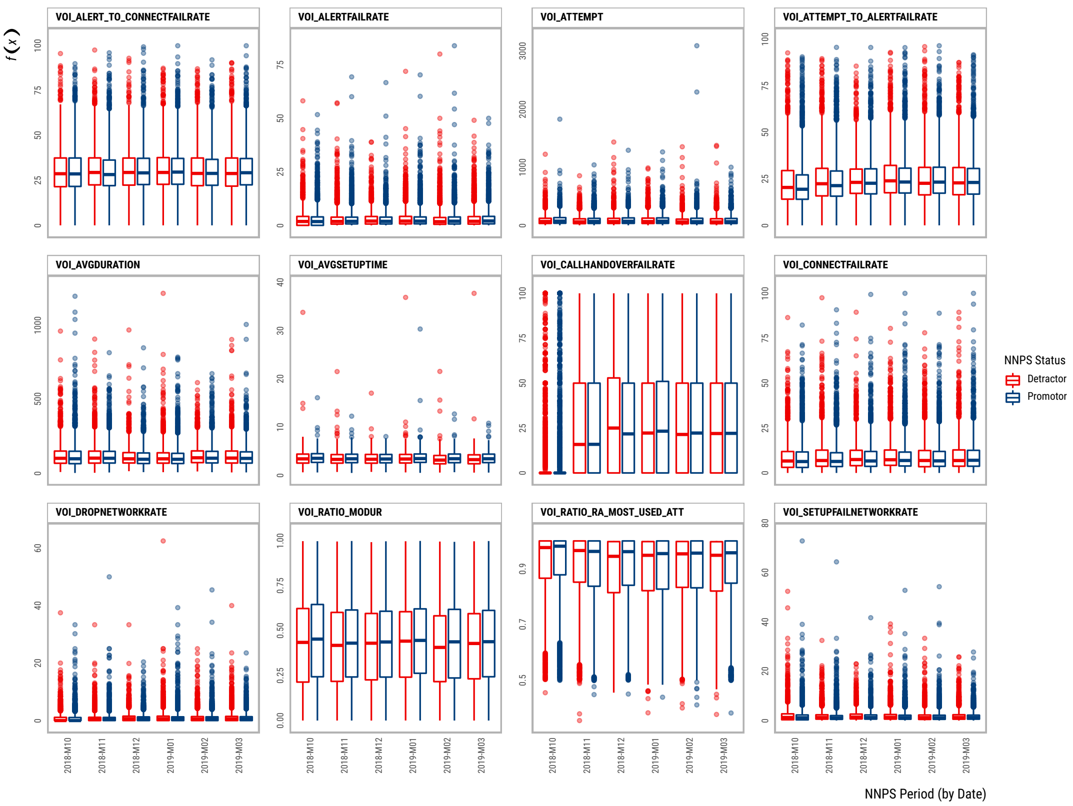


FIGURE 4.14: Box plot trends of the original VOI variables for each month. The plots show distribution box plots of each continuous variable per month prior to the application of normalisation transformations. A 10% sample of the complete VOI dataset was used for plotting.

Mean-Value Linear Trends

Figure 4.15 shows monthly trends of the weighted-mean values of each numerical voice service features, broken down by NPS class. We note that the mean values of most variables remain stable during the analysis period, except for individual cases, such as `CALLHANDOVERFAILRATE` which indicate a drastic shift from October to November. This could be attributed to a platform issue, as noted and discussed in Section 4.4, Box-Plot Trends.

We notice a clear seperation between *Detractors* and *Promotors* in the mean monthly trends of some variables, for example `ALERTFAILRATE`, `DROPNETWORKRATE`, `SETUPFAILRATE` and `CONNECTFAILRATE` possibly suggesting that these might be valuable predictor variables in our models to distinguish between NPS classes. The specific variables and the location of the *Detractor* trend in relation to the *Promotor* trend also aligns with informed knowledge which concurs with the prior statement.

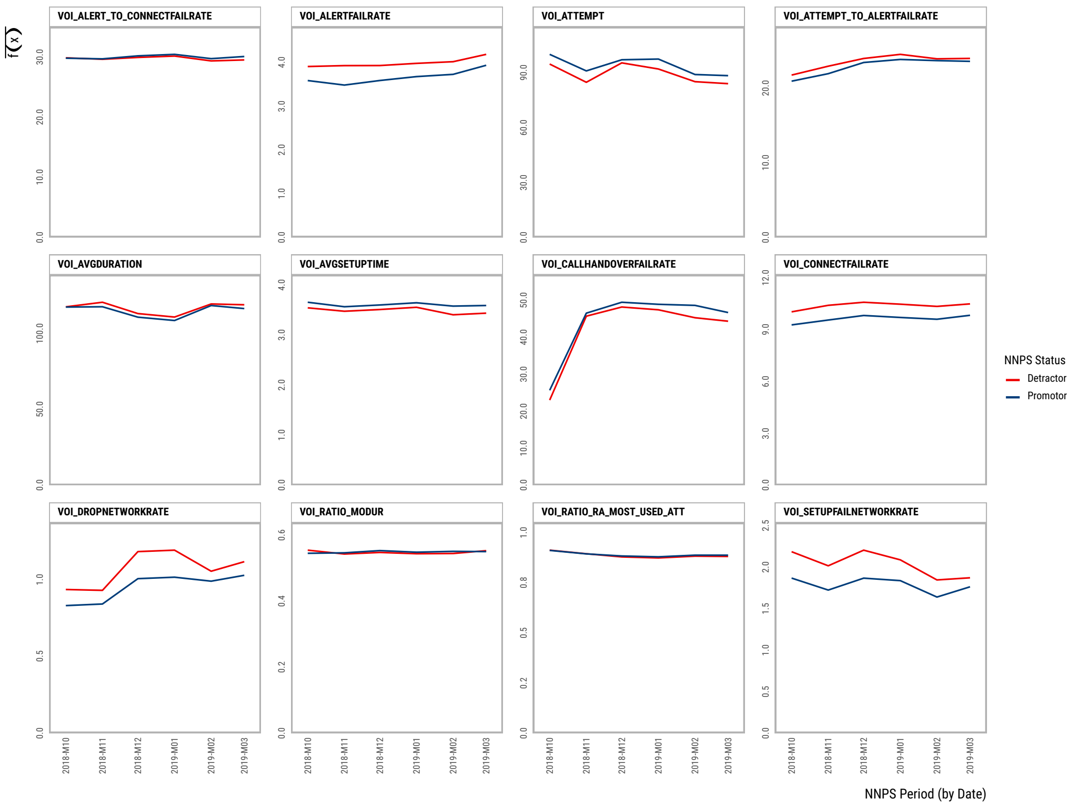


FIGURE 4.15: Linear trends of the monthly weighted-means of the continuous VOI predictor variables broken down by NPS category.

Processed Variables Visualised

Normalised Distributions

Figure 4.16 shows kernel density estimate plots of the continuous voice features, broken down by NPS class, *after* applying normalisation transformations. We notice that the applied transformations improved the shapes of skewed variables, but outliers affect the proper, valid range of variables, such as `ATTEMPT_TO_ALERTFAILRATE`. Outliers can impact the accuracy of models like logistic regression, especially if the samples are influential, and they affect the estimated coefficient values obtained during model fitting. Further diagnostics are therefore needed to mitigate the effect of outliers on our models.

During modelling, covered in Chapter 5, we use the *Mahalanobis* distance measure during input data preparation as a pre-step to identify outliers. We subsequently test their impact on model performance by systematically *"dropping"* identified outlier samples with score values above a specific threshold value during model training and validation. By using cross-validation and iterating through a given range of outlier score threshold values, we can establish the optimal *"dropout"* value for discarding outliers, consequently achieving more stable and reliable model performance.

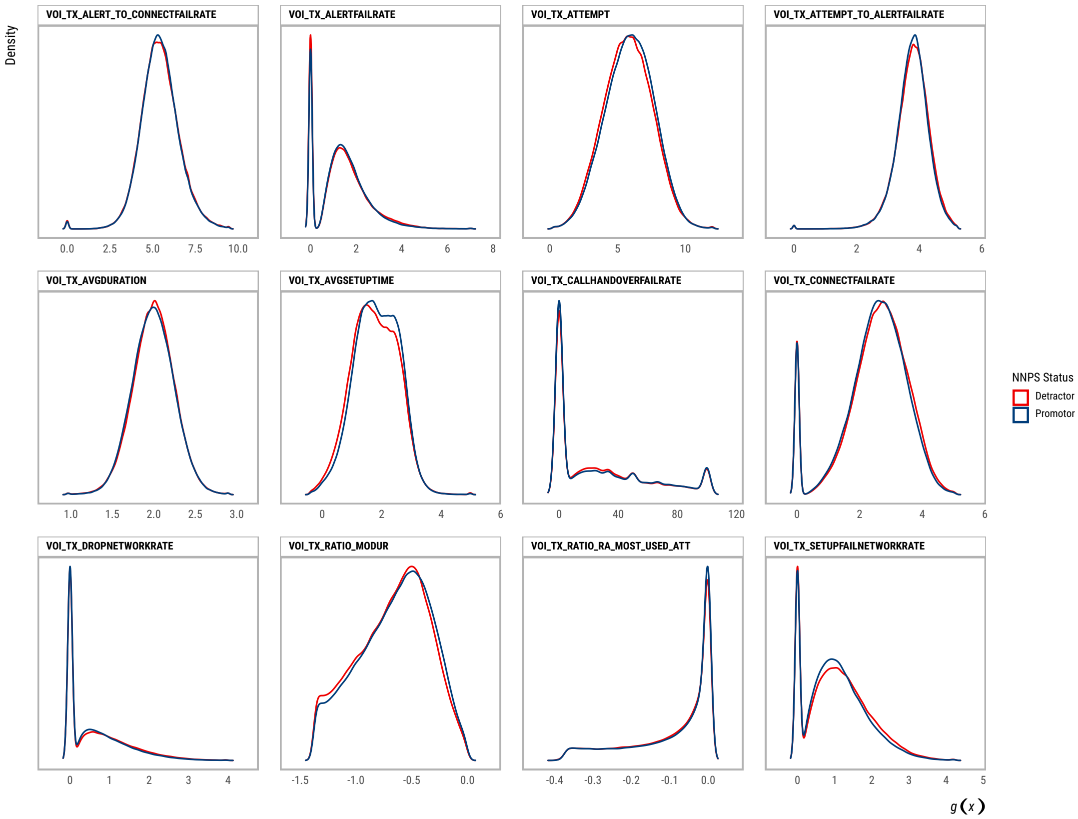


FIGURE 4.16: Kernel density estimate plots of the numeric VOI variables after normalisation transformation was applied.

Standardised Distributions

Figure 4.17 shows the kernel density estimate plots of the continuous voice service features, broken down by NPS class, *after* the application of standardisation transformations. Shapes remain similar to the *normalised* variable distributions in Figure 4.16, but value ranges of variables are scaled to one standard deviation.

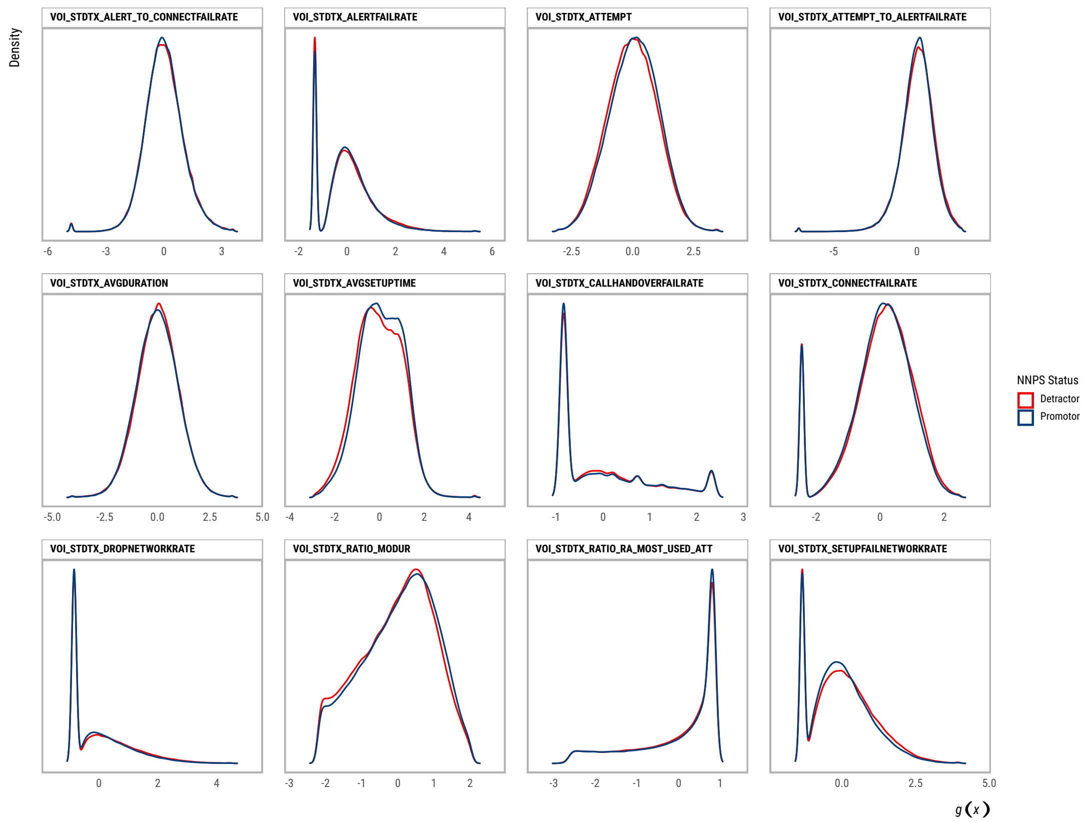


FIGURE 4.17: Kernel density estimate plots of the standardised VOI variables over the complete analysis period. The plots show numeric variable distributions after normalisation transformations and standardisation.

Correlation Analysis

Correlation analysis is a technique used to identify potential correlation issues in a dataset. One such issue is *collinearity*, which occurs when there is a strong correlation between predictor variables expressed by a *linear relationship* in a regression model. In such a scenario, the predictor variables cannot independently predict the value of the outcome variable, as they explain some of the same variance in the dependent variable. This reduces the statistical significance of the predictors when fitting a linear model to the data, making it challenging or even impossible to estimate their regression coefficients reliably. Collinearity becomes a concern when a high association exists between two potential predictor variables, and the inclusion of one predictor variable in the model reduces the significance level of the other predictor. The extreme case of collinearity, where the variables are perfectly correlated, is called *singularity*. Another problem is *multicollinearity*, which occurs when more than two predictor variables are related, and a decrease in statistical significance is observed when all are included in the model. *Variance inflation factors* (VIF) can be used as a diagnostic tool to detect both scenarios, and a common guideline is that the VIF value should not exceed 10, suggesting a high degree of collinearity or multicollinearity. Multicollinearity may not always be predictable before observing its effects on a multiple regression model, as any two predictor variables may have only a low degree of correlation. Another detection mechanism is *visualisation* of the correlation matrix using a correlation plot.

In Figure 4.18 we show the correlogram of the features from the VOI dataset prior to cleanup of correlation problems. We note that there are multiple cases of variables that have extreme absolute correlation values, in excess of 0.9, meaning that some variables are highly correlated – in some instances we notice correlation values of 1 or -1 indicating *singularity* cases. We make use of the `R caret` package, specifically the `findCorrelation` and `findLinearCombos` functions to identify possible variables to consider for exclusion, or using only one where two variables may be giving the same information.

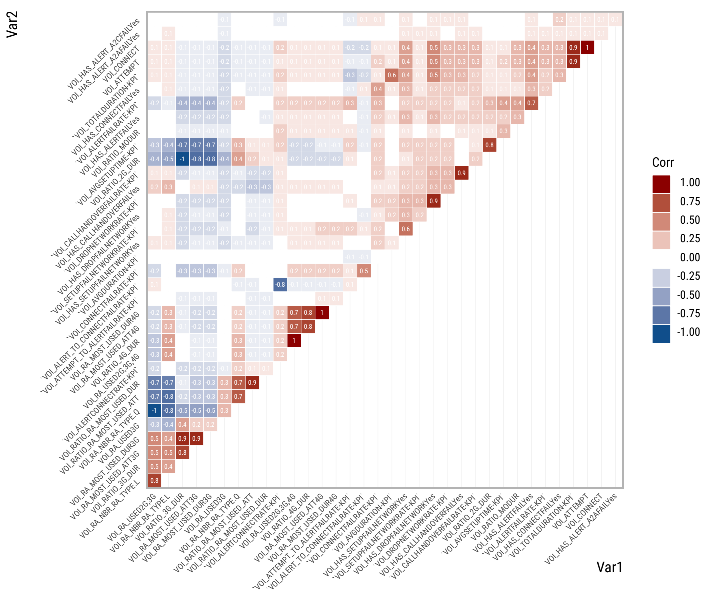


FIGURE 4.18: Correlation matrix plot of the VOI dataset prior to removal of highly-correlated variables. Correlation calculation was done making use of Spearman's ranked correlation ρ.

Some examples of variables detected by these two functions are listed in Table 4.11, showing the variable pairs with Spearman's ranked correlation ρ values. Considering the variables on a case-by-case basis, combined with informed knowledge of the meaning of each metric, allowed us to select unnecessary variables that could be removed from the data, as they either repeat the same or provide inverse information. For example, `CONNECTSUCCRATE` and `CONNECTFAILRATE` convey the same information, which is the inverse of the other. The process is iterative, and the impact on the correlation matrix needs to be verified each time after changing the predictors.

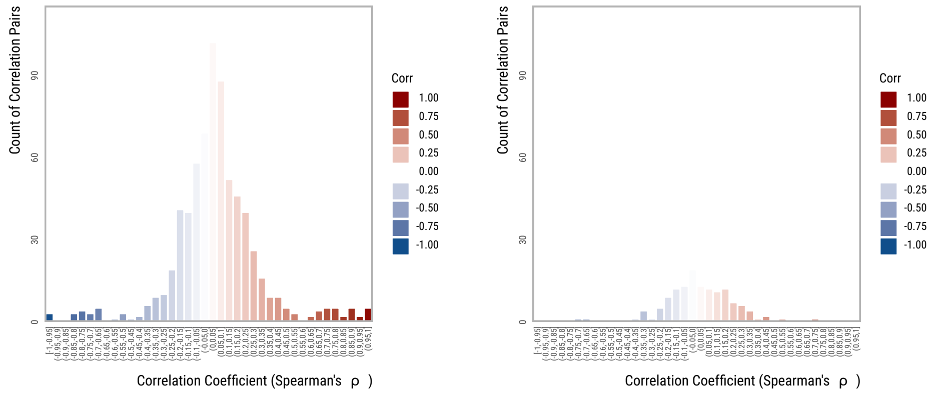
Variable 1	Variable 2	ρ	p	N
VOI_CONNECT	VOI_CONNECTSUCCESS	1.00	0.000	562 681
VOI_CONNECT	VOI_SUCCESS	0.98	0.000	562 681
VOI_ALERTATTEMPT	VOI_ALERTSUCCESS	0.97	0.000	562 681
VOI_ALERTATTEMPT	VOI_CONNECTATTEMPT	0.97	0.000	562 681
VOI_ALERTATTEMPT	VOI_CONNECTSUCCESS	0.97	0.000	562 681
VOI_ALERTATTEMPT	VOI_ATTEMPT	0.96	0.000	562 681
VOI_ALERTATTEMPT	VOI_CONNECT	0.96	0.000	562 681
VOI_CONNECT	VOI_CONNECTATTEMPT	0.94	0.000	562 681
VOI_CALLHANDOVERFAILRATE-KPI	VOI_INTRARATCALLHNDVRSUCCESSRATE-KPI	−0.90	0.000	562 681
VOI_CALLHANDOVERSUCCESSRATE-KPI	VOI_INTRARATCALLHNDVRFAILRATE-KPI	−0.90	0.000	562 681
VOI_CALLHANDOVERSUCCESSRATE-KPI	VOI_CALLHANDOVERFAILRATE-KPI	−1.00	0.000	562 681
VOI_INTERRATCALLHNDVRSUCCESSRATE-KPI	VOI_INTERRATCALLHNDVRFAILRATE-KPI	−1.00	0.000	562 681
VOI_INTRARATCALLHNDVRSUCCESSRATE-KPI	VOI_INTRARATCALLHNDVRFAILRATE-KPI	−1.00	0.000	562 681
VOI_ALERTSUCCESSRATE-KPI	VOI_ALERTFAILRATE-KPI	−1.00	0.000	562 681
VOI_CONNECTSUCCRATE-KPI	VOI_CONNECTFAILRATE-KPI	−1.00	0.000	562 681

TABLE 4.11: Examples of highly correlated variables from the original VOI dataset. Considering the variables on a case-by-case basis, combined with informed knowledge of the meaning of each metric, allowed us to select unnecessary variables that could be removed from the data, as they either repeat the same or inverse information.

One of the primary objectives of the present use case is to investigate the driving factors behind customer *dissatisfaction*. To aid the model output's interpretation, we made the decision to retain the *negative connotation* variables, such as `CONNECTFAILRATE`. By doing so, we have simplified the process of understanding the correlation between *network failures* and the *likelihood of customer dissatisfaction*. This approach has proven to be more effective than expressing the *the opposite of success* metrics.

Figure 4.19 shows a histogram plot of the count of variable pairs within binned ranges of correlation values before versus after the highly correlated variables were removed. Clean bins at the extremes of the histogram is aparent in Figure 4.19b after removing the selected variables.

Figure 4.20 shows the correlogram after the cleanup of highly correlated variables from the VOI dataset is complete. The most extreme remaining correlation is 0.75 which is between factor `RA_MOST_USED_ATT`, level *4G*, and factor `RA_USED` level *2G,3G,4G* – meaning that subscribers that have 4G as their most used technology, has a high positive correlation with subscribers that are active on all three radio technologies. This is inline with expected network behaviour, as ideally subscribers that are 4G enabled in terms of coverage and device capability, should ideally be active on 4G most of the time.



(a) Before removal of collinear variables. (b) After removal of collinear variables.

FIGURE 4.19: Distribution of the Spearman's ρ correlation coefficient between predictor variables of the VOI dataset before and after the removal of highly correlated variables.

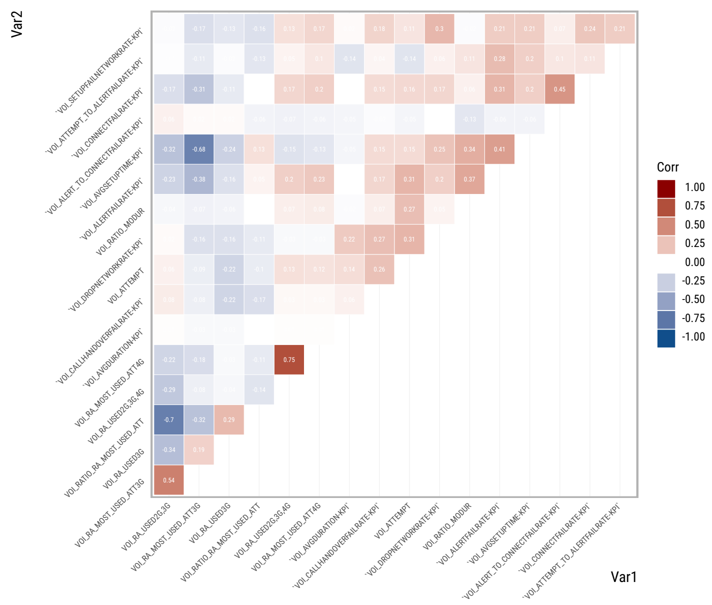


FIGURE 4.20: Correlation plot of the VOI dataset post removal of highly-correlated variables. Correlation calculation was done making use of Spearman's ranked correlation ρ.

4.5 Special Service Cases

In Section 3.4, Methods for Managing Sparse Mobile Network Data, we discussed the feature-engineering methods applied to mitigate the general occurrence of sparse data when combining data from multiple probing sources. Examples covered at the time included the *radio technology* and *directionality* attributes of the voice service. There are, however, special cases related to packet-switched data services we will present here to showcase the use of EDA to inform the feature-engineering process.

The CNA dataset contains subscriber experience data from a dedicated CEM application that focuses on analysing packet-switched application usage in mobile telecommunications networks. The application can capture and analyse user-application traffic down to the protocol message level as data is collected by probing systems from the live network, as discussed in Section 2.1, Network Probing.

Probing measurements of different applications can be grouped based on similar traffic characteristics consumed or generated on the mobile network. Some applications, for example, video or audio *streaming applications*, are prone to create high volumes of downlink traffic over extended periods at a constant bit-rate, while other applications, like *web browsing*, are generally used *ad hoc* and mainly consist of small up-link transmissions in combination with larger downlink transmissions at varying *bit-rates*. Specific applications, like *gaming* and *voice-over-IP* (VoIP), are sensitive to transmission delay (*latency*). Radio technology impacts user applications due to physical constraints imposed on data transmissions over the air interface. The generation of implemented network equipment determines the air-interface characteristics – the newer the technology, the higher the throughput (*bit-rate*) that can be achieved and the shorter the delay (*latency*).

To cater for performance analysis of the probing data, the analytical application that post-processes the probing measurements, also groups applications based on its general network transmission behaviour into so-called *Application Categories*.

Re-factoring of the CNA Application Category

Initially it was planned to include engineered features like *the most used application category* (`APP_MOST_USED`) and the *proportions* and *ranked usage* of each application category, for example a factor like: `APP_USE_web_applications` for each category, and a numerical feature like: `APP_RATIO_web_applications`) per subscriber to attempt capturing the combinations, as well as the diversity of applications used by each subscriber into our modelling dataset.

In Table 4.12 we show that many of the variables obtained from this type of scheme is not ideal and produces variables with near-zero variance which would not make usefull predictors.

To reduce the total number of categories and subsequently avoid *high cardinality* factor variables, the application categories with low occurrence were identified and grouped together into a new application category level *other*. The new `APP_RATIO_other` variable was also created to show the ratio of usage per subscriber in this new category.

Following, Figure 4.21 shows the ranked usage factor variables that were created per application category – based on how much each application is used per subscriber. Grouping and discards of low-use levels was subsequently performed, as shown in the plot, for each application-use and radio-type-use factor variable.

Feature	FreqRatio	% Unique	% Zero	Mean	SD	MAD
`CNA_APP_RATIO_test_applications`	368 459.0	0.11	99.89	0.000	0.002	0.000
`CNA_APP_RATIO_peer_to_peer`	367 090.0	0.48	99.52	0.001	0.015	0.000
`CNA_APP_RATIO_gaming`	357 956.0	2.96	97.04	0.001	0.013	0.000
`CNA_APP_RATIO_voice_over_ip`	353 396.0	4.20	95.81	0.001	0.010	0.000
`CNA_APP_RATIO_updates`	34 195.6	7.30	92.70	0.002	0.019	0.000
`CNA_APP_RATIO_app_stores`	15 111.4	13.96	86.03	0.005	0.027	0.000
`CNA_APP_RATIO_file_transfer`	14 484.6	37.17	62.83	0.006	0.027	0.000
`CNA_APP_RATIO_email`	11 094.8	30.82	69.18	0.008	0.032	0.000

Unique: Unique values, **Zero**: Zero values, **GT-Zero**: Greater than zero values

FreqRatio: Frequency ratio - ratio of frequencies for the most common value over the second most common value

Mean: Mean value, **SD**: Standard Deviation, **MAD**: Mean Absolute Deviation

TABLE 4.12: Near zero variance numerical predictor variables that were dropped from the CNA dataset. These features have very few unique values relative to the number of samples, and the ratio of the frequency of the most common value to the frequency of the second most common value is large. The `nearZeroVar` function (`freqCut = 1000, uniqueCut = 38`) from the `R` package `caret` Jed Wing et al., 2019 was used to identify the near zero variance predictors.

In Figure 4.22 we show a heatmap plot that was generated by using the ranked usage factor variables created in Figure 4.21, combined with the existing application categories attribute. The count of samples in each application and ranked-usage bin, determines the ratio of total subscribers that fall in a particular application versus ranked-use zone.

From the heatmap plot, we are able to identify usage *hot-spots* and subsequently define new re-factored application categories based on the ranked-use per application. The new categories are defined below:

web-applications – Includes any web applications that are using generic *http* and *https* to transfer content (like browsing, Facebook, Google, Flickr, Flipboard, LinkedIn and many more), was the most popular application category based on a ranking of total volume for 81.7% of our survey group, and only 1.16% of the subscribers active on packet-switched data services did not use any web_applications at all.

The next application categories competing between the 2^{nd} and 5^{th} ranked positions in descending order of popularity are: **messaging** – including applications like Hangouts, iMessage, Snapchat, Telegram, WeChat and WhatsApp, **security** – any applications using *ssl* and *dtls*, **network-operation** – applications using *dns, nat, netbios, ppp, pptp, radius* and *snmp*, and lastly **streaming-applications** – video and audio streaming applications (like Deezer, Google Play, Icecast, iTunes, Movshare, Napster, Netflix, Adobe, RTP Audio / Video, Soundcloud, Spotify, Vimeo and Youtube).

Following the very popular application categories are those that are generally more related to remote working individuals, like **file-transfer** and **email**, followed by the general device applications like **app-stores**, **updates** and **voice-over-ip**.

The least used application categories are **gaming**, **peer-to-peer** – Including applications like BitTorrent, eDonkey, Gnutella and Jabber, and lastly **test-applications** – like Ookla speed test.

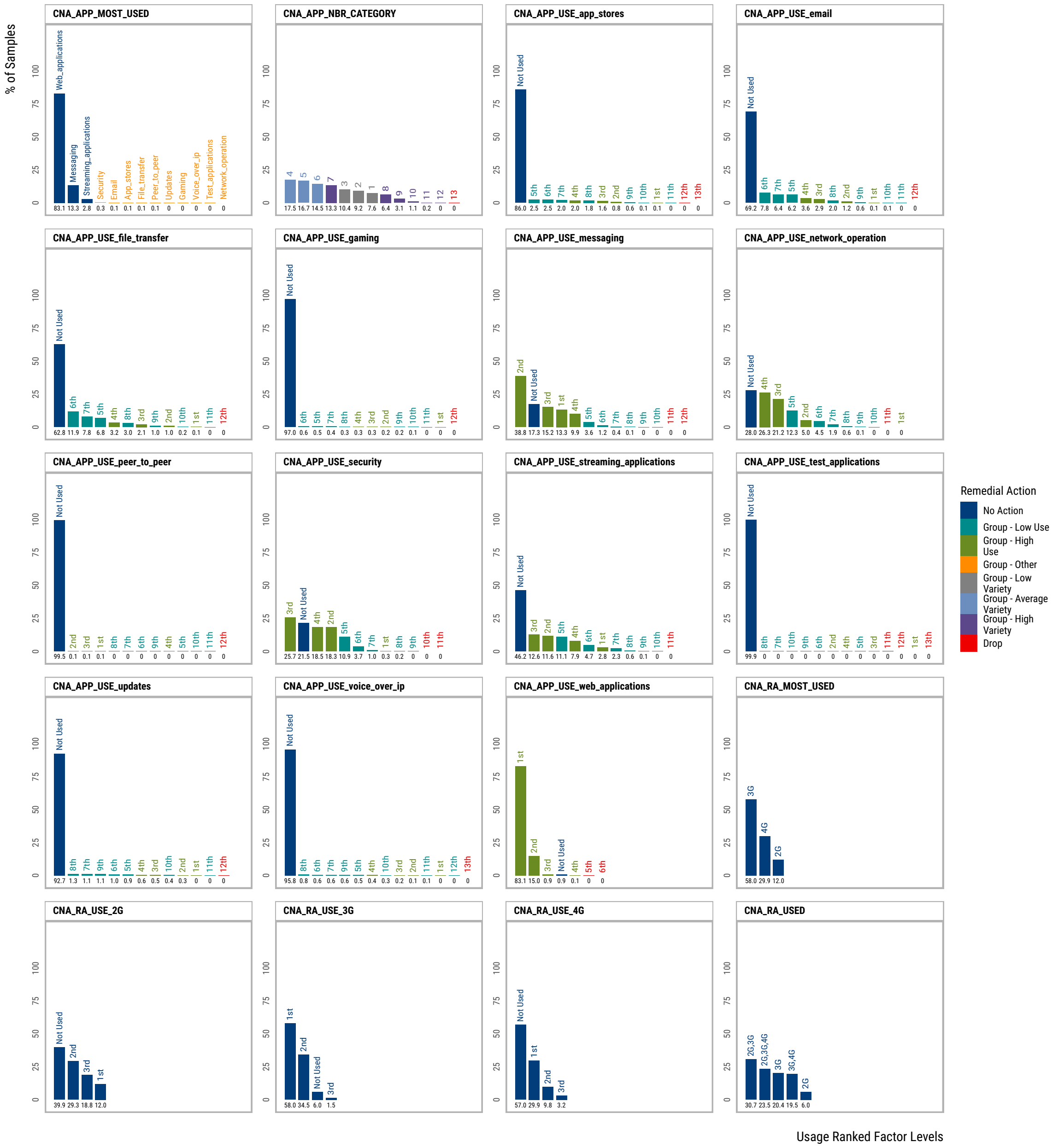


FIGURE 4.21: Sample ratio in each level of the CNA factor variables. The low contributing factor levels marked as `Action='Group'` will be re-grouped into new factor levels `'High Use'`/`'Low Use'` to indicate the ranked level of usage. Samples associated with the remaining factor levels that have low representation, marked as `Action='Drop'`, will be removed from the dataset for simplifying the analysis.

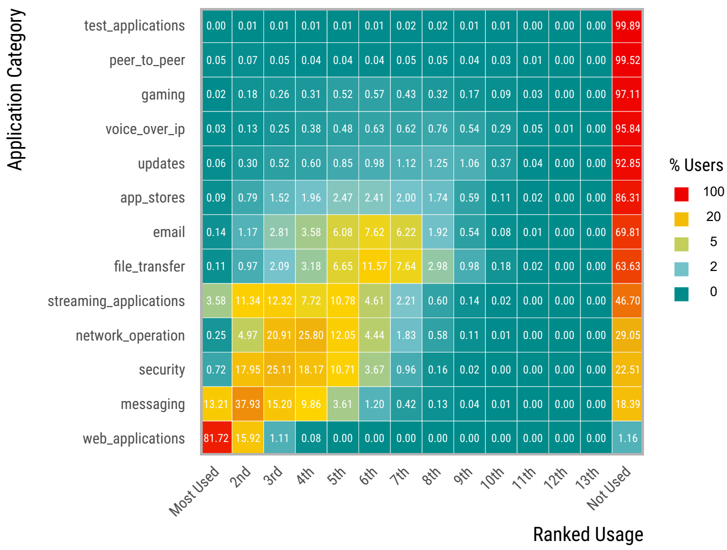


FIGURE 4.22: The percentage of data users by their ranked usage of application categories from the CNA user-plane dataset.

Chapter 5

Analysis

Modelling Approach

In our search for *'the best model'*, we compared different configurations of data sources, data pre-processing operations, and types of classification models to determine which methods produce the best results based on performance measured by some pre-defined criteria. In the present analysis prediction *accuracy* was selected as primary performance measurement criteria, as a basic starting point for purposes of simplifying the analysis. If models tie on performance based on accuracy, *ROC AUC – Receiver Operating Characteristic* (ROC) *Area Under the Curve* (AUC), will be used as deciding metric.

Some of the data preparation steps in this section might have been repeated manual work already done during the data cleanup process in Chapter A, Probing Data Preparation. The repeat is due to a late discovery of the `tidymodels` framework, better suited for repeating similar data preparation steps across multiple datasets and models. The `tidymodels` framework is a collection of packages for modelling and machine learning using tidyverse priciples; where all software packages – in this case specifically for the `R` language, share an underlying design philosophy, grammar, and data structures (Kuhn and Wickham, 2020). The *tidymodel* approach helped us conduct various experiments to effectively search an ample solution space to identify focus areas for further tuning a few performant models.

Summary of the modelling approach employed:

1. Combination of probing datasets, based on subscriber groups that use similar mobile services. We create ten distinct modelling ready datasets containing combinations from one up to all mobile network services, as discussed in Section 3.4.
2. Conduct a correlation analysis on each modelling dataset, to identify potential new collinear features that may emerge when merging data from multiple services.
3. Repeated steps for each selected type of classification model, using the concepts of *workflows* and *recipes* from the `tidymodels` framework :
    - Preprocessing and checking of variables:
        - Basic data preparation steps:
            * Check for *novel* factor levels and auto-assign previously unseen values to a *"new"* level. This step is unnecessary, due to the thorough clean-up analysis that was performed, but was included as a safety net to make our modelling pipelines future-ready.

    * Dummy variable allocation – converting factor variables to *“dummy”* numerical variables that models can use as input.
    * Checking for zero-variance predictors.
  - Variable transformations:
    * *Simple scaling* – Normalise numeric variables by scaling to one standard deviation.
    * *Normalising* – Transforming variables using regular Yeo-Johnson across all variables, as well as `bestNormalize` to select the optimal *“Normality”* transformation.
    * Outlier removal – We also test the effect of outlier removal by using a method that calculates an outlier score using the *Chisquare* distribution of the *Mahalanobis* distances (Ghorbani, 2019). A *“dropout”* parameter allows specifying a threshold value of the outlier score, beyond which samples matching criteria will be dropped. We perform grid search to find optimum values of dropout based on model performance metric *accuracy*, and we analyse the effect by plotting the model *Beta* coefficients versus *dropout score*.
- Data split into training-, validation- and test sets.
- Model training and parameter tuning by grid search and cross-fold validation of model parameters.
- Collection of performance measures.
- Select best performing model based on model performance metric *accuracy*.
- Plot performance results and analyse model behaviour by using ROC curves and confusion matrix.
- Interpret models by plotting *Variable Importance* plots. This allows us to meet our secondary objective of the project, namely to understand which features in the data are most important to drive model prediction of negative customer satisfaction.

4. Plot comparitive performance results for all models on the same performance metric axis to show relative performance of selected models compared to alternative models.

5.1 Modelling Datasets

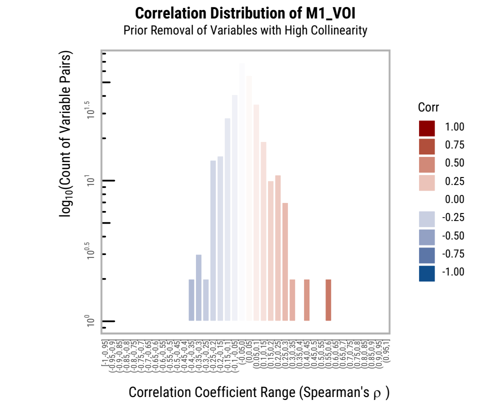
Table 5.1 shows the number of observations per modelling dataset that are available for training and testing of models. The ten modelling datasets were constructed making use of the data preparation steps discussed in Section 3.4, Separate Modelling Data.

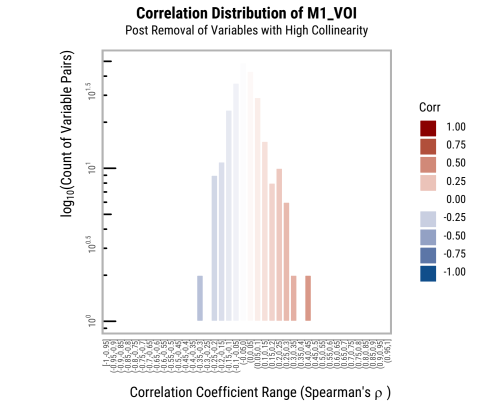
Model Id	Included Services	Dataset Label	Count	%
1	VOI	`MOD_M1_VOI`	457 840	100.00
2	SMS	`MOD_M2_SMS`	457 840	100.00
3	BERa	`MOD_M3_BERa`	406 874	88.90
4	BERb	`MOD_M4_BERb`	162 989	35.60
5	CNA	`MOD_M5_CNA`	333 064	72.70
6	VOI, SMS	`MOD_M6_VOI-SMS`	457 840	100.00
7	VOI, SMS, BERa	`MOD_M7_VOI-SMS-BERa`	406 874	88.90
8	VOI, SMS, BERb	`MOD_M8_VOI-SMS-BERb`	162 989	35.60
9	VOI, SMS, BERa, CNA	`MOD_M9_VOI-SMS-BERa-CNA`	333 064	72.70
10	VOI, SMS, BERb, CNA	`MOD_M10_VOI-SMS-BERb-CNA`	153 419	33.50

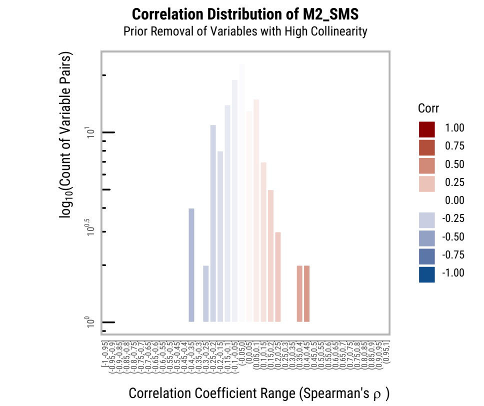
Table 5.1: The number of observations per dataset that are available after modelling data preparation.

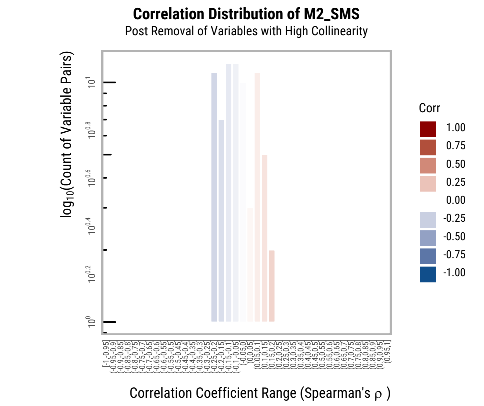
5.2 Correlation Analysis

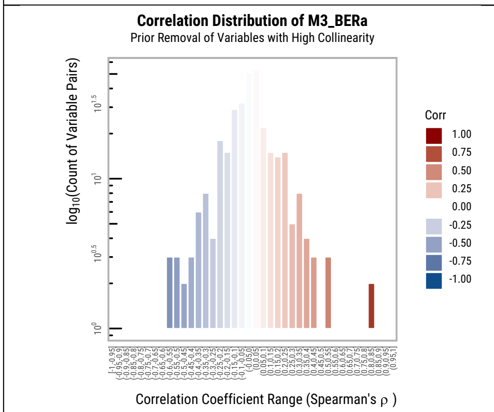
We conduct a correlation analysis on each modelling dataset to identify any potential new collinear features that may emerge when merging data from multiple services. We employed functionality from the `R caret` package, `findCorrelation` function, to identify possible variables with high correlation to other predictors before removing selected cases that made sense (Jed Wing et al., 2019). Table 5.2 compares the histograms showing the *count of variable pairs* within bins of *correlation ranges* before versus after removing selected variables that correlate highly with other variables. Table 5.3 and Table 5.4 show correlation plots for the `M9_VOI-SMS-BERa-CNA` and `M10_VOI-SMS-BERb-CNA` datasets, which are the two datasets that contain most services and features – before and after highly collinear predictor variables were avoided.

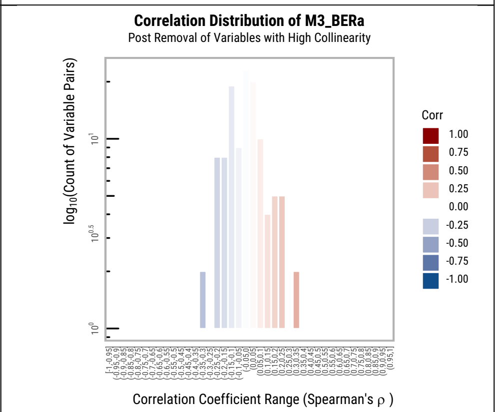
<table>
<tr><th>Before</th><th>After</th></tr>
<tr><td>


</td><td>


</td></tr>
<tr><td>


</td><td>


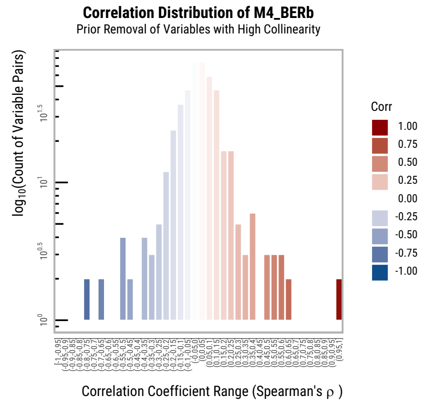
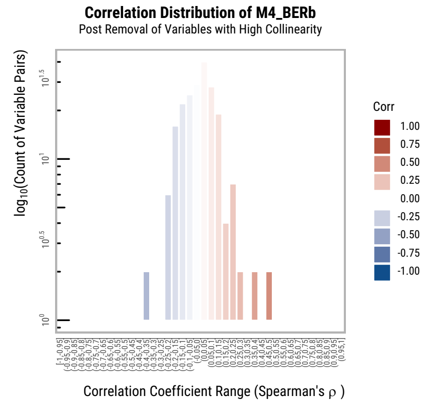
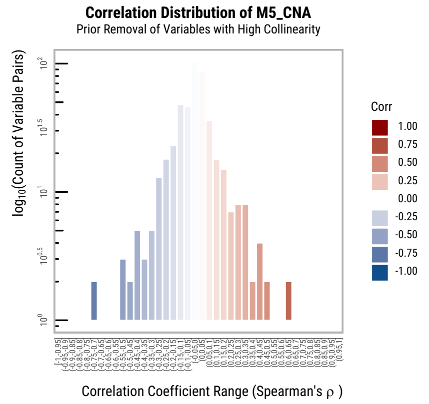
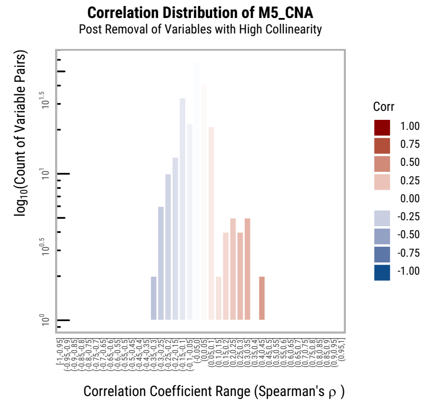
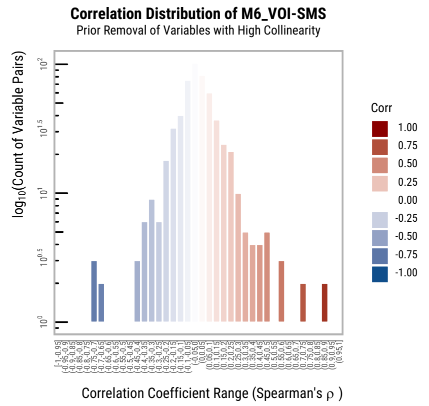
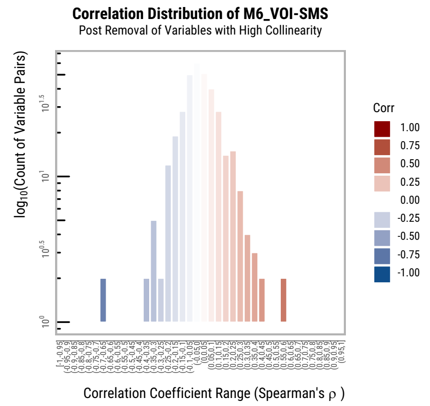
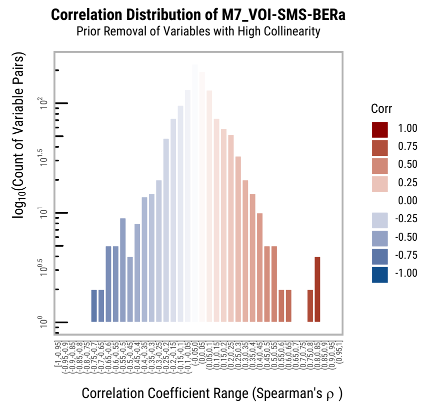
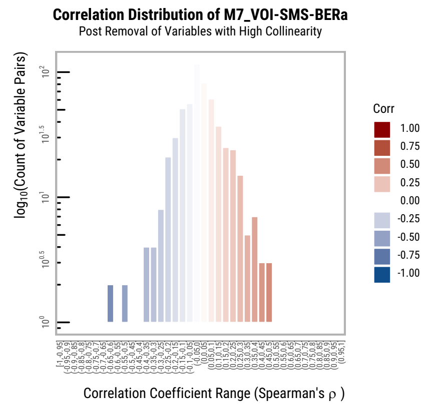
</td></tr>
<tr><td>


</td><td>


</td></tr>
</table>

Continued on next page

Table 5.2 – continued from previous page

Before	After
	
	
	
	

Continued on next page

Table 5.2 – continued from previous page

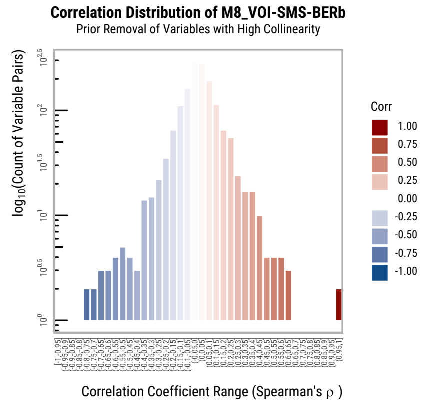
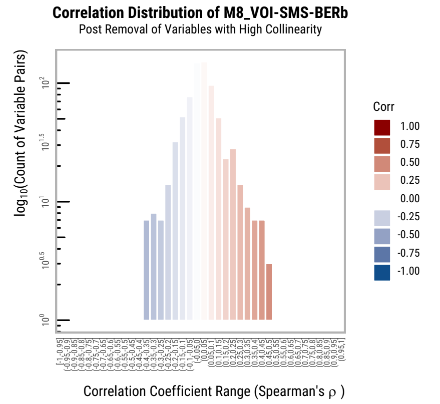
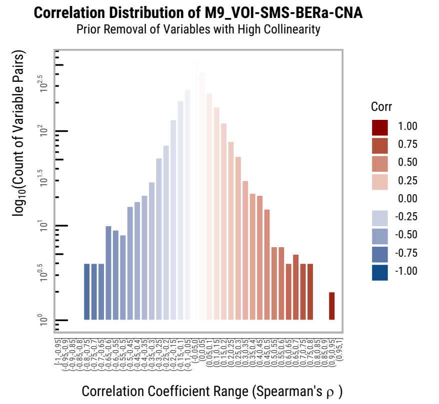
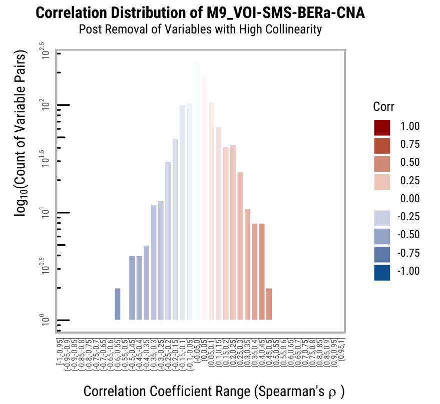
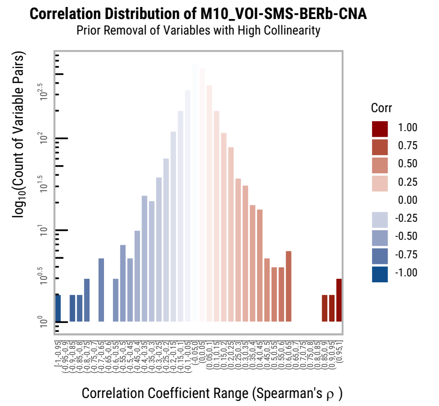
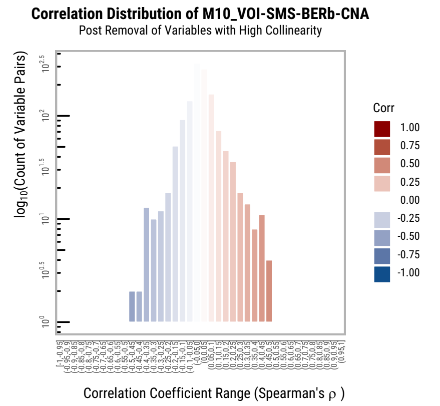
Before	After
	
	
	

Table 5.2: Distribution of the Spearman's ρ correlation coefficient between predictor variables of the ten modelling datasets before and after the removal of highly correlated variables. *Left*: Before removal of highly correlated variables. *Right*: After removal of highly correlated variables.

Before

After

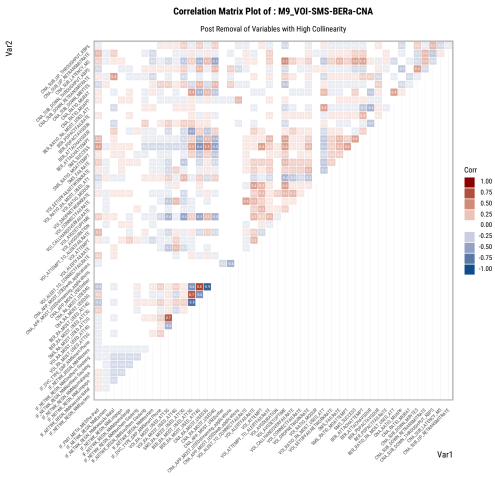
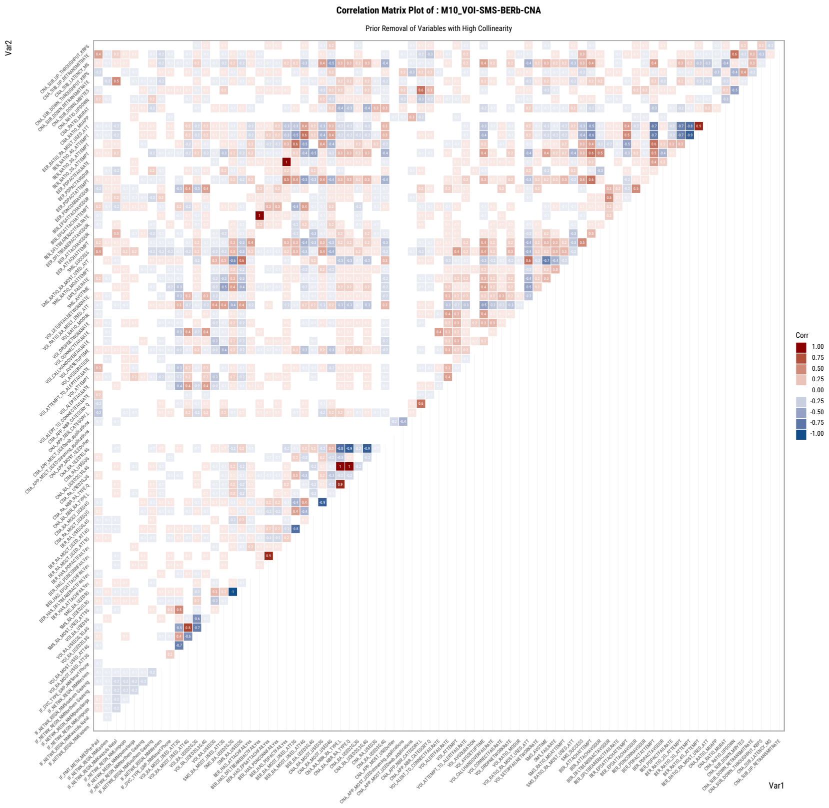


TABLE 5.3: Comparison of the Spearman's ρ correlation plots of model dataset `MOD_M9_VOI-SMS-BERa-CNA` before vs. after the removal of variables that are highly correlated. *Top*: Before removal of highly correlated variables. *Bottom*: After removal of highly correlated variables.

Before



After

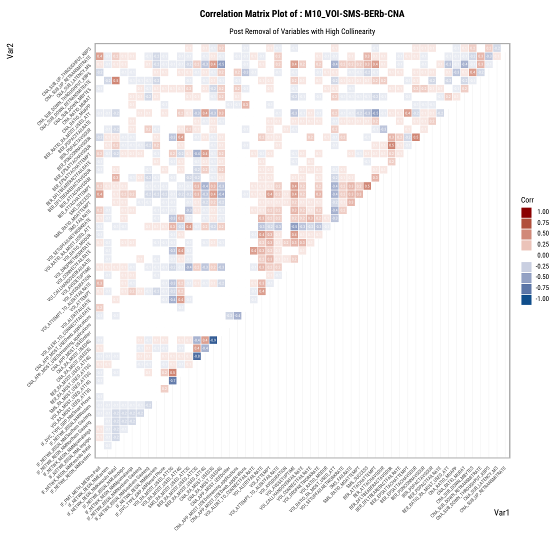


TABLE 5.4: Comparison of the Spearman's ρ correlation plots of model dataset `MOD_M10_VOI-SMS-BERb-CNA` before vs. after the removal of variables that are highly correlated. *Top*: Before removal of highly correlated variables. *Bottom*: After removal of highly correlated variables.

5.3 Variable Transformation Detail

Model performance was weighed up against various preprocessing options – including basic scaling by standardisation, as well as *"Normalising"* transformations of numerical variables by making use of the `R` software package `bestNormalize`. We specifically provide detail on the normalising transformations that were used, as they produced better performing models during training and validation.

Normalising Transformations

During the Exploratory Data Analysis (EDA) of the Customer Experience (CE) metrics, covered in Section 4.3, we noted that some of the numerical variables in their unprocessed form have highly skewed distributions, or are on very different scales compared to the rest of the variables in the dataset. Such variations in model features can lead to difficulties during the training of certain model types, or may lead to biased model performance due to feature bias toward variables with large ranges or orders of magnitude differences. Therefore, prior to model training we perform feature scaling to standardise the ranges of independent variables and we make use of the `bestNormalize` `R` package as an additional pre-processing step to select the best normalising transformations to use for each variable.

The `bestNormalize` package is built to estimate the best normalising transformation for a vector consistently and accurately. The package will look at a range of possible transformations and return the best one that makes the data look the most 'normal' (Gaussian distribution). The package introduces a new adaptation of a normalisation technique, which is called *Ordered Quantile normalisation* (ORQ). ORQ transforms the data based off of a rank mapping to the normal distribution and normally distributed transformed data is guaranteed – conditional that ties are not present in the data (Peterson and Cavanaugh, 2019). Finally, the functionality of `bestNormalize` can be integrated into a machine learning workflow via recipes, which forms part of the `tidymodels` framework, which we employed to build machine learning pipelines for the present project.

Package `bestNormalize`

The `bestNormalize` package selects from the following list of transformations to perform normalisation, and selects the best option by performing repeated cross-validation to estimate the Pearson's P statistic divided by its degrees of freedom to calculate the *"Normality statistic"*, and selecting the transformation that produces the lowest value:

Box-Cox

$$g(x;\lambda) = \mathbf{1}_{(\lambda\neq 0)}\frac{x^\lambda - 1}{\lambda} + \mathbf{1}_{(\lambda=0)}\log x, \tag{5.1}$$

Yeo-Johnson

$$\begin{aligned} g(x;\lambda) = {} & \mathbf{1}_{(\lambda\neq 0, x\geq 0)}\frac{(x+1)^\lambda - 1}{\lambda} \\ & + \mathbf{1}_{(\lambda=0, x\geq 0)}\log(x+1) \\ & + \mathbf{1}_{(\lambda\neq 2, x<0)}\frac{(1-x)^{2-\lambda} - 1}{\lambda - 2} \\ & + \mathbf{1}_{(\lambda=2, x<0)} - \log(1-x), \end{aligned} \tag{5.2}$$

Arcsinh

$$g(x) = \text{arcsinh}(x) = log(x + \sqrt{x^2 + 1}), \tag{5.3}$$

Square Root

$$g(x) = \sqrt{x + a} \text{ , with } a = \max(0, -\min(x) + \epsilon), \tag{5.4}$$

Exponential

$$g(x) = \exp(x) = e^x, \tag{5.5}$$

Log

$$g(x) = \log_{10}(x + a) \text{ , with } a = \max(0, -\min(x) + \epsilon). \tag{5.6}$$

Normalising the Numeric VOI Predictors

Figure 5.1 shows the *Log Estimate Normality Statistic* results obtained from using the `bestNormalize` function to find the optimal transformation for each numerical predictor variable in the VOI dataset. The transformation producing the lowest *Normality* statistic value was selected by performing cross validation (CV) using 10 folds and 3 repeats, with an out-of-sample minimum value of the Pearson's P/df normality statistic for each variable. Table 5.5 contains a summary of the transformations and parameters selected through CV.

Predictor Variable	Transformation	Parameter	Value	$g(x)$	$\overline{g(x)}$	σ
`VOI_ALERT_TO_CONNECTFAILRATE`	*sqrt(x+a)*	a	-	$\sqrt{x}$	5.420	1.130
`VOI_ALERTFAILRATE`	*sqrt(x+a)*	a	-	$\sqrt{x}$	1.420	1.060
`VOI_ATTEMPT`	*Box-Cox*	λ	0.153	$\frac{x^{0.153}-1}{0.153}$	5.770	1.800
`VOI_ATTEMPT_TO_ALERTFAILRATE`	*arcsinh(x)*	-	-	$\log\left(x + \sqrt{x^2+1}\right)$	3.770	0.531
`VOI_AVGDURATION`	*Log(x)*	-	-	$\log_{10}(x)$	1.990	0.253
`VOI_AVGSETUPTIME`	*Box-Cox*	λ	0.602	$\frac{x^{0.602}-1}{0.602}$	1.780	0.751
`VOI_CALLHANDOVERFAILRATE`	*no-transform*	-	-	x	26.700	31.600
`VOI_CONNECTFAILRATE`	*arcsinh(x)*	-	-	$\log\left(x + \sqrt{x^2+1}\right)$	2.500	1.020
`VOI_DROPNETWORKRATE`	*arcsinh(x)*	-	-	$\log\left(x + \sqrt{x^2+1}\right)$	0.629	0.743
`VOI_RATIO_MODUR`	*Box-Cox*	λ	0.725	$\frac{x^{0.725}-1}{0.725}$	-0.664	0.323
`VOI_RATIO_RA_MOST_USED_ATT`	*Box-Cox*	λ	2.000	$\frac{x^{2.000}-1}{2.000}$	-0.089	0.108
`VOI_SETUPFAILNETWORKRATE`	*arcsinh(x)*	-	-	$\log\left(x + \sqrt{x^2+1}\right)$	1.070	0.792

TABLE 5.5: The best normalising transformations and parameter values for the numerical predictor variables of the VOI dataset. The optimal transformations were obtained with `R` package `bestNormalize` (Peterson and Cavanaugh, 2019) through the comparison of out-of-sample values of the Pearson's P/df normality statistic using cross validation with 10 folds and 3 repeats (`k=10, r=3`).

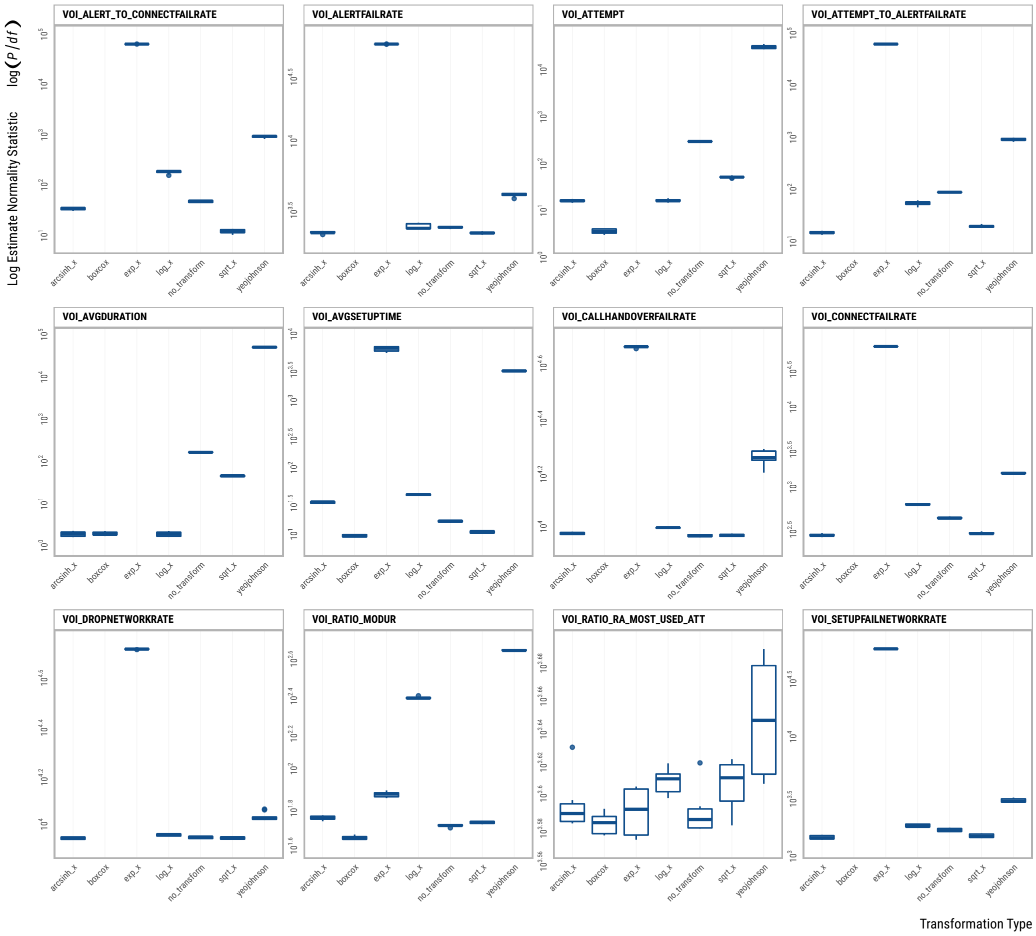


FIGURE 5.1: Selection of the best normalizing transformation for each numerical variable of the `VOI` dataset. The transformation with a minimum value of the Pearson's P/df normality statistic is used as selection criteria.

5.4 Results: Logistic Regression Model

We provide results obtained from using 3 000 random samples from the `M10_VOI-SMS-BERb-CNA` modelling dataset, which contains representative data across all services and radio technologies. We perform training and validation of two *Logistic Regression* model workflows using all predictors from the dataset. We compare model performance by training a simple model without any variable transformations, and a second model where we make use of the Yeo-Johnson transformation to normalise numeric predictors. The resampling strategy that we use is 10-fold cross validation with 3 repeats, with 70% training, 20% validation and 10% testing data.

We “drop” *Neutral* respondents from the data before sampling of the training, validation and test datasets. We do this under the current hypothesis, that the neutral group of respondents are indifferent in their response to the NNPS survey and that the neutral group does not contain information regarding the input predictor relations to *detractor* probability. We are of the opinion that the neutral responses are essentially noise and that the stable *Neutral* sample proportions we see in graphs like Figure 4.2a, are simply the natural proportions of respondents that are indifferent while filling in the NNPS survey.

5.4.1 Simple Logistic Model without Variable Transformation

The ROC curve in Figure 5.2 indicates weak separabilitly between predicted classes as evidenced by the curves being close to the 45 degree line. The *Area Under the Curves* (AUCs) for the in sample data are estimated to be 0.647, and 0.592 for the out of sample observations. The AUC values are close together, so the model is likely not overfitting. If there were substantial differences between the two values such that the out of sample value was much lower that the in sample value the model would not be generalising well to new unseen data, and therefore match the training data too closely – thereby overfitting.

Overall, this suggests that the inference we can draw is weak. We must conclude that either this model is not capturing the structure of the data-generating process correctly, or that any signal is being dominated by noise. We suspect that it may be the latter. Despite detecting significant associations after accounting for the presence of all variables, prediction performance remains relatively weak under this model.

Model-Fit Output

The `R` model-fit output of the simple Logistic Regression model is provided in Table 5.6.

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
3006	2168	-1423	2952	3253	2846	2116	2169

Table 5.6: Model fit output for a simple Logistic Regression model containing all predictor variables from the M10 dataset without any variable transformations.

Model Interpretation

The model-fit output indicates that there are at least one significant association in the model based on the χ^2 test.

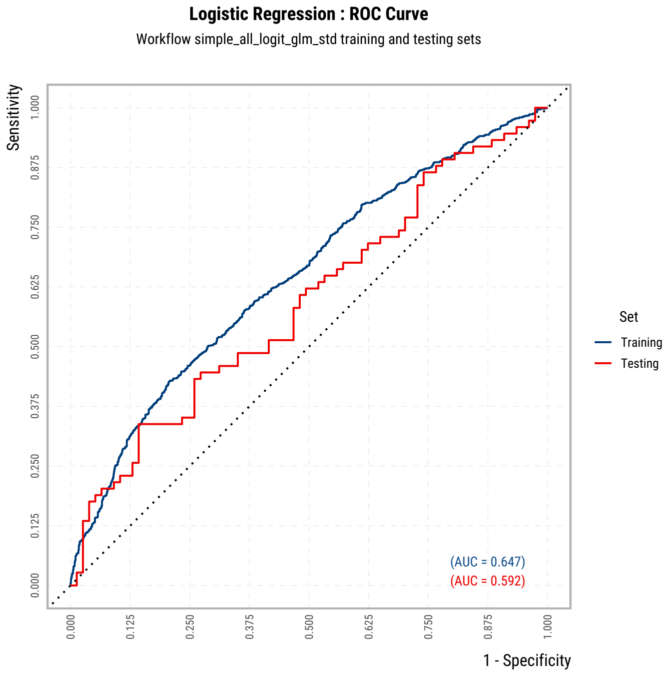


FIGURE 5.2: ROC curves for the simple logistic regression model without variable transformation. The model was trained by doing 10-fold cross validation with 3 repititions on a training set of 2 249 samples, and testing set of 151 samples from the M10 dataset covering all services and subscriber segmentation features.

χ^2 Test:

We get the below from the `R` model-fit output based on the fitting of 3 000 random samples to the `M10_VOI-SMS-BERb-CNA` modelling dataset.

Difference in degrees of freedom between a null model and the proposed model:

- $df_{null} = 2168$ (Null deviance degrees of freedom),
- $df_{resid} = 2116$ (Residual deviance degrees of freedom),

$$\Delta df = (df_{null} - df_{resid}) = (2168 - 2116) = 52.$$

Difference in deviance between a null model and the proposed model:

- $deviance_{null} = 3006$ (Null deviance),
- $deviance_{resid} = 2846$ (Residual deviance),

$$\Delta deviance = (deviance_{null} - deviance_{resid}) = (3006 - 2846) = 160.$$

The critical value for a χ^2 distribution with $\alpha = 0.05$ (95% confidence interval) and $df = 52$ is approximately equal to $\chi^2(\alpha = 0.05, df = 52) \approx 70$. The $\Delta deviance$ is thus large compared to the χ^2 critical value, and therefore we conclude that at least one significant association exists in the data.

Variable Importance

The variable importance plot in Figure 5.3 and the model output in Table 5.7 indicates that the top five *significant predictors* in our model are `IF_PMT_METH_MED_Pre.Paid`, `CNA_SUB_DOWN_MBYTES`, `BER_RA_MOST_USED_ATT_X3G`, `IF_NETWK_REGN_NM_Eastern` and `VOI_ATTEMPT`. The model output indicates low p for the top five variables which indicates that statistically there is strong evidence for the variables to be influential in the model prediction.

This suggests that:

1. The *Pre-Paid* subscribers are more likely to be *Detractors*. This does not align with our observations in Figure 4.5. Migh be due to the fact that the *Pre-Paid* subscriber base is the largest group as substantiated by Figure 4.4.

2. Subscribers with *higher download volumes* are more likely to be *Promoters*. This does intuitively make sense if we make the assumption subscribers need a good stable network in order to achieve higher download volume. We don't have statistical evidence in the EDA to substantiate.

3. Subscribers that are *more active on 3G* technology for data access are more likely to be *Promoters*. This does intuitively make sense because 3G would imply better data experience compared to 2G. The proportion of subscribers that are ever active on 3G is 20% less than those that are active on 2G as indicated in Table 4.8.

4. Subscribers from *Eastern region* are more likely to be *Detractors*. This does not align with our observations in Figure 4.8 and Figure 4.9.

5. Subscribers with *more voice call attempts* are more likely to be *Detractors*. Intuitively this does make sense, because multiple call attempts may imply re-connection attempts due to calls being dropped, or subscribers not being able to successfull setup calls. The `SETUPFAILNETWORKRATE` and `DROPNETWORKRATE` variables may substantiate the claim, but these two variables are not significant in the prediction of *Detractors* and their direction of association is negative.
   An alternative explanation might be higher call costs associated with more calls, which may be annoying to subscribers.

Insignificant cases of interest includes:

1. `VOI_DROPNETWORKRATE` with association direction *negative* – opposite to the expected. The significance statistic is not high (-1.82), and the p value is high – not significant (0.07).
2. `VOI_SETUPFAILNETWORKRATE` with association direction *negative* – opposite to expected.
3. `VOI_CALLHANDOVERFAILRATE` with association direction *negative* – opposite to expected.

All three these variables are informed by experience to be key metrics to measure user experience in mobile networks.

term	estimate	std.error	statistic	p.value	Association
IF_PMT_METH_MED_Pre.Paid	0.419	0.12	3.49	0.000489	Positive
CNA_SUB_DOWN_MBYTES	-0.0000596	0.0000175	-3.41	0.000648	Negative
BER_RA_MOST_USED_ATT_X3G	-1.09	0.322	-3.39	0.000705	Negative
IF_NETWK_REGN_NM_Eastern	0.827	0.278	2.97	0.00297	Positive
VOI_ATTEMPT	0.00143	0.000532	2.69	0.00705	Positive

TABLE 5.7: Coefficient summary of the Logistic Regression model with no transformation of the predictor variables.

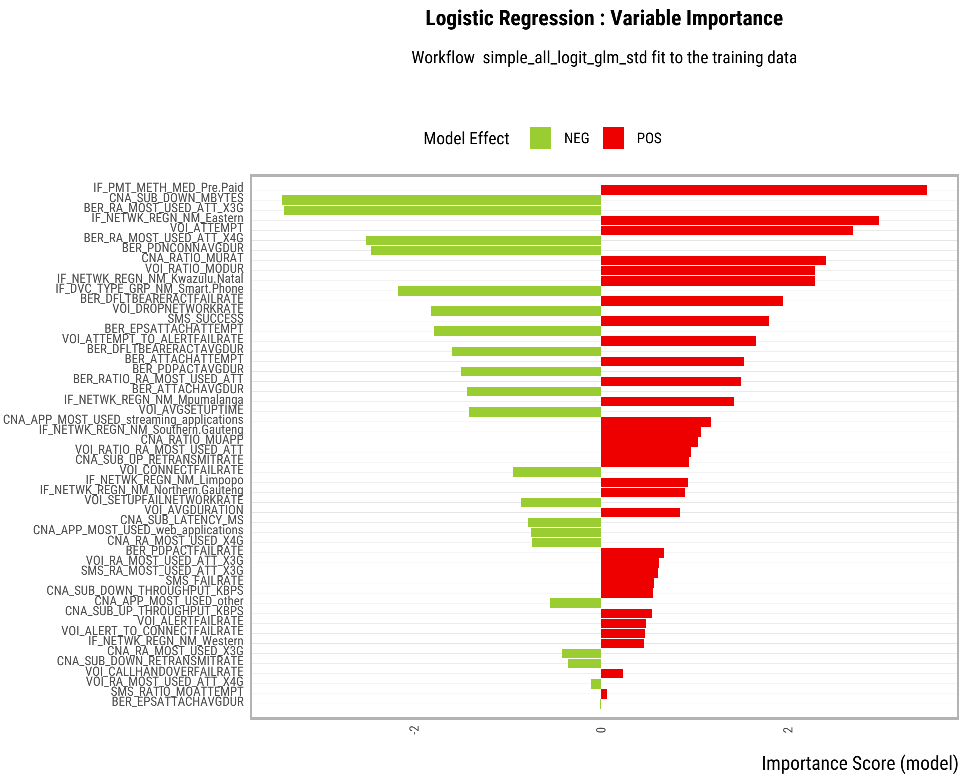


FIGURE 5.3: Variable importance plot of the simple Logistic Regression model without transformation of the input predictors.

5.4.2 Logistic Model with Yeo-Johnson Transformation

The ROC curve in Figure 5.4 indicates weak separabilitly between predicted classes as evidenced by the curves being close to the 45 degree line. The AUCs for the in sample data are estimated to be 0.654, and 0.605 for the out of sample observations. The AUC values are very similar to the fit of the simple logistic model, and models are therefore performing similar.

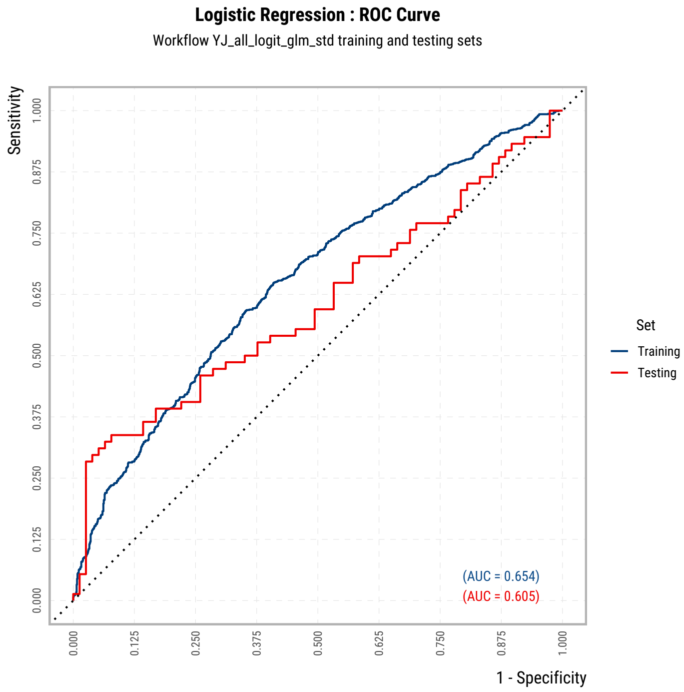


FIGURE 5.4: ROC curves for the logistic regression model with Yeo-Johnson transformation of the numerical variables. The model was trained by doing 10-fold cross validation with 3 repititions on a training set of 2 249 samples, and testing set of 151 samples from the M10 dataset covering all services and subscriber segmentation features.

Overall, this suggests that the inference we can draw is still weak. We must conclude that the variable transformation did not improve the model performace and must therefore conclude that either this model is not capturing the structure of the data-generating process correctly, or that any signal is being dominated by noise. We suspect that it may be the latter. Despite detecting significant associations after accounting for the presence of all variables, prediction performance remains relatively weak under this model.

Model-Fit Output

The `R` model-fit output of the Logistic Regression model with the Yeo-Johnson transformation applied to all predictors is provided in Table 5.8.

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
3016	2175	-1425	2957	3258	2851	2123	2176

TABLE 5.8: Model fit output for a simple Logistic Regression model containing all predictor variables from the M10 dataset with Yeo-Johnson transformation applied to all numerical predictor variables.

Model Interpretation

The model-fit output indicates that there exists at least one significant association in the model based on the χ^2 test.

χ^2 Test:

Difference in degrees of freedom between a null model and the proposed model:

- $df_{null} = 2175$ (Null deviance degrees of freedom),
- $df_{resid} = 2123$ (Residual deviance degrees of freedom),

$$\Delta df = (df_{null} - df_{resid}) = (2175 - 2123) = 52.$$

Difference in deviance between a null model and the proposed model:

- $deviance_{null} = 3016$ (Null deviance),
- $deviance_{resid} = 2851$ (Residual deviance),

$$\Delta deviance = (deviance_{null} - deviance_{resid}) = (3016 - 2851) = 165.$$

The critical value for a χ^2 distribution with $\alpha = 0.05$ (95% confidence interval) and $df = 52$ is approximately equal to $\chi^2(\alpha = 0.05, df = 52) \approx 70$. The $\Delta deviance$ is thus large compared to the χ^2 critical value, and therefore we conclude that at least one significant association exists in the data.

The variable importance plot in Figure 5.5 and the model output in Table 5.9 indicates that the top five *significant predictors* in our model are `CNA_SUB_DOWN_MBYTES`, `VOI_ATTEMPT`, `BER_RA_MOST_USED_ATT_X3G`, `IF_PMT_METH_MED_Pre.Paid` and `CNA_RATIO_MURAT`.

This suggests that:

1. Subscribers with *higher download volumes* are more likely to be *Promoters*.
2. Subscribers with *more voice call attempts* are more likely to be *Detractors*.
3. Subscribers that are *more active on 3G* technology for data access are more likely to be *Promoters*.
4. *Pre-Paid* subscribers are more likely to be *Detractors*.
5. Subscribers with greater proportions of data use on their *most used radio type* are more likely to be *Detractors*.

Insignificant cases of interest includes:

1. `VOI_DROPNETWORKRATE` with association direction *negative* – opposite to the expected. Surprisingly the significance statistic is reasonably high (-2.66), but the p value is not very significant (0.008).

2. `VOI_SETUPFAILNETWORKRATE` with association direction *negative* – opposite to expected.

3. `VOI_CALLHANDOVERFAILRATE` with association direction *negative* – opposite to expected.

All three these variables are informed by experience to be key metrics to measure user experience in mobile networks.

Comparing results to the simple Logistic model without variable transformations, we notice that four of the top five ranked variables re-appear, and in all cases the directions of associations are the same and p values comparable. The `CNA_SUB_DOWN_MBYTES` stands out as it has a similar coefficient value to the simple model and a lower p value which does imply that there is substantial statistical evidence to support the claim that the variable is influential in both these models.

term	estimate	std.error	statistic	p.value	Association
CNA_SUB_DOWN_MBYTES	-0.0762	0.0203	-3.76	0.000171	Negative
VOI_ATTEMPT	0.157	0.0493	3.19	0.00142	Positive
BER_RA_MOST_USED_ATT_X3G	-0.873	0.298	-2.93	0.0034	Negative
IF_PMT_METH_MED_Pre.Paid	0.365	0.127	2.87	0.00408	Positive
CNA_RATIO_MURAT	0.928	0.337	2.75	0.00588	Positive

TABLE 5.9: Coefficient summary of the Logistic Regression model with Yeo-Johnson transformation of the predictor variables.

Variable Importance

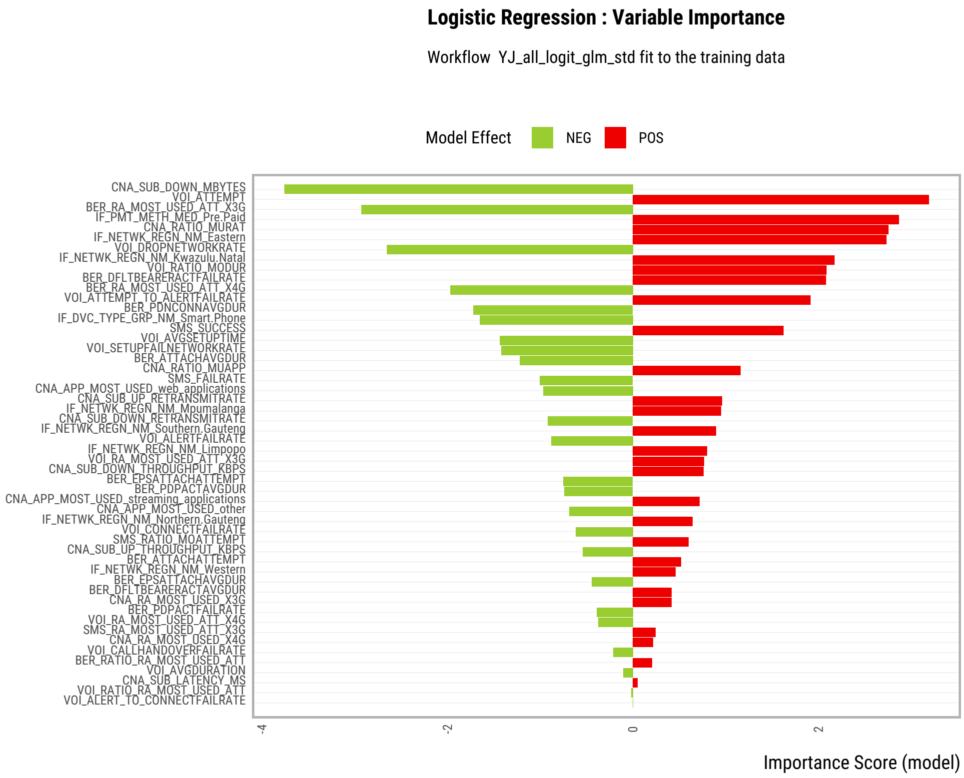


FIGURE 5.5: Variable importance plot of the simple Logistic Regression model with Yeo-Johnson transformation of the numerical variables.

5.5 Results: Tree Based Models

We provide results obtained from using 3 000 random samples from the `M10_VOI-SMS-BERb-CNA` modelling dataset, which contains representative data across all services and radio technologies. We perform training and validation of *Classification Tree* models using the `rpart` package, and we inlcude all predictors from the dataset. We compare validation graphs and tree model structure for different values of the `cp` value. The training dataset consists of 70% of the data.

In Figure 5.6 we plot the validation curve of the model fit with *seed(1)* and $cp = 0.01$

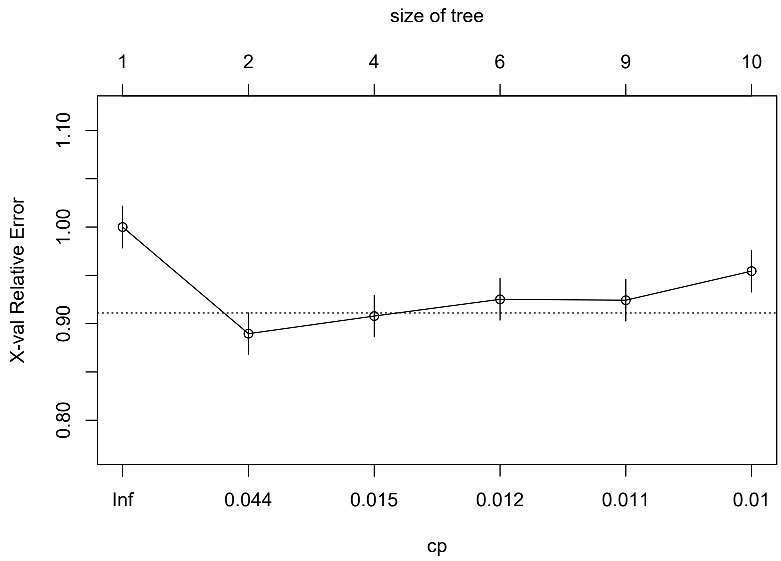


FIGURE 5.6: Classification Tree validation plot with *seed(1)* and $cp = 0.01$

In Figure 5.7 we plot the validation curve of the model fit with *seed(2)* and $cp = 0.01$

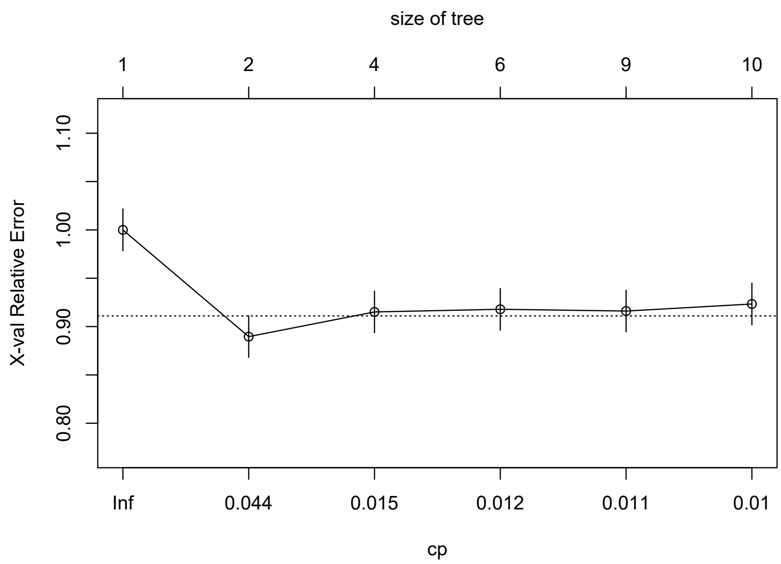


FIGURE 5.7: Classification Tree validation plot with *seed(2)* $cp = 0.01$

We notice that a change of the random seed value with similar model *cp* fit parameters produces *different* validation curves every time that the model fits. This points to the random nature of the tree algorithm that performs splits of the data using a greedy algorithm.

In Figure 5.8 we plot a complex Classification Tree model with *seed(1)* and $cp = 0.01$. We notice that the primary determinant for tree branching is the *Payment method* = "Post-Paid", attribute at the root node of the tree, followed by `DROPNETWORKRATE` and `VOI_ATTEMPT`. *Payment Method* was a significant variable in the Logistic Regression models analysed in Section 5.4 as well.

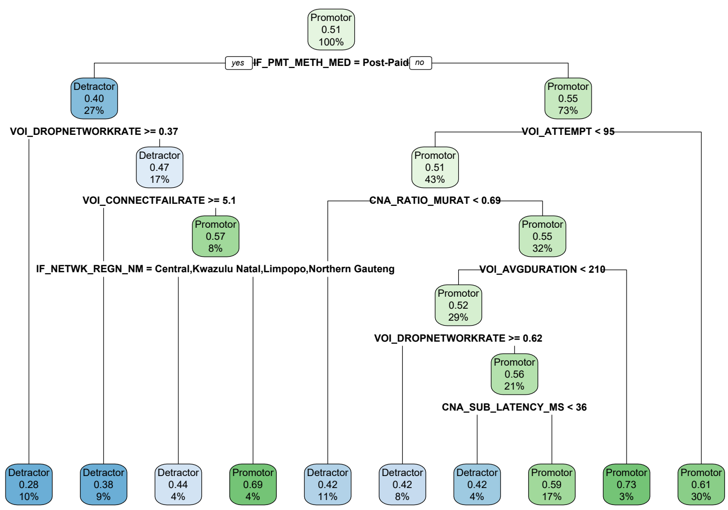


FIGURE 5.8: Complex Classification Tree with a high *cp* value of 0.01

We perform pruning of the complex tree by lowering the *cp* value to the first minimum value of the curve that falls below the relative error line on the plot. We note that this will produce a tree with two branches. In Figure 5.10 we show the pruned tree with *cp* value of 0.044 producing a tree with 2 branches. We have *Payment method* = "Post-Paid" at the root branch indicating that it's the most important variable to split on based on the greedy random algorithm.

In Figure 5.9 we plot the validation curve of the model fit with *seed(1)* and $cp = 0.016$

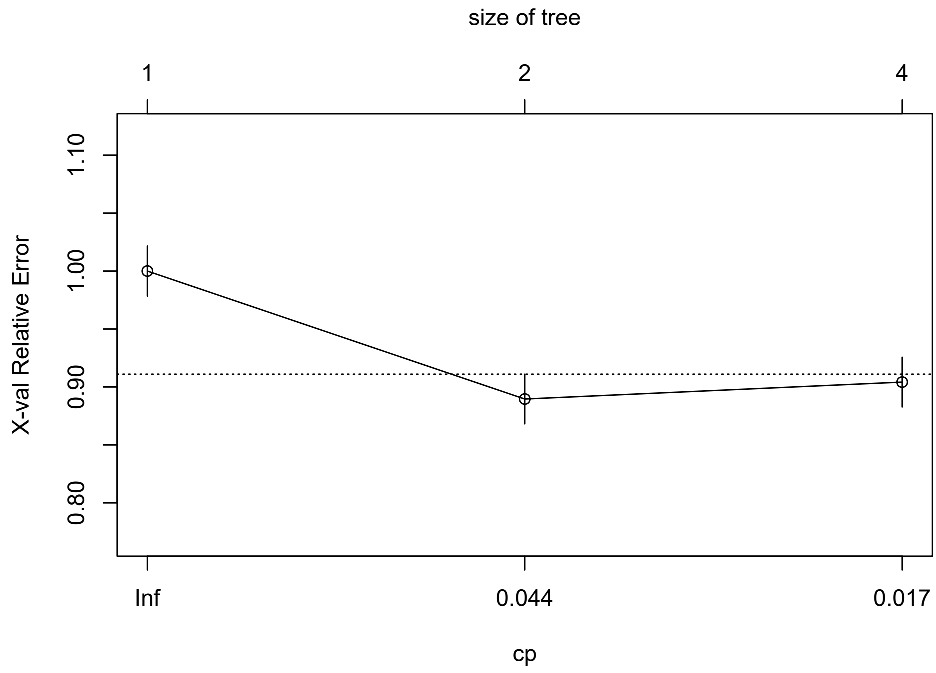


FIGURE 5.9: Classification Tree validation plot with *seed(1)* and *cp* value of *cp* = 0.016

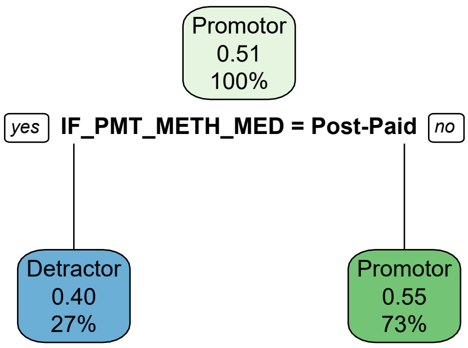


FIGURE 5.10: Pruned Classification Tree with a *cp* value of 0.016

Chapter 6

Conclusion

This thesis aimed to determine whether classification machine learning models could accurately predict the Net Promoter Score (NPS) of mobile network subscribers, using their cellular network experience data from probing as a primary data source. However, the study concluded that it was inconclusive in meeting its primary objective. None of the machine learning algorithms managed to detect meaningfull signal within the data to reliably and consistently predict NPS detraction. We establish that there are only weak associations between the present dataset and the NPS outcome variable and no meaningfull inference can be made from the data analysis. This suggests that NPS may instead be driven by other factors, such as pricing or the interaction of customers with other processes that are more important and not represented within the present data.

The practical implementation of a machine learning pipeline that can use probing data as input and provide NPS class predictions with reasonable accuracy proved challenging, mainly due to the structure of the current dataset at its primary source. The present probing experience measurements are not ideally structured for modelling and machine learning. As a result, much time is spent preparing data that could be used for modelling. The cost of collecting and restructuring the data to be usable for modelling does not justify the present levels of accuracy obtained. However, more time spent on modelling might prove otherwise. To further explore the viability of the present use case in a real-life implementation, we therefore propose using differently structured data with machine learning as an end goal in mind from the start. An ideal implementation would be a machine learning pipeline that continuously collects data from all sources and structures them into an aggregated form at the subscriber level based on typical usage patterns.

Further, exploring whether additional data can improve current model accuracy would be beneficial. Some ideas for future improvements to input data include:

- Include additional profiling attributes from customer billing data, like their total spend on voice and data services.
- Include network alarms about site outages.
- Include site geographic information (latitude, longitude) of a subscriber's most used cell towers (sites).
- Perform sentiment analysis on verbatim responses to isolate network-related versus other factors.
- Include network element metrics like site availability (impact of load shedding) at the primary used location.

- Include call centre interactions, for example, customer complaints and reported network issues.
- Include radio network quality measures, like signal strength and radio signal quality, distance to the primarily used site, and indoor versus outdoor use.

Also, it would be beneficial to explore whether other types of models that were not considered for the present project, primarily due to the interpretability objective, may perform better at detecting patterns in the data. Some ideas for future improvements to models include:

- Consider the temporal aspect. The system (cell network) within which subscribers are generating the modelling data is changing over time, and this might have positive or negative effects on their experience, which can further result in changing patterns with regards to the observed responses – as we saw during EDA where NPS does fluctuate at the regional level. The aspect of changing patterns over time was not in the scope of the present project. Still, it can be modelled with time series prediction models, a possible future improvement to the current modelling approach. To presently mitigate this problem, we propose to regularly update the currently selected model with new training data of more recent NNPS survey response subjects, which entails the collection of new data from all sources, including probing, if we continue using the same collection strategy of a recent period before survey responses.
- Experiment with more complex modelling techniques like multi-layer neural networks.

We further acknowledge that our research is based on data from a single MNO within the South African market with a particular CEM system implementation. It makes the situation unique, and results might not generalised well to other markets.

Appendix A

Probing Data Preparation

The initial probing dataset collected for this project consisted of eight primary probing sources spanning two analytical platforms, three mobile technologies and eleven network interfaces, as shown in Table A.1.

Index	Source	Services	Interfaces	Technologies	Platform
1	`VOI`	CS Voice & VoLTE	A, IuCS, S11, S1-U	2G, 3G, 4G	Touchpoint
2	`SMS`	CS Short Message Service *SMS*	A, IuCS	2G, 3G	Touchpoint
3	`BER`	PS Bearer Access	Gb, IuPS, S1-MME	2G, 3G, 4G	Touchpoint
4	`CNA`	PS User-Plane Applications	Gn-UP, S1-U	2G, 3G, 4G	CNA
5	`GN-CP`	PS Control-Plane	Gn-CP	2G, 3G	Touchpoint
6	`GN-UP`	PS User-Plane	Gn-UP (HTTP)	2G, 3G	Touchpoint
7	`CSFB`	CS Fallback	SGs	4G	Touchpoint
8	`GTPv2`	PS Evolved Control-Plane	S11	4G	Touchpoint

CS: Circuit-switched, **PS**: Packet-switched

TABLE A.1: The eight major probing sources that were identified as possible data sources for this project.

Each probing source provides information about the individual subscriber experiences with a particular *network service*, for example, the *VOI* dataset covering the voice service and *BER* covering packet-switched network bearer access.

Within each service, we have a logical separation of measurements into subsets, which may cover up to three network technologies, for example, *2G*, *3G* and *4G* technologies for voice. Chapter 2, Section 2.1, details all probing sources, services and technologies identified as potential data sources for this project. The combination of eight services with their associated interfaces spread over multiple technologies resulted in seventeen individual probing datasets that had to be merged per subscriber to get a full view of all network activity.

The first iteration of EDA on the raw collected dataset pointed out several issues regarding incomplete or unstable data from some sources. The data also pointed out a fair amount of complexity associated with subscribers using varying combinations of services and underlying technologies. What evolved from this initial EDA was that two datasets, namely *GN-CP* and *GN-UP*, were identified as being too unstable and untrustworthy for use in the analysis and were therefore excluded from the project. In addition, there were two more datasets, namely *CSFB* and *GTPv2*, which were found to include activity from only a tiny subset of the survey responders (approximately 32%) – due to the recency of the technology and it not being used by many subscribers at the time. The *GTPv2* service also had missing data for one month out of the six months used for this project.

The first section of the present chapter, Section A.1 EDA 1st Iteration – Cleanup Raw Extracted Data, was included for completeness – to show the findings which resulted in the mentioned datasets being dropped from the analysis in subsequent EDA and modelling sections.

A.1 EDA 1st Iteration – Cleanup Raw Extracted Data

Samples per Service

Table A.2 shows the observations for each service obtained from our probing sources. Out of the 627 156 survey responders, 627 112 subscribers had network activity captured by probing sources on at least one network interface during the observation period. However, due to subscribers with different usage patterns, devices with varying capabilities, and the network not providing the same radio coverage layers everywhere, we find that the subscribers from our survey responder group have varying representations within each data source.

Service	Count	%
NNPS	627 156	100.0
VOI	624 299	99.6
SMS	623 308	99.4
BER	560 708	89.4
CNA	492 286	78.5
GN-UP	405 519	64.7
GN-CP	400 128	63.8
CSFB	226 625	36.1
GTPv2	197 941	31.6

UP: User-plane
CP: Control-plane

Table A.2: The number of probing observations that were collected per service before the data was cleaned.

We notice that the most used services within the targeted subscriber segments are circuit-switched *voice* and *SMS*. These two services, represented by the VOI and SMS interfaces in our dataset, have the best coverage of greater than 99%. Data services, represented by interfaces BER, CNA and GN, have lower than 90% coverage. This is due to the fact that the majority of subscribers from the targeted market segments are not extensive data users. BER service has a higher coverage than CNA and GN as a result of subscribers that were connected to the mobile data network control-plane (CP), but did not use any data on the user-plane (UP).

Finally, we see very low coverage of some services that are specific to 4G. Less than 37% of the surveyed subscriber base had activity on the CSFB and GTPv2 interfaces. Based on this observation these two services will be disregarded during the modelling process for purposes of simplifying this analysis.

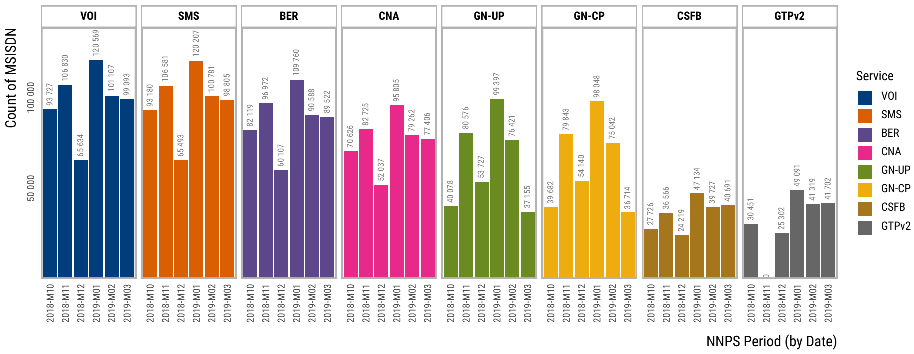


FIGURE A.1: Trend of the monthly probing samples per service during the observation period.

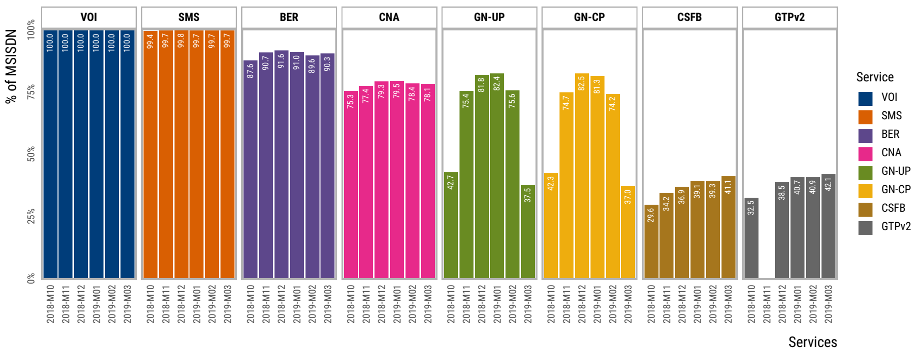


FIGURE A.2: Trend of the monthly probing sample ratio per service of the counts shown in Figure A.1.

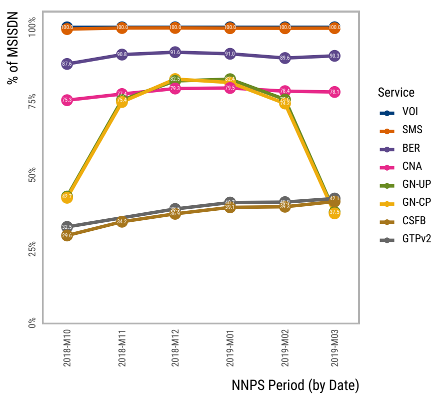


FIGURE A.3: Trend of the monthly probing sample ratio per service of the counts shown in Figure A.1.

Identify Small Sample Groups

Samples per Combination of Services

Table A.3 shows groupings of the various combinations of services that were used by the surveyed subscribers, arranged in descending order by the count of subscribers in each group. Combined the top three groups represent a total of 618 592 or 98.6% of the surveyed subscriber base, and all subscribers in these groups used both voice and SMS services. The bulk of the subscribers, represented by the top group with 515 102 or 82.1% of the samples, made use of all network services, i.e. voice, SMS and data and will therefore be represented in most of the probing datasets.

Following the above are two much smaller groups, representing a total of 103 490 or 16.5% of the total samples. These two groups represent those subscribers that did not make use of data services. The larger of the two groups, with 53 211 or 8.5% of the samples, represents the subscribers that did not even connect to the data network *(Data CP = 'N', Data UP = 'N')*. The exact reason for the behaviour is not known at this stage, but it is not uncommon and most probably as a result of low end devices that do not support data services. The smaller group of 8.0%, represents the subscribers that did connect to the data network, but did not use any data *(Data CP = 'Y', Data UP = 'N')*. This can be due to various reasons, for example subscribers that deliberately deactivated mobile data on their devices, or those subscribers that were not provisioned to use data services on the network at the time.

The highlighted groups, which have small observation counts below 1% of the total sample, have unusual permutations of services that seem to be out of the norm. The fourth group of 5 583 subscribers for example, based on the available probing information used data *(Data UP = 'Y')*, but did not establish any data connections *(Data CP = 'N')* which is technically impossible. Therefore we make the assumption that these smaller groups, which have a combined representation of only 1.4% of the total base, most probably have incomplete or missing data from the probing sources and therefore will be disregarded for the remainder of this analysis.

Voice	**SMS**	**Data** *(CP)*	**Data** *(UP)*	**Count**	**%**	**% Cumulative**	**Descending**	**% Descending**
Y	Y	Y	Y	515 102	82.14	82.14	627 112	100.00
Y	Y	N	N	53 211	8.49	90.62	112 010	17.86
Y	Y	Y	N	50 279	8.02	98.64	58 799	9.38
Y	Y	N	Y	5 583	0.89	99.53	8 520	1.36
N	Y	Y	Y	1 400	0.22	99.75	2 937	0.47
N	Y	N	N	556	0.09	99.84	1 537	0.25
N	Y	Y	N	495	0.08	99.92	981	0.16
Y	N	Y	Y	243	0.04	99.96	486	0.08
N	Y	N	Y	86	0.01	99.97	243	0.04
N	N	N	Y	70	0.01	99.99	157	0.03
Y	N	Y	N	52	0.01	99.99	87	0.01
N	N	Y	Y	21	0.00	100.00	35	0.01
N	N	Y	N	8	0.00	100.00	14	0.00
Y	N	N	Y	5	0.00	100.00	6	0.00
Y	N	N	N	1	0.00	100.00	1	0.00

CP: Control-plane, **UP**: User-plane

TABLE A.3: The number of observations in groups of services that were used in combination.

Samples per Radio Technology

Table A.5 below shows the count and percentage of the 618 592 (98.6%) subscribers from the top three groups in Table A.3, that were active on each radio type during the observation period. We see that on average the most used radio type is 2G; with 95%, and secondly 3G; with 83% of subscribers. The least used technology is 4G; used by only 42% of subscribers.

Radio Type	**Count**	**%**
2G	596 945	95.2
3G	520 252	83.0
4G	259 937	41.4

Table A.4: The number of probing observations per radio technology.

Radio Type	**Count**	**%**
2G	589 805	95.3
3G	514 088	83.1
4G	258 337	41.8

Table A.5: The number of probing observations per radio technology.

Figure A.4 shows the monthly trend of sample ratios for each radio technology. We see that the sample ratios of all radio types increased from October to November. In the following periods the sample ratios on 2G and 3G technologies show slight declines, but still continued to increase on 4G. This can be an indication that the probing systems that harvest data for 2G and 3G were not performing optimal between October and November, whereas the increasing ratio of 4G activity can probably be explained, as it aligns with the increase we saw for the penetration of *Smartphone* devices in Section 4.2 Network NPS by Device Type Figure 4.10. We therefore make the assumption that the increasing ratio of 4G active subscribers is caused by a combination of an increase in the number of 4G capable devices being introduced into the mobile operator's network; as subscribers switch over to newer high-end smart devices, together with an overlapping network expansion of 4G radio coverage over time.

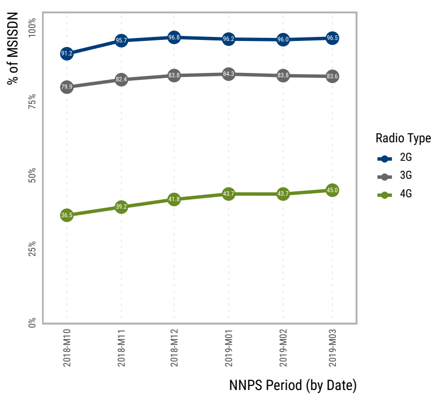


Figure A.4: Trend of the monthly Network Customer Experience probing sample ratio for the counts per radio technology.

The change in device composition is also visible in the ratio of samples per service as shown in Figure A.3. We see that there was a gradual increase in activity of packet-switched data services BER and CNA, which makes sense as these services are generally used more by Smartphone devices. We also see increased activity for CSFB and GTPv2 services that are specific to 4G. Section A.1 will give us insight as to how the use of radio types changed over time for each service.

Samples per Combination of Radio Technologies

In table A.6 below we show the number and percentage of subscribers for each combination of radio technologies that were used during the observation period ranked in descending order by count.

We see that there were mainly three subscriber groups that dominated based on the use of radio technologies. The top group; of 39.4%, represents subscribers that made use of all three radio technologies. The second group has approximately the same size as the first, and represents subscribers that used a combination of 2G and 3G. The third group is less than half the size of the former two; with only 16.4% samples, and represents 2G-only subscribers. The remaining highlighted groups represent uncommon combinations of radio technologies and are much smaller than the first three; with the largest one 2.7%. The combined size of these uncommon radio groups represent only 5.1% of the total base, and will be disregarded for purposes of simplifying this analysis.

Radio Types			Count	%
2G	3G	4G	243 814	39.4
2G	3G		241 509	39.0
2G			101 754	16.4
	3G		16 992	2.7
	3G	4G	11 773	1.9
2G		4G	2 728	0.4
		4G	22	0.0

TABLE A.6: The number of probing observations for each combination of radio technologies used by subscribers.

After removal of the low RA types we are left with 587 077 samples.

Of this 487 041 are using data (where *USEDATA* = '*Y*').

Filtering of cases where CNA and BER data are not complete leaves 472 573 samples.

Radio Type	Count	%
2G	587 077	100.0
3G	485 323	82.7
4G	243 814	41.5

TABLE A.7: The number of probing observations per radio technology after filtering top3.

Figure A.6 shows trends of the monthly ratio for each combination of radio technologies used by subscribers after filtering for the top 3 cases. We observe higher ratios for levels *3G only* and *3G/4G* in October. This was caused by degraded data captured from probing due to system issues at the time.

Radio Types			Count	%
2G	3G	4G	243 814	41.5
2G	3G		241 509	41.1
2G			101 754	17.3

TABLE A.8: The number of probing observations for each combination of radio technologies after filtering for top3.

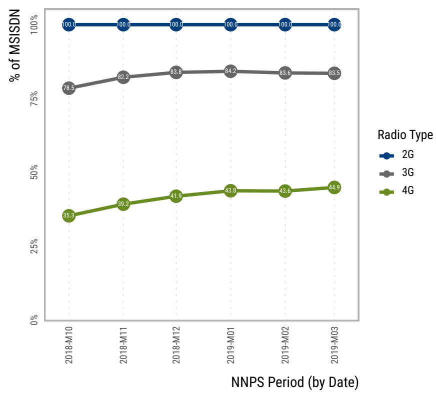
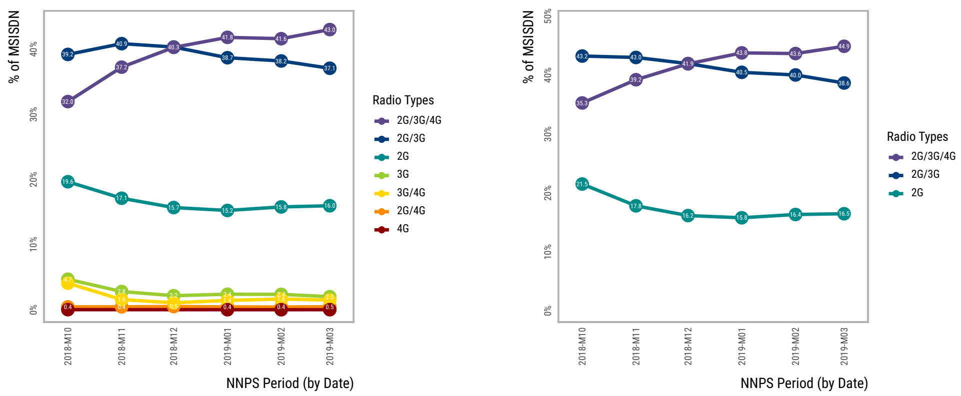


FIGURE A.5: Trend of the monthly probing sample ratio per radio technology after filtering for top3.



(a) Before filtering (b) After filtering

FIGURE A.6: Trend of the monthly ratio for each combination of radio technologies used together. Before vs. after filtering for top3 cases.

Samples per Radio Technology and Service

The trends below are obtained per radio technology and service after filtering for the top 3 cases.

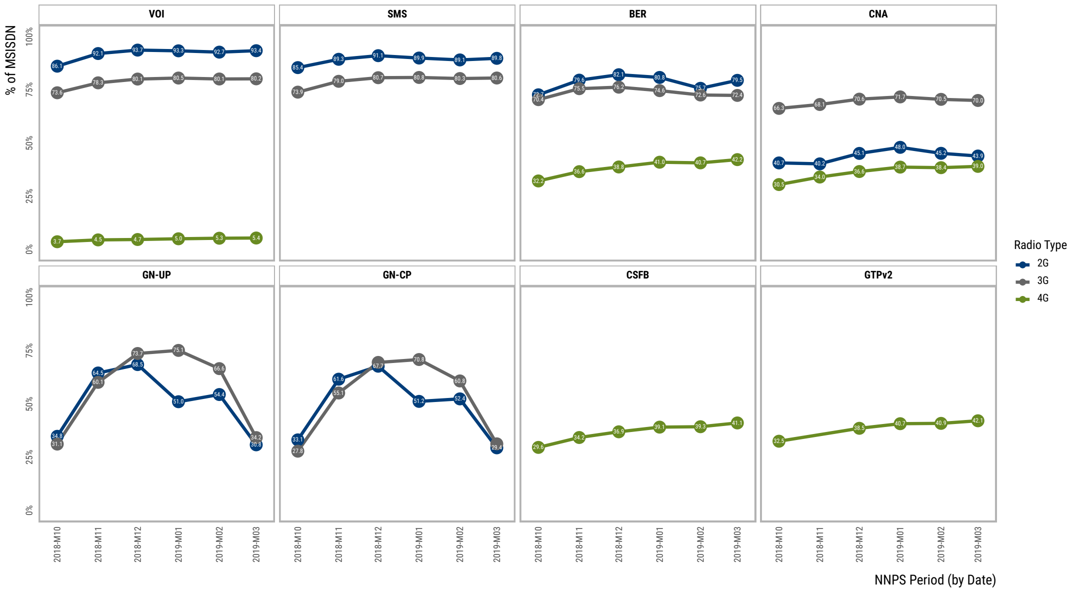


FIGURE A.7: Trend of the monthly Network Customer Experience probing sample ratio for the counts per service and radio technology.

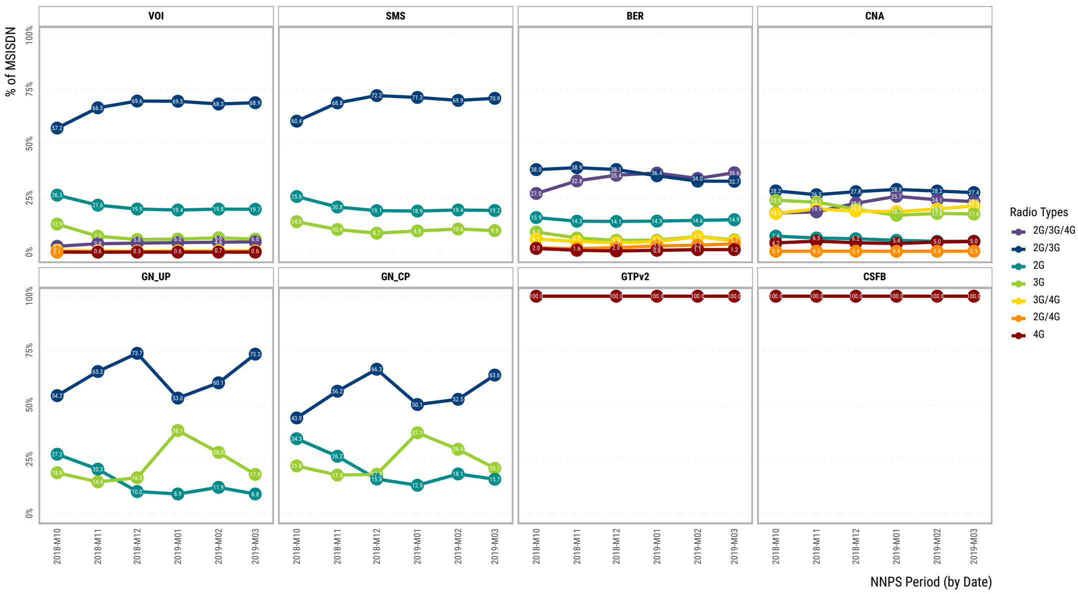


FIGURE A.8: Trend of the monthly probing sample ratio for combinations of radio technologies per service.

A.2 EDA 2nd Iteration – Isolate Clean Data for Modelling

Removal Based on Segmentation and Incomplete Probing

Start with 627 156 survey samples.

Disregarded Probing Services

627 095 samples remain after aligning with the remaining probing sources which will be used for modelling. We are disregarding these datasets from the original data sourced due to incomplete probing and suspected untrustworthy data as pointed out in Section A.1:

- GN-CP
- GN-UP
- GTPv2
- CSFB

Alignment of Sources

- Align survey and probing data sources: 627 095
- Check for complete device-information (type and radio capability): 626 821

Small Segmentation Groups

- Enterprise removed: 612 890
- Tablet & Basic Phone removed: 584 810

584 810 samples remain after checking for probing activity (on VOI, SMS, BER and CNA sources), confirming the device-type and device-capability, and removing the small segmentation groups: Enterprise, Tablet & Basic Phone.

Samples per Service

After removal of the disregarded probing datasets and re-aligning with the reduced survey data, we are left with the samples per probing source as shown in Table A.9.

Service	**Count**	**%**
VOI	582 313	99.6
SMS	581 577	99.4
BER	522 249	89.3
CNA	456 585	78.1

Table A.9: The number of probing observations per service for the reduced dataset. Sample counts as well as the percentage coverage against the NNPS survey dataset are shown for each service.

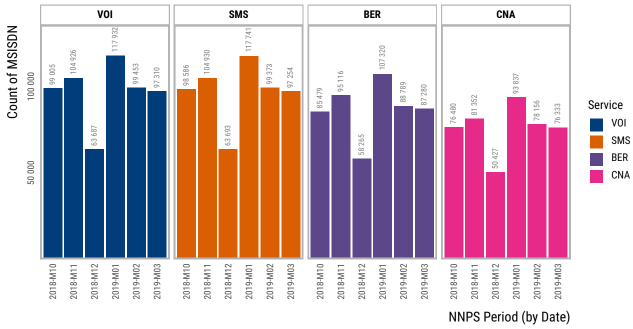


FIGURE A.9: Trend of the monthly probing samples per service during the observation period.

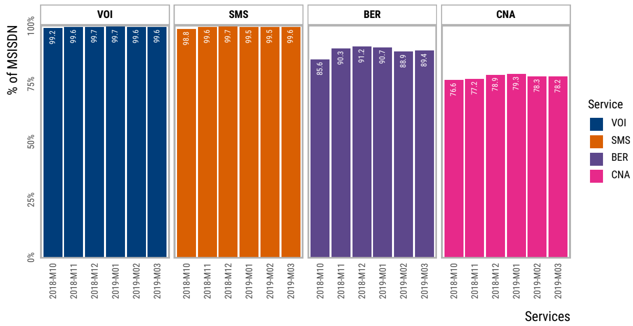


FIGURE A.10: Trend of the monthly probing sample ratio per service of the counts shown in Figure A.9.

Samples per Combination of Services

We start with 584 810 samples and determine the top cases by grouping of the data samples by levels of service combinations used.

Voice	SMS	Data *(CP)*	Data *(UP)*	Count	%	% Cumulative	Descending	% Descending
Y	Y	Y	Y	443 663	75.86	75.86	584 810	100.00
Y	Y	Y	N	73 812	12.62	88.49	141 147	24.14
Y	Y	N	N	52 856	9.04	97.52	67 335	11.51
Y	Y	N	Y	8 895	1.52	99.05	14 479	2.48
Y	N	Y	Y	2 695	0.46	99.51	5 584	0.95
N	Y	Y	Y	1 039	0.18	99.68	2 889	0.49
N	Y	Y	N	662	0.11	99.80	1 850	0.32
N	Y	N	N	553	0.09	99.89	1 188	0.20
Y	N	Y	N	301	0.05	99.94	635	0.11
N	Y	N	Y	97	0.02	99.96	334	0.06
Y	N	N	Y	70	0.01	99.97	237	0.04
N	N	N	Y	69	0.01	99.98	167	0.03
N	N	Y	Y	57	0.01	99.99	98	0.02
Y	N	N	N	21	0.00	100.00	41	0.01
N	N	Y	N	20	0.00	100.00	20	0.00

CP: Control-plane, **UP**: User-plane

TABLE A.10: The number of observations in groups of services that were used in combination.

570 331 samples (97.52%) will remain after filtering for the *top 3* cases.

The largest group of 443 663 or 75.86% had activity on all of the probed services.

The two smaller groups directly following have a combined total of 126 668 or 21.66% (12.62% + 9.04%) samples, and represent subscribers that did not have mobile packet-switched activity on the user-plane. The aggregate remaining 14 479 or 2.48% highlighted cases, represent uncommon combinations of services, that will be removed from our data for porposes of simplifying this analysis.

Samples per Radio Technology

Samples per technology without any filtering of the 584 810 subscribers.

Radio Type	Count	%
2G	553 239	94.6
3G	479 283	82.0
4G	240 614	41.1

Table A.11: The number of probing observations per radio technology for the full dataset after removal of small analytic attribute groups and checking for device information.

Remaining samples per technology for the 570 331 subscribers included within the top 3 cases. We see that the distribution of samples between radio technologies have not been affected much by the removal of the uncommon service combinations.

Radio Type	Count	%
2G	542 502	95.1
3G	467 511	82.0
4G	235 860	41.4

Table A.12: The number of probing observations per radio technology for the top service combinations.

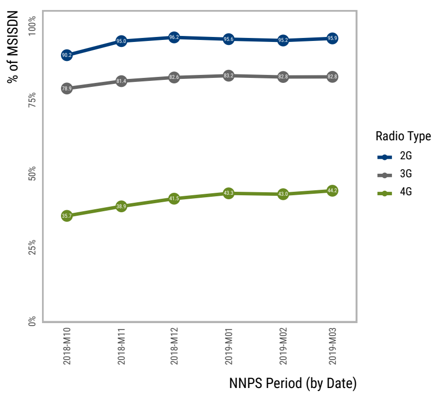


Figure A.11: Trend of the monthly Network Customer Experience probing sample ratio for the counts per radio technology prior to filtering of the top cases and radio types.

Samples per Combination of Radio Technologies

Samples per combination of radio technologies for the full reduced dataset.

Radio Types			Count	%
2G	3G		224 384	38.4
2G	3G	4G	223 386	38.2
2G			102 413	17.5
	3G		17 399	3.0
	3G	4G	14 114	2.4
2G		4G	3 056	0.5
		4G	58	0.0

Table A.13: The number of probing observations for each combination of radio technologies used by subscribers for the full dataset after removal of small analytic attribute groups and checking for device information.

Samples per combination of radio technologies after filtering for the top cases.

Radio Types			Count	%
2G	3G	4G	220 912	38.7
2G	3G		218 770	38.4
2G			100 162	17.6
	3G		15 539	2.7
	3G	4G	12 290	2.2
2G		4G	2 658	0.5

Table A.14: The number of probing observations for each combination of radio technologies used by subscribers within the top service combinations. The highlighted bottom three cases consisting of 30 487 or 5.4% of the total samples will be removed for purposes of simplifying this analysis.

After removal of the low RA types we are left with 539 844 samples.

Radio Type	Count	%
2G	539 844	100.0
3G	439 682	81.4
4G	220 912	40.9

Table A.15: The number of probing observations per radio technology for the new dataset.

Radio Types			Count	%
2G	3G	4G	220 912	40.9
2G	3G		218 770	40.5
2G			100 162	18.6

Table A.16: The number of probing observations for each combination of radio technologies after filtering for top3.

Trend after filtering for the top 3 cases.

Samples per Radio Technology and Service

Trends of the sample ratio of each radio technology used per service, prior to filtering of the top combinations.

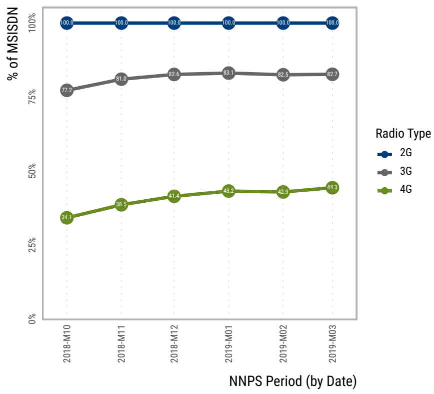
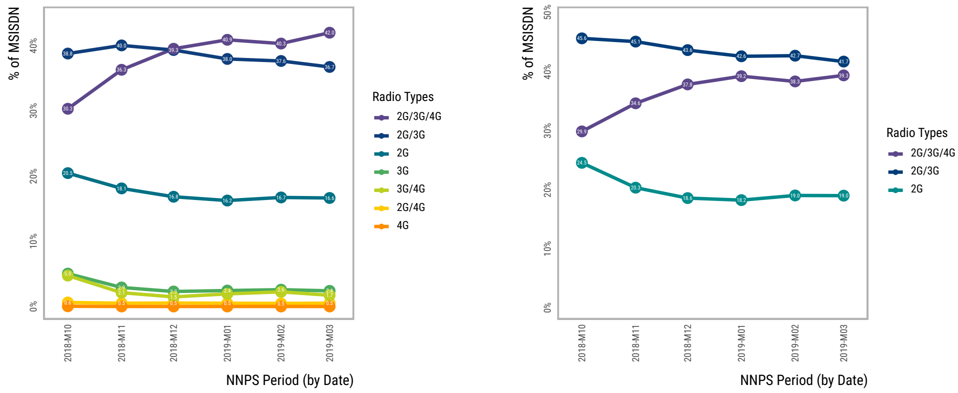


FIGURE A.12: Trend of the monthly probing sample ratio per radio technology after filtering for top3.



(a) Before filtering (b) After filtering

FIGURE A.13: Trend of the sample ratio for each combination of radio technologies used together. Before vs. after filtering for top3 cases.

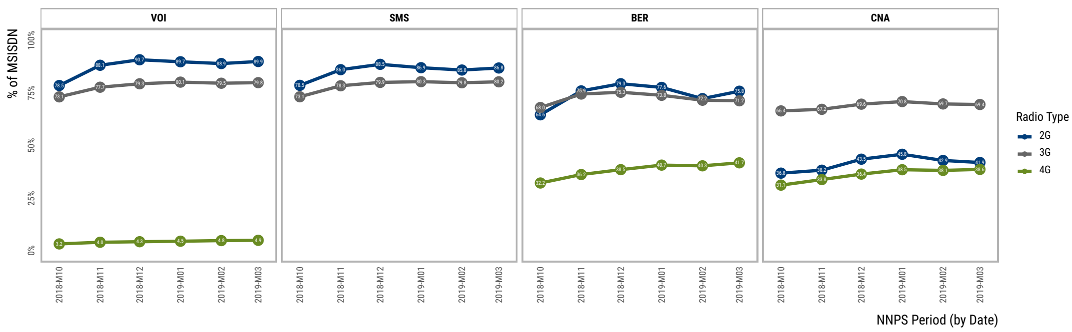


FIGURE A.14: Trend of the monthly Network Customer Experience probing sample ratio for the counts per service and radio technology.

Figure A.15 is the same as Figure A.13, but showing the unique combinations of radio types used per service.

Radio Types and Services with Low Representation

After removal of these additional 16 samples we are left with 539 828 samples.

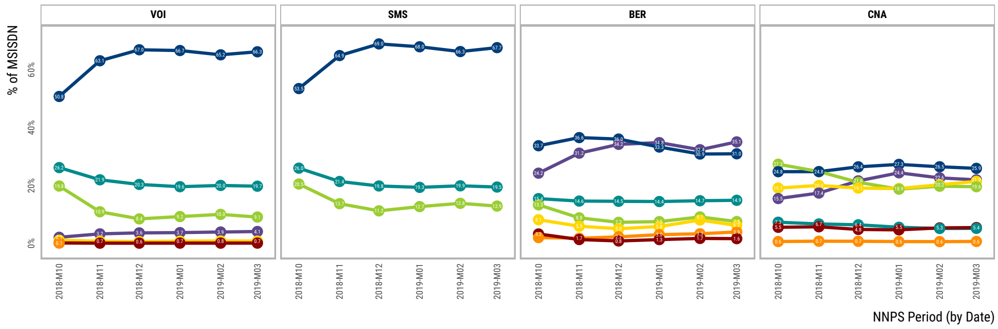


FIGURE A.15: Trend of the monthly probing sample ratio for combinations of radio technologies per service prior to filtering of the top cases.

Voice	SMS	Data (CP)	Data (UP)	Probing Services	Radio Types	Count	%	% Cumulative
Y	Y	Y	Y	VOI SMS BER CNA	2G 3G 4G	211 578	39.19	39.19
Y	Y	Y	Y	VOI SMS BER CNA	2G 3G	186 828	34.61	73.80
Y	Y	N	N	VOI SMS	2G	48 714	9.02	82.82
Y	Y	Y	N	VOI SMS BER	2G	32 936	6.10	88.92
Y	Y	Y	N	VOI SMS BER	2G 3G	28 513	5.28	94.21
Y	Y	Y	Y	VOI SMS BER CNA	2G	18 512	3.43	97.64
Y	Y	Y	N	VOI SMS BER	2G 3G 4G	9 318	1.73	99.36
Y	Y	N	N	VOI SMS	2G 3G	3 429	0.64	100.00
Y	Y	N	N	VOI SMS	2G 3G 4G	16	0.00	100.00

CP: Control-plane, **UP**: User-plane

TABLE A.17: The count of unique subscribers for each combination of services and radio types that were used by the NNPS respondents over the period October 2017 to May 2018. The last case highlighted at the bottom of the list is a small cluster of 16 subscribers that used a combination of voice and SMS services on all radio technologies. SMS is only probed on 2G and 3G technologies in our service provider's network and therefore the 4G activity reflected for these samples will be due to 4G voice, or VoLTE activity. It is expected that activity on 4G voice will always be accompanied by activity on the BER service, because a subscriber using VoLTE first has to setup a dedicated bearer session before being able to use VoLTE. Based on the collected data it is not the case with these 16 subscribers and we therefore suspect incomplete data which will be disregarded from this analysis.

Per Service Low Representation Clusters

Within service small clusters of samples making use of uncommon combinations of radio types were identified which will also be removed from the data for purposes of simplifying the analysis.

After removal of an additional 47 579 within service small radio type clusters, we are left with 502 884 samples that will be used for modelling, which reflects an overall 85.99% retention of our dataset containing 584 810 samples.

Samples per Service Remaining

After the removal of the service and radio technology groupings with low representation we are left with the number of observations per service as shown in Table A.19.

Service	Radio Types			Count	%	% Cumulative	Descending	% Descending
VOI	2G	3G		363 516	67.3	67.3	539 828	100.0
VOI	2G			118 115	21.9	89.2	176 312	32.7
VOI		3G		35 417	6.6	95.8	58 197	10.8
VOI	2G	3G	4G	19 673	3.6	99.4	22 780	4.2
VOI		3G	4G	2 840	0.5	100.0	3 107	0.6
VOI	2G		4G	198	0.0	100.0	267	0.0
VOI			4G	69	0.0	100.0	69	0.0
SMS	2G	3G		373 124	69.1	69.1	539 828	100.0
SMS	2G			114 355	21.2	90.3	166 704	30.9
SMS		3G		52 349	9.7	100.0	52 349	9.7
BER	2G	3G		174 481	32.3	32.3	539 828	100.0
BER	2G	3G	4G	166 055	30.8	63.1	365 347	67.7
BER	2G			74 803	13.9	76.9	199 292	36.9
BER				52 143	9.7	86.6	124 489	23.1
BER		3G		30 205	5.6	92.2	72 346	13.4
BER		3G	4G	24 317	4.5	96.7	42 141	7.8
BER	2G		4G	13 121	2.4	99.1	17 824	3.3
BER			4G	4 703	0.9	100.0	4 703	0.9
CNA				122 910	22.8	22.8	539 828	100.0
CNA	2G	3G		116 955	21.7	44.4	416 918	77.2
CNA	2G	3G	4G	93 621	17.3	61.8	299 963	55.6
CNA		3G	4G	80 585	14.9	76.7	206 342	38.2
CNA		3G		79 724	14.8	91.5	125 757	23.3
CNA	2G			25 629	4.7	96.2	46 033	8.5
CNA			4G	18 596	3.4	99.7	20 404	3.8
CNA	2G		4G	1 808	0.3	100.0	1 808	0.3

Table A.18: A per service breakdown of the unique subscriber count for each combination of radio technologies that were used by the NNPS respondents over the period October 2018 to March 2019. A few additional cases considered for removal due to low sample representation in the population are highlighted in blue. It includes an additional 1.7% of cases which are not covered by the proposed filtering of service type as shown in Table A.17, or by filtering of radio access type as shown in Table A.13. Cases highlighted in yellow represent samples that didn't have any activity for the particular service.

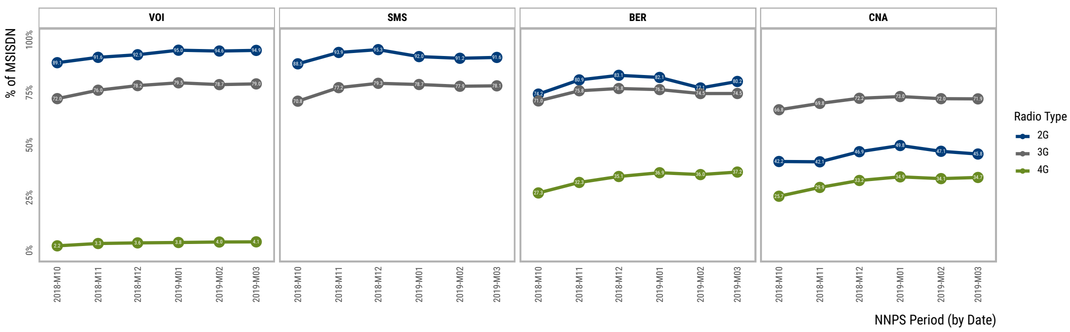


Figure A.16: Trend of the monthly Network Customer Experience probing sample ratio for the counts per service and radio technology after filtering.

A.3 Handling of Incomplete Data

Complete Cases and Missing Values per Service

VOI

Table A.20 below shows the number of missing values for each numerical feature of the VOI dataset.

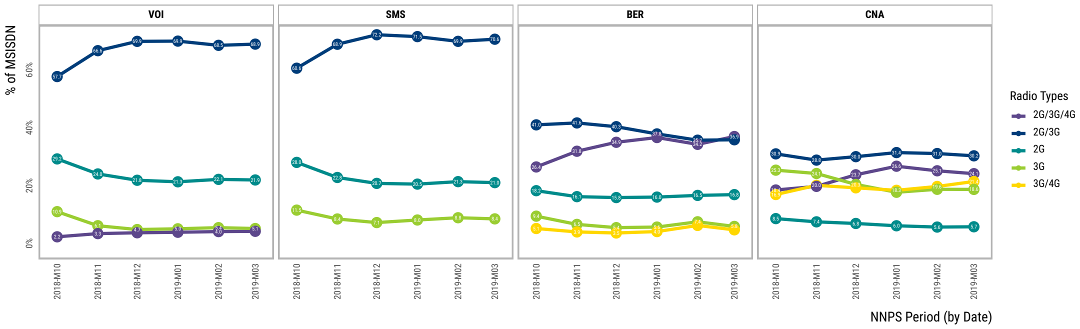


FIGURE A.17: Trend of the monthly probing sample ratio for combinations of radio technologies per service after filtering.

Source	Before	Retained	%
NNPS	584 810	502 884	86.0
VOI	582 313	502 884	86.4
SMS	581 577	502 884	86.5
BER	522 249	450 741	86.3
CNA	456 585	382 233	83.7

TABLE A.19: The number of observations per data source for the reduced dataset. Sample counts before the cleanup process are shown against the retained samples and percentage for each source.

Feature	Complete	%	NA	%	Zero	%	GT-Zero	%
VOI_ATTEMPT	502 882	100.00	0	0.00	0	0.00	502 882	100.00
VOI_CONNECT	502 882	100.00	0	0.00	2 688	0.54	500 194	99.47
VOI_AVGDURATION	500 194	99.47	2 688	0.54	0	0.00	500 194	99.47
VOI_AVGSETUPTIME	500 983	99.62	1 899	0.38	0	0.00	500 983	99.62
VOI_TOTALDURATION	500 194	99.47	2 688	0.54	0	0.00	500 194	99.47
VOI_ALERT_TO_CONNECTFAILRATE	502 882	100.00	0	0.00	6 886	1.37	495 996	98.63
VOI_ALERTCONNECTRATE	500 979	99.62	1 903	0.38	1 204	0.24	499 775	99.38
VOI_ALERTFAILRATE	502 882	100.00	0	0.00	109 862	21.85	393 020	78.15
VOI_ATTEMPT_TO_ALERTFAILRATE	502 882	100.00	0	0.00	4 111	0.82	498 771	99.18
VOI_CALLHANDOVERFAILRATE	502 882	100.00	0	0.00	224 022	44.55	278 860	55.45
VOI_CONNECTFAILRATE	502 882	100.00	0	0.00	42 156	8.38	460 726	91.62
VOI_DROPNETWORKRATE	502 882	100.00	0	0.00	220 402	43.83	282 480	56.17
VOI_SETUPFAILNETWORKRATE	502 882	100.00	0	0.00	93 437	18.58	409 445	81.42
VOI_RATIO_2G_DUR	502 882	100.00	0	0.00	44 398	8.83	458 484	91.17
VOI_RATIO_3G_DUR	502 882	100.00	0	0.00	125 129	24.88	377 753	75.12
VOI_RATIO_4G_DUR	502 882	100.00	0	0.00	485 606	96.56	17 276	3.44
VOI_RATIO_MODUR	502 882	100.00	0	0.00	21 663	4.31	481 219	95.69
VOI_RATIO_RA_MOST_USED_ATT	502 882	100.00	0	0.00	0	0.00	502 882	100.00
VOI_RATIO_RA_MOST_USED_DUR	502 882	100.00	0	0.00	2 573	0.51	500 309	99.49

Complete: Complete cases, **NA**: Missing values (Not Available), **Zero**: Zero values, **GT-Zero**: Greater than zero values

TABLE A.20: Summary statistics indicating the number of observations and missing values for each of the numerical VOI features. The variables highlighted in blue were identified as highly collinear variables and was removed from our final modelling dataset.

SMS

Table A.21 below shows the number of missing values for each numerical feature of the SMS dataset.

Feature	Complete	%	NA	%	Zero	%	GT-Zero	%
SMS_ATTEMPT	502 882	100.00	0	0.00	0	0.00	502 882	100.00
SMS_SUCCESS	502 882	100.00	0	0.00	34	0.01	502 848	99.99
SMS_AVGTIME	502 828	99.99	54	0.01	0	0.00	502 828	99.99
SMS_FAILRATE	502 882	100.00	0	0.00	132 208	26.29	370 674	73.71
SMS_SUCCESSRATE	502 882	100.00	0	0.00	34	0.01	502 848	99.99
SMS_RATIO_2G_ATTEMPT	502 882	100.00	0	0.00	43 986	8.75	458 896	91.25
SMS_RATIO_3G_ATTEMPT	502 882	100.00	0	0.00	112 349	22.34	390 533	77.66
SMS_RATIO_MOATTEMPT	502 882	100.00	0	0.00	47 665	9.48	455 217	90.52
SMS_RATIO_MTATTEMPT	502 882	100.00	0	0.00	69	0.01	502 813	99.99
SMS_RATIO_RA_MOST_USED_ATT	502 882	100.00	0	0.00	0	0.00	502 882	100.00

Complete: Complete cases, **NA**: Missing values (Not Available), **Zero**: Zero values, **GT-Zero**: Greater than zero values

Table A.21: Summary statistics indicating the number of observations and missing values for each of the numerical SMS features. The variables highlighted in blue were identified as highly collinear variables and removed from the modelling dataset.

BER

Table A.22 below shows the number of missing values for each of the numerical features in the BER dataset. We see that the features that are specific to LTE or 4G, namely EPSATTACH, DFLTBEARERACT and PDNCONN have the highest percentages of missing values. The reason for the high amount of missing data for the LTE features is because most of the surveyed subscribers, 59.1% (40.5% 2G/3G + 18.6% 2G - see table A.16) were making use of only the older 2G and 3G technology types, and subscribers that only use 2G or 3G will have missing values for all features that are specific to LTE.

Feature	Radio Types	Complete	%	NA	%	Zero	%	GT-Zero	%
BER_ATTACHATTEMPT	2G 3G	450 736	100.00	0	0.00	5 327	1.18	445 409	98.82
BER_DFLTBEARERACTATTEMPT	4G	172 048	38.17	278 688	61.83	1 605	0.36	170 443	37.81
BER_EPSATTACHATTEMPT	4G	172 048	38.17	278 688	61.83	196	0.04	171 852	38.13
BER_PDNCONNATTEMPT	4G	172 048	38.17	278 688	61.83	41	0.01	172 007	38.16
BER_PDPACTATTEMPT	2G 3G	450 736	100.00	0	0.00	21 142	4.69	429 594	95.31
BER_ATTACHAVGDUR	2G 3G	444 109	98.53	6 627	1.47	0	0.00	444 109	98.53
BER_DFLTBEARERACTAVGDUR	4G	155 063	34.40	295 673	65.60	0	0.00	155 063	34.40
BER_EPSATTACHAVGDUR	4G	170 675	37.87	280 061	62.13	0	0.00	170 675	37.87
BER_PDNCONNAVGDUR	4G	170 842	37.90	279 894	62.10	0	0.00	170 842	37.90
BER_PDPACTAVGDUR	2G 3G	395 052	87.65	55 684	12.35	0	0.00	395 052	87.65
BER_ATTACHFAILRATE	2G 3G	445 409	98.82	5 327	1.18	345 474	76.65	99 935	22.17
BER_DFLTBEARERACTFAILRATE	4G	170 443	37.81	280 293	62.19	127 541	28.30	42 902	9.52
BER_EPSATTACHFAILRATE	4G	171 852	38.13	278 884	61.87	100 239	22.24	71 613	15.89
BER_PDNCONNFAILRATE	4G	172 007	38.16	278 729	61.84	102 000	22.63	70 007	15.53
BER_PDPACTFAILRATE	2G 3G	429 594	95.31	21 142	4.69	238 568	52.93	191 026	42.38
BER_RATIO_2G_ATTEMPT	2G 3G 4G	450 736	100.00	0	0.00	50 699	11.25	400 037	88.75
BER_RATIO_3G_ATTEMPT	2G 3G 4G	450 736	100.00	0	0.00	74 634	16.56	376 102	83.44
BER_RATIO_4G_ATTEMPT	2G 3G 4G	450 736	100.00	0	0.00	278 688	61.83	172 048	38.17
BER_RATIO_RA_MOST_USED_ATT	2G 3G 4G	450 736	100.00	0	0.00	0	0.00	450 736	100.00

Complete: Complete cases, **NA**: Missing values (Not Available), **Zero**: Zero values, **GT-Zero**: Greater than zero values

Table A.22: Summary statistics of the numerical features in the BER datset, indicating the radio types on which each feature is available, as well as the number of observations and missing values for each feature.

Factor Level Re-grouping and Filtering of Incomplete Measurements

Feature	Radio Types	Compl	%	NA	%	Zero	%	GT-Zero	%
BER_ATTACHATTEMPT	2G 3G	439 915	100.00	0	0.00	0	0.00	439 915	100.00
BER_DFLTBEARERACTATTEMPT	4G	167 932	38.17	271 983	61.83	0	0.00	167 932	38.17
BER_EPSATTACHATTEMPT	4G	167 932	38.17	271 983	61.83	0	0.00	167 932	38.17
BER_PDNCONNATTEMPT	4G	167 932	38.17	271 983	61.83	0	0.00	167 932	38.17
BER_PDPACTATTEMPT	2G 3G	439 915	100.00	0	0.00	18 002	4.09	421 913	95.91
BER_ATTACHAVGDUR	2G 3G	439 915	100.00	0	0.00	0	0.00	439 915	100.00
BER_DFLTBEARERACTAVGDUR	4G	152 859	34.75	287 056	65.25	0	0.00	152 859	34.75
BER_EPSATTACHAVGDUR	4G	167 932	38.17	271 983	61.83	0	0.00	167 932	38.17
BER_PDNCONNAVGDUR	4G	167 932	38.17	271 983	61.83	0	0.00	167 932	38.17
BER_PDPACTAVGDUR	2G 3G	388 070	88.22	51 845	11.79	0	0.00	388 070	88.22
BER_ATTACHFAILRATE	2G 3G	439 915	100.00	0	0.00	341 684	77.67	98 231	22.33
BER_DFLTBEARERACTFAILRATE	4G	167 932	38.17	271 983	61.83	125 625	28.56	42 307	9.62
BER_EPSATTACHFAILRATE	4G	167 932	38.17	271 983	61.83	98 125	22.30	69 807	15.87
BER_PDNCONNFAILRATE	4G	167 932	38.17	271 983	61.83	99 713	22.67	68 219	15.51
BER_PDPACTFAILRATE	2G 3G	421 913	95.91	18 002	4.09	232 821	52.92	189 092	42.98
BER_RATIO_2G_ATTEMPT	2G 3G 4G	439 915	100.00	0	0.00	47 462	10.79	392 453	89.21
BER_RATIO_3G_ATTEMPT	2G 3G 4G	439 915	100.00	0	0.00	70 391	16.00	369 524	84.00
BER_RATIO_4G_ATTEMPT	2G 3G 4G	439 915	100.00	0	0.00	271 983	61.83	167 932	38.17
BER_RATIO_RA_MOST_USED_ATT	2G 3G 4G	439 915	100.00	0	0.00	0	0.00	439 915	100.00

Compl: Complete cases, **NA**: Missing values (Not Available), **Zero**: Zero values, **GT-Zero**: Greater than zero values

TABLE A.23: Summary statistics indicating the number of observations and missing values for each of the numerical BER features after dropping the samples from the insignificant factor levels and those without **ATTACH** activity on the most used radio access technology (RAT) type. The variables highlighted in blue were identified as either being highly collinear variables, or having close to zero variance and was removed from the modelling dataset. The variables that are higlighted in yellow are the remaining variables that have low missing values, and the one's highlighted in orange, are 4G specific KPI's that have a high number of missing values due to only about 35% of the subscribers using 4G at the time.

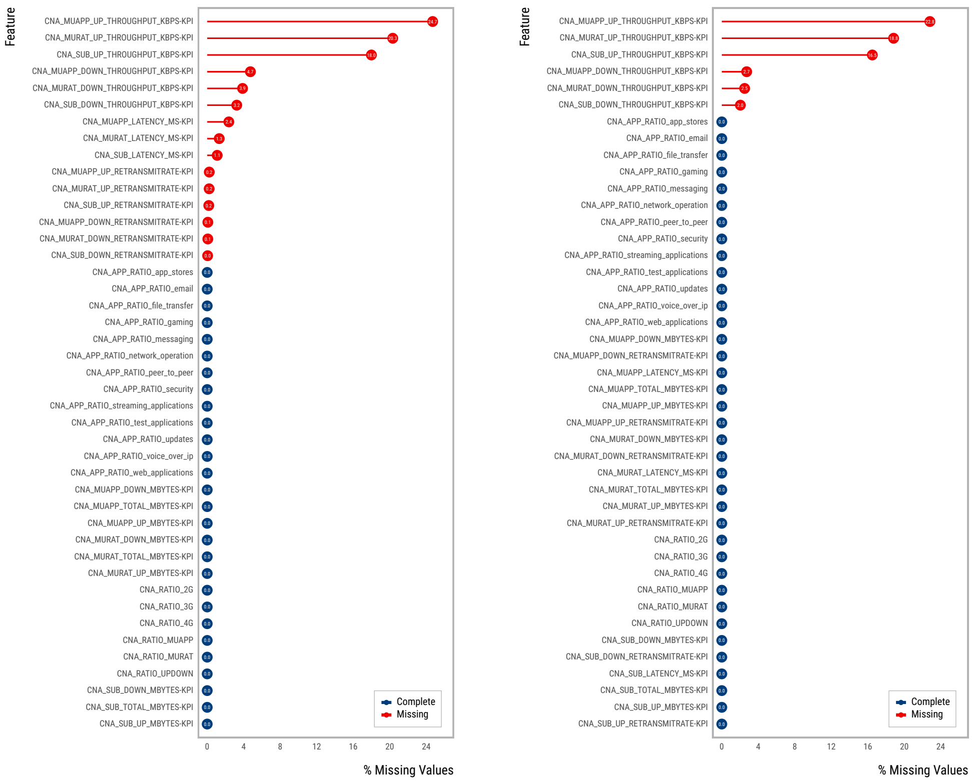
(a) Before filtering

(b) After filtering

FIGURE A.18: The percentage of missing values for each numerical feature of the BER dataset. Before vs. after filtering of the *no-Attach* activity, low representation within factor levels after grouping, and the incomplete packet-switched control-plane vs. user-plane cases.

CNA

Missing Data



(a) Before filtering (b) After filtering

Figure A.19: Missing data for numerical features of the CNA user-plane dataset. Before vs. after filtering of the low representation cases and dropping the small number of samples with features that have low missingness.

A.4 Summary of Data Preparation

In summary, during a first iteration the data of each probing source was prepared for modelling separately with the aim to drop as few records in each dataset as possible during the pre-processing. Due to a large number of missing values within the raw probing data, techniques like aggregation needed to be used to combine the individual indicators that measure opposite legs of the same process, into single bi-directional measurements that are more complete. For example; the *mobile-originating* and *mobile-terminating* legs of each metric for voice calls were combined into bi-directional measurements that are more complete, as explained in Section 3.4. This aggregation solves the issue of incomplete voice data with lots of null values due to subscribers that only have measurements in a single direction; for example someone that only received calls during the measurement period.

During the data preparation we also identified various scenarios of small sample groups that we were able to discard with two objectives in mind:

- To create 'stable' and complete datasets without any gaps that would result in sparse modelling data, as discussed in Section 3.4, Feature Engineering and Combining Datasets.
- To simplify the present analysis, as catering for the discarded scenarios would simply involve creating additional models each focussing on a smaller subset of

Feature	Complete	%	NA	%	Zero	%	GT-Zero	%
CNA_SUB_UP_MBYTES	406 406	100.00	0	0.00	742	0.18	405 664	99.82
CNA_SUB_DOWN_MBYTES	406 406	100.00	0	0.00	260	0.06	406 146	99.94
CNA_SUB_TOTAL_MBYTES	406 406	100.00	0	0.00	0	0.00	406 406	100.00
CNA_MURAT_UP_MBYTES	406 406	100.00	0	0.00	920	0.23	405 486	99.77
CNA_MURAT_DOWN_MBYTES	406 406	100.00	0	0.00	341	0.08	406 065	99.92
CNA_MURAT_TOTAL_MBYTES	406 406	100.00	0	0.00	0	0.00	406 406	100.00
CNA_MUAPP_UP_MBYTES	406 406	100.00	0	0.00	967	0.24	405 439	99.76
CNA_MUAPP_DOWN_MBYTES	406 406	100.00	0	0.00	364	0.09	406 042	99.91
CNA_MUAPP_TOTAL_MBYTES	406 406	100.00	0	0.00	0	0.00	406 406	100.00
CNA_SUB_LATENCY_MS	401 593	98.82	4 813	1.18	0	0.00	401 593	98.82
CNA_MURAT_LATENCY_MS	400 692	98.59	5 714	1.41	0	0.00	400 692	98.59
CNA_MUAPP_LATENCY_MS	396 109	97.47	10 297	2.53	0	0.00	396 109	97.47
CNA_SUB_UP_RETRANSMITRATE	405 664	99.82	742	0.18	1 278	0.31	404 386	99.50
CNA_SUB_DOWN_RETRANSMITRATE	406 146	99.94	260	0.06	2 386	0.59	403 760	99.35
CNA_MURAT_UP_RETRANSMITRATE	405 486	99.77	920	0.23	1 431	0.35	404 055	99.42
CNA_MURAT_DOWN_RETRANSMITRATE	406 065	99.92	341	0.08	2 742	0.68	403 323	99.24
CNA_MUAPP_UP_RETRANSMITRATE	405 439	99.76	967	0.24	5 485	1.35	399 954	98.41
CNA_MUAPP_DOWN_RETRANSMITRATE	406 042	99.91	364	0.09	6 788	1.67	399 254	98.24
CNA_APP_RATIO_web_applications	406 406	100.00	0	0.00	5 166	1.27	401 240	98.73
CNA_APP_RATIO_messaging	406 406	100.00	0	0.00	76 254	18.76	330 152	81.24
CNA_APP_RATIO_security	406 406	100.00	0	0.00	95 493	23.50	310 913	76.50
CNA_APP_RATIO_network_operation	406 406	100.00	0	0.00	122 561	30.16	283 845	69.84
CNA_APP_RATIO_streaming_applications	406 406	100.00	0	0.00	191 952	47.23	214 454	52.77
CNA_RATIO_2G	406 406	100.00	0	0.00	174 915	43.04	231 491	56.96
CNA_RATIO_3G	406 406	100.00	0	0.00	41 162	10.13	365 244	89.87
CNA_RATIO_4G	406 406	100.00	0	0.00	220 200	54.18	186 206	45.82
CNA_RATIO_MURAT	406 406	100.00	0	0.00	0	0.00	406 406	100.00
CNA_RATIO_MUAPP	406 406	100.00	0	0.00	0	0.00	406 406	100.00
CNA_RATIO_UPDOWN	406 406	100.00	0	0.00	742	0.18	405 664	99.82
CNA_SUB_UP_THROUGHPUT_KBPS	331 161	81.48	75 245	18.52	735	0.18	330 426	81.30
CNA_SUB_DOWN_THROUGHPUT_KBPS	392 766	96.64	13 640	3.36	86	0.02	392 680	96.62
CNA_MURAT_UP_THROUGHPUT_KBPS	321 986	79.23	84 420	20.77	704	0.17	321 282	79.05
CNA_MURAT_DOWN_THROUGHPUT_KBPS	390 255	96.03	16 151	3.97	99	0.02	390 156	96.00
CNA_MUAPP_UP_THROUGHPUT_KBPS	303 383	74.65	103 023	25.35	970	0.24	302 413	74.41
CNA_MUAPP_DOWN_THROUGHPUT_KBPS	386 456	95.09	19 950	4.91	113	0.03	386 343	95.06

Complete: Complete cases, **NA**: Missing values (Not Available), **Zero**: Zero values, **GT-Zero**: Greater than zero values

TABLE A.24: Summary statistics indicating the number of observations and missing values for each numerical feature, before cleanup of the CNA dataset.

data, similar to the principle we discuss in Section 3.4, Separate Modelling Data.

A summary of the primary data cleanup outcomes we achieved in Chapter A are:

- Dropping of subscriber, device and activity groups with low representation.
  - Small subscriber segmentation groups.
- Dropping of complete services with incomplete or contaminated data due to probing system stability issues.
- Dropping of services that have overlap with an alternative source, which is more reliable.
  - GN-CP.
  - GN-UP.
  - GTPv2.
  - CSFB.

- Dropping of irregular usage patterns.
  - Samples per Combination of Services - Top cases.
  - Samples per Combination of Radio.
  - Samples per Radio Technology and Service.
  - Samples with Low Representation when we split by radio types and service.
- Dropping of features or samples within service
  - Complete Cases and Missing Values per Service
  - Near zero variance numerical predictor variables

A.5 Processed Variables Visualised

SMS: Normalised Distributions

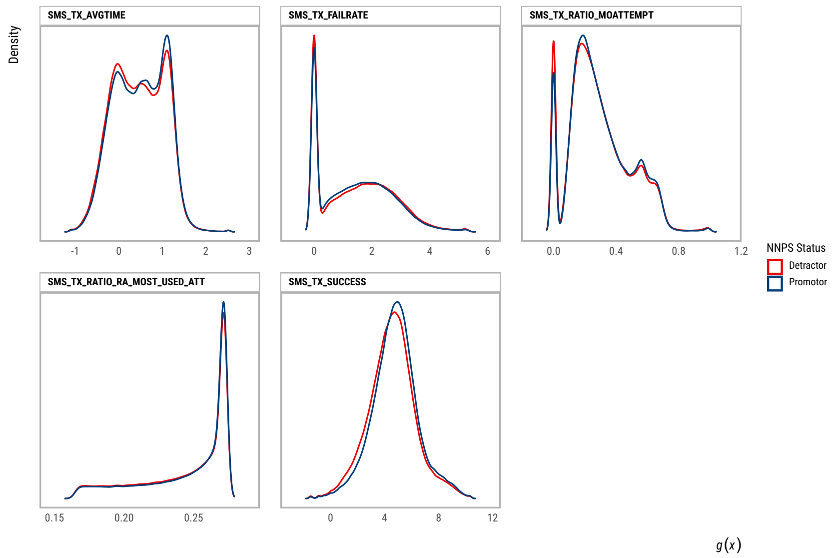


FIGURE A.20: Kernel density estimate distribution plots of the transformed SMS variables over the complete analysis period. The plots show variable distributions after normalisation transformations.

BER: Normalised Distributions

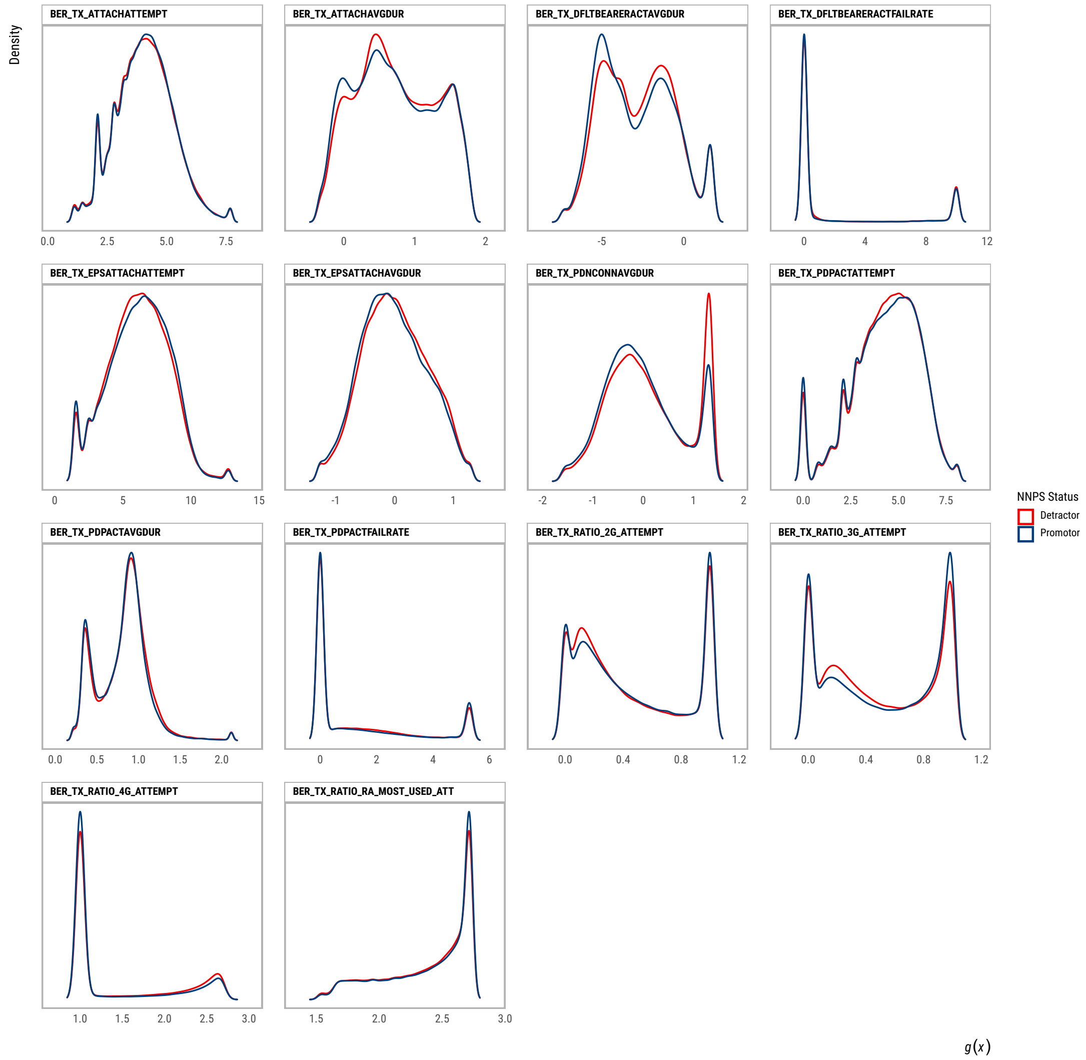


Figure A.21: Kernel density estimate distribution plots of the transformed BER variables over the complete analysis period. The plots show variable distributions after normalisation transformations.

CNA: Normalised Distributions

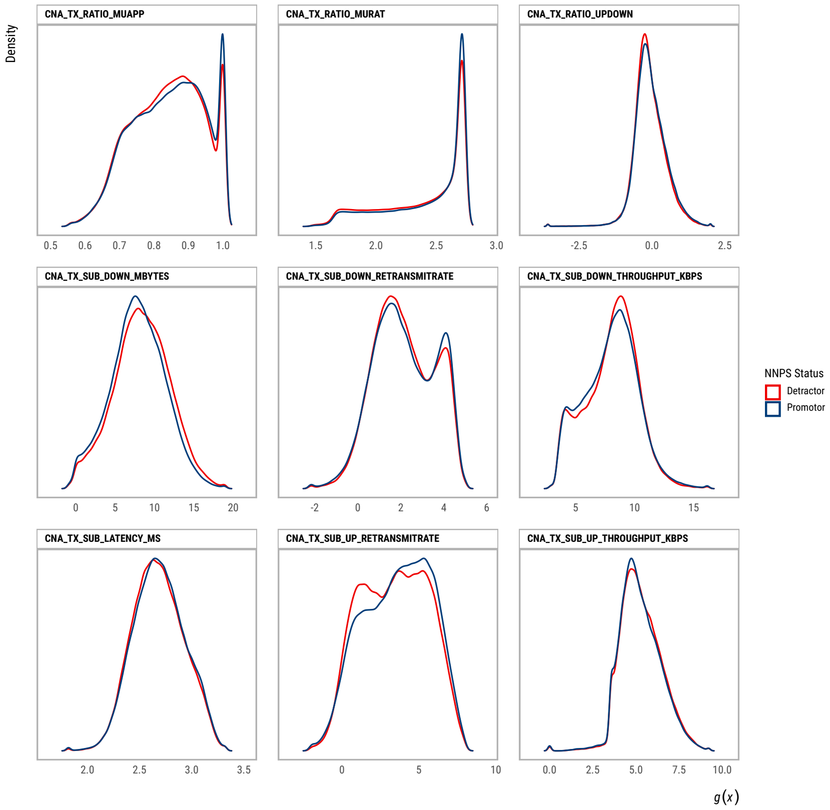


FIGURE A.22: Kernel density estimate distribution plots of the transformed CNA user-plane variables over the complete analysis period. The plots show variable distributions after normalisation transformations.

Appendix B

Data Dictionary

B.1 Customer Analytics Data

Network NPS (NNPS) and **Subscriber Attributes**

Field Name	Description	Example Values
NNPS_PERIOD	Survey period short text	2018-M01
START_DATE	Start date of aggregation period type	2018/01/01
SUB_MSISDN	Mobile Subscriber Identity (MSISDN)	27821234567
DATETIME_CREATED_RESULT	Date and time of survey results record creation	2018/01/12 11:45:57
DAYS_PRIOR_IN_PERIOD	Number of days in aggregation period prior to survey result recording	11
NNPS_LTR_A	Likelyhood to Recommend score	0 - 10
LTR_A_PROMOTER	Promoter flag	Yes, No
LTR_A_NEUTRAL	Neutral flag	Yes, No
LTR_A_DETRACTOR	Detractor flag	Yes, No
IF_BUS_CONS_CLASS_NM	Business Consumer Class name	Consumer, Enterprise
IF_CHANNEL	Customer communication channel	Network
IF_DATA_USAGE	Data usage flag	Yes, No
IF_DVC_GSM_GEN_CD	Device GSM generation code	2G, 3G, 4G
IF_DVC_NM	Device marketing name	Samsung SM-J320FN, Huawei P8 Lite
IF_DVC_TYPE_GRP_NM	Device Type group name	Smart Phone, Feature Phone
IF_MARKET	Country	South Africa
IF_NETWK_REGN_NM	Network Region name	Central, Eastern, Western
IF_PMT_METH_MED	Payment method	Pre-Paid, Post-Paid
IF_PMT_METH_NM	Payment Method name	Prepaid, Top Up, Contract
IF_PRICE_PLAN_NM	Price Plan name	Prepaid150c, SmartXL
IF_PROV_NM	Resident province name	Western Cape, Gauteng
IF_SEGMENT	Market segment	Mass Market, High Value, Large Enterprises
IF_SOLUTION	Type of service solution used	
IF_SUB_SEGMENT	Sub Market Segment	Aspiration Seeker, Emerging
IF_SUBS_AGE	Subscriber age in years	25, 41
IF_SUBURB_NM	Resident Suburb name	Knysna, Rooihuiskraal
IF_TOWN_NM	Resident town name	Knysna, Centurion, Howick
IF_VOICE_USAGE	Voice usage flag	Yes, No

Table B.1: Data dictionary for the Network NPS and customer analytics attributes.

B.2 Customer Experience Raw Counters

B.2.1 Circuit-switched Services

Voice (VOI)

Field Name	Description	Units
PERIOD_TYPE	Time aggregation period type. {D = Daily, W = Weekly, M = Monthly}	
NNPS_PERIOD	Survey period, e.g. 2018-M01 = January 2018	
START_DATE	Start date of aggregation period type	
SUB_MSISDN	Mobile Subscriber Identity (MSISDN) of a Subscriber	
RA_TYPE	Radio Access Type as per last seen Create Session / Modify Session. GERAN: (2G) UTRAN: (3G) E-UTRAN: (4G)	
CALLHANDOVERATTEMPT	Number of call legs that started with a handover	Events
CALLHANDOVERSUCCESS	Number of call legs that started with a handover and terminated normally	Events
HNDVRCORRELATIONATTEMPT	Number of handover events for which correlation was attempted|	Events
HNDVRCORRELATIONSUCCESS	Number of successfully correlated handover events	Events
INTERRATCALLHNDVRATTEMPT	Number of call legs that started with a handover which came from a different bearer	Events
INTERRATCALLHNDVRSUCCESS	Number of call legs that started with a handover which came from a different bearer and terminated normally	Events
MOVOICEALERTATTEMPT	Total number of MO Voice Calls which reach RAB assignment	Events
MOVOICEALERTSUCCESS	Total number of MO Voice Calls with an alert	Events
MOVOICEALERTSUCCESSRATE_D	Count of events used to generate MOVoiceAlertSuccessRate - Denominator	Events
MOVOICEALERTSUCCESSRATE_N	The number of MO Voice Calls with an alert versus calls which reach RAB assignment - Numerator	Events
MOVOICEATTEMPT	Total Number of UE Originated Normal Calls	Events
MOVOICEAVGDURATION_D	Average Call Duration for UE Originated Normal Calls Count - Denominator	Events
MOVOICEAVGDURATION_N	Total Call Duration for UE Originated Normal Calls - Numerator	Time (s)
MOVOICEAVGSETUPTIME_D	Average Setup Time for UE Originated Normal Calls Count - Denominator	Events
MOVOICEAVGSETUPTIME_N	Total Setup Time for UE Originated Normal Calls - Numerator	Time (ms)
MOVOICECNNCTSUCCESSRATE_D	Count of events used to calculate MOVoiceConnectSuccessRate - Denominator	Events
MOVOICECNNCTSUCCESSRATE_N	Total number of MO Voice Calls with an alert which connect versus calls with an alert - Numerator	Events
MOVOICECONNECT	Total Number of UE Originated Voice Calls Connect	Events
MOVOICECONNECTATTEMPT	Number of MO calls which alert and connect or alert and fail due to an error	Events
MOVOICECONNECTSUCCESS	Total number of MO Voice Calls with an alert which connect	Events
MOVOICEDROPNETWK	Total Number of UE Originated Normal Calls where a CONNECT message was recorded and an Network Error Occurred	Events
MOVOICEDROPPED	Total Number of UE Originated Normal Calls where a CONNECT message was recorded and an Error Occurred	Events
MOVOICEFAIL	Number of MO Voice Calls that failed due to any error	Events
MOVOICEFAILNETWORK	Number of MO Voice Calls that failed due to a network error	Events
MOVOICESETUPFAIL	Number of Failed UE Originated Normal Calls where Failure occurred during Call Setup	Events
MOVOICESETUPFAILNETWK	Number of Failed UE Originated Normal Calls where Network Failure occurred during Call Setup	Events
MOVOICESUCCESS	Total Number of Successful UE Originated Normal Calls	Events
MOVOICESUCCESSRATE_D	Total Number of UE Originated Normal Calls Attempts - Denominator	Events
MOVOICESUCCESSRATE_N	Total Number of Successful UE Originated Normal Calls - Numerator	Events
MTVOICEALERTATTEMPT	Total number of MT Voice Calls which reach RAB assignment	Events
MTVOICEALERTSUCCESS	Total number of MT Voice Calls with an alert	Events
MTVOICEALERTSUCCESSRATE_D	Count of events used to generate MTVoiceAlertSuccessRate - Denominator	Events
MTVOICEALERTSUCCESSRATE_N	Count of MT Voice Calls with an alert versus calls which reach RAB assignment - Numerator	Events
MTVOICEATTEMPT	Total Number of UE Terminated Normal Calls	Events
MTVOICEAVGDURATION_D	Total number of UE Terminated Normal Calls Count - Denominator	Events
MTVOICEAVGDURATION_N	Total Call Duration for UE Terminated Normal Calls - Numerator	Time (s)
MTVOICEAVGSETUPTIME_D	Total number of UE Terminated Normal Calls Count - Denominator	Events
MTVOICEAVGSETUPTIME_N	Total Setup Time for UE Terminated Normal Calls - Numerator	Time (ms)
MTVOICECNNCTSUCCESSRATE_D	Count of events used to calculate MTVoiceConnectSuccessRate - Denominator	Events

Continued on next page

Table B.2 – continued from previous page

Field Name	Description	Units
MTVOICECNNCTSUCCESSRATE_N	Total number of MT Voice Calls with an alert which connect versus calls with an alert - Numerator	Events
MTVOICECONNECT	Total Number of UE Terminated Voice Calls Connect	Events
MTVOICECONNECTATTEMPT	Number of MT calls which alert and connect or alert and fail due to an error	Events
MTVOICECONNECTSUCCESS	Total number of MT Voice Calls with an alert which connect	Events
MTVOICEDROPNETWK	Total Number of UE Terminated Normal Calls where a CONNECT message was recorded and an Network Error Occurred	Events
MTVOICEDROPPED	Total Number of UE Terminated Normal Calls where a CONNECT message was recorded and an Error Occurred	Events
MTVOICEFAIL	Total number of MT Voice Calls with an error during conversation	Events
MTVOICEFAILNETWORK	Total number of MT Voice Calls with a network error during conversation	Events
MTVOICESETUPFAIL	Number of Failed UE Terminated Normal Calls where Failure occurred during Call Setup	Events
MTVOICESETUPFAILNETWK	Number of Failed UE Terminated Normal Calls where Network Failure occurred during Call Setup	Events
MTVOICESUCCESS	Total Number of Successful UE Terminated Normal Calls	Events
MTVOICESUCCESSRATE_D	Total Number of Attempted UE Terminated Normal Calls - Denominator	Events
MTVOICESUCCESSRATE_N	Total Number of Successful UE Terminated Normal Calls - Numerator	Events
VOICEFAILURERATE_D	Total number of Voice Attempts - Denominator	Events
VOICEFAILURERATE_N	Total number of voice Failures - Numerator	Events

Table B.2: Data dictionary for the voice service customer experience dataset.

Note: VoLTE (Voice over LTE) is technically a packet-switched service, but the signalling information from probing allows VoLTE to be modelled in a similar fashion to circuit-switched voice on 2G and 3G. Subsequently it therefore allows the overall customer experience of voice to be modeled agnostic of the underlying technology used to provide the service, and that all voice datasets can be managed as a single end-to-end service with similar experience metrics across all technologies.

Short Message Service (SMS)

Field Name	Description	Units
PERIOD_TYPE	Time aggregation period type. {D = Daily, W = Weekly, M = Monthly}	
NNPS_PERIOD	Survey period, e.g. 2018-M01 = January 2018	
START_DATE	Start date of aggregation period type	
SUB_MSISDN	Mobile Subscriber Identity (MSISDN) of a Subscriber	
RA_TYPE	Radio Access Type as per last seen Create Session / Modify Session. GERAN: (2G) UTRAN: (3G) E-UTRAN: (4G)	
MOSMSATTEMPT	Total Number of MO SMSs	Events
MOSMSAVGTIME_D	Average MO SMS Delivery Time (ms) Count - Denominator	Rate
MOSMSAVGTIME_N	Average MO SMS Delivery Time (ms) - Numerator	Rate
MOSMSSUCCESS	Number of Successful MO SMSs	Events
MOSMSSUCCESSRATE_D	MO SMS Success Rate Count - Denominator	Rate
MOSMSSUCCESSRATE_N	MO SMS Success Rate - Numerator	Rate
MTSMSATTEMPT	Total Number of MT SMSs	Events
MTSMSAVGTIME_D	Average MT SMS Delivery Time (ms) Count - Denominator	Rate
MTSMSAVGTIME_N	Average MT SMS Delivery Time (ms) - Numerator	Rate
MTSMSSUCCESS	Number of Successful MT SMSs	Events
MTSMSSUCCESSRATE_D	MT SMS Success Rate Count - Denominator	Rate
MTSMSSUCCESSRATE_N	MT SMS Success Rate - Numerator	Rate

Table B.3: Data dictionary for the Short Message Service customer experience dataset.

B.2.2 Packet-switched Control-plane

Data Access Bearer (BER)

Field Name	Description	Units
PERIOD_TYPE	Time aggregation period type. {D = Daily, W = Weekly, M = Monthly}	
NNPS_PERIOD	Survey period, e.g. 2018-M01 = January 2018	
START_DATE	Start date of aggregation period type	
SUB_MSISDN	Mobile Subscriber Identity (MSISDN) of a Subscriber	
RA_TYPE	Radio Access Type as per last seen Create Session / Modify Session. GERAN: (2G) UTRAN: (3G) E-UTRAN: (4G)	
ATTACHATTEMPT	Count of Attach Requests attempted in the time period	Events
ATTACHAVGDURATION_D	Count of successful Attach message types - Denominator	Rate
ATTACHAVGDURATION_N	Total duration for Attach message types - Numerator	Rate
ATTACHFAILNETWORK	Count of failed Attach Requests due to network error	Events
ATTACHSUCCESS	Count of successful Attach Requests in the time period	Events
ATTACHSUCCESSRATE_D	Count of Attach Requests attempted in the time period - Denominator	Rate
ATTACHSUCCESSRATE_N	Count of successful Attach Requests attempted in the time period - Numerator	Rate
PDPACTATTEMPT	Count of PDPActivates attempted in the time period	Events
PDPACTAVGDURATION_D	Count of all successful PDPActivate message types - Denominator	Rate
PDPACTAVGDURATION_N	Total duration for all PDPActivate message types - Numerator	Rate
PDPACTFAILNETWORK	Count of failed PDPActivates in the time period, due to network error	Events
PDPACTSUCCESS	Count of successful PDPActivates attempted in the time period	Events
PDPACTSUCCESSRATE_D	Count of PDPActivates attempted in the time period - Denominator	Rate
PDPACTSUCCESSRATE_N	Count of successful PDPActivates in the time period - Numerator	Rate
EPSATTACHATTEMPT	Count of EPS Attach Requests	Events
EPSATTACHAVGDURATION_D	Count of successful EPS Attach attempts - Denominator	Rate
EPSATTACHAVGDURATION_N	Total EPS Attach duration - Numerator	Rate
EPSATTACHFAILNETWORK	Count of failed EPS Attach Requests due to network error	Events
EPSATTACHSUCCESS	Count of Successful EPS Attach Requests	Events
EPSATTACHSUCCESSRATE_D	Count of EPS Attach Requests - Denominator	Rate
EPSATTACHSUCCESSRATE_N	Count of successful EPS Attach Requests - Numerator	Rate
DFLTBEARERACTATTEMPT	Count of Default Bearer Activation Requests	Events
DFLTBEARERACTAVGDUR_D	Count of successful Default Bearer Activation attempts - Denominator	Rate
DFLTBEARERACTAVGDUR_N	Total Default Bearer Activation duration - Numerator	Rate
DFLTBEARERACTFAILNETWORK	Count of failed Default Bearer Activation Requests due to network error	Events
DFLTBEARERACTSUCCESS	Count of Successful Default Bearer Activation Requests	Events
DFLTBEARERACTSUCCESSRATE_D	Count of Default Bearer Activations - Denominator	Rate
DFLTBEARERACTSUCCESSRATE_N	Count of successful Default Bearer Activations - Numerator	Rate
DDCTDBEARERACTATTEMPT	Count of Dedicated Bearer Activation Requests	Events
DDCTDBEARERACTAVGDUR_D	Count of successful Dedicated Bearer Activation attempts - Denominator	Rate
DDCTDBEARERACTAVGDUR_N	Total Dedicated Bearer Activation duration - Numerator	Rate
DDCTDBEARERACTFAILNETWORK	Count of failed Dedicated Bearer Activation Requests due to network error	Rate
DDCTDBEARERACTSUCCESS	Count of Successful Dedicated Bearer Activation Requests	Events
DDCTDBEARERACTSUCCRATE_D	Count of Dedicated Bearer Activations - Denominator	Rate
DDCTDBEARERACTSUCCRATE_N	Count of successful Dedicated Bearer Activations - Numerator	Rate
PDNCONNATTEMPT	Count of PDN Connectivity Requests	Events
PDNCONNAVGDUR_D	Count of successful PDN Connectivity Request attempts - Denominator	Rate
PDNCONNAVGDUR_N	Total PDN Connectivity Request duration - Numerator	Rate
PDNCONNFAILNETWORK	Count of failed PDN Connectivity Requests due to network error	Events
PDNCONNSUCCESS	Count of Successful PDN Connectivity Requests	Events
PDNCONNSUCCESSRATE_D	Count of PDN Connectivity Requests - Denominator	Rate
PDNCONNSUCCESSRATE_N	Count of successful PDN Connectivity Requests - Numerator	Rate

Table B.4: Data dictionary for the data access bearer service customer experience dataset.

B.2.3 Packet-switched User-plane

Mobile data **Customer Network Analytics** (CNA).

Field Name	Description	Units
PERIOD_TYPE	Time aggregation period type. {D = Daily, W = Weekly, M = Monthly}	
NNPS_PERIOD	Survey period, e.g. 2018-M01 = January 2018	
START_DATE	Start date of aggregation period type	
SUB_MSISDN	Mobile Subscriber Identity (MSISDN) of a Subscriber	
RA_TYPE	Radio Access Type as per last seen Create Session / Modify Session. GERAN: (2G) UTRAN: (3G) E-UTRAN: (4G)	
APP_CATEGORY	Category of automatically detected applications and user-defined applications. E.g. 'Web Applications', 'File Transfer', 'Email', 'Messaging', 'Streaming Applications'.	
LATENCY_MS	The calculated Avg Round Trip Time for all TCP message pairs for the above dimensions.	ms
LATENCY_MS_N	The sum total of Round Trip Time for all TCP message pairs for the above dimensions. Used to calculate "Avg. TCP SRTT (smoothed)" metric.	ms
LATENCY_MS_D	The number of TCP Round Trip events for the above dimensions. Used to calculate "Avg. TCP SRTT (smoothed)" metric.	Count
UP_BYTES	The total volume of bytes uplink for the above dimensions in the aggregation period.	bytes
UP_RETRANSMIT_RATE	Ratio of retransmitted uplink packets to total uplink packets.	rate
UP_RETRANSMIT_PACKETS	The total count of retransmitted packets uplink for the above dimensions in the aggregation period.	Count
UP_PACKETS	The total count of packets uplink for the above dimensions in the aggregation period.	Count
UP_THROUGHPUT_MBPS	The average observed upload throughput for an application in this aggregation period.	Mbps
UP_THROUGHPUT_MBIT	The total volume of simultaneous bytes up for an application in this aggregation period, measuring application network usage. Used to calculate the "Avg. UL Application User Throughput" metric.	Mbit
UP_THROUGHPUT_SECONDS	The total active up time for an application in this aggregation period. Used to calculate the "Avg. UL Application User Throughput" metric.	Second
DOWN_BYTES	The total volume of bytes down-link for the above dimensions in the aggregation period.	bytes
DOWN_RETRANSMIT_RATE	Ratio of retransmitted down-link packets to total down-link packets.	rate
DOWN_RETRANSMIT_PACKETS	The total count of retransmitted packets down-link for the above dimensions in the aggregation period.	Count
DOWN_PACKETS	The total count of packets down-link for the above dimensions in the aggregation period.	Count
DOWN_THROUGHPUT_MBPS	The average observed download throughput for an application in this aggregation period.	Mbps
DOWN_THROUGHPUT_MBIT	The total volume of simultaneous bytes down-link for an application in this aggregation period, measuring application network usage. Used to calculate the "Avg. DL Application User Throughput" metric.	Mbit
DOWN_THROUGHPUT_SECONDS	The total active down-link time for an application in this aggregation period. Used to calculate the "Avg. DL Application User Throughput" metric.	Second

Table B.5: Data dictionary for the mobile network application customer experience dataset.

B.3 Customer Experience Indicator Calculations

B.3.1 Circuit-switched Services

Voice Service (VOI) experience indicator calculations.

CEI	Type	Description and Formula
VOI_ATTEMPT_TO_ALERTFAILRATE	Rate	Attempt to alert fail rate calculated as the ratio of calls that are failing between the attempt and alerting phases, versus the total number of call attempts. $= \frac{\texttt{ATTEMPT\_TO\_ALERTFAIL}}{\texttt{ATTEMPT}} \times 100$
VOI_ALERTFAILRATE	Rate	Alert fail rate calculated as the ratio of calls that are failing during the alerting phase, versus the total number of alert attempts. $= \left(1 - \frac{\texttt{ALERTSUCCESS}}{\texttt{ALERTATTEMPT}}\right) \times 100$
VOI_ALERT_TO_CONNECTFAILRATE	Rate	Alert to connect fail rate calculated as the ratio of calls that are failing between the alerting and connection phases, versus the total number of successful alert attempts. $= \frac{\texttt{ALERT\_TO\_CONNECTFAIL}}{\texttt{ALERTSUCCESS}} \times 100$
VOI_ALERTCONNECTRATE	Rate	Alert connect rate calculated as the ratio of calls that attempt to connect, versus the total number of successful alert attempts. $= \frac{\texttt{CONNECTATTEMPT}}{\texttt{ALERTSUCCESS}} \times 100$
VOI_SETUPFAILNETWORKRATE	Rate	Setup failure network rate calculated as the ratio of calls that failed to setup due to network related error causes, versus the total number of call attempts. $= \frac{\texttt{SETUPFAILNETWORK}}{\texttt{ATTEMPT}} \times 100$
VOI_CONNECTFAILRATE	Rate	Connect fail rate calculated as the ratio of calls that failed to connect, versus the total number of connect attempts. $= \left(1 - \frac{\texttt{CONNECTSUCCESS}}{\texttt{CONNECTATTEMPT}}\right) \times 100$
VOI_DROPNETWORKRATE	Rate	Call drop network rate calculated as the ratio of calls that dropped due to network related error causes, versus the total number of calls that successfully connected. $= \frac{\texttt{DROPNETWORK}}{\texttt{CONNECTSUCCESS}} \times 100$
VOI_CALLHANDOVERFAILRATE	Rate	Call handover fail rate calculated as the ratio of calls that failed to handover, versus the total number of calls that attempted to handover. $= \left(1 - \frac{\texttt{CALLHANDOVERSUCCESS}}{\texttt{CALLHANDOVERATTEMPT}}\right)$
VOI_INTERRATCALLHNDVRFAILRATE	Rate	Inter-rat handover fail rate calculated as the ratio of calls that failed to handover between 3G and 2G, versus the total number of calls that attempted to handover between 3G and 2G. $= \left(1 - \frac{\texttt{INTERRATCALLHNDVRSUCCESS}}{\texttt{INTERRATCALLHNDVRATTEMPT}}\right) \times 100$
VOI_INTRARATCALLHNDVRFAILRATE	Rate	Intra-rat handover fail rate calculated as the ratio of calls that failed to handover on 2G and 3G, versus the total number of calls that attempted to handover on 2G and 3G. $= \left(1 - \frac{\texttt{INTRARATCALLHNDVRSUCCESS}}{\texttt{INTRARATCALLHNDVRATTEMPT}}\right) \times 100$
VOI_RATIO_RA_MOST_USED_ATT	Rate	Ratio RA most used calculated as the ratio of call attempts on the most used radio access type, versus the total number of call attempts. $= \frac{\texttt{RA\_MOST\_USED\_ATTEMPT}}{\texttt{ATTEMPT}} \times 100$

MO: Mobile-originating, MT Mobile-terminating

Table B.6: Customer experience voice service normalised indicator calculations and descriptions.

CEI	Type	Description and Formula
VOI_ATTEMPT	Count	Total number of MO and MT voice calls that where attempted on the 2G, 3G and 4G radio layers. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{ATTEMPT})_{i,j}$
VOI_ATTEMPT_TO_ALERTFAIL	Count	Total number of MO and MT voice failures between the attempt and alert phase on the 2G, 3G and 4G radio layers. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{ATTEMPT\_TO\_ALERTFAIL})_{i,j}$
VOI_ALERTATTEMPT	Count	Total number of MO and MT voice alert attempts on the 2G, 3G and 4G radio layers. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{ALERTATTEMPT})_{i,j}$
VOI_ALERTSUCCESS	Count	Total number of successful MO and MT voice alert attempts on the 2G, 3G and 4G radio layers. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{ALERTSUCCESS})_{i,j}$
VOI_ALERT_TO_CONNECTFAIL	Count	Total number of MO and MT voice failures between the alert and connect phase on 2G, 3G and 4G. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{ALERT\_TO\_CONNECTFAIL})_{i,j}$
VOI_SETUPFAILNETWORK	Count	Total number of MO and MT voice setup failures on 2G, 3G and 4G due to network related error causes. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{SETUPFAILNETWORK})_{i,j}$
VOI_CONNECTATTEMPT	Count	Total number of MO and MT voice connect attempts on the 2G, 3G and 4G radio layers. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{CONNECTATTEMPT})_{i,j}$
VOI_CONNECTSUCCESS	Count	Total number of successful MO and MT voice connect attempts on the 2G, 3G and 4G radio layers. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{CONNECTSUCCESS})_{i,j}$
VOI_DROPNETWORK	Count	Total number of MO and MT dropped calls on the 2G, 3G and 4G radio layers due to network related error causes. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{DROPNETWORK})_{i,j}$
VOI_SUCCESS	Count	Total number of successful MO and MT voice calls on the 2G, 3G and 4G radio layers. $= \sum_{i \in \{\texttt{MO},\texttt{MT}\}} \sum_{j \in \{\texttt{2G},\texttt{3G},\texttt{4G}\}} (\texttt{SUCCESS})_{i,j}$
VOI_RA_MOST_USED_ATTEMPT	Count	Number of call attempts on the most used radio access type, based on the maximum attempt count for all radio technologies. $= \max\left[\text{sum}(\texttt{ATTEMPT})_{\texttt{2G}}, \text{sum}(\texttt{ATTEMPT})_{\texttt{3G}}, \text{sum}(\texttt{ATTEMPT})_{\texttt{4G}}\right]$
VOI_CALLHANDOVERATTEMPT	Count	Total number of call handover attempts on the 2G and 3G radio layers. $= \sum_{j \in \{\texttt{2G},\texttt{3G}\}} (\texttt{CALLHANDOVERATTEMPT})_{j}$
VOI_CALLHANDOVERSUCCESS	Count	Total number of successful call handovers on the 2G and 3G radio layers. $= \sum_{j \in \{\texttt{2G},\texttt{3G}\}} (\texttt{CALLHANDOVERSUCCESS})_{j}$
VOI_INTERRATCALLHNDVRATTEMPT	Count	Total number of inter-rat handover attempts between the 3G and 2G radio layers. $= \sum_{j \in \{\texttt{2G},\texttt{3G}\}} (\texttt{INTERRATCALLHNDVRATTEMPT})_{j}$
VOI_INTERRATCALLHNDVRSUCCESS	Count	Total number of successful inter-rat handovers between the 3G and 2G radio layers. $= \sum_{j \in \{\texttt{2G},\texttt{3G}\}} (\texttt{INTERRATCALLHNDVRSUCCESS})_{j}$
VOI_INTRARATCALLHNDVRATTEMPT	Count	Total number of intra-rat handover attempts on the 2G and 3G radio layers. $= \sum_{j \in \{\texttt{2G},\texttt{3G}\}} (\texttt{INTRARATCALLHNDVRATTEMPT})_{j}$
VOI_INTRARATCALLHNDVRSUCCESS	Count	Total number of successful intra-rat handovers on the 2G and 3G radio layers. $= \sum_{j \in \{\texttt{2G},\texttt{3G}\}} (\texttt{INTRARATCALLHNDVRSUCCESS})_{j}$

MO: Mobile-originating, MT Mobile-terminating

Table B.7: Customer experience voice counter aggregations and descriptions.

CEI	Type	Description and Formula
VOI_AVGSETUPTIME_N	Time *(ms)*	Total time in milliseconds that calls were in the setup phase for MO and MT voice calls on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MO,MT}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{AVGSETUPTIME\_N})_{i,j}$
VOI_AVGSETUPTIME_D	Count	Total number of MO and MT voice call attempts on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MO,MT}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{AVGSETUPTIME\_D})_{i,j}$
VOI_AVGSETUPTIME	Time *(ms)*	Average call setup time in milliseconds calculated as the total time calls are in the setup phase divided by the total number of call attempts. $= \frac{\texttt{AVGSETUPTIME\_N}}{\texttt{AVGSETUPTIME\_D}}$
VOI_AVGDURATION_N	Time *(s)*	Total time in seconds that connected voice calls were active on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MO,MT}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{AVGDURATION\_N})_{i,j}$
VOI_AVGDURATION_D	Count	Total number of voice connected calls that were active on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MO,MT}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{AVGDURATION\_D})_{i,j}$
VOI_AVGDURATION	Time *(s)*	Average voice call duration in seconds calculated as the total time calls were active divided by the total number of connected calls. $= \frac{\texttt{AVGDURATION\_N}}{\texttt{AVGDURATION\_D}}$
VOI_RATIO_MODUR	Rate	Ratio of mobile originating call duration versus total call duration across 2G, 3G and 4G radio layers. $= \frac{\texttt{MOAVGDURATION\_N}}{\texttt{MOAVGDURATION\_N} + \texttt{MTAVGDURATION\_N}} \times 100$
VOI_MOAVGDURATION_N	Time *(s)*	Total time in seconds that connected MO voice calls were active on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MO}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{MOAVGDURATION\_N})_{i,j}$
VOI_MOAVGDURATION_D	Count	Total number of MO voice connected calls that were active on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MO}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{MOAVGDURATION\_D})_{i,j}$
VOI_MOAVGDURATION	Time *(s)*	Average MO voice call duration in seconds calculated as the total time MO calls were active divided by the total number of connected MO calls. $= \frac{\texttt{MOAVGDURATION\_N}}{\texttt{MOAVGDURATION\_D}}$
VOI_MTAVGDURATION_N	Time *(s)*	Total time in seconds that connected MT voice calls were active on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MT}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{MTAVGDURATION\_N})_{i,j}$
VOI_MTAVGDURATION_D	Count	Total number of MT voice connected calls that were active on the 2G, 3G and 4G radio layers. $= \sum_{i\in\{\texttt{MT}\}} \sum_{j\in\{\texttt{2G,3G,4G}\}} (\texttt{MTAVGDURATION\_D})_{i,j}$
VOI_MTAVGDURATION	Time *(s)*	Average MT voice call duration in seconds calculated as the total time MT calls were active divided by the total number of MT connected calls. $= \frac{\texttt{MTAVGDURATION\_N}}{\texttt{MTAVGDURATION\_D}}$
VOI_TOTALDURATION	Time *(s)*	Total voice call duration in seconds calculated as the total time MO and MT calls were active on 2G, 3G and 4G radio layers. $= \texttt{VOI\_MOAVGDURATION\_N} + \texttt{VOI\_MTAVGDURATION\_N}$

MO: Mobile-originating, MT Mobile-terminating

Table B.8: Customer experience voice service time-related indicator calculations and descriptions.

Short Message Service (SMS) experience indicator calculations.

CEI	Type	Description and Formula
SMS_ATTEMPT	Count	Total number of SMS message send or receive attempts on the 2G and 3G radio layers. $= \sum_{i \in \{\mathtt{MO,MT}\}} \sum_{j \in \{\mathtt{2G,3G}\}} (\mathtt{SMS\_ATTEMPT})_{i,j}$
SMS_SUCCESS	Count	Total number of successful SMS message send or receive attempts on the 2G and 3G radio layers. $= \sum_{i \in \{\mathtt{MO,MT}\}} \sum_{j \in \{\mathtt{2G,3G}\}} (\mathtt{SMS\_SUCCESS})_{i,j}$
SMS_FAILRATE	Rate	SMS fail rate calculated as the ratio of SMS that failed, versus the total number of SMS attempts. $= \left(1 - \frac{\mathtt{SUCCESS}}{\mathtt{ATTEMPT}}\right) \times 100$
SMS_AVGDURATION_N	Time *(ms)*	Total SMS delivery time in milliseconds of SMS messages that were sent or received on the 2G and 3G radio layers. $= \sum_{i \in \{\mathtt{MO,MT}\}} \sum_{j \in \{\mathtt{2G,3G}\}} (\mathtt{AVGDURATION\_N})_{i,j}$
SMS_AVGDURATION_D	Count	Total number of SMS messages that were sent or received on the 2G and 3G radio layers. $= \sum_{i \in \{\mathtt{MO,MT}\}} \sum_{j \in \{\mathtt{2G,3G}\}} (\mathtt{AVGDURATION\_D})_{i,j}$
SMS_AVGDURATION	Time *(ms)*	Average SMS delivery time in milliseconds calculated as the total SMS delivery time divided by the total number of SMS messages sent or received. $= \frac{\mathtt{AVGDURATION\_N}}{\mathtt{AVGDURATION\_D}}$
SMS_RA_MOST_USED_ATTEMPT	Count	Number of SMS attempts on the most used radio access type, based on the maximum attempt count for all radio technologies including 2G and 3G. $= \max\left[\mathrm{sum}(\mathtt{ATTEMPT})_{\mathtt{2G}}, \mathrm{sum}(\mathtt{ATTEMPT})_{\mathtt{3G}}\right]$
SMS_RATIO_RA_MOST_USED	Rate	Ratio RA most used calculated as the ratio of SMS attempts on the most used radio access type, versus the total number of SMS attempts. $= \frac{\mathtt{RA\_MOST\_USED\_ATTEMPT}}{\mathtt{ATTEMPT}} \times 100$
SMS_RATIO_MOATTEMPT	Rate	Ratio MO attempt calculated as the ratio of MO SMS attempts, versus the total number of SMS attempts. $= \frac{\mathtt{MOATTEMPT}}{\mathtt{MOATTEMPT} + \mathtt{MTATTEMPT}} \times 100$

MO: Mobile-originating, MT Mobile-terminating, SMS: Short Message Service

Table B.9: Customer experience SMS service indicator calculations and descriptions.

Data Access Bearer (BER) experience indicator calculations.

CEI	Type	Description and Formula
BER_ATTACHATTEMPT	Count	Total number of Attach Requests attempted on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{ATTACHATTEMPT})_j$
BER_ATTACHFAIL	Count	Total number of unsuccessful Attach Requests on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{ATTACHFAIL})_j$
BER_ATTACHFAILRATE	Rate	Attach fail rate calculated as the ratio of unsuccessful Attach Request messages, versus the total number of Attach Request messages. $= \frac{\texttt{ATTACHFAIL}}{\texttt{ATTACHATTEMPT}} \times 100$
BER_ATTACHAVGDURATION_N	Time *(ms)*	Total message duration for Attach Request message types on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{ATTACHAVGDURATION\_N})_j$
BER_ATTACHAVGDURATION_D	Count	Total number of Attach Request message types on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{ATTACHAVGDURATION\_D})_j$
BER_ATTACHAVGDURATION	Time *(ms)*	Average Attach Request time in milliseconds calculated as the total Attach Request time divided by the total number of Attach Request messages sent. $= \frac{\texttt{ATTACHAVGDURATION\_N}}{\texttt{ATTACHAVGDURATION\_D}}$
BER_PDPACTATTEMPT	Count	Total number of PDP Activations attempted on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{PDPACTATTEMPT})_j$
BER_PDPACTFAIL	Count	Total number of unsuccessful PDP Activation Requests on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{PDPACTFAIL})_j$
BER_PDPACTFAILRATE	Rate	PDP Activation fail rate calculated as the ratio of unsuccessful PDP Activation Request messages, versus the total number of PDP Activation Request messages. $= \frac{\texttt{PDPACTFAIL}}{\texttt{PDPACTATTEMPT}} \times 100$
BER_PDPACTAVGDURATION_N	Time *(ms)*	Total message duration for PDP Activation Request message types on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{PDPACTAVGDURATION\_N})_j$
BER_PDPACTAVGDURATION_D	Count	Total number of PDP Activation Request message types on the 2G or 3G radio layers. $= \sum_{j \in \{2G,3G\}} (\texttt{PDPACTAVGDURATION\_D})_j$
BER_PDPACTAVGDURATION	Time *(ms)*	Average PDP Activation Request time in milliseconds calculated as the total PDP Activation Request time divided by the total number of PDP Activation Request messages sent. $= \frac{\texttt{PDPACTAVGDURATION\_N}}{\texttt{PDPACTAVGDURATION\_D}}$

ATTACH: Attach Request, PDPACT: PDPActivation Request

TABLE B.10: Customer experience data bearer indicator calculations and descriptions for 2G or 3G radio access type.

CEI	Type	Description and Formula
BER_EPSATTACHATTEMPT	Count	Total number of EPS Attach Requests attempted on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{EPSATTACHATTEMPT})_j$
BER_EPSATTACHFAIL	Count	Total number of unsuccessful EPS Attach Requests on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{EPSATTACHFAIL})_j$
BER_EPSATTACHAVGDURATION_N	Time *(ms)*	Total message duration for EPS Attach Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{EPSATTACHAVGDURATION\_N})_j$
BER_EPSATTACHAVGDURATION_D	Count	Total number of EPS Attach Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{EPSATTACHAVGDURATION\_D})_j$
BER_DFLTBEARERACTATTEMPT	Count	Total number of Default Bearer Activation Requests attempted on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DFLTBEARERACTATTEMPT})_j$
BER_DFLTBEARERACTFAIL	Count	Total number of unsuccessful Default Bearer Activation Requests on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DFLTBEARERACTFAIL})_j$
BER_DFLTBEARERACTAVGDUR_N	Time *(ms)*	Total message duration for Default Bearer Activation Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DFLTBEARERACTAVGDUR\_N})_j$
BER_DFLTBEARERACTAVGDUR_D	Count	Total number of Default Bearer Activation Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DFLTBEARERACTAVGDUR\_D})_j$
BER_DDCTDBEARERACTATTEMPT	Count	Total number of Dedicated Bearer Activation Requests attempted on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DDCTDBEARERACTATTEMPT})_j$
BER_DDCTDBEARERACTFAIL	Count	Total number of unsuccessful Dedicated Bearer Activation Requests on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DDCTDBEARERACTFAIL})_j$
BER_DDCTDBEARERACTAVGDUR_N	Time *(ms)*	Total message duration for Dedicated Bearer Activation Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DDCTDBEARERACTAVGDUR\_N})_j$
BER_DDCTDBEARERACTAVGDUR_D	Count	Total number of Dedicated Bearer Activation Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{DDCTDBEARERACTAVGDUR\_D})_j$
BER_PDNCONNATTEMPT	Count	Total number of PDN Connectivity Request attempted on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{PDNCONNATTEMPT})_j$
BER_PDNCONNFAIL	Count	Total number of unsuccessful PDN Connectivity Request on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{PDNCONNFAIL})_j$
BER_PDNCONNAVGDUR_N	Time *(ms)*	Total message duration for PDN Connectivity Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{PDNCONNAVGDUR\_N})_j$
BER_PDNCONNAVGDUR_D	Count	Total number of PDN Connectivity Request message types on the 4G radio layer. $= \sum_{j\in\{\texttt{4G}\}}(\texttt{PDNCONNAVGDUR\_D})_j$

EPSATTACH: EPS Attach Request, DFLTBEARERACT: Default Bearer Activation Request
DDCTDBEARERACT: Dedicated Bearer Activation Request, PDNCONN: PDN Connectivity Request

Table B.11: Customer experience data bearer counter calculations and descriptions for 4G radio access type.

CEI	Type	Description and Formula
BER_EPSATTACHFAILRATE	Rate	EPS Attach fail rate calculated as the ratio of unsuccessful EPS Attach Request messages, versus the total number of EPS Attach Request messages. $= \frac{\texttt{EPSATTACHFAIL}}{\texttt{EPSATTACHATTEMPT}} \times 100$
BER_EPSATTACHAVGDURATION	Time *(ms)*	Average EPS Attach Request time in milliseconds calculated as the total EPS Attach Request time divided by the total number of EPS Attach Request messages sent. $= \frac{\texttt{EPSATTACHAVGDURATION\_N}}{\texttt{EPSATTACHAVGDURATION\_D}}$
BER_DFLTBEARERACTFAILRATE	Rate	Default Bearer Activation fail rate calculated as the ratio of unsuccessful Default Bearer Activation Request messages, versus the total number of Default Bearer Activation Request messages. $= \frac{\texttt{DFLTBEARERACTFAIL}}{\texttt{DFLTBEARERACTATTEMPT}} \times 100$
BER_DFLTBEARERACTAVGDUR	Time *(ms)*	Average Default Bearer Activation Request time in milliseconds calculated as the total Default Bearer Activation Request time divided by the total number of Default Bearer Activation Request messages sent. $= \frac{\texttt{DFLTBEARERACTAVGDUR\_N}}{\texttt{DFLTBEARERACTAVGDUR\_D}}$
BER_DDCTDBEARERACTFAILRATE	Rate	Dedicated Bearer Activation fail rate calculated as the ratio of unsuccessful Dedicated Bearer Activation Request messages, versus the total number of Dedicated Bearer Activation Request messages. $= \frac{\texttt{DDCTDBEARERACTFAIL}}{\texttt{DDCTDBEARERACTATTEMPT}} \times 100$
BER_DDCTDBEARERACTAVGDUR	Time *(ms)*	Average Dedicated Bearer Activation Request time in milliseconds calculated as the total Dedicated Bearer Activation Request time divided by the total number of Dedicated Bearer Activation Request messages sent. $= \frac{\texttt{DDCTDBEARERACTAVGDUR\_N}}{\texttt{DDCTDBEARERACTAVGDUR\_D}}$
BER_PDNCONNFAILRATE	Rate	PDN Connectivity fail rate calculated as the ratio of unsuccessful PDN Connectivity Request messages, versus the total number of PDN Connectivity Request messages. $= \frac{\texttt{PDNCONNFAIL}}{\texttt{PDNCONNATTEMPT}} \times 100$
BER_PDNCONNAVGDUR	Time *(ms)*	Average PDN Connectivity Request time in milliseconds calculated as the total PDN Connectivity Request time divided by the total number of PDN Connectivity Request messages sent. $= \frac{\texttt{PDNCONNAVGDUR\_N}}{\texttt{PDNCONNAVGDUR\_D}}$

EPSATTACH: EPS Attach Request, DFLTBEARERACT: Default Bearer Activation Request
DDCTDBEARERACT: Dedicated Bearer Activation Request, PDNCONN: PDN Connectivity Request

Table B.12: Customer experience data bearer indicator calculations and descriptions for 4G radio.

Mobile Data Applications (CNA) network analytics indicator calculations.

CEI	Type	Description and Formula
CNA_LATENCY_MS_N	Time *(ms)*	The total TCP Round Trip Time (RTT) in milliseconds for all TCP message pairs on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} (\texttt{LATENCY\_MS\_N})_j$
CNA_LATENCY_MS_D	Count	The total number of TCP Round Trip events on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} (\texttt{LATENCY\_MS\_D})_j$
CNA_UP_MBYTES	Count	Total volume in Megabytes of data sent on the uplink for all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{UP\_BYTES}_j / 1 \times 10^6$
CNA_DOWN_MBYTES	Count	Total volume in Megabytes of data received on the down-link for all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{DOWN\_BYTES}_j / 1 \times 10^6$
CNA_UP_RETRANSMIT_PACKETS	Count	Total count of retransmitted uplink packets on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{UP\_RETRANSMIT\_PACKETS}_j$
CNA_UP_PACKETS	Count	Total count of uplink packets on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{UP\_PACKETS}_j$
CNA_DOWN_RETRANSMIT_PACKETS	Count	Total count of retransmitted down-link packets on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{DOWN\_RETRANSMIT\_PACKETS}_j$
CNA_DOWN_PACKETS	Count	Total count of down-link packets on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{DOWN\_PACKETS}_j$
CNA_UP_THROUGHPUT_MBIT	Count	Total volume of Megabits uplink measuring application network usage on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{UP\_THROUGHPUT\_MBIT}_j$
CNA_UP_THROUGHPUT_SECONDS	Time *(s)*	Total active uplink time for an application measuring network usage on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{UP\_THROUGHPUT\_SECONDS}_j$
CNA_DOWN_THROUGHPUT_MBIT	Count	Total volume of Megabits down-link measuring application network usage on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{DOWN\_THROUGHPUT\_MBIT}_j$
CNA_DOWN_THROUGHPUT_SECONDS	Time *(s)*	Total active down-link time for an application measuring network usage on all radio technologies. $= \sum_{j \in \{\texttt{2G,3G,4G}\}} \texttt{DOWN\_THROUGHPUT\_SECONDS}_j$

MBYTE: Megabyte, MBIT: Megabit

Table B.13: Mobile applications network customer experience counter and aggregation descriptions.

CEI	Type	Description and Formula
CNA_LATENCY_MS	Time *(ms)*	Average Round Trip Time in milliseconds calculated as the total TCP Round Trip Time for all TCP message pairs divided by the total number of TCP Round Trip events. $= \frac{\texttt{LATENCY\_MS\_N}}{\texttt{LATENCY\_MS\_D}}$
CNA_UP_RETRANSMITRATE	Rate	Ratio of retransmitted uplink packets to total uplink packets. $= \frac{\texttt{UP\_RETRANSMIT\_PACKETS}}{\texttt{UP\_PACKETS}} \times 100$
CNA_DOWN_RETRANSMITRATE	Rate	Ratio of retransmitted down-link packets to total down-link packets. $= \frac{\texttt{DOWN\_RETRANSMIT\_PACKETS}}{\texttt{DOWN\_PACKETS}} \times 100$
CNA_RETRANSMITRATE	Rate	Ratio of retransmitted uplink and down-link packets, to total uplink and down-link packets. $= \frac{\texttt{UP\_RETRANSMIT\_PACKETS} + \texttt{DOWN\_RETRANSMIT\_PACKETS}}{\texttt{UP\_PACKETS} + \texttt{DOWN\_PACKETS}} \times 100$
CNA_UP_THROUGHPUT_KBPS	Rate *(kbps)*	Average uplink throughput for applications measuring network usage calculated as the total upload volume in Kilobits divided by the total active upload time in seconds. $= \frac{\texttt{UP\_THROUGHPUT\_MBIT}}{\texttt{UP\_THROUGHPUT\_SECONDS}} \times 1 \times 10^3$
CNA_DOWN_THROUGHPUT_KBPS	Rate *(kbps)*	Average down-link throughput for applications measuring network usage calculated as the total download volume in Kilobits divided by the total active download time in seconds. $= \frac{\texttt{DOWN\_THROUGHPUT\_MBIT}}{\texttt{DOWN\_THROUGHPUT\_SECONDS}} \times 1 \times 10^3$
CNA_TOTAL_MBYTES	Count	Total combined uplink and down-link volume in Megabytes. $= \texttt{UP\_MBYTES} + \texttt{DOWN\_MBYTES}$
CNA_TOTAL_PACKETS	Count	Total count of combined uplink and down-link transmitted packets on all radio technolgies. $= \texttt{UP\_PACKETS} + \texttt{DOWN\_PACKETS}$
CNA_RATIO_MURAT	Rate	Ratio RA most used calculated as the ratio of transmitted data packets on the most used radio access type, versus the total number of data packets transmitted on all radio types. $= \frac{\texttt{PACKETS}_{\texttt{RA\_MOST\_USED\_PACKETS}}}{\texttt{TOTAL\_PACKETS}} \times 100$
CNA_RATIO_MUAPP	Rate	Ratio RA most used calculated as the ratio of transmitted data packets on the most used radio access type, versus the total number of data packets transmitted on all radio types. $= \frac{\texttt{PACKETS}_{\texttt{RA\_MOST\_USED\_PACKETS}}}{\texttt{TOTAL\_PACKETS}} \times 100$

MBYTE: Megabyte, MBIT: Megabit, MBPS: Megabit per Second, KBPS: Kilobit per Second

Table B.14: Mobile applications network customer experience indicator calculations and descriptions.

***

Bibliography

"3rd Generation Partnership Project" (n.d.). *3GPP*. URL: https://www.3gpp.org/about-us/introducing-3gpp.

Anaman, Michael and Mark Lycett (Jan. 2010). "Toward a Model of Customer Experience". In: *In Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pp. 385–390. DOI: 10.5220/0002888003850390.

Becker, Larissa and Elina Jaakkola (Jan. 2020). "Customer Experience: Fundamental Premises and Implications for Research". In: *Journal of the Academy of Marketing Science* 48, pp. 630–648. DOI: 10.1007/s11747-019-00718-x.

Casas, Pedro et al. (May 2017). "Predicting QoE in cellular networks using machine learning and in-smartphone measurements". In: *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. DOI: 10.1109/QoMEX.2017.7965687.

CG (2023). *Net Promoter Score Benchmarks for Fortune 500 Companies*. [Accessed: 2023-01-11]. Customer Guru. URL: https://customer.guru/net-promoter-score/fortune-500.

Cheng, Yu et al. (2017). "A Survey of Model Compression and Acceleration for Deep Neural Networks". In: *CoRR* abs/1710.09282. URL: http://arxiv.org/abs/1710.09282.

Consulta (n.d.). *The South African Customer Satisfaction Index*. [Accessed: 2023-02-06]. URL: https://consulta.co.za/the-sacsi.

Department of Telecommunications and Postal Services (Oct. 2016). *Electronic Communications Act: National Integrated ICT Policy White Paper*. ICT Policy - Electronic Communications. [Accessed: 2023-01-11]. URL: https://www.gov.za/sites/default/files/gcis_document/201610/40325gon1212.pdf.

Ghorbani, Hamid (June 2019). "Mahalanobis Distance and its Application for Detecting Multivariate Outliers". In: *Facta Universitatis*. Mathematics and Informatics 34.3, pp. 583–595. ISSN: 2406-047X. DOI: https://doi.org/10.22190/FUMI1903583G.

Grisaffe, Douglas B. (2007). "Questions about the Ultimate Question: Conceptual Considerations in Evaluating Reichheld's Net Promoter Score (NPS)". In: *The Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior* 20, p. 36.

Hamilton, David et al. (May 2014). "Assessing Treatment Outcomes Using a Single Question: The Net Promoter Score". In: *The Bone & Joint Journal* 96-B, pp. 622–628. DOI: 10.1302/0301-620X.96B5.32434.

Hinton, Geoffrey E., Oriol Vinyals, and Jeffrey Dean (2015). "Distilling the Knowledge in a Neural Network". In: *ArXiv* abs/1503.02531.

Ickin, Selim et al. (2019). “On Network Performance Indicators for Network Promoter Score Estimation”. In: pp. 1–3. DOI: 10.1109/QoMEX.2019.8743206.

International Organization for Standardisation (2019). “ISO8601-1:2019 Date and time – Representations for information interchange – Part 1: Basic rules”. In.

Jed Wing, Max Kuhn. Contributions from et al. (2019). *caret: Classification and Regression Training*. R package version 6.0-84. URL: https://CRAN.R-project.org/package=caret.

Kuhn, Max and Hadley Wickham (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* URL: https://www.tidymodels.org.

Kulas, John T. and Alicia A. Stachowski (2009). “Middle Category Endorsement in Odd-numbered Likert Response Scales: Associated Item Characteristics, Cognitive Demands, and Preferred Meanings”. In: *Journal of Research in Personality* 43.03, pp. 489–493. ISSN: 0092-6566. DOI: 10.1016/j.jrp.2008.12.005.

Laghari, Khalil Ur Rehman and Kay Connelly (Apr. 2012). “Toward Total Quality of Experience: A QoE Model in a Communication Ecosystem”. In: *IEEE Communications Magazine* 50.4, pp. 58–65. ISSN: 0163-6804. DOI: 10.1109/MCOM.2012.6178834.

Lemon, Katherine N. and Peter C. Verhoef (2016). “Understanding Customer Experience Throughout the Customer Journey”. In: *Journal of Marketing* 80.6, pp. 69–96. DOI: 10.1509/jm.15.0420.

Moldovan, Christian and Florian Metzger (Sept. 2016). “Bridging the Gap between QoE and User Engagement in HTTP Video Streaming”. In: DOI: 10.1109/ITC-28.2016.122.

Möller, Sebastian and Alexander Raake, eds. (2014). *Quality of Experience: Advanced Concepts, Applications and Methods*. Cham: Springer. ISBN: 978-3-319-02680-0. DOI: 10.1007/978-3-319-02681-7.

Peterson, Ryan A. and Joseph E. Cavanaugh (2019). “Ordered quantile normalisation: a semiparametric transformation built for the cross-validation era”. In: *Journal of Applied Statistics*, pp. 1–16. DOI: 10.1080/02664763.2019.1630372.

Reichheld, Frederick (n.d.). *Companies That Use Net Promoter*. [Accessed: 2023-02-08]. URL: https://www.netpromotersystem.com/about/companies-that-use-net-promoter/.

– (Dec. 2003). “The One Number You Need to Grow”. In: *Harvard Business Review* 81, pp. 46–54, 124.

South African Government (2013). *Protection of Personal Information Act 4 of 2013*. URL: https://www.gov.za/documents/protection-personal-information-act.

Spiess, Jeffrey et al. (2014). “Using Big Data to Improve Customer Experience and Business Performance”. In: *Bell Labs Technical Journal* 18.4, pp. 3–17. DOI: 10.1002/bltj.21642.

The GSM Association (2017). *Safety, Privacy and Security Across the Mobile Ecosystem*. Tech. rep. URL: https://www.gsma.com/publicpolicy/resources/safety-privacy-security-across-mobile-ecosystem.

Tong, Lili et al. (Nov. 2017). “The Research of Customer Loyalty Improvement in Telecom Industry Based on NPS Data Mining”. In: vol. 14, pp. 260–268. DOI: 10.1109/CC.2017.8233665.

van Buuren, Stef and Karin Groothuis-Oudshoorn (2011). “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3, pp. 1–67. DOI: 10.18637/jss.v045.i03.