Politecnico di Milano

Dipartimento di Elettronica, Informazione e Bioingegneria
Master of Science in Telecommunication Engineering

Prediction of cellular customers satisfaction with network measurements at 4G radio access

Supervisor:
Prof. Alessando E. C. Redondi
Co-Supervisor:
Ing. Andrea Pimpinella

Master Thesis by:
Andrea Marabita
Matr. 903420

Academic Year 2019 - 2020

Ringraziamenti

Innanzitutto desidero ringraziare il prof. Redondi per essere stato oltre che un ottimo professore un vero e proprio mentore. Da professore è riuscito a trasmettermi la curiosità necessaria e la passione per la ricerca sul campo. Da relatore, invece, il prof. Redondi mi ha dato la possibilità di affrontare e approfondire questo progetto di tesi, fornendomi, nonostante le difficoltà incontrate in questo particolare periodo, supporto, continua disponibilità e senso critico finalizzati all'analisi e al perfezionamento della ricerca. Un particolare ringraziamento va ad Andrea per aver creduto in me ogni giorno, avermi seguito con costanza durante la sperimentazione ed, in particolar modo, durante la stesura. Il loro supporto tecnico e la loro esperienza sono stati fondamentali per lo svolgimento di questo studio.

Ringrazio tutti i miei amici che mi sono stati vicini e, nonostante le distanze, mi hanno sempre sostenuto durante questo percorso, permettendomi di dare il meglio nella carriera e nella vita. Ringrazio particolarmente chi è sempre riuscito ad essere presente in questi anni, dagli amici di infanzia Simone T., Gianluca C., Roberta G., a quelli che si sono rivelati come fratelli, Giorgio D., Andrea G.

Ringrazio Daniele M., Giovanni S., Paolo F., che più di tutti mi hanno trasmesso la loro continua voglia di imparare e migliorarsi e mi hanno spronato nei momenti più difficili.
Un ulteriore ringraziamento va a Carmelo F., Med E., Lorenzo M., Alessandro M., Alessandro S. che hanno condiviso con me, durante gli anni di questo percorso, gioie e dolori, e con cui ho stretto un legame che resterà negli anni.
Ringrazio tutti i colleghi del patio, una seconda famiglia a Milano, con cui ho passato giorni e notti e dove ho trovato sempre reciproco supporto, tecnico e personale.

Ringrazio chi mi ha sostenuto sempre in questo percorso di crescita, un'amica speciale, una sorella acquisita. Questo nostro indissolubile rapporto fatto di allegria, pazzie e leggerezza mi ha dato la forza e mi ha accompagnato dolcemente in questi anni, la certezza di poter sempre contare su di te mi rassicura. Questo ringraziamento va a te, Giulia.

Ringrazio chi mi è stato più vicino durante questo periodo di tesi, e più in generale durante tutto il mio percorso accademico, a partire dalla mia famiglia che ha sempre creduto in me, che mi ha permesso la realizzazione di ciò in cui credevo, che mi ha sostenuto nei momenti più difficili e gioito con me in quelli più felici. Ringrazio mia madre, da cui imparo ogni giorno ad affrontare la vita, che dispensa da sempre ottimi consigli personali, maestra di tenacia. Ringrazio mio padre, fonte inesauribile di pazienza e di perseveranza nel lavoro e guida essenziale ai valori familiari.

Infine un ringraziamento speciale va alla persona che mi ha visto crescere e che, sapendo di dovermi lasciare prima degli altri, mi ha donato finché ha potuto il più grande amore. Sei riuscita a trasmettermi i valori in cui credevi, tenendo salda la famiglia, amando incondizionatamente i nipoti. Resterai nel mio cuore e festeggerai al mio fianco.
Questa tesi la dedico a te, Nonna.

Abstract

The expected characteristics that the mobile network connectivity market will assume with the well-established adoption of 4G technology and the imminent deployment of 5G technology, together with the rising global demand for related services, will lead to a challenging scenario for mobile network operators. Operators need to face the challenge by reshaping their investments in all growing network domains and focusing in particular on providing customers with the highest Quality of Service (QoS). These objectives, however, must necessarily be supported by strategic plans aimed at reducing the number of churners, i.e. those customers who, dissatisfied with the service offered, interrupt their subscription, and at recovering their consent and trust in the service subscribed. The detection of dissatisfied users will therefore represents, for mobile operators, the key point for the subsequent strategic planning and the tracking of the customer's degree of satisfaction related to the services offered. The method that allows this monitoring will be the measurement of the Quality of Experience (QoE) perceived by users. Due to the high cost of conducting survey campaigns and the problems associated with them, such as poor customer cooperation, it is crucial for operators to invest in research into the relationship between QoS and QoE. An effective solution to these problems is through the use of Machine Learning models, which can predict QoE directly from data rather than collecting customer feedbacks. This thesis proposes a method for the detection of potential churners, predicting their level of satisfaction for three classes of service: video streaming quality, network speed, network coverage. In addition, the detection of dissatisfied customers can be easily extended to the identification of under-performing radio cells, since for each user are used the data of the most visited radio cell only. Assuming that users who spend most of their time in poorly performing cells experience low quality service, the major objective is to understand whether it is possible to identify the network services or procedures that cause the perceived poor quality of service. The different network aspects that most affect the user experience are: Accessibility, Retainability, Mobility, Integrity and Availability of the mobile service. The proposed

method has been validated on the entire Italian LTE network of a large European mobile operator. In fact, we will collect, through customer responses to a satisfaction survey campaign on service classes that contains when and which cells are visited by customers, the network side measurements that the operator stores in the Operational Support System. Afterwards we will proceed to a composition of Key Performance Indicators (KPIs) that can significantly describe the different network aspects mentioned above, and finally we will implement the engineering of these KPIs in order to extract as much information as possible from the customer visit days. Therefore, we will assess the impact of the application of different Machine Learning algorithms to predict the level of user satisfaction for the different classes of service. Finally, we will evaluate how certain KPIs impact on the perceived quality of the different services on the models. The results suggest that i) it is possible to predict a customer's satisfaction with the video streaming service by using network side data, as a matter of fact the measures of the success rate of handover procedures between different Radio Access Technology (RAT), the maximum number of users connected to the cell and the overall traffic volume have proved to have a high value in predicting satisfaction related to this service, in fact we reach improvements of 34% for the F1 score, and 12. 3% for the prediction accuracy with respect to the reference case; ii) measures about the success rate of the handover procedure between different communication frequencies in the same RAT and the connection to the radio cell, as well as the volume of download traffic have a direct impact on the quality of the network speed experience, allowing us to raise the accuracy of the model by 17.9% with respect to the reference threshold; (iii) time measurements of full and limited mobile service activity have a homogeneous impact on the perceived experience for all classes of service; (iv) the greater the number of network performance descriptors and survey responses which can be leveraged in the training of supervised machine learning models, the more accurate they are in recognizing the level of satisfaction of customers.

Sommario

Le caratteristiche che il mercato della connettività di rete mobile andrà assumendo con la già consolidata diffusione della tecnologia 4G e con l'imminente distribuzione della tecnologia 5G, contestualmente all'incremento della domanda globale dei servizi ad essa connessi, apriranno uno scenario nuovo e ricco di sfide che gli operatori mobile dovranno affrontare rimodulando i loro investimenti in tutti i crescenti domini di rete ed orientandosi sempre più a fornire ai clienti la migliore Qualità del Servizio (QoS). Tale finalità, tuttavia, dovrà necessariamente associarsi a strategie operative che mirino a contrarre il numero dei churners, cioè di quei clienti che, insoddisfatti del servizio offerto, interrompono la loro sottoscrizione, ed a recuperare il loro consenso e la loro fiducia nel servizio sottoscritto. L'individuazione degli utenti insoddisfatti rappresenterà pertanto, per gli operatori mobile, il punto chiave per la successiva pianificazione strategica ed il monitoraggio del grado di soddisfazione dell'utente relativamente al servizio offerto. Il metodo che permetterà questa azione di controllo sarà la misurazione della Qualità dell'Esperienza (QoE) percepita dagli utenti. A causa dei costi elevati per condurre campagne di sondaggi e dei problemi legati ad esse, come la scarsa attitudine dei clienti a partecipare, è fondamentale per gli operatori investire nella ricerca riguardo la relazione tra QoS e QoE. Una soluzione efficace a questi problemi è l'uso di modelli di Machine Learning, in grado di predire direttamente la QoE dai dati piuttosto che collezionare feedbacks dei clienti. Questa tesi propone un metodo per la rilevazione di potenziali churners, predicendo il loro livello di soddisfazione per tre classi di servizio: qualità dello streaming video, velocità di rete, copertura di rete. Inoltre la rilevazione di clienti insoddisfatti è facilmente estendibile all'identificazioni di celle radio sotto performanti in quanto per ogni utente vengono usati i dati della sola cella radio più visitata. Assumendo che, gli utenti che passano la loro maggior parte del tempo in celle con prestazioni scadenti sperimentino un servizio di bassa qualità, l'obiettivo principale è capire se è possibile identificare i servizi o le procedure di rete che causano la scarsa qualità di servizio percepita. A tal proposito sono oggetto di studio di questa tesi i diversi aspetti di rete che influiscono maggior-

mente sull'esperienza utente: Accessibilità, Mantenibilità, Mobilità, Integrità e Disponibilità del servizio mobile. Il metodo proposto è stato validato sull'intera rete LTE Italiana di un grande operatore mobile Europeo. Raccoglieremo infatti, attraverso le risposte dei clienti ad una campagna di sondaggi sulla soddisfazione relativa alle classi di servizio contenente, tra l'altro, quando e quali celle vengono visitate dai clienti, le misurazioni lato rete che l'operatore immagazzina nell'Operational Support System. Procederemo poi ad una composizione di Key Performance Indicators (KPIs) che possano descrivere significativamente i diversi aspetti di rete sopra menzionati, per infine attuare l'ingegnerizzazione di questi KPI in modo da estrarre più informazione possibile dai giorni di visita del cliente. Valuteremo quindi l'impatto dell'applicazione di diversi algoritmi di Machine Learning atti a predire il livello di soddisfazione utente per le diverse classi di servizio. Infine, valuteremo sui modelli l'impatto di determinati KPI sulla qualità percepita per i diversi servizi. I risultati suggeriscono che i) è possibile predire la soddisfazione di un cliente riguardo al servizio di video streaming sfruttando i dati lato rete, infatti, le misure del tasso di successo delle procedure di handover tra diversi Radio Access Technology (RAT), il massimo numero di utenti connessi alla cella e il volume del traffico complessivo hanno dimostrato di avere un'alta valenza nella predizione della soddisfazione relativa a questo servizio, per questi motivi raggiungiamo migliorie del 34% per lo score F1, e del 12.3% per l'accuratezza di predizione rispetto al caso di riferimento; ii) misure sul tasso di successo della procedura di handover tra frequenze di comunicazione diverse nella stessa RAT e della connessione alla cella radio, oltre che il volume di traffico in scaricamento impattano direttamente sulla qualità dell'esperienza relativa alla velocità di rete, permettendoci di incrementare l'accuratezza del modello del 17,9% rispetto alla soglia di riferimento ; iii) le misurazioni temporali di pieno e limitato funzionamento del servizio mobile hanno un impatto omogeneo sull'esperienza percepita per tutte le classi di servizio; iv) maggiore è il numero dei descrittori delle prestazioni di rete e delle risposte ai sondaggi che è possibile sfruttare per l'addestramento di modelli di Machine Learning supervisionati, maggiore è la loro accuratezza di riconoscimento del livello di soddisfazione.

Contents

1 Introduction **1**
1.1 Thesis Outline 4

2 State of the art **7**
2.1 4G Network Performance Indicators 7
2.2 User satisfaction assessment 9
2.2.1 Short-term customers QoE prediction 9
2.2.2 Long-term customers QoE prediction 10

3 Datasets **13**
3.1 Data Retrieval 14
3.2 LTE Network Measurement Dataset 15
3.2.1 Accessibility KPI 16
3.2.2 Retainability KPI 16
3.2.3 Mobility KPI 16
3.2.4 Integrity KPI 17
3.2.5 Availability KPI 17
3.2.6 Traffic KPI 17
3.3 User Satisfaction Dataset 18
3.4 Final Dataset Building and Feature Engineering 19

4 Data Analysis **23**
4.1 Data Transformation 23
4.2 Satisfaction Threshold 24

5 Prediction and results **31**
5.1 Prediction Pipeline 31
5.1.1 Feature Selection 32
5.2 Model evaluation 36
5.2.1 Evaluation metrics 36
5.2.2 Driving factors in the choice of metrics 38

5.3 Model comparison 38
5.4 Network-side vs. user-side based prediction models 47

6 Concluding remarks **49**
6.1 Limitations and issues 49
6.2 Analysis of Results and Future Works 50

Bibliography **53**

List of Figures

1.1 Global Mobile subscriber growth (Source: Cisco Annual Internet Report, 2018-2023). 1

1.2 Global Mobile Connections growth (Source: Cisco Annual Internet Report, 2018-2023). 2

2.1 4G/LTE Key performances indicators 8

3.1 KPI Analyzer GUI screenshot 14

3.2 LTE Architecture 15

3.3 Satisfaction grades distributions. 19

3.4 Average relative time spent in the 20 most visited cells for each user. 20

4.1 Minimum RRC Setup Success Rate distribution and transformed distribution 24

4.2 Minimum Downlink Volume distribution and transformed distribution 24

4.3 Network Coverage relative minimum Inter frequency Handover out Success Rate Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8 26

4.4 Network Coverage relative No Service Time Ratio Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8 27

4.5 Video Quality relative minimum Inter frequency Handover out Success Rate Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8 27

4.6 Video Quality relative No Service Time Ratio Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8 28

4.7 Data Speed relative minimum Inter frequency Handover out Success Rate Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8 28

4.8 Data Speed relative No Service Time Ratio Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8 29

5.1 Data flow and operations in the pipeline. The raw dataset is divided into train set and test set, the train set is used to identify the optimal model (hyper parameter tuning and RFE) and to train it, finally the test set is used to evaluate the prediction performance. 34
5.2 Confusion Matrix 36
5.3 ROC curve for video streaming QoE prediction, $T = 6$ 40
5.4 ROC curve for video streaming QoE prediction, $T = 8$ 41
5.5 ROC curve for data speed QoE prediction, $T = 6$ 42
5.6 ROC curve for data speed QoE prediction, $T = 8$ 43
5.7 ROC curve for coverage QoE prediction, $T = 6$ 45
5.8 ROC curve for coverage QoE prediction, $T = 8$ 46

List of Tables

3.1 Columns and description of LTE Network measurement dataset . 21
3.2 Columns and description of User satisfaction dataset 22

5.1 Video Quality relative tuned hyper parameters 33
5.2 Features selected for each Service Item and Threshold 35
5.3 Baseline performance per Threshold T 39
5.4 Performance obtained for video streaming QoE prediction 39
5.5 Improvements from baseline obtained for video quality QoE prediction 40
5.6 Performance obtained for data speed QoE prediction 43
5.7 Improvements from baseline obtained for data speed QoE prediction 43
5.8 Performance obtained for coverage QoE prediction 44
5.9 Improvements from baseline obtained for coverage QoE prediction 44
5.10 Model comparisons with respect to [1] 47

Chapter 1

Introduction

According to recent Cisco Annual Internet Report (2018,2023)([2]), a global analysis that estimates digital transformation on several business segments, the total number of mobile subscribers, i.e., users that have a mobile device and subscribe to a cellular service, is going to increase from 5.1 Billion (66% of the global population) in 2018 to 5.7 billion(71% of the global population) by 2023 (Figure 1.1)

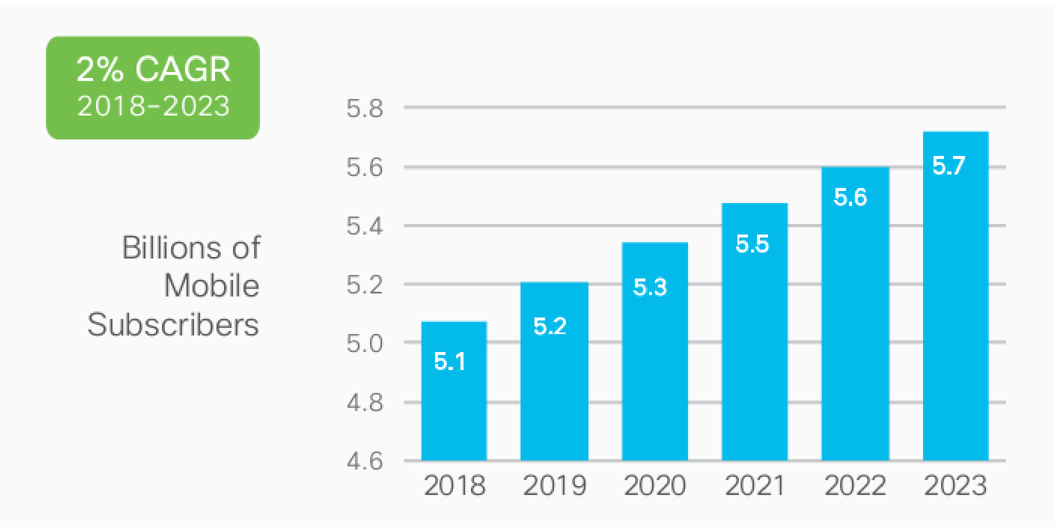

Figure 1.1: Global Mobile subscriber growth (Source: Cisco Annual Internet Report, 2018-2023).

In relation to mobile subscriber, Mobile devices are projected to grow from 4.9 billion in 2018 to 6.7 billion in 2023. A consequence of the growth in the number of mobile subscribers is the explosion of mobile applications and an increased scope of mobile connectivity. In particular, many popular applications in recent years require high bandwidth, such as video streaming, cloud storage, etc. As a consequence Mobile broadband speed is predicted to triple by 2023: as the report estimates, the average network connection speed grows from 13.2 Mbps in 2018 to 53.9 Mbps in 2023. The massive estimated growth in the average speed of connection in the mobile network has resulted in the need to optimize bandwidth management, as can also be seen in the utilization growth of 4G among broadband cellular network technologies. (Figure 1.2) As the report ([2]) states, global 4G connections will be 46% of total mobile connection by 2023, with growing from 3.7 million in 2018 up to 6 million by 2023, as shown in Figure 1.2.

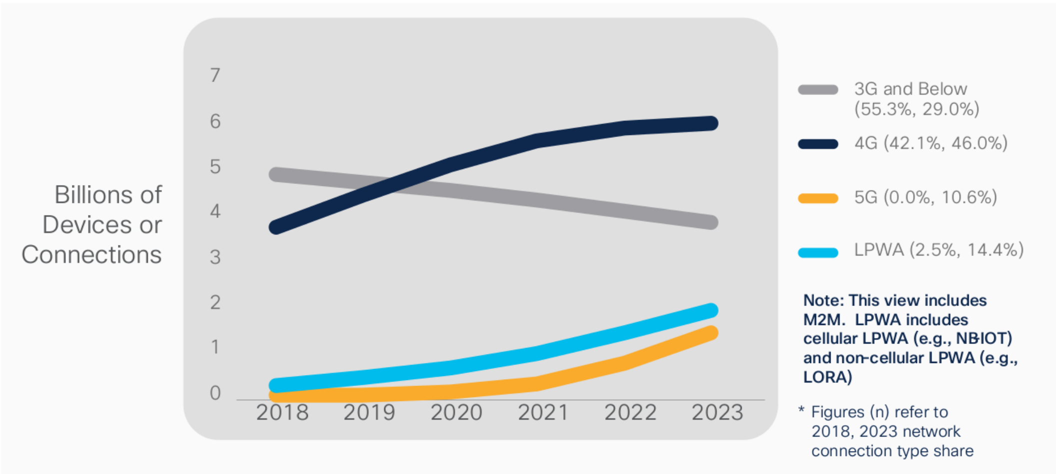

Figure 1.2: Global Mobile Connections growth (Source: Cisco Annual Internet Report, 2018-2023).

Transition from 3G and previous mobile connection technologies to 4G and now 5G deployment follow a global trend which shows a forecast of 60% mobile device with 4G+ connection capability by 2023. Aware of these projections, mobile network operators have to constantly invest in all the growing network domains, especially in 4G and 5G network technologies and architectures. MNOs also constantly monitor and optimize their access networks to ensure the highest quality of service aimed at attracting new customers and limit the number of churners, i.e., dissatisfied users who interrupt their contract for subscribing to another operator. In addition to monitor the satisfaction level of their customers (QoE), i.e, to receive a feedback on experienced QoS, MNOs continuously survey their customers. There are several tools to gather the users' satisfaction level. Some operators use standard tools such as the Net Promoter Score (NPS) survey, which collects the likelihood that the user recommends the operator to friends or coworkers on a scale from 0 to 10. Other operators instead, in addition to this generic score, often ask the user to answer specific questions related to a specific mobile network service (video quality, network data speed, voice quality, network coverage, etc.) that allow to identify possible problems in the network related to the service item. Based on these surveys responses, operators have much information about which service should be optimized and possibly which one should start: for example, the operator can then invest in improving Radio Access Technology or in increasing the available bandwidth. However, collecting feedback from users through surveys is a complex task. On the one hand, relying on users feedbacks is risky as users, in general, have poor cooperative attitudes and also feedback can be influenced by several user subjective factors and also by other elements such as customer age, sex, level of education, etc. On the other hand, MNOs have typically low visibility of network performance observed by end devices. In the QoE domain, several significant researches have focused on QoE modelling, exploring the relationship between QoE and mobile network performance. Most studies generally focus on specific applications ([3, 4, 5, 6]) or a specific service item, e.g. video quality, and frequently monitor the network through measurement platforms such as Netradar ([7, 4]). However, mobile operators have several ways to capture objective measurements from their customers: they can capture both radio access network data and Packet Deep Inspection (DPI) measurements. They can also collect data through specific mobile applications under the user's prior consent ([1]). As an example, in ([8]) authors deployed an automatic churn prediction and retention system, which, leveraging a giant volume of data from the Business Support System(BSS) and Operational Support System(OSS), can detect with high accuracy future churners. However, only a few studies explore the relationship between subjective and objective factors impacting on users satisfaction. Understand this relationship is fundamental to allow operators to detect which part of the network (objective side) causes dissatisfaction (subjective side) among customers and may address the problem by investing in root causes. In this thesis work, we explore the possibility to predict the customer satisfaction related to three different service item: video quality, data speed and network coverage. Predicting satisfaction with the video streaming service is paramount as it constitutes the majority of mobile traffic, particularly in recent years. Besides, due to the rise of HD/4K streaming and high-bandwidth applications, it has become more and more necessary in recent years to maintain a high level of network speed. Furthermore, network coverage is an essential service for users since no network service can clearly be provided without radio coverage. Satisfaction with these three network services has a significant impact on the user's decision to subscribe with a better mobile network operator. Differently from other related studies in the domain which analyze the network on the users' side, in this work, using empirical network-side data (i.e. collected from the operator's OSS), we place the end-user and the specific service as close as possible to their real daily use, providing a more representative evaluation. This is done through the introduction of several key concepts that relate the QoS, i.e. the quality of network service expressed in KPIs, to the QoE, i.e. the quality of the experience perceived by the end-user. We based our study on country-wide dataset obtained from one of the biggest European MNO, containing both ground truth satisfaction dataset (i.e. perceived QoE) and network-side measurement(i.e. quality of service KPI). We describe and study the features extracted from those datasets, and we report the prediction results obtained using these features to train different machine learning models. Finally, investigating the relevance of the features for the different service items, we outline those that reflect a relationship between the performance on the network side, i.e. the objective one, and the quality of the user experience, i.e. the subjective one.

1.1 Thesis Outline

The remainder of this thesis is structured as follows:

In **Chapter 2**, the most relevant related works about 4G mobile network performance indicators, assessment of user satisfaction through different temporal analysis and techniques are described.

In **Chapter 3**, the used LTE network measurement dataset and the categories concepts which compose it as well as their impact on the customers' QoE are explained. Also, a study of survey responses and the procedures we used to engineer the features and to compose the final dataset is illustrated.

In **Chapter 4**, machine learning techniques are employed to perform data analysis on the dataset, also examining the information that each feature brings to the prediction task when the binarization threshold changes.

In **Chapter 5**, the core machine learning techniques and the built prediction pipeline is described in detail. In addition, the main machine learning metrics are presented, and the driving factors in the choice of these are explained. To conclude the chapter, a comprehensive comparison of the ML models performance for each service item is illustrated and argued.

Finally, in **Chapter 6**, we conclude discussing the obtained results and proposing some possible future work to extend the thesis.

Chapter 2

State of the art

This chapter overviews the literature investigated for this thesis work. Section 2.1 presents the works related to the identification of network KPIs reflecting customers QoE in 4G wireless systems, while Section 2.2 reports on research works about the assessment of users satisfaction in mobile networks.

2.1 4G Network Performance Indicators

In order to guarantee the quality of service, Mobile Network Operators (MNOs) continuously monitor their infrastructure through network counters placed at base stations premises, i.e. at the access of the network. Such counters provide the operator with raw measurements that are finally aggregated to compose specific Key Performance Indicators (KPIs) that summarise network performance. The latest standards from European Telecommunications Institute (ETSI) [9][10] recognize six different KPIs categories to describe the performance of Evolved UMTS Terrestrial Radio Access Networks (E-UTRANs), as those that mostly impact on end-users network experience [10]. Also, the standard describes in detail the network services over which the KPIs have to be measured. The list of KPIs categories and corresponding services are:

- **Accessibility** and **Retainability**, referring to E-UTRAN Radio Access Bearer (E-RAB) service,
- **Integrity**, referring to IP packets delivery,
- **Mobility** KPIs referring to the capability of Evolved Node-Bs (eNBs) to successfully prepare and execute handovers,
- **Availability** KPIs referring to Evolved Node-Bs (eNBs) capability to provide the service (i.e. the E-UTRAN Radio Access Bearer between the EU and the Core Network) in their area,

- **Energy Efficiency**, referring to the ratio between the performance indicator (typically the Data Volume in Uplink or Downlink) and the energy consumption during the same time frame.

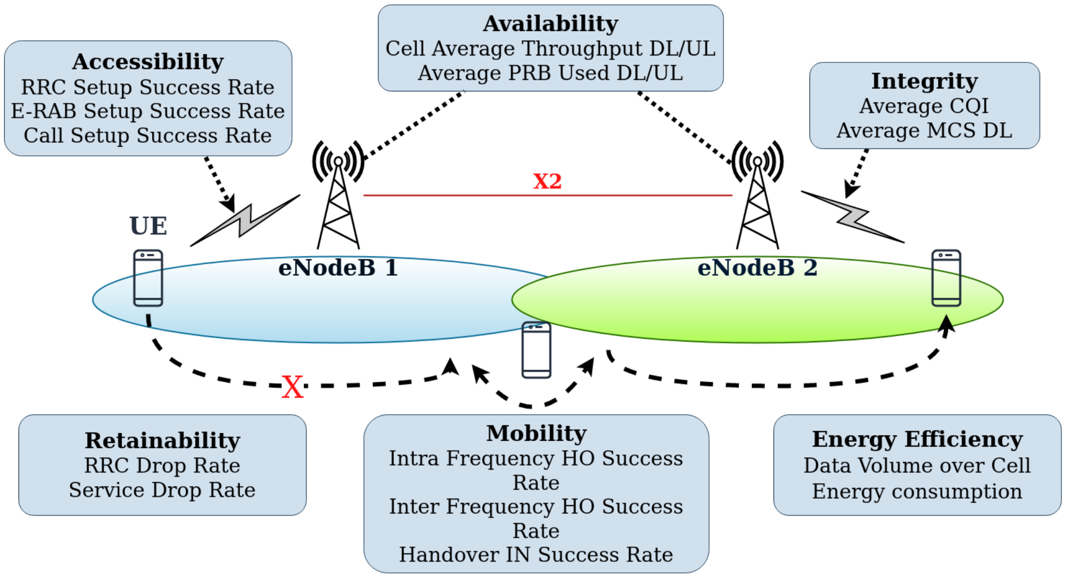

Figure 2.1: 4G/LTE Key performances indicators

Given the above KPIs categories, authors in literature ([11, 12]) widely investigated which KPI can better capture end-user Quality of Experience. In [11], authors focus their work on analyzing accessibility and retainability KPIs for 4G Network in Kosovo. Measuring KPIs through terminals installed by different manufacturers and monitoring 4G network traffic for three months in 2017, they were able to observe a correlation between the growing demand for the 4G service and the performance offered. Specifically, the increase in demand for the network service (from 80 million requests per day to 120 million, with traffic volumes doubling) was reflected in the degradation of 4G network performance, causing a growing number of failed access requests from 0.31% (at the beginning of the observation period) to 0.5% (at the end of the observation period). In [12] the authors verify the feasibility of extracting Key Quality Indicators (KQI) from a real LTE network and propose a KQI-driven anomaly detection and diagnosis framework. Exploiting a group of 19 KPIs, the authors generate a Fault Cause Codebook, which can be adopted by an operator to find the fault cause type. Their analysis shows that the degradation of QoE due to a drop in radio access network KPIs accounts for 80% of overall user QoE and other reasons are typically due to core network failures or service provider issues.

2.2 User satisfaction assessment

Monitoring cellular network customers satisfaction has become a primary task for MNOs, who need to limit the number of churners. In order to monitor users QoE, MNOs continuously survey their customers. However, MNOs have typically low visibility of network performance observed by end devices. Improving the visibility of network performance is a daunting challenge because it requires several non-network factors. To achieve this goal in [3], the authors collected extensive, continuous, and large-scale measurements (over a 17-month time span) from various devices and networks. They studied network performances in three major US cities through device-based measurement applications, turning out that there are significant differences in mobile network performance among different operators, access technologies, geographical regions and over time.

Due to the difficulty of collecting users feedbacks, many authors during the last decade investigated and evaluated the feasibility of predicting short-term ([4, 5]) and long-term ([7, 6, 1]) customers' QoE concerning different network services and mobile applications. Moreover, temporal effects such as recency of experiences affect directly the evaluation. As the authors of [13] state, the peak-end phenomena suggests that an evaluation depends on the peak experience (negative experience with the highest magnitude) and the end experience (last experience of the episode). Therefore, the absence of the adaptation effect and the capture of the end experience could potentially have had a positive impact on prediction performance. On the one hand, short-term QoE concerns individual and time-limited sessions in which users are instructed to use a service (e.g. watching a video content on YouTube) under controlled network environments and then are asked about the quality of their experience. On the other hand, long-term QoE refers to the experience of users in the cellular network for periods spanning over several weeks or months, which is typically composed of many, uncontrolled network events.

2.2.1 Short-term customers QoE prediction

A reliable approach for assessing the performance of networks and services is conducting controlled laboratory experiments, since they rely on the full control of the evaluation process. However, laboratory experiments miss several important QoE influencing factors such as the specific user context or many user subjective factors, limiting the overall view of the services offered by the network. In [5], authors combine subjective controlled lab tests and passive end-device measurements with QoE user feedback on five different applications collected through a field trial. They show that downlink bandwidth fluctuations are crucial in determining the QoE of a service, especially for high-interactive ones. Furthermore, they also state that end-user involvement in the determination process of QoE in mobile devices is essential to achieve reliable QoE ground truths.

In comparison to lab-based experiments, field trial experiments place the end-user and the analyzed aspects of the network as close as possible to their real and everyday scenario, providing more representative evaluations.

In [4] Authors, using a rich QoE dataset taken from field trials in operational cellular networks and applications built to monitor network passively, benchmarking satisfaction prediction performance of different machine learning models.
They studied network performance during specific applications (YouTube, Facebook and Google Maps), and the user reported QoE (Mean Opinion Score and Acceptability) related to a single application.

Out of all the models employed, the best was the one based on decision trees, which predicted the overall experience and acceptability of the service with 91% and 98% accuracy.

2.2.2 Long-term customers QoE prediction

Only a few studies have exploited both measurement applications and QoE surveys to find effective predictors of perceived end-user QoE.

Finley, et al. [7] By Aalto Finley University combined network, non-network data and surveys collected through Netradar platform to study significant predictors of user satisfaction. This client-server based platform provides a suite of mobile applications for various mobile platforms that allows to perform measurements on demand and to take other information of the device, such as location, MNO and in-use mobile platform for the specific user-initiated measurement (Active Measurements).

The authors implemented a custom pop-up survey to collect QoE data from users with five statements, which are classified into the following classes (Class: 'cited original statement'):

- Device QoE: 'I am satisfied with the performance of my mobile device in general.'
- Coverage QoE: 'My mobile connection (current operator) is available when I need it.'
- Speed QoE: 'I am satisfied with the speed of my mobile connection.'

- Video QoE: 'My mobile connection is good enough for watching online videos.'
- Recommendation: 'I would recommend my current operator (current operator) to my friends.'

They used a five-point Likert scale to quantify user satisfaction. Furthermore, analyzing the responses to the polls, they found the largest correlation (0.78) between Speed and Video QoE and also a very high correlation (0.74) between Speed QoE and Recommendation.
After dichotomization of responses (1,2) for unsatisfaction and (3,4,5) for satisfaction and data preprocessing they built a series of ordinal logistic regression models to determine which features are good predictors for Coverage and Speed QoE.

As a result, they found minimum download goodput, number of frequently measured locations, network operator, and device type as good predictors for general user satisfaction.

In [6], the authors studied the performance of the mobile network during the use of specific applications and in specific contexts of use, asking the user to report QoE degradation during their use. They found that even though many users had similar network conditions, the context of use strongly influences the QoE.

In [1] the authors, leveraging user-side networks measurements taken from passive monitoring application, trained many ML algorithms to predict network coverage and video streaming user satisfaction. Explaining the difficulties in the long-term cellular user satisfaction prediction, they propose several action plans to improve the results of the prediction. Since the survey response is also influenced by non-network factors, they state that it is possible to improve the prediction results by including in the model commercial features (e.g. data plan type, fee, etc.) as well as subjective user factors such as age, gender or customer type. A further improvement can also be obtained by increasing data availability provided by operators. They were limited by lack of responses (overall 15% responses for coverage and 5% for video satisfaction).

As pointed out by the authors, including the commercial data from the Business Supporting System (BSS), the prediction performances are improved, as the authors of [8] demonstrated.

Regarding the monitoring of QoE in cellular networks, a variety of tools are used to measure network performance ([14],[15],[7]). Although most of these focus on monitoring the application-specific QoE factors (such as video stalls, downlink bandwidth, round trip time), rather than the whole set of network descriptors that could affect the user experience on a broader class of services.

In [8], authors deployed an automatic churn prediction and retention system for prepaid customers using 9 Month of measurements by one of the biggest mobile operator in China. They exploited for the first time a giant volume of data taken from both operator's BSS and OSS (Operating Supporting System), proving that bigger data quantity can be more valuable assets for improving predictive performances of customers churn. Authors assessed the performances of their churn prediction system through the 3V's perspective, i.e. Volume, Variety and Velocity: results further demonstrate that Variety (i.e. the diversity and quantity of features used in the training phase) plays a more significant role in churn prediction with respect to Volume and Velocity. As a matter of fact, their system reached 0.96 precision for the top 50000 predicted churners in the list, overcoming previous researches and achieving a big business value.

Chapter 3

Datasets

This chapter describes the methods and procedures used for data collection and preprocessing (i.e., data retrieval, data cleaning, data integration, data selection).
This work leverages two datasets referring to the LTE network of one of the major European mobile operator which is currently active in a mid-sized European city. The two datasets consist respectively of country-wide and cross-vendor network data measured at the access of the network through counters installed at base stations premises and of visiting users related information, where users sensible details (e.g. MSISDN) have been properly anonymized by the operator. On the one hand, the former dataset collects raw data measuring different network performance of $75,5k$ different cells (e.g., cell average DL/UL throughput, average PRB utilization, average number of visitors, etc.) referring to six months from November 2019 to April 2020. On the other hand, the latter dataset contains the ground-truth satisfaction feedbacks of 10k users, collected by the MNO through a surveying campaign in a four months period from August to November 2019. Moreover, this dataset contains other users-level information, about i) which network cells each user has visited on 60 days before the response to the survey and ii) for how long (i.e., cell visit times). In the following sections, we provide a detailed description of the two datasets. Section 3.1 presents procedure used in automation of data retrieval process while sections 3.2 and 3.3 describe the generic content of the two datasets. Finally, section 3.4 provides a description of the methods used to join the raw datasets in a final dataset which will be then used in the experiments.

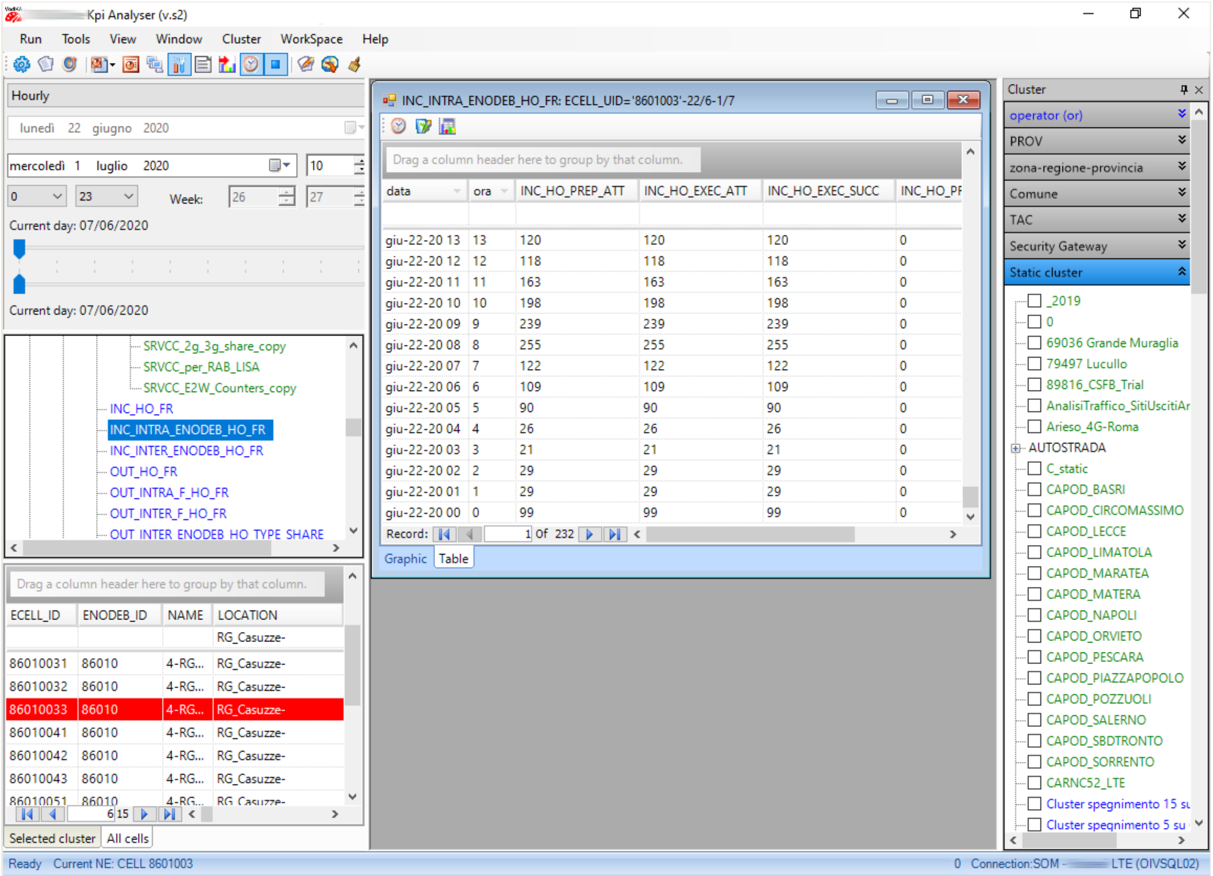

Figure 3.1: KPI Analyzer GUI screenshot

3.1 Data Retrieval

The first contribution provided by this thesis work is the automation of the data retrieval process, previously approached through manual, laborious and time-consuming operations. The data retrieval process was in fact performed through the use of a Graphical User Interface (GUI) proprietary software of the operator, shown in Figure 3.1, which allows to perform some pre-set SQL queries on the OSS server allowing to choose the time granularity, the start and end date of observation and the cell or set of cells (clusters). However, several limitations were present, such as the limited number of cells in a cluster, the complexity in composing new custom SQL queries, and the inability to automate the process for multiple base station vendors. Therefore we decided to automate this process by programming a Python algorithm that takes in input the different custom SQL queries, the start and end date of desired observation period and the subset of cells from which retrieve the data, filters the cells for specific vendor, executes the queries on the dedicated server for each vendor by saving the data in about 60 comma-separated values files (csv) of 250MB, and finally merges the different files providing in output a single csv file for each vendor. This process is repeated for each desired month of observation, producing a large file size (14-18 GB/Month) for each vendor. Finally this raw data will be uploaded to a local SQL server for further data aggregation and merging operations, as described in Section 3.4.

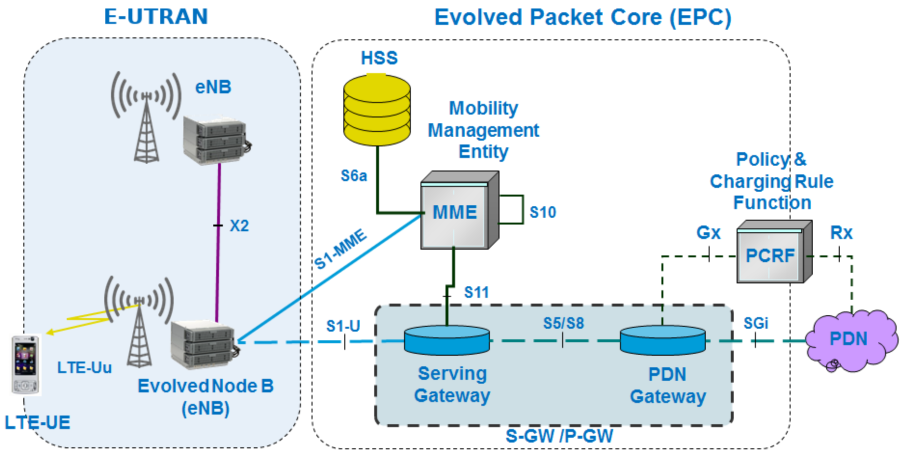

Figure 3.2: LTE Architecture

3.2 LTE Network Measurement Dataset

The first dataset was collected from the Operations Support System (OSS) of the operator according to a preliminary study of the most relevant E-UTRAN KPIs that explain end-user QoE in a 4G network. ([11],[12]) Considering the reference architecture depicted in Figure 3.2, Network performance has been measured countrywide and cross-vendor though base station counters that refer to the following network aspects:

1. Accessibility
2. Retainability
3. Mobility
4. Integrity
5. Availability
6. Traffic

In the following, a description of each KPI category is provided.

3.2.1 Accessibility KPI

Accessibility KPIs provide the network operator with information about whether the services requested by the user can be accessed. Moreover, if a user often cannot access the provided service, he might change his wireless subscription provider due to dissatisfaction. Hence, have good accessibility is fundamental from a QoE and business point of view. The service provided by E-UTRAN is defined as E-RAB, whereas RRC, E-RAB and Call setup are the fundamental procedures for accessibility KPIs. We collected counters for the following accessibility KPIs: RRC Setup Success Rate, RRC Connection Reestablishment Success Rate, E-RAB Setup Success Rate, Call Setup Success Rate.

3.2.2 Retainability KPI

Retainability KPIs provide the capacity of the system to handle request during user use of services and perform its intended function, avoiding interruptions of service. Besides if an end-user is interrupted often during use of the provided service, he might perceive poor QoS. Hence, have a good Retainability is fundamental from a QoE and business point of view. The service provided by E-UTRAN for this KPI is defined as E-RAB. I collected counters for the Service Drop Rate Retainability KPI.

3.2.3 Mobility KPI

Mobility is an essential function that provides a continuous service to users who move across the area covered by the mobile network. Mobility KPIs are about handovers (Hos), i.e. the transfer of an active EU connection from one cell to another. Measurements include both intraE-UTRAN and interRAT Hos, i.e. intraE-UTRAN refer to a connection relocation between different E-NB in the same mobile network. In contrast, InterRAT refers to a connection transfer between different Access Technologies (e.g. LTE to CDMA, etc.). Measurements are performed at cell and cluster level. All Hos in LTE are hard, i.e. the connection between the EU and the RAN is temporarily broken during HOs. We collected counters for the following Mobility KPIs: Intra-Frequency Handover Out Success Rate, Inter-Frequency Handover Out Success Rate, Handover In Success Rate, Inter-RAT Handover Out Success Rate (LTE to CDMA), Inter-RAT Handover Out Success Rate (LTE to WCDMA), Inter-RAT Handover Out Success Rate (LTE to GSM)

3.2.4 Integrity KPI

Service integrity is defined from ITU-T [16] as the degree to which a service is provided without excessive impairments, once obtained. Integrity in Mobile Radio Networks refers to the level of acceptability of service quality provided to the user. The service provided by E-UTRAN for this KPI is defined as the delivery of IP packets. We collected counters for the following Integrity KPIs: Cell Downlink Average Throughput, Cell Uplink Average Throughput

3.2.5 Availability KPI

Availability KPIs measure the percentage of time that the wireless service is available, in LTE networks refer to the percentage of time in which eNB can provide services between UE and Core Network. We collected counters for Radio Network Unavailability Rate KPI.

3.2.6 Traffic KPI

Traffic KPIs provide a measure of traffic volumes and correlated factors such as number of active user on LTE RAN. We collected counters for the following Traffic KPIs: Downlink Traffic Volume, Uplink Traffic Volume, Average User Number, Maximum User Number.

3.3 User Satisfaction Dataset

The second dataset collects the responses of the operator's mobile subscribers to directed surveys which ask the users feedback regarding the quality of their experience in the network. In particular, the considered surveys are comprised of two sections:

- Recommendation: this section asks users to indicate the likelihood of recommending the network operator to a friend or colleague on a scale from 0 to 10 (similarly to Net Promoter Score (NPS) surveys).

- Service Specific: this section asks customers to rate on a scale from 0 to 10 their satisfaction or Quality of Experience (QoE) relative to specific network services. This work considers three different services, namely Network Coverage, Video Quality and Data Speed: each user is randomly asked to provide feedback about one of such services.

Considering that the customers were required to answer at least to the previous question (whereas the latter could be optionally answered), out of 10k answering customers at the end of the surveying campaign only 17% answered the more specific question for one of the considered services. Therefore we have 1.8k customer responses for satisfaction regarding data speed and network coverage and 1.7k customer responses for satisfaction about video quality. Figure 3.3 plots the distribution of users responses for both Recommendation and each Service Specific item. As one can see, distributions are highly skewed towards high grades, with the largest number of users answering with positive feedback. It is reasonable ([7, 1]) to divide users satisfaction labels in 2 classes, with respect to a predefined satisfaction threshold T: users whose grade is less or equal than T are considered as Unsatisfied, while the opposite happens for those whose vote is strictly greater than T. We will detail the choice of threshold T in Section 4.2. To conclude, we remark that in this work, we will perform the prediction of the service-specific survey responses solely, regardless of users responses to the recommendation section. This is because service-specific questions can better capture the quality of customers experience, which might be different for different type of services.

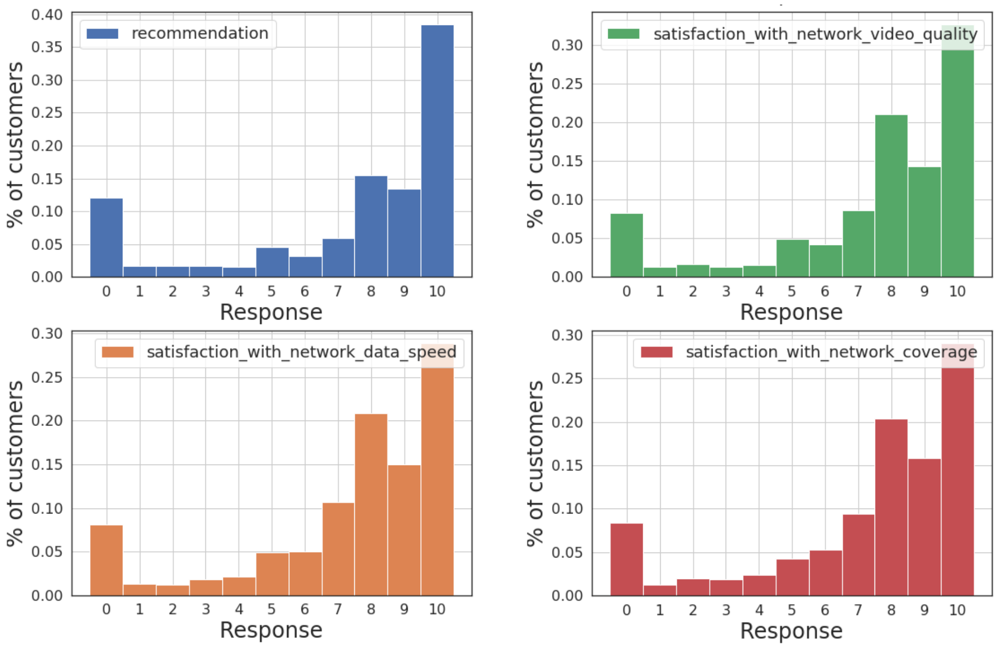

Figure 3.3: Satisfaction grades distributions.

3.4 Final Dataset Building and Feature Engineering

This section shows the activities done to create the final features that will be used to predict the QoE of mobile customers. We extracted all the necessary counters from provider databases through SQL queries to take only visited cells data, using a time granularity of one hour to have the most detailed information possible about the network performances. Given the datasets described above, as a first step, we joined the two datasets through a table join in SQL, these have the entries presented in Tables 3.1 and 3.2.

Analyzing the visit times of all users per cell in Fig. 3.4 using the three counters of the times spent, it is quite clear that on average all users spend more than 50% of their total time in the most visited cell, which we are going to call *home cell*.

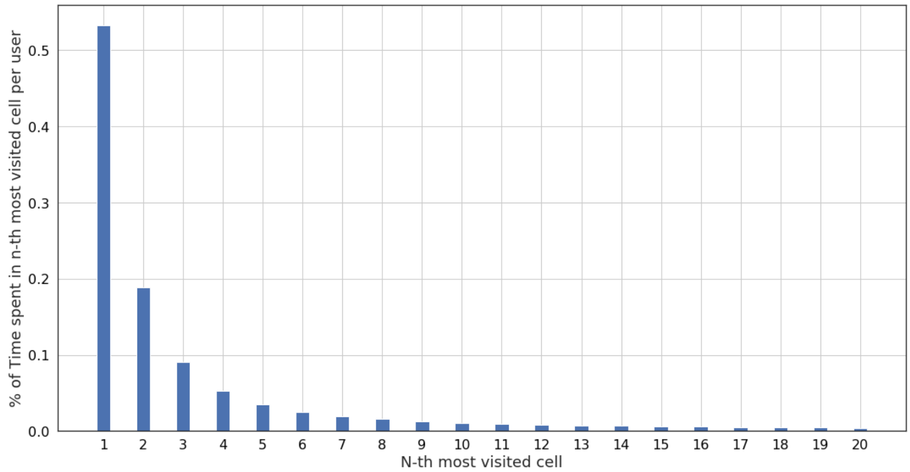

Figure 3.4: Average relative time spent in the 20 most visited cells for each user.

Therefore we only used the home cell of each user to join the tables, keeping all the measurements of the user's visit days in order to have a more detailed insight into customer-perceived network quality, and hereby creating 18 Features.

Hence for sake of dataset consistency, i.e. to have only one entry per user and to extract the most extensive information, we computed the *Minimum, Maximum and Average value* on all cell visits per user and for each $i-th$ KPI f_i as:

$$f_n^{\min} = \min_{i=d-n}^{d} f_i \tag{3.1}$$

$$f_n^{\max} = \max_{i=d-n}^{d} f_i \tag{3.2}$$

$$f_n^{\text{mean}} = \frac{\sum_{i=d-n}^{d} f_i}{m} \tag{3.3}$$

where n is the assumed user memory, d is the response date and m is number of samples in observation period (24 samples per day). Computing these triplets for each KPI results in a total number of 13 x 3 = 39 features per user. Moreover, for the same reason as above and since we suppose that a customer's overall satisfaction strongly depends on the fraction of time he spends in full, limited and no service as well as network performances experienced, we also calculate for each user the *Cumulative Full Service Time Ratio*, F_n as:

$$F_n = \frac{\sum_{i=d-n}^{d} f_i}{\sum_{i=d-n}^{d} f_i + l_i + n_i} \tag{3.4}$$

Similarly, we compute the *Cumulative Limited Service Time Ratio* (L_n and the *Cumulative No Service Time Ratio* (Z_n) by adjusting the numerator in (3.4). Since being in a linear relation, the sum of the three value is equal to 1, and then we consider only 2 of them in the same model. Adding these features, along with full, limited and no service times, we obtain a total of 39 + 5 = 44 features.

Dataset	**Column**	**Description**
LTE Network Measurement Dataset	Data	Date and Hour of measurements
	CELL ID	Unique Cell Identifier
	RRC Setup SR	RRC Connection Estalishment Success Rate
	RRC Connection Reestablishment SR	RRC Reconnection Estalishment Success Rate
	Intra Frequency HO out SR	Intra Frequency Handover out Success Rate
	Inter Frequency HO out SR	Inter Frequency Handover out Success Rate
	HO In SR	Handover In Success Rate
	Inter RAT HO out SR E2W(LTE to WCDMA)	Inter Radio Access Technology Handovers Success Rate
	Inter RAT HO out SR E2G(LTE to GSM)	
	DownLink Volume	Downlink and Uplink Volumes
	UpLink Volume	
	Max DownLink Throughput	Maximum Downlink and Uplink Throughput
	Max UpLink Throughput	
	Average User Number	Average number of connected users
	Maximum User Number	Maximum number of connected users
	Cell Location Name	Cell Location Name
	Cell Longitude	Exact Cells Geographical position
	Cell Latitude	

Table 3.1: Columns and description of LTE Network measurement dataset

Dataset	Column	Description
User Satisfaction Dataset	MSISDN	Mobile Station International Subscriber Directory Number
	Response Date	Date on which the user Responded to the survey
	Response Recommendation	Reported Likelihood of Recommending the network operator
	Response QoE Network Data Speed	Reported Quality of Experience Regarding Network Data speed
	Response QoE Network Video Quality	Reported Quality of Experience Regarding Video Quality provided
	Response QoE Network Coverage	Reported Quality of Experience Regarding Network Coverage
	Visit Date	Date of customers visit on Cell ID
	CELL ID	Unique Cell Identifier
	Full Service Time	Time spent with Full Service on Visit Date
	Limited Service Time	Time spent with Limited Service on Visit Date
	No Service Time	Time spent with No Service on Visit Date

Table 3.2: Columns and description of User satisfaction dataset

Chapter 4

Data Analysis

This chapter outlines the datasets analysis tasks conducted in this work. In particular, the engineered features are studied in this chapter through some preliminary analyses aimed at understanding the informative power they will provide to the QoE prediction. The reminder of the chapter is as follows: Section 4.1 provides with a descriptive analysis of the considered features, while Section 4.2 describes the process of satisfaction response binary discretization.

4.1 Data Transformation

In this work, we considered a supervised learning approach, i.e. the prediction algorithms are trained to estimate the (known) desired output (i.e., users satisfaction label) which corresponds to each training sample, exploiting the features included in the model (i.e., access network KPIs and users cell visit time). During the learning phase, the algorithms search for (and learn) patterns in the training data that correlate with the desired output. If the learning complete successfully, given the availability of new and never observed input samples, the prediction algorithm can output the corresponding satisfaction class. This mainly depends on the model that is adopted to perform prediction, i.e. to the set of selected input features. Regardless of the meaning of the features, it is crucial to check their statistical distributions, as the majority of ML methods assume that input features are distributed like a Gaussian random variable. Let us define σ_k as the *skewness* of the distribution of the $k-th$ feature f_k, i.e., the direction and relative magnitude of the distribution's deviation from the Gaussian distribution. In particular, $\sigma_k = 0$ if f_k is distributed as a Gaussian random variable. However, considering that most of the selected features were not distributed as a Gaussian (i.e. $\sigma_k \neq 1$ for most of the features), we implemented a log-like transformation in order to obtain the desired statistical characterization as it follows:

$$f_k^{\text{tr}} = \begin{cases} -\log(1 - f_k), & \text{if } \sigma_k > 1 \\ \log(f_k), & \text{if } \sigma_k < -1 \end{cases} \tag{4.1}$$

For the sake of clarity, we underline that $f_k^{\text{tr}} = f_k$ if $\sigma_k \in [-1, 1]$. As an example, we show in Figures 4.1 and 4.2 the distribution of Minimum RRC Setup Success Rate ($\sigma > 1$) and Minimum Downlink Volume ($\sigma < -1$) respectively, before (left) and after (right) the log-like transformation. As it can be seen, in both the cases raw distributions are not Gaussian, as most users either experience very high RRC setup success rates or very low minimum downloaded volumes. However, after the transformation, both the distributions look more like Gaussian bells.

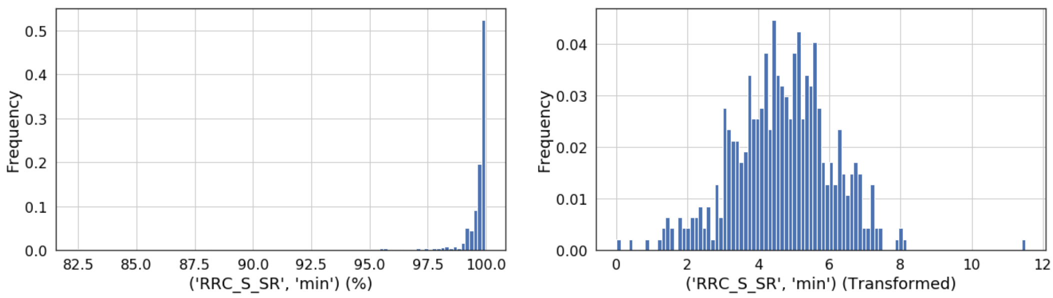

Figure 4.1: Minimum RRC Setup Success Rate distribution and transformed distribution

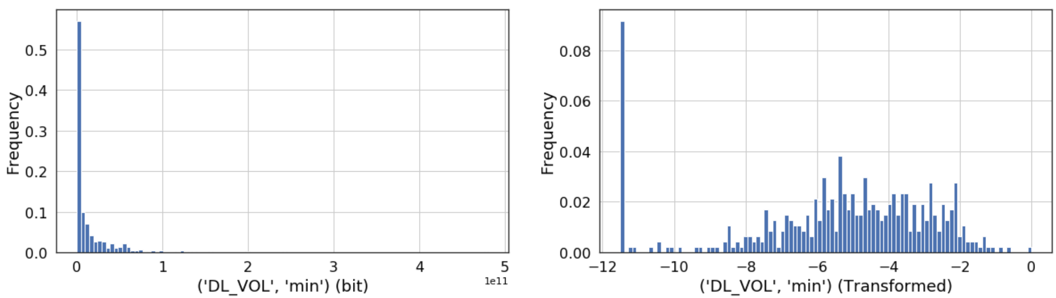

Figure 4.2: Minimum Downlink Volume distribution and transformed distribution

4.2 Satisfaction Threshold

Many machine learning algorithms are intrinsically developed for binary decision-making problems, and generally the majority perform better for binary prediction. For this reason, we choose to consider a binary problem where each user can either be satisfied or dissatisfied according to a satisfaction threshold T, whose tuning is described in this Section. Considering that for a given satisfaction item each user expresses a vote between 0 and 10, on the one hand setting T to a value lower than 6 would consider as satisfied all those users reporting a grade equal to 6, which is too far from the maximum grade (10). On the other hand, assuming $T = 9$ we would include in the class of satisfied users all those reporting whether a 9 or a 10 only, which is a too strict approach. Therefore, for each of the three specific service items, we let T take values in the range $[6, 7, 8]$, and we analyzed the information that a prediction algorithm can extract from the considered features conditioned to the couple of considered satisfaction classes. Figures 4.3, 4.4, 4.5, 4.6, 4.7 and 4.8 show the class-conditional Cumulative Distribution Functions (CDFs) for each value of the threshold T and for all service items of two different features taken as example, namely minimum Inter frequency Handover out Success Rate and No service time ratio. In general, for a given feature, the wider the separation between the blue and red curves the more the information associated to that feature is conditioned to the observations class label. In other words, the observation of a feature's class-conditional CDFs is crucial for a preliminary features selection process, as when the gap between the conditioned curves is little then we expect that a classifier will extract little or no useful information from such feature to perform prediction.

As an example, referring to Figures 4.3 and 4.4, we observe that users satisfied about the network coverage (blue curves) have experienced (in their home cell) a lower minimum handover out rate of success and shorter No Service periods than dissatisfied users about the same service item.

Figures 4.5 and 4.6 shows that users satisfied about the video streaming quality (blue curves) have experienced (in their home cell) a lower minimum handover out rate of success and shorter No Service periods than dissatisfied users about the same service item, disregarding the satisfaction threshold. As a matter of fact, in Fig. 4.5 considering $T = 6$ it can be seen that almost 40% of satisfied users had a minimum handover out success rate greater than 98% (corresponding to $SR^{\text{tr}} = 2$, while this is true for only 15% of unsatisfied users. Similarly, in Fig. 4.6 for $T = 6$ we can see that about only 2% of satisfied users had a No time service ratio greater than 0.7%(corresponding to $NSTR^{\text{tr}} = 5$, while it is true for 5% of unsatisfied ones. Although such differences are minimal, when used in combination, they are very informative for the classifiers, enhancing their prediction performance.

As can be seen by observing the class-conditional CDFs, when the threshold T changes, different features have a wider separation. This is why we decided to analyze the QoE prediction cases for $T = 6$ and $T = 8$, also observing that for $T = 7$ there are no significant variations from the case $T = 6$. Moreover, for each service item, there is a T which on average maximizes the amount of information that can be extracted from the features.

Usually, the T threshold value is fixed a priori, but in this work, we wanted to take into account the two values that respectively maximize the QoE prediction performance of the different service items.

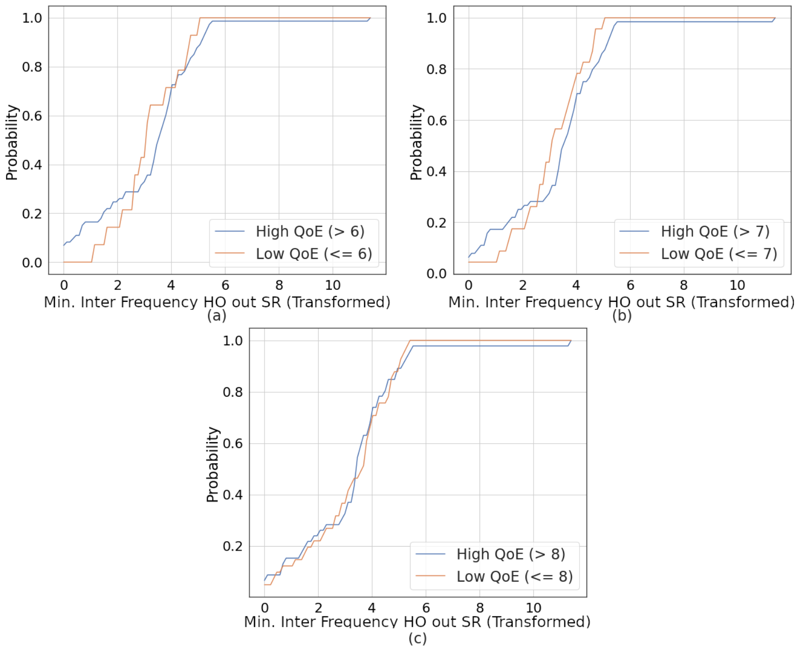

Figure 4.3: Network Coverage relative minimum Inter frequency Handover out Success Rate Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8

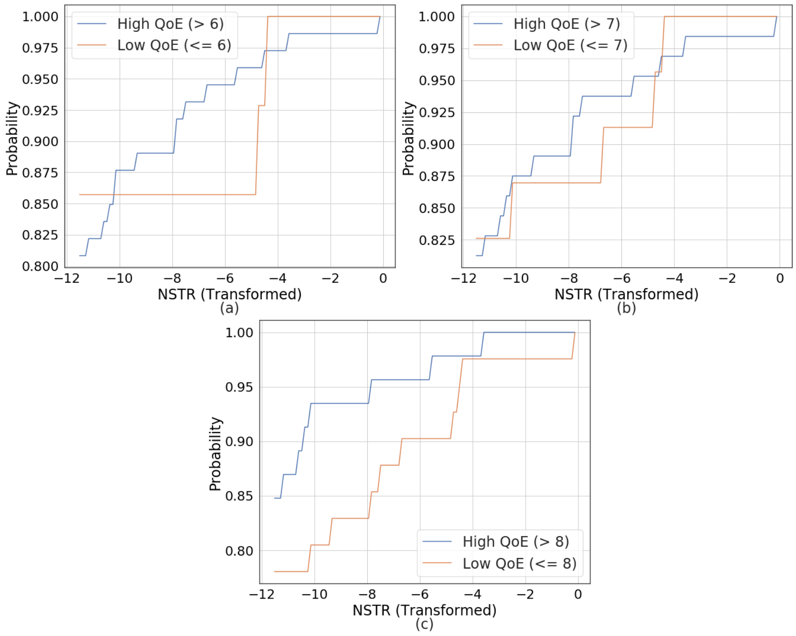

Figure 4.4: Network Coverage relative No Service Time Ratio Class-conditional CDF at threshold T = (a)6,(b)7,(c)8

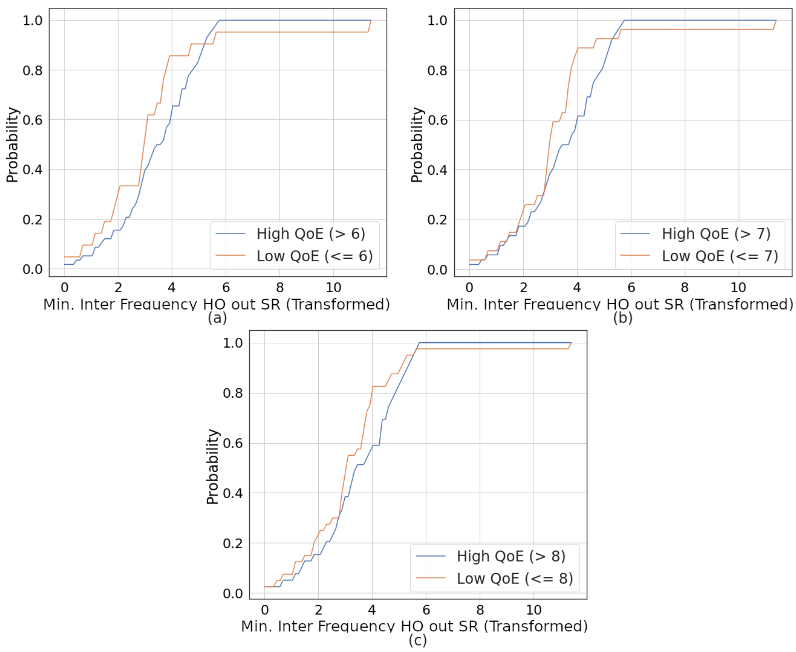

Figure 4.5: Video Quality relative minimum Inter frequency Handover out Success Rate Class-conditional CDF at threshold T = (a)6,(b)7,(c)8

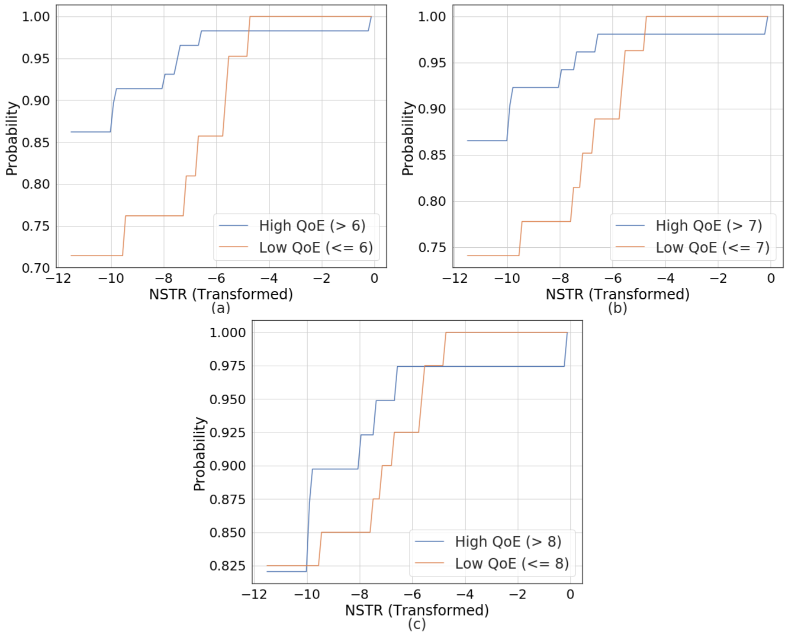

Figure 4.6: Video Quality relative No Service Time Ratio Class-conditional CDF at threshold $T = $ (a)6,(b)7,(c)8

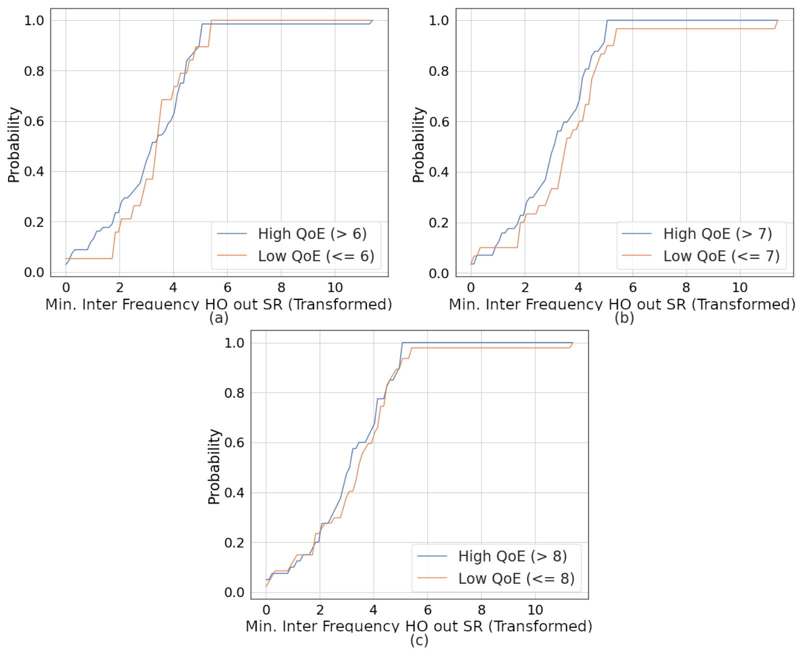

Figure 4.7: Data Speed relative minimum Inter frequency Handover out Success Rate Class-conditional CDF at threshold $T = $ (a)6,(b)7,(c)8

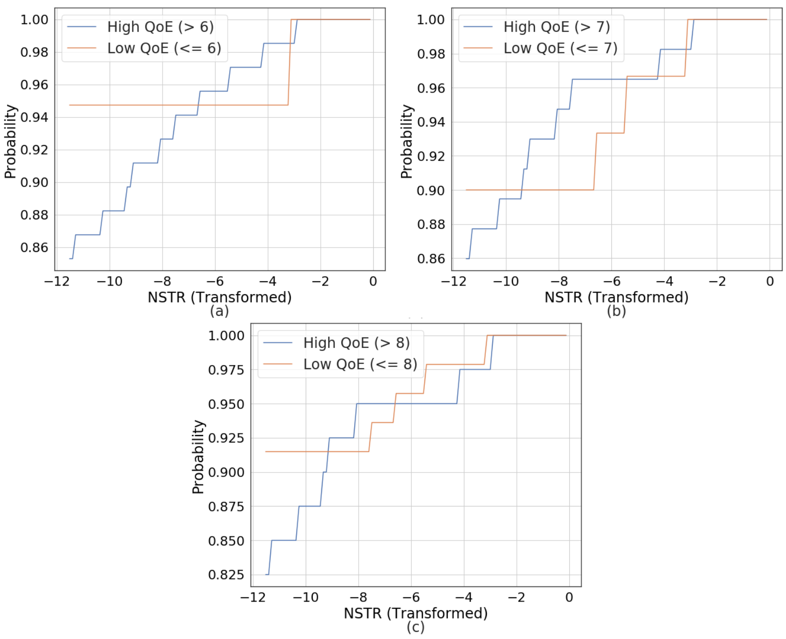

Figure 4.8: Data Speed relative No Service Time Ratio Class-conditional CDF at threshold $T =$ (a)6,(b)7,(c)8

Chapter 5

Prediction and results

In this chapter, we describe the techniques used to perform prediction of cellular users satisfaction regarding the considered service items and we show the obtained results. First, in section 5.1 we explain in detail the construction of the prediction pipeline adopted in this work. Secondly, in section 5.1.1 the techniques used to perform features selection are shown. Then, section 5.2 explains the driving factors considered to choose the most suitable performance metric for the different use cases while, in Section 5.3 we show the prediction results and compare the performance of several machine learning models. Finally, in the Section 5.4 we made a comparison between this thesis work and a previous similar research, showing our contribution.

5.1 Prediction Pipeline

In this section, the methodology and techniques used for building a predictive model of customer satisfaction related to a specific service is presented. As discussed in Chapter 4, it is possible to discriminate the subsets of features that are most informative about users satisfaction concerning the different specific network services. The selected features are used in this work as input to feed the following supervised machine learning models: i)Regularized Logistic Regressor (RLR), ii) Gaussian Naive Bayes (GNB), iii) Linear Discriminant Analysis(LDA), iv) Support Vector Classifier(SVC), v) Decision Trees (DT), vi) Random Forest (RF), vii) eXtreme Gradient Boosting (XGB). All classifiers except GNB and LDA need different hyper-parameters whose tuning is not trivial and need optimization. For instance, Regularized Linear Regression and Support Vector Classifier require the inverse regularization factor C (Equation 5.1) to reduce the generalization error at fixed training errors.

$$C = \frac{1}{\alpha} \tag{5.1}$$

In other words, the regularization is used to let a classifier better generalize on unseen data, preventing the algorithm from overfitting the training samples. In this thesis, we choose a *Lasso* (also called L2) regularization that also works as a features selector, penalizing features that increase the generalization error. Similarly, tree-based models (DT, RF) require to setup hyper-parameters such as trees' maximum depth and splitting criteria. In order to tune each classifier's hyper-parameters, we perform a *Grid Search* over predefined hyper-parameters values using a k-Fold cross-validation strategy with $k = 10$ as shown in Figure 5.1. Each filtered dataset is first divided into 10 folds with splitting ratios 90% (Training Set) and 10% (Test Set). Secondly, focusing on the Training set and using a further 10-Fold cross-validation, we split it into a sub-training set (80% of the overall dataset) and a Validation Set (10% of the overall dataset), to tune algorithms' hyper-parameters. In particular, a grid search is performed within the inner cross-validation loop to select for each classifier the best hyper-parameters (i.e., those maximizing the classifier performances on validation set). Table 5.1 shows the hyper-parameters values that are finally used to train the prediction models in the outer cross-validation loop. Then, each classifier's prediction performance are evaluated on the corresponding test set for each fold and finally results are averaged among all the folds. In the following Section we detail the algorithm that is used within the inner cross-validation loop to perform the features selection process.

5.1.1 Feature Selection

Feature selection is a core process of machine learning and aims at selecting those features that best contribute to the prediction process. The main reason for performing features selection is to exclude irrelevant features from the prediction pipeline, as they would decrease the performance of the prediction model. A further reason for excluding irrelevant features from the model is to to avoid the *curse* of dimensionality, to shorten the computation times and improve the model generalization capability. In this work is used a backward, wrapper-type feature elimination technique, called Recursive Feature Elimination (RFE). Wrapper-type selection algorithms use a predictive model to select feature subsets, in contrast to filter-based feature selections that give a score to each feature and select those features with the highest (or lowest) score. In particular, RFE works by searching for the subset of the features set that maximizes the performance on the validation set, removing features until the best subset is recognized. First, the considered classifier is trained on the sub-training set. Secondly, the features are ranked by importance and less important features are

Model	Description	Tuned Value
Regularized Logistic Regression	Regularization factor (C = {0.1:10:1e3})	1
Support Vector Classificator	Regularization factor (C = {0.1:10:1e3})	1
Decision Tree	Splitting criterion ({'gini','entropy'})	gini
	Splitting strategy ({'best','random'})	best
Random Forest	Number of estimators ({100:200:1000})	100
eXtreme Gradient Boosting	Maximum depth of tree ({3,4,5})	4
	Minimum loss reduction required to make a further partition ({0.5,1,1.5,2,5})	0.5
	Minimum sum of instance weight in a child ({1,5,10})	1
	Subsample ratio of the training instances. ({0.6,0.8,1})	0.8
	Subsample ratio of columns when constructing each tree ({0.6,0.8,1})	0.8

Table 5.1: Video Quality relative tuned hyper parameters

discarded. Finally, the model is re-fitted on the sub-training set. This process is repeated until the optimal number of features is reached. The number of desired features is a hyper-parameter of the RFE algorithm, whose setup is not trivial. Therefore, through an exhaustive search, a specific subset of features has been selected for each service item and each relative threshold, as shown in Table 5.2. In the following Section, we introduce the metrics used to evaluate the performance of the considered classifiers.

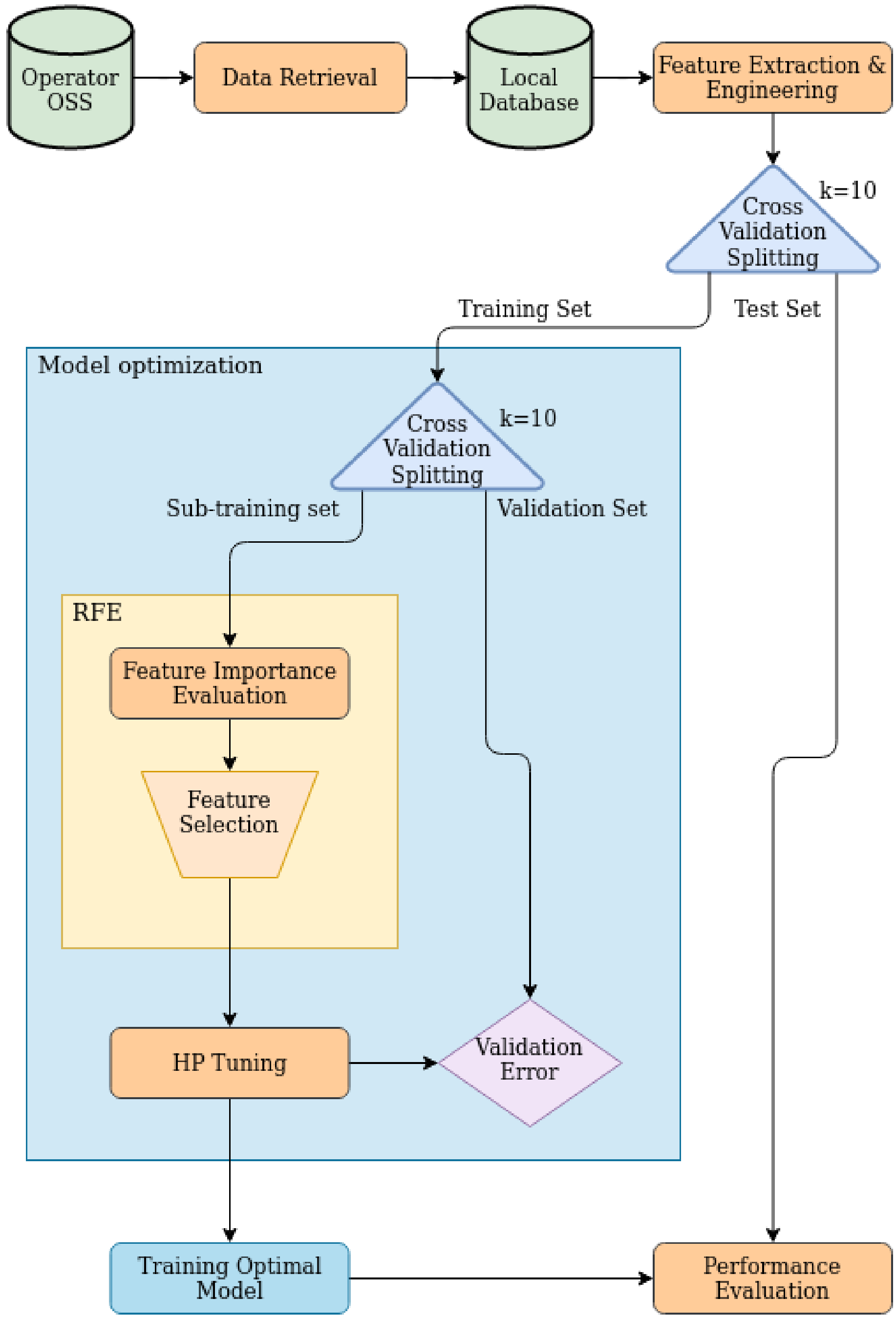

Figure 5.1: Data flow and operations in the pipeline. The raw dataset is divided into train set and test set, the train set is used to identify the optimal model (hyper parameter tuning and RFE) and to train it, finally the test set is used to evaluate the prediction performance.

Service Item	Video		Speed		Coverage	
Threshold	6	8	6	8	6	8
Feature						
Average DownLink Volume	Yes	Yes	Yes	Yes	Yes	No
Average HO In SR	Yes	Yes	Yes	No	Yes	Yes
Average Inter Frequency HO out SR	No	Yes	Yes	Yes	Yes	Yes
Average Inter RAT HO out SR E2G(LTE to GSM)	Yes	Yes	No	No	No	Yes
Average Intra Frequency HO out SR	No	Yes	Yes	Yes	Yes	Yes
Average RRC Connection Reestablishment SR	No	Yes	No	No	No	Yes
Average RRC Setup SR	No	No	Yes	No	Yes	Yes
Average UpLink Volume	Yes	Yes	Yes	Yes	Yes	Yes
Average of Max DownLink Throughput	Yes	Yes	Yes	Yes	Yes	Yes
Average of Max UpLink Throughput	Yes	Yes	Yes	Yes	Yes	Yes
Average of Maximum User Number	Yes	Yes	Yes	Yes	Yes	Yes
Full Service Time	No	Yes	No	No	Yes	Yes
Limited Service Time	Yes	Yes	Yes	Yes	Yes	Yes
Limited Service Time Ratio	Yes	Yes	Yes	Yes	Yes	Yes
Maximum DownLink Volume	Yes	Yes	Yes	Yes	Yes	Yes
Maximum HO In SR	Yes	Yes	Yes	No	Yes	Yes
Maximum Inter Frequency HO out SR	No	Yes	Yes	Yes	No	Yes
Maximum Inter RAT HO out SR E2G(LTE to GSM)	Yes	Yes	Yes	No	No	No
Maximum Intra Frequency HO out SR	No	Yes	No	Yes	No	Yes
Maximum RRC Connection Reestablishment SR	No	Yes	No	Yes	No	Yes
Maximum RRC Setup SR	No	Yes	Yes	Yes	No	Yes
Maximum UpLink Volume	No	Yes	Yes	Yes	Yes	Yes
Maximum of Max DownLink Throughput	Yes	Yes	Yes	Yes	Yes	Yes
Maximum of Max UpLink Throughput	Yes	Yes	Yes	Yes	Yes	Yes
Maximum of Maximum User Number	No	Yes	Yes	Yes	Yes	Yes
Minimum DownLink Volume	No	Yes	Yes	Yes	Yes	No
Minimum HO In SR	No	Yes	Yes	No	Yes	Yes
Minimum Inter Frequency HO out SR	No	Yes	Yes	Yes	No	Yes
Minimum Inter RAT HO out SR E2G(LTE to GSM)	Yes	Yes	No	No	No	Yes
Minimum Intra Frequency HO out SR	No	Yes	Yes	Yes	Yes	Yes
Minimum RRC Connection Reestablishment SR	No	Yes	No	Yes	No	Yes
Minimum RRC Setup SR	No	Yes	No	No	Yes	No
Minimum UpLink Volume	Yes	Yes	Yes	Yes	Yes	Yes
Minimum of Max DownLink Throughput	No	Yes	Yes	Yes	Yes	Yes
Minimum of Max UpLink Throughput	No	Yes	Yes	Yes	Yes	Yes
Minimum of Maximum User Number	Yes	Yes	Yes	Yes	Yes	Yes
No Service Time	Yes	Yes	Yes	Yes	Yes	Yes
No Service Time Ratio	Yes	Yes	Yes	Yes	Yes	Yes

Table 5.2: Features selected for each Service Item and Threshold

5.2 Model evaluation

5.2.1 Evaluation metrics

Evaluating the prediction performance of a supervised classification method with respect to unseen data is crucial in machine learning. A nice and comprehensive way to show the performance of a machine learning model is through the *Confusion Matrix*. As shown in Figure 5.2, considering a binary classifier, the Confusion Matrix is characterized by 4 entries. On the one hand, the True Positives *TP* and True Negatives *TN* represent the correct classified examples, in detail they indicate respectively how many dissatisfied customers have been correctly detected and how many are recognized as satisfied. On the other hand, the False Positives *FP* and False Negatives *FN* report the cases of false classification: *FP* comprises cases in which a satisfied customer has been detected as unsatisfied, also called false alarms, while *FN* indicates the number of opposite cases, i.e. those in which an unsatisfied user has been identified as satisfied. From the knowledge of the Confusion Matrix, it is possible to compute three important performance metrics, namely i) Area under the ROC Curve (AUC), ii) *F1 Score* and iii) *Accuracy*, which are detailed in the following.

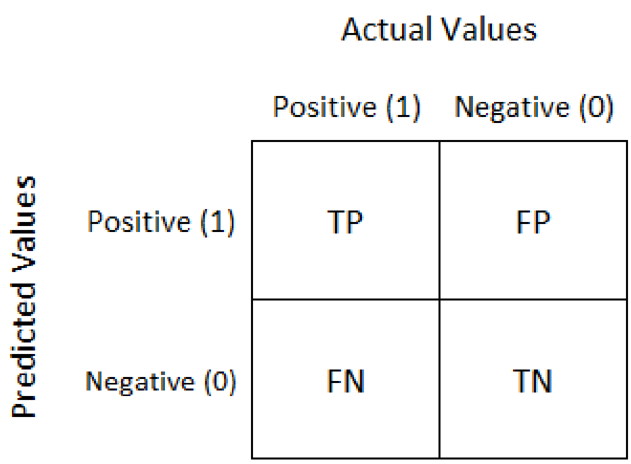

Figure 5.2: Confusion Matrix

AUC

AUC stands for *Area Under the Curve*, also referred to as *AUROC*, or *Area Under Receiver Operating Characteristic*. The *Receiver Operating Characteristic curve* (or ROC curve) studies the relationships between correctly recognized instances and false alarms, analyzing the *True Positive Rate* (*TPR*, that is the fraction of true positives) and the *False Positive Rate* (*FPR*, that is the fraction of false positives). *TPR* and *FPR* are calculated from the confusion matrix in Fig. 5.2:

$$TPR = \frac{TP}{FN + TP} \tag{5.2}$$

$$FPR = \frac{FP}{TN + FP} \tag{5.3}$$

The *ROC curve* is generated by plotting the distribution function relative to the probability of recognition (*TPR*) with respect to the false alarm probability distribution function (*FPR*): it is a curve that allows to analyze the binary result provided by models, in order to choose the best one.In relation to the *ROC curve*, the *AUC* corresponds to the area underneath that curve, and we use it to summarize. Note that, for a random guessing classifier, the *AUC* is 0.5.

F1 Score

The *F1 score*, also called the *F score* or *F measure*, is a measure of a classification accuracy. The *F1 score* is defined as the weighted harmonic mean of the classification precision and recall.
This score is calculated according to:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.4}$$

Where Precision and Recall are computed from the confusion matrix as:

$$Precision = \frac{TP}{TP + FP} \tag{5.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.6}$$

Combining these 3 equations it can be seen that the *F1 score* gives more weight to the true positives, making it suitable for our study (enhancing unsatisfaction accuracy) :

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{5.7}$$

Accuracy

The *Accuracy* is the proportion of correct prediction among the total number of instances examined and measures how well a classifier correctly identifies an unknown sample.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.8}$$

5.2.2 Driving factors in the choice of metrics

In section 5.1 we illustrated the method used to optimize the choice of the models hyper-parameters. In particular, for each classifier, we choose through a grid search cross-validated process the hyper-parameters values that maximize the prediction performance on the validation set according to a selected prediction metric. The choice of the reference prediction metric to use in this process is not trivial and depends on the numerosity of the prediction classes. On the one hand, we noticed that when the binarization threshold T equals 6, the satisfaction classes are strongly unbalanced, such that 25% of users are labeled as Unsatisfied and 75% as Satisfied. On the other hand, satisfaction classes turn to be completely balanced (i.e. 50%-50%) when T is equal to 8. Considering that one of the major problems in machine learning relates to unbalanced prediction classes, we need to rely on evaluation metrics that take this issue into account. For this reason, we chose Accuracy as the reference performance metric for balanced datasets whereas we chose F1 Score as the reference metric when prediction classes are unbalanced. Note that AUC will be used for both cases as a further way to evaluate the classifier's prediction performance.

5.3 Model comparison

In this Section we evaluate the prediction performance of the introduced classifiers. Considering that network measurements are available for the period between November 2019 and April 2020 while ground truth satisfaction data refer to the period between August and November 2019, the intersection of the two data sources described in Chapter 3 outputs a dataset comprised of 253 observations. Each observation corresponds to a customer who has voted on one of the three considered service items, which are video quality (79 users), network data speed (87 users) and network coverage (87 users). Such three datasets are used to train the different classifiers, whose performance with respect to each service item is finally assessed according to the metrics defined in previous Section. Finally, in order to understand the effective capability of each classifier to perform correct prediction, we will compare the prediction performance with the (baseline) performance of a dumb classifier forced to label each observation as belonging to the majority class (i.e., to the class of the satisfied users). Table 5.3 summarises for T equal to 6 and 8 the baseline performance with respect to the considered evaluation metrics.

Threshold	AUC	F1 Score	Accuracy
6	0.50	0.40	0.75
8	0.50	0.66	0.50

Table 5.3: Baseline performance per Threshold T

Threshold	Model	F1	ACC	AUC
6	Decision Tree	0.272	0.571	0.497
	Gaussian NB	0.231	0.536	0.420
	LDA	**0.74**	**0.873**	**0.797**
	Logistic Regression	0.437	0.784	0.760
	Random Forest	0.197	0.721	0.667
	XGBoost	0.263	0.723	0.690
8	**Decision Tree**	**0.664**	**0.657**	**0.650**
	Gaussian NB	0.240	0.507	0.518
	LDA	0.532	0.586	0.552
	Logistic Regression	0.474	0.520	0.581
	Random Forest	0.525	0.520	0.597
	XGBoost	0.474	0.495	0.506

Table 5.4: Performance obtained for video streaming QoE prediction

Satisfaction with network video quality

Predicting satisfaction with the video streaming service is paramount as it constitutes the majority of mobile traffic, particularly in recent years. According to the feature selection process, the prediction model for this service item is composed by 19 features when $T = 6$ and 38 features when $T = 8$, as shown in first two columns of Table 5.2. Figures 5.3 and 5.4 show the average ROC curves of the considered classifiers while Table 5.4 summarises the classifier's average performance of each model for the two binarization thresholds. As reported in the Table, the best performance for T equal to 6 and 8 is achieved by LDA and DT, with AUC values of 0.797 and 0.65 respectively. We also report in Table 5.5 the performance improvement yielded by LDA and DT with respect to the baseline performance for the three evaluation metrics. Recalling that when $T = 6$ the reference metric is the F1 score whereas the Accuracy is considered as reference metric for $T = 8$, we observe that LDA and DT yields 34% better F1 Score and 16% better Accuracy than the baseline respectively.

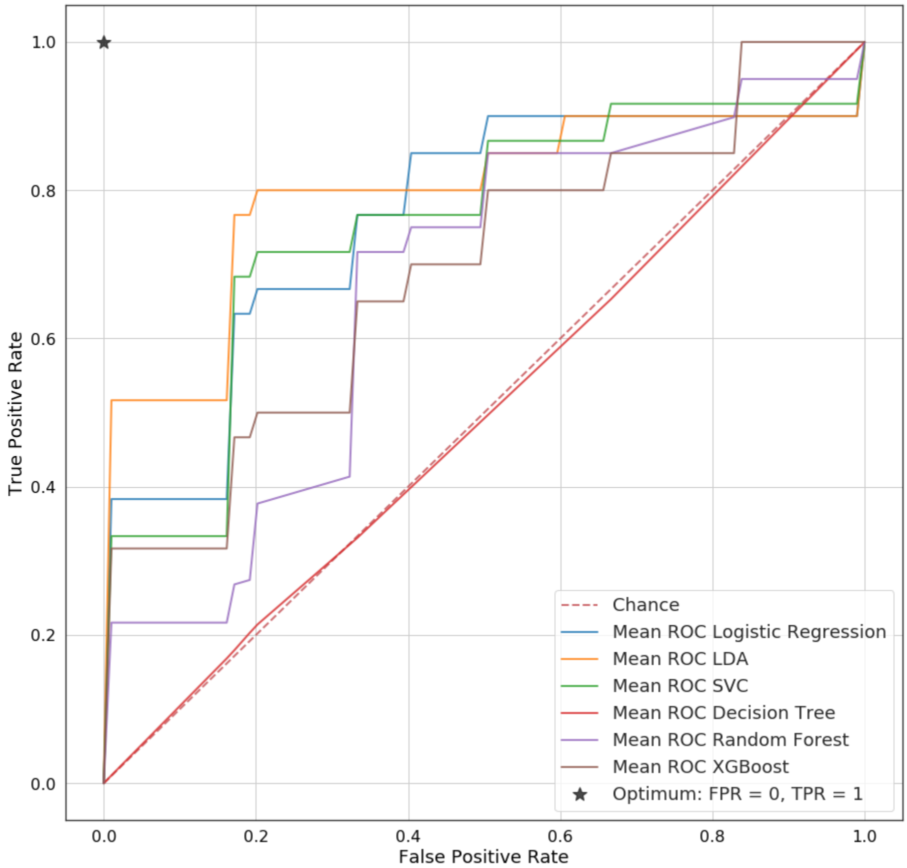

Figure 5.3: ROC curve for video streaming QoE prediction, $T = 6$

Model	Threshold	F1	ACC	AUC
LDA	6	+34%	+12.3%	+29.7%
Decision Tree	8	+0.4%	+15.7%	+15%

Table 5.5: Improvements from baseline obtained for video quality QoE prediction

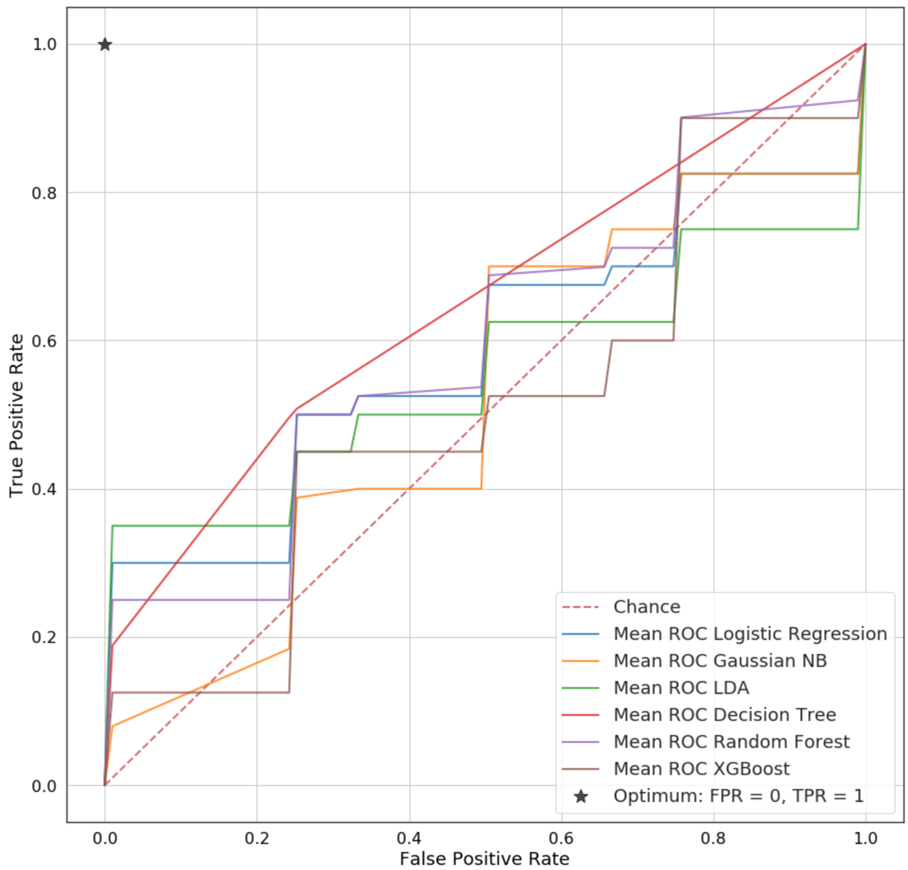

Figure 5.4: ROC curve for video streaming QoE prediction, $T = 8$

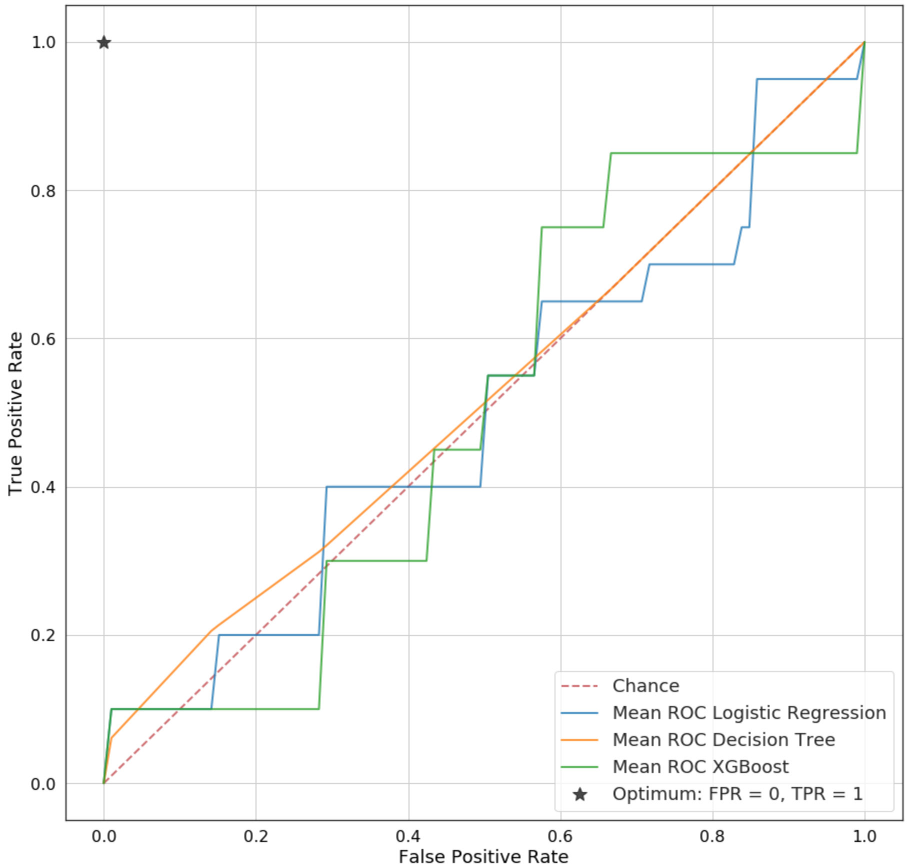

Figure 5.5: ROC curve for data speed QoE prediction, $T = 6$

Satisfaction with network data speed

Besides network video quality, we also focus on predicting user satisfaction related to QoE with network data speed. In this case, the RFE algorithm selected 30 features when $T = 6$ and 28 features when $T = 8$, as shown in third and fourth column of Table 5.2. Similarly to the case of video quality, we show in Figures 5.5 and 5.6 the average ROC curves of the considered classifiers for the two T while Table 5.6 summarises the average prediction performance. For data speed service item, while classifiers perform at par and quite poorly for $T = 6$, we observe in 5.6 that RLR is the best performing classifier when $T = 8$, reporting an AUC value of 0.709. Moreover, as shown in Table 5.7, RLR improves the accuracy by 18% with respect to the baseline.

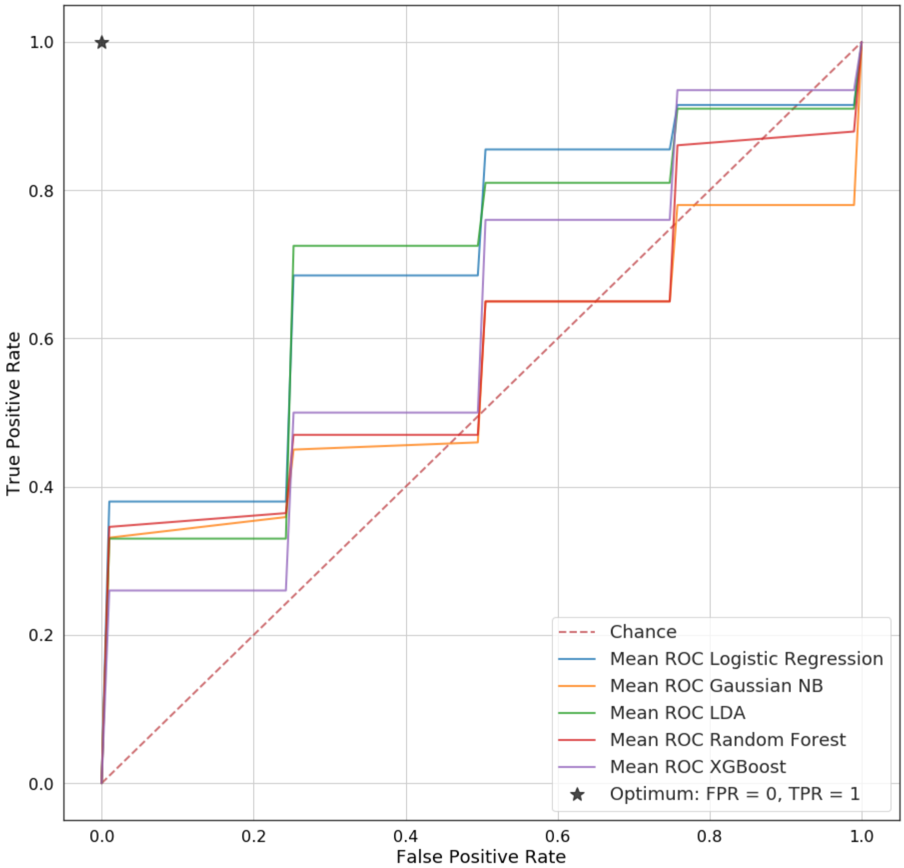

Figure 5.6: ROC curve for data speed QoE prediction, $T = 8$

Threshold	**Model**	**F1**	**ACC**	**AUC**
6	Decision Tree	0.235	0.679	0.520
	LDA	0.140	0.664	0.430
	Logistic Regression	0.090	0.712	0.498
	XGBoost	0.133	0.746	0.498
8	Decision Tree	0.467	0.474	0.472
	LDA	0.594	0.621	0.694
	Logistic Regression	**0.691**	**0.679**	**0.709**
	XGBoost	0.616	0.562	0.614

Table 5.6: Performance obtained for data speed QoE prediction

Model	**Threshold**	**F1**	**ACC**	**AUC**
Logistic Regression	8	+0.31%	+17.9%	+20.9%

Table 5.7: Improvements from baseline obtained for data speed QoE prediction

Threshold	**Model**	**F1**	**ACC**	**AUC**
6	Decision Tree	0.273	0.710	0.546
	SVC	0.100	0.475	0.461
8	**Decision Tree**	**0.522**	**0.514**	**0.512**
	SVC	0.298	0.432	0.584

Table 5.8: Performance obtained for coverage QoE prediction

Model	**Threshold**	**F1**	**ACC**	**AUC**
Decision Tree	8	No impr.	+1.4%	+0.12%

Table 5.9: Improvements from baseline obtained for coverage QoE prediction

Satisfaction with network coverage

Finally, we focus on predicting user satisfaction related to QoE with network coverage. For this use case, the feature selection algorithm selected 28 features when $T = 6$ and 3 features 4 for $T = 8$, as reported in the last two columns of Table 5.2. Similarly to the previous use cases, we plot in Figures 5.7 and 5.8 the average ROC curves of the considered classifiers while Table 5.8 shows the corresponding average prediction performance of each model for the two binarization thresholds. As one can observe from both the ROC curves, all the classifiers perform likely the baseline for both the binarization thresholds, yielding average AUC close to 0.5. While for T=6 no improvement is reported in terms of F1 score, we observe that DT yields 1.4% better Accuracy than the baseline when T=8, as reported in 5.9.

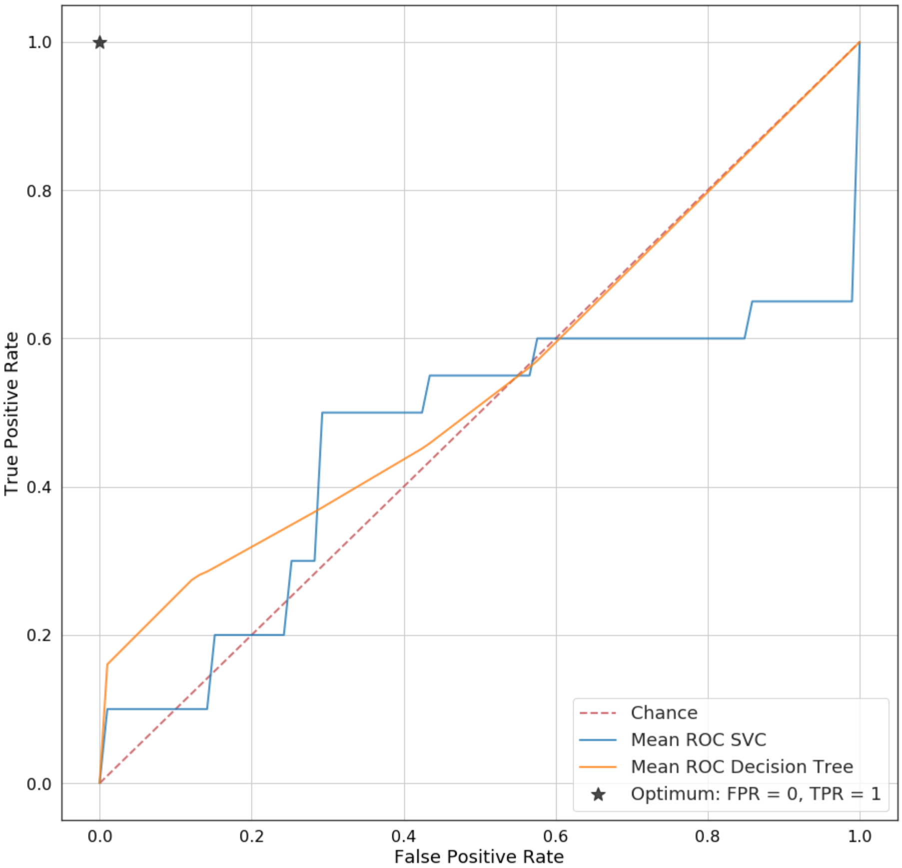

Figure 5.7: ROC curve for coverage QoE prediction, $T = 6$

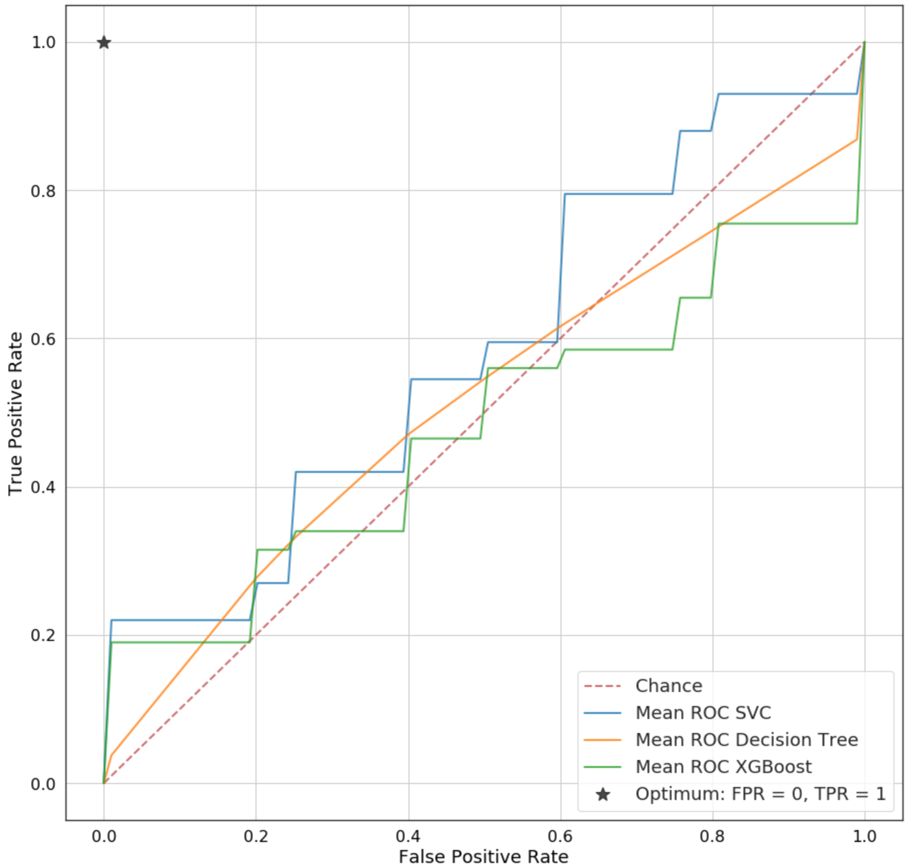

Figure 5.8: ROC curve for coverage QoE prediction, $T = 8$

	This Thesis				Previous Research			
Response	Model	AUC	F1 Score Benchmark	F1 Score Improvement	Model	AUC	F1 Score Benchmark	F1 Score Improvement
Video Quality	LDA	0.79	0.40	+34%	RF	0.58	0.37	+1.5%
Network Coverage	DT	0.51	0.66	No impr.	RLR	0.60	0.35	+3%

Table 5.10: Model comparisons with respect to [1]

5.4 Network-side vs. user-side based prediction models

In order to illustrate and explain the improvements made by our study, it is worth to make a comparison with the counterpart of the research, i.e. the user-side based prediction models proposed in [1]. Similarly, the authors in [1] leveraged both user-side activity measurements and ground-truth satisfaction feedbacks to train different Machine Learning models to predict users satisfaction. Differently from this thesis work, the authors i) focused on long-term satisfaction, i.e., the satisfaction relative to a period of time spanning several weeks and ii) did not consider network performance measured at the access of the cellular network (i.e. at base station side). Table 5.10 compares the result reached in this thesis work with those reported in [1]. On the one hand, considering video quality, our model outperforms that presented in [1], improving the AUC by 21% and the F1 Score by 32.5%. Interestingly, while in [1] best results for video quality use case were achieved for T=6 and for RF classifier, in this work we observe that best performance are achieved by LDA classifier. On the other hand, regarding network coverage service item, our model does not improve the prediction performance. This probably means that access network data (e.g., success rates of handovers, number of connected users, volume of network traffic, etc. referred to the home cell) turn to be more informative to predict users satisfaction on video quality, whereas users satisfaction about network coverage can be better inferred by user-side measurements (e.g., users relative Full Service time, Signal to Noise Ratio (SNR), geographic location of the users, etc.). Moreover, authors in [1] point out that predicting long-term satisfaction is much more challenging than estimating short-term QoE. It suggests that users' memory is conditioned by adaptivity to QoS, i.e. users in a long period adapt to network QoS, thus affecting the correct prediction of their satisfaction. To conclude, as observed in [1] and shown in [8], including commercial data from the Business Supporting System (BSS) to the prediction model leads to an improvement of classification performance.

Chapter 6

Concluding remarks

In this thesis, a methodology to find significant predictors of user satisfaction in 4G cellular network is proposed. This methodology aims at identifying the level of satisfaction of mobile customers, thus detecting potential churners, using information related to the performance of the mobile network. Since we forecast the customers QoE using only the measurements related to their home cells, our methodology can also be employed to find under-performing cells, allowing the operator to perform upgrades or maintenance on those cells. These measurements were collected by the operator's OSS leveraging previous works on customer QoE relation with 4G network performances. In addition to the described objective measures of the network performance (KPI), the prediction process also leverages subjective information regarding user satisfaction (Survey responses), provided voluntarily by users that are requested by the network operator as concerns the quality of their experience in the network through targeted surveys.

6.1 Limitations and issues

We note the following limitations in this work. First of all, we had available a small set of samples to train the classifiers due to practical and privacy-related issues which limited the MNO in sharing BSS data. Therefore, we need further experiments to test the effective generalization capabilities of the prediction models, considering larger customers populations. Moreover, it is widely accepted in literature [8] that bigger data volumes can improve the performance of supervised machine learning models. Secondly, due to same reasons, we could neither investigate long-term satisfaction prediction nor analyse the relationship between prediction performance and the length of the users activity observation window (which in this work is fixed to 6 days). Due to such limited observation window, it is not clear whether the prediction model recognize long-term temporal phenomena such as users adaptation effects [13]. Additionally, the dataset does not include commercial-related features, which typically improve the classification performance [8].

6.2 Analysis of Results and Future Works

Despite the limitations that characterize this work, our results led to the following conclusions:

- Our methodology of feature engineering of LTE network measurements, described in Chapter 3, is effective in creating an informative set of features that enhances the detection of the End Users Quality of Experience although they describe the only network side.

- Different measures related to the specific home cell chosen have a high informative value in predicting user satisfaction. As a matter of fact, the measures concerning Inter-RAT handover out (LTE to GSM) and Handover IN are selected for Video Streaming, indicating that when the handover procedure fails the user experiences a block in the streaming and consequent buffering, resulting in a bad quality of service and therefore bad QoE. Similarly, the selection of features about the maximum number of users and the traffic volume highlights that a congested or a busy cell with many connected users is related to the poor quality of the users' experience, causing buffering due to saturation of network resources.

- It is possible to observe how the Intra and Inter Frequency handover out, the successful Setup and Re-connection Rate to the RRC and those related to the Downlink Volume directly impact the Quality of the experience related to the network speed.

- The "user-side" measures used are very effective in improving user satisfaction prediction performance. In fact, for all models the feature selection technique select the Limited and No Service Time Ratio, suggesting that measures "closer" to the user side provide useful information for the satisfaction prediction process, as already highlighted in other researches ([1]).

- Machine Learning algorithms are efficient in predicting user satisfaction regarding the various service items when features marginal distributions and prediction performances are studied by varying their binarization threshold of satisfaction labels. In fact, for different service items, we found better performance for different binarization thresholds .

Future research activities will focus on the use of larger dataset with longer observation periods (long-term prediction) to examine whether the users' memory has a significant impact on the evaluation using network-side data. Moreover, future works will investigate the use of users' n-th ($n > 1$) most visited cells, such to define a model that covers a larger part of the total users connection time. Finally, as suggested in [1], it is worth to observe that the prediction of users satisfaction can be used to recognize network areas or elements that mostly caused users dissatisfaction. Considering that our model predicts users satisfaction analysing their activity in the home cells, i.e. in each user's most visited cell, it can be used to identify under-performing or malfunctioning network cells. The impact of individual user's prediction errors on the under-performing network elements detection task when predicted feedbacks are grouped on single network elements has not yet been investigated.

Bibliography

[1] A. Pimpinella, A. E. C. Redondi, I. Galimberti, F. Foglia, and L. Venturini. Towards long-term coverage and video users satisfaction prediction in cellular networks. In *2019 12th IFIP Wireless and Mobile Networking Conference (WMNC)*, pages 146–153, Sep. 2019.

[2] Cisco. Cisco annual internet report (20182023). Technical report, 2020.

[3] Katz-Bassett E. Mao Z.M. Welsh M. Nikravesh A., Choffnes D.R. Mobile network performance from user devices: A longitudinal, multidimensional analysis. *International Conference on Passive and Active Network Measurement*, 8362:12–22, 2014.

[4] P. Casas, A. D'Alconzo, F. Wamser, M. Seufert, B. Gardlo, A. Schwind, P. Tran-Gia, and R. Schatz. Predicting qoe in cellular networks using machine learning and in-smartphone measurements. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, May 2017.

[5] P. Casas, M. Seufert, F. Wamser, B. Gardlo, A. Sackl, and R. Schatz. Next to you: Monitoring quality of experience in cellular networks from the end-devices. *IEEE Transactions on Network and Service Management*, 13(2):181–196, June 2016.

[6] F. Kobayashi, G. Kawaguti, H. Nojiri, and A. Takahashi. Crowdsourced qoe evaluation of mobile device usage with consideration on user dependence. In *2015 17th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pages 388–391, Aug 2015.

[7] B. Finley, E. Boz, K. Kilkki, J. Manner, A. Oulasvirta, and H. Hämmäinen. Does network quality matter? a field study of mobile user satisfaction. *Pervasive and Mobile Computing*, 39:80–99, 2017.

[8] Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, and Jia Zeng. Telco churn prediction

with big data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 607–618, New York, NY, USA, 2015. Association for Computing Machinery.

[9] European Telecommunications Standards Institute. Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions (3GPP TS 32.450 version 15.2.0 Release 15). Technical report, ETSI, October 2019.

[10] European Telecommunications Standards Institute. Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Requirements (3GPP TS 32.451 version 15.1.0 Release 15). Technical report, ETSI, April 2019.

[11] F. Krasniqi, A. Maraj, and E. Blaka. Performance analysis of mobile 4g/lte networks. In *2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)*, pages 1–5, 2018.

[12] X. Liu, G. Chuai, W. Gao, K. Zhang, and X. Chen. Kqis-driven qoe anomaly detection and root cause analysis in cellular networks. In *2019 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6, 2019.

[13] Benjamin Weiss, Dennis Guse, Sebastian Möller, Alexander Raake, Adam Borowiak, and Ulrich Reiter. *Temporal Development of Quality of Experience*, pages 133–147. 03 2014.

[14] Florian Wamser, Michael Seufert, Pedro Casas, Ralf Irmer, Phuoc Tran-Gia, and Raimund Schatz. Poster: Understanding youtube qoe in cellular networks with yomoapp: A qoe monitoring tool for youtube mobile. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, pages 263–265, New York, NY, USA, 2015. Association for Computing Machinery.

[15] Pedro Casas, Michael Seufert, and Raimund Schatz. Youqmon: A system for on-line monitoring of youtube qoe in operational 3g networks. *SIGMETRICS Perform. Eval. Rev.*, 41(2):44–46, August 2013.

[16] ITU-T. Terms and definitions related to quality of service and network performance including dependability, sept 2008.