

**NOVA**

**IMS**

Information  
Management  
School

# MAAA

---

**Mestrado em Métodos Analíticos Avançados**

Master Program in Advanced Analytics

**CDR-based location analytics**

**&**

**Gender prediction from subscribers'**

**list of installed mobile applications**

*Double Projects Internship Report*

Dahmane Sheikh

Internship report presented as partial requirement for obtaining  
the Master's degree in Data Science & Advanced Analytics

NOVA Information Management School  
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

**NOVA Information Management School**  
**Instituto Superior de Estatística e Gestão de Informação**  
Universidade Nova de Lisboa

**CDR-BASED LOCATION ANALYTICS**  
**&**  
**GENDER PREDICTION FROM SUBSCRIBERS' LIST OF**  
**INSTALLED MOBILE APPLICATIONS**

by

Dahmane Sheikh

Internship report presented as partial requirement for obtaining the Master's degree in Data Science & Advanced Analytics

**Advisor: Mauro Castelli**

January 2020

# Acknowledgements

Foremost, I would like to express my sincere gratitude to Carlos Santos, head of Big Data and Advanced Analytics at Vodafone Portugal for giving me the opportunity to be part of his team for an internship. A very special gratitude to my supervisor at Vodafone, Fernando Goulart da Silva, for his invaluable guidance and help throughout my entire journey. To the rest of the team for their friendship and companionship.

I would also like to thank my wife María Arranz for all the care, encouragement and understanding, she provided during my entire academic path. To my parents and closest friends for their support, no matter the distance and time.

To my academic supervisor and teacher, Prof. Mauro Castelli, for all the help in reviewing this present internship report.

# Abstract

Big Data is big news in most industries, and telecommunication is no exception. Over the last decades, telecom operators experienced numerous changes in their business models, driven by technological innovations. Although, telecom operators have long had access to substantial bits of data, the scenario has radically evolved with the advent of smartphones, mobile broadband, rapid development of internet, growth of mobile services and Big Data Analytics capabilities (BDA). In today's data intensive world of communications, tremendous amount of diverse type of data are generated by telecom, bringing both challenges and opportunities to the table. This present internship report summarises my contribution part of the Big Data & Advanced Analytics team of Vodafone Portugal with two research projects; The first one consisted in studying human mobility from cellular network-based data, considering the so-called *Call Detail Records* (CDR) as a core proxy to extract spatiotemporal density distribution at finer geospatial granularity levels. The second consisted in conducting an observational study of the predictability of mobile subscribers' demographic traits from their installed mobile applications. The latter has the use-case of predicting the gender of mobile subscribers. Both research projects draw attention to the particular ubiquity aspect of connected mobile devices, being widely available and used all over the world.

## Keywords

Big data; Location analytics; Human mobility; Mobile phone location data; Call detail records; Demographics prediction; Gender; Mobile devices; Mobile applications; Machine learning.

# Resumo

A área de Big Data é uma grande novidade para a maioria das empresas, incluindo as companhias de telecomunicação. Durante as últimas décadas, e graças às inovações tecnológicas, os operadores de telecomunicações viveram muitas mudanças nos seus modelos de atividade comercial. Embora as empresas de telecomunicação já tinham acesso a uma quantidade considerável de dados (bits), o cenário mudou por completo com a chegada dos smartphones, a banda larga, o rápido desenvolvimento de internet, um grande crescimento dos serviços móveis e o Big Data Analytics Capabilities (BDCA). A frenética realidade atual do mundo das comunicações, cria uma grande e diversa quantidade de dados, gerada pelas empresas de telefonia, supondo ao mesmo tempo novos desafios e oportunidades. No seguinte relatório de estágio, resume-se a minha contribuição à equipa de Big Data e Advanced Analytics de Vodafone com dois projetos de investigação: O primeiro projeto consistiu em estudar a mobilidade dos humanos baseando-se nos dados extraídos da rede móvel, considerando o chamado Call Detail Records (CDR) como principal variável para poder obter informação mais detalhada sobre a densidade espaço-temporal em níveis de granularidade. O segundo projeto é um estudo observacional sobre a previsibilidade das características demográficas dos utentes tendo em conta as aplicações instaladas nos seus telemóveis. O caso prático deste último pretende prever o género dos clientes da rede móvel. Estes dois projetos de investigação pretendem chamar a atenção para a posição onipresente que ocupam os dispositivos móveis ligados à rede na nossa sociedade, estando disponíveis e sendo utilizados no mundo inteiro.

# Index

List of Figures .....	ix
List of Tables.....	xi
List of Acronyms.....	xii
<b>Chapter 1 Internship description.....</b>	<b>1</b>
1. Academic context.....	2
2. Internship context.....	2
3. Research projects.....	3
4. Internship report outline.....	5
5. Place of mobile phone in the modern era .....	6
6. Role of Big Data in Telecom companies .....	7
<b>Chapter 2 CDR-based location analytics .....</b>	<b>9</b>
Abstract .....	10
1. Introduction.....	11
2. Literature Review .....	12
3. Essential Telecom GIS concepts .....	14
3.1. Call Detail Records.....	14
3.2. Cell sites and Cell towers.....	16
3.3. Additional subscribers data .....	17
3.4. International Mobile Subscriber Identity - IMSI.....	18
3.5. Shapefile spatial data format .....	19
3.6. Unit association between people and mobile devices.....	19
4. Research definition.....	20
4.1. Motivations.....	20
4.2. Application, goal and objectives .....	20

4.3. Business applications .....	23
4.4. KPIs – What statistics to be extracted?.....	24
5. Data Sources.....	25
5.1. Call detail records dataset .....	25
5.2. Vodafone cell sites dataset .....	26
5.3. Lisbon parishes shapefile .....	28
5.4. Mobile Network operator dataset .....	29
6. Methodology, tools and approach .....	30
6.1. Approach.....	30
6.2. Tools .....	31
6.3. CDR dataset analysis .....	33
6.3.1. GDPR & Vodafone regulation.....	33
6.3.2. CDR dataset pre-processing .....	34
6.4. Geospatial Upper scaling approach .....	35
6.4.1. Basic spatiotemporal capabilities of call detail records .....	36
6.4.2. Spatial assumption to perform the geospatial upper scaling .....	37
6.4.3. Cell sites network coverage attribution.....	38
6.4.4. From cell sites network coverage to Parishes.....	46
6.5. KPIs – Key Performance Indicators .....	48
6.5.1. Density distribution count per parish.....	49
6.5.2. Density distribution count hourly-wise per parish.....	51
6.5.3. Nationality parsing.....	51
6.5.4. Top X nationality .....	51
6.6. Towards population extrapolation .....	51
6.7. Code testing.....	52
7. Experimental results .....	54
8. Conclusions.....	56

## **Chapter 3 Gender prediction from subscribers’ list of installed mobile apps ..... 57**

Abstract .....	58
1. Introduction .....	59
2. Literature review .....	60
3. Research definition .....	63
3.1. Problem statement .....	63
3.2. Study area.....	64
3.3. Goals and objectives .....	65
3.4. Research architecture .....	67
4. Methodology, tools and approach .....	69
4.1. Big Data Platform – BDP.....	69
4.2. Tools .....	70
4.3. Approach .....	71
4.4. Data exploration .....	73
4.4.1. Primary considerations.....	73
4.4.2. Target quantification .....	74
4.5. Data pre-processing .....	75
4.6. Data collection.....	78
4.7. Data preparation .....	80
4.8. Data analysis and visualization .....	84
4.9. Data modelling .....	90
4.9.1. Binary classification.....	90
4.9.2. Classification algorithms.....	90
4.10. Models evaluation.....	95
4.10.1. Model training and Testing sampling.....	95
4.10.2. Evaluation metrics.....	97
5. Experimental results .....	100



6. Conclusions.....	105
Bibliography.....	106

# List of Figures

Figure 1: Vodafone global presence (shaded in red) .....	3
Figure 2: Cell sites collocated on a cell tower .....	16
Figure 3: IMSI hierarchy - Example .....	18
Figure 4: Location of Lisbon, Portugal, in Europe .....	21
Figure 5: Map of Lisbon with its 24 parishes .....	21
Figure 6: Cell sites coordinates.....	27
Figure 7: Cell sites coordinates on top of the map of Lisbon.....	27
Figure 8: Map of Lisbon: Olivais parish with red bordered .....	28
Figure 9: QGIS Software graphic user interface with visualization of cell sites (yellow points) onto the map of Lisbon.....	32
Figure 10: Upper scaling approach of one cell site.....	35
Figure 11: Cell site location, a mobile subscriber connected to access the cellular network..	36
Figure 12: Map of Lisbon with representation of cell sites' antennas signal ranges in blue ..	37
Figure 13: Voronoi diagram application into an Euclidean plane of 4 cell sites.....	39
Figure 14: Step 1 – Load cell sites as inputs on QGIS.....	42
Figure 15: Step 2 – Create Voronoi Diagram based on selected cell sites (one per coordinate) .....	42
Figure 16: Downscaling of Lisbon city from parishes to cell sites network coverage .....	43
Figure 17: Assumed location of a subscriber based on its call detail record (reded polygon)	44
Figure 18: Map of Lisbon with three cases of cell sites' network coverage (in red, brown and blue).....	46
Figure 19: Example of KPI Density Distribution Count per parish.....	50
Figure 20: 2-hour intervals density distribution across 24 parishes of Lisbon.....	54
Figure 21: Research Architecture for the comparison analysis .....	68
Figure 22: Research Architecture: Objective 1 & 2.....	78
Figure 23: Outlook of the values of the 6 generated datasets.....	79
Figure 24: Format 1 – Apps name as predictors.....	81
Figure 25: Format 2 – Apps aggregated to their Google Play Category.....	81
Figure 26: Format 3 – Truncated Singular Value Decomposition example with 2 dimensions .....	82

Figure 27: Plot for selecting k number of TSVDimensions .....	83
Figure 28: Research Architecture: Objective 1, 2 and 3 .....	83
Figure 29: Count of mobile apps in the mobile subscriber's device (in percentage): .....	86
Figure 30: Percentage of mobile apps regarding their recurrence in the users' device .....	86
Figure 31: Frequency count of apps per Google Play Category .....	87
Figure 32: Frequency count of apps per Google Play Category and Gender .....	88
Figure 33: Percentage of mobile apps per Google play category and gender .....	88
Figure 34: Count of mobile apps per mobile subscriber's gender for category: TOOL (Frequency).....	89
Figure 35: Count of mobile apps per mobile subscriber's gender for category: TOOL (Percentage) .....	89
Figure 36: Logit Model (example with cut-off at 0.5) .....	91
Figure 37: Visual Example of a Decision Tree .....	93
Figure 38: Random Forest .....	94
Figure 39: Hold-out method (Train/Test dataset split) .....	95
Figure 40: K-fold Cross-Validation .....	96
Figure 41: AUC – ROC Curve .....	99
Figure 42: Model's performance considering the number of apps subscribers have in their device .....	104

# List of Tables

Table 1: Typical attributes of Call Detail Records .....	14
Table 2: Example of a Call Detail Record log.....	26
Table 3: Toy example of a cell site from the cell site dataset .....	26
Table 4: Data representation of parish Olivais from the shapefile .....	28
Table 5: Mobile Network operator dataset.....	29
Table 6: Toy example of a Call Detail Record log .....	33
Table 7: CDR log crossed with cell site dataset to localize the connected cell site .....	36
Table 8: Example of cell sites' antennas with in red the signal ranges.....	38
Table 9: Shapefile of cell sites dataset – Example of its content.....	41
Table 10: Example of Voronoi diagram output saved as a shapefile.....	43
Table 11: Mapping between Voronoi cell sites and the other cell sites.....	44
Table 12: CDR log crossed with cell site dataset to approximate the location of a mobile subscriber.....	44
Table 13: Updated cell site dataset with polygonal shape of each cell sites .....	45
Table 14: Three cell sites with proportions overlapping the parish(es).....	47
Table 15: Final Cell sites dataset with overlapping proportion over parish(es).....	47
Table 16: Description of the Function: Density distribution count per parish .....	49
Table 17: Toy cell site dataset and call detail records for a practical example using the function Density distribution count in Ajuda .....	50
Table 18: Toy example of merging CDR with cell site dataset.....	50
Table 19: Description of the stages of the data workflow.....	65
Table 20: Target Quantification – Basic statistics .....	74
Table 21: APK along with apps title and category if belonging to Google Play Store .....	75
Table 22: Google Play Store: 49 categories for its applications.....	76
Table 23: Generated table 'Google Play Android Apps' .....	77
Table 24: BDP temporary table of mobile subscribers with APK and gender known .....	79
Table 25: Confusion Matrix .....	97
Table 26: Experimental results with mobile apps title as predictors .....	102
Table 27: Experimental results with Google Play categories as predictors .....	103
Table 28: Experimental results with the TSVD method (500 dimensions) .....	103

# List of Acronyms

<b>API</b>	Application Programming Interface
<b>APK</b>	Android Package
<b>APPS</b>	Applications
<b>AUC</b>	Area Under the ROC Curve
<b>BDA</b>	Big Data Analytics
<b>BDP</b>	Big Data Platform
<b>CDR</b>	Call Detail Records
<b>CSP</b>	Communication Service Provider
<b>CV</b>	Cross-Validation
<b>DT</b>	Decision Tree
<b>ESRI</b>	Environmental Systems Research Institute
<b>FP</b>	False positive
<b>FN</b>	False negative
<b>GDPR</b>	General Data Protection Regulation
<b>GIS</b>	Geographic Information System
<b>GPS</b>	Global Positioning System
<b>GSM</b>	Global System for Mobile Communication
<b>HDFS</b>	Hadoop Distributed File System
<b>HQL</b>	Hive Query Language
<b>IG</b>	Information Gain
<b>IMS</b>	Information Management School
<b>IMSI</b>	International Mobile Subscriber

<b>IoT</b>	Internet of Things
<b>KPI</b>	Key Performance Indicator
<b>MCC</b>	Mobile Country Code
<b>MNC</b>	Mobile Network Code
<b>MNO</b>	Mobile Network Operator
<b>MSIN</b>	Mobile Subscriber Identity Number
<b>ML</b>	Machine Learning
<b>MLE</b>	Maximum Likelihood Estimation
<b>MVNO</b>	Mobile Virtual Network Operator
<b>LAC</b>	Location Area Code
<b>LR</b>	Logistic Regression
<b>L2P</b>	Location To Profile
<b>RF</b>	Random Forest
<b>ROC Curve</b>	Receiver Operating Characteristic Curve
<b>SIM</b>	Subscriber Identity Module
<b>SQL</b>	Structured Query Languages
<b>TP</b>	True positive
<b>TSVD</b>	Truncated Singular Value Decomposition
<b>TN</b>	True negative
<b>TPR</b>	True positive rate
<b>UMTS</b>	Universal Mobile Telecommunication System
<b>YARN</b>	Yet Another Resource Negotiator
<b>VF</b>	Vodafone

# **Chapter 1**

## **Internship description**

### **1. Academic context**

This internship report is part of the second-year Master's degree in Advanced Analytics & Data Science of the faculty IMS (Information Management School) of the University Nova of Lisbon in Portugal. As part of the Master's requirements, students are to choose to complete either a thesis, a project or an internship with along its written report. As per, this report summarises my contribution part of the Big Data & Advanced Analytics team of Vodafone Portugal. My internship lasted 5 months, starting from early October 2018 till end of February 2019 and took place in the office based in the modern area of Lisbon, Parque das Nações.

### **2. Internship context**

The promise of Big Data Analytics (BDA) has come as a data-centric solution for corporate to strive in this highly competitive and quantitative world. For the last decade, Big Data Analytics has proven its potential to remodel how companies manage and enhance high value businesses performance (Pugna, Dutescu, & Stănilă, 2019). Nowadays, Data is everywhere, mostly, with the advent of Internet of Things (IoT) and Web 2.0 technologies. In 2018, the world created 33 zettabytes of data, and is forecasted to create 20 times more data in 2030 (Armstrong, 2019). In fact, Big Data is big news in most industries, and telecommunication is no exception. Although, Big Data is still in early phase of deployment in telecom, it became a core topic for most telecom executives (Bughin, Reaping the benefits of big data in telecom, 2016). As a fact, telecom operators are capturing more and more data volume and are benefiting from a larger variety of sources than ever before.

Vodafone Portugal, originally born as Telecel in 1992, became a full subsidiary of the British multinational telecommunication conglomerate Vodafone Group since 2003. As one of the world's largest mobile communication providers, Vodafone Group, has mobile operation in 24 countries, including Portugal, and has partner networks in over 43 more. As of 30 June 2019, Vodafone Group counts a total of approximately 640 million customers across the world, including 4 million in Portugal, which represents around one third of the Portuguese market share (Vodafone, 2019). Vodafone provides to both consumers and businesses a wide range of products such as handsets and communication services such as voice, messaging, internet data and fixed broadband through a series of diverse solutions. Besides that, Vodafone delivers cloud, security and carrier services to enterprise customers.



## CHAPTER 1. Internship description

---

The Portuguese telecom market counts four main operators: Vodafone, NOS and MEO are Mobile Network Operators (MNO) as they own and run on their infrastructures while NOWO is a Mobile Virtual Network Operator (MVNO) as it doesn't have its own infrastructure and thus rely on MNO's one.



Figure 1: Vodafone global presence (shaded in red)

### 3. Research projects

The Big Data & Advanced Analytics team of Vodafone Portugal is composed of 5 to 6 senior data scientists and 1 senior manager. Their main challenge is to extract and leverage valuable insights from the tremendous amount of data generated and collected by Vodafone. Although, the team has a focus on their own big data projects, they also work closely with several departments to assist them in their analytical tasks (e.g. marketing campaigns, reducing churn rate) and with the data engineering team to improve the data pipeline processes as well as developing and maintaining cloud-based solutions. Vodafone Group counts many teams of data scientists based in several countries, in which the company is operating, such as in Portugal, Italy, India, Egypt, U.K and constantly works on improving the cohesion and knowledge-sharing across the different teams.

My 5-months internship can be described as an immersion into the data science area as being into the boots of a data scientist at Vodafone. Core goals of my internship were of learning, discovering and apprehending the daily life, challenges, accomplishments and struggles of a Data scientist at Vodafone Portugal. Throughout my stay in the team, I was supported and

## CHAPTER 1. Internship description

---

guided closely by my supervisor and mentor, Fernando Silva, a senior data scientist whom assigned me projects, goals, objectives, guidelines and advice. In this report, I describe in detail the two projects I have been proposed to work on, that interested me. I invested more or less the same duration for both research projects, which was in between 2 to 3 months per project: from mid-October till end of December and from early January till end of February.

My first assigned research project consisted in studying human mobility from cellular network-based data considering the so-called *Call Detail Records* (CDR) as a core data source. A call detail record is generated every time a person sends/receives a voice call, a text message or even connects to the internet through its mobile operator. Primarily used for billing purposes and network capacity improvement, these records contain valuable spatiotemporal information, connected cell site and timestamp, respectively. The research use-case consisted in analysing anonymized call detail records that took place in Lisbon. In short, the achieved objectives were to develop from CDR, (1) KPIs that extract spatiotemporal density distribution insights, (2) a geospatial upper scaling approach to determine subscribers' locations at specific spatial scale and (3) Code testing to ensure and monitor the functionality of the developed functions and KPIs.

My second assigned research project consisted in conducting an observational study of the predictability of mobile subscribers' demographic traits from their installed mobile applications. To do so, the application focused on the gender of mobile subscribers. Few published papers demonstrate the capability to predict certain demographic traits such as gender or age based on the users' installed mobile applications. This approach has gained sufficient interest to open this topic considering that Vodafone lacks accurate demographic traits for a subset of its mobile subscribers and motivated by scientific curiosity. In the era of data-driven solution, customer demographic traits such as gender or age play a crucial role for customer centric strategies (e.g. campaign development, market quantification). Therefore, this research project was most of all an exploratory work to evaluate Vodafone capabilities to predict the gender of its mobile subscribers based on their installed mobile applications. This project encompasses the following steps of the data science workflow: Problem understanding, data exploration, data pre-processing, data collection, data preparation, data analysis & visualization, data modelling and models evaluation.

### 4. Internship report outline

This written report includes at its core the two projects I have been working on, during my five months internship, part of the Big Data & Advanced Analytics team of Vodafone Portugal.

The report follows a structure in three main chapters:

- I. Internship description
- II. Project 1: CDR-based location analytics
- III. Project 2: Gender prediction from subscribers' list of installed mobile applications.

The first chapter includes the aforementioned three sections (*Academic Context*, *Internship Context* and *Research projects*) as well as this same section (*Internship report outline*) and finally the two following sections (*Place of Mobile phone in the modern era* and *the role of Big Data in telecom companies*). This primary chapter has an introductory role. It defines the background, context and directions of the whole report as well as giving a first overview of the study area.

The two following chapters are each describing one of the two projects. Both chapters follow a similar structure, which is inspired from the academic model of writing a Master thesis and basically contain the below sections:

- Abstract
- Introduction
- Literature Review
- Research definition
- Methodology, tools and approach
- Experimental Results
- Conclusions

The bibliography comes at the end of the report to not surcharge each chapter.

Regarding the specific examples and visualizations of this report, they are all based on synthetic data and not on real data collected by Vodafone.

### **5. Place of mobile phone in the modern era**

The global economy has drastically evolved over the past century and a half and the telephone is known as one of the world's most transformative inventions by mankind. It has revolutionized the way humans communicate across any kind of distances, passing from landlines with fixed locations to connected mobile devices. Nowadays, mobile phones are everywhere. As of today, more than 60% of the world population has a mobile phone connection (Statista, 2019), and since 2015, the number of mobile subscriptions has outnumbered the world population (The World Bank, 2018). In most developed countries, its penetration rate has almost reached 100%, and even in the remote places of developing countries, the use of connected devices is not unusual. Besides that, since 2017, more than half of worldwide website traffic is generated through mobile phones (Statista, 2019).

As one of the most ubiquitous technologies of the modern society, the mobile phone has literally become part of our daily life. From communication to entertainment, mobile devices switching towards "Smartphones" are more and more used as an unavoidable tool to stay connected to others and to instant information. With more than 3 million apps available across Google and Apple digital marketplace, such handy devices are used for a vast number of versatile tasks, making them unique personal assistants equipped with tons of sensors capable of doing a myriad of applications.

In today's modern society, it is normal to carry your mobile phone everywhere all day and night. Each and every connected device, including the simplest cell phones generate data. Given their ubiquity, increasing functionalities and powerful sensors, mobile devices are ever-increasingly employed in diverseness ways to collect data. Mobile phone datasets have even grown into a stand-alone topic and have already found numerous contributions across a wide number of fields such as in social network, human mobility, urban planning, smart cities, disaster and emergency management. More precisely, a tremendous amount of mobile data is generated and collected by mobile operators and can unveil a lot about the users' behaviour, activities habits and demographic traits. To name some sources, every text, phone call, internet search, app use or even usage of one's device's sensor conduce to the development of big data. It is predicted that by 2020, for every person in the world, more than 6 gigabytes of new information per hour will be created (Marr, 2015).

Since more than a decade, the place of mobile phone has opened the door to various empirical studies, as the essence of mobile phones have revealed to be a source of valuable data. Namely, with the use of CDR – Call Detail Records – collected by telecom operators. As already said, CDR contain information about every call, text message and internet connection carried by the operator, listing features such as caller, callee, date, time, duration, interaction type and the geo-location of the cell tower(s) handling the interaction with the cellular network along more features.

## **6. Role of Big Data in Telecom companies**

With the rise in the use of mobile devices, Communication Service Providers (CSP) see massive amount of data at their disposal. With more users and connected devices than ever before, telecom companies constantly need to keep the pace with their environment, so that they can harness the Volume, Variety and Velocity of data coming into their organization. Mobile data has grown over 17-fold over the past 7 years. By 2022, smartphones are expected to surpass 90% of mobile traffic with an annual traffic reaching almost one zettabyte, a zettabyte equal a trillion gigabytes (Cisco, 2019). Becoming a data-driven telecom company requires major cultural shift. Main bottlenecks encountered in big data projects are due to constraints related to the data (in terms of quality, quantity and permissions), talents with lack of required domain knowledge and organization's culture not sufficiently embracing big data (Bughin, Reaping the benefits of big data in telecom, 2016).

As Carlos Santos, Head of Big Data of Vodafone Portugal said to the Portuguese Journal 'Jornal de Negócios', the big challenge stem from the way to extract and take advantage of all this amount of information that is generated. Any big data projects must start with a "smart data plan". The latter, following a top-down approach, begins with the business problem. From that point, one can start collecting useful data, cross all known techniques in the area of artificial intelligence, statistics and machine learning with the support of distributed computing infrastructures. The recipe needed, to explore and bring out the interpretable insights from these raw data. Finally, a team of data scientists with mix valences matters to be able to transform those insights into actionable business decisions. (Santos, 2017)

As previously mentioned, telecom companies are daily collecting large volumes of various type of data such as from call detail records, mobile phone usage, network equipment, billing and so on, providing lots of information about their customers and network.

## CHAPTER 1. Internship description

---

Big Data enablers are the ways to handle efficiently the data pipelines, namely, the rapidity to process the data, secure it, store it and derive insights from it. After all, how much can telecommunication companies really benefit from big data? Hereafter are few of the key benefits telecom has proven to benefit Big Data Analytics:

- ❖ Enhanced targeted marketing: Proactively establish customer centric KPIs and develop more personalized and adequate offers (Make use of meaningful customer data);
- ❖ Customer prediction (Churn): As one of the biggest challenges in the telecom industry, it aims at identifying customers who are most likely to leave;
- ❖ Improvement of Network services: ongoing improvement in the network capabilities;
- ❖ Data Driven improvement of security systems: To prevent fraud detection, ensure payment processing, data compliance and protection.
- ❖ Location Analytics: Telecom operators collect a ton of spatiotemporal data which enable location analytics to be applied in a wide range of applications such as in urban planning by identifying how subscribers are moving across a given area (e.g. most visited location), identify locations which require 4G services improvement and so on... Besides that, it opens the door to data monetization opportunities by selling to third-parties location intelligence insights.

The industry is awash in information, but only a few companies manage it effectively. (Bughin, Telcos: The untapped promise of big data, 2016)

# **Chapter 2**

## **CDR-based location analytics**

### **Abstract**

For most of human history, people had limited options for exchanging information with one another or accessing new information. In the past, people had to physically deliver messages, write letters and/or send them with the help of birds or more recently with the use of telegrams. In the last decades, we saw the technology dramatically revolutionizing the way people communicate locally and globally as well as enabling access to instant information. Today, real-time communication is made possible almost anywhere around the globe, and the internet and mobile devices are responsible for a big part. In the modern age, the telecom industry plays a undeniable role in the communication sector. With the advent of Big Data, telecom operators see at their disposal tremendous amount of data from call detail records, mobile phone usage, network, server logs, billing to social networks. Whenever someone makes or receives a call, a text message or turns on his mobile data, a call detail record (CDR) is automatically generated by the cellular carrier. CDR have the particularity of having spatiotemporal references about people's interactions with the cellular network. Originally collected for billing and network improvement purposes, their usage have taken applications to a wide range of research fields to study different aspects of human mobility. In short, telecom operators see huge amount of data about how, when, with whom (anonymized) and where communications are made each day on a network. This primary research project summarises the applied process of leveraging call detail records to extract spatiotemporal density distribution at specific spatial scale. More concretely, the use-case application consisted in extracting from call detail records such spatiotemporal insights across the 24 parishes of the city of Lisbon, Portugal.



### 1. Introduction

The advent of Big Data technologies, the rise in Internet of Things (IoT) and the widespread adoption of ubiquitous computing by very large portions of the world population, have enabled large-scale human mobility researches (Zhao, et al., 2016). In fact, billions of personal connected devices such as, mobile phones, smartwatch, fitness bracelets or self-driving cars connect us to the cloud, providing rich information to measure people daily mobile routine (Cuttone, Lehmann, & González, 2018). The analysis of mobility data obtained from tracking moving objects, have replace the need of relying on non-scalable and expensive methods like surveys or direct observation to get a glimpse of people movements (Isaacman, et al., 2012). Subsequently, in the last decade, there have been a surge in the number of studies related to human mobility patterns, given their applications into a variety of societal topics, such as in urban planning, traffic forecasting, epidemiology, policy-making, disaster and emergency management. In the literature, a multitude of proxies extracted from mobile phone with spatiotemporal characteristics have been considered to study human mobility, such as Call Detail Records data from mobile operators, GPS and/or WiFi-based data from location-based social networks or third-party mobile apps data such as Twitter or Uber.

In this research project, I address the process of leveraging Call Detail Records data to study human mobility with the particular use-case of developing KPIs which identify the population density distribution across the 24 parishes of Lisbon city, Portugal. Originally collected by mobile operators for billing and network analysis purposes, Call Detail Records (CDR) contain timestamped and geo-referenced logs on all of their subscribers' interactions with the cellular network, i.e. calls, text message or internet data connection. The main challenges with this data source are that CDR temporal resolution differs for each subscriber according to the mobile phone communication patterns and CDR spatial resolution doesn't refer to the subscriber location but rather to the handling cell site(s) location. However, given the large availability of CDR for a considerable part of the population, in recent years, several anonymized CDR logs datasets have been widely investigated to perform large-scale human mobility studies.

This primary project describes in detail all the theoretical and practical steps taken throughout the research study with the core goal of using call detail records logs as a proxy to estimate the population density distribution across the 24 parishes of Lisbon, Portugal. Generating such spatiotemporal insights have numerous business applications both for Vodafone internal purposes as well as for external purposes as explained throughout this chapter.

### 2. Literature Review

The project presented in this second chapter ‘CDR-based location analytics’ is related to a considerable number of studies on the analysis of network-based data to study human mobility and more precisely, the extraction of spatiotemporal insights. The exploitation of cellular network data for studying human movements has been an active area of research in the last decade, given the ubiquitous aspect of mobile phones and their worldwide availability (Kujala, Aledavood, & Saramäki, 2016). Indeed, the study of human mobility has particularly attracted the scientific community given its applications across a wide range of fields. Call detail records data have been to the fore, when searched for a proxy to study human mobility. Collected by mobile operators, those records contain non-continuous traces of people when interacting with the cellular network and include enormous amount of information on how, when, where and with whom people communicate on a daily basis.

To name specifically some contributions considering call detail records, there are, post-earthquake geospatial study in Haiti ( Bengtsson, Lu, Thorson, Garfield, & Von Schreeb, 2011), mapping malaria in Kenya (Wesolowski, et al., 2012), optimizing transport networks in Abidjan (Berlingerio, et al., 2013), CDR analysis for assisting policy intervention to control Ebola epidemic in Guinea (Shibasaki, 2017) and many more which study the process of translating CDR into spatiotemporal insights (Blondel, Decuyper, & Krings, 2015). These contributions include diverse empirical methods, framework and results for extracting spatiotemporal insights from mobility data.

Given the low spatial resolution of call detail records, most of them have in common the process of reconstructing the resolution to a more fine-grained spatial scale. Call detail records do not ‘observe’ people but rather connected devices, and at the handling cell sites spatial resolution when interacting with the cellular network. Besides that, another factor affecting the experimental approaches of those studies are related to the use of real or synthetic data. Indeed, using real data brings up privacy matters of subscribers and enforced data usage laws in the given country where the data is collected. The current European (EU) legislation for personal data and their usage is the General Data Protection Regulation (GDPR), which took effect and was enforced since the 25<sup>th</sup> May 2018. In its scope, GDPR encompasses pseudonymization of personal data among much more regulations as to safeguard individual privacy (Jones, Daniels, Heys, & Ford, 2019). Although, large-scale human mobility studies usually implies anonymizing and aggregating the call detail records of all mobile subscribers for a given area.

## CHAPTER 2. CDR-based location analytics

---

In 2016, (Zhao, et al.), investigated the possible caveats and limitations of using CDR for studying human mobility and, although, their conclusions are not black or white, they confirm that the underlying nature of CDR introduce a certain degree of bias given the uneven distribution of people's communication activities in space over time. In other words, some people interact much more with the cellular network than some others. Their experiments demonstrate that CDR tend to underestimate some mobility indicators such as the total travel distance. CDR scarcity of information is a problematic approach as well by (Fiadino, Ponce-López, Antonio, Torrent-Moreno, & D'Alconzo, 2017), whom studied the situation comparing two datasets collected by Nation-wide operator in 2014 and 2016. Living in the 'always connected era', their conclusions are that the quality and volume of CDR data has drastically changed, providing higher temporal accuracy for users' locations in 2016 compared to 2014. They believe that this path will even continue and grow in the future years, given the increase in mobile communication flat plans from mobile operators.

Considering that CDR underlying spatial characteristic refers to a low resolution, handling cell site instead of subscriber location, researchers demonstrate across their publications different approaches and methodological processes to reach a finer-grained resolution to perform their analysis. The Joint Research Centre Technical Report proposes a systematic methodological framework using mobile phone network-based data, call detail records (CDR) and visitor location register (VLR), using data of multiple mobile network operators (MNOs) for the task of estimating population density distribution at pan-European level (Ricciato, Widhalm, Craglia, & Pantisano, 2015). The ITU report demonstrates the usage of CDR for approaching human mobility to tackle Ebola outbreak in Guinea, by performing a human movement analysis at two scales; city-to-city movement in Guinea and then transboundary movement across neighbouring countries, Guinea, Sierra Leone and Liberia (Shibasaki, 2017). Usage of call detail records to study specifically urban mobility for urban-planning has as well been a hot-topic. There is a study, which takes as example Singapore city, to use call detail records to generate urban micro simulations of individual daily activities and travel for urban and transportation planning purposes (Jiang, Shan, Ferreira, Jr., & Gonzalez, 2017). Another paper, on which experiments were performed in the Greater Maputo metropolitan area in Mozambique, proposes a method to translate CDR into origin-destination trips. As such, computation methods creating visualization of trip generation maps, attraction maps and the Origin-Destination matrix to reconstruct the population distribution and understand their flow pattern (Batran, Mejia, Kanasugi, Sekimoto, & Shibasaki, 2018).

### 3. Essential Telecom GIS concepts

Before going further in this project, it is important to get a good grasp of a few concepts about mobile networks and mobile phone datasets. As such, this section describes fundamentally the following technical concepts: Call detail records, the distinction between cell sites and cell towers, these location-aware data coupled with additional demographic data, IMSI identifier, the shapefile spatial data format and the unit assumption between mobile devices and people.

#### 3.1. Call Detail Records

In telecom, talking about mobile phone dataset to study human mobility, is synonym of using CDR logs. Call detail records (CDR) are metadata about how phone numbers use the cellular network. Whenever a mobile subscriber makes/receives a call, a text message or even simply turns on his mobile data, a call detail record about the interaction itself is generated and collected by the concerned cellular carrier. Call detail records contain both spatial and temporal information. Content excluded, a record typically includes data for the following attributes about the interaction, among many other details:

Table 1: Typical attributes of Call Detail Records

Attributes	Description
<b>Caller ID</b> (e.g. IMSI identifier, Described later in this same section)	ID of user whom triggered the event i.e. mobile subscriber made a call, send a text message or turned on his internet data
<b>Callee ID (e.g. IMSI identifier)</b>	ID of User whom received a call or text message.
<b>Event ID</b>	Action that triggered the event. i.e. call, text message, internet data connection
<b>Direction (Inbound, outbound, none)</b>	Specifies if call detail record was incoming, outgoing or none for internet data connection.
<b>Duration</b>	Time interval of the entire event.
<b>Cost</b>	Monetary cost of the event.
<b>Timestamp</b> (Temporal attribute)	Datetime when event was recorded.
<b>Cell site ID or coordinates</b> (Spatial attribute)	ID or coordinates of cell site which handled the event with the cellular network.

## CHAPTER 2. CDR-based location analytics

---

Primarily used for billing purpose and network capacity improvement, telecom operators see at their disposal a huge amount of data about how, when, with whom (anonymized) and where communications are made each day on a network. Through the ever-evolving big data capabilities, CDR huge amount of daily generated data have opened the door to a rich number of researches across different fields such as in social interaction, economic activity or more particularly human mobility, both on individual and population level using disaggregated or aggregated CDR data. Telecom mobile phone data for studying human mobility has been a highly popular topic for the last two decades, thanks to its large-scale data collection process outperforming the traditional non-scalable techniques (e.g. survey) and its ubiquitous capacity to capture the population patterns and movements.

Despite that CDR are biased by the mobile subscribers' activity degree and that they only pin the location of the connected cell site(s) instead of the subscriber's exact location at the time of the interaction, several techniques have surged to more accurately unveil human movements and location. Indeed, there is a fairly rich number of published papers about the use of those non-continuous location traces such as for urban planning (e.g. population density), traffic forecasting, commuting patterns or for social good studies such as for disaster and emergency management (e.g. Ebola epidemics in Guinea).

As a matter of fact, CDR by themselves are usually of minimal information for human mobility analysis, as they typically only contain ID or coordinates of the connected cell site. In order to perform a comprehensive analysis, technical information about all cell sites of the operator in the region of interest must be gathered as to properly identify and map out specific network coverage area of all cell sites.

For this reason, gathering for the region of interest (e.g. Lisbon) all cell sites information from the concerned mobile operator is as important as CDR, in order to pursue efficiently the human mobility analysis.

### **3.2. Cell sites and Cell towers**

A cell site is a cellular-enabled mobile device which house electronic communications equipment such as transmitter/receiver antennas, GPS, base transceiver station and more. Usually located on top of elevated structures (buildings, hills...) or cell towers built by telecom or tower companies, to ensure a proper elevation of the antennas. It is common to see several cell sites sharing the same structure to enhance the network coverage. Cell sites are more particularly designed to (1) efficiently receive and transmit radio-frequency signals from/to connected devices like cell phones or radio and to (2) handle a multitude of devices, ranging from a dozen to more than a hundred simultaneously, by operating on different radio frequencies. The below figure shows a cell tower with several cell sites collocated on it. Those cell sites can belong to one operator or in some other cases, to several telecom operators.

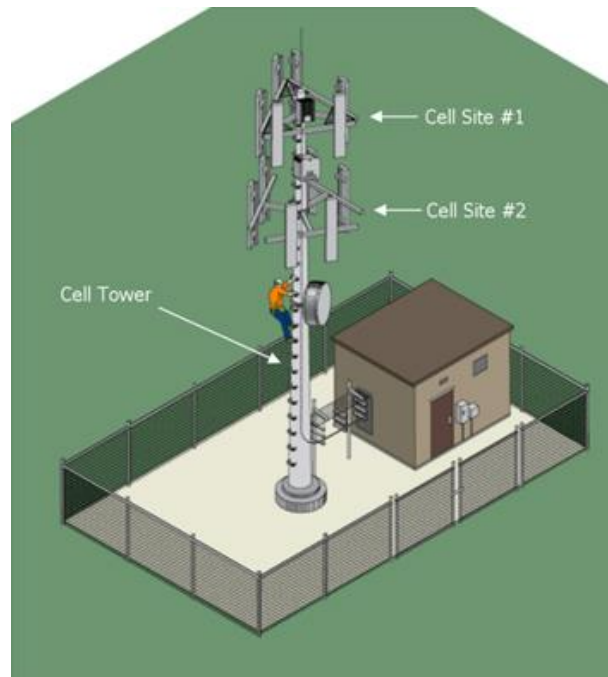


Figure 2: Cell sites collocated on a cell tower

Whenever someone use his mobile phone to make a call, send a text or even access internet via his mobile service provider, his device emits electromagnetic radio waves that is to be theoretically received by the nearest cell site or another close one to access the cellular network. Signals are then going back and forth through the network to reach and connect the other mobile phone through the callee's nearest cell site or instead connects to the internet. It is around that time that the call detail record about the interaction is generated and collected by the telecom operator(s). For example, Vodafone collects data for the caller and NOS for the callee.

## CHAPTER 2. CDR-based location analytics

---

Providing a good network coverage to all mobile subscribers implies having numerous and well-located cell sites. In practice, most are placed in urban areas or in other densely populated geographical locations (cities, suburbs, main roads...), but with their proliferation around the earth, they can be found in lots of rural areas, especially of developed countries. Agreements between operators of different countries are commonly made to allow mobile subscribers to use the network abroad. For example, mobile subscribers with Vodafone Portugal connecting to the cell sites of the operator Orange in Belgium to access the cellular network (i.e. Roaming).

Mobile phones are specifically designed to constantly scan nearest cell sites and usually connect to the one with best connectivity, which is shown in the form of signal strength on the mobile phone screen. It is always to the network to make the final call to which cell site a mobile phone connects, based on a set of technical measurements to improve the network connectivity. To name some, the distance between mobile phone and cell site, the connecting technology, landscape features such as buildings or hills, antennas' signal strength or even that some cell sites are set with reduced antennas' signal to not interfere with neighbouring cells. As such, a device may usually connect to the nearest cell site of its operator but not necessarily as depending of the overall connectivity among other parameters. Being on the move, mobile phones switch from one cell site to another in order to constantly send/receive information with the cellular network. As said above, although cell sites collocated on a same structure means they share identical coordinates, their antennas often differ in their features, namely, in terms of *azimuth* (direction), *beam width* (signal openness) or *radius* (operating signal distance).

- ✓ **Azimuth** – Antenna's direction. It is specified in the units of degrees (i.e. North is 0°);
- ✓ **Beam width** - Antenna's signal openness (i.e. omnidirectional openness is 360°);
- ✓ **Radius** – Maximum distance signal can reach (i.e. 1000 meters from antennas).

### 3.3. Additional subscribers data

Although, call detail records do not include demographic variables (i.e. age, gender) of mobile subscribers generating them, it is theoretically possible for the operator to merge their CDR with personal customer data. Thus, generating more diverse statistics, such as the gender or age distribution of its subscribers across specific locations and timespan (e.g. shopping mall). However, all data must be anonymized, encrypted before any analysis happen, such approach should be looked in much depth as to respect enforced data regulations and laws related to the use of customer data in the concerned country (e.g. General Data Protection Regulation, GDPR).

### **3.4. International Mobile Subscriber Identity - IMSI**

The International Mobile Subscriber Identity (IMSI) is a 15 maximum length unique digit code used to identify any mobile phone subscriber across the globe. It is stored in every SIM card (Subscriber Identity Module). To ensure uniqueness, this code follows a hierarchical structure in the following three components: MCC for Mobile Country Code, MNC for Mobile Network Code and MSIN for Mobile Subscriber Identity Number. Finally, IMSI is associated with all Global System for Mobile Communications (GSM) and Universal Mobile Telecommunication System (UMTS).

- ❖ **Mobile Country Code (MCC)** – The MCC has always the primary three digits to identify the country;
- ❖ **Mobile Network Code (MNC)** – The MNC follows from one to three digits to identify the network operator;
- ❖ **Mobile Subscriber Identity Number (MSIN)** – The MSIN at last has from nine to ten digits to identify the mobile subscriber. (This part is usually encrypted/ anonymized before any analysis happen). Hereafter is a toy example to understand the IMSI:



Figure 3: IMSI hierarchy - Example

In most cases, IMSI is included in all generated call detail records as the identification number to identify every mobile subscribers. Although, the MSIN is usually anonymized to ensure the non-recognition of subscribers. The MCC and MNC are kept as they are, to enable further analysis to identify the subscribers' country of origin and network operator, respectively.

In this project research, IMSI was predominantly used for aggregating CDR per mobile subscriber and for aggregating or filtering CDR per country (e.g. MCC = 268 for Portuguese mobile subscribers or != 268 for all non-Portuguese mobile subscribers/tourists).



### **3.5. Shapefile spatial data format**

Shapefile is a popular geospatial vector data format used in Geographic Information System (GIS) and is mostly known for storing location, shape and attributes of geographic features.

This format can spatially describe vector features: points, lines and polygons. Those vectors can create the shapes of spatial objects, and together create a representation of the geographic data. For example, points for trees, lines for roads and polygons for lakes. More concretely, points for Vodafone cell sites and polygons for Lisbon parishes or for attributing a network coverage to Vodafone cell sites. The shapefile format is made up of at least four files that are recognized by their extensions: .shp contains the feature geometry, .shx the spatial index linked to each feature, .dbf contains each feature attributes and .prj which has information about projection and coordinate system.

In this project research, the shapefile format was required at several stages of the research project, namely, to spatially represent the 24 parishes of Lisbon and for the so-called Voronoi mapping technique which administrated polygonal network coverage shapes to Vodafone cell sites. (It is explained in much depth later in the report).

### **3.6. Unit association between people and mobile devices**

Strictly speaking, the cellular network does not ‘observe’ people, but rather mobile phones. Although, the easy assumption between mobile phones and people being 1:1 is made, it is not always the case. This assumption is a source of error which can potentially create ‘noise’ in the analysis when leveraging call detail records to estimate the population density in a given area.

Nowadays, some people have multiple phones (e.g. one personal and another one for business), and some others do not have any mobile phone. Besides that, it happens that some mobile phones keep fixed location and do not belong to any person (e.g. machine to machine communications).

In this initiated research, the assumption made is that each mobile phone belong to one person, and that each person has exactly one mobile phone. This research focuses more in the whole process of leveraging call detail records to generate spatiotemporal insights. This aspect is a limitation that must be approached in much depth in future works.

### 4. Research definition

#### 4.1. Motivations

The core research motivations stemmed from the interest in approaching the current telecom capabilities to geo-localise mobile subscribers as well as in exploiting cellular network-based data to extract human mobility patterns. That is to say that those motivations are mainly coupled with the following societal, technological and academic enablers:

- ✓ Ubiquitous place of mobile phone in the modern society;
- ✓ Telecom operators collecting daily spatiotemporal logs of subscribers' interactions with the cellular network (*Call Detail Records* being the core data source);
- ✓ Big Data telecom capabilities to collect, store and process huge amount of daily data;
- ✓ Telecom operators aware of the richness in their collected customer location data;
- ✓ Literature on the exploitation of telecom spatiotemporal data to study human mobility.

The central components of this location analytics project turned around Call Detail Records (CDR) that contains timestamp and geo-referenced logs of mobile subscribers' interactions (calls, text messages and internet connection) and the Vodafone cell sites scattered across Lisbon city.

#### 4.2. Application, goal and objectives

Arise from the above motivations, my supervisor oriented me with a concise research baseline encompassing a use-case application with along a goal and objectives that were to be further defined throughout the study.

The use-case application consisted in analysing a week of anonymized call detail records collected by Vodafone that took place in the municipality of Lisbon, Portugal. With the research purpose of learning and applying the process of translating call detail records into spatiotemporal insights.

To give some geospatial context, Lisbon is the capital and largest city of Portugal and is composed of 24 civil parishes, as shown in the two illustrations on the next page.

## CHAPTER 2. CDR-based location analytics

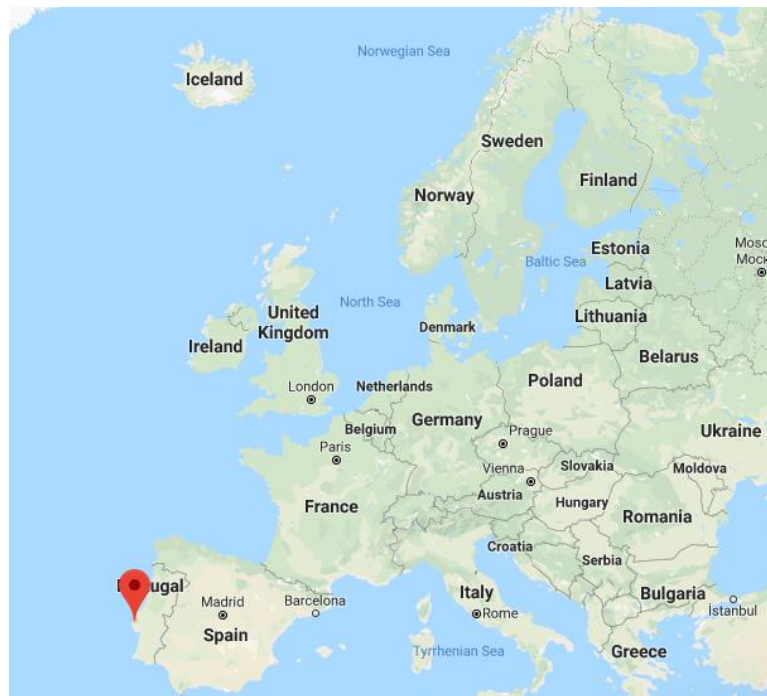


Figure 4: Location of Lisbon, Portugal, in Europe

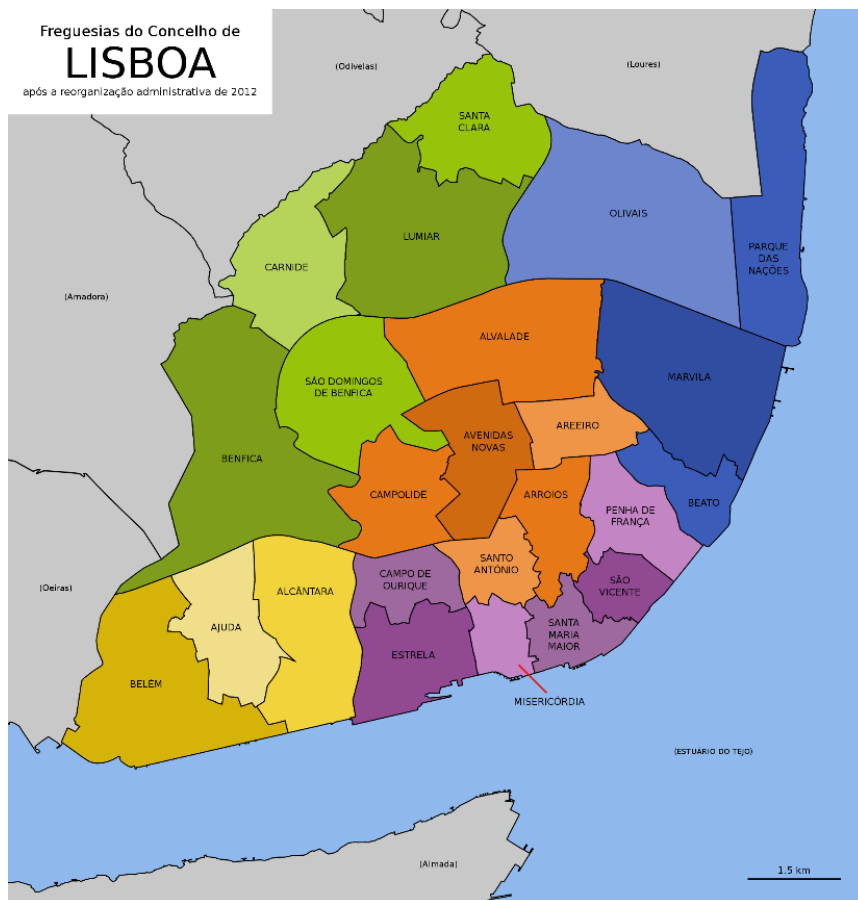


Figure 5: Map of Lisbon with its 24 parishes

## CHAPTER 2. CDR-based location analytics

---

The research goal consisted in extracting spatiotemporal density insights in Lisbon and more specifically across the 24 civil parishes of Lisbon. Throughout the study, several objectives derived from the goal as to measure the progress toward the goal and to clarify important steps of the analysis:

✓ **Define a Geospatial upper scaling approach:**

This objective referred to the low spatial resolution of call detail records. Indeed, every time a subscriber interacts with the cellular network (i.e. call, text message or internet connection), the spatial information collected by the concerned operator only refers to the connected cell site, instead of the actual subscriber location. As per, this objective encompassed a two-step process, which firstly identifies the subscriber location as being within the connected cell site network coverage and secondly, enlarge that spatial resolution to the parish scale. It is described in much depth later in the subsection 6.4. Geospatial upper scaling.

✓ **Develop KPIs that translate CDR into spatiotemporal density distribution insights:**

This second objective consisted in developing KPIs through functions that translate call detail records into spatiotemporal density distribution of mobile subscribers across the 24 parishes of Lisbon. It includes as well other functions to filter or parse specific call detail records, namely, to filter CDR based on the assumed nationality of mobile subscribers (e.g. Total of n Spanish were in Lisbon on 31<sup>st</sup> October 2019, and total of n Belgian were in Lisbon on 7<sup>th</sup> November).

The developed statistics only refers to mobile subscribers connecting to Vodafone cell sites. Although, the research project didn't reach that point, it was thought to extrapolate these statistics to more 'real/plausible' counts by considering the telecom operators market share statistics. For example, in Portugal, Vodafone has a market share of around 30% of mobile subscribers and by identifying country and network operator of all subscribers whom connected to Vodafone cell sites, it is possible to extrapolate that count to an estimate of the population count.

✓ **Code testing the functions to ensure their functionality as expected:**

This third objective consisted in ensuring that the developed functions worked as expected. Performed in parallel with the development of the functions, this process consisted in manual testing and automated testing with the Python library called Pytest.

### **4.3. Business applications**

Location analytics has proven to be a key success factor for telecom operators to ensure their competitiveness in the sector. Since 2017, 70% of telecom companies believe that Location Intelligence plays an undeniable role in their business ongoing growth revenue strategies (Columbus, 2017). With numerous business applications to both internal and external purposes, telecom operators see new lucrative opportunities by generating meaningful location analytics. Telecom operators have tremendous data about how, when, with whom (anonymized) and where communications are made each day on their network.

To name some examples, internally, leveraging location data helps supporting marketing decisions (e.g. spatial variable to optimize marketing campaigns), network and infrastructure optimization (e.g. expand network capacity), market segmentation (e.g. cluster areas based on customers revenue), and so on. External purposes are usually categorized under the *Data monetization* hood, and refers to the process of providing spatiotemporal insights of people to third-parties (private and public organizations). As a fact, more and more organizations understand the value to take advantage of, in understanding the spatial distribution of people.

By anonymizing and aggregating customer location data and respecting enforced regulations related to the usage of such data (e.g. GDPR), telecom operators are able to freely monetize their data for location intelligence. The primary research project shows the process of leveraging spatiotemporal density distribution of people across the 24 parishes of Lisbon. Such spatiotemporal insights can have a wide array of useful applications to organizations. For example, businesses willing to know better their market in terms of spatial distribution, socio-demographic traits and travel patterns. To the public sector, location intelligence can assist or help developing smart cities strategies, tourist mining (e.g. where do tourists come from? How long do they stay? Where do they stay? How many are they? Where are their point of interest?), in measuring the number of people for a given event (e.g. football match at the stadium), or even in evaluating how people are moving in urban areas (e.g. by bus, by train, during weekday, weekend). This primary project goes in this direction; In the process of leveraging telecom data to be able to extract spatiotemporal density distribution at a specific spatial scale.

### **4.4. KPIs – What statistics to be extracted?**

The first step of the study, way before approaching the data, was about the KPIs that were to be developed to extract spatiotemporal insights from CDR along with available data sources. Before having clarified the goal of extracting specifically spatiotemporal density distribution, several directions were opted, based on the literature review, to retrieve human mobility statistics from mobile phone data. This initial step gave me a better sense of the data and its capabilities to extract spatiotemporal insights.

Considering the spatiotemporal capabilities of CDR, the time constraint to work on this project (around 2 months) and the recommendation of my supervisor, we decided to focus my learning in studying the process of extracting spatiotemporal density distribution across the 24 civil parishes of Lisbon. Key Performance Indicators (KPIs) that translate CDR input into spatiotemporal statistics that answer, more generally the following type of questions:

- ✓ Where? (Location-wise - e.g. per parish, city center, Lisbon airport, football stadium)
- ✓ When? (Temporal-wise - e.g. Day-wise, hour-wise)
- ✓ Whom? (anonymized and aggregated)? (e.g. nationality-wise, tourist-wise)
- ✓ Why? ( Assumption related to the context, e.g. watch a football match at the stadium)

More specifically, the KPIs were thought to be (pythonic) functions parametrically designed to answer the following type of questions:

- ✓ Where are people during a time period?
  - E.g. at 8pm, a total of n tourists arrived at the Lisbon airport
  - E.g. At 6pm, a total of n people are at the stadium to watch a football match
- ✓ How is the population distributed geographically across time?
  - E.g. Count of people across the 24 parishes of Lisbon across days or hours
- ✓ Where are the residential and working areas?
  - E.g. residential areas considering interactions late in the night and working areas considering interactions during weekday during daytime
- ✓ Where are the points of interest per time period? (e.g. Shopping mall)
- ✓ What was the impact of special event? (e.g. Football match at the Stadium)
- ✓ Where do tourists arrive and leave? (e.g. train station, Lisbon airport)
- ✓ Where do tourists sleep, go and stay most time?
- ✓ What is the national diversity across specified locations (e.g. parishes)

### 5. Data Sources

This section describes in detail the four specific data sources, provided by my supervisor, that were necessary to perform the human mobility analysis and the development of the KPIs and functions. Below are listed the data sources, followed by a description of each of them:

- ✓ A week of anonymized call detail records (CDR) logs in a csv file format;
- ✓ List of Vodafone’s cell sites along with their configuration detail in a csv file format;
- ✓ List of country and operator names along with their MCC and MNC in a csv file format;
- ✓ List of 24 Lisbon parishes with their respective geometric shapes as a shapefile format;

#### **Important notice related to the data description in this report:**

Although Call Detail Records and Vodafone cell sites are described as they were in the studied datasets, synthetic toy data are used for all specific examples and visualizations.

#### **5.1. Call detail records dataset**

The studied dataset contained the anonymized mobile phone activity records (CDR) carried by Vodafone that took place in the city of Lisbon during a time span of 7 days. The CDR logs concerned the following two type of mobile subscribers that connected to Vodafone cell sites: (1) Mobile Subscribers with a Portuguese Vodafone SIM card and (2) Mobile subscribers with a non-Portuguese SIM card but roaming in Portugal and accessing the cellular network through Vodafone cell sites (collaboration between different telecom operators aka roaming).

The dataset size was of approximately 6.5 gigabytes, with around +60 million logs/interactions and 11 features to describe every interaction. Although, out of the 11 features, only 8 were of interest in the human mobility analysis and are listed and described below: Start datetime, End datetime, Caller IMSI, Event\_ID, Site name, Site ID, City and Cell ID.

**Start and end datetime** determine when the interaction started and ended with this temporal hierarchy: *YearMonthDayHourMinutesSeconds*. **IMSI** being the subscriber ID, it has three components: MCC for the country, MNC for the operator, and MSIN was pseudonymized to anonymize the mobile subscribers. **Event\_ID** relates to the six following types of interaction:

- Inbound call (ID 1);
- Outbound call (ID 2);
- Location update (ID 3);
- Internet data connection (ID 4);
- Inbound text message (ID 5);
- Outbound text message (ID 6).

## CHAPTER 2. CDR-based location analytics

---

**Location update** is an interaction type that is made at each specific interval hour to update the location of every mobile subscribers. **Site ID** refers to the named location area (e.g. Ajuda), while **City** in our case remains static as being Lisbon. Finally, the **Cell ID** refers to connected cell site. When a call detail record is generated, the location of the connected cell site can be determined by crossing, the cell ID, with the one of the cell site dataset. (It is the following data source explained). Below is a toy example of a CDR log as it appeared in the source file:

Table 2: Example of a Call Detail Record log

Start Datetime	End Datetime	IMSI	Event ID	Site Name	Site ID	City	Cell ID
20180924202256	20180924202340	26801123456789	1	Ajuda	3GA	Lisbon	3DF1

As shown in the example, CDR by themselves do not contain sufficient information to perform any spatiotemporal analysis, as records lack the spatial aspect by only referring the connected cell site by their ID. As such, information about all cell sites of Lisbon have been gathered in another dataset to properly identify and further map out the network coverage of each cell sites.

### **5.2. Vodafone cell sites dataset**

Along with the CDR, the Vodafone cell sites dataset was essential to perform the human mobility analysis, as together they gather for each subscriber's interaction, the datetime and location of connected cell site. This studied dataset listed thousands of active cell sites scattered across the city of Lisbon, that are described by a set of 9 features. Hereafter is a description of each feature, a toy example of a cell site as it appeared in the dataset and finally the visualization of those cell sites (toy data for visualization).

- Cell ID;
- MCC;
- MNC;
- LAC;
- Longitude;
- Latitude;
- Radius\_Real;
- Azimuth;
- Hbeam.

Table 3: Toy example of a cell site from the cell site dataset

Cell ID	MCC	MCN	LAC	LON	LAT	Radius_real	Hbeam	Azimuth
3DD2000A	268	1	350	-9.15	38.72	500	359	0

**Cell ID** is the reference used to identify each cell site and as a cross-reference with the CDR to be able to determine the location of connected cell sites. **MCC and MNC** are both components used to identify the country and network operator of the cell sites. Although, as all cell sites belong to Vodafone and are located in Lisbon, those two values were fixed and thus



## CHAPTER 2. CDR-based location analytics

discarded as of no further use. **LAC**, also called Location Area Code refers to the code given to set of cell sites that are grouped together based on their location. **Latitude and Longitude** are geographic coordinates to pinpoint cell sites' exact location. **Radius\_real** refers to cell site antenna's network distance coverage (e.g. 2 kilometres from Antenna), **Azimuth** corresponds to the faced direction (e.g. North as  $0^\circ$  or South as  $180^\circ$ ) and finally, **Hbeam** refers to the signal openness which usually vary between  $60^\circ$  to  $359^\circ$  being omnidirectional. Below are two visualizations using toy data, to represent the cell sites by their longitude and latitude. The first visualization is of all cell sites, while the second visualization is of all cell sites onto the map of Lisbon.

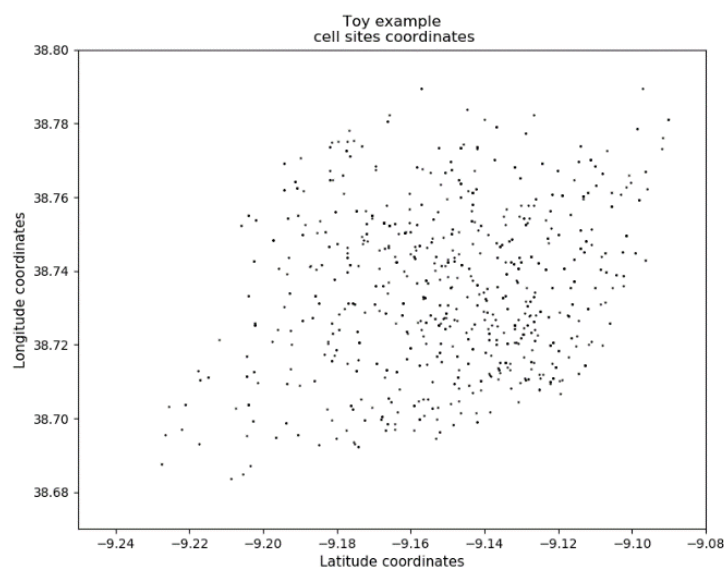


Figure 6: Cell sites coordinates

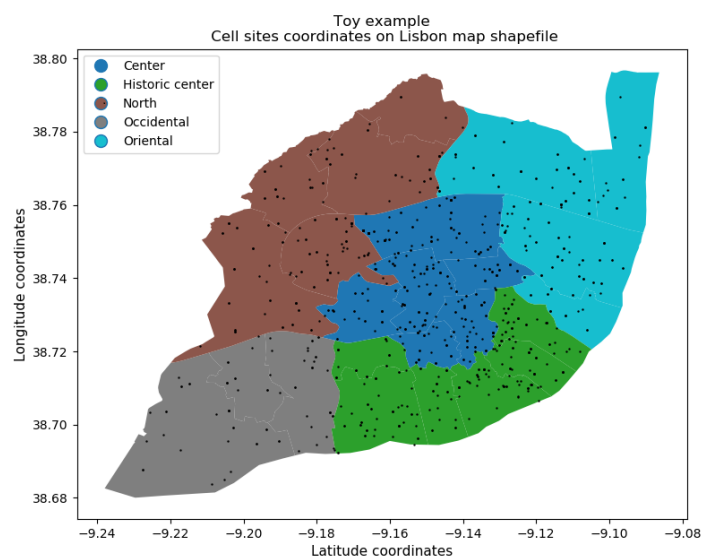


Figure 7: Cell sites coordinates on top of the map of Lisbon

## CHAPTER 2. CDR-based location analytics

It is important to highlight that the thousands cell sites are scattered across hundreds of locations (i.e. cell towers, on top of building). Indeed, it is common for several cell sites to be collocated on a same structure and thus implies that, in our case, most cell sites share identical coordinates with few others. Besides that, urban areas such as Lisbon, counts numerous cell sites scattered across the city. It is why, all cell sites cannot be clearly distinguished on the above two graphics, as many cell sites share identical coordinates or are closely located to others. Lastly, all cell sites are thought as a network and thus are usually set with different and/or complementary parameters to not interfere with each other, e.g. facing direction.

### 5.3. Lisbon parishes shapefile

As a shapefile format, this dataset contains the geospatial data representation of the 24 Lisbon parishes. Every parish of the dataset is represented by a set of points (latitude/longitude) grouped together to form the corresponding parish as a polygonal shape. In this research, this dataset was used for visualization purposes as well as for determining the overlapping proportion between the cell sites network coverage and the Lisbon parishes. (It is explained in much depth later in the report). Below is a visualization of all parishes, with the example of the data representation of the parish “Olivais”.

Table 4: Data representation of parish Olivais from the shapefile

Lisbon Parish	Polygon geographic representation
Parish_Name	POLYGON ((Tuple of latitude longitude,.....))
Olivais	POLYGON((-9.13377777 38.78487227, -9.13138887 38.7843483,...))

Lisbon counts the following 24 civil parishes in its municipality:

Belém, Ajuda, Alcântara, Estrela, Campo de Ourique, Misericórdia, Santa Maria Maior, São Vicente, Penha de França, Benfica, São Domingos de Benfica, Carnide, Lumiar, Santa Clara, Campolide, Santo António, Arroios, Areeiro, Avenidas Novas, Alvalade, Beato, Olivais, Marvila and Parque das Nações.

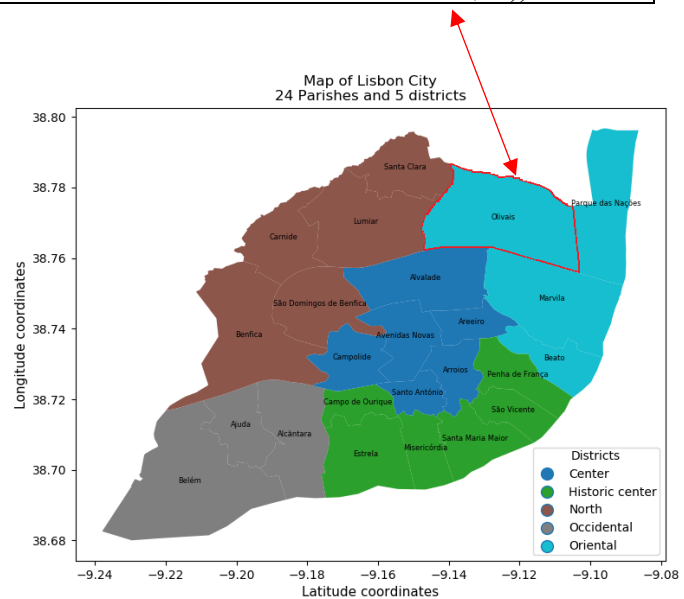


Figure 8: Map of Lisbon: Olivais parish with red bordered

## **5.4. Mobile Network operator dataset**

This dataset gathers an exhaustive list of Mobile Country Code (MCC) and Mobile Network Code (MNC) with along the corresponding country and mobile network operator. Below are 8 examples of the dataset:

Table 5: Mobile Network operator dataset

<b>Country</b>	<b>Network Operator</b>	<b>MCC</b>	<b>MNC</b>
Portugal	Vodafone	268	01
Portugal	MEO	268	06
Belgium	Proximus	206	01
Belgium	Base	206	20
Italy	Vodafone	222	10
Italy	Wind	222	88
Spain	Orange	214	03
Spain	Vodafone	214	01

As already explained in the section *Essential Telecom GIS concepts*, IMSI identifier has three components: MCC (Mobile Country Code), MNC (Mobile Network Code) and MSIN (Mobile Subscriber Identity Number). Every call detail record contains the IMSI to identify each mobile subscriber. As shown above, MCC is the code that relates to the country, while MNC relates to the network operator.

Besides that, as already explained, IMSI is a unique identifier stored in every SIM card used to identify any mobile subscriber, as for any generated call detail records. With the assumption that mobile subscribers are in possession and in use of the SIM card from the country in which they live, it is possible to derive from their generated CDR, where they come from (e.g. MCC of 206 for Belgium) or even to which network operator they belong to (e.g. MNC of 01 for Vodafone).

This dataset was used as an additional data source to develop the KPIs and functions, mostly for filtering and aggregating mobile subscribers by country. Let's consider the tourist mining example, one could filter the CDR by their MCC as not being equal to 268, where 268 corresponds to Portugal, as to extract the statistics of interest for this subset of mobile subscribers.

## 6. Methodology, tools and approach

### 6.1. Approach

The methodological flow started with the research definition, which encompassed goal, objectives and a primary understanding of the ‘business domain’ related to the usage of call detail records to study human mobility and generate location intelligence insights. It is followed by four main stages that are described hereafter:

**1. CDR dataset analysis** – This primary stage describes the regulations related to the use of CDR for analysis purposes as well as the required pre-processing step due to the underlying characteristics of CDR dataset being heavy in size. This dataset including more than 60 million logs, and of 6.5 gigabytes passed through a filtering step to ease further explorations and analysis. The dataset included CDR logs of a timespan of 7 days and were thus splitted into 7 smaller csv file, each gathering the logs that took place during a day of the week.

**2. Geospatial upper scaling approach** – Given the low spatial resolution of CDR referring to connected cell sites instead of subscribers’ actual location, this core stage of the project research relates to the process of filling this gap by performing twofold: (1) attributing a coverage network to all cell sites, assuming that subscribers connecting to a cell site were within its coverage network at the time of the call detail record, and then (2) identifying subscribers’ location at the parish spatial scale to reach the core goal of extracting spatiotemporal density distribution across the 24 parishes of Lisbon.

**3. Subscribers’ density distribution KPIs and functions development** – Now that the CDR dataset has been pre-processed and that the spatial resolution of CDR has been adjusted to the parish scale, KPIs and additional functions were developed to extract from those call detail records, the statistics of interest. This stage describes 4 of the functions developed. Two related to the extraction of spatiotemporal density distribution of mobile subscribers whom connected to Vodafone cell sites across the 24 parishes of Lisbon. The two others are additional functions to retrieve further insights about the mobile subscribers nationality-wise.

**4. Code testing to ensure the functions functionality** – This stage was performed in parallel with the development of the functions and consisted in ensuring that they work as expected by use of manual testing and automated testing.

### **6.2. Tools**

First of all, the entire project was carried out using the computer that I was provided. All steps of the analysis were performed using a single local machine. The data analytics tools used to perform the entire research relied entirely on Python, a general-purpose programming language, for all tasks of data analysis and QGIS, a geographic information system software, specifically for analysing, editing and visualizing geospatial type of data. Both software are known as free, open-source, cross-platform and with a thriving user-based support giving them a considerable edge over most data analytics tools in their genre. No tools to extract the data are listed. The different data sources needed for the analysis have been delivered by my supervisor, whom has collected the needed data from the Vodafone Big Data Platform with the use of Apache Hive big data tool, commonly used for data query and analysis.

Python programming language has taken its place for almost all data analysis tasks, as for reading, exploring, wrangling, automating, visualizing, testing and much more thanks to its rich ecosystem including vast and diverse libraries and simplicity for writing/reading codes. Libraries such as *pandas* and *NumPy* were considerably used for most of data wrangling, *geopandas*, *geopy* and *shapely* to make geospatial data analysis easier in python, *matplotlib* and *seaborn* to represent the data visually with high-quality figures, *pytest* to write automated test and ensure code functionality or even *yaml*, a data serialization language to write structured data and well-suited for configuration files.

QGIS, a geographic information system (GIS) application that gives users a lot of flexibility when it comes to work with geospatial data. It includes a plethora of GIS functionality to perform a variety of functions such as analysing or editing spatial information, producing or exporting maps or supporting a multitude of vector data formats (e.g. ESRI shapefile). Vector data is used to represent geographic objects by means of points, lines and polygons that are set of geographical coordinates. In this research, QGIS has been used to perform some visualization but principally to calculate distances between geographic locations, area attribution and determining overlapping proportion between cell site network coverage and parishes of Lisbon.

On the next page, the figure shows the QGIS interface with the visualization of cell sites scattered across the city of Lisbon.

## CHAPTER 2. CDR-based location analytics

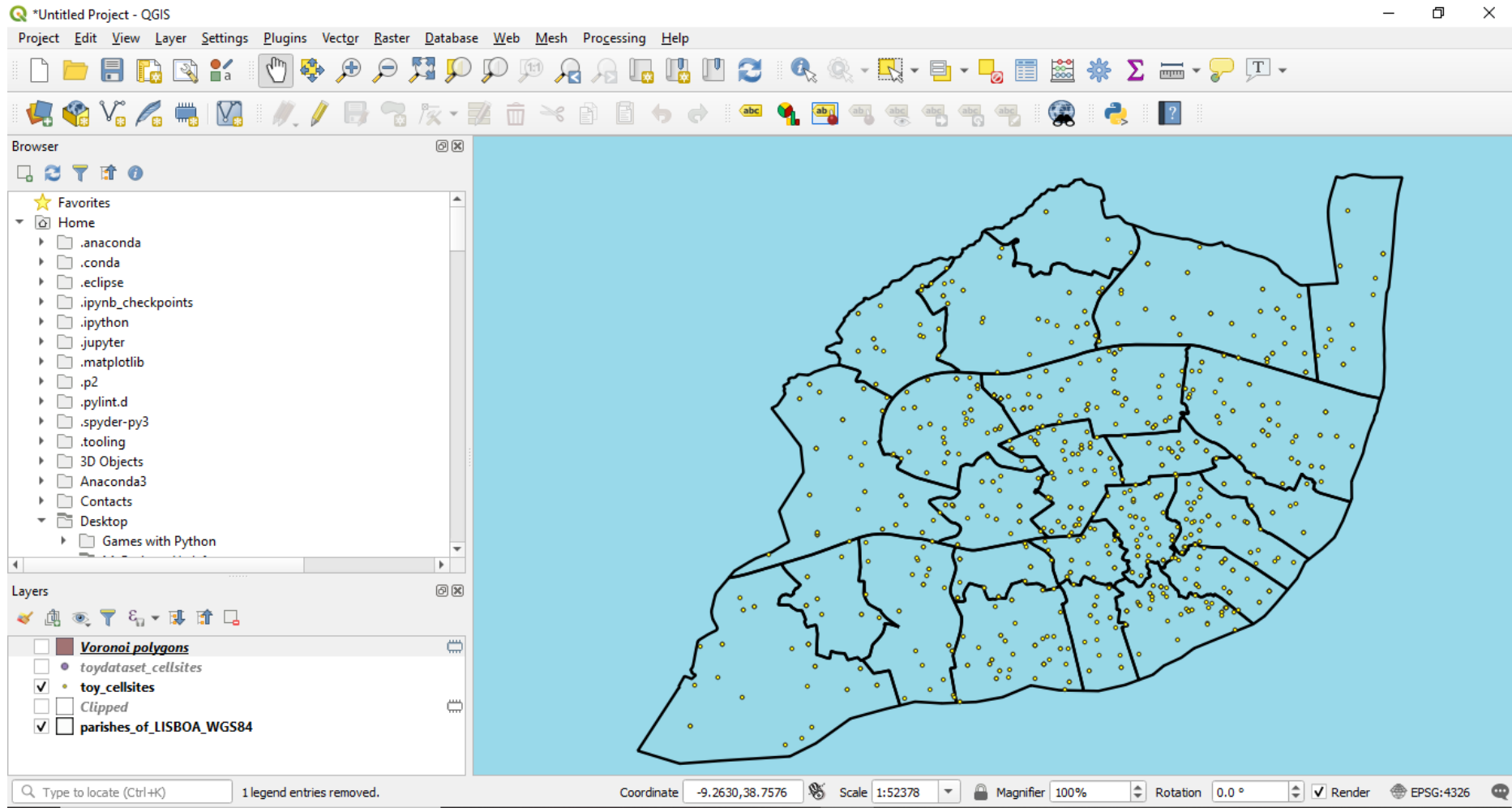


Figure 9: QGIS Software graphic user interface with visualization of cell sites (yellow points) onto the map of Lisbon

### **6.3. CDR dataset analysis**

This sub-section describes the primary steps related to the CDR dataset which encompasses two main parts: (1) GDPR & Vodafone regulation and (2) its required pre-processing step.

#### **6.3.1. GDPR & Vodafone regulation**

Call detail records (CDR) are instances of mobile subscribers phone activity. It is why, the EU General Data Protection Regulation (GDPR) has enforced limitations related to the analysis of CDR logs to ensure anonymization and data privacy of mobile subscribers. GDPR regulations have been ensured by Vodafone way before I get to see the data. The IMSI (International Mobile Subscriber Identity) is used to specifically identify each mobile subscriber's call detail records. As already explained, IMSI has three components: MCC and MNC are kept as they are, as they both refer to the SIM-card originated country and activated network operator. Although the third component, MSIN, which refers to the identifier number has passed through an encryption, anonymization step to ensure the non-recognition of any mobile subscribers.

Besides that, the data privacy officer of Vodafone Portugal has added other rules related to the analysis of call detail records. The temporal granularity of call detail records (start/end datetime features) passes from *YearMonthDayHourMinutesSeconds* to *YearMonthDayHour*. In other words, for any call detail records generated, minutes and seconds are to be removed and not used for analysis purpose. This limitation restrains the profoundness in the spatiotemporal analysis of CDR. Let's consider a toy example:

Table 6: Toy example of a Call Detail Record log

<b>Start Datetime</b>	<b>End Datetime</b>	<b>IMSI</b>	<b>Event ID</b>	<b>Site Name</b>	<b>Site ID</b>	<b>City</b>	<b>Cell ID</b>
20180924202256	20180924202340	2681...	1	Ajuda	3GA	Lisboa	3DF1

From this toy example, the temporal information retrievable are that this mobile subscriber started the interaction at 8pm and finished it as well at 8pm, the 24<sup>th</sup> September 2018. In other words, the temporal granularity of CDR can be performed across hours and not anymore across minutes nor seconds. It implies that actual event time cannot be determined nor its duration. As explained in the next part, End datetime with along other features are removed from the CDR dataset, to lightweight its current file size (6.5 gigabytes).

### 6.3.2. CDR dataset pre-processing

Call Detail Records data are undoubtedly a very valuable source of information for telecom operators, but they also require Big Data technologies for collection, storage, processing and analysis. Indeed, ordinary computers or database systems are unable to process efficiently within an acceptable timeframe ongoing generated call detail records. Vodafone knows that and process their data coming into the organization with the Hadoop Big Data technology. In short, the Hadoop framework enables to process the data faster and efficiently by distributing the storage and processing across clusters of physical machines.

In this research, the studied CDR dataset is saved as a csv file, and the entire research is performed on a local machine. This dataset counts more than 60 million logs for a rough file size of 6.5 gigabytes. Subsequently, as working on a single machine, this CDR dataset had passed through a primary pre-processing step to ease processing and analysis of its content.

This pre-processing step consisted in three main operations:

1. Based on the enforced Vodafone regulation regarding the spatial granularity of call detail records, delete for all CDR the minutes and seconds of the start datetime feature;
2. Removal of features that aren't to be used in further steps of this research:
  - ✓ End Datetime, Site Name, Site ID and City
3. Splitting the CDR dataset into 7 smaller dataset, each for a day of the week.
  - ✓ The core CDR dataset contains logs throughout the span of 7 days in Lisbon.

This pre-processing step considerably ease computations for all further steps of the research by splitting the main CDR dataset into 7 smaller csv.file, each compiling all CDR of a day of the week.



## 6.4. Geospatial Upper scaling approach

This sub-section is directly in line with the spatial limitation of call detail records, which do not refer to the actual location of subscribers but rather to the location of connected cell site. To recap, the research goal consisted in extracting spatiotemporal density distributions of mobile subscribers across the 24 parishes of Lisbon, based on their interactions with the cellular network, saved as call detail records (calls, text messages, internet data connection or location update). As such, the objective of the upper scaling approach consisted in improving the spatial scale of call detail records, passing from connected cell site location to the parish location. To do so, the upper scaling approach was performed by means of two successive steps:

1. Determining the cell sites' network coverage (assumption that subscribers are within it, when connecting to cell sites, for accessing the cellular network);
2. Upper scaling alignment between cell sites' network coverage and overlapping parishes (Probabilities to be in certain parishes based on connected cell sites).

To ease the reading, this sub-section is divided into the following four parts:

- 6.4.1. Review of basic spatiotemporal capabilities retrievable from call detail records;
- 6.4.2. Core spatial assumption to perform the geospatial upper scaling;
- 6.4.3. 1<sup>st</sup> step Geospatial upper scaling - attributing a network coverage to all cell sites by applying the Voronoi diagram mathematical concept;
- 6.4.4. 2<sup>nd</sup> step Geospatial upper scaling - from cell sites' network coverage to Parishes.

A picture is worth a thousand words. On the right, is a visualization of the entire upper scaling process of one cell site. In yellow is the cell site location, dashed red its attributed network coverage and in blue the delimitation of the parish Belém. In a nutshell, mobile subscribers only connecting to this cell site have a probability of 1 (100%) to be in Belém at CDR timestamp.

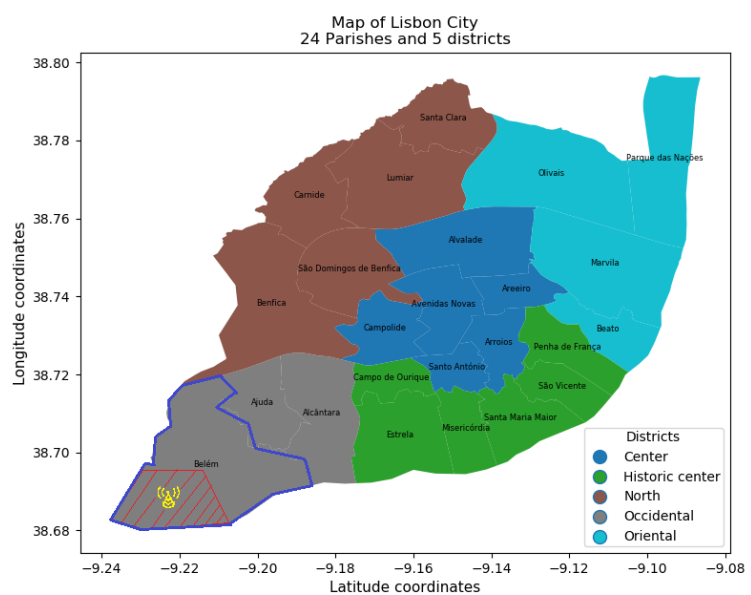


Figure 10: Upper scaling approach of one cell site

### 6.4.1. Basic spatiotemporal capabilities of call detail records

Theoretically, call detail records contain spatial and temporal information, connected cell ID and start datetime, respectively. Although, CDR spatial aspect only refers to connected cell sites by their ID instead of coordinates. Crossing the CDR with the cell site dataset, through the common cell ID, gives the basic spatial resolution to determine coordinates of connected cell sites, as illustrated below with a toy example:

Table 7: CDR log crossed with cell site dataset to localize the connected cell site

Toy Call Detail Record				
Start datetime	IMSI	Event_ID	Cell_ID	Other features...
20180924202256	26801123456789	2 (Outbound call)	<b>3D0A</b>	....

Toy cell site from cell sites dataset			
Cell ID	Longitude	Latitude	Other features...
<b>3D0A</b>	-9.1575..	38.74...	....

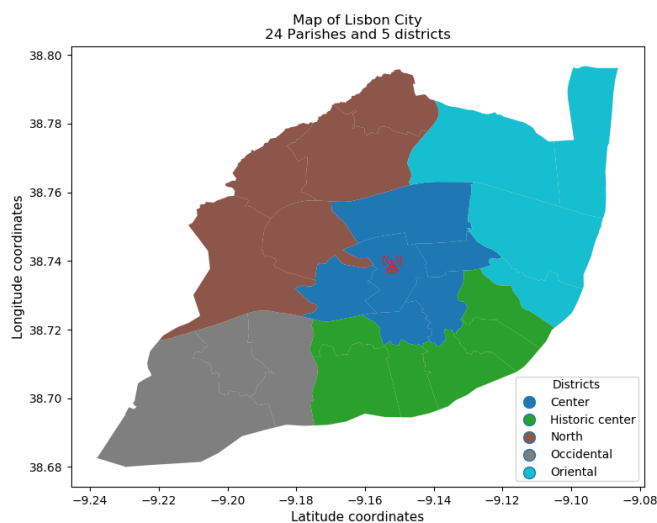


Figure 11: Cell site location, a mobile subscriber connected to access the cellular network

From this toy example, we can pinpoint that subscriber with IMSI 26801123456789 connected to the cell ID **3D0A** to make a call, at 8pm on the 24<sup>th</sup> September 2018. Regarding the enforced Vodafone regulations for the analysis of call detail records, minutes and seconds cannot be analyzed and thus imply that end datetime is of no use and that, interaction duration cannot be determined. Based on the IMSI, we can highlight that this subscriber may come from Portugal considering that its SIM-card MCC is 261, which is the country code for Portugal.

### 6.4.2. Spatial assumption to perform the geospatial upper scaling

As said, the creation of a call detail record implies that a subscriber has connected to the cellular network and thus had to pass through a cell site. This fact led to the assumption of a spatial relation between mobile phones and potential cell sites they could connect to, for accessing the cellular network. As such, two main assumptions came up and are described hereafter:

#### Based on cell sites technical parameters:

One assumption was based on the technical parameters of Vodafone cell sites and their attached antennas. Predominantly, antennas can have different signal ranges varying from a hundred meters to more than a few kilometres. Although, this assumption was left out because of its vague estimations with overlapping cell sites network coverage, considering that urban cities counts highly numerous cell sites. For Lisbon itself, Vodafone counts thousands of active cell sites scattered across hundreds of location in the city. Thus, with this optic, mobile subscribers could hypothetically be as far as a few kilometres when connecting to a cell site.

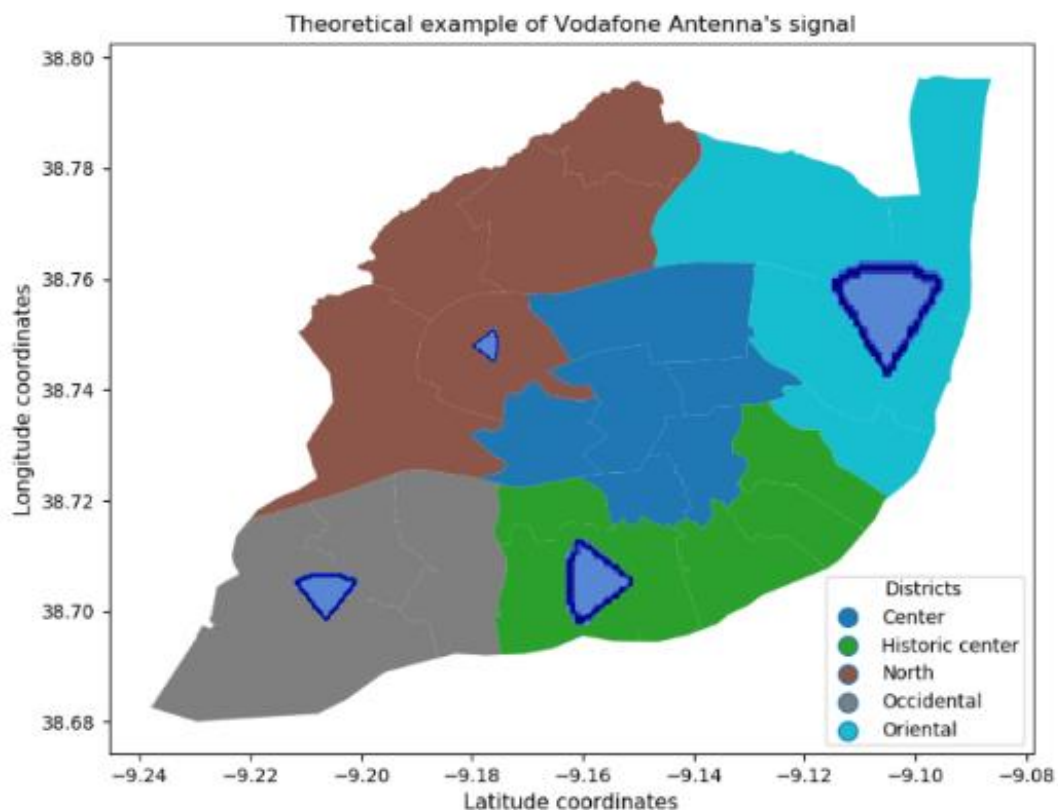




Figure 12: Map of Lisbon with representation of cell sites' antennas signal ranges in blue

**Based on mobile devices technical parameters :**

Looking at the other way around, based on the technical assumption that cell phones typically connect to antennas of the nearest cell site, and if not, to another close one, subscribers' locations can be approximated from CDR, as being in an extent area around the connected cell site. Besides that, even if cell sites' antennas are directional (azimuth) and have different signal openness (Hbeam), they, by default send/receive signals omni-directionally to all nearest connecting devices. In our research case, Lisbon is an urban area which counts numerous cell sites scattered across Lisbon. These aspects are in accordance with the approach of defining to every cell sites, an omnidirectional network coverage encompassing all coordinates with less distance than to any other cell sites. Some published papers and researches hold on this assumption to determine the subscribers' potential location based on connected cell sites. As such the considered approach was of dividing Lisbon into smaller regions, where each region is the theoretical network area of each cell site. Below are two examples of two different antennas' signal openness, on the left, being at 90° and on the right being 360°. As shown, even the one with 90° openness still send/receive signals in all direction to nearest devices.

Table 8: Example of cell sites' antennas with in red the signal ranges

Directional antennas' cell site (90°)	Omnidirectional antennas' cell site (360°)
	

This second assumption stemmed as the basis of further steps for the geospatial upper scaling approach to attribute a theoretical coverage network to all cell sites. The main assumption was that mobile subscribers connect to the nearest cell site for accessing the cellular network.

**6.4.3. Cell sites network coverage attribution**

As explained above, the first step towards a finer geospatial granularity level for the CDR was of attributing a network coverage to all cell sites. We can assume that any subscribers connecting to a cell site, are within the cell site network coverage at the time of the interaction with the cellular network. The applied technique to put into application the former assumption was performed with the mathematical concept called Voronoi diagram. On the next page, the theoretical aspect of this concept is primarily explained, before describing its application.

### Voronoi diagram in theory:

The Voronoi diagram is the mathematical concept used to put into application the former theoretical idea of attributing to every cell site a network coverage that encompasses all coordinates with minimal distance to reach its cell site than to any other site. This technique is based on the minimal distance to reach a landmark (here, a cell site).

More formally, a Voronoi diagram is the partitioning of a plane (e.g. Lisbon) into convex polygons (e.g. network coverage) based on a set of finite number of seeds (e.g. cell sites), where each polygon contains exactly one seed (e.g. non overlapping network coverage). The line segments of the below diagram are all the points that are equidistant (equally far) to the two closest seeds. It implies that any point inside a region is less distant to the inside seed than to any other seed of the plane. In our application, distances are measured using the Euclidean distance.

The illustration below shows the process of transforming a plane with four cell sites (left) into a Voronoi diagram (right) with the underlying pattern of adding line segments to the plane (coloured in black). The red and yellow lines highlight that the black line segments are equidistant and perpendicular to the two closest cell sites, respectively.

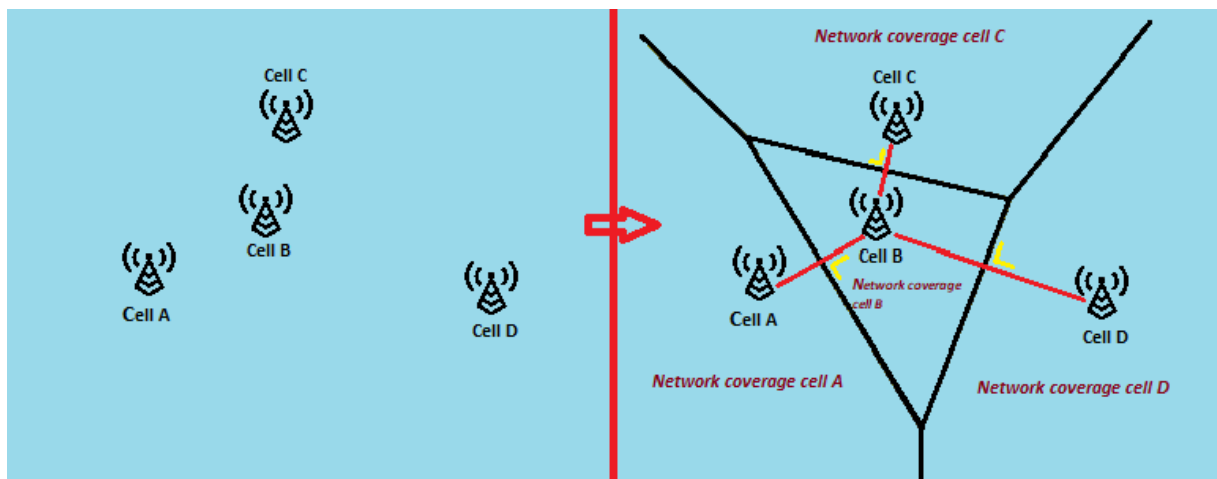


Figure 13: Voronoi diagram application into an Euclidean plane of 4 cell sites

This process changes the geospatial granularity of CDR, passing from the cell site location knowledge to the approximate subscribers' location knowledge. On the left side of the figure, location of mobile subscribers cannot be estimated based on their CDR, while on the right side, it is assumed that mobile subscribers are within the network coverage of the connected cell site at the interaction time with the cellular network.

### **Voronoi diagram into application:**

Now that the assumption has been established and the process of applying it has been understood, it is time to apply it on our data. In our application, applying the Voronoi diagram based on the thousands Vodafone cell sites scattered across hundreds locations around Lisbon has passed through a series of steps on Python and QGIS, as explained below:

#### **1. Cell sites are moved in their facing direction (azimuth) by a distance of 10 meters:**

The first step was performed on Python and concerned the fact that the thousands Vodafone cell sites are collocated on hundreds of structures across Lisbon. It implies that many cell sites are collocated on structures with a few others, and thus, that those cell sites share identical coordinates. The limitation encountered was that the Voronoi diagram is constructed based on Euclidean distances between the cell sites. Developing the Voronoi diagram like that would imply that for each structure counting more than a single cell site, only one of them would be used to construct the Voronoi diagram and only this one will be attributed a *Network coverage*. Although, it was not possible to attribute unique coordinates to every cell sites, it was possible to maximize this count. Cell sites are usually collocated on same structure to improve the cellular coverage and implies that they often have complementary settings, mostly in terms of, signal strength, facing direction (azimuth), signal openness (Hbeam).

With the objective of attributing distinct coordinates to most cell sites, all of them were moved by a distance of 10 meters in their facing direction (azimuth). This distance was decided based on past experimentations performed by my supervisor. Moving all cell sites by such minimal distance didn't create noise in the attribution of network coverage and allowed developing a finer geospatial granularity. This step attributed distinct coordinates to around hundreds more cell sites. As it is explained later in much detail, for the cell sites still sharing identical coordinates, they were attributed the same network coverage. This manner enabled identifying the subscribers' location when connecting to any of the Vodafone cell sites, whether used to construct the Voronoi diagram or not.

#### **2. Save the updated cell site dataset as a shapefile, a geospatial vector data format:**

To pursue with the process, the second step consisted in saving as a shapefile, a geospatial vector data format, the updated cell site dataset (with all cell sites moved by a distance of 10 meters in their facing direction). This data format change was necessary as the construction of

## CHAPTER 2. CDR-based location analytics

---

the Voronoi diagram was performed on QGIS and this software reads geospatial data format such as shapefile.

Thus, this step consisted in three main operations: (1) Create the feature *Coordinates* for all cell sites, being the combination of two other features latitude and longitude. (2) With the Geopandas python library, create the geospatial vector feature, called *Point*, which corresponds to the newly created feature *Coordinates* but under specific geospatial format. (3) Finally saving the cell site dataset as a shapefile. Hereafter is a toy example of all features listed with along example of plausible values:

Table 9: Shapefile of cell sites dataset – Example of its content

<b>Cell_ID</b>	3DD2000A
<b>LAC</b>	350
<b>Longitude</b>	-9.15
<b>Latitude</b>	38.72
<b>Radius_Real</b>	500
<b>Hbeam</b>	359
<b>Azimuth</b>	0
<b>Coordinates</b>	(-9.18843262560578, 38.74452778)
<b>Point (required to create shapefile)</b>	POINT (-9.18843262560578, 38.74452778)

### 3. Perform the Voronoi diagram on QGIS software:

QGIS can take as input shapefile format, to read the coordinates of the cell sites with the geospatial feature, *Point*. This software has the feature to create the Voronoi diagram based on calculating distances between those *Point*, each being a cell site, in our case. Internally, it calculates the Euclidean distances between every cell points to create the boundaries and thus attribute to those cell sites a *Network coverage*, with a given geospatial polygonal shape. The latter implies clearly that for all cell sites with identical coordinates with others (collocated on a same structure), only a single one of them will be attributed automatically a *Network coverage*. The solution to fill this gap is explained later in the process, by creating a mapping between cell sites used for the Voronoi diagram with others through their common coordinates.

A picture is worth a thousand words. To illustrate the process performed on QGIS software, below are the visualizations with along comments to explain the different steps. The initial step consisted in loading and visualizing the inputs on QGIS; the cell sites coordinates are represented by the geospatial feature *Points*. Thus, this first visualization is the QGIS desktop,

## CHAPTER 2. CDR-based location analytics

with on the right side, on the blue canvas, the cell sites. All cannot be easily distinguished as many cell sites still share identical coordinates or are very near with other cell sites.

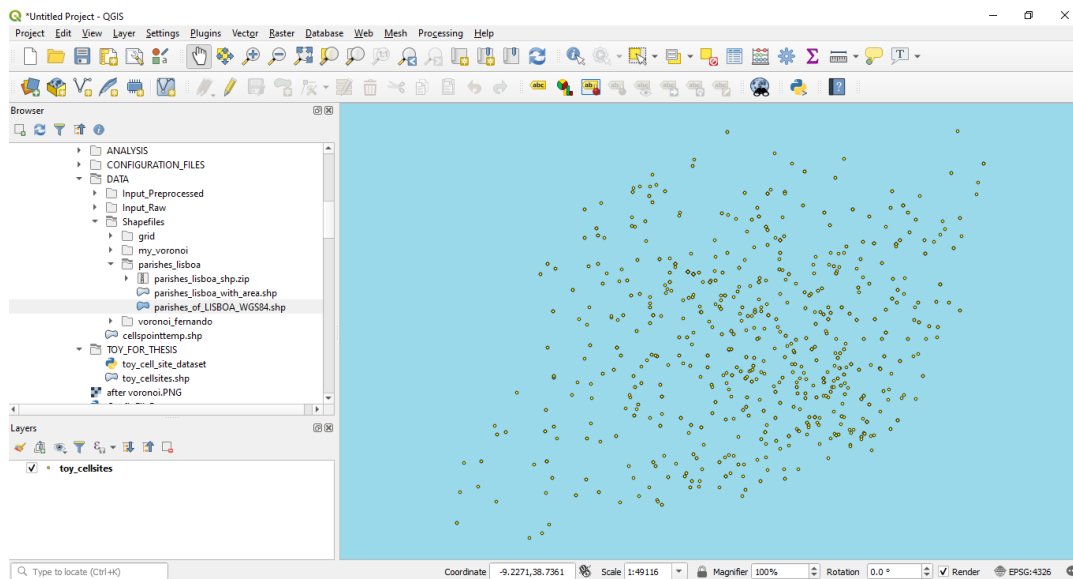


Figure 14: Step 1 – Load cell sites as inputs on QGIS

The following visualization shows the output Voronoi diagram applied on chosen cell sites. As said earlier, for cell sites sharing identical coordinates, only one of them was chosen to be attributed a *Network coverage*. Thus, for the chosen cell sites, they were attributed a geospatial *Polygonal* shape. These polygonal shape were simply combination of all coordinates defining the boundaries of the cell sites' network coverage. The external boundary of the visualization was user-defined to delimit the shape of Lisbon city.

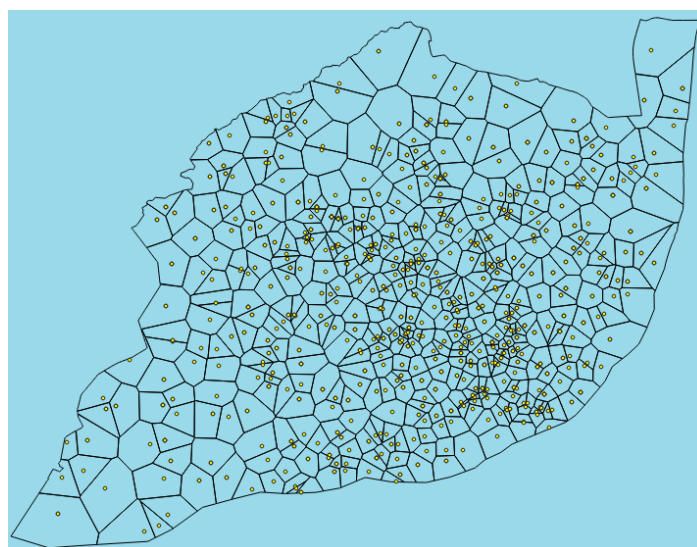


Figure 15: Step 2 – Create Voronoi Diagram based on selected cell sites (one per coordinate)



Looking at the other way around, the below graphic demonstrate the process of dividing Lisbon into smaller regions based on the cell sites attributed *Network coverage*.

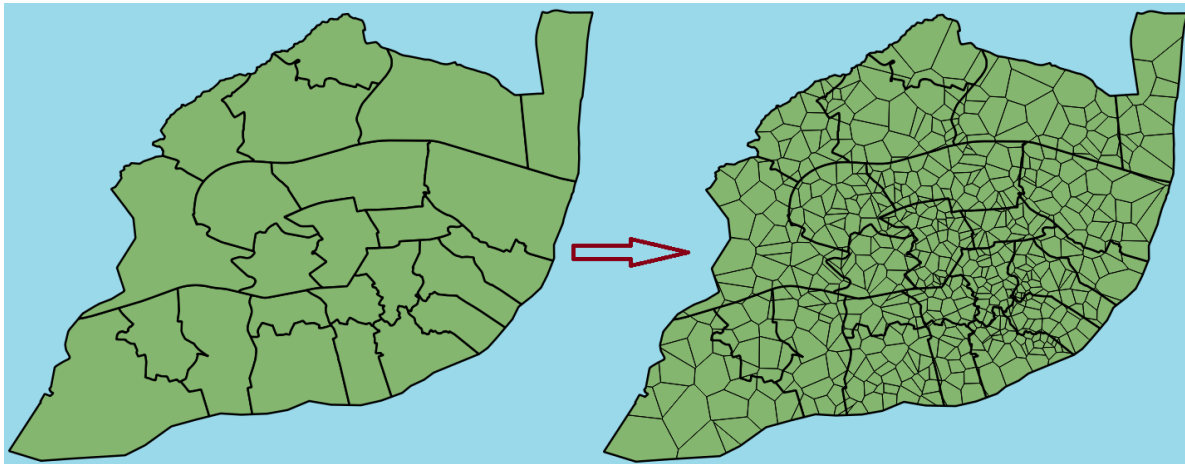


Figure 16: Downscaling of Lisbon city from parishes to cell sites network coverage

#### 4. Save the Voronoi diagram output as a shapefile format file:

Thus, after having performed the Voronoi diagram on QGIS based on Vodafone cell sites, the output was saved as a shapefile, which basically contained four features as illustrated below:

Table 10: Example of Voronoi diagram output saved as a shapefile

Voronoi_Cell_ID	Latitude	Longitude	Geometry
3AD1	38.69	-9.20	POLYGON((-9.188, 38.744,...,series of tuples...))

This output shapefile didn't keep features other than these four. Out of the thousand cell sites, this shapefile counted hundreds of cell sites with a newly attributed network coverage polygon (geometry).

Indeed, it is explained by the fact that the thousand cell sites were scattered across hundreds of locations, and that only a single cell site per coordinates was used to construct the Voronoi diagram. As such, this shapefile includes the ID of the cell sites used to develop the Voronoi, with along their correspondent latitude, longitude and geospatial feature called Geometry, which refer to polygonal shape representing the boundaries of their attributed network coverage.

The issue encountered at that stage was that we have the *Network coverage* only for the cell sites used to generate the Voronoi and not for all the others. The solution to fill this gap is explained in the next point.

**5. Mapping between Voronoi cells sites and the other cell sites:**

A mapping was created between cell sites used to construct the Voronoi diagram, with the non-selected cell sites based on identical latitude/longitude. The output mapping looks as follow:

Table 11: Mapping between Voronoi cell sites and the other cell sites

Voronoi Cell_ID	Non_Voronoi Cell_ID	Latitude	Longitude	Geometry
3AD1	1ED4	38.69	-9.20	POLYGON((( -9.18848, 38.74458,...))

This step performed; a network coverage was attributed to all cell sites of Vodafone located in the city of Lisbon. With the assumption that mobile subscribers connect to the cellular network passing through the nearest cell site, it was then possible to estimate their location as being within the network coverage of the connected cell site at interaction time (CDR).

**6. New spatiotemporal capabilities retrievable from call detail records:**

Here, we reach the 1<sup>st</sup> geospatial upper scaling step which allows to determine the locations of mobile subscribers based on their call detail records, as being within the *Network coverage* of each connected cell site at interaction time. Below, a toy example is illustrated:

Table 12: CDR log crossed with cell site dataset to approximate the location of a mobile subscriber

Toy Call Detail Record				
Cell ID	Start datetime	IMSI	Event_ID	Other features...
3D0A	20180924202256	26801123456789	2 (Outbound call)	....

Toy cell site from latest updated cell sites dataset			
Cell ID	Longitude	Latitude	Geometry
3D0A	-9.1575..	38.74...	POLYGON((-9.188, 38.744,...))



Figure 17: Assumed location of a subscriber based on its call detail record (reded polygon)

## CHAPTER 2. CDR-based location analytics

---

Now from this toy example considering a single call detail record, we are able to assume that this subscriber was within the *Network coverage* of cell ID **3D0A** (reded polygon), at 8pm on the 24<sup>th</sup> of September 2018.

In summary, the first step of the upper scaling approach can be summarised in this newly cell site dataset, which contains the following features for all of the cell sites as shown below:

Table 13: Updated cell site dataset with polygonal shape of each cell sites

<b>Cell site with network coverage polygon shapefile dataset</b>	
<b>Features</b>	<b>Example of a toy cell site</b>
<b>Cell_ID</b>	3DD2000A
<b>Longitude</b>	-9.188...
<b>Latitude</b>	38.744...
<b>Geometry</b>	POLYGON((-9.188..., 38.744,...))

At this stage, it is already possible to retrieve a series of insights about the way people move around the city based on their interactions with the cellular network, saved as call detail records. Although, the geospatial upper scaling approach didn't stop at this stage. Those *Network coverage* polygons attributed to cell sites cannot directly determine in which parish subscribers were based on their call detail records.

It is why, the next page describes the process of probabilistically determining in which parish mobile subscribers were based on the newly developed spatial scale. In a nutshell, this following process is based on the overlapping proportion of cell sites' network coverage with the Lisbon parishes.

#### 6.4.4. From cell sites network coverage to Parishes

The process of determining in which parish mobile subscribers were, based on (1) their call detail records and (2) connected cell sites' network coverage was mainly developed on probability. Indeed, by looking at the below Voronoi diagram overlapping the map of Lisbon, one can distinguish two possible scenarios in which it is possible to infer in which parish subscribers were based on the knowledge of connected cell sites' network coverage. These two scenarios are cited hereafter:

- ✓ A cell site' network coverage is fully within a parish;
- ✓ A cell site' network coverage overlaps at least two parishes (two or more).

Thus, the applied idea was of attributing probabilities to all cell sites based on the overlapping proportion between the cell sites' network coverage and the parishes with which they overlap. This stage was performed on Python, by calculating overlapping proportions between cell sites' network coverage polygons and the 24 parishes polygons of Lisbon. Below, three possible cases are exemplified to understand the probability attribution: overlap 1, 2 or 3 parishes.

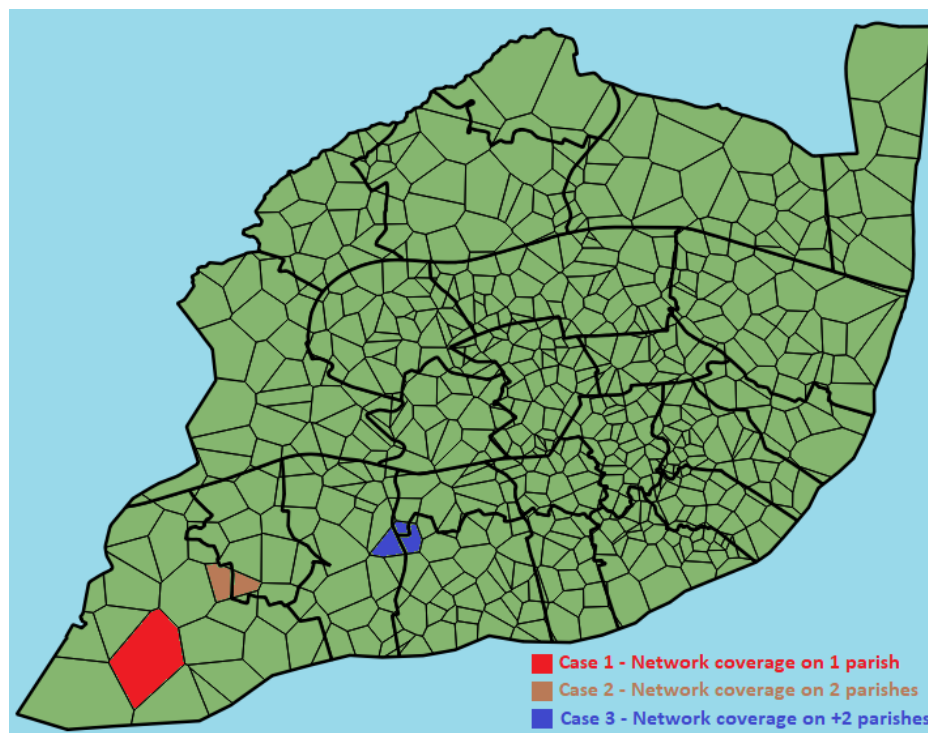


Figure 18: Map of Lisbon with three cases of cell sites' network coverage (in red, brown and blue). For these three cell sites called CASE1, CASE2 and CASE3, they are attributed a proportion as overlapping the corresponding parish(es) based on their network coverage (polygon), as shown in the below board:

## CHAPTER 2. CDR-based location analytics

Table 14: Three cell sites with proportions overlapping the parish(es)

Cell ID	Parish of Lisbon city	Overlapping proportion
<b>CASE 1</b>	<b>Belém</b>	<b>1 ( 1 being 100%)</b>
<b>CASE 2</b>	<b>Belém</b>	<b>0.6542155...</b>
<b>CASE 2</b>	<b>Ajuda</b>	<b>0.3457845..</b>
<b>CASE 3</b>	<b>Alcântara</b>	<b>0.5786542...</b>
<b>CASE 3</b>	<b>Campo de Ourique</b>	<b>0.2665608...</b>
<b>CASE 3</b>	<b>Estrela</b>	<b>0.4213458...</b>

This process is applied for all cell sites of the dataset. In a nutshell, a proportion is calculated for all cell sites as overlapping the concerned parish(es). Case 1 in red shows that this cell site is fully within the parish of Belém, which implies that subscribers connecting to this cell site are assumed to be in Belém at interaction time. Case 2 in orange shows that this cell site overlaps two parishes: Belém and Ajuda, with an overlapping of 65,42% and 34.57%, respectively. Case 3 in blue shows that this cell site overlaps 3 parishes: Alcântara, Campo de Ourique and Estrela, for an overlapping proportion of 57,86%, 26.65% and 42.13%, respectively.

Thus, at this point, all cell sites are attributed an overlapping proportion calculated between their network coverage and parish(es) of Lisbon. Reaching the end of the geospatial upper scaling approach, we have the below cell sites dataset, which has for each cell site the proportion overlapping the corresponding parish(es). As it can be understood, these overlapping proportion are to be used as probability to be in a given parish at specific time.

Table 15: Final Cell sites dataset with overlapping proportion over parish(es)

Features	Description
<b>Cell_ID</b>	Cell ID identifier - i.e. 3DD2000A
<b>Parish</b>	Name of the parish with which cell site network overlaps i.e. Olivais
<b>Overlapping proportion</b>	Overlapping proportion of cell site network coverage and parish - i.e. 0.8545234

This final cell sites dataset was necessary to develop the KPI functions for retrieving spatiotemporal density distribution insights of mobile subscribers moving across the 24 parishes of Lisbon. It is explained in much depth in the following part being about the KPIs.

### **6.5. KPIs – Key Performance Indicators**

Retrieving spatiotemporal insights from call detail records required some adjustments with the data, namely, in terms of temporal granularity given their usage regulations (GDPR and Vodafone Data officer), as well as in terms of spatial granularity given that CDR refer to connected cell sites instead of mobile subscribers. This sub-section describes the defined functions as KPIs to retrieve spatiotemporal density distribution of Vodafone mobile subscribers across the 24 parishes of Lisbon. The developed functions are generic and do not necessarily require Vodafone's data input but any telecom data satisfying the functions' criterias. Hereafter is the primary description of each of the four functions:

**Density distribution count per parish** – This primary function calculates the count of mobile subscribers for the user-specified parish, based on their generated call detail records. In a nutshell, it requires as input, the CDR logs, the cell site data with overlapping proportion, the user-defined parish name of interest and optionally a list of cell sites ID that are to be discarded. Compared to the next function, this one simply takes the input and calculates the count of mobile subscribers for the specified parish and yield it.

**Density distribution count per parish hourly-wise** – Modularity programming advice to cut programs into smaller pieces, each for a single task. Compared to the former function, this one first filter the CDR by the hours of interest through the *StartDatetime* feature. For example, keeping only the CDR that took place from 4am till 6am. Only after that step, it applies the as an inner function the former function 'Density distribution count per parish' to calculates the count of mobile subscribers for the user-specified parish.

**Nationality Parsing** – This function parses the CDR by the Mobile Country Code (MCC) of mobile subscribers' SIM card. To recap, the MCC is derived from the IMSI identifier attributed to each mobile subscriber. Assuming that mobile subscribers are in possession of a SIM-card from the country in which they live, it is possible to filter CDR per the nationality of mobile subscribers. Practically, it is useful to analyse a subset of subscribers (e.g. tourist mining).

**Top X Nationality** – This function yields the top countries with most mobile subscribers in the given area, here in Lisbon, based on their generated call detail records. The X component of the function is user-defined and refers to the number of countries to yield as a result. For example, to determine the top 15 countries with most mobile subscribers for a given time in Lisbon.

### 6.5.1. Density distribution count per parish

This KPI function calculates the count of mobile subscribers based on their CDR for the user-specified parish. Its core task is to take as input the below arguments and to yield the count of subscribers for the user-defined parish, i.e. Olivais. It takes four arguments as input:

Table 16: Description of the Function: Density distribution count per parish

Input arguments	Description
<b>CDR dataset file path</b>	This CDR file must include IMSI and cell ID for each event; Example of a field: IMSI: 26805123456789, cell ID: 3DAE1, ...
<b>Cell sites overlapping parishes file path</b>	This file must include Cell ID, ,parishes name and corresponding overlapping proportion between cell sites and parishes; Example of a field: Cell ID: 3DA1, Parish: Ajuda, proportion: 1
<b>Parish Name</b>	The parish of interest must be user-defined; Example of parish name: Campolide
<b>Cells ID to exclude gathered in a list (optional argument)</b>	Cells ID to be discarded before calculating subscribers count for the user-specified parish; This argument is optional. Example: List of cell ID deserving only Lisbon Airport (located in Olivais parish) such as (AE1D, EDA3,...)

Prior giving a simple practical example of this function, I briefly explain the workflow of how it behaves internally. First it loads the CDR and checks for the optional parameter (cell sites to be discarded). Then it keeps only IMSI and cell ID features and checks for duplicates (remove rows where subscribers connected more than once to the same cell site). As said, this function’s core task is to take the input and yields the count. After that it loads the cell site dataset, filters it to keep only rows for the user-specified parish name and merge it with the CDR object through the common Cell ID, in such a way that we end up with only CDR that connected to cell sites overlapping with the parish of interest.

Finally, it calculates the count of mobile subscribers for the user-specified parish. To do so, it aggregates the probability of each of the concerned subscriber with CDR connecting to cell sites overlapping the parish. Primarily, it attributes to each cell site of the former created object, the *proportion non-overlapping* the parish, called *area neighbour*. After that, it calculates for each subscriber the following:  $\text{Prob}(\text{in Parish}) = 1 - (\text{area neighbour} \times \text{area neighbour} \times \dots)$ . For each subscriber, we obtain a probability between  $[0,1]$  as being in the concerned parish. Finally, we aggregate the probability value of all concerned subscribers to obtain the rounded count of mobile subscribers for the concerned parish.

## CHAPTER 2. CDR-based location analytics

**Example** - A mobile subscriber whom connected to the following cell sites: A1, A2, A3. We would like to know his probability to have been in Ajuda.

Table 17: Toy cell site dataset and call detail records for a practical example using the function Density distribution count in Ajuda

Toy cell site dataset			Toy call detail records		
Cell ID	Parish of Lisbon city	Overlapping proportion	IMSI	Cell ID	Other features
A1	Ajuda	1 (100%)	20601123456789	A1	...
A2	Ajuda	0.20	20601123456789	A2	...
A2	Belém	0.80	20601123456789	A3	...
A3	Ajuda	0.30			
A3	Alcântara	0.50			
A3	Benfica	0.20			

Thus, this mobile subscriber has 3 CDR that connected to three different cell sites with proportion overlapping the parish of interest: AJUDA. Hereafter, I skip the primary part of the function to go straight to the calculation.

Table 18: Toy example of merging CDR with cell site dataset

IMSI	Cell ID	Parish of Lisbon city	Overlapping proportion
20601123456789	A1	Ajuda	1
20601123456789	A2	Ajuda	0.20
20601123456789	A3	Ajuda	0.30



Figure 19: Example of KPI Density Distribution Count per parish

After merging both datasets, with Ajuda as the parish, the proportion of each cell site *not overlapping* the parish is attributed to each row. Thus, an *area neighbour* of 0 for A1, 0.80 for A2 and 0.70 for A3. Then for each mobile subscriber, it calculates the following: Probability(In Parish) =  $1 - (\text{area neighbour} \times \text{area neighbour} \times \dots)$ . In this example, we need to calculate just for one person and considering 3 cell sites:  $P(\text{Ajuda}) = 1 - (0 \times 0.80 \times 0.70) = 1$  (100%). Considering that this subscriber connected to cell site A1, being fully within Ajuda parish, it directly attributes a probability of 1 (100%). If connecting to only cell site A2 and A3, his probability wouldn't be of 1 but the following:  $P(\text{Ajuda}) = 1 - (0.80 \times 0.70) = 0.44$  (44%). Thus a probability value of 0.44. If considering more mobile subscribers, it aggregates the probability values of all of them to obtain the count (rounded) of mobile subscribers in Ajuda.



### **6.5.2. Density distribution count hourly-wise per parish**

This KPI function does the same job as the former one, although this one considers as well the temporal aspect of CDR. Therefore, it primarily filters the mobile subscribers' CDR by the hours of interest given the *Start Datetime* feature. For example, if we want to analyse human mobility only from 4am till 6 am. After having filtered the CDR time-wise, this function runs as an inner function the *Density distribution count per parish*. This function has numerous applications considering specific time or schedule. For example, analysing the count of people at an event (football stadium from 6pm till 9pm) or the density distribution of people across 24 parishes of Lisbon at each hour of the day.

### **6.5.3. Nationality parsing**

Assuming that mobile subscribers are in possession of a SIM-card from the country in which they usually live, it is possible to group them country-wise through the MCC code of IMSI. This function consists in filtering the CDR logs given the user-specified country/countries.

### **6.5.4. Top X nationality**

In the same logic as the former function, this function groups mobile subscribers country-wise based on their generated CDR. It then arrange them in decreasing order and yields the top X countries with most mobile subscribers. E.g. given a file of CDR logs, it yields top 5 countries with most mobile subscribers. Mostly insightful for tourist mining analysis.

## **6.6. Towards population extrapolation**

The spatiotemporal insights retrieved from the above functions only relate to mobile subscribers whom connected to Vodafone cell sites, and not for any other mobile subscribers. To recap, we are talking about mobile subscribers with a Portuguese Vodafone SIM-card and mobile subscribers with a non-Portuguese SIM-card but roaming in Portugal and accessing the cellular network through Vodafone cell sites (collaboration of different telecom operators).

However, it is possible to extrapolate those results to the population level, by considering the telecom market share statistics. Vodafone has 30% of market share in Portugal and can potentially know the market share of other telecom operators with which it collaborates. Thus, based on the MCC of mobile subscribers connecting to its cell sites and market share statistics it is theoretically possible to extrapolate the count of mobile subscribers to the population.

### **6.7. Code testing**

As Donald Knuth, computer scientist, whom wrote ‘The art of Computer Programming’ said, Beware of bugs in the above code, I have only proved it correct, not tried it. (Knuth, s.d.).

After having developed the above mentioned functions, it was crucial to ensure that all were error free and worked in conformance with the functional requirements. Code testing is the adequate solution to help find existent bugs easily and verify that each function work end-to-end as intended with expected output under all possible scenarios. Testing code is as important as writing applications themselves and are primordial before releasing codes into production as it increases code robustness and avoid delivering sub-standard code with issues. Software testing is broadly categorized into two approaches: manual testing with test cases executed manually by a human and automation testing performed with the help of tools and pre-scripted test cases to be executed automatically. The type of testing depends of various factors, namely, what is being tested, project budget, timeline and human expertise. In this research, both type of testing were performed on each function. The two types of are briefly explained hereafter.

Manual testing is the process of manually executing functions and features of an application (e.g. KPI) as an end-user to evaluate its behaviour. This approach is best suited when a person’s judgment is required, for non-repetitive tasks or dynamic conditions test cases (e.g. evaluate user experience). Usually performed on the fly while writing the code, manual testing always requires a tester for each test call and can quickly become time consuming, cost expensive and prone to human error. Automated testing compensates where manual testing fails. With the help of testing tools, pre-scripted tests do the job automatically by comparing actual results with expected results. Once the test scripts are recorded, it makes it easy to fast execute automated test suite and thus save time, cost and manpower. It is principally best suited for repetitive testing, quick code verification after changes are made or even as supplement documentation to help contributors get familiar with the application and its potential pitfalls.

Code testing was highly encouraged by my supervisor as being a fundamental aspect when writing functions. Manual testing was performed on the fly while writing and experimenting the functions, and automated testing was written as side scripts for each function. Python has several test runners available with the most popular being unittest, nose and Pytest. As of choice, all testing were written with the Pytest library. Pytest is a python testing framework that makes it easy to write test-cases and develop suites of them for automation.

Therefore, side scripts were written for each of the function to ensure their functionality. These scripts are external data free and thus are each time creating synthetic data before applying the Pytest on the functions to evaluate their functionality by comparing expected output with actual output. For each function, several case situations are tested. Below I explained the developed script's structure for testing the function 'Density distribution count per parish':

Primarily, it is important to understand what is to be tested. For this function, I wrote automated testing considering 4 different cases where one would retrieve the count of mobile subscribers for one or more parish(es). Thus the pytest runs tests to evaluate that the function 'Density distribution count per parish' yield correct outputs for the following four cases:

- **Case 1** – Retrieve count of mobile subscribers for a parish considering that no CDR passed through a cell site overlapping the parish of interest. For this situation the expected result should be 0.
  1. **Case 2** – Retrieve count of mobile subscribers for a parish where CDR passed through a cell site with a network coverage fully within the parish of interest
  2. **Case 3** – Retrieve count of mobile subscribers for a parish were CDR passed through cell sites not fully overlapping the parish of interest AND where mobile subscribers connected only one time to each of these concerned cell sites.
  3. **Case 4** - Retrieve count of mobile subscribers for a parish were CDR passed through cell sites overlapping the parish of interest AND where mobile subscribers connected several times to each of these concerned cell site

After that part clarified, this script includes primarily the data to be created so the script becomes external data free. For this function, a toy CDR and toy cell sites dataset were created. After that, an object with expected values for each of the concerned parish is created. Only then a pytest function asserts the expected values with the actual values that are obtained by running the function 'Density distribution count per parish'.

Script written, it can easily be run from the command-line and usually takes a few seconds to evaluate if the function works as expected or if issues arise from the defined test cases.

## 7. Experimental results

In this section, I demonstrate and describe some results one can obtain from passing telecom data on the above developed functions. To do so and makes the explanation practical, let's begin with the below toy figure which depicts the density distribution of mobile subscribers whom connected to Vodafone cell sites, across the 24 parishes of Lisbon at 2-hour intervals during a whole day.

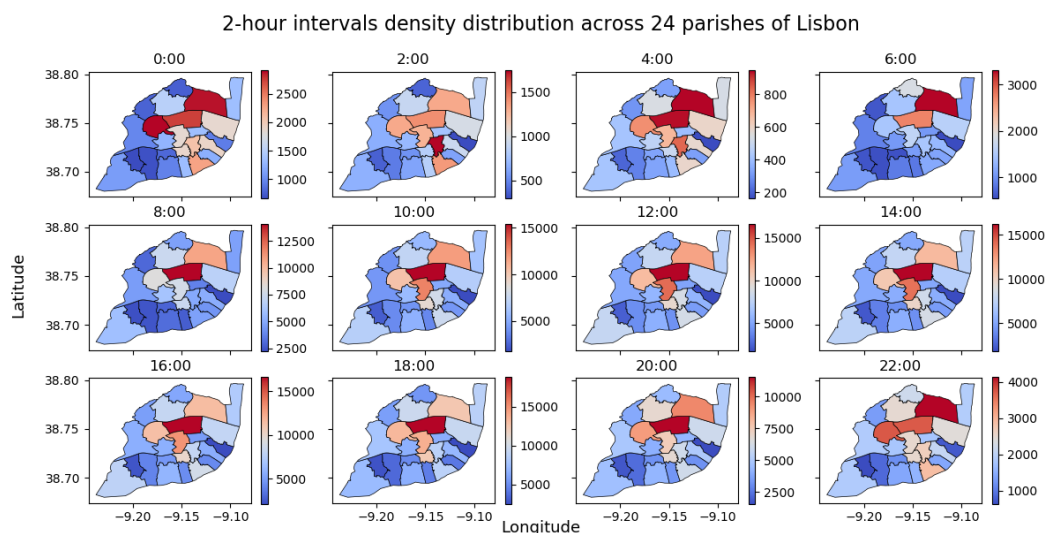


Figure 20: 2-hour intervals density distribution across 24 parishes of Lisbon

Each of these 12 visualizations of the figure shows the aggregated count of the mobile subscribers per parish at a specific hour of the day. To avoid redundancy, this figure shares the same *x-label* (Latitude) and *y-label* (Longitude) for all visualizations. Although, each plot has its own gradient colour legend with the according count of mobile subscribers, which can vary from 200 (4:00) to 15000 (from 08:00 till 18:00). With such visualization, it is possible to get in one picture the total count of these mobile subscribers for all parishes (spatial) and across time (temporal). Besides that, one can make the approach more dynamic by generating small videos of these density distributions across time.

Although, the above figure contains only synthetic data, the idea behind such visualization is to depict the location of people across the 24 parishes of Lisbon and across time. For example, at this spatial scale one can discover the point of interests, residence areas, working areas, and so on. On the next page, I describe more components to consider when using such functions.

## CHAPTER 2. CDR-based location analytics

---

As already said, the functions yield only the count of mobile subscribers whom connected to Vodafone cell sites. However, it is possible to infer the actual population count per parish. A potential solution would be to consider telecom market share statistics.

On one hand, Vodafone knows the national market share it has in each of the country it operates (e.g. around 30% in Portugal) as well as the national market share of the telecom operators with which it has roaming network agreement with. For example, Belgian mobile subscribers with the Orange network operator would connect to Vodafone Portugal cell sites to connect to the cellular network. On the other hand, assuming that mobile subscribers are in possession of a SIM-card from the country in which they usually live, Vodafone can determine their country through the MCC (Mobile Country Code) and their network operator through the MNC (Mobile Network Operator) which is stored in every generated call detail records. Based on the aforementioned information, it is possible to pass from the count of mobile subscribers whom connected to Vodafone cell sites to an approximate count of the population.

Although, the previous figure shows the possibility to specify the parish(es) and timespan of interest, one can also include specific mobile subscribers and/or exclude specific cell sites.

Filtering out mobile subscribers can be performed with the use of the ‘Nationality parsing’ function. As said, every call detail records contain the Mobile Country Code (MCC) component which identifies the country in which mobile subscribers are assumed to come from. Based on that feature, it is possible to keep mobile subscribers from specific countries. For example, for tourist mining, excluding all mobile subscribers with a SIM-card originating from Portugal. Concerning the cell sites, the core function ‘Density distribution count per parish’, has an optional argument which accepts a list of Cell\_ID that are to be discarded. Practically, this argument can be used to focus on specific cell sites that are located in a parish. For example, if one wish to determine the approximate count of people in a stadium during a football game or in a shopping mall. The other way around, Olivais is a parish where is located the airport of Lisbon. By filtering out all cell sites deservng the airport, it is possible to retrieve more accurate statistics about the population being in Olivais.

The function Top X nationality can be used to retrieve even more insights about the concerned mobile subscribers from their generated CDR. By aggregating the count of mobile subscribers to their respective country in descending order, and yielding the top X countries, one can have an insightful snapshot about where tourists come from with along their count. As for the KPI functions, such insights can be analysed through visualizations or through tabular format.

### 8. Conclusions

While location data alone means little, Location Analytics has the power to transform these raw data into meaningful insights. In today's fast-paced changing world, the power of where has gained countless applications across industries and organizations. Rather it be in retail, healthcare, disaster management, government and more, everybody can benefit from location analytics. This primary research project focuses in the process of leveraging telecom call detail records to generate meaningful spatiotemporal insights. The application consisted in developing KPI functions that would generate spatiotemporal density distribution of mobile subscribers whom connected to Vodafone cell sites across the 24 parishes of Lisbon, Portugal. Considering the basic spatial and temporal resolution of call detail records, a series of analytical steps were necessary to prepare, clean and pre-process the given telecom data to turn it into a suitable format to perform the subsequent spatiotemporal analysis. Beyond the benefit of this research project to my personal learning, I believe it can lead to real impacts by revealing people mobility patterns to organizations.

After having worked on this project for an approximate duration of 2 months, I can highlight a few of the points that are to be looked in much depth in future work. Primarily, the biased unit assumption between mobile phone and people. Nowadays, many people carry more than a single mobile phone with them on a daily basis. Secondly, it would be interesting to pursue this study considering scalable infrastructure instead of relying on the computational capabilities of a single machine. Finally, approach the problem through specific real-life situations such as in understanding transport patterns (e.g. how people commute on a daily basis) or in trajectory pattern mining (e.g. trajectories and people flows at different time identification) or mapping areas (e.g. clustering work areas, residency areas, point of interests). My present work has undoubtedly just scratched the tip of the iceberg but shows my great interest in the topic.

Reaching the end of this primary research project had me take a step back and reflect on the lessons learned from this location analytics project. Primarily, working on real data can be challenging and uncertain. It is primordial to clearly understand the problem to be solved, the available data and develop a 'smart data plan'. Reading the literature has been a fundamental step towards getting a good perception of what has already been proven and achieved to follow solid analytic paths. With the considerable development of Internet of Things (IoT) and the proliferation of large amount of geospatial type of data, I believe I will continue to nurture my knowledge into the location analytics field.

## **Chapter 3**

# **Gender prediction from subscribers' list of installed mobile apps**

Your mobile apps tell who you are

### **Abstract**

Whether small or large, organizations require targeted approaches to consumers. In today's highly competitive world, understanding the customers is important and their demographics are crucial to help improve customized services and targeted advertising. However, demographics are often missing due to privacy and/or other reasons. In telecommunication, operators do have demographics of their mobile subscribers but rarely for all of them due to inaccuracy or missingness in the data collection process. This second research project presents the roadmap, technologies and algorithms to build gender prediction classifiers based on the list of mobile subscribers' installed mobile applications. Smartphones being the most personal devices, the type of mobile applications people install could therefore be closely linked to one's habits, demographics and personality traits. To be more precise, this project is an observational study that consists in evaluating several factors throughout the analytic roadmap such as in the recommended quantity of data, apps usage, predictors type and supervised-learning algorithms to develop robust and performant predictive models.



### 1. Introduction

Is marketing art, science or both? For most of the past century, marketing was seen more as art than science and marketers were mostly taking decisions based on intuition. Although, models have markedly shifted across organizations with the advent of big data. In the age of data-driven solution, data came to the fore as an indispensable asset and tool of the toolbox. As Shep Hyken highlighted last year in a Forbes article, customer experience is the new brand for companies to thrive in this highly competitive and globalized economy (Hyken, 2018). Undoubtedly, demographics play a key role for organizations to target better their already made customers and increase their count. As a fact, retaining customers is considered as one of the most critical challenges in the telecom industry (Ahn, Han, & Lee, 2006). However, in practice, demographic traits such as age or gender are usually unavailable due to privacy and/or other reasons. In telecommunication, operators do have demographics of their mobile subscribers but rarely for all of them due to inaccuracy or missingness in the data collection process.

In this research project, I address a solution with along its methodological process to tackle a real problem in telecom companies: lack of accurate demographics of mobile subscribers. More precisely, this project is an observational study of the predictability of mobile subscribers' gender based on their installed mobile applications, collected by the mobile operator Vodafone. Previous contributions demonstrate the predictive power held within installed mobile applications to infer their users' demographics.

Smartphones usage and the associated mobile app market ecosystem are expanding rapidly (Seneviratne, Seneviratne, Mohapatra, & Mahanti, 2014). As of 2019, Google Play Store leads the mobile apps market with more than 2.5 million available apps for Android users, followed by Apple App Store with more than 1.8 million apps for iOS users (Clement, 2019). Besides that, in 2018, more than 194 billion mobile apps were downloaded on connected mobile devices (Clement, 2019). These aspects contribute considerably to the today's Smartphone ubiquitous and highly personalized nature, making them the most personal devices people own. Nowadays, there are mobile applications for almost any situation of our daily life. Many researchers approached the topic by studying potential correlations between individuals and their adoption of connected devices as well as their usage of mobile apps and generated mobile phone metadata. Part of the group, Vodafone is as well present in both Google Play store and App store with its mobile app 'MyVodafone', which allows customers to manage both mobile and/or fixed line tariff. Therefore, can your mobile apps tell who you are behind your screen?

### 2. Literature review

Although complex, the reality of today's economy could be described in a few fundamental aspects: globalized, digitally disrupted and highly competitive. To keep the head above the water and develop important competitive edge, organizations are constantly required to deeply understand who their customers are and what they need, to find adequate ways to meet these needs. To do so, demographics are key factors in getting that edge and develop customer centric strategies. As product preferences vary across group of consumers, segmentation can provide crucial insights into the market and customers. However, as most organizations do not easily have access to customers' personal data such as gender, age, income, marital status and so on, different approaches are considered to fill this gap. This section briefly presents the different contributions related to the demographic prediction.

Inferring demographics has been investigated in academia for a while, and numerous proxies have been considered to perform this predictive task. Predicted demographic traits varied across the read literature, but gender, age and marital status are attributes that come up more often. Early researches on demographic prediction considered linguistic, psychological and sociologic features (Hu, Zeng, Li, Niu , & Chen, 2007). Considering linguistic features, previous contributions focused on modelling the association between demographic attributes and diversity in the linguistics writing and speaking style. As early as 1997, Eckert classified users' gender based on spoken language differences including intentional, phonological and conversational indications (Eckert, 1997). Another early contribution studied demographics predictability based on people's answers to specific psychometric test (Costa & McCrae, 1992).

More recent contributions took advantage of the proliferation of digital communication and Internet, which brought more diverse opportunities for inferring demographic attributes. To be even more precise, considering online behavioural data (e.g. social networks, internet browsing), location check-ins (location to profile) or even mobile phone usage (e.g. mobile networks communication patterns, metadata or installed mobile applications). Indeed, given the prominent place and ubiquitous aspect of such devices, mobile phone datasets have grown into a stand-alone topic with a brand new opportunity with numerous applications such as in inferring users' demographic traits. In telecommunication, organizations constantly generate and collect tremendous amount of mobile data that have already proven their capabilities to unveil a lot about the users' behaviour, activities, habits or in this case demographics.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

In today's business model of developing customized web applications, demographic details play particularly a crucial role. Increasing web services such as search engines, websites and so on, started to give more attention to personalized services for improving the user experience. For example, (Hu, Zeng, Li, Niu, & Chen, 2007), made an approach to predict users' gender and age based on their web browsing behaviours, considering webpage click-through data. In their paper, they highlight that prior contributions show that there exists a correlation between users' browsing behaviour and some of their demographic traits. Their experimental results indicate that the proposed algorithms can achieve up to 79.7% on gender and 60.3% on age in terms of Macro F1-score. Another research performed by (Zhong, Yuang, Zhong, Zhang, & Xie, 2015), consider location check-ins for inferring demographics. They propose a general 'location to profile' (L2P) framework, considering users' check-ins in terms of spatiality, temporality and location knowledge.

The recent telecom's capabilities to handle huge amount of data opened the door to consider interpersonal communication networks (call detail records) to study social interactions. In short, call detail records are transactional records that contain information related to calls and/or text messages such as origin, destination, duration, network. A considerable number of contributions relate to the use of this proxy to infer the gender and/or age of mobile subscribers. Remarkably, the principle of 'homophily' has been demonstrated across different publications. This principle suggests that people tend to have ties with people who are similar to themselves in socially significant ways. For instance, (Brea, Burrioni, & Sarraute, 2015), performed an observational study that approach the relation between mobile phone usage and customers' age. They affirm that the strong age homophily is responsible for the success of their presented algorithm. In another contribution, researchers present their process to automatically infer users' age and gender based on their daily mobile communications (Dong, Yang, Tang, Yang, & Chawla, 2014). Based on real-world data counting more than 7 million users and 1 billion communication records (CDR), they highlight several interesting social strategies, such that young people are more active to broaden their social circles, while seniors tend to keep close with more stable connections.

Finally, few recent academic researchers have investigated if users' adoption of different mobile services could be influenced by their demographics and/or personality traits. Smartphones being the most personal devices people own, the kind of apps installed and/or used could reflect to a certain extent their users' interest, demographics and personality.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

As already said earlier, the smartphones usage and the associated mobile app market ecosystem are expanding rapidly (Seneviratne, Seneviratne, Mohapatra, & Mahanti, Your installed Apps Reveal Your Gender and More!, 2014). It is why, a number of researchers have started investigating how people use their smartphones and how they are affected by them. Very few contributions demonstrate how some users' demographic traits can be inferred based on a snapshot of their installed mobile applications. A principal contribution related to this task performed by, (Seneviratne, Seneviratne, Mohapatra, & Mahanti, 2014), provides several insights on different features related to installed mobile applications to infer the gender of users. They notice gender-specific tendencies for specific mobile applications, such as Pinterest among female users or Reddit among male users. These researchers emphasize that some of their observations corroborate with the market reports about the gender popularity among mobile applications. Besides that, this paper highlights potential privacy loss regarding the collection of information from users' mobile phone such as the list of their mobile applications. A more recent publication addresses the predictability of users' demographics based on the list of their installed mobile applications (Malmi & Weber, You are what you use: Demographic prediction based on user's apps, 2016). These two researchers analysed few particular aspects being, the predictability results of different demographics (gender, age, race, income, married, children), the effects of the training size set and number of users' apps and the effect of three dimensionality reduction techniques. In their conclusions, they point out that there are large differences in the predictability between the studied demographics. Gender being the most predictable and income the least.

## 3. Research definition

### 3.1. Problem statement

In the era of data-driven solution, customer demographic traits such as gender or age play an important role that may enable enterprises enhance their offers adequately and target the right customer in the right time and at the right place. In short, demographic information is a key component for customer centric strategies, e.g. marketing campaign. The problem being, Vodafone face missing and/or unreliable demographic data for a part of its mobile subscribers.

Vodafone, as for most telecommunication operators, usually offer to choose between two mobile plans: Prepaid or Postpaid. As the name suggests, the difference is about when clients pay for them. Prepaid plans tend to be no contract and pay before using the service, whereas for Postpaid, clients receive the bill at the end of each contractual period (e.g. month). From the telecom operator standpoint, the difference remains as well in the data collected from its mobile subscribers. Indeed, a client with a Postpaid mobile tariff needs to share more of its personal details (name, age, gender, house address...) as the operator will have to send the bill at the end of each contractual period. While, prepaid mobile clients are usually much free of such data collection requirements. Besides that, it happens that clients have more than a single phone number active on their account (e.g. mother account with only her demographics but including phone numbers for all family members), leading to potential inaccuracy in terms of demographics per phone number.

In summary, Vodafone may lack accurate demographics data per mobile subscriber occurring due to the following potential situations:

- Mobile subscribers whom did not share demographic details on their account or didn't even create an account (e.g. Prepaid tariff);
- Mobile subscribers with more than just one single phone number registered on their account. Vodafone cannot know if all phone numbers are used by the same person or are used by other people (e.g. family account or business account).

The aforementioned aspects lead to a lack of accurate demographics traits per mobile subscriber, rather due to its missingness or inaccuracy. The core research motivations initiated from the interest in filling the gap of missing accurate gender and/or age details of mobile subscribers, with the use of machine learning techniques on available telecom customer data.

### **3.2. Study area**

Problem statement established, potential solutions took roots and influence from published papers. As explained in much depth in the previous section *Literature Review*, several works discussed gender and/or age identification across different areas and using different techniques. Based on telecom data, two approaches have at first sight attracted my supervisor's interest as well as mine to study the predictability of mobile subscribers' demographic traits:

- ✓ Based on mobile subscribers' daily mobile communication patterns (call detail records)
- ✓ Based on mobile subscribers' list of installed mobile applications

As seen in the literature, both approaches demonstrate significant results and interesting experimental approaches to infer the mobile subscribers' demographics, i.e. gender and/or age. While, the first approach is based on subscribers' daily mobile communication, considering call detail records as the proxy, the second considers their installed mobile applications.

In this research project, the chosen approach was of conducting an observational study of the predictability of mobile subscribers' demographics based on their installed mobile apps over the other approach for several reasons:

- Vodafone has already tackled the topic of mobile communication patterns to predict the demographics of subscribers in the past (Call detail records);
- Curiosity-driven to analyse till what point installed mobile applications can infer users' demographics;
- Vodafone Big Data team' interest in opening this topic for the first time as to make sense of those data to solve one of Vodafone challenges;
- Personal interest in working with another source of data, as for my first project, call detail records data was the core studied proxy.

Considering that Vodafone Portugal had not yet approached the prediction of their mobile subscribers' demographics based on their installed mobile applications, this present research project is considered as a primary observational study in this topic with their data. Therefore, the use-case application focused in opening the topic, by only evaluating the predictive power held within the installed mobile apps to infer the subscribers' gender, being a binary target. Although, past contributions demonstrate that installed mobile apps can as well be used as an interesting proxy to predict the age of mobile subscribers (Classifying age into age range). In this research, only mobile subscribers of Vodafone Portugal were considered.

### 3.3. Goals and objectives

As said, the research purpose stemmed from the interest in filling the gap of missing accurate mobile subscribers' demographics and led to study the predictability of their demographic traits based on their installed mobile apps. From that purpose, goals and objectives have successively unfolded to set the research baseline and architecture. They as well shaped the directions and limitations throughout the research. Practically, the use-case application focused in evaluating the predictability of only the mobile subscribers' gender. As an initiate study, it is easier to predict their gender being a binary target than age and ease the process of defining the analytic roadmap and evaluate the models' performance. The research goals are the following threefold:

- ✓ Conduct an observational study of the predictability of mobile subscribers' gender based on the list of their installed mobile applications;
- ✓ Define a 'roadmap solution', which describes the applied steps of the data science workflow to perform this predictive task;
- ✓ Perform a comparative analysis of several stages of the defined roadmap solution to provide a series of best practices.

The primary goal concerns the research as a whole and in conducting it with a focus on the gender demographic of mobile subscribers. The second goal represents the skeleton of the methodology section of this report. Although, explained in much depth later in the report, the research was performed through two platforms: on Vodafone cloud solution (Big Data Platform) where customer data of interest were stored, and then on my local machine. Hive big data tool was used to perform queries on the Big Data Platform, and Python was used for all other analytical tasks on the local machine. Although, the research required lots of going back and forth throughout the stages of the workflow, below it is simplified by a logical linear flow:

Table 19: Description of the stages of the data workflow

<b>Workflow stages</b>	<b>Location</b>	<b>Core task in brief</b>
<b>Data exploration</b>	Big Data Platform (BDP)	Understanding the data
<b>Data pre-processing</b>	Local Machine	Web-crawling Google Play Store
<b>Data collection</b>	Big Data Platform (BDP)	From BDP create 6 datasets as csv.file
<b>Data preparation</b>	Local Machine	Prepare data formatting for prediction
<b>Data analysis</b>	Local Machine	Visualization and basic statistics
<b>Data modelling</b>	Local Machine	Supervised-learning algorithms
<b>Models evaluation</b>	Local Machine	Metrics to evaluate the results obtained

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

The third goal consisted in performing a comparative analysis concerning several stages of the project workflow to end-up with a series of best-practices and ultimately few baseline models. The concerned stages of the workflow were: *data collection*, *data preparation* and *data modelling*. To do so, different experimental approaches were applied to finally be able to compare them based on their processes, basic statistics and predictive results. To ease this goal, it was deconstructed into four core objectives that were achievable and measurable steps throughout the workflow. These four objectives are thus described below, each being related to a stage of the workflow along with its research question(s).

### **Data collection Stage:**

- Objective 1 – Comparison between using 1, 3 or 6 months of available data?

Determine the recommendable quantity of customers' data to obtain better predictive results. Vodafone keeps customers' data for a maximum duration of 6 months before being deleted. Considering more months directly implies considering more clients and more mobile apps.

- Objective 2 – Comparison in between considering per subscriber, all of mobile apps installed in their device or only the ones which generated internet data?

This question aimed at determining if considering only mobile apps generating internet data could be an interesting proxy to filter mobile apps installed but never or not used. Thus, removing mobile apps that would not really represent the subscriber. Vodafone has metadata about which mobile apps generated internet data through its cellular network. This aspect was evaluated, although keeping in mind that some mobile apps do not require internet to be used.

### **Data preparation Stage:**

- Objective 3 – comparison between using the following three predictors types:

(1) Title of installed mobile apps such as Gmail, Facebook, and so on ... (2) or two dimensionality reduction techniques that are: (2.1) aggregate mobile apps to the Google Play category. This app store has 49 categories for their apps (e.g. Gmail for Communication). (2.2) Truncated Singular Value Decomposition (TSVD), which consists in performing a linear dimensionality reduction. These three approaches are explained in much depth later in the report in the "Data preparation" sub-section and why aggregating mobile apps to the Google Play store is as well explained in the "Data pre-processing" stage. These three techniques are supported by past contributions which show encouraging results.



### **Data modelling & Models evaluation Stage:**

➤ Objective 4 – comparison between using the following three classification algorithms: Logistic Regression (LR), Decision trees (DTs) and Random Forest (RF). The choice of these three algorithms was supported based on previous contributions which demonstrate interesting and concluding results using them. As an initiate observational study and considering the time constraint to work on this research project, no hyper-parameter tuning were performed to set the baseline of the different classifiers.

### **3.4. Research architecture**

The figure 22, *Research Architecture*, that is on the next page intends helping readers understand more clearly the process of achieving the above four defined objectives. It visually summarises the process of generating the 6 datasets (Considered quantity of data & apps usage). The following step demonstrates that on each of the 6 datasets, the three predictors are adapted through different data formatting, as to ultimately perform the comparative analysis. Thus, the methodological process to perform the comparative analysis is explained as follow:

- 1. Select all mobile subscribers with 1 phone number and for whom Vodafone has data about their gender and list of mobile applications – Ground Truth:**

The primary step concerned the subset of mobile subscribers representing the 'Ground Truth' to be considered in the process for analysis and modelling purpose.

- 2. Objective 1 & 2: Data Collection – Creation of the 6 datasets as csv.file:**

The 1<sup>st</sup> objective relates to the quantity of information to keep in each dataset (1,3, 6 months) and is coupled with the objective 2 which considers for each month, the apps usage per subscriber (all of their installed apps or keep only the apps that generated internet data). Thus, by considering these 2 objectives, the 6 datasets were created as csv.file to pursue on Python.

- 3. Objective 3 – Predictors data formatting on each of the 6 datasets:**

For each of the 6 datasets, adapt the 3 data formatting, thus having 18 tables to evaluate (6x3).

- 4. Objective 4 – Feed the 18 tables to the 3 algorithms to evaluate their predictions:**

Finally, feed the 18 tables to the three algorithms and evaluate them based on a set of metrics.

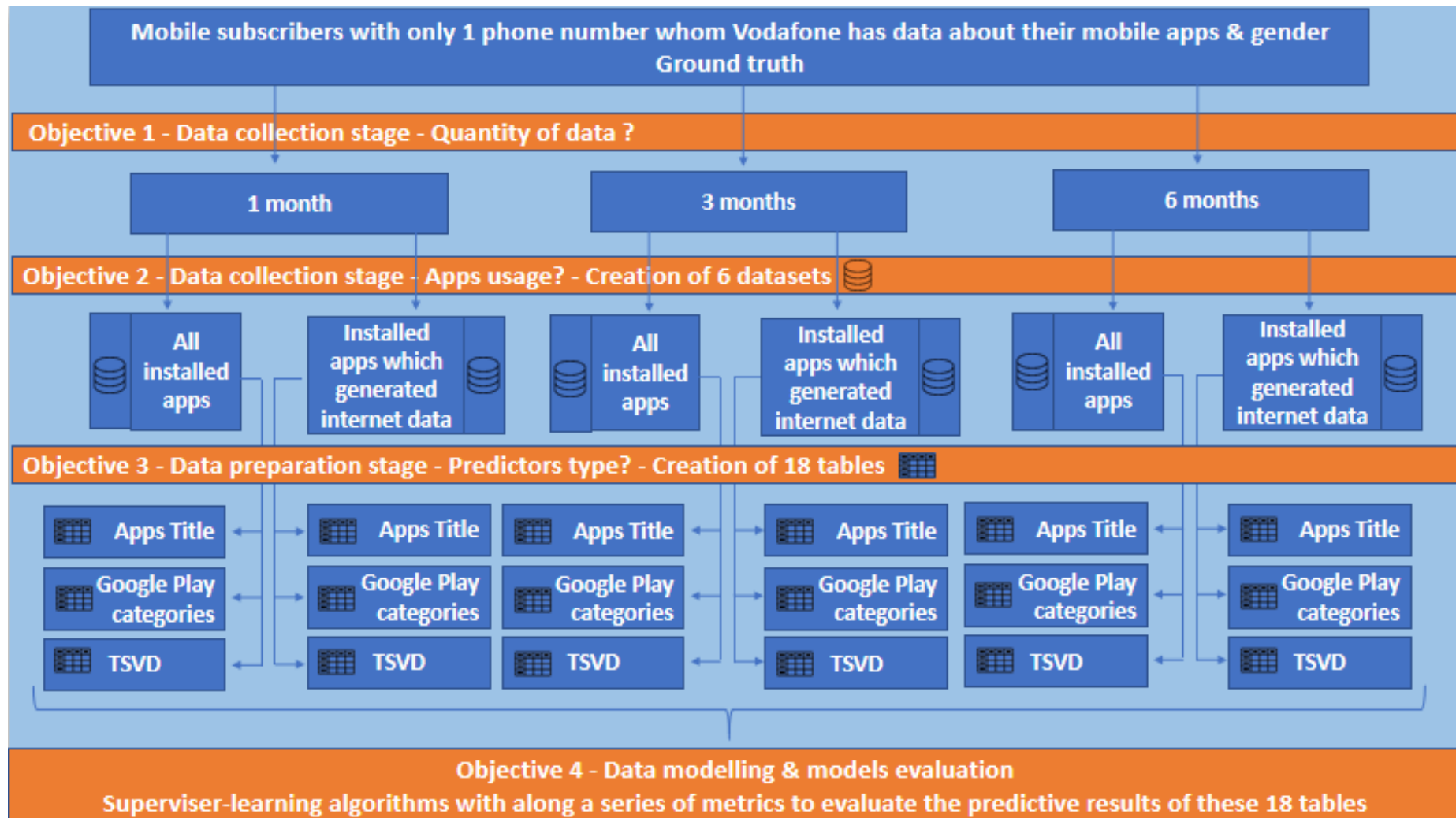


Figure 21: Research Architecture for the comparison analysis

## **4. Methodology, tools and approach**

### **4.1. Big Data Platform – BDP**

Vodafone established its own on-premises data centres as a private cloud hosting solution. Main differences with public cloud are that data are not hosted by a third party (e.g. AWS) and besides having data located in the cloud, Vodafone is also responsible for management, control and security of the data centres. The solution includes a Big Data Platform (BDP) that runs under Apache Hadoop ecosystem, which encompasses big data components as well as tools for managing efficiently huge datasets and Hue, a web interface for analysing data with Hadoop.

Apache Hadoop is an open-source, scalable and fault tolerant framework used to distribute storage and processing of data in a distributed computing environment. In short, Hadoop consist of four key components: HDFS for data storage, Map-Reduce for data processing, YARN for resource management and Hadoop Common as a collection of common utilities and libraries. Alongside is the suite of Hadoop tools to solve big data problems such as Hive, an open-source data warehouse system that facilitate data summarization, queries and analysis of large datasets stored in Hadoop files. The Big Data Platform secured behind firewalls is to be entered through a user web interface portal to access data stored in Vodafone data centres. Data are stored across sparsely populated tables that can scale to billions of rows and thousands of columns, enabling to store tremendous amount of data, reaching the petabytes.

In this research, the data stored in the BDP were accessed, analysed and extracted by using the Apache Hive big data tool through the Hue interface. Apache Hive has the particularity to give a SQL-like user interface to write queries to process the data. Data analytics tools used are explained in much detail in the following section "*Tools*".

### **4.2. Tools**

Two main data analytics tools were used to perform the research: Apache Hive and Python. The entire analysis was carried out using the computer that I was provided by Vodafone.

Apache Hive is an open-source data warehouse tool built on top of Hadoop, which is designed to ease data summarization, ad-hoc queries and analysis of large data stored in Hadoop. Its purpose is to simplify writing complex Hadoop MapReduce jobs for ad-hoc requirements. Indeed, it provides a SQL dialect called Hive Query Language (HQL) for querying and translating them into MapReduce jobs. In this research, this big data tool was used to perform all data queries on the Vodafone Big Data Platform, to access customers data stored in the cloud. Hive big data tool was considerably used to perform exploration, analysis and data retrieval tasks.

Python, being a general-purpose programming language, was used after having collected the data from the BDP to perform a myriad of data analytics tasks that are explained thoughtfully in the methodology section. To be more precise, Python was used in the data pre-processing stage for web-crawling Google Play Store, and in the stages from the data preparation to the models evaluation (For more clarity, see Table 19). Python wouldn't be what it is today without its rich ecosystem counting numerous libraries that enable hundreds to even thousands of different tasks. In this research, we took advantage of a series of libraries; *Pandas*, an easy-to-use tool for data structures and analysis, *NumPy* as the fundamental module for scientific computing, *Dask* for providing advanced parallelism when reading heavy files. Concerning the visualization tasks, *Matplotlib* and *Seaborn* modules. *Play-scraper* for the specific task of scraping and parsing apps from the Google Play Store and finally *scikit-learn* library for developing and evaluating the predictive models.

### **4.3. Approach**

The entire methodological process encompasses the following stages of the data science workflow: Data exploration, data pre-processing, data collection, data preparation, data analysis and visualization, data modelling and models evaluation. The analysis was performed on two platforms, on the Big Data Platform with use of Apache Hive queries through the Hue web interface and on my local machine with the Python programming language. Although, the flow consisted of constant back and forth throughout the different stages of the project workflow, below it is explained as following a logical process, for sake of clarity. The approach of each stage is briefly introduced to give an overview of the methodology:

- **Goals and objectives setting the experimental directions of the research:**

Redefining the objectives throughout the entire research was necessary in this observational study to reach the defined goals. Indeed, that stage contributed considerably in the process of orienting the different experimental approaches to create the different suitable datasets as to compare their predictive results and coming up with a series of best practices. The experimental approaches took in parallel several directions in terms of quantity of data (1, 3 or 6 months of data), apps usage (all installed apps or only the ones generating internet data) to create the 6 datasets. On each of these 6 datasets, the data formatting was transformed to fit the following three type of predictors (mobile apps title or 2 dimensionality reduction techniques: (1) mobile apps aggregated at their google play category, (2) TSVD technique). In this logic, 18 tables were created and on which the three chosen algorithms (Logistic Regression, Decision trees and Random forest) were performed, to be then evaluated thanks to a series of metrics.

- **Data exploration:**

This first stage consisted in drawing the initial picture of the studied problem statement (lack of accurate demographic traits) by exploring data stored on the Big Data Platform with HIVE queries. This stage includes two main parts: (1) first considerations based on this primary exploration and (2) retrieval of basic statistics about the target population (quantification task).

- **Data pre-processing:**

Based on the data exploration performed on the BDP, it was noticed that Vodafone collects all subscribers' mobile apps and in their APK format (Android Package). This pre-processing step consisted in scraping the Google Play Store website for two reasons: (1) filter APK that are not

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

installable mobile apps (i.e. Download) and (2) retrieve details for the recognized APK. Finally, the output list of APK along with their details was loaded to the BDP as a reference table.

- **Data collection:**

The data collection stage describes the process of retrieving the data from the BDP to generate the 6 primary suitable datasets. Looking back at the research architecture (See Fig. 22), one can see that this stage encompasses the primary and second objective that were required to perform the comparative analysis. The 1<sup>st</sup> objective relates to the quantity of data to be considered (1, 3 or 6 months) and the 2<sup>nd</sup> objective relates to the apps usage considered (if considering all subscribers' installed mobile apps or only the ones which generated internet data). Based on these two aspects, the 6 datasets were created through HIVE queries and saved as CSV file to be further worked on with Python directly on my local machine.

- **Data preparation:**

The data preparation stage describes the three data formatting transformation related to the predictors: (1) title of installed mobile apps (i.e. Gmail, Facebook and so on...); Two dimensionality reduction techniques: (2) installed mobile apps aggregated to their Google Play category (i.e. Google in TOOLS), (3) TSVD (Truncated Singular Value Decomposition).

As shown on the Research Architecture (See Fig. 22), the data preparation stage includes the Objective 3 of the comparative analysis flow. As the data formatting transformation concerned all the 6 former created datasets, it led to ultimately create 18 tables. (Example of a table: 1 month of data, with all subscribers' installed mobile apps and with the apps title as predictors).

- **Data analysis & visualization:**

Prior to the data modelling, the data analysis & visualization stage aimed at better understanding the data, potential similarities and differences across the 6 generated datasets. A series of basic statistics and example of visualizations are included in this stage of the report.

- **Data modelling & models evaluation:**

This stage describes the 3 classification algorithms, the model evaluation process and the different metrics employed to assess the performance of all of the generated models. To recap, there are 18 different tables which differ in terms of quantity of data, apps usage and predictor data formatting.

### **4.4. Data exploration**

This primary data exploration stage consisted in exploring the data stored on the Big Data Platform (BDP), to get a first good grasp of how the data was structured across the tables, as well as to analyse the customers' data of interest, and finally retrieve a series of basic statistics about these mobile subscribers. This sub-section has two parts; the first one summarises the key-aspects considerations while the second part describes the basic statistics retrieval process.

#### **4.4.1. Primary considerations**

Vodafone has 'MyVodafone' mobile app, available on Google Play Store for Android users and Apple Play Store for iOS users.

In this research, only mobile subscribers with an Android system were considered, representing around 90% of all mobile subscribers. Advice by my supervisor, iOS users were discarded, as representing less than 10 percent and considering that the retrieval process of mobile apps from iOS device was not as straight forward as for Android devices due to advanced privacy settings.

Besides that, Vodafone doesn't collect the mobile apps of all of its mobile subscribers. At some point, mobile subscribers having the 'MyVodafone' app have all of their mobile apps collected by Vodafone. It implies that mobile subscribers whom did not have this app at the time of the study were not concerned at all. Concerning the ground truth, it concerns only mobile subscribers respecting the following three criterias: (1) having only 1 phone number on their account, (2) gender known by Vodafone, (3) mobile apps collected by Vodafone. Indeed, mobile subscribers for whom Vodafone didn't collect their mobile apps are not even concerned by this observational study and the ones with more than 1 phone number on their account fall under the 'inaccuracy' category, and the ones for whom Vodafone doesn't know their gender fall under the 'missing' category. The ground truth represents the subset of mobile subscribers on which we have all the necessary data to analyse and build the predictive models.

Finally, mobile apps collected by Vodafone required to pass through a pre-processing step, as strictly speaking, Vodafone doesn't just collect the installed mobile apps of its mobile subscribers but literally all applications present in their device, from apps installable from Apps stores (e.g. Gmail) to part-of-device apps (e.g. Download, radio). Moreover the apps are collected under a specific Android Package format known as APK. Although explained later in the 'data pre-processing stage', this step consisted in web-scraping Google Play Store.

### 4.4.2. Target quantification

The target quantification was a crucial step as it helped better picture the lack of accurate demographic data of Vodafone mobile subscribers as well as determining the ground truth that was to be later on, used to build the predictive models. This quantification step was performed for the two following targets to get the entire picture of the problem: (1) All mobile subscribers and (2) All mobile subscribers for whom Vodafone has data about their mobile applications.

As said in the previous sub-section, Vodafone does not have data about the mobile apps of all of its mobile subscribers, but only for a subset of it. Although, it was necessary and interesting to put the problem and research into perspective by looking at the problem at two certain levels. The first one, being general, concerned the entire problem of lack of accurate gender details of mobile subscribers while the second one concerned only the mobile subscribers for whom Vodafone had data about their mobile applications.

For both targets, the same and following statistics were retrieved to get a better sense of the problem. The below multi-level board helps understand the research through different angles; Thus, per tariff (Postpaid, Prepaid) and per gender (missing data, women or men), the total count of mobile subscribers' account and then more precisely, the aggregate count of mobile subscribers' account with 1,2,3 or more phone numbers per account (e.g. Count of all accounts with 2 phone numbers). Accounts for which Vodafone does not know the mobile subscriber's gender (missing) or has more than 1 phone number (inaccurate) cannot be in the ground truth.

Table 20: Target Quantification – Basic statistics

Tariff	Gender	Count of all accounts	Count of all accounts pn_1	Count of all accounts pn_2	Count of all accounts pn_3	Counts of all accounts pn_x
Postpaid	Missing	Missing				
	Women	inaccuracy	Ground truth			
	Men		Ground truth			
Prepaid	Missing					
	Women		Ground truth			
	Men		Ground truth			

This table, normally filled with aggregated count, classifies mobile subscribers into 3 groups. **In green**, mobile subscribers' account as ground truth, **red** with inaccurate data (at least 2 phone numbers per account) and **yellow** with missing gender data.



### **4.5.Data pre-processing**

This stage relates to the discoveries made throughout the data exploration stage about the mobile apps that Vodafone collected from its mobile subscribers, as explained in the previous sub-section 'Primary considerations'. The objective of this pre-processing stage was to 'clean' the mobile apps collected by Vodafone that were under the Android Package format (APK).

As said, Vodafone does not appropriately have the list of installed mobile apps of its mobile subscribers. Instead, it has the list of all applications present in the device of its mobile subscribers and under the APK format. It includes applications such as Netflix that could potentially be installed from an App store or even non-installable part-of-device applications such as radio or downloads. The Android Package or mostly referred by its acronym APK, is an Android package file format used by the Android operating system for distribution and installation of mobile apps and middleware. Below are four examples of APK; Two being from Google Play Store and two others being part-of-device applications.

Table 21: APK along with apps title and category if belonging to Google Play Store

<b>Android Package Format (APK)</b>	<b>Apps title</b>	<b>Google Play category</b>
Android :com.google.android.gm	Gmail	COMMUNICATION
Android :com.android.providers.downloads	/	/
Android :com.vodafone.mCare	MyVodafone	PRODUCTIVITY
Android :com.android.settings	/	/

The objective of this pre-processing stage was to collect all APK from the BDP and web scrap the Google Play Store to create a 'Reference table' that includes for all recognized APK, their details such as their app title and google play category. The Web scrapping did the following:

- Retrieve recognized apps details from Google Play Store such as title and category;
- Filtering out APK that are part-of-device applications such as downloads or settings.

Only Google Play Store was considered for the web scrapping task. It is recognized as being the leading and largest distribution for Android mobile applications and counts more than 2 million apps available. Before describing this cleaning process, let's see briefly about Google Play Store. This digital App store has for each of its application, a series of features describing them, such as their APK, title, category, version, price, developer details and more. As shown on the next page, Google Play has 49 categories to group their apps by their respective genre.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

Table 22: Google Play Store: 49 categories for its applications

ART_AND_DESIGN	GAME_CARD	MAPS_AND_NAVIGATION
AUTO_AND_VEHICLES	GAME_CASINO	MEDICAL
BEAUTY	GAME_CASUAL	MUSIC_AND_AUDIO
BOOKS_AND_REFERENCE	GAME_EDUCATIONAL	NEWS_AND_MAGAZINES
BUSINESS	GAME_MUSIC	PARENTING
COMICS	GAME_PUZZLE	PERSONALIZATION
COMMUNICATION	GAME_RACING	PHOTOGRAPHY
DATING	GAME_ROLE_PLAYING	PRODUCTIVITY
EDUCATION	GAME_SIMULATION	SHOPPING
ENTERTAINMENT	GAME_SPORTS	SOCIAL
EVENTS	GAME_STRATEGY	SPORTS
FINANCE	GAME_TRIVIA	TOOLS
FOOD_AND_DRINK	GAME_WORD	TRAVEL_AND_LOCAL
GAME_ACTION	HEALTH_AND_FITNESS	VIDEO_PLAYERS
GAME_ADVENTURE	HOUSE_AND_HOME	WEATHER
GAME_ARCADE	LIBRARIES_AND_DEMO	
GAME_BOARD	LIFESTYLE	

Google Play Store has for each of its application, a long list of features to describe them. Below few features are listed for the example of “Google Chrome: Fast & Secure” mobile app:

- **APK** – com.android.chrome
- **Title** – Google Chrome: Fast & Secure
- **GenreId** – COMMUNICATION
- **Free** – Yes
- **Installs** – 1,000,000,000+
- **Review** – 17333662
- **Score** – 4.3
- **Updated** – November 1, 2019

Now, let see about the process of retrieving such data from the Google Play Store through the APK. On the next page, the web scrapping process is described and explained following three main steps:

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

- **Step 1 – APK collection from the Big Data Platform:**

The primary step consisted in collecting all of the APK appearing in the BDP as being present in the device of the mobile subscribers. As such, more than 200,000 APK were collected and saved in a csv.file to pursue with the web crawling process on python. Although, most APK were appearing in very few different devices (less than 10 devices), they were all collected with the objective of creating a reference table on the BDP including all details for these APK.

- **Step 2 – Web crawling Google Play Store with 'Play Store Scraper':**

Play Store Scraper is a free and open source framework for Python which provides APIs to easily scrapes and parses applications from the Google Play Store. Thus, on Python with this library, mobile apps details were fetched for the recognized APK from the Google Play store.

The function iterates across the lines of the list of APK and checks for each of them, if the APK is being recognized from the Google Play Store. If it did, the apps details are fetched (app title, category, version, price, developer and so on...), otherwise the APK is being discarded from the list. Out of the +200,000, around 40,000 were recognized from the Google Play Store.

- **Step 3 – Load the list of mobile apps output into the BDP as a new table:**

The list of 40,000 APK recognized from the Google Play Store contains for each APK their corresponding apps details. To pursue with the research, this list was loaded into the BDP as a new table 'Google Play Android Apps' to be used as a reference for Android mobile apps. Below is an example of two rows of this newly-created table:

Table 23: Generated table 'Google Play Android Apps'

APK	Title	Category	Is_Free	Other features
Android :com.google.android.gm	Gmail	Communication	Yes	...
Android :com.airbnb.android	Airbnb	Travel and local	Yes	...

To be more specific, this new table was required in the following stage being the 'data collection'. This table was necessary to select for the concerned mobile subscribers, only their google play installable apps, and thus discard all other APK. As explained in much depth in the next page (data collection stage), the 6 primary datasets are to be created (considering the quantity of data: 1, 3, 6 months and the apps usage: all installed mobile apps or only the ones generating internet data).

## 4.6. Data collection

The data collection stage relates to the process of gathering the data of interest from the BDP to generate the 6 datasets as csv file and thus continue the research in Python with the data preparation, data analysis & visualization, data modelling & models evaluation stages.

To recap with what is meant by data of interest, it is the list of mobile subscribers with along their installed mobile apps (title, google play category) and gender. Let's first look why 6 datasets were created before looking at how they were extracted from the BDP.

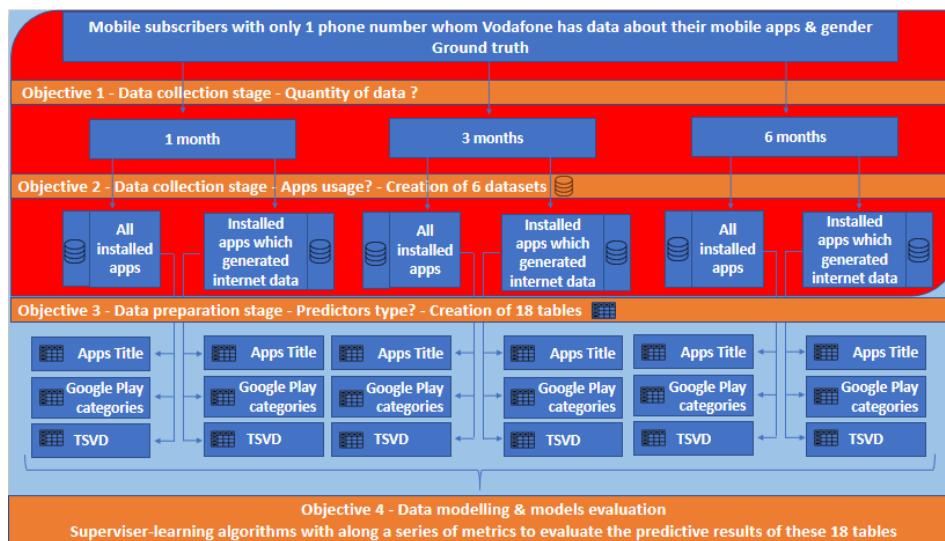


Figure 22: Research Architecture: Objective 1 & 2

As shown on the above figure of the research architecture, the data collection stage concerns the 2 primary objectives (red background). These 6 created datasets as csv are listed hereafter:

- 1 month of data with all subscribers' installed apps;
- 1 month of data with only subscribers' installed apps that generated internet data;
- 3 months of data with all subscribers' installed apps;
- 3 months of data with only subscribers' installed apps which generated internet data;
- 6 months of data with all subscribers' installed apps;
- 6 months of data with only subscribers' installed apps which generated internet data.

Considering all installed mobile apps, count varied roughly from around 80.000, 90.000 and 100.000 subscribers and 6500, 7500, 8500 apps, for 1,3 and 6 months of data respectively. Considering only installed mobile apps which generated internet data, count varied roughly from around 14.000, 18.000 and 23.000 subscribers and 600, 900, 1200 apps, for 1,3 and 6 months of data respectively.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

For sake of clarity, a new temporary table was created on the BDP to ease the data extraction for creating the 6 datasets. This table includes for the concerned subscribers (only having 1 phone number on their account), their respective APK, gender, month in which each APK was collected and Boolean values telling for each app if it generated or not internet data. Hereafter is an example of this temporary Big table. As shown, per row there is just one APK. Thus a mobile subscriber with 35 APK will be repeated across 35 rows.

Table 24: BDP temporary table of mobile subscribers with APK and gender known

Client ID	Gender	APK	Month	Is_generating_traffic
Clientid1	Male	Android :com.google.android.gm	09	False
Clientid1	Male	Android :com.vodafone.mCare	09	True
...	...	...	...	...
Clientid2	Female	Android :com.airbnb.android	11	True
...	...	...	...	...

Therefore, this temporary table was crossed with the new table 'Google Play Android Apps' to generate the 6 datasets. Below is an example of how values of each of the 6 datasets look like:

Index	ClientID	Title	Category	Gender
0	Clientid1	Facebook	SOCIAL	Female
1	Clientid1	Instagram	SOCIAL	Female
2	Clientid1	Activobank	FINANCE	Female
3	Clientid1	Huji Cam	PHOTOGRAPHY	Female
4	Clientid2	Facebook	SOCIAL	Male
5	Clientid2	8 Ball Pool	GAME_SPORTS	Male
6	clientid2	Score! Hero	GAME_SPORTS	Male
7	clientid3	Facebook	SOCIAL	Male
8	clientid3	Fifa Soccer	SPORTS	Male
9	clientid3	Santander	FINANCE	Male
10	...	...	...	...

Figure 23: Outlook of the values of the 6 generated datasets

As shown on the above figure, the basic formatting of each of these 6 datasets lists subscribers' mobile apps across several rows. This basic format wouldn't enable the classification algorithms to study the interrelations of all apps installed in each subscribers' device with their gender but would rather analyse it mobile app per mobile app with the gender.

The next step of the workflow, being the data preparation stage, consisted in adapting adequately these 6 datasets to the three predictors format: apps title, aggregated to the google play category or TSVD, and thus generating 18 different tables on Python. This step represents the Objective 3 of the research architecture (See Fig. 23).

### **4.7.Data preparation**

Before describing the applied tasks of the data preparation, let's review briefly what has been achieved till now. Goals and objectives have considerably shaped the directions and lines of the research to define a data workflow (See Table 19) and an architecture (See Fig 22) to perform the comparative analysis. In the primary data exploration performed in the BDP, a series of considerations and basic statistics have set the baseline of the research. After that, a new table 'Google Play Android Apps' was added to the BDP, including the corresponding apps details for APK being recognized by the Google Play Store. Finally, in the data collection stage, the 6 primary datasets were extracted from the BDP and saved as csv.file to pursue the research in Python. This data preparation stage corresponds to the objective 3 of the research architecture (See Fig. 23). This stage consisted in adapting the 6 datasets formatting to the following three predictors and thus led to create 18 suitable tables:

**1) Apps title as predictors;**

- This format uses the apps title (e.g. Gmail) to predict the subscribers' gender.

**2) Apps aggregated to their google play categories as predictors;**

- This format aggregates subscribers' apps to their corresponding google play category. (e.g. Facebook and Instagram to SOCIAL with a value of 2).

**3) Truncated Singular Value Decomposition (TSVD) as predictors.**

- This format performs a linear dimensionality reduction.

Implementing these 3 predictors format to each of the 6 datasets led to create 18 tables in Python. As shown in the Figure 24 (previous page), the basic data formatting doesn't enable the Classifiers to study the underlying interrelations of all subscribers' apps with their gender. Indeed, with this basic format, subscribers' apps are listed individually across several rows. Thus a subscriber with 15 apps would be repeated across 15 rows along with the gender. With such formatting, the algorithms would study the relation mobile app per mobile app with the gender, which is not what is wanted. Instead, what we want is for the algorithms to study the interrelations of mobile apps installed in each mobile subscriber's device with their gender. More precisely, allow the algorithms capture the relation between each subscribers' installed apps (apps title, aggregated to Google Play category or TSVD) with their gender.

The following three predictors formatting solve this issue by arranging accordingly subscribers a single time row-wise and predictors (title, categories or TSVD) column-wise.

1) **Format 1 - Apps name as predictors:**

	Index	Facebook	Instagram	Activobank	Huji Cam	8 Ball Pool	Score! Hero	Fifa Soccer	Santander	Gender
Clientid1 →	0	1	1	1	1	0	0	0	0	Female
Clientid2 →	1	1	0	0	0	1	1	0	0	Male
Clientid3 →	2	1	0	0	0	0	0	1	1	Male
	3	...	...	...	...	...	...	...	...	...

Figure 24: Format 1 – Apps name as predictors

A picture is worth a thousand words; As for the three formats, the core step consists in transposing the table (See Fig. 24), in such a way that subscribers are listed a single time row-wise and predictors column-wise. For this first format, apps title are listed column-wise, with binary values: 1 if the client has the app in his mobile phone, 0 otherwise. Although, this data format shows promising predictive results, it comes as well with heavy computational costs. Indeed, the size of the table is not only proportional to the number of mobile subscribers (rows) but as well to the total number of mobile apps (columns). Thus, if the dataset contains 10.000 different apps, then each of these 10.000 apps would become a column with values of 0 or 1 for each subscriber. In short, the main disadvantage of format 1 is that it doesn't scale easily.

2) **Format 2 – Apps aggregated at the Google Play Category:**

	Index	SOCIAL	FINANCE	PHOTOGRAPHY	GAME_SPORTS	Gender
Clientid1 →	0	2	1	1	0	Female
Clientid2 →	1	1	0	0	2	Male
Clientid3 →	2	1	1	0	1	Male
	3	...	...	...	...	...

Figure 25: Format 2 – Apps aggregated to their Google Play Category

This second format, being one of the two dimensionality reduction techniques, solves the computational costs issues of the former format. This method being directly performed from the basic format (See Fig. 24), consists in aggregating per subscriber, all of their apps to their respective google play category. To recap, Google Play Store has 49 categories to group all of its mobile apps. Therefore, the table size is limited to a maximum of 49 predictors. Following the values of the table with the basic format (Fig. 24), Clientid3 has 1 SOCIAL app (Facebook), 1 FINANCE app (Santander), 1 GAME\_SPORTS app (FIFA Soccer). However, the main cons. of this approach is that algorithms do not take advantage of the 'predictive power' held within specific mobile apps title. For example, based on the analysis, "Pinterest" has a tendency to be used mostly by women and "Meus Resultados" by men. Instead, this format considers the aggregated count of subscribers' apps per Google Play Category.

3) Format 3 – Truncated Singular Value Decomposition (TSVD):

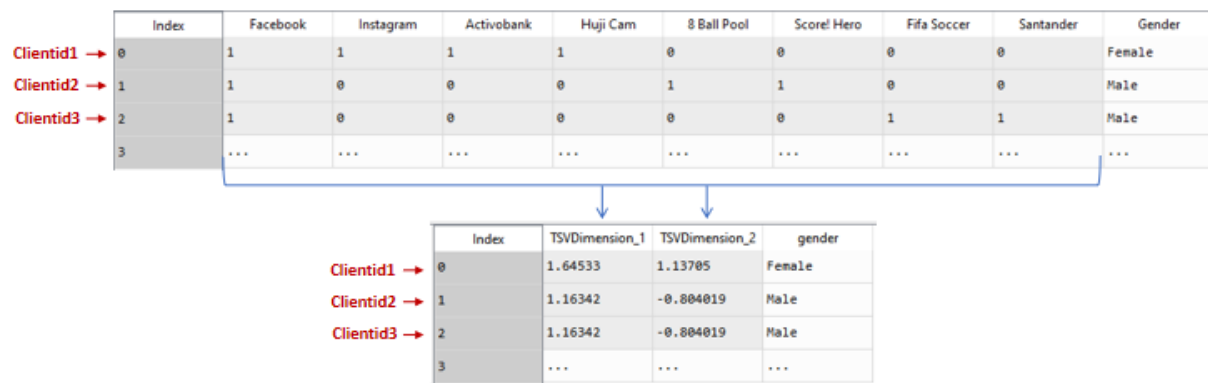


Figure 26: Format 3 – Truncated Singular Value Decomposition example with 2 dimensions

This third and last predictor format consists in applying the TSVD, a matrix factorization method that performs linear dimensionality reduction. In this context, it is used to reduce a high-dimensionality dataset into fewer dimensions while keeping most information/variability. Although, very similar to PCA (Principal Component Analysis), this technique does not center the data nor perform factorization on the covariance matrix but rather on the data matrix. It implies that it works fine with sparse matrices, which is what is wanted. In a nutshell, The SVD is a matrix decomposition method used to reduce a matrix into its constituent matrices:  $A = USV^T$ . Where A is an  $m \times n$  matrix; U is an  $m \times n$  orthogonal matrix; S is an  $n \times n$  diagonal matrix and V is an  $n \times n$  orthogonal matrix.

On the contrary of the two former predictor formats, this method cannot be applied on the basic format of the 6 datasets. As shown on the above figure 27, this method requires to transpose first the dataset in a table with the apps predictors column-wise (format 1). Only then the TSVD can take place to perform the linear dimensionality reduction. It is explained by the fact that, we need to arrange accordingly all mobile apps per subscriber single row-wise instead of repeating them across several rows. Performing a linear dimensionality reduction technique on dataset having the basic format (See Fig. 24), wouldn't capture the interrelations between all installed apps per device, as users' apps are listed across several rows.

The number of retained dimensions should be considered as a trade-off between number of dimensions and cumulative explained variance. The above figure shows the theoretical process of passing from 8 dimensions (8 predictors) to 2 TSVDDimensions. The main drawback of this method is that, it does not counter the computation cost of the format 1 (Apps title as predictors), as it relies on it when performing the TSVD.



## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

As an example, the below graph shows the explained variance at each number of retained dimensions in the situation of a dataset with 5000 different mobile apps, thus with 5000 columns (each mobile app is a predictor). In this example, with  $k=500$ , it is possible to retain 80% of the total variance of the original data while passing from 5000 predictors to only 500. During the modelling process, tables with different number of TSVD were compared to find the 'optimal' number of TSVD for improving the models' performance while reducing the dimensionality.

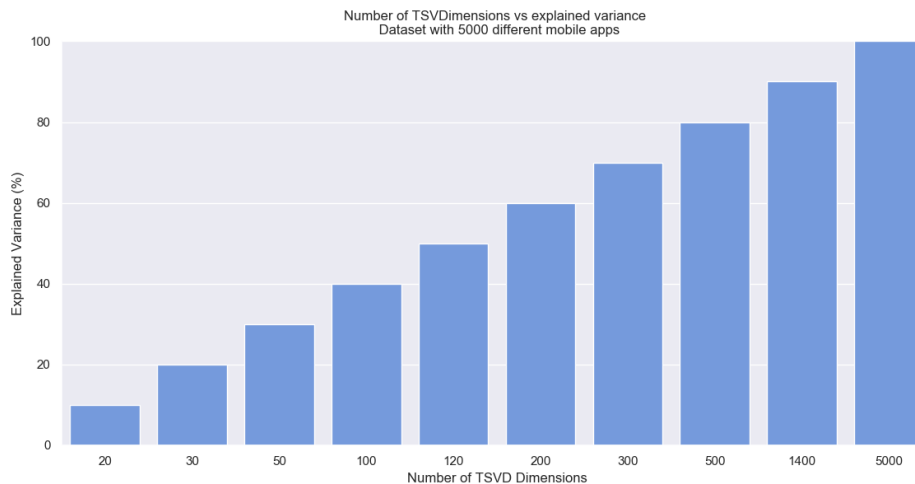


Figure 27: Plot for selecting  $k$  number of TSVD Dimensions

Reaching this point, and applying the three above explained predictor formatting on the 6 datasets, we pass to 18 tables on Python. As shown on the below figure of the Research Architecture, the objectives 1,2 and 3 are achieved. The next stage being the data analysis and visualization describes the processes applied to better understand the data and differences across the 6 generated datasets and highlight interesting insights hidden in the data.

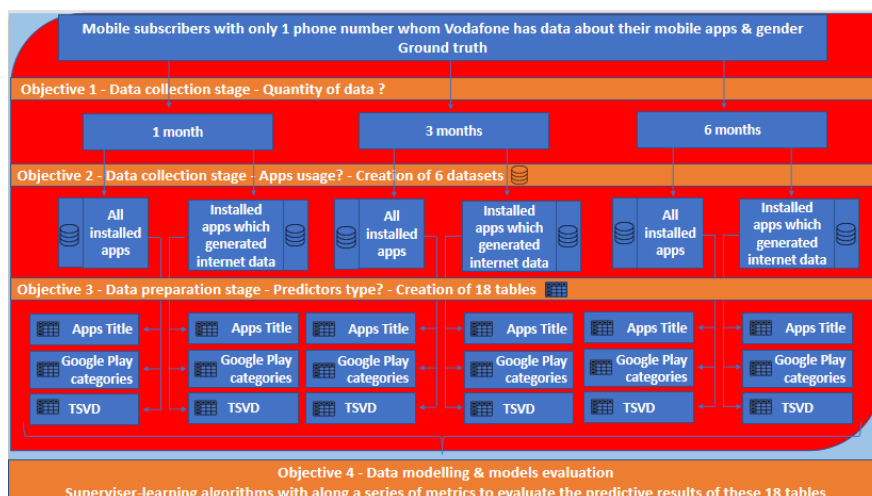


Figure 28: Research Architecture: Objective 1, 2 and 3

### **4.8. Data analysis and visualization**

To pursue with the comparative analysis, the 6 generated datasets were primarily compared based on a series of basic statistics and visualizations to grasp potential similarities and differences. Further visualizations were performed after having compared the 18 tables based on models' performance. Although explained in much depth later, it has been confirmed that the quantity of data did not affect the models performance whether with 1,3 or 6 month of data and that datasets including all subscribers' installed mobile apps led to models with 10% better results than datasets with only subscribers' installed mobile apps which generated internet data.

To begin, the 6 datasets were compared based on the below series of basic statistics:

- Total mobile subscribers
- Total mobile apps
- Average apps count per user
- Maximum apps count per user
- Standard deviation count per user
- Average categories count per user
- Maximum categories count per user
- Standard deviation count per user

In a nutshell, basic statistics for the 3 datasets with all installed apps differed considerably with the 3 datasets with only installed apps which generated internet data. Primarily, by comparing monthly-wise (1,3,6 months) the datasets with all installed apps have in general 5 times more mobile subscribers and 10 times more appearing apps than the datasets with only installed apps which generated internet data. (These aspects are also described in the data collection stage).

Secondly, basic statistics across months per same apps usage type datasets don't have striking differences. For the 3 datasets with all apps, mobile subscribers' device have roughly on average 45 apps, a standard deviation of 10 apps with a maximum of 550 apps. The counterpart 3 datasets with only apps which generated internet data, mobile subscribers' device have roughly on average 15 apps, a standard deviation of 8 apps and a maximum of 170 apps. Regarding the categories, the 3 datasets with all apps, mobile subscribers' device have roughly on average mobile apps belonging to 16 categories and a maximum of 40 categories. Its counterpart dataset with only apps which generated data, mobile subscribers' device have roughly on average mobile apps belonging to 8 categories and a maximum of 30 categories.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

Besides these aforementioned aspects, it was important to understand the role of each predictor in the prediction of the mobile subscribers' gender. To gain more insights about these predictors, I used machine learning techniques to retrieve the feature importance. In many businesses, models performance is equally important as models interpretability and knowing the importance of the chosen predictors can benefit through multiple angles; Rather it be in the model's logic understanding, model's performance improvement (removing useless features), and in interpreting why models identified such patterns based on the given data (trustworthy?).

Thus, I used the concept of Entropy to determine and attribute a value to each predictor (i.e. apps title or category) as being important in explaining the target variable (i.e. gender). In a nutshell, Entropy evaluates how much each feature contributes in decreasing the impurity in the dataset as to separate observations into sub-groups that are more similar to each other regarding the target variable (i.e. Gender). From this feature selection method, some of the mobile apps and categories with highest entropy and thus 'strongest' predictive power to infer the gender of mobile subscribers are listed hereafter:

- **Mobile apps:** Meus Resultados, Pinterest, OLX PT, Snapchat, Tinder, Standvirtual Carros, Steam, MB Way, Period Tracker,...
- **Google Play categories:** Tools, Productivity, Sports, Communication, Social, ...

Inferring the gender of mobile subscribers based on their mobile apps aggregated to the Google Play categories is mainly influenced by their count of apps per category. In short, algorithms find gender-wise patterns in the count of apps subscribers have across the Google Play categories. From the next page, I describe a series of visualizations that were developed to better understand the patterns within the data, namely, in terms of distributions, relation between apps/categories and gender and so on... Hereafter are listed few the main generated visualizations that are exemplified on the following pages using synthetic data:

- Count of mobile apps in the subscribers' device (Percentage);
- Percentage of mobile apps regarding their recurrence in the users' device;
- Total count of mobile apps per Google Play category;
- Total count of mobile apps per Google category and gender;
- Percentage of mobile apps per Google Play category and gender;
- Number of mobile apps gender-wise for one specific category: TOOL (Frequency)
- Number of mobile apps gender-wise for one specific category: TOOL (Percentage)

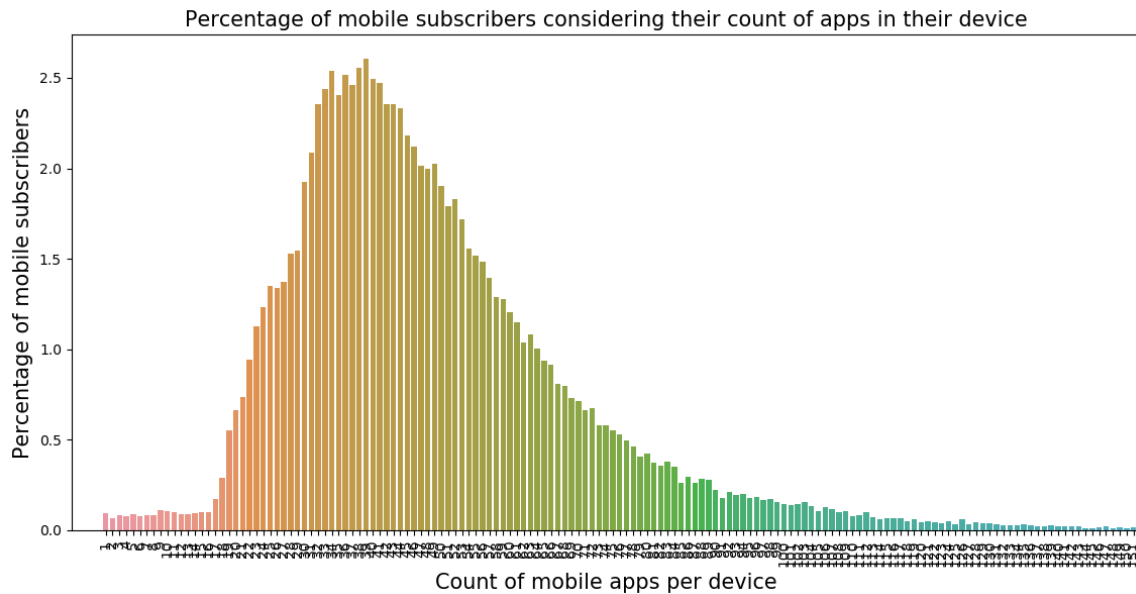


Figure 29: Count of mobile apps in the mobile subscriber's device (in percentage):

This primary figure depicts the count of mobile apps, subscribers have in their mobile device. The  $x$ -axis range from the minimum to the maximum count of mobile apps. In the above illustration, it ranges from 1 to 151. The  $y$ -axis represents the percentage of mobile subscribers. Along with the basic statistics, this visualization led to better understand the patterns in terms of count of mobile apps per subscriber's device. As demonstrated later, predictive models inferring the gender of subscribers having 5 mobile apps doesn't have the same performance as for subscribers with 100 mobile apps.

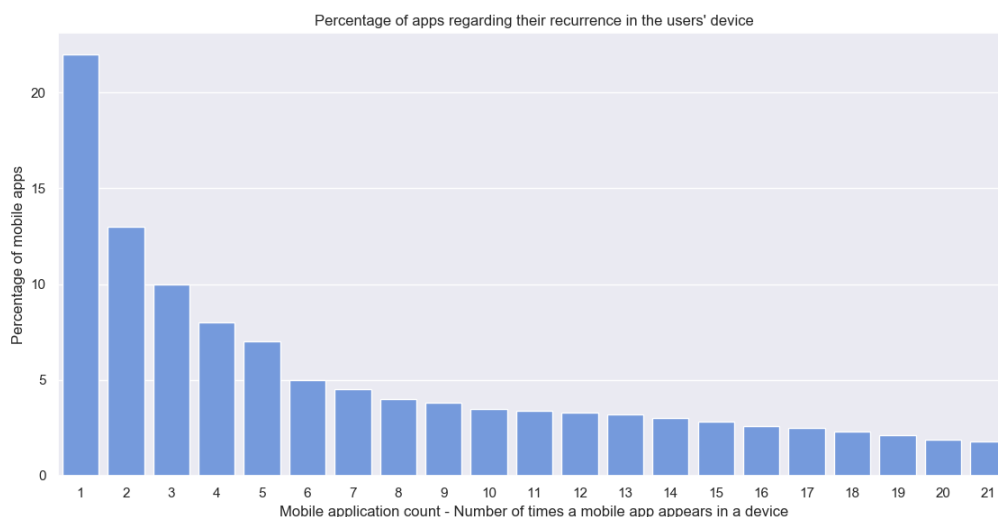


Figure 30: Percentage of mobile apps regarding their recurrence in the users' device

While the primary figure highlights the number of mobile apps subscribers have in their device, this figure focuses on the recurrence of mobile apps in a dataset (appears in how many devices).

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

The 6 datasets included from a few hundreds to a few thousands of different mobile apps. This figure shows the percentage of apps (y-axis) per recurrence count in the users' devices (x-axis). Practically, one can see that around 20% of mobile apps appear only in a single device and more than 50% of mobile apps appear in less than 5 different devices. It implies that these less recurrent mobile applications would not provide as much information when building predictive models as other mobile apps being present in much more devices. This insight led to think of a threshold to filter least recurrent mobile apps in the datasets. Considering the above illustration, filtering mobile apps appearing in less than 6 different devices would remove around 50% of mobile apps. Such threshold would decrease the computational cost when considering apps title as predictors (Format 1). To remind, the main disadvantage of this format relies on its non-scalability, as each app becomes a column.

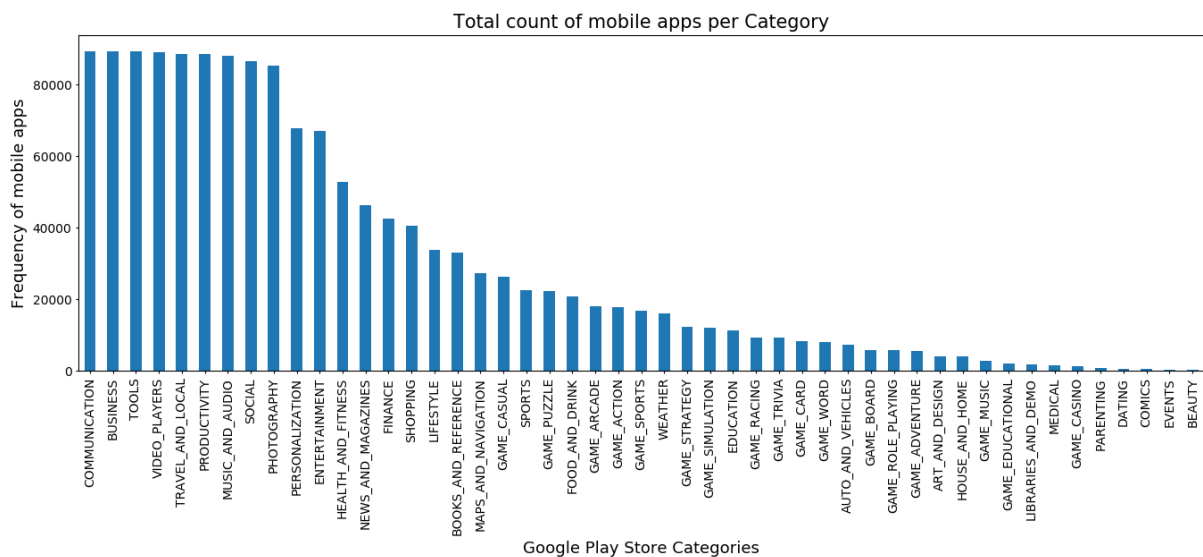


Figure 31: Frequency count of apps per Google Play Category

This figure shows the total count of mobile apps appearing in a dataset grouped to their respective google play category. Based on the above plot, one can see that most appearing mobile apps of this toy dataset belong to COMMUNICATION, BUSINESS or TOOLS categories while the least belong to BEAUTY, EVENTS or COMICS categories. Besides that, this plot exhibits in a ranking fashion, how each category is popular compared to each other.

While, this visualization is more general, the two following plots highlight the total count of mobile apps per category and gender-wise.

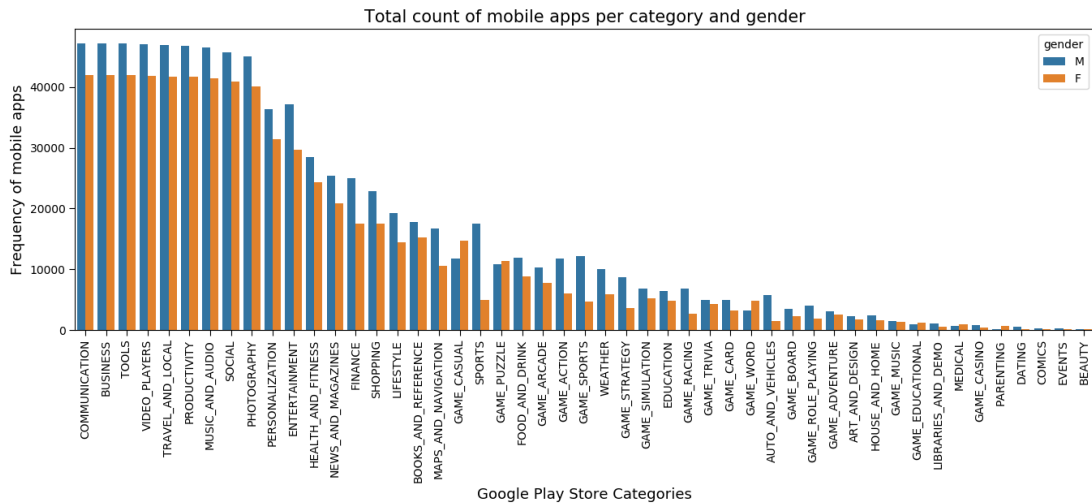


Figure 32: Frequency count of apps per Google Play Category and Gender

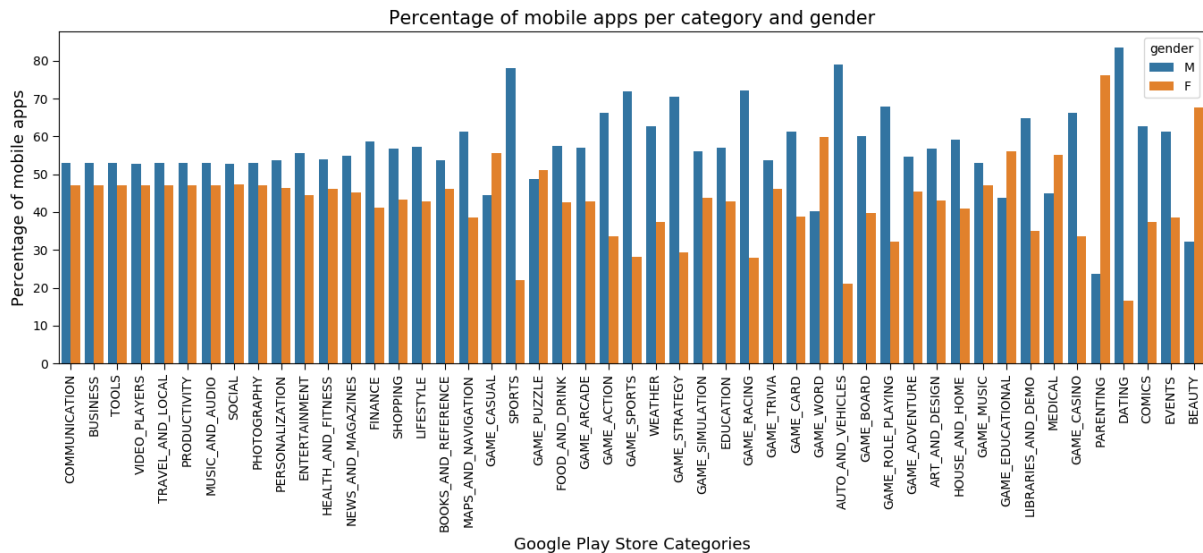


Figure 33: Percentage of mobile apps per Google play category and gender

The two above illustrations depict the frequency and percentage of mobile subscribers per category and gender. Indeed, it is appropriate to look at such figures in terms of frequency count and of proportion to not misunderstand the data. Let's take an example with the category BEAUTY. Looking at the frequency plot, it shows that relatively very few mobile subscribers of the dataset have mobile apps belonging to this category. The second plot shows clearly that around 70 % of apps belonging to this category are in the women's devices. Thus the second plot shows the percentage of male and female having apps in each category independently of other categories.

The following two visualizations show a more profound approach where we analyse per specific category the frequency and percentage of number of apps gender-wise.

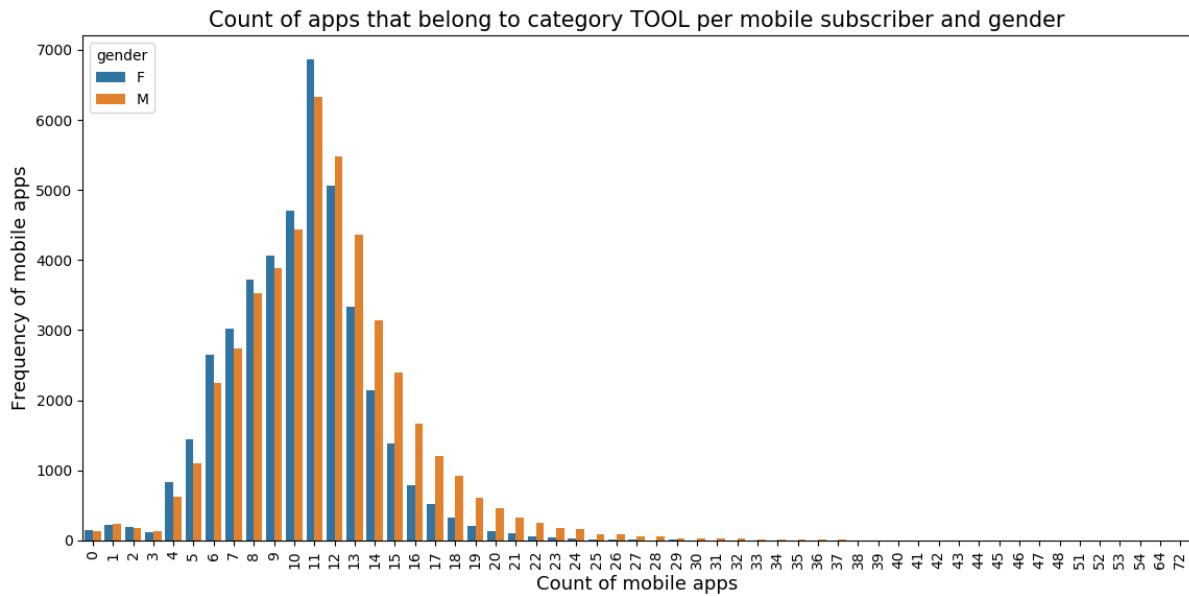


Figure 34: Count of mobile apps per mobile subscriber's gender for category: TOOL (Frequency)

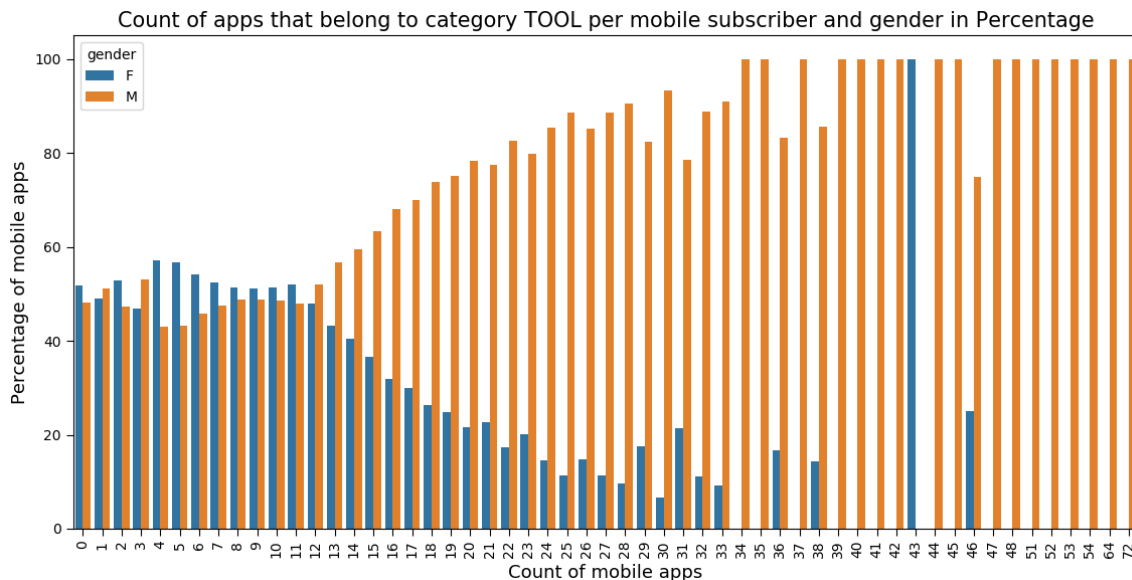


Figure 35: Count of mobile apps per mobile subscriber's gender for category: TOOL (Percentage)

These two plots show the frequency (former) and the percentage (latter) of count of mobile applications considering the subscribers' gender for the Google Play Category: Tool. Based on the above two illustrations, one can retrieve more profound insights and potential patterns about the count of mobile apps between male and female mobile subscribers across categories. For instance, the primary visualization shows that most mobile subscribers (male and female) have 11 apps belonging to the TOOL category (Mode). Analysing both the two visualizations, one can see on the former that fewer people have more than 20 mobile apps belonging to this category and that the later shows that in terms of percentage, mostly male subscribers have more than 20 mobile apps.

### **4.9. Data modelling**

The data modelling stage includes the definitions, components and reasoning behind the modelling techniques applied in this research. The primary part clarifies the concept of binary classification while the second part describes the three classification algorithms.

#### **4.9.1. Binary classification**

The essence of a classification predictive model is to approximate a 'mapping function'  $f$ , from a set of input features  $x$ , also referred as predictors, to a discrete output target variable  $y \in 1, \dots, C$ , where  $C$  represents the number of classes. Fundamentally, the objective is to develop a statistical model that finds the 'best' mapping function from historical data to make predictions on new unseen data where the same target variable is unknown. Binary or dichotomous classification is referred when the target variable has two mutually exclusive possible outcomes,  $y \in \{0,1\}$ . In this research project, we aim at solving a binary classification problem; Predict the gender of mobile subscribers, Male or Female, based on their installed mobile applications.

#### **4.9.2. Classification algorithms**

The choice of the three classification algorithms is based on previous contributions relating to a similar problem. As an initiate observational study, the core objective was to define the 'baseline' models as reference point about how these algorithms can learn from the given data. Considering the time-constraint to work on this study, no specific hyper-parameter tuning were performed throughout the data modelling stage on the following algorithms: Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF).

Looking back at the Fig. 21 – Research Architecture, one can see that this stage corresponds to the objective 4 of the data workflow performed to reach the 'comparative analysis' goal. As it can be understood, the 18 different tables, being the different experimental tables passed through the 3 aforementioned classifiers to evaluate the different predictive results.

Besides that, as explained earlier in the chapter 'Tools', Python was used to perform most steps of this data analytics workflow. More specifically, the library *Scikit-Learn* was necessary for the data modelling and the models evaluation stage.



- **Logistic Regression:**

Logistic Regression, also known as Logit Regression or Logit Model, is one of the most popular supervised-learning algorithm to solve binary classification problem. This statistical method is employed to estimate, how the probability  $P$  of a particular discrete outcome is affected by one or more explanatory variables (predictors). Logistic Regression is named after the function it uses, *the logistic function*, or also referred as *sigmoid function*, for its S-shaped curve.

In this project scenario, this algorithm's model estimates a probability for each mobile subscriber to be a *Male* or *Female*, given their installed mobile applications. To do that, it first needs to 'fit' the predictive model on given data (training set), by finding a mathematical relationship between the installed mobile applications of mobile subscribers and their gender (Male/Female). Unlike the 'Linear Regression', this algorithm outputs probability values bounded between 0 and 1, where 0 represents a class and 1 the other one. The classification threshold is a scalar-value criterion used to classify each observation into one of the two classes. As a rule of thumb, obtained probabilities are rounded to the nearest whole number (threshold value of 0.5), but the latter can vary given the problem's approach. Besides that, Logistic Regression model comes with the advantage of being somehow interpretable. The probability that the output is 1 given its input can be represented as follow:  $P(Y = 1 | X = x)$ . In practice, the value of  $P$  for each mobile subscriber is approached by the following logistic distribution:

$$p(Y = 1 | X)_i = \frac{1}{1 + e^{-(z_i)}}$$

Where:  $z_i = (\beta_0 + \beta_1 x_i + \dots + \beta_n x_n)$

$i = 1, 2, \dots, n$

When fitting the Logistic Regression to given data, its parameters (betas) can be estimated by the probabilistic framework called: Maximum Likelihood Estimation (MLE). In a nutshell, it consists in finding the coefficient values (beta 0, beta 1, till beta n) that maximize the log-likelihood function, in such a way that they 'best' explain the observed data  $X$ . In other words, estimating the parameters that maximize the probability of observing  $Y$  values based on given  $X$  data.

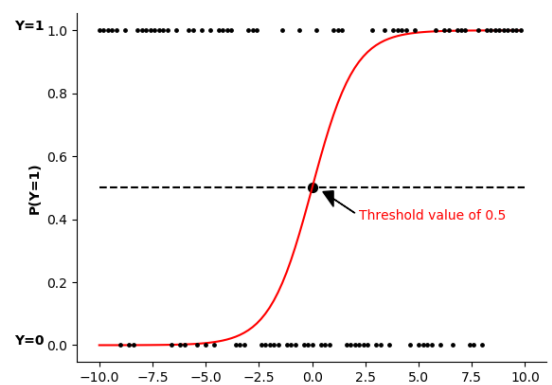


Figure 36: Logit Model (example with cut-off at 0.5)

- **Tree-based algorithms: Decision tree & Random Forest:**

The two other chosen supervised-learning algorithms for this classification task were tree-based methods: Decision Tree (DT) and Random Forest (RF). As implied by their name 'tree' and 'forest', a Random Forest is basically a collection of decision trees. Both can build classification and regression models. Let's break down these two methods to understand how they work as well as their pros. and cons.

Decision tree (DT) is a decision support tool represented as a schematic, tree-shaped diagram that maps out all possible solutions to a decision based on certain conditions. It is built top-down, starting from the root node and involves splitting the dataset into subsets that contain instances with similar values regarding the explanatory variables (e.g. apps title) with respect to the target variable (i.e. gender). A decision tree has three main parts: a root node, leaf nodes and branches. The tree starts with a single node (root), which contains the entire population of the dataset. Thus, the process consists in successively splitting the population: from a node into two or more sub-nodes, through alternative branches, each representing a value option of a chosen attribute. The length of the tree can be user-defined and/or pruned. In a Decision Tree, a leaf can mean two things: a terminal node representing an classification decision (i.e. Male or Female) or another decision node that contains a subset of the dataset that can continue to be splitted into more homogenous subsets. Path(s) from root to terminal leaf(s) represent the series of rules to potentially classify subset(s) of the population. In other words, a Decision Tree (DT) is built by breaking down a dataset into smaller and smaller subsets, while incrementally developing the model through series of rules, with set of nested if-then-else decision rules.

Regarding the splitting decision criteria, its objective is to split the root node and decision nodes into more homogenic sub-nodes to increase their 'purity' with respect to the target variable; Gini index and Entropy are such selection criteria to calculate the *information gain*, as they essentially 'determine' which feature is the best classifier (and split point for continuous feature) that would lead to more purity within the generative sub-nodes. In short, at the root node and at each decision nodes, based on the instances, it quantifies the usefulness of each attribute to ultimately determine the one that has the 'best' information gain to perform the split using the chosen feature's values. Thus, it is based on the information gain that one can decide which attribute goes into the root node or decision node. Hereafter, I describe briefly the Entropy, sometimes referred as a 'measure' of the disorder in the data and the Information gain.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

Primarily, the below Entropy formula is calculated for each feature (predictors):

$$\mathbf{Entropy}(S) = \sum_{i=1}^c -P_i \log_2(P_i)$$

Where:

- $S$  is the dataset
- $P_i$  is the proportion of instances in  $S$  in which the value of the feature is equal to  $i$ , and considering that it can have  $c$  possible values.

As an example, Entropy is calculated as follow for an attribute having two possible values considering a binary target (Female: F and Male: M):

$$\mathbf{Entropy}(S) = -P_M \log_2 P_M - P_F \log_2 P_F$$

Then, to determine the information gain of each feature, the following formula is applied:

$$\mathbf{Information\ Gain}(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where:

- $Entropy(S)$ : entropy of the dataset (regarding entropy of the target)
- $\sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$ : entropy relative to one feature
- $Values(A)$  is the set of all possible values for feature  $A$
- $S_v$  is the subset of observations of  $S$  for which feature  $A$  has value  $v$

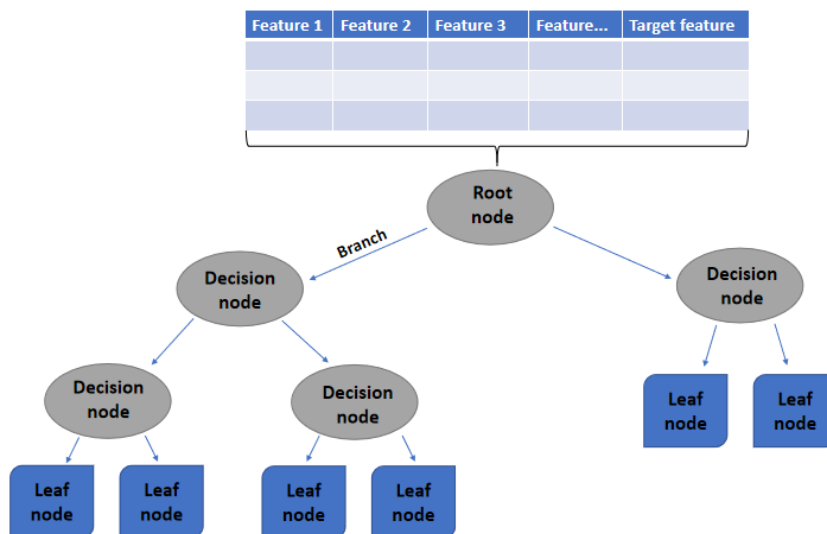


Figure 37: Visual Example of a Decision Tree

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

The Decision tree method is known to be a powerful classification algorithm given its sequential and hierarchical yet logical decision process to map out paths and outcomes, and its easiness in interpretability and visualization. Besides that, it can deal with missing values and/or outliers and is non-parametric effective as it does not hold underlying assumptions about the distribution of the data. Although, it is prone to overfitting, especially when being deep and has issues with the variance error. Indeed, the model could potentially 'memorize' given data, thus avoiding it to 'generalize' its learning to predict the target on unseen data.

Random Forest (RF), is as it sounds, a forest of trees, and to be precise of Decision trees (DTs). As an Ensemble method, its core idea is that it includes a set of individually trained 'weak learners', that are the decision trees present in the forest. Although, it surpasses or succeeds to a certain extent where simple decision tree fail, it comes with the cost of losing in interpretability, being a kind of black box. Its random component makes that there is (1) a random sampling of observations when building trees, also known as 'bagging' and (2) a random subset of features to be considered when splitting the root or a decision node. These aspects adds more randomness and diversity when building the Random Forest model. Indeed, it is demonstrated that, compared to a single decision tree, it provides a more robust model, with decreased variance and better predictive performance. Finally comes the 'Majority-vote' method, where each tree of the forest 'vote' for a classification. At the end, the forest chooses the classification having the most votes to build a 'strong learner', being more robust.

In our application, Random Forest runned considering 100 decision trees and with the entropy as a measure of purity within the dataset.

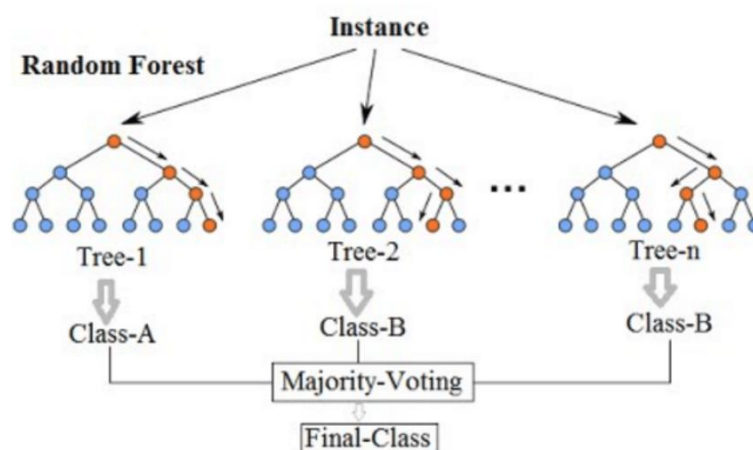


Figure 38: Random Forest

### 4.10. Models evaluation

Model evaluation is a crucial and integral part of any model development process. It consists in controlling the learning procedure and in measuring the performance of the predictive models by means of statistical metrics. Although, being only a part of the overall process, it remains a cornerstone in developing accurate, robust and reliable statistical models. The core objective being not in building models but rather in evaluating them. More concretely, this component was important to compare the predictive capabilities of the 18 tables fed on each of the three chosen classifiers. This model evaluation section describes both the learning procedure and the different evaluation metrics employed in this project.

#### 4.10.1. Model training and Testing sampling

Fitting the model on a set of data and then testing its prediction on the same data is a methodological mistake. In machine learning, it is common to split the dataset into two parts: *Training set*, where the model is trained to ‘learn’ the given data and develop a mapping function, and the *Test set*, where the model performs on ‘unseen data’ to assess its predictive capabilities on ‘unseen data’. Sometimes it has even a third part, *Validation set*, which is mostly used to fine-tune the model hyperparameters. The simple approach is commonly referred as **Hold-out method**. It is the action of dividing the dataset, with usually around 70% that goes in the Training set and the remaining 30% in the Test set. With such an approach, it is possible to assess the ‘generalization’ ability of the model when performing on unseen data. Although being fast and simple, this partitioning method has several flaws given that it relies on a unique perspective of the data (one training set) to learn the underlying patterns, leading to potential low accuracy, selection bias and lack of generalization (overfitting or underfitting) with new “unseen” data. Indeed, there could be high variability between all data points of the training set and the data points of the test set, resulting in meaningful differences when assessing the model.

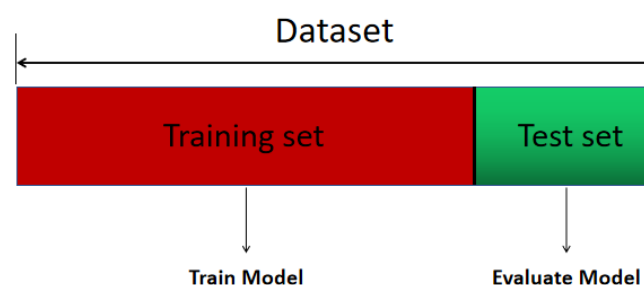


Figure 39: Hold-out method (Train/Test dataset split)

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

The **(K-fold) Cross-Validation method** came up to counter the aforementioned limitations and add more certainty to the learning process. As shown on the below figure, this partitioning method consists in:

- Splitting the dataset into  $k$  equal size subsets, called folds, where  $k$  is user-defined and where the number of iterations to perform is equal to  $k$ ;
- The Hold-out method is applied at each iteration, using  $k-1$  of the folds as training data;
- Finally, the performance measure is the average results of all test sets of the loop.

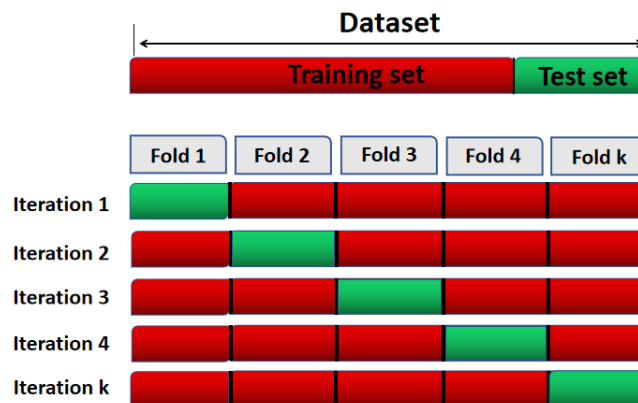


Figure 40: K-fold Cross-Validation

Compared to the simple Hold-out method, the advantage here is that every data points gets to be in a test set one time and in the training set  $k-1$  times. As a rule of thumb and empirical evidence, a value of 5 or 10 for  $k$  is preferred, although nothing is fixed. With the K-fold Cross-Validation method, the error estimation is the average of all iterations to obtain the model's performance. Considering these  $k$  iterations, this method can be  $k$  times more computationally expensive than a single Hold-out method. In a nutshell, the pros. of this method are that it reduces bias and variance, as most of the data goes into the training set and test set, at different iteration.

Finally, comes an advance version referred as **Stratified K-fold Cross-Validation**, which is the method used to evaluate all the developed models. Besides applying all the above, it ensures that the relative class frequencies is “at best” preserved in each of the generated fold. More concretely, in this binary classification problem, this method ensures that in each fold, each class (male and female) comprises approximately half of the instances.

### 4.10.2. Evaluation metrics

As already mentioned, model evaluation is closely tied to the machine learning task. Evaluation metrics are the ways to measure the performance of a statistical model. This part describes the different statistical metrics employed to measure the performance of the different Classifiers built to predict the binary target: Gender of mobile subscribers (Female or Male). These metrics were necessary to assess both training set and test set of each model and were the following ones: Classification accuracy, precision, recall, F1-score, and the AUC ROC Score.

To start, let's primarily pass through the **Confusion matrix**, which gives an intuitive understanding of how calculating the aforementioned metrics of interest. The Confusion matrix provides a summary of the correctly and incorrectly classified instances for each class (e.g. Male and Female). In binary class problems, it is a 2x2 matrix representation of 4 different combinations of actual and predicted values. Besides being useful to get an overview of the correctness of the model, its 4 inner values (TP, FP, FN and TN) are necessary to calculate other insightful metrics.

Table 25: Confusion Matrix

		Predicted Values (of statistical model)	
		Positive Class (1 or Female)	Negative Class (0 or Male)
Actual Values	Positive Class (1 or Female)	True Positive (TP)	False Positive (FP)
	Negative Class (0 or Male)	False Negative (FN)	True Negative (TN)

**True Positive (TP):** Number of instances with a positive class that were correctly classified (e.g. Number of Female person, where the model classified them correctly as Female).

**True Negative (TN):** Number of instances with a negative class that were correctly classified (e.g. Number of Male person, where the model classified them correctly as Male).

**False Positive (FP):** Number of instances with a negative class that were incorrectly classified; (e.g. Number of Male person, where the model misclassified them incorrectly as Female).

**False Negative (FN):** Number of instances with a positive class that were incorrectly classified (e.g. Number of Female person, where the model misclassified them incorrectly as Male).

The classification metrics of interest are calculated based on these four simple statistical concepts as shown on the next two pages.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

**Accuracy:** This metric determines the fraction of predictions that were correctly classified by the statistical model. For a binary target, its formula is as follow:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number Of Instances}}$$

This basic metric was used considering that the different datasets were balanced regarding the proportion of male and female person. It gives a ratio about the correct predictions made by the model. Correct predictions for both classes are treated equally, as this metric doesn't have distinction between classes (Numerator is an addition of all correctly classified instances).

**Precision:** Compared to Accuracy, this one identifies the frequency where the model correctly classified the positive class (True positive). Its definition is as follow:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positives} + \text{False Positives}}$$

This metric evaluates the model's performance at predicting the positive class (e.g. Female). In other words, it answers the question: 'Out of all positive labels, how many did the model correctly identify? Thus, a low precision score can indicate a large number of instances predicted as positive that were actually negative (False Positive).

**Recall:** This measure calculates out of all actual positive classes, how many of them were predicted correctly by the model and not wrongly as negative. Thus instead of Precision, which consider the subset of negative instances that were incorrectly classified as positive, Recall focuses on the subset of positive instances that were incorrectly classified as negative.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positives} + \text{False Negatives}}$$

This measure is also referred as Sensitivity or True Positive Rate (TPR) which is used to determine the metric ROC Curve/AUC Score. It is explained in the later metric AUC Score.

**F1 score:** It is the 'harmonic' balance between the precision and recall score of the model.

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is bounded by the [0,1] range, where 1 is ideal and 0 worst. Considering that it can be potentially difficult to compare models looking at both precision and recall, the F1-score helps measuring both at the same time through a unique metric. In short, this measure considers both false positives and false negatives. Consequently, the F1-score will be small if either precision or recall is low.



### AUC ROC score (Area Under the ROC Curve):

AUC stands for the “Area Under the ROC curve”, and to understand this scoring metric, we primarily need to grasp what is the ROC (Receiver Operating Characteristic) curve. As a recap, Classification algorithms such as Logistic Regression outputs for each predicted instance a probability value between [0,1] to belong to class 0 or 1. Although, the remaining question concerns the ‘best’ classification threshold to separate both classes. The ROC curve simplifies answering to such question. ROC is a probability curve that shows the performance of a binary Classifier (i.e. Test set) at all classification thresholds on a graph. The ROC and AUC metrics are based upon two specific metrics from the confusion matrix: Specificity and Sensitivity. To do that, the ROC curve plots these two parameters considering the model’s probability values bounded between [0,1] in a two-dimensional graphical representation:

- Along the y axis: Sensitivity or also referred as Recall or True Positive Rate (TPR):

$$TPR/Recall/Sensitivity = \frac{True\ Positive}{True\ Positives + False\ Negatives}$$

- Along the x axis: Specificity or also referred as False Positive Rate (FPR):

$$FPR = \frac{False\ Positive}{False\ Positives + True\ Negatives}$$

In a nutshell, the ROC curve of a Classifier is a summary of all confusion matrices that each classification threshold generated. It shows the trade-off between the predicted instances as false positive and true positive at various classification threshold. The dashed line representing a random Classifier implies that the model has as many true positives as false positives. For such model, the respective computed

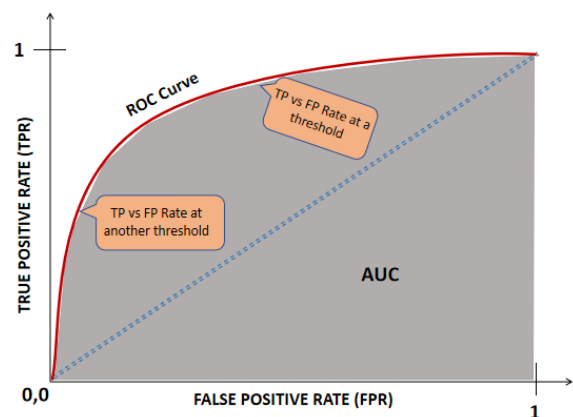


Figure 41: AUC – ROC Curve

AUC score is of 0.5. A perfect Classifier with an AUC score of 1, would have the ROC curve going along the y axis till the upper-left corner and then going towards the right-side till 1.

AUC comes in to provide an aggregated score of performance across all possible classification thresholds. It is an ideal metric to use when comparing at the same time the performance of several Classifiers. In short, AUC represents the measure of a model’s separability.

### 5. Experimental results

In this section, I describe the experimental results obtained from the 18 tables fed to the 3 supervised-learning algorithms. In line with the comparative analysis goal and thus the Research Architecture, a series of aspects were analysed throughout the study and in the data modelling stage, namely, in terms of quantity of data, apps usage, predictors and classifiers. Hereafter, I review these aspects point by point before highlighting the best results obtained. Finally, I recommend potential directions for improving processes and models' performance to predict the gender of mobile subscribers based on their installed mobile applications.

The primary aspect consisted in evaluating the relation between dataset size and models' performance. To do so, 1, 3 and 6 months of data were considered for comparison analysis. For more clarity about the differences between the datasets in terms of quantity of data, I advise to review the first page of the data collection stage. (Malmi & Weber, 2016) demonstrate in their paper that increasing the training set led to better models performance, passing from a 100 users to 2300 users. Although, the experimental results didn't yield the same conclusion. It is explained by the fact that the dataset with least number of mobile subscribers (1 month of data with only apps which generated internet data) already included roughly 14.000 subscribers. Conclusion is that models performance are not affected differently when using 1,3 or 6 months of data as experimental results yield similar performance whether using 1,3,6 months of data.

The second aspect concerned mobile apps that could be installed but not used and thus not being really 'representative' of the gender of the mobile subscriber. Considering that nowadays increasing mobile apps are accessing mobile internet when used, the proxy approached was of removing mobile apps installed but not generating internet data and considering those as 'not used'. For this experimentation, we kept in mind that there are still many mobile applications that do not necessarily require internet to be used. The experimental results demonstrate that models using the datasets with only mobile apps which generated internet data are 10% less performant across the different metrics (accuracy, precision, recall, F1 and roc) than the datasets with all installed mobile apps per subscriber. Conclusions of this experimentation are that this proxy did not improve the models' performance compared to models which consider all installed apps of subscribers and that some apps that work offline are good predictors.

In a nutshell, models have better performance when considering all installed mobile apps of mobile subscribers but are not influenced whether with 1,3,6 months of data.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

The third aspect concerned the predictors rather it be in the titles of installed mobile apps or in their aggregated count per Google play category or in the TSVD, a linear dimensionality reduction (applied on top of sparse table with apps title as predictors). To recap, when retrieving the data from the BDP, the format was not adequately arranged. Indeed, mobile apps per subscriber were listed across several rows. With such a formatting, the algorithms would not study the interrelations of all installed apps subscribers have in their mobile device but instead would study one-by-one the relation between each mobile app and the subscribers' gender. To fix that, it was required to transpose each table towards sparse format: mobile subscribers row-wise and predictors column-wise. These three predictors and the formatting approach are based on previous contributions which demonstrate encouraging results. Each of these three methods comes up with pros. and cons. in terms of computation and model's performance.

Experimental results yield best models performance when considering all installed mobile apps of subscribers and independently of the quantity of data (1,3,6 months) as follow:

1. Mobile apps title as predictors: F1: 0.75, ROC: 0.73
2. TSVD with 500 dimensions (+/- 80% explained variance): F1: 0.74, ROC: 0.73
3. Google play categories: F1: 0.70, ROC:0.69

Although mobile apps title yield the best results, this method is not easily scalable. Indeed, predictors are listed column-wise in each of the sparse table. It implies that the sparse table is directly proportional to the number of unique apps of a dataset. (e.g. 10 apps -> 10 columns). This scalability limitation comes up with the need in adapting the memory and the computation. As of personal experience, my computer could run locally sparse tables with a maximum of roughly 5000 mobile apps column-wise before crashing. To remind, the dataset with 6 month of data and all installed apps had more than 7000 mobile apps. To counter that, I had to set thresholds to remove mobile apps that would appear in very few different mobile devices (e.g. less than 6 devices) to decrease the computation.

The pros. of the Google Play store relies in its scalability as mobile apps per subscriber are easily aggregated to their google play category (49 categories). It implies that tables/dataframes have a maximum of 50 columns (49: categories + 1: gender). The main disadvantages of this method is that the dimensionality reduction is limited to 49 categories and that predictive power held in specific mobile apps is lost (i.e. Meus Resultados has a higher popularity among male while Pinterest among female). Instead with the Google Play categories, algorithms study the relation between the target (i.e. gender) and count of apps per category.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

Compared to the Google Play categories method, the TSVD has the advantage that the dimensionality is not reduced to a maximum of 49 categories but is rather user-defined. Let's take the example of the dataset with 6 months of data with all installed mobile apps. This dataset included around 7000 mobile apps. Experimental results shows that reducing the 7000 mobile apps to a minimum of 500 TSVD dimensions yield a F1-score of 0.74 and ROC-score of 0.73. Retaining 500 TSVD dimensions allow to keep 80% total variance of the original dataset with 7000 mobile apps. However, the cons. of this method is that it can only be applied on top of sparse tables/dataframe and more precisely with the apps title as predictors (i.e. TSVD works only on sparse matrices). This disadvantage implies that this method doesn't easily and entirely counter the computation cost of having apps title as predictors.

Finally, the last aspect concerned the three different supervised-learning algorithms chosen to study the relation between the set of predictors with the target variable (i.e. subscribers' gender) The Classifiers are Logistic Regression, Decision Tree and Random Forest. They were chosen based on previous studies which rely on these three specific algorithms. Given the time constraint to work on this research project (around 2 months), no hyper-parameter tuning were performed but only the basic parameters of each of the Classifiers to set the baseline. Besides being already used in the literature, they were chosen because of being known algorithms used both in the academia and in the industry. Logistic Regression has the pros. of generating fast, performant models and yet interpretable. Decision Tree as well but with the specific advantaged of the possibility to be visually analysed. Random Forest is mainly known for its specific performance in reducing variance and creating robust models. For both decision tree and random forest, the measure of purity chosen was the 'Entropy'. Random Forest was runned always considering a 100 trees.

Below are the best results obtained during the learning phase (training set) and evaluation on 'unseen data' (test set) considering 6 months of data and all installed apps:

- **Mobile apps title as predictors:**

Table 26: Experimental results with mobile apps title as predictors

	Training set					Test set				
	Accuracy	Precision	Recall	F1	Roc	Accuracy	Precision	Recall	F1	Roc
LR	0.77	0.80	0.75	0.78	0.77	0.73	0.76	0.71	0.74	0.73
DT	0.98	0.99	0.98	0.98	0.98	0.64	0.67	0.65	0.66	0.64
RF	0.98	0.99	0.98	0.98	0.98	0.73	0.74	0.75	0.75	0.73

- **Google Play categories:**

Table 27: Experimental results with Google Play categories as predictors

	Training set					Test set				
	Accuracy	Precision	Recall	F1	Roc	Accuracy	Precision	Recall	F1	Roc
LR	0.69	0.73	0.65	0.69	0.69	0.69	0.73	0.66	0.69	0.69
DT	0.97	0.99	0.96	0.97	0.97	0.61	0.63	0.62	0.62	0.61
RF	0.97	0.98	0.97	0.97	0.97	0.69	0.71	0.70	0.70	0.69

- **TSVD with 500 dimensions (explained variance of 80% of the original dataset):**

At the time of the report redaction, I didn't have any more all of the obtained results using the TSVD method. Although, below are the best results obtained when retaining 500 dimensions for a total explained variance of 80% of the original dataset counting around 7000 mobile apps. Hereafter are the metrics score on the test set using the Logistic Regression:

Table 28: Experimental results with the TSVD method (500 dimensions)

Algorithms	Test Set				
Metrics	Accuracy	Precision	Recall	F1-score	Roc Score
Logistic Regression	0.74	0.74	0.75	0.74	0.73

Based on the two first figures, one can see that Decision trees and Random forest tend to overfit during the learning phase compared to when performing on 'unseen data'. Logistic Regression doesn't show this flaw during the learning phase. This 'memorization' issue could be countered through hyper-parameter tuning (e.g. prune or limit the profoundness of trees models). Although, such steps weren't performed as reaching the end of the internship (end of February).

Based on the performed experiments, I can highlight that Logistic Regression distinguishes itself from the two other algorithms for its good performance, fast computation, interpretability and interesting generalization capabilities. However, at this stage of the study, it is important to consider that no hyper-parameter tuning were performed, nor other algorithms were considered. I believe that the aspects to be looked in much depth for future work should be in the data modelling stage. Besides that, applying machine learning models on infrastructure(s) with more computational capabilities (e.g. distributed computing environment or powerful local computer).

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

Finally the below visualization depicts how well the models can predict the gender of the mobile subscribers considering the count of mobile apps they have in their device. Indeed, the performance of the models vary at predicting their gender depending if the mobile subscribers has installed five mobile apps or a hundreds mobile apps. The below visualization considered the Logistic regression as classifier.

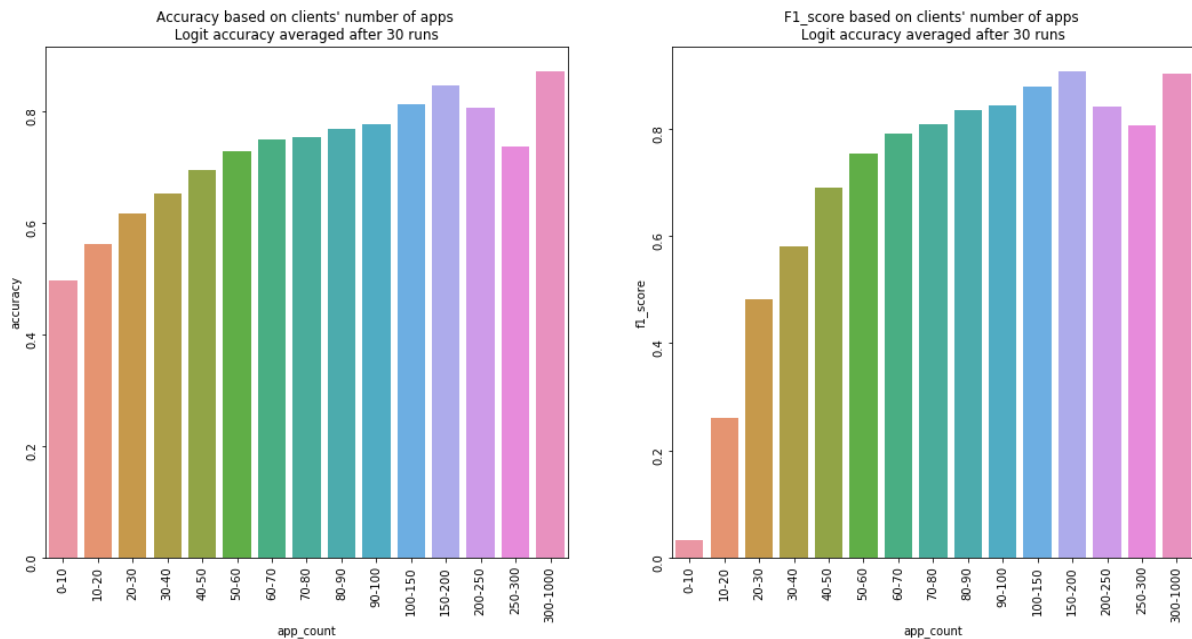


Figure 42: Model's performance considering the number of apps subscribers have in their device

The above figure demonstrates both the model performance in terms of accuracy (left) and F1-score (right). Performance is at its best for mobile subscribers having in between 150-200 mobile apps (Accuracy and F1-score of 0.80) and is at its worst for mobile subscribers having in between 0 to 10 apps (Accuracy: 0.50, F1-score: 0.05) in their device. One can see a decrease in performance for mobile subscribers having in between 200 to 300 mobile apps before seeing another small increase. It may be explained by potential noise in the subscribers' device in terms of predictors. The count of mobile apps, subscribers have in their device is an important factor to consider when predicting their gender, as it affects considerably the performance of the model.

### 6. Conclusions

This second research project being about the predictability of Vodafone mobile subscribers' gender based on their installed mobile applications retrace most steps of the data science lifecycle. Originally motivated by the interest in filling the gap of missing demographic traits of mobile subscribers, it led to perform this initiative observational study. In today's highly competitive world, understanding the customers is important and their demographics are crucial components to help improve customized services and targeted advertising. This observational study let me put a number of aspects in perspective, namely, in terms of scientific curiosity, domain knowledge, data understanding, data reliability, data privacy, work reproducibility, models scalability and in the importance of visualizing the 'big picture' of any data science project.

Real-work research projects do not always follow logical nor linear workflow. Instead, they require lots of going back and forth to understand and define clear and robust paths along the continuity of the project workflow. Estimating the time required for each step of the project workflow can quickly become complex as challenges arises at any point throughout the research. In this study, I performed a comparative analysis of several factors that would affect the models performance in inferring the gender of the mobile subscribers. Defining the data workflow for the comparative analysis and preparing the data in suitable datasets took most time of this study. Finally, during the defined timeframe, I defined a roadmap solution which encompasses a series of best practices to perform such predictive task. Besides that, the BDP counts a new table with details about more than 40.000 mobile apps from Google Play Store.

I would put to the fore the following recommendations for further work. Primarily, the experimental results defined a primary baseline in terms of models performance one can achieve. Improving models could go in a number of ways such as in considering tuning the parameters of the chosen classifiers. Besides that, I advise to go towards 'hybrid models' which would consider both 'quality' and 'quantity' in terms of predictors. Indeed, considering both apps title with the highest entropy (i.e. good predictors) and the count of mobile apps per category/dimension (e.g. 49 google play categories or user-defined number of TSVDimensions). Finally, the last visualization demonstrate that models performance vary considerably when inferring the gender of mobile subscribers having 10 mobile apps in their device or having more than 100 mobile apps.

## Bibliography

- Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunication service industry. *Elsevier*.
- Armstrong, M. (2019, April 16). *All of the data created in 2018 is equal to....* Retrieved from Statista: <https://www.statista.com/chart/17723/the-data-created-last-year-is-equal-to/>
- Batran, M., Mejia, M. G., Kanasugi, H., Sekimoto, Y., & Shibasaki, R. (2018). Inferencing Human Spatiotemporal Mobility in Greater Maputo via Mobile Phone Big Data Mining. *International Journal of Geo-Information*.
- Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved Response to Disasters and Outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study Haiti. *PLoS Medicine*.
- Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., & Sbodio, M. L. (2013). AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. *IEEE Transactions on Visualization and Computer Graphics*.
- Blondel, D. V., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science* .
- Brea, J., Burrioni, J., & Sarraute, C. (2015). Inference of Users Demographic Attributes based on Homophily in Communication Networks. *Netmob 2015 - Fourth Conference on the Scientific Analysis of Mobile Phone Datasets*. Cambridge: NetMob 2015.
- Bughin, J. (2016). Reaping the benefits of big data in telecom. *Journal of Big Data*, 5.
- Bughin, J. (2016, June). Telcos: The untapped promise of big data. *McKinsey Quarterly*.
- Cisco. (2019, February 18). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper*. Retrieved from Cisco: <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html>



## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

- Clement, J. (2019, September 18). *Annual number of global mobile app downloads 2016-2018*. Retrieved from Statista: <https://www.statista.com/statistics/271644/worldwide-free-and-paid-mobile-app-store-downloads/>
- Clement, J. (2019, October 09). *Number of apps available in leading app stores 2019*. Retrieved from Statista: <https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/>
- Clement, J. (2019, October 9). *Number of apps available in leading app stores as of 3rd quarter 2019*. Retrieved from statista: <https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/>
- Columbus, L. (2017). What's New In Location Intelligence For 2018? *Forbes*.
- Costa, P., & McCrae, R. R. (1992). Reply to Ben-Porath and Waller. *Psychological Assessment*.
- Cuttone, A., Lehmann, S., & González, C. M. (2018). Understanding predictability and exploration in human mobility. *EPJ Data Science*.
- Dong, Y., Yang, Y., Tang, J., Yang, Y., & Chawla, N. V. (2014). Inferring User Demographics and Social Strategies in Mobile Social Networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM.
- Eckert, P. (1997). Gender and sociolinguistic variation. (i. J. ed., Ed.) *Oxford: Blackwell*.
- Fiadino, P., Ponce-López, V., Antonio, J., Torrent-Moreno, M., & D'Alconzo, A. (2017). Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era". *The Workshop*, (pp. 43-48). doi:10.1145/3098593.3098601
- Frías-Martínez, V., Frías-Martínez, E. , & Oliver, N. (2010). A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records. *Artificial Intelligence for Development, Technical Report SS-10-01*. Stanford: 2010 AAAI Spring Symposium.
- Hu, J., Zeng, H.-J., Li, H., Niu , C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. *WWW 2007*.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

- Hyken, S. (2018, July 15). Customer Experience Is The New Brand. *Forbes*. Retrieved from <https://www.forbes.com/sites/shephyken/2018/07/15/customer-experience-is-the-new-brand/#3c8aa53e7f52>
- Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. (2012). Human Mobility Modeling at Metropolitan Scales. *MobiSys '12 Proceedings of the 10th international conference on Mobile systems, applications, and services*.
- Jiang, Shan, Ferreira, J., Jr., & Gonzalez, C. M. (2017). *Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A case Study of Singapore*. IEEE Transactions on Big Data.
- Jones, K. H., Daniels, H., Heys, S., & Ford, D. V. (2019). Toward an ethically founded framework for the use of mobile phone call detail records in health research. *JMIR Mhealth Uhealth*.
- Knuth, D. E. (n.d.). *Frequently Asked Questions*. Retrieved from cs-faculty-standford: <https://www-cs-faculty.stanford.edu/~knuth/faq.html>
- Kujala, R., Aledavood, T., & Saramäki, J. (2016). Estimation and monitoring of city-to-city travel times using call detail records. *EPJ Data Science* .
- Malmi, E., & Weber, I. (2016). You are what Apps you use: Demographic prediction based on User's Apps. *International Conference on Web and Social Media (ICWSM)*. Cologne: AAAI Press.
- Malmi, E., & Weber, I. (2016). You are what you use: Demographic prediction based on user's apps. *ICWSM*.
- Marr, B. (2015, September 30). *Big Data: 20 Mind-Boggling Facts Everyone Must Read*. Retrieved from Forbes: <https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#787a606917b1>
- Pugna, I. B., Dutescu, A., & Stănilă, O. G. (2019). Corporate Attitudes towards Big Data and Its Impact on Performance Management: A Qualitative Study. *sustainability*.
- Ricciato, F., Widhalm, P., Craglia, M., & Pantisano, F. (2015). *Estimating population density distribution from network-based mobile phone data*. Joint Research Centre.

## CHAPTER 3. Gender prediction from subscribers' installed mobile apps

---

- Santos, C. (2017, July 06). Big Data: O grande desafio é "extrair valor" dos dados. (A. Laranjeiro, Interviewer) Retrieved from <https://www.jornaldenegocios.pt/negocios-iniciativas/portugal-digital-awards/detalhe/big-data-o-grande-desafio-e-extrair-valor-dos-dados>
- Seneviratne, S., Seneviratne, A., Mohapatra, P., & Mahanti, A. (2014). Your installed Apps reveal your gender and more! *Mobile Computing and Communications Review*, 1-8.
- Seneviratne, S., Seneviratne, A., Mohapatra, P., & Mahanti, A. (2014). Your installed Apps Reveal Your Gender and More! *ACM SIGMOBILE Mobile Computing and Communication Review*.
- Shibasaki, R. (2017). *Call Detail Records (CDR) analysis: Republic of Guinea*. International Telecommunication Union (ITU).
- Statista. (2019). *Number of mobile phone users worldwide from 2015 to 2020 (in billions)*. Retrieved from Statista: <https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/>
- Statista. (2019, July 22). *Percentage of all global web pages served to mobile phones from 2009 to 2018*. Retrieved from Statista: <https://www.statista.com/statistics/241462/global-mobile-phone-website-traffic-share/>
- The World Bank. (2018). *Mobile cellular subscriptions (per 100 people)*. Retrieved from data.worldbank: <https://data.worldbank.org/indicator/IT.CEL.SETS.P2>
- Vodafone. (2019). *About Vodafone Group*. Retrieved from Vodafone: <https://www.vodafone.com/about>
- Wesolowski, A., Eagle, N., Tatem, A., Smith, L. D., Noor, A., Snow, W. R., & O Buckee, C. (2012). Quantifying the Impact of Human Mobility on Malaria. *Science*.
- Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*.
- Zhong, Y., Yuang, N. J., Zhong, W., Zhang, F., & Xie, X. (2015). You are where you go: Inferring demographic attributes from location check-ins. *8th ACM International Conference on Web Search and Data Mining* (pp. 295-304). ACM .

