Vera Lúcia Guerra Estanqueiro

WHEN AND HOW TO CONTACT YOUR CUSTOMER

1 2 9 0
UNIVERSIDADE D
COIMBRA

1 2 9 0

UNIVERSIDADE D
COIMBRA

Vera Lúcia Guerra Estanqueiro

WHEN AND HOW TO CONTACT YOUR CUSTOMER

Dissertation in the context of the Master in Informatics Engineering, Specialization in Intelligent Systems, advised by Professor Penousal Machado and Professor Nuno Lourenço presented to Faculty of Sciences and Technology / Department of Informatics Engineering.

October 2021

Faculty of Sciences and Technology

Department of Informatics Engineering

When and How to contact your customer

Vera Lúcia Guerra Estanqueiro

Dissertation in the context of the Master in Informatics Engineering, Specialization in Intelligent Systems advised by Professor Penousal Machado and Professor Nuno Lourenço presented to the
Faculty of Sciences and Technology / Department of Informatics Engineering.

October 2021

This page is intentionally left blank.

Acknowledgements

First of all, I would like to thank my advisors, Professor Penousal Machado and Professor Nuno Lourenço, for all the support throughout this year. Their collaboration, opinions and constructive criticism were very important for my learning process, both professionally and personally.

I would also like to thank all the people I had the chance to meet while working at this project, especially Ricardo Paiva for your availability and cooperation throughout this entire project

Alexandre, Martinho, Sara and Caio thank you for being with me during the endless hours spent in the department working on projects. With your support and teamwork I was able to improve my self and overcome all adversities. It was your friendship that created so many good memories which I will remember for years.

An heartfelt thank you to Sofia Lebreiro, for being my companion on this journey. You always found a way of encouraging me when I felt down or stressed. Your patience, humor and optimism have made these last years an enjoyable experience that I will always be grateful for.

To my little sister, Lídia, thank you for being on the other side of the phone whether it was to exchange good laughs or to let off steam about our academic life and personal events. Thank you for your love and advices.

Last but not the least, a very special thanks to my parents who always believed in me. Your support help me to maintain my head up even when the strengths started to fail. Without you this path would be impossible to achieve. Thank you for everything I am today!

This work is funded by national funds through the FCT — Foundation for Science and Technology, I.P., within the scope of the project CISUC - UID/CEC/ 00326/2020 and by European Social Fund , through the Regional Operational Program Centro 2020, and the contract MEO-UC-ALB in the context of the protocol between Altice and the University of Coimbra.

This page is intentionally left blank.

Abstract

Marketing campaigns in companies are vital to communicate with customers. Regardless their operational contexts, companies rely on Customer Relationship Management (CRM) frameworks based on data mining methods to maintain and nurture the connection with their clients. Data mining techniques have gained popularity due to its wide applicability in several CRM systems and its predictive models are used to support decisions. When comes to the time to contact a customer, the availability and the channel of contact are two important components for its success.

This work aims at tackling two issues that emerge when an organization needs to contact a client. In concrete, we developed two intelligent systems: the first to predict the best hour to contact a client, and the second to select which channel (e.g., voice call, Short Message Service (SMS), Electronic Mail (E-MAIL)) is more appropriate. Both challenges make use of historical data communications to predict customers' availability to receive a contact from a service provider and the best way to establish the contact. We use real data from a telecommunications company. We developed approaches based on unsupervised and supervised learning algorithms in order to compare their results. Regarding the problem of identification of the best contact hour, this work shows gains in almost every hour of the day when compared with the current success rate achieved by the company. In turn, the identification of the best contact channel problem obtained better results than randomly selecting a channel to establish the contact.

Finally, the practical applicability of the study led us to discuss not only the methodology adopted but also the approaches from a trade-off perspective. We think this analysis achieved relevant results to balance speed and success rate.

Keywords

Data mining; Customer Relationship Management (CRM); CRISP-DM; Telemarketing; Decision support systems (DSS).

This page is intentionally left blank.

Resumo

As campanhas de marketing são formas essenciais de comunicação entre as empresas e os seus clientes. Independentemente dos sectores operacionais, estas dependem de sistemas de Gestão de Relacionamento com Clientes (GRC), baseadas em métodos de mineração de dados, de forma a manter uma boa relação com seus clientes. Atualmente, a técnica de mineração de dados está a ganhar bastante popularidade devido à sua aplicabilidade em diversos sistemas de GRC. Esta utiliza modelos preditivos para auxiliar o processo de suporte à decisão. No momento de iniciar um contato com o cliente, tanto a disponibilidade como o canal de contato são dois componentes importantes para o sucesso da comunicação.

Este trabalho visa abordar duas questões que surgem quando uma organização necessita de entrar em contacto com os clientes. Concretamente, desenvolvemos dois sistemas inteligentes: o primeiro para prever a melhor hora para contactar um cliente e o segundo para selecionar qual é o canal (por exemplo, chamada de voz, sms, e-mail) mais apropriado. Ambos os desafios fazem uso de dados baseados nas comunicações dos clientes de forma a prever tanto a sua disponibilidade para a recepção de um contato, como o melhor canal para o fazer. Através da utilização de dados de uma empresa de telecomunicações desenvolvemos várias abordagens usando algoritmos de aprendizagem supervisionada e não supervisionada de forma a comparar os seus desempenhos. Em relação ao problema de identificação do melhor horário de contacto, este trabalho mostra ganhos em quase todas as horas do dia quando comparado com a percentagem de sucesso atualmente alcançada pela empresa. Por sua vez, a identificação do melhor canal de contato obtém melhores resultados do que uma seleção aleatória.

Por fim, a aplicabilidade do estudo leva-nos a adicionar uma descrição da metodologia adotada e, também, promove uma discussão acerca da análise das abordagens tendo em conta uma perspectiva de *trade-off*. Consideramos que esta avaliação alcançou resultados relevantes que demonstram um equilíbrio entre a velocidade dos modelos e a sua taxa de sucesso.

Palavras-Chave

Mineração de dados; Gestão de Relacionamento com o Cliente (GRC); CRISP-DM; Telemarketing; Sistemas de Suporte à Decisão.

This page is intentionally left blank.

Contents

This page is intentionally left blank.

Acronyms

ANN Artificial Neural Network. 7

AUC Area under the ROC Curve. 13

CDRs Call Detail Records. 16, 27

CRISP-DM Cross Industry Standard Process for Data Mining. 11

CRM Customer Relationship Management. v, 1, 2, 14, 23, 50

DSS Decision support systems. v

E-MAIL Electronic Mail. v

GPRS General Packet Radio Service. 16

GRC Gestão de Relacionamento com Clientes. vii

MLPNN Multilayer Perception Neural Network. 13

NIC Client identification number. 16

NN Neural Network. 13

ROC Receiver Operating Characteristic. 13

SMS Short Message Service. v

This page is intentionally left blank.

List of Figures

List of Tables

This page is intentionally left blank.

Chapter 1

Introduction

Telecommunication companies depend on marketing campaigns to let customers know the services they have to offer. However, the increasingly number of advertisements over time has reduced their effect on the general public [3]. Not only does the pressure to compete in a demanding market have an impact on the approach marketing managers have to decide upon, but also, timing is everything. There are several challenges regarding when and how to contact the customer. Depending on the strategy defined, interactions between the company and clients can be considered as intrusive from the point of view of the latter. Both timing and channel are two important factors to take into consideration when defining the best approach for marketing campaigns.

Companies must know their costumers and how to reach them to enhance their business. Furthermore, a good relationship with their clients must be maintained. And to give an answer to these needs, the concept of Customer Relationship Management (CRM) emerges. The CRM is a strategy used by companies to manage all information system related to customer interactions where the client is the center of all business processes [4]. It was designed to anticipate consumers' needs, and to define management strategies to maintain the relationship between the customer and the organization. This technique should be able to help them to match or surpass the competition in the industry.

The vast amount of information collected through CRM can be used to support managerial decision making [5] and a decision based on the right information is influenced by all the companies' available data. Depending on their operation sector, information collected will contain properties about the organization's background, and it must be analyzed according to its nature and context. For instance, a dataset from a telecommunications company as various sources of information include, but are not limited to: Call Detail Records, which contain information about the success of the call, as well as duration, caller, and receiver.

Data mining is a powerful tool capable of transforming data into valuable information with a business context application. It is defined as the field of discovering novel and potentially useful information from large amounts of data by [6, 7]. The same reference also adds that the term data mining covers a wide variety of data analysis procedures with roots in several domains, including statistics, machine learning, pattern recognition, information retrieval, and others. So, data mining can be used to address the primary challenges raised when communication with the client must be established: **How and when should contact a customer?**

In the scope of this work, we are interested in determining the best time to contact customers and the best channel of contact. For instance: Should we contact a specific customer during working days? And at what time of the day? Should it be through a voice call, SMS or through an e-mail?

We will use a data mining approach to create a system to meet our goal and try to answer the previously raised questions.

1.1 Motivation

This work has three main motivations. Two of them are related to a practical business context. And the other is motivation from an academic point of view.

First, there is the need for a mechanism to solve communication problems between telecommunications companies and its customers. An efficient contact management reduces the feeling of intrusiveness commonly caused by the excessive number of contacts. Therefore, the goal behind this motivation is to improve the number of answers per contact attempts, in such a way that the process of reaching out to a client can be considered less stressful for both parties.

Then, the development of a framework capable of answering questions such as "When?" and "How?" when it comes to contacting a client has major practical application to the business world. It is true this work is based on data collected directly from a telecommunication company and all the initial analysis of understanding the whole data had to take into account its context and its nature. Nevertheless, the CRM is a strategy commonly used by companies, independently of their operational market to support managerial decision making. This study proposes a possible approach to efficiently extract valuable information from data collected from a CRM to determine the best time to contact customers and the best channel of contact. So, our motivation is in fact the development of an approach that, in general, can be considered as a worthwhile asset to add to the information systems of an organization in such a way they will retain potential customers and maximize their customer value.

Finally, the results presented in this work could be relevant to the field of Decision Support Systems. Answering a set of secondary questions such as: **Which means a successful contact? How can we represent our output? Which predictive models are appropriate for this problem?** will add value to this study and hopefully be a starting point to further researches.

1.2 Objectives

This work contains two goals. We are interested in determining the best time to contact a customer and we want to discover which channel is the most appropriate to establish the contact. Particularly, both problems require the development of a ranking system in order to help companies to better manage their resources.

The practical applicability of this study is a heavy concern during all stages of this project as we aim to produce models that can be used in a business context. Therefore, the two main objectives were sub-divided into the ones described bellow. By responding to them, we are obtaining answers to our primary target.

- Investigate how can we use historical communications' information to make predictive hour and channel models;

- Compare the gains/losses between the current process of contacting clients with an implementation based on predictive models.

1.3 Contributions

The present work contains contributions related to both problems Identification of the Best Contact Channel and Identification of the Best Hour of Contact. Additionally, we discuss a possible methodology to solve the challenges we are interested in. So, a summary of the main contributions is presented below.

- A set of predictive models to make an hour ranking per client;
- A predictive model to make a channel ranking per client;
- A description of an adapted CRISP-DM methodology that can be applied to a data mining development process based on data extracted from a CRM context.
- A set of intermediary code deliveries that are being integrated into the organization's systems to use this works's discoveries in their products.

1.4 Structure

This document is structured as follows.

Chapter 2 provides to the reader an understanding of the Context of this study and the concepts that emerge from it. Also, it aims to approach some methodologies that will be used during this project. Some work reviews will be addressed as well, and finally, the limitations of related studies developed in the area will be discussed.

Chapter 3 describes all development phases involved in this work, how they were connected, and which methodology was adopted. Additionally, a description of the source dataset and its transformations are provided in this section.

Chapter 4 of this document presents a formal definition to the Identification of the Best Hour of Contact problem. More specifically, it aims to provide a presentation of the elaborated approaches, their results, and a critical analysis.

Chapter 5 has the same purpose as Chapter 4 but for the problem Identification of the Best Contact Channel.

Chapter 6 is the last chapter and summarizes the conclusions of the analysis documented through other sections. Particularly, it provides the reader with both an opinion about the success of achieving the predefined objectives and which future work can be developed from the results of this study.

Appendices can be found at the end of this document, and they contain complementary information regarding the Exploratory Data Analysis phase.

This page is intentionally left blank.

Chapter 2

Background

This chapter aims to provide the reader with the main concepts needed to understand both the context and methodologies adopted in the development process of this work. First, a definition of data mining will be given, followed by its applicability in a business environment. Then, the concept of CRM will be introduced and we will see how it is related with data mining. Next, a data mining methodology for software development will be described step by step. And finally, we present a literature review about the prediction of customer's behaviour, ending with a discussion on that subject.

2.1 Data Mining for Customer Relationship Management

The concept of data mining is often linked to Customer Relationship Management but before we can discuss their importance together we will first approach them, separately, to understand their individual contributions.

2.1.1 The Data Mining Concept

Data mining, also known as Knowledge Discovery in Databases (KDD), refers to the non-trivial extraction of potential useful information from databases. Although, data mining is frequently treated as KDD, the author of [8] argues that it is actually part of the knowledge discovery process. The enormous amount of data stored in several repositories led to the need to develop powerful means, which enable analysis and, perhaps, the interpretation of such data, could help in the decision-making process.

Data mining has its applicability mainly in business contexts. Its activities can be divided in 3 general categories [9]:

- **Discovery** - the process of looking onto a database to find hidden patterns without a predetermined idea or hypothesis about what the patterns may be.
- **Predictive Modeling** - the process of taking patterns discovered from the database and using them to predict the future.
- **Forensic Analysis** - the process of applying the extracted patterns to find anomalous or unusual data elements.

Accordingly to [10], data mining is used to construct 6 types of models aimed at solving business problems: classification, regression, time series, clustering, association analysis, and sequence discovery. Classification and regression are commonly used to make predictions; association and sequence discovery are methods for discovering relations between variables and to find statistically relevant patterns; clustering can be used for forecasting or description.

These models can be divided into two main types of learning approaches: supervised or unsupervised.

In supervised modeling, the goal is to predict an event or estimate the values of numeric or categorical attributes, making use of labeled datasets. In these models there are two types of fields, input and output. The input is known as predictor and the output is called the target field. Based on this two variables, the model identifies the prediction function. Other possible representation is thinking about this as a function, where the predictor is the X part and the target field is the Y part. The question of learning is reduced to the question of estimating a functional relationship of the form $\mathrm{f}: \mathrm{X} \to \mathrm{Y}$, that is a relationship between input and output [11]. It is supervised because pattern recognition takes into consideration previously labelled fields. The mapping function associates predictors with the output.

In an unsupervised modeling, there are just inputs fields. The learning process does not depend on a defined output field and, therefore, the pattern recognition is not guided by a specific target attribute. The goal is to find similar groups between data patterns (clustering); determine data distribution in the space (density estimation) [12], or to make a dimensionality reduction.

From the 6 models discussed above, classification, regression and time series are supervised, and clustering, association analysis, and sequence discovery are considered as unsupervised methods.

Inside the context of data mining, there are several algorithmic techniques that can be applied to the construction of a model. Commonly, during the modeling phase of a project, the choice of the technique is determined by factors such as the type of data or the project's requirements. In detail, we will approach Decision tree-based algorithms (Random Forest and XGBoost), Linear Regression, Artificial Neural Networks (ANNs) and Clustering (KModes and Agglomerative clustering) because it will help the reader to both understand the section 2.3 and the study presented in this work.

Decision tree-based algorithms are supervised learning models. By taking into account predefined classes (i.e., labels), they use a recursive structure to represent a procedure, on which the final decision is based to classify a future instance. The tree has the upper layer which is designated as the root node and represents the input value. The lower layers are the internal or decision nodes used to test each attribute. The final layer is made up of terminal nodes and determines the classification of the input. From the root, a test defined by a condition is applied, and depending on the result it will go to the next layer number, which may be terminal or not. Through the decision trees it is possible to extract If-Then rules for classification [13, 14]. An example of a decision tree representing a binary decision about whether to play outside or not can be seen in Figure 2.1.

The next two described techniques are based in decision tree models: Random Forests and XGboost (in this case we will consider that the weak predictor is a decision tree). In particular, we would like to enhance the importance of these models since they are used in the present work. They were selected from the set of techniques that are usually used by studies with a similar context for three main reasons: they can be used for any type of data

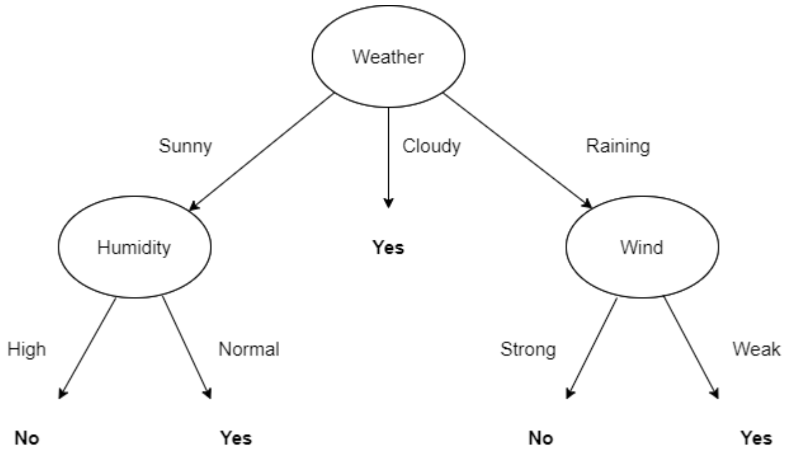


Figure 2.1: An example of a decision tree representing a decision about whether to play outside or not.

(numerical or categorical); They can handle data that are not normally distributed; And they are easy to interpret. In addition, the XGboost has speed benefits that will promote a discussion about speed/performance trade-off.

Random Forest is a collection of decision trees. In short, it builds multiple decision trees and merges them together to get a more accurate and stable predictions. While growing the trees, this algorithm adds additional randomness to the model by searching for the best feature among a random subset of features while is splitting a node. Generally, the diversity added through this technique results in better models. [15]

The XGboost algorithm is a decision-tree-based ensemble Machine Learning algorithm that uses gradient boosting. The gradient boosting has been particularly successful when applied to tree models and its goal is to combine many weak learners to come up with one strong learner. In this case, the weak learners are the individual decision trees. They are all connected in series and each one tries to minimise the error of the previous tree [16]. This approach supports both regression and classification. One problem associated to the gradient boosting is the overfitting since the selection of the number of trees means the amount of weak learners in the model and too many of them may lead to overfitting. To void not only algorithmic but also performance problems, the XGboost was designed to be highly efficient, flexible and portable and provides a parallel tree boosting to solve many data science problems in a fast and accurate way [17]. Unlike other boosting algorithms where weights of misclassified branches are increased, in Gradient Boosted algorithms the loss function is optimised [18].

Linear regression allows for the establishment of a statistical relationship between one or more variables (x) and the defined target (y) [19]. It finds the best equation (line) that represents the input variables in relation to the output variables. To estimate the value y, an equation is used that determines the relationship between the variables x and y. It is used for predicting the dependent variable with the help of independent variables.

Artificial Neural Network (ANN) are inspired by biological neural networks [20]. The principal processing elements of an ANN are called artificial neurons. Several of these units are connected to each other, forming a network, and this architecture can be represented as a directed weighted graph.

Its basic architecture consists of three types of neuron layers: input, hidden, and output

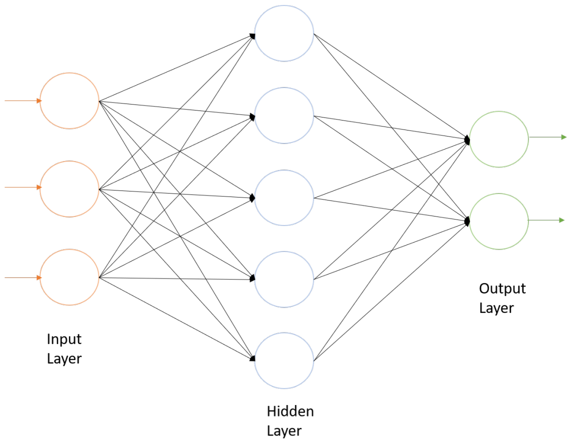


Figure 2.2: An example of a feedforward ANN with a single hidden layer.

layers [21], illustrated in the Figure 2.2. The signal can flow strictly from input to output units and when it does we are in a presence of feed-forward networks. Or, in another scenario, it can contain, also, feedback connections, which turn out to be designated as Recurrent Networks.

The neuron output signal is given by the weighted sum of its inputs, which are computed by a non-linear activation function. And the learning process is based on the adjustment of the weights between the connections of neurons as a way to minimize the loss function [21].

KModes clustering is one of the unsupervised Machine Learning algorithms that is used to cluster categorical variables. It uses the dissimilarities(total mismatches) between the data points. The lesser the dissimilarities the more similar our data points are. The dissimilarity metric used for K-Modes is the Hamming distance [22]. The K-Modes clustering process consists of the following steps:

1. Randomly select k unique objects as the initial cluster centers (modes).
2. Calculate the distances between each object and the cluster mode; assign the object to the cluster whose center has the shortest distance.
3. Repeat until all objects are assigned to clusters.
4. Select a new mode for each cluster and compare it with the previous mode. If different, go back to Step 2; otherwise, stop.[23]

Agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It works in a “bottom-up” manner, that is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root) [24]

2.1.2 The CRM and Data Mining

According to the paper [25], a CRM definition is quite divergent. That work presents an analysis of different perspectives of CRM. Its results show that 50% of definitions were considered as a business strategy, 17% associated CRM to philosophy, 22% were related with technology and the rest of them were a mixed of this three point of views. Therefore, the following definition of CRM is proposed:

“CRM is the building of a customer-oriented culture by which a strategy is created to acquiring, enhancing the profitability of, and retaining customers, that is enabled by an IT application; for achieving mutual benefits for both the organization and the customers” [25].

We consider that this definition includes all different perspectives in order to be a base for agreement for the CRM implementation.

CRM was first developed in the middle of 1990s. One of the most recent works was made by Buttle [26], who stated that there are defensive reasons and offensive reasons behind the adoption of a CRM model by an organization. The offensive reasons are about the desire to improve the profitability by reducing costs. The defensive reasons are related to the fears of a successful implementation of a CRM by competitors, which would lead to the loss of customers for the company. Accordingly, Geib, Kolbe, and Brenner [27] and Liou [28] refer that CRM was a response to the decreasing of customer loyalty in different industries. They are supported by Schierholz, Kolbe, and Brenner [29] who says that it is fundamental an implementation of a CRM to increase the customer´s loyalty.

These ideas are a good starting point for discussing data mining in a CRM framework.

A Literature review of The Current Applications of Machine Learning Techniques in CRM [2] divides the process of building up a customer relationship into four dimensions: **customer identification** (it targets possible customers by identifying the most profitable ones), **customer attraction** (it allocates resources to attract the identified target groups), **customer retention** (it intents to meet the customer expectations) and **customer development** (it focuses on the individual customer profitability). Each dimension was sub-divided into elements (activities) with goals related to marketing, sales and customer relationship. They can be seen in more detail in Table 2.1.

A compilation of data mining techniques used in several works can be found in [2]. The Table 2.1 aims at giving an overview about the context on which techniques are applied.

It is clear that several techniques were applied to each CRM element. Customer identification and customer retention are the two dimensions with the highest number of techniques studied by works. So, we can consider there is a special interest in solving problems related with them, this could be an indicator of what a certain market needs. In general, we can conclude there is both interest and investment in finding different techniques which can be applied to CRM.

From all CRM elements, concepts related with **marketing goals** are directly connected with the work presented in this document. For this reason, we will focus our discussion on that subject.

The book [4] defines 3 different **marketing activities**, customer segmentation, direct marketing campaigns and market basket analysis, where data mining is commonly used.

Customer Segmentation is the division of potential customers in a given market into discrete groups. This division is based on both client’s need and behavior traits such

Dimensions	CRM Elements	Techniques
Customer Identification	Target Customer Analysis	Neural Network; Multiple Logistic Regression; Weighted Random Forest; RUSBoost; Genetic Algorithm; Decision Tree; SVM (Support Vector Machine).
Customer Identification	Customer Segmentation	Self-organizing Maps; Support Vector Clustering; XG-Boost; Logistic Regression; Naïve Bayes; Neural Network; C-MK-SVM (Collaborative Multiple Kernel Support Vector Machine); Random Forest; SVM; PCA (Principal Component Analysis).
Customer Attraction	Direct Marketing	Logistic Regression; Projected Gradient Descendent Algorithm; Fuzzy.
Customer Retention	Loyalty Program	Random Forest; Decision Tree; Logistic Regression; Fuzzy.
Customer Retention	One-to-one Marketing	SVM; Exhaustive Algorithm, Genetic Algorithm, Covering Algorithm; Decision Tree, Logistic Regression; Gradient Boosting; Naïve Bayes; Random Forest; PCA; Neural Network.
Customer Retention	Complaint Management	Logistic Regression; Random Forest, Gradient Boosting; XG-Boost, AdaBoost, Neural Networks, Extra Trees; Case-based Reasoning; SVM; Naïve Bayes; Decision Tree; NBTree (Naïve Bayes Tree); Co- EM Ensemble; Co-EM SVM; Co-training; Co-EM Baysian.
Customer Development	Customer Lifetime Cycle	Random Forest; SVM; Naïve Bayes; MK-SVR.
Customer Development	Up/Cross Selling	MK-STM (Multi-kernel Support Tensor Machine); Neural Network; Fuzzy Algorithm.
Customer Development	Market Basket Analysis	Neural Network; Logit Model; KNearest Neighbor.

Table 2.1: Table to represent the data models techniques per dimension and CRM elements. Adapted from [2]

as buying characteristics, responses to messaging, marketing channels. Data mining can create data-driven behavioral segments. It also can uncover groups with distinct profiles and lead to rich segmentation schemes with business meaning and value.

Direct Marketing Campaigns is a type of advertising campaign that seeks to achieve a specific action in a selected group of clients. The communication can be established by e-mail, SMS, internet, telemarketing or other channel. Two types of campaigns can be identified based on their goals. For one hand, the acquisition campaigns aim at drawing new potentially clients away from the competition. While campaigns to prevent churn and to drive customer to purchase more products are also practiced by companies. Retention campaigns intent to prevent valuable customers from terminating their relationship with the company. Data mining can build models to identify the right customers to contact.

Market Basket and Sequence Analysis is related with the identification of products usually purchased together. Data mining can be used to determine it. In addition, sequences of events or purchases can be identified through sequential models.

These marketing activities can be associated with the 4 dimensions related to the process of building up a customer relationship. Customer Segmentation is one element of Customer Identification; Direct Marketing Campaigns are related to Customer Attraction; and Market Basket belongs to Customer Development. Inside Direct Marketing Campaigns, it became clear that the concept of Customer Retention was mentioned. So, by merging the information contained in these two works we can see that marketing activities are performed in every dimension. Therefore, they are vital to support all CRM processes inside a company.

Marketing strategies are defined by organizations considering its context. And, for that reason, it is important to understand how the knowledge discovery process can be applied and what resources are available to accomplish such a task. According to [9], data mining can be found in four different backgrounds, such as retail, banking, telecommunications and other industries. The telecommunication area is particularly interesting to be further discussed.

Telecommunication companies around the world face escalating competition, forcing them to aggressively market special pricing programs aimed at retaining existing customers and attracting new ones. Knowledge discovery in telecommunications include the following [9]:

- **Call detail record analysis**—Telecommunication companies use call records to identify customer segments in order to develop attractive pricing and feature promotions.
- **Customer loyalty**— Usually customers “churn”, to take advantage of attractive incentives by competing companies. The companies can use data mining to identify the characteristics of customers who are likely to remain loyal once they switch to retain the most profit ones.

2.2 CRISP-DM, a Data Mining Methodology

The success of a project depends on the steps followed during its implementation. Projects involving real world data and business usually raise some difficulties because there is no clear methodology to support all required activities. The book [4] defines the basic phases to develop a data mining project according to the Cross Industry Standard Process for Data Mining (CRISP-DM) process model and the Figure 2.3 is its graphical representation. The aim of this methodology is to standardize data mining processes across industries and it has since become the most common methodology for data mining, analytics, and data science projects. The main steps can be described as follows:

1. **Business understanding:** This phase involves the problem definition and objectives for the project. A clear definition of the problem will add meaning to the purpose of the analysis. By the end, it is expected as output a set of narrow questions. The success criteria should be established as well.
2. **Data understanding:** It includes all activities since determining the required data, up to verifying its quality. It includes data extraction and data exploration steps. The main goal is the identification of potential problems regarding availability and quality of the data.

3. **Data Preparation:** The raw data must be prepared for the modeling phase. In this phase processes of data cleaning, the generation of new attributes required for analysis and data integration (merge and aggregate) might be necessary. The consolidated data should be properly transformed according to the requirements of the models to be applied.

4. **Modeling:** This phase starts by selecting modeling techniques, that is the same as answering the question "Which models should we try?". The experimental design will define if there is the need to split data into training, test and validation sets.

5. **Evaluation:** The results need to be interpreted based on domain knowledge and compared to the pre-defined success criteria. Generally, multiple models are tested against each other and the chosen model is approved and prepared for deployment.

6. **Deployment:** The complexity of this phase depends on the complexity of the requirements. A model is not particularly useful unless the customer can access its results, which must be compiled and summarized in a report. In addition, not only a plan should be developed for deploying the model but also, a monitoring and maintenance plan is required to avoid issues during the post project phase. Finally, a review of the project can be conducted to know what could be improved in the future.

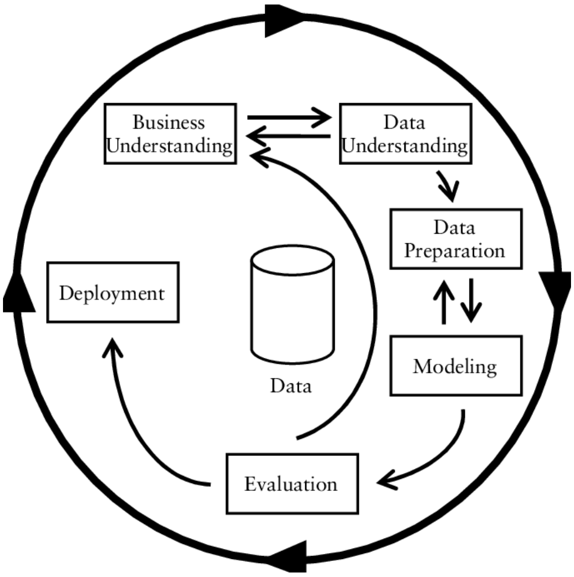


Figure 2.3: CRISP-DM Process flow Diagram [1]

2.3 Predicting Customers Behaviour

The development of a framework dedicated to analyse data from a CRM system is important to help defining business strategies and to identify valuable customers. Companies want to predict the future of the relationship between both parties. Understand client's behaviors will contribute to the management of measurements such as loyalty to keep a profitable business.

From the point of view of Femina Bahari e Sudheep Elayidom [5], data mining can be used to create learning models based on the classification of the client's behavior to assist in the decision making process. They suggest the building of a framework capable of working with data from a CRM of a Portuguese bank. The dataset is composed by 17 marketing campaigns performed between May 2008 and November 2010, only made through phone contacts. There are 2 possible outcomes, either the customer subscribes the proposed campaign or rejects it. 16 input variables were considered: 8 features related to the customer; 4 dedicated to characterize the last contact; and 4 variables regarding the campaign itself. This work had as primary goal the comparison of a *Multilayer Perception Neural Network (MLPNN)* efficiency against a *Naive Bayes* model. The results of both classifiers were shown through the analysis of 3 statistical measurements: accuracy, sensitivity and specificity. It was concluded that the MLPNN model's performance was superior than the *Naive Bayes* model with an accuracy of 88,63%.

Usually, marketing campaigns are performed through phone calls. According to [3] the increasingly vast number of marketing campaigns over time has reduced its effect on the general public. This study intends to highlight the importance of marketing in banks and thus the importance of phone calls. Making use of machine learning algorithms, its goal is to predict if a customer subscribes the deposit or not. Both dataset and targets chosen for these work are the same as in [5]. Another common point between the two works are the features used to develop the project. Regarding the predictive model, it was selected a two-layer feed-forward network with 15 inputs and 20 sigmoid hidden neurons and linear output neurons. Training was done using scaled conjugate gradient back propagation network developed by Moller [30]. The results shown that the model succeeded to classify 89% of the simulation samples. Both studies present promising performance values and they validate each other.

Still within the bank marketing area, there is a research made by Sérgio Moro, Paulo Cortez, Paulo Rita [31] that is worth to mention. According to them, there is the necessity of a client's segmentation when direct marketing is the chosen approach. In addition, they believe that the task of selecting the best set of clients, i.e., the ones that are more likely to subscribe a product, is considered NP-hard [32]. In that study, an intelligent DSS is proposed in order to predict the result of a call for the sale of deposits in the long term, using a DM approach. This tool aims to help prioritizing and selecting the customers to be contacted, with the direct consequence of reducing the time and costs of each campaign. The effective management of the number of calls leads to a decrease in customer stress, making the telemarketing process less intrusive for the customer. For the development of the project, the dataset used was from a bank, whose data were collected between 2008 and 2013. The dataset was unbalanced since there were only 12.38% of successful cases. 150 features were analyzed but only 22 were considered relevant, within which attributes related to telemarketing, product details and customer information can be observed. These features were further enriched with social and economic data. The models used in the study were linear regression, decision trees, neural networks and support vector machines. These models were compared using two metrics, Area under the ROC Curve (AUC) and Alift. The AUC metric represents the measure of the ability of a classifier to distinguish between classes and it is computed from the Receiver Operating Characteristic (ROC) curve that plots the true positive rate as a function of the false positive rate [33]. The Lift metric compares the models' performance with the average of the population [34] and the Alift is the area of the Lift cumulative curve. The Neural Network (NN) was the model that obtained the best results in both metrics, AUC of 0.80 and an Alift of 0.67. It was also concluded that a percentage of 79% of successful sales can be achieved only by contacting half of the customers instead of contacting all customers. Regarding the features used, the study was able to determine which ones had a greater influence on the results: Euribor rate, direction of the call (inbound, outbound), the experience of the agent and the duration of previous calls, which needed to be rescheduled to obtain a final customer response.

2.4 Discussion

The literature is rather scarce in the area of our study. Although the aforementioned studies do not belong to the telecommunications area, they refer to organizations providing Customer Relationship Management (CRM) services. It is true that the type of service can influence the customers' behaviour, but the mentioned works are promising. They aim at offering a tool capable of extracting useful information from marketing calls data and, therefore, they offer relevant insights for this study.

In addition, we notice limitations regarding the works that focused on the problematic we were dealing with. Questions such as *When* and *How to contact a client* are not addressed when it comes to make an analysis about customer retention and satisfaction. Companies tend to make use of traditional methods such as concentrating contacts at 10am or 16pm, avoiding Mondays, believing that Thursdays and Wednesdays are better days to contact and even and never calling at lunch times (12pm - 14pm) [35, 36]. Some of these organizations have means to prioritize contacts by filtering them regarding a customer's segmentation and the type of marketing campaign they are advertising. But, usually, the method does not take into consideration the the client's availability, which can be identified through an historical analysis of the customer's interactions records.

In addition, the choice of the means of contact commonly follows simple methods such as, if the client has email, they will be contacted through email, if not, they can be added either to the list of phone calls or be sent an SMS. This approach contributes neither for an efficient resources management nor for the success of the campaigns.

The combination of data mining with a CRM model can make interactions between companies and clients more personalized. The goal is to view the customer as a single individual who has common traits with his segment group, but also, to identify his unique needs to specialize the treatment. Both availability and means of contact are important components in the referred approach.

With the increase of both the means of communication and the amount of marketing campaigns, we believe that hitting on the right time through the right channel is crucial to build an efficient client's management system.

Chapter 3

Methodology

Both problems presented in this study followed a development methodology similar to CRISP-DM. In this chapter we will start by describing our projects' stages and how they were developed considering each problem's background. Then, to summarize the overall view of the integration of each phase into the project's development process we will finish with a representation of the dynamic between each step.

3.1 Business Understanding

The business understanding stage involves a clear presentation and formulation of the problem to be solved and the definition of our success criteria. So, we approached this phase by discussing the following set of questions:

- What is the definition of a successful contact? What will be our models' target?
- Are the metrics that define a successful contact the same for all communication channels?
- How can we represent our output?
- What features can be relevant for this study?
- What could our success criteria be?

By answering these questions, we gathered valuable information not only to define our problem but also to give us an idea about the challenge's dimension. Both the identification of the best contact hour and the identification of the best contact channel contain a segment dedicated to the problem formulation.

3.2 Data Understanding

The source data was provided by the organization we are working with, and it is described in this section. It is composed of 16 csv files containing associated information between each other. These records were collected from August 2020 to November 2020 for the region of Coimbra.

One of the files contains information about the client's company. There are 33547 unique clients with 8 features about account details such as Client identification number (NIC), activation account date, deactivation account date, telephone's number, and code's market, which indicates to which market the customer belongs (particular or PME).

In addition, we have access to information about card's clients. There are 43904 records with 8 features about the card by itself, like card's code and telephone's number, but also, it contains information about the client to which the card is linked. In this file some personal features were provided as well: NIC, career's code, number of children, age and sex. At a later date, more data was received. It contained clients' records pertaining to the amount of accepted campaigns in the year of 2020 and other relevant information such as the customer's device, their tariff, if they ever accessed the client's area and if they subscribed to the direct debit.

Interaction between the CRM and the customer is available as well. This file contains 108088 interactions described by 16 attributes. In general, the information stored in it is concerned with the date when the interaction was initiated, the date when the interaction was finished, the channel used, the direction of the interaction (inbound or outbound) and which client the interaction is related to. A solicitation identification number is also needed to link each one of these records to the correspondent solicitation.

The relation between interactions and solicitations is not trivial to understand, but it is a central component of this study. Each interaction is expected to generate a solicitation. However, some solicitations can initiate interactions too. For instance, a client report an issue which opens a solicitation. Then, the client is contacted by the company to solve the problem. This interaction was initiated due to an opened solicitation. So, one can understand that the relationship between these two entities is many to many. There is one special case to take into account: it is not mandatory for an interaction to generate a solicitation and vice versa.

The file containing information about solicitations has 45355 records described by 20 attributes. Some of them are related to the context of the solicitation: its start date and finish date; its direction (inbound or outbound); the solicitation's duration; channel and sub channel used. Other fields are included to make the relation with the client table. Finally, features such as process and typification give us information about the business context of this solicitation. Its nature can be classified as a breakdown, a commercial solicitation or other type of issue.

The Call Detail Records (CDRs) contain information about call/SMS traffic and customer's data usage (or General Packet Radio Service (GPRS)). In total there are 6 files: 2 of them related to calls traffic; 2 for SMS traffic; and the other 2 for GPRS. Each pair of files represent two types of clients, pre-paid users and post-paid users (this distinction is based on mobile contracts). The number of records per file is between 335014 and 4701968, composed of 10 fields (approximated). Regarding the call/SMS traffic files, there is information about the date of communication, the number which made the communication, the number receiver and the communication's duration (not applicable for SMS). On the other hand, files related to the user's data usage contain, once more, information about the record's date, but also, it provides the session's duration and the amount of kb used by the user. All records have information to make the connection with the client through the card's code and market code. The field month old is given in these files and it is associated with the card. It represents the age of the card.

The rest of the dataset is composed of auxiliary files to add information about some variables that were coded by an integer, such as the fields career and client's market.

The data source needs to contain information to help us achieve our goal, especially when related to the number of clients and interaction we have at our disposal. So, the data understanding phase raised a set of questions regarding data availability:

- How many clients had interacted with CRM?
- Will we have access to enough interactions to support our model? And how many of them are inbound? And outbound?
- How is the distribution of interactions per contact channel?
- How many contacts do we have with a contact duration higher than 0 seconds? Or 5 minutes?
- Does the dataset has outliers considering different variables?

In the section *Exploratory Data Analysis* presented in both studies, the reader can find not only the answer to these questions, but also, the method used to analyze the data quality.

3.3 Extract, Transform and Load

When dealing with vast amounts of data, structure is a key aspect of the process. We received the data in csv files, and each file had information about entities which complement each other. As such, we decided to create a Relational Data Base to organize all information contained in the files in such a way that relationships between the data were perceptible. Our main concern was about maintaining the correct connection between the entities to avoid changing the meaning of the information during the process.

The final model is not presented to the reader due to confidentiality reasons. Although all data has been processed and organized, just a part was considered relevant for our purpose. This design was made in collaboration with another research that uses the same source of data for a different analysis.

The database was implemented in PostgresSQL. The architecture suggested was planned to support all queries about both clients' interactions and solicitations. The main goal of this study is strongly connected with the information contained in these two entities. Features such as the starting and ending date of an interaction, its duration, the channel used for contact and the type of contact (inbound or outbound) are valuable to analyze temporal patterns and to detect a possible recurrent behavior regarding the questions raised in this chapter. The rest of entities contain information to support additions to primary queries to give the opportunity to make more complex analysis which is required to answer the secondary set of questions.

The source dataset had to be rendered anonymous. All fields containing confidential information were transformed into a hash. The first task was to change all attributes coded with hash to integers and create auxiliary tables in database to keep records of the new correspondences. This action not only will allow us to simplify all queries that involve these fields, but also, it will help to reduce the complexity of data visualization.

Then, it was necessary to change the fields sex, age, number of children and career to the table client. We believe it makes more sense if a customer possesses such attributes. In the process it was discovered some customers that did not have correspondence with the clients' table. In other words, there was data in use on table cards that did not match with any known client. As there were so few, we decided to add them to the client's table with all attributes empty. Our goal was to create a coherent data representation. A similar process was reproduced to the card's table. The field month old had to be added and the previously mentioned features were moved to the client's table.

Depending on the data codification used to write the files and the type of reading we were making, some values had an erroneous type. One example was the field age from the file client, which was read as a float. So, another script was created to change the type of the variables in a column and used when needed.

The main challenges we faced during this stage were concerned with data redundancy and missing data. We decided to deal with null values during the exploratory data analysis phase. We could not discard the entire entry because it could contain valuable information related to other tables, neither change its value as it would corrupt the data integrity.

To conclude this process, the last step was to load our database with the resulting data from the pre-processing stage.

3.4 Modeling

Although the CRISP-DM methodology has been mentioned in the *Context* chapter, it was described in a general way. In other words, each project has its own details, and its workflow does not always fit in standard models.

Up to this point, we explained the process of data transformation from its raw stage to logical data as it is a key step to prepare data to support the two problems presented in this study.

In this section we want not only to guide the reader through our development process but also to specify how we adapted the CRISP-DM methodology to our project.

Figure 3.1 represents all project's stages. We would like to draw attention to the presence of a data transformation phase in the scheme. Its addition will be justified bellow as it is not included in the CRISP-DM methodology description.

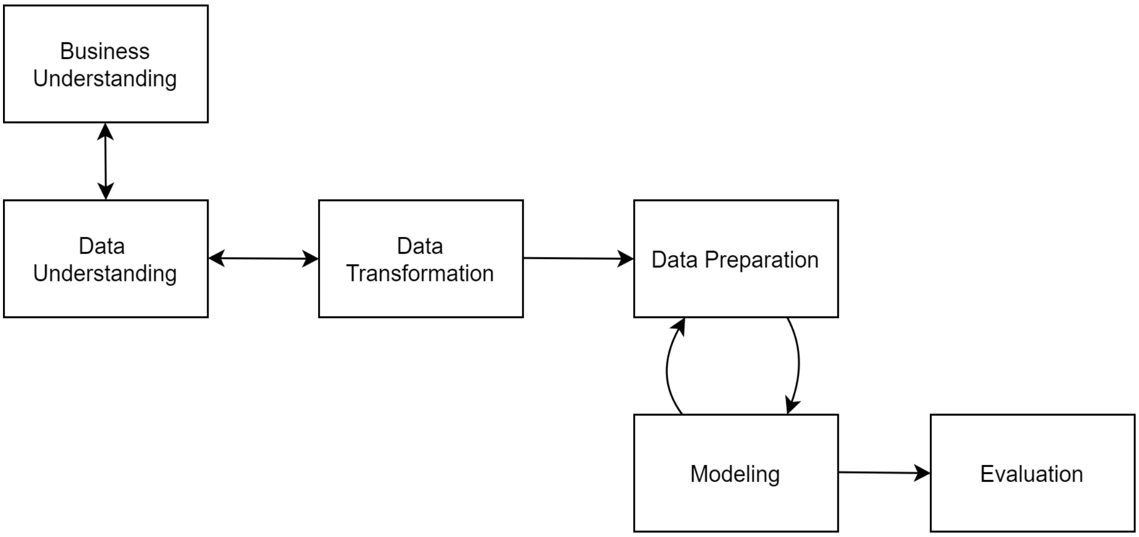


Figure 3.1: Project's Stages

Both business and part of the data understanding phases were developed simultaneously. While we were defining our problem, we had to start by knowing if the available data could give an answer to our challenges. This process was done with the help of a group of representatives of the organization as they contained the business knowledge that we were missing of.

The data transformation stage was added to the previous ones. As we obtained more insights about the core business and started to better understand the relations between the available data, we defined our data model to represent the information in a logical way. This step was composed by the extraction of data from the csv files, followed by its transformation and loading, as it is described in section 3.3. When the process is concluded, the data understanding phase can answer the set of questions raised in section 3.2 and the business understanding can formally define the problem.

In the data preparation stage, we used information stored in our database to construct a data source capable of supporting all operations related to the cleaning, filtering, selecting, and labelling of our samples.

Our approaches were defined in the modeling phase. We used data mining knowledge to develop predictive models to meet our goals. The experimental scenarios were delineated in this stage, as well.

The last phase is the models' evaluation. Our development process ends with the interpretation of the results considering our success criteria. In addition, we compare different approaches according to the most suitable analysis tools.

From all described stages, only the activities of data understanding considering the business context and data transformation were common phases to solve our two problems. For that reason, both were explained in the present chapter. The other stages are specified in the respective chapter of each problem.

3.5 Planning

As to achieve our objectives, this work was structured by tasks. We had two different sprints: one that was related with our weekly internal meetings where we both discussed the results obtained during the week and planned the following tasks; and another that was a meeting with the organization that happened fortnightly. In these reunions we presented our results to the company, accessed our risks, and defined our next steps considering the feedback obtained. In addition, we had hard deadlines to met on the 6th (M6) and 9th (M9) months of the project. In each one of them we had to provide both the code and a report to the organization and, we also had to present the work done in report presentations. In order to help the company understand the code and how to integrate it into a testing environment, we met with them in what we called code visiting meetings.

The Figure 5 of the Appendix B represents a high-level Gantt's diagram with the initial project's work plan. All tasks were planned based on the objectives specified for the hard deadlines at the 6th and 9th months. The main tasks identified in the diagram follow the approach described in section 3.4, but we want to draw attention to the time allocation of the task *Dissertation*, that was planned accordingly to the project's investigation component that culminates with the writing of all discoveries into a dissertation.

Regarding the Figure 6 of the Appendix B, it shows the real time allocation of each task during the project. As we were expecting, more tasks were performed than the ones initially planned. Some difficulties considering the approach's definition of the identification of the best hour to contact a customer led us to develop four different models. In addition, it was proposed to us the writing of an article for the company's magazine, which resulted into the inclusion of the task *Article writing.* In order to deal with this unplanned work, we increased our effort per sprint and we adjusted the tasks related to the channel problem to start at a later time. Finally, the *validation stage* was suppressed in an attempt to minimize this delay, and due to difficulties related with the integration of the code into the organization's workflow.

Overall, all tasks were completed with the exception of *Validation*, although more work than the one initially defined was done. None of the hard deadlines were rescheduled meaning that we managed to met the goals for the M6 and M9 project stages.

This page is intentionally left blank.

Chapter 4

Identification of the best contact hour

This chapter is dedicated to the description of the components included in the problem of identifying of the best contact hour. Its goal is to formally define the problem, discuss the findings of the exploratory data analysis, describe the process of feature engineering, clarify the proposed approaches and discuss the results of the evaluation phase.

4.1 Problem definition

Customer segmentation can be made based on his availability for being contacted, by dividing the set of clients into groups with similar behavioral patterns. Discovering the best hour to contact a customer can be seen as a classification problem composed of fifteen categories (each category corresponding to an hour between 8am and 10 pm). In particular, the goal is to rank those hours per client. This division is particularly relevant for voice communications as both actors need to be available and engaged in the action. As such, in the study described in this chapter we will only consider voice interactions between the customer and the organization. In addition, some priori knowledge was used to support the problem and the following assumptions were taken into account:

- (A1) Past communications between CRM and the client are useful to group clients based on the interaction hour.
- (A2) Clients usually have a set of hours of availability.
- (A3) Clients with similar availability share common characteristics.
- (A4) Some contact hours can be popular between clients with similar characteristics.

4.2 Exploratory Data Analysis

The main goal of this analysis is to detect the presence of outliers and make a temporal distribution of the number of interactions between the CRM and the customer per week, to have better understanding of the data we are working with. The method used to inspect the data was mainly based on graphical representations. The set of questions we purposed

Channel	Amount of interactions made with CRM
Voice	5703
SMS	13847
Email	1148
Writting	1825
In person	3376
Web	3050
IVR	692

Table 4.1: The table represents the amount of interactions made with CRM per client.

to answer are the ones mentioned in chapter 3, section 3.2 regarding data information to support our goal.

Table 4.1 and Figure 4.1 display two types of data representation. Their goal is to give information regarding the following questions:

- How many clients have interacted with CRM?
- How is the distribution of communications per contact channel?
- Will we have access to enough communications to support our model? And how many of them are inbound? And outbound?
- How many contacts do we have with a contact duration higher than 0 seconds? Or 5 minutes?

From Table 4.1 we can observe that the channel SMS is the most selected to make interactions between the customer and the CRM, independently of the contact's direction (inbound or outbound) or its duration (further analysis show that SMS is mostly used for outbound interactions). Voice channel comes right next and *In person* and WEB channels have a similiar usage. Overall, we benefit from a large number of clients to support the predictive model to be developed.

The histograms in Figure 4.1 show the distribution of the number of interactions per number of clients. This representation also includes a distinction related to the communications' direction (inbound, outbound and without direction). As a filtering parameter it is used the contact's duration (seconds). The values of the variable number of clients, y, are displayed as log(y) to make it more readable. Each graph represents a combination of histograms regarding to the contact's direction. Different histograms are added to the existent one, for instance, first it is drawn the Inbound histogram, then on top of that, the values related to outbound variable are added and at last, the values labeled as *without direction* are attached.

In order to collect information about the interactions' duration, its minimum value was defined as 0 (0min), 180 (3min) and 300 (5min), Figure 4.1. We are interested in analyze the amount of clients that are available to be engaged in a contact considering different duration thresholds. The only channel which supports this approach is Voice type. The others have neither any record considering this variable nor enough samples to make a meaningful analysis.

We can observe in the histogram of Figure 4.1 that when the number of interactions increases, the number of clients decreases, meaning that each client usually has not the need to make many contacts with the CRM (few samples per client). For the voice channel, we also conclude that the number of inbound interactions is considerably higher than the number of outbound ones, which can indicate that the clients frequently initiate the contact by voice. In addition, we can see that there are not significant differences between all three histograms, which means we have access to many clients with the availability to make calls whose duration is at least 5 minutes.

Regarding the outliers analysis, Figures 4.2, 4.3, 4.4 show us the values for interaction's duration diverge the ones expected. We detected a high number of outliers that must be discarded when making the future preprocessing phase to define the training dataset, as they represent unusual values in our dataset and, in this case, they distort the reality by adding information that was originated due to a technical error (for instance, calls duration of 4M of seconds). On the other hand, the median has acceptable values, especially for the voice channel, so this is a favorable result for the use of the variable interactions duration as an indicator for the voice interaction success. Its duration value can be defined based on the median value, and, in order to do that, we propose analyzing calls duration depending on the reason why communication was established.

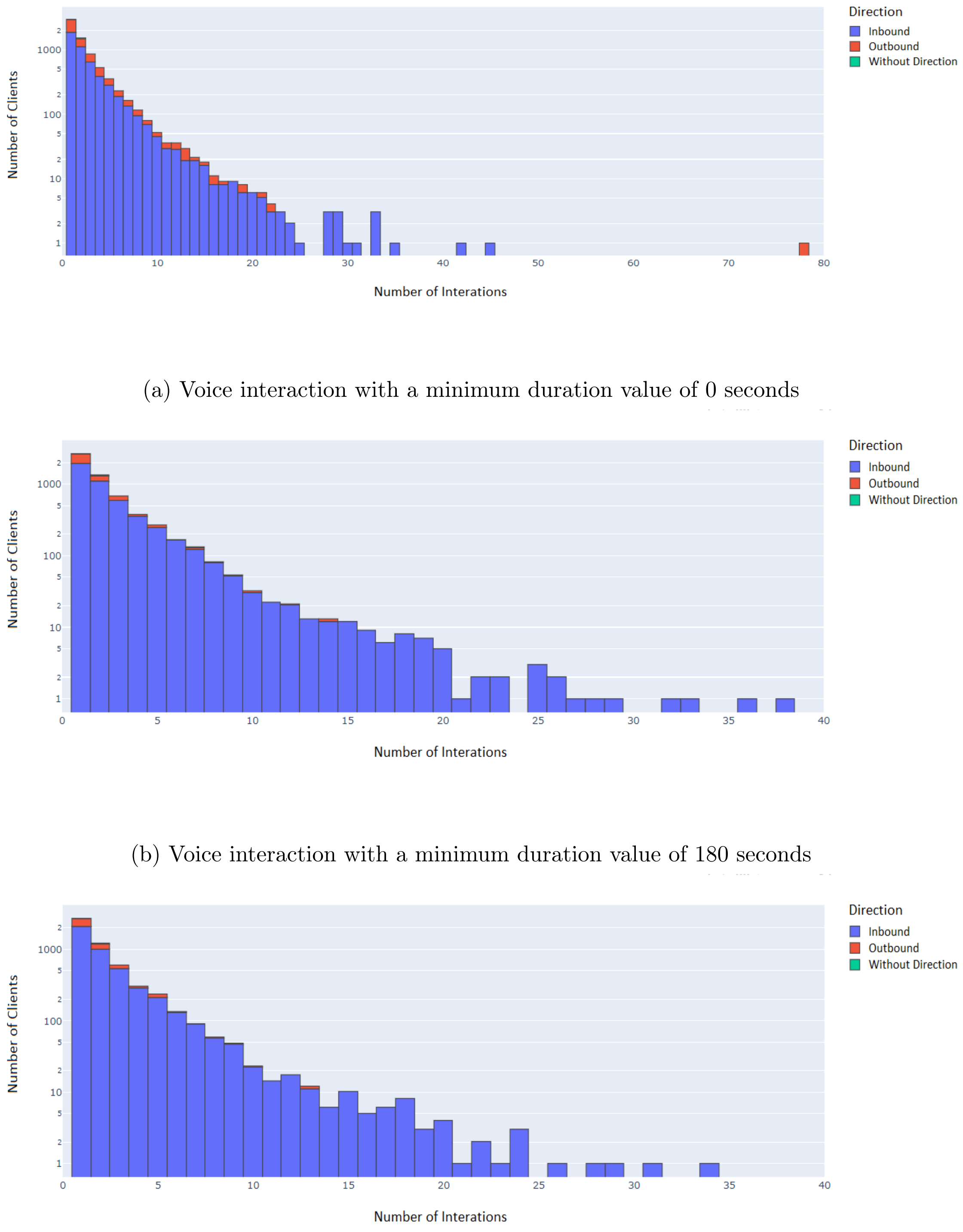
The motives behind each contact can be deducted by taking into consideration one more dimension: the sub-channel. In addition, we assumed that *Agentes*, *Técnico*, *Outros* e *Suporte ao Técnico* sub-channels do not represent the client's availability and, therefore, we will not take them into account to determine an acceptable value for calls duration. From *Comercial*, *Customer care* and *Retenção* sub-channels, we can observe that the median is between 473 and 897 seconds (about 15 minutes). These values offer a good start point to define limits for the patterns that should be considered for our training dataset. So, from this analysis we can conclude that special attention between different sub channels will not be required when designing our model. Particularly, we can discard the idea of making a personalized model per sub channel.

Finally, in order to analyze possible temporal patterns hidden inside the data, density heatmaps were created. Each graph represents the distribution of the number of communications per hour of the day, shown through a temporal period corresponding to a week, starting on Monday, and finishing on Sunday. The filters direction of contact and communications' duration were added to make a more complete data exploration.

The temporal analysis of interactions can be seen through figures 1, 2, 3 and 4 from Appendix A. We are interested in not only observe differences between inbound and outbound interactions but also, the differences when the communications duration values are set from 180 seconds to 3000 seconds. The reason for the upper threshold restriction is to make a simple elimination of some detected outliers, as mentioned in the conclusions of the box-plot representation.

Figures 1 and 2 from Appendix A show us all interactions made when the minimum value for its duration is 0 seconds and the maximum is 3000 seconds. For the inbound case, we can see that there is a high concentration of contacts at 11am and in the period between 2pm to 5pm. Saturdays and Sundays have the lowest number of communications. Regarding the outbound representation, it can be observed that there is an elevated number of interactions at 12am and 3-5pm. Once more, Saturdays and Sundays are avoided to make interactions with the client. It also seems that Mondays are not a common choice.

Figures 3 and 4 from Appendix A represent inbound and outbound communications with a minimum duration of 180 seconds and a maximum duration of 3000 seconds. We pretend to discover the weekly interaction distribution for communications that lasts at least 3 minutes, as we believe that this value is an acceptable threshold to indicate the client's



(a) Voice interaction with a minimum duration value of 0 seconds

(b) Voice interaction with a minimum duration value of 180 seconds

(c) Voice interaction with a minimum duration value of 300 seconds

Figure 4.1: Graphical representation of the distribution of customers that had interactions with CRM through Voice and a duration of 0, 180 and 300 seconds

availability to be engaged in the contact. Our intent was to filter communications where the customer declines right after the contact is established. Regarding the inbound graphic, it continues to be notorious that some peaks are concentrated at 12am and between 2pm and

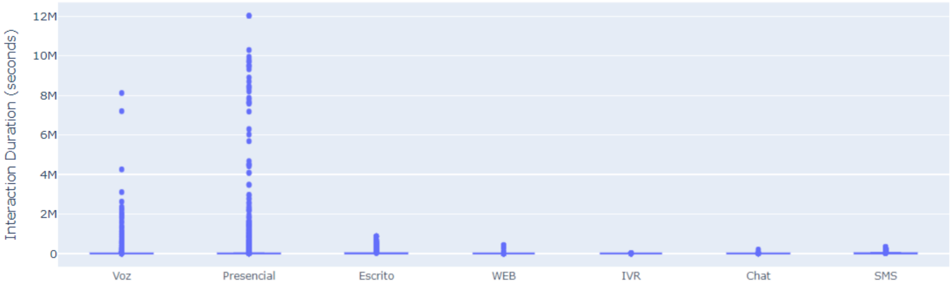


Figure 4.2: Box-plot Channel per interactions' duration

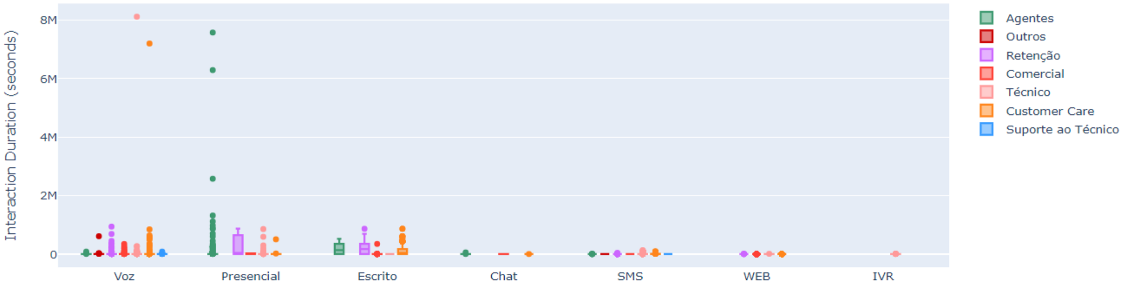


Figure 4.3: Box-plot Channel and sub-channel per interactions' duration

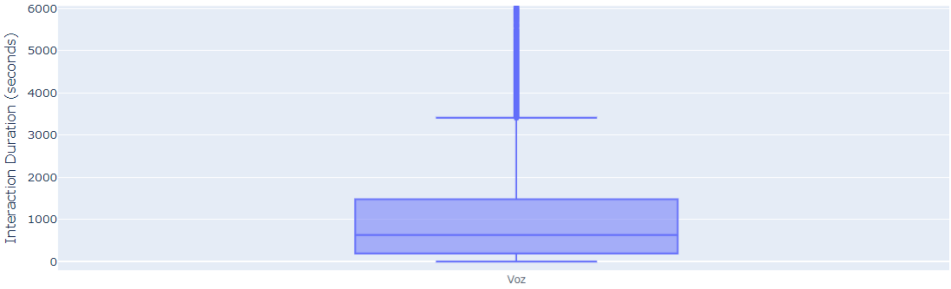


Figure 4.4: Box-plot Channel Voice per interactions' duration

5pm. But the distribution of interactions is more spread from 9am to 10pm. The days with a lower number of interactions are the weekend days. On the other hand, the outbound interactions analysis does not show big differences regarding the representation where the minimum duration was set to 0 seconds. The only point that is worth to mentioning is that the highest number of interactions decreased by 10-20%, which can indicate that clients tend to remain in the communication at least 3 minutes.

4.3 Feature Selection and Engineering

The construction of the dataset that will support our model is a key step. Our goal was to transform and select data that can represent successful voice interactions made between the CRM and the client.

To achieve that, all communications registries from the table interactions were filtered based on calls duration, accepting only the ones with 180s to 3000s. Each entry of the dataset contains information about the interaction itself, such as the contact hour. Client features were added as well to allow the data to have the desired characterization. Although the original dataset contains more features about the client, not all of them had a good data volume to be considered for this project.

To improve the dataset and to have a better characterisation of the clients we engineered additional features, based on domain knowledge provide by the organization. The goal was to include more information about the client that could indicate availability per hour. Interactions with CRM are not the only source of data of this project. We also have access to call records through the CDRs. Out of them, we generated the following features: total of calls in the morning, total of calls in the afternoon, total of calls in the evening, average calls per day, average calls per week and total calls per hour from 8am to 10pm. They represent the calls behavior of a client, and we believe it can indicate a propensity for receiving/making phone calls at certain hours.

In the end, our dataset aggregates 20 fields from clients calls behavior, 3 from its own characteristics and 1 from each interaction between the customer with the CRM. A summary can be seen in Table 4.2. The new data is loaded into the database according to the data model representation.

Feature	Data Type	Origin
Month old	Numerical	Client
Sex	Categorical	Client
Interaction Hour	Numerical	Interaction
Total of calls in the morning	Numerical	Calls behavior
Total of calls in the afternoon	Numerical	Calls behavior
Total of calls in the evening	Numerical	Calls behavior
Average calls per day	Numerical	Calls behavior
Average calls per week	Numerical	Calls behavior
Total calls at 8am to 10pm	Numerical	Calls behavior

Table 4.2: Datasets features summary

Besides the feature selection based mostly on data volume (for instance, more than 95% of the values of the feature *number of children* were null values, which do not add information to the classification), a graphical correlation visualization was made between each pair of features. This data exploration did not show any possible redundancy on the selected data, which could indicate that each field contributes with unique information for the model. On top of the described data exploration, some data transformations were made as well. Regarding to the sex field, it was coded into integer values that represents the two categories Male and Female.

Finally, numeric features were transformed into categorical ones by dividing each field into intervals and then, assigning them a category. Month old was transformed into years and each year represents a category. For example, a client with 3 month belongs to the interval [0, 12[month old, thus being represented by the category 1. The next described step was applied to all features related to calls behavior. Each field was divided by quantile percentage into 5 categories (1-far below average, 2-below average, 3-average, 4-above average, 5-far above the average). The chosen division criterion helps to isolate outliers into their own category. In this case, discarding the outliers could lead to a misrepresentation of the reality, especially when it comes to each clients calls behavior.

This last step was a preparation for the grouping phase and a detailed description this transformation will be added in the next section. Conceptually, making groups based on values intervals helps to interpret reality as the categories above mentioned. There is no difference between a costumer that makes 1 call per week and another that makes 2 calls per week, but it is more relevant to know if the number of calls is on average or above average.

4.4 Proposed Approaches

The client's availability to be contacted depends not only on the time of the day but also on the day of the week. The temporal data analysis described in the section *Exploratory Data Analysis* shows that some weekdays and some periods of the day have a higher concentration of phone calls. So, taking this into account, two types of approaches were defined: the general ones, where the model does not consider a distinction based on weekdays, and the models that are weekday based.

4.4.1 Non weekday-based approaches

One way to address the problem is to assume that clients with similar availability share common characteristics and, as such, customers can be divided into groups depending on the hour they interact with the company. So, it is useful to make a categorical representation of generic clients that can be contacted at each hour. For this purpose, we applied a cluster algorithm to find the center of each clients' group to obtain a generalized features vector (i. e., prototype) that characterizes the cluster. This process uses historical information to make the generalization (A1).

Initially, the dataset was divided by contact hour to group all interactions by a time indicator, which results in 15 sub-datasets containing information about all clients that interacted with CRM at each hour.

These groups represent clusters, and their unity can be seen as a point (a vector of features that characterize the client that made an interaction) on a 17 dimensions representation. The next step is to find the center of each cluster to make a generalization of the costumer characteristics that typically make interactions with CRM at the correspondent hour of the cluster. For this purpose, for each sub-dataset, we used a Kmodes algorithm to get the center of the group. Furthermore, this algorithm was selected because it can be adapted to work with categorical data.

A 3D visualization was made to help follow the process. All 17 dimensions were reduced to 3 with the help of a PCA algorithm. For example, the cluster corresponding to 12pm (light blue) and its center (dark blue) is represented in Figure 4.5. Each axis is a combination of dimensions in order to make a dimensional reduction to obtain the 3D representation.

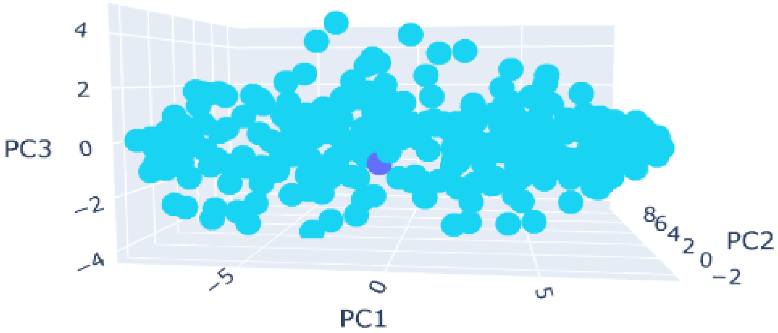


Figure 4.5: A 3D visualization of the cluster corresponding to 12am and its center.

This step is not crucial for the construction of the model, but this type of visualization helps to verify the result. From Figure 4.5 we can consider that the implemented model is correctly identifying the center of the cluster.

After finding the centers of each cluster and making their 3D representation, Figure 4.6, it is notorious that some centers are very close, almost overlapped (points marked with a red circle). This result could indicate that certain hours share similar features.

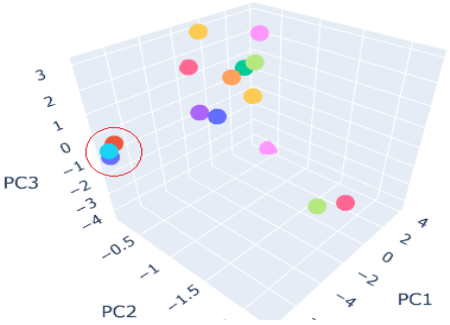


Figure 4.6: A 3D visualization of the centers of the 15 clusters

So, a clustering algorithm was applied to the centers' dataset to discover which hours can be grouped. The algorithm returns the label of each center given as input. This label is a numeric value that indicates to which group the center was categorized. Considering that we want to group the 15 centers into 8 clusters, the result of this operation is the transformation of the initial 15 hours into 8 available patterns to represent the same 15 hours. For the purpose of this study, we selected two types of clustering algorithms to analyze and compare their results. One of them was the Kmodes algorithm with the number of clusters equal to a pre-determined value between 8 and 15 (this value was one of the experimental variables). The other was a hierarchical agglomerative clustering algorithm with the number of clusters equals to a value between 8 and 15, affinity equals to cosine and, as linkage, we chose both complete and single to make our experimental case scenarios. They can be read in more detail in the next section *Evaluation*.

The described process uses historical information to make the generalization (assumption A1). At this point, we introduced into our model the assumption that clients usually have a set of hours of availability (A2). Although some customers had past contacts at certain hours, we intend to rearrange the clusters in such a way that clients are placed in a cluster where they have more similarity among them.

Using algorithms for distance calculation between points represented by categorical features, we calculated the distance between each client features and the centers of each cluster. The goal is to find the cluster with the minimum distance to the customer and replace this one into that group. The minimum distance represents the minimum of differences between the 2 points, in other words, we want to minimize the difference and maximize the similarity. The two algorithms used to accomplish such task were Gower and Jaccard algorithms. The Gower distance is a metric that measures the dissimilarity of two items with mixed numeric and non-numeric data [37]. It computes the distance as the the average of all elements. If the element is numeric, its absolute value is the difference between the two values divided by the range. For non-numeric elements it assigns the value 1 if the element is different and 0 if the element is the same. In turn, the Jaccard Similarity measures similarities between sets. It is calculated by dividing the number of equal observations in both sets by the number of observations in either set [38].

Finally, an interception calculation algorithm is applied to all clients that belong to each cluster. If all clients share the same category for the feature X, then this category is used to categorize the cluster. If any client breaks the previous condition, then the feature X will have no value and it will not be considered as a characteristic of the cluster. The result of this operation is a matrix with the common features between all the points of a cluster. This step is also responsible for the transformation of the numerical features into categorical. It will not be possible to make an interception to characterize each hour if we had a huge set of values per feature. So, an interval division was the solution adopted.

In this phase the goal is to return a vector per client that represents the ranking of the top 5 best hours to make voice contact.

Using a supervised learning algorithm, we can train a model with the resulting matrix from the clustering phase and test it with a dataset that contains the clients and their features. The algorithm should learn to identify each hour based on their characteristics (training phase) and then, when confronted with new clients, it must label the customer with the most suitable hour.

We use the Random Forest algorithm to accomplish such a task. In the training phase, it receives as input the matrix of features of each hour and their correspondent hours as the label parameter. For the testing phase, it is necessary to give it the clients features, and it will return just one hour (label).

In order to make the top 5 ranking hours per client, the described process needs to run 5 times per client. In each run, the output (hour assigned to the client) will be erased from the training dataset, which invalidates that hour to be chosen again as the best hour to contact the client. Figure 4.7 represents a flowchart that summarizes the logic of the ranking process.

The previously described process can be a time/resources consumer, especially as we are dealing with a huge amount of data in a business context. So, and after analysing the results we created a simplified model by skipping the clustering phase. Instead of making a categorical representation of generic clients that can be contacted at each hour, it is possible to directly use the clients' features as the input to the supervised learning algorithm and the hours when an interaction with the CRM was made as the target.

The classification algorithm used in this approach was the XGboost as it is a decision tree-based algorithm like the Random Forest algorithm, but it has the advantage of being designed for speed and performance.

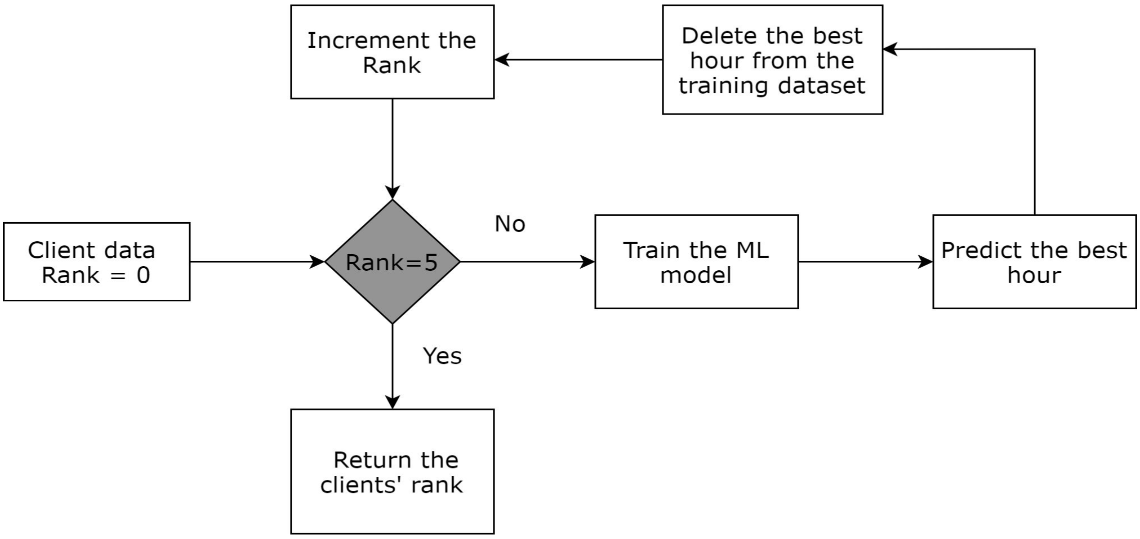


Figure 4.7: Representation of the ranking process

4.4.2 Weekday-based approaches

Considering the results from the temporal data analysis, the weekday dimension will now be taken into account. We identified two ways to add that component into our model: by making an hour ranking per client per weekday or by first discovering the best day to contact a client and then, making the hour ranking for that weekday. It is worth mentioning that only working days (Monday, Tuesday, Wednesday, Thursday, Friday) were considered as valid weekdays to contact a client.

The first approach was to develop a model that returns an hour ranking per client per weekday. This can be achieved by splitting the initial dataset by the day of the week of each contact. The result is a sub-dataset for each working day. Then, these sub-datasets will be used as the input of a supervised Machine Learning algorithm in such a way that five models are trained (one per working day). Each client's features are given to the models to obtain the best hour to be contacted per day. Once more, our goal is to find the top 5 best hours to contact a client, so the described process needs to run 5 times per customer, considering that in each run the previously assigned hour is deleted from the sub-dataset. The described process is represent in Figure 4.8.

Another approach to the problem is to take into account the weekday dimension. Instead of creating an hour ranking per client per weekday, it is possible to first find the best working day to contact the customer and then, based on that information, make the hour ranking. This model is similar to the previous one, containing one more step. In this case, the five sub-datasets (labeled with the correspondent weekday) are merged together to be the input of a supervised learning algorithm, whose purpose is to assign a working day to each client. Then, the ranking hour is obtained by giving the client's features to a second supervised learning algorithm trained only with the samples of the sub-dataset corresponding to the weekday attributed in the previous step. The Figure 4.9 is an scheme of the classification process.

All the supervised Machine Learning algorithms mentioned in this section were the XGboost algorithm.

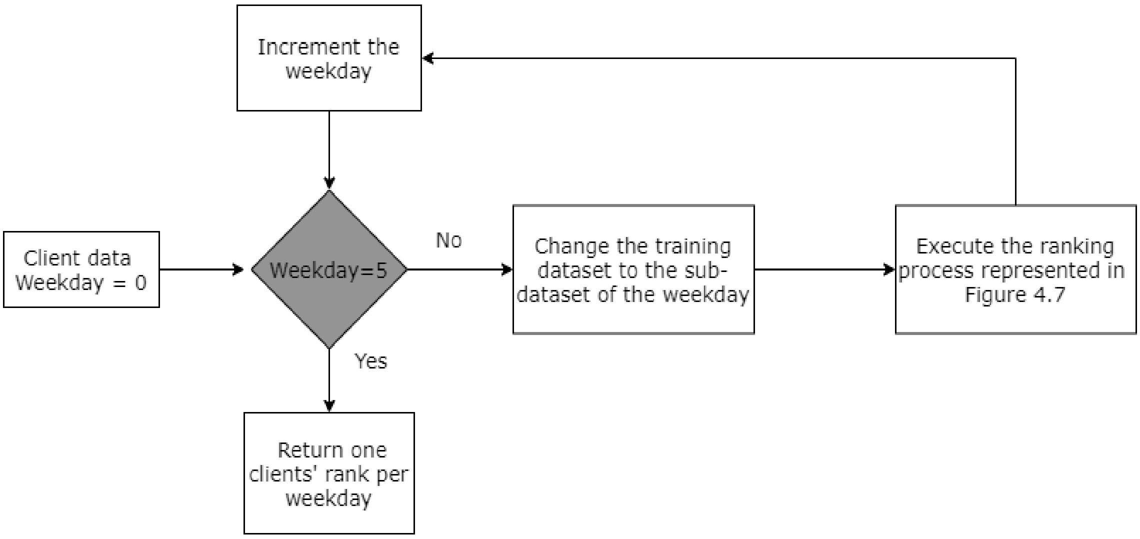

Increment the weekday
Client data Weekday = 0
Weekday=5
No
Change the training dataset to the sub-dataset of the weekday
Execute the ranking process represented in Figure 4.7
Yes
Return one clients' rank per weekday

Figure 4.8: Representation of the weekday-based approached to make a client's hour ranking per weekday

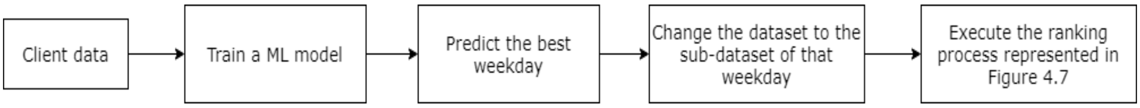

Client data
Train a ML model
Predict the best weekday
Change the dataset to the sub-dataset of that weekday
Execute the ranking process represented in Figure 4.7

Figure 4.9: Representation of the weekday-based approached to identify the best weekday and make the client's hour ranking

4.5 Evaluation

The models described in the section *Proposed Approaches* were submitted to an experimental testing phase, followed by an analysis of the results. All experiments were designed with the dataset division into training and testing by the ratio of 80%-20%, respectively. It is relevant to mention that this technique was used only in the Hour Ranking phase of each model. During the clustering Phase, the goal was to make an approximate characterization of groups of clients, so there was no need to train or test models.

One more point that needs to be clarified is the way we evaluate our approaches' final result. The success rate of each model was determined at the hour level. Each client belonging to the testing set can have up to 15 true labels as the client can have one successful interaction with the CRM per hour of the day. Taking this information into consideration, we divided the classification result by hour, and, for each group, we applied the following strategy: if the client had into their true label set the correspondent hour assigned by the classifier, we marked them as a success, otherwise they would not be accounted for. Finally, we divided the result for the number of clients of each group. This success rate model was created as an answer to the question: Using the predictive model to make a clients contact attempt queue per hours, what could it be our answering success rate per hour?

Tables 4.3 summarizes the experimental scenarios designed to test all approaches non weekday based.

We start by presenting and evaluating the results obtained in the hour characterization phase. For each clustering technique applied, we show an analysis of that process. In addition, we will interpret the formed groups considering prior knowledge about the customer's daily life patterns. Also, we would like to point out that although different distance

Hour Characterization		Ranking hour per Client
Clustering Algorithm	**Distance Algorithm**	**ML Algorithm**
Kmodes	Jaccard	Random Forest
	Gower	
Hierarchical agglomerative clustering	Jaccard	Random Forest
	Gower	
-	-	XGboost

Table 4.3: The table summarizes the experimental scenarios designed for non weekday-based approaches.

algorithms were taking into account to the calculation of the features interception, the results were very similar between the Jaccard algorithm and the Gower. So, their results comparison will be omitted as we think they do not add value to the present analysis.

Regarding the clustering implementation based on the Kmodes algorithm, we created an elbow plot to determine the optimal number of clusters into which the data may be clustered, Figure 4.10. The y-axis named as cost represents the amount of dissimilarity per number of clusters. The lower the cost value, the more homogeneous the data inside each cluster is. As we can see the curve does not show an evident elbow, which suggests that there is no optimal value for the number of clusters. However, we can consider that the number of clusters equal to 13 is a suitable candidate since, from that value on wards, the dissimilarity is no longer relevant.

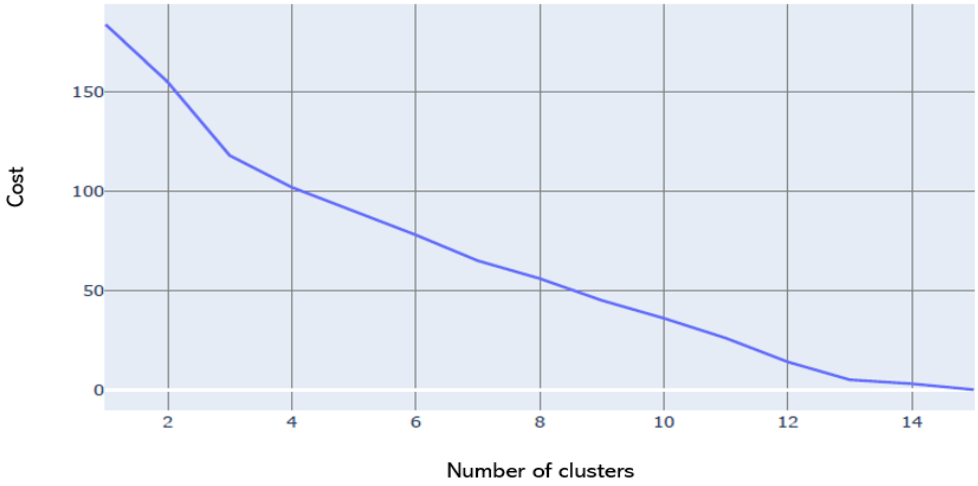


Figure 4.10: Elbow's plot

We believe that it is important to analyze the clusters' composition when the number of clusters varies. For that purpose, we ran the clustering algorithm with different numbers from 8 up to 13. The minimum value was chosen, considering that the goal of this study is to make a rank of hours and if we decreased the number of available labels to our classification, we could lose specificity, leading to an ineffective classification of all customers. The maximum threshold is the value obtained from the elbow's analysis.

In Figure 4.11, we can identify two strong clusters as they are the last to be separated. The one composed of 13, 17 and 22 hours and the group formed by 16 and 19 hours. Partially, they can be explained through the results of the temporal data analysis. The hour 16

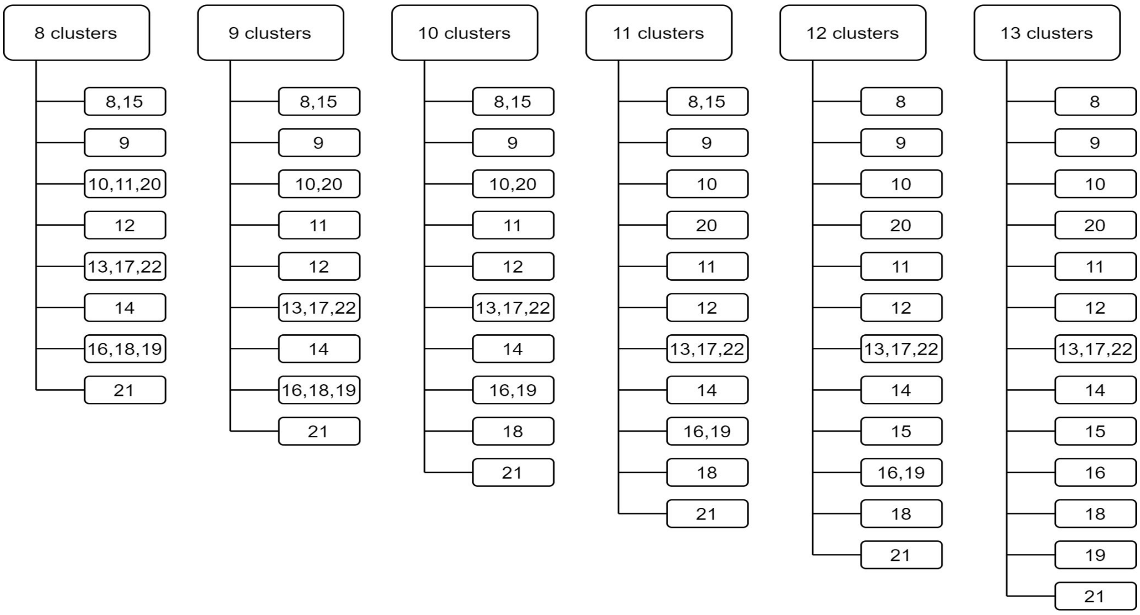


Figure 4.11: Clusters' composition using a Kmodes algorithm

showed a high concentration of interactions, while 13 and 22 represented hours with few contacts, which indicate that clients in these time slots share similar characteristics.

Now, considering the clustering implementation based on hierarchical agglomerative algorithms, we made a dendrogram analysis to discover the similarity between groups and a clusters' composition analysis like the one that was previously discussed. Once more, the results of the two experimentations with hierarchical agglomerative clusters were very identical. So, to simplify this evaluation we only present the results from the complete linkage. Also, at the same time, we take the opportunity to analyze the clusters' composition as both representations show correlated information.

Regarding Figure 4.12, we can see that consecutive hours are similar between each other. Also, we notice a clear division into 4 clusters: 8-9, 10-11-12-13, 14-15-16-17-18, 19-20-21-22. If we assign them a part of the day and considering this data was taken from months with few available daylight, we could get the following result: 8-9 (early morning period), 10-11-12-13(morning period), 14-15-16-17-18(afternoon period), 19-20-21-22(night period). Although our goal is not ranking day periods, this result is very interesting as we can interpret it and give it a logical meaning. In addition, the organization's services can benefit from these conclusions since it allows them to see which clients fit into each periods.

Taking into account the previous analysis, it is no surprise the composition of each cluster presented in Figure 4.13. They are formed by consecutive hours. Additionally, we can identify the afternoon period as the strongest cluster, particularly the hours 14, 15 and 16. While the results of the previous clustering method could be partially explained by the temporal data analysis, these results can be inserted between the range of hours with the highest number of calls (2pm - 5pm).

Now we are going to analyze the results of the ranking process. For that reason, we need to introduce the baseline model used to compare the results. Its purpose is to give us an estimation of the interactions' success rate obtained by the company by simulating the process of attempting to contact customers. At the end, we want to discover the success

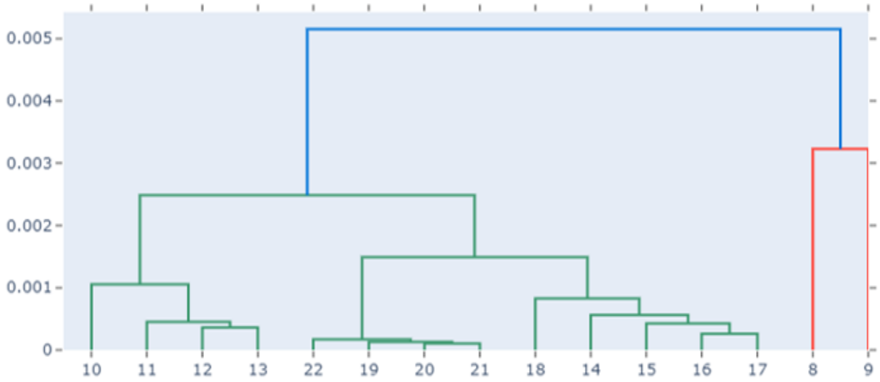


Figure 4.12: Dendrogram of the hierarchical agglomerative cluster algorithm

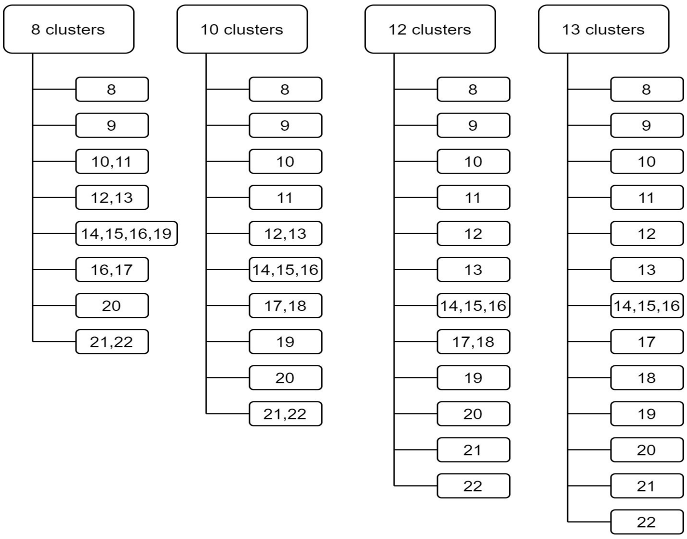


Figure 4.13: Clusters' composition using an hierarchical agglomerative cluster algorithm

rate per hour of the day.

So, from all sets of available clients we randomly selected 100 to be contacted at a determined hour of the day (between 8am and 10pm). Then, we checked if communication was successful by searching into our historical available contacts attempts. If the client had answered the call at that hour, they would be marked as a success and removed from the queue of attempting clients. Otherwise, the client was placed into the final of the queue to be contacted later. Last, we added new elements to the queue matching the same number of clients marked as success. This process is executed one time per hour of the day, and it is repeated 100 times to obtain the average of each hour's success rate.

After defining our base model, we can use it to compare each hour's success rate with the results of our approaches.

Up to this point we only discussed the results of our clustering-based approaches. It remains to analyze their ranking phase and the performance of the XGboost classifier. For that purpose, we produced Figure 4.14 which represents our gain/loss per hour compared to the described base model. This representation was made by calculating the success rate of each hour per ranking position, and then merging their results into a bar chart. In other words, we considered just the top 1 hours of each client's rank. The same operation was made regarding the second-best hour, then the third, fourth and fifth best hour. Finally, the five generated plots were overlapped, meaning that each bar represents the minimum value of success rate if the company tries to call 5 times to each customer (taking into account their hour rank).

The results of both cluster-based approaches show a very similar overall success rate, on average, so we omitted one of them to simplify this analysis. However, it is worth mentioning that they diverge when it comes to the individual values of the success rate per hour. These differences can be related with the clusters' composition of each approach as they did not agree on that point.

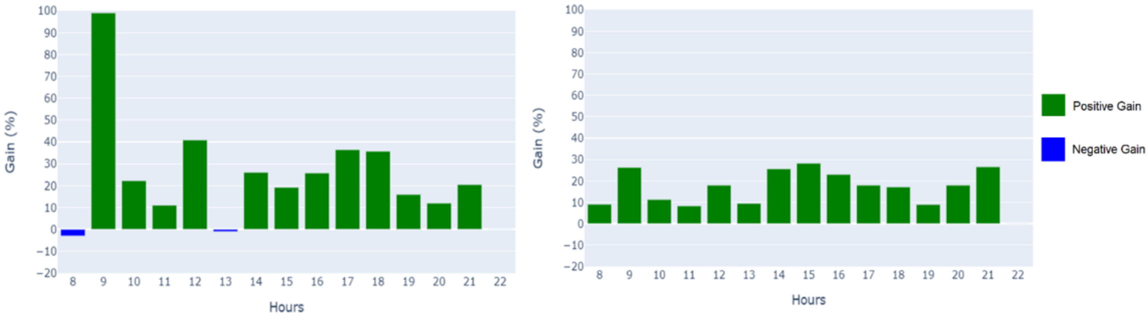


Figure 4.14: Test results of the model implemented with clustering phase (left graphic). Test results of the model implemented without clustering phase (right graphic)

Regarding the left panel of Figure 4.14, the approach shows an overall average gain of 24%. Hours 8 and 13 represent losses, where the model could not correctly predict the clients. Looking at the right panel of Figure 4.14, we can see that the model shows gains at every hour. The smallest gain value is 9% at 11am and the maximum gain value is 28% at 3pm. In general, the proposed approach has a gain of 17% on average, comparatively to the success rate of the one currently used to establish a communication with the clients. This is a positive indicator that the company could benefit from the implementation of an approach, as suggested in this section, to both define which clients should be contacted at each hour and to increase the success rate of their answering.

As a side note, we would like to point out that these results should be analysed with caution as the success' rate of the base model is a reality's estimation and they require a field validation.

To conclude the analysis of the non-week-based approaches, we want to emphasize that both models show gains at almost every hour. There are differences between the two approaches regarding computation times and complexity. The simplified model is about twenty times faster than the other to rank the same number of clients. These factors could influence the companies' decision when it comes to choosing which one should be implemented.

The evaluation of the week-based approaches will be presented by comparing the two models' success rates per hour per day of the week.

A brief perusal of the results in Figure 4.15 shows that we have a superior number of hours with gains than the ones with losses on all weekdays. Tuesday, Wednesday, and Thursday have a more uniform gain per hour, with only five hours without gains. Particularly, 8am and 8pm to 10pm could be considered difficult to predict. We would like to point out that

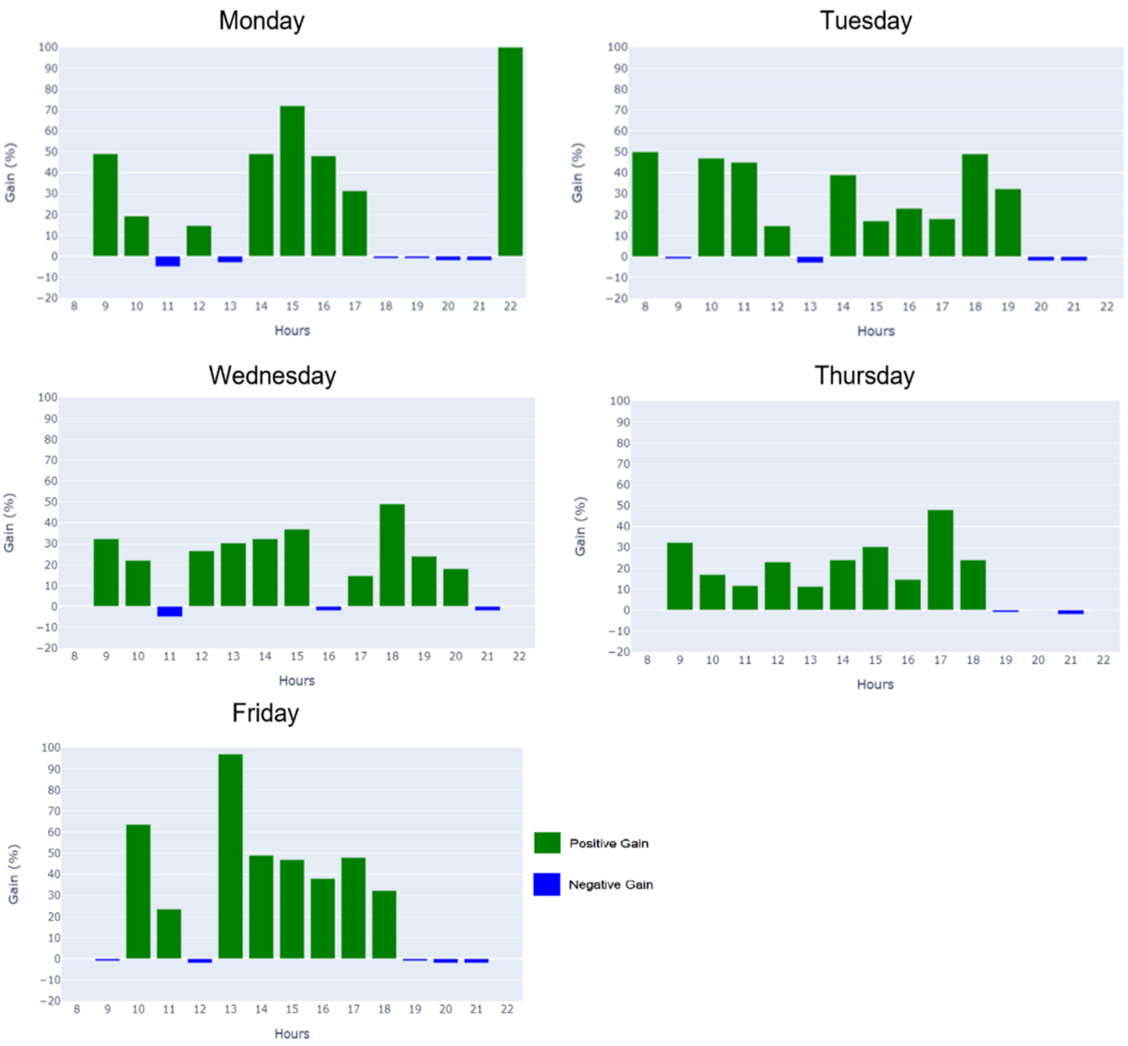


Figure 4.15: Results of the Machine Learning Model to the hour identification for weekday based approach

the hours with a perfect gain rate had few representative samples in the testing dataset and this factor contributed to their score. Nevertheless, globally, the results obtained show a gain of 23% on average, which shows that the implementation of a model as described in this chapter could support the decision-making process when it comes to making a calls' schedule based on working days.

Regarding the results of Figure 4.16 referring to the weekday-based approach where we first predict the best workday followed by the hour ranking, the results indicate a general average gain of 4%, with no gain consistency per hour between all weekdays. We expected these types of results as the model had the challenge of determining the best weekday to contact a client. In the testing phase its success rate was about 30%, meaning that the next classification step (the hour prediction) was made on top of untrustworthy results. This, in turn, could indicate that clients do not have a preferred weekday to make/receive voice calls. This conclusion can be supported by the temporal pattern analysis presented in section 4.2 that shows no evidence of a high concentration of inbound interactions regarding the workdays.

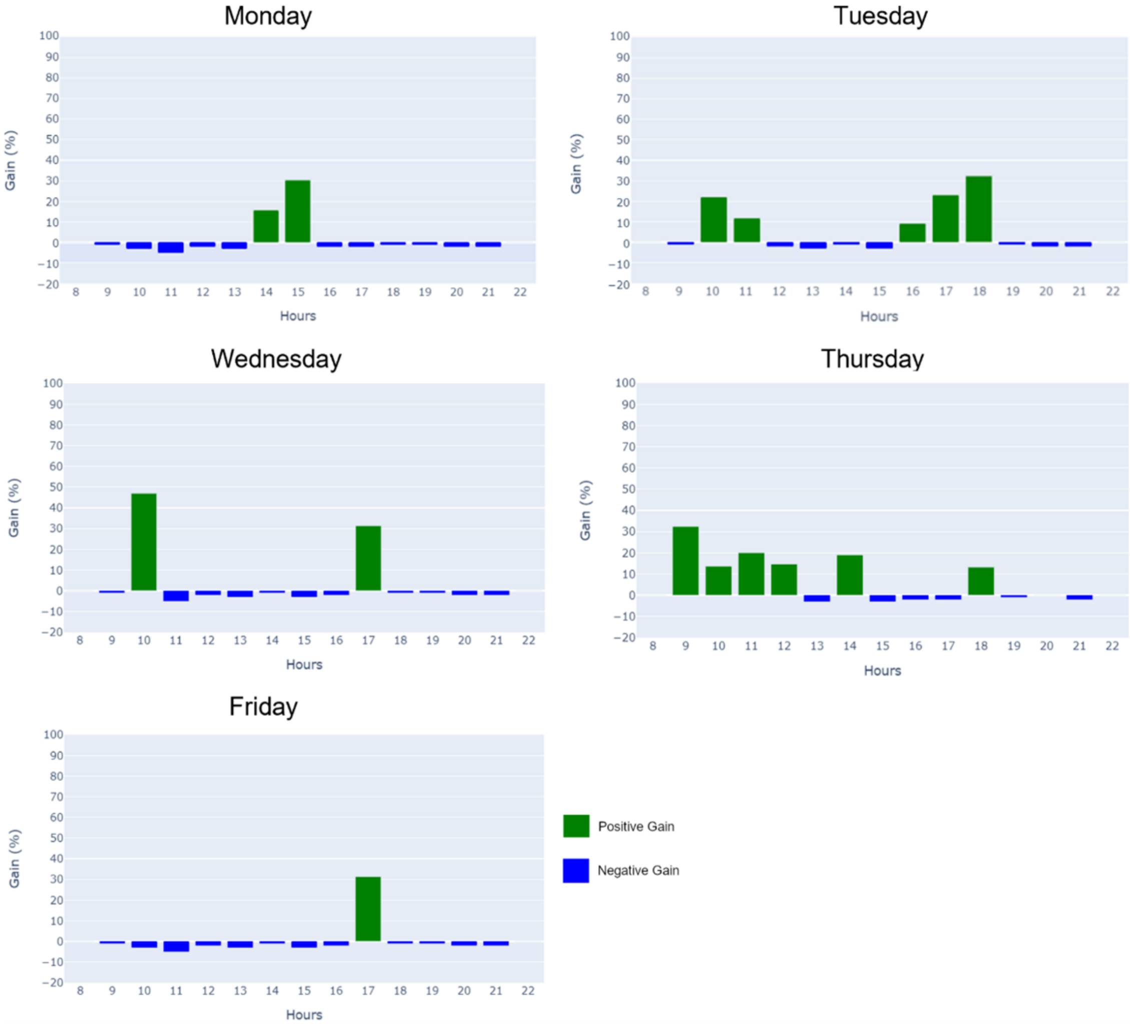


Figure 4.16: Results of the model that predicts the rank of hours per client based on the predicted working day

Non weekday-based approaches		Weekday-based approaches	
1 - Clusters based	**2 - Non clusters based**	**3 - Without the best weekday**	**4 - With the best weekday**
- General average gain of 24%	- General average gain of 17%	- General average gain of 23%	- General average gain of 4%
- Only 3 hours show no gains	- All hours show gains	- 8pm, 9pm and 10pm show no gains through all weekdays	- No gain consistency per hour between all weekdays
- Medium models' complexity: 2 steps of computation	- Low complexity	- Low complexity	- Medium complexity: 2 steps of computation
- Slow computation times (about 4 hours to rank 50 clients)	- Acceptable computation times (about 40 minutes to rank 200 clients)	- Acceptable computation times (about 3 hours to rank 200 clients)	- Acceptable computation times (about 1h 30 min to rank 200 clients)

Table 4.4: The table summarizes evaluation points for the 4 proposed approaches

Once more, these results should be considered with caution as the percentage of success of the base model is a reality's estimation.

The evaluation section ends with a summary of the main conclusions about the success rate of each approach and the discussion about both models' complexity and computation times in Table 4.4.

As this study was made from a business context, we think it is relevant to point out some information related to technicalities. Although approach 1 shows the best gains, from a company's perspective, it could be considered a time/resources intensive compared to approach 3 that shows almost the same gains and is about 5 times faster. At the end, the choice of the best model to be implemented will depend on the company's internal goals.

4.6 Results and Discussion

The results of each model were positive considering that the challenge was a 15-label classification problem. One of our big concerns was related to the number of available clients per class, particularly in the week-based approaches. Our dataset needed to have enough interactions per hour of every weekday, that lasted 3 minutes minimum and whose motive was related with costumer care or marketing. So, the success' rate evaluation model by class helped to decrease the impact of the unbalanced dataset.

Overall, each approach, excluding the last presented, shows gains at almost every hour. As it was mentioned, one problem that requires a detailed study is the determination of the best weekday to contact a client. Clearly it would be an advantage in the implementation of the last presented approach. Although the temporal data analysis did not show conclusive results to support this idea, meaning that, it is likely that clients have not a preferred weekday to make interactions with the CRM, more information about this problematic could be an interesting addition to future approaches.

Also, we would like to point out that our models could benefit from the addition of more features associated with the client itself as the studies presented in the state of art suggested. At the end, we can conclude that the prediction of the hours ranking per client was almost exclusively based on their daily life pattern calls.

The discussion regarding which approach is the best solution depends on the companies' intentions. All we can add is that the hour ranking model per weekday seems to offer a good trade-off between speed and success rate.

Nevertheless, an overall increase of any percentage in the amount of reach out clients by a company means a possible increase in the marketing campaign's success.

Chapter 5

Identification of the Best Contact Channel

The present chapter aims to provide the reader a description of the problem of identifying of the Best Contact Channel. It requires its own exploratory data analysis to help defining success metrics and evaluate the available data. We will use its results to explain both the feature engineering process and the approach's development. Finally, we will discuss the results obtained in the evaluation phase.

5.1 Problem definition

Deciding the communication channel is a key factor when contacting a customer. Our goal is to create a rank of channel preferences per client as it can help manage the resources without compromising efficiency.

Voice, SMS, and E-Mail are the most common means of interaction and, taking this into account, we can see the identification of the best contact channel as a classification problem with three labels. The challenging part was the labeling process, as the thresholds of a successful contact must be well defined, since each customer can fit into more than one category. We decided to consider that a customer prefers to communicate via voice calls if they have voice interactions with the CRM that last between 180 seconds and 3000 seconds. These threshold values were defined with the help of the Exploratory Data Analysis and the reason for this choice is presented in the sub-chapter 4.2. The email label was assigned to all clients that have sent at least one email to the CRM as it shows the customer's predisposition to initiate an interaction through that channel. It is relevant to point out that each interaction with the CRM has a motive associated. We only valued communications related to marketing, customer care or customer retention to exclude interactions originated by an emergency such as technical support, that could not indicate the customer's preference. Finally, the definition of a successful SMS communication was based on the clients' weekly SMS traffic. A customer should send a minimum of 8 SMS per week and a maximum of 60 to be classified as a potential SMS user. Once more the Exploratory Data Analysis of weekly SMS traffic, presented in the sub-chapter 5.2, was instrumental to decide acceptable values to consider that a client usually communicates through SMS.

Features	Number of null values	Amount of distinct values
device's description	736	11
tariff	0	159
bundle	12708	15
amount of accepted campaigns through portal in 2020	18998	12
amount of accepted campaigns through ivr in 2020	18998	29
amount of accepted campaigns through sms in 2020	18998	19
access to clients' area	12708	3
direct debit	12708	3

Table 5.1: The table represents a summary of the datasets' constitution.

5.2 Exploratory Data Analysis

The exploratory data analysis phase aims at answering to a set of relevant questions regarding not only the amount of data we had at our disposal to support the problem but also to know the consumption's behaviors of our clients and to use it to clarify the definition of the success criteria for each class.

From a total of 21042 clients, we created Table 5.1 to summarize two important variables and analyze the data volume: the presence of null values and the number of distinct values each feature can assume. The first one helps us to decide if the field is relevant or if it can be discarded for lack of information. The second one gives us an indication of possible feature transformations that can be made considering business understanding.

Starting with the null values analysis we could identify the features related to the number of accepted campaigns as a potential problem. About 90% of their information are non-useful information. So, we considered that our model would not benefit from these types of features, given that this feature was present in a small amount of samples.

Regarding the number of distinct values that each field can assume, we see that tariff's description (it represents the mobile tariff contracted by the customer) show a huge number of different values, which can indicate that the model will not consider this feature as important to the classification as the amount of correlated information between it and the label (channel) could be despised. So, taking this result into account, we designed two experimental scenarios where, in one of them, the feature tariff has 159 values and in the other this field only contains two categories, pre-paid and pos-paid.

To define the success criteria per channel, we analyzed the amount interactions with the CRM. We found that about 1272 clients made voice interactions with the CRM for motives related to marketing, customer care or customer retention with a duration between 180s and 3000s, which indicates that this definition can be used to label these clients as channel voice users. Considering email interactions, we discovered that only 448 clients had sent emails to the CRM for the same reasons. Nevertheless, we accepted this restriction as a valid success criterion for labeling the samples for the email channel. Although the SMS channel is the second most used to communicate with the client most of the traffic is outbound and there is no way of knowing if the customer was engaged in this type of interaction. So, we filtered the SMS communications by the inbound direction. There are only 30 clients that meet those conditions. It is clear that another label system must be defined for the SMS channel. To do this we decided to analyse the number of SMS that the client send per week. This can be achieved by looking into the CDR data.

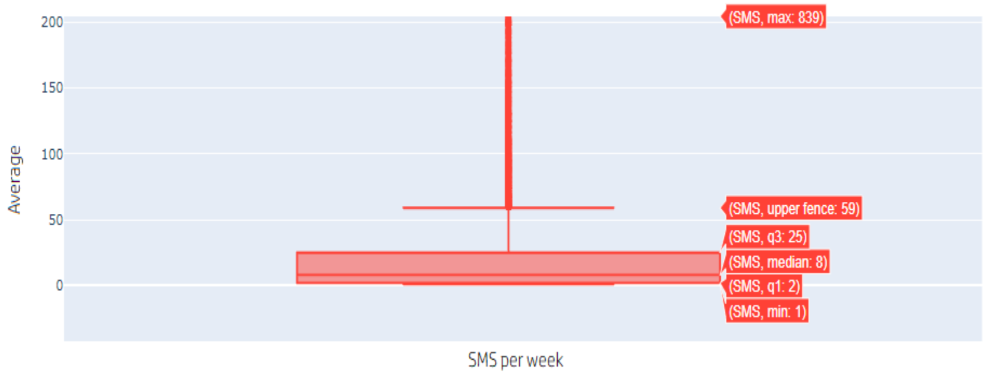


Figure 5.1: Weekly SMS traffic box-plot

The box plot represented in Figure 5.1 shows an analysis of the SMS traffic by week. It is relevant to mention that this traffic is related with the amount of SMS sent by clients, which means that the customer chooses the SMS channel to communicate on a daily life basis. The median and the upper bound are 8 and 59, respectively, and they appear to be acceptable values to characterize a potential SMS user. In addition, we also know that 50% of the customers are between these two values, so we have sufficient data volume to define the SMS success criteria as a user that sends a minimum of 8 SMS and a maximum of 60 SMS per week.

5.3 Feature Selection and Engineering

The dataset created for our problem was composed of samples extracted not only from the set of interactions' records between the client and the CRM but also from the voice, SMS and GPRS consumption records.

From the clients' component we selected the sex feature to be part of our dataset as an analysis of call durations showed that men and women have a clear difference when it comes to the amount of time they are engaged into a voice conversation. This feature was transformed into a numerical one with the help of the one hot encoder.

Regarding the card associated with each client we extracted information related to the type of device where the card is plugged in, if the customer has ever accessed the client area and if the customer has activated direct debit. These features can be viewed as the clients' technological characterization as they could indicate the propensity to adapt to modern technologies/services. In addition, we selected data related to the marketing component such as information about the customer's tariff and their historical records of accepting a campaign through SMS, voice or the online portal. All these features were transformed from categorical to numerical.

As a complement to this information, we engineered new features with the customers' voice, SMS and GPRS consumption records. The main goal is to provide our model with the importance of each component in the daily life of each client. So, we generated both daily and weekly amount of voice calls, SMS, data usage and the time spent during a data usage action per customer. Only the weekly granularity was chosen to be part of our dataset to avoid data redundancy.

Although we had identified the set of features presented in this sub-section as valid for the resolution of our problem, we made a feature selection based on the importance of these components for the classification model. This process is further detailed in sub-section 5.5.

Finally, we had to label our samples. The goal was to transform and select data that can represent successful voice and email interactions with the CRM and add samples of clients that choose the SMS channel to daily communications. This step was distinct for all three categories. Voice communications were filtered based on the call's duration; just the client-to-CRM email interactions were considered for this problem; and the customers with SMS communications between the defined thresholds were labelled as potential SMS users. Each category was represented by an identification number.

5.4 Proposed Approach

The identification of the best way to establish a communication with the client can be obtained designing a supervised learning algorithm for a three-class classification (Voice, Email and SMS). Our goal is to return the channel ranking per client and, to complete such task, the ranking model follows a similar approach to the previous problem. The process runs two times per customer, where the first iteration is a three-class classification, whose output is the best channel to contact that client and the next iteration is a binary classification between the other two classes. The last label is added to the final of the ranking as it is shown in Figure 5.2.

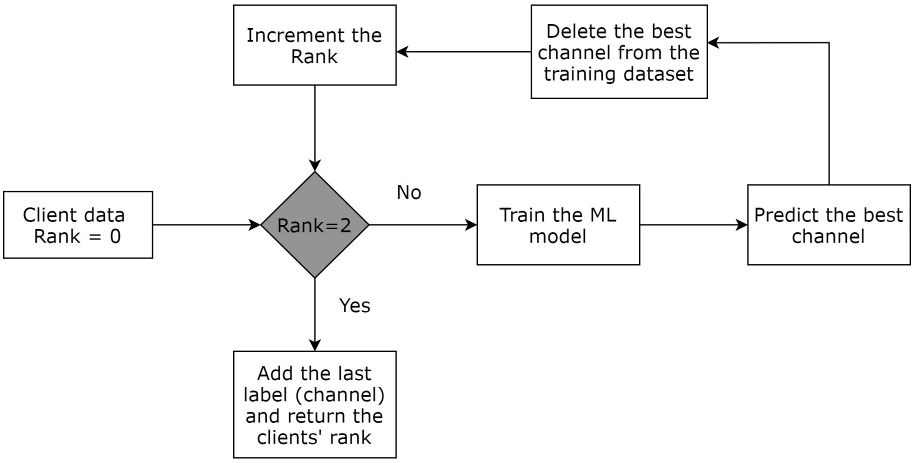

Increment the Rank
Delete the best channel from the training dataset
Client data Rank = 0
Rank=2
No
Train the ML model
Predict the best channel
Yes
Add the last label (channel) and return the clients' rank

Figure 5.2: Representation of the ranking process

5.5 Evaluation

The model evaluation presented in this sub-section is divided into two main goals: the comparison between models with the feature *traffic* processed differently (resulting from the exploratory analysis of this feature) and the results evaluation of the proposed approach in general.

All experimental scenarios were designed with the same data balancing technique (undersampling) to maintain an equal number of clients per class. Also, we used the ratio of

80%-20% to divide the dataset into training set and testing set, respectively.

The feature selection technique was made with the help of a tree-based classifier. Each model ran 6 times, one per set of features composed by a value between 5 and 10 (maximum value of features). Each run selected the most key features for the classification. It is important to mention that the raking stage includes both a tree-label classification and a two-label classification, and, for that reason, the features importance values were calculated considering the combination of 3 and 2 labels.

The success rate of each model was determined at the channel level. The same logic applied to the identification of the best hour problem was used here. The only difference is the number of available labels as each client can have up to three true labels (the client can meet the three labeling criteria at the same time).

Considering that the output of this approach is a channel ranking per client composed by three channels, the results of this process are shown per ranking position, Figures 5.3 and 5.4. Due to the feature selection technique, we can add that the model created with 159 tariff values shows the best results (on average) when it has 6 features, meanwhile the model with a binary traffic only needed 5.

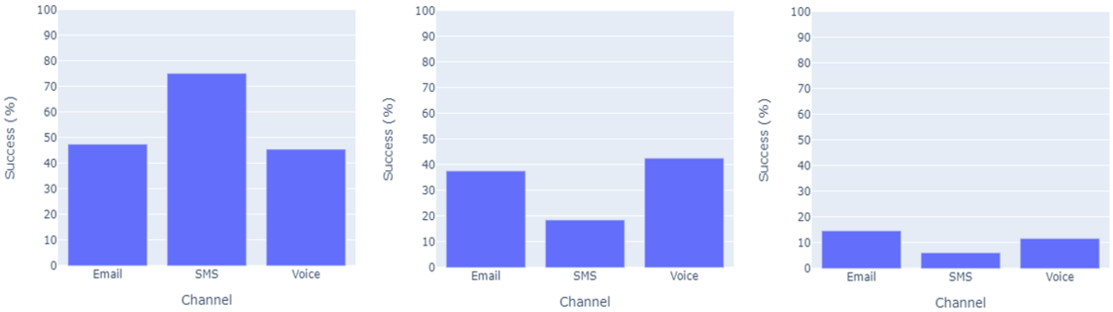


Figure 5.3: Success results of the model created with 159 tariff values. The first chart from left to right is related to the best ranked channel, the middle one represents the second best ranked channel and the last is the third best ranked channel.

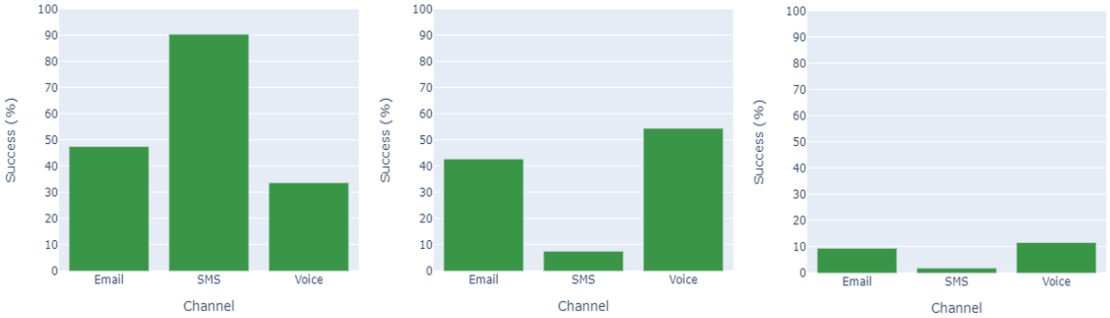


Figure 5.4: Success results of the model created with binary tariff values. The first chart from left to right is related to the best ranked channel, the middle one represents the second best ranked channel and the last is the third best ranked channel.

Although both representations show equivalent results, the main differences between Figures 5.3 and 5.4 can be found in the success rate of the SMS channel and the Voice channel regarding the best and the second-best ranked channels. While Figure 5.3 shows a Voice

	Metrics	Voice	Email	SMS
Model created with 159 tariff values (Model 1)	Average number of calls	30.2	16.87	16.38
	Rate of clients with number of calls higher than median (=10)	70	55	58
Model created with 2 tariff values (Model 2)	Average number of calls	24.44	14.92	24.62
	Rate of clients with number of calls higher than median (=10)	71	55	78

Table 5.2: The table represents the results analysis based on the frequency of calls on a daily basis.

success rate higher than Figure 5.4, this last one seems to have a higher success rate on the SMS channel.

To stress out which experimental scenario was more accurate, we decided to analyze if customers classified as Voice make frequent calls on a daily basis. For that reason, we defined two metrics: the average number of calls of the set of customers classified as Voice and the rate of clients with number of calls higher than the median. The last one was extracted from the exploratory data analysis made with the voice traffic data. Table 5.2 shows the results.

Looking at Table 5.2, we can observe that in Model 1, when considering the set of clients labeled as Voice, both metrics have higher values than the set of clients classified as Email or SMS. The same does not happen in Model 2, as it shows similar values for clients classified as Voice and SMS, which means that there is no difference in the voice call patterns between a customer that prefers the Voice channel and the one that prefers SMS. This analysis could indicate that the results of the model created with 159 tariff values are more reliable comparatively to the Model 2, since we expected clients classified as potential voice users to demonstrate a predisposition to use voice calls in their daily lives.

In order to finish the comparison between these two models, we would like to point out that the graphics related to the features' importance were evaluated as well. To simplify this analysis, we omitted the features' importance plot of the model with a binary tariff representation and we summarized their results. The other model will be analysed in detail when we discuss its results in general. So, related to this subject, we discovered that when the feature tariff was transformed to *pre-paid* and *post-paid*, the model did not consider it as important to make the classification. We observed that it was ranked on average as the 10th most important feature in a set of 10 features. On the other hand, when the feature tariff had 159 values, it was placed between the first 4 most key features for the classification. So, from this analysis, we can conclude that pre-paid customers and post-paid customers do not show a defined channel preference to communicate.

The second part of this analysis focuses on the results of the proposed approach. For this purpose, we selected to use the model created with 159 tariff values as it showed a better voice customer's classification in what concerns the call's behavior of a client that prefers the voice channel to communicate.

Overall, the results represented in Figure 5.3 show that a high percentage of clients are correctly classified considering just the preferred channel. We have almost 50% of success rate in both Email and Voice channels and 75% for the SMS channel. Globally, taking into account the first graphic, we have about 56% of success rate, which means that the channel assigned to the clients matches with their communication preferences. Additionally, it can be observed that these values are better than randomly selecting a channel, where each class will have an approximate success of 33% (considering a balanced dataset).

The three graphics show progressive losses from the best channel success rate to the third best channel success rate. This behavior confirms that the model can correctly classify a higher rate of clients when confronted with three labels comparatively with the two-label classification (represented by the middle chart) and the addition of the last label (represented by the last graphic). Thus, we can conclude that it will take a small number of iterations to successfully reach out to a customer, as the probability of choosing their preferred channel decreases with the increase of the channel rank position.

Regarding the features' importance analysis of this approach, the results are displayed in Figure 5.5. They show the importance of each feature when the target's model is the combination of these classes: Voice, SMS, and Email.

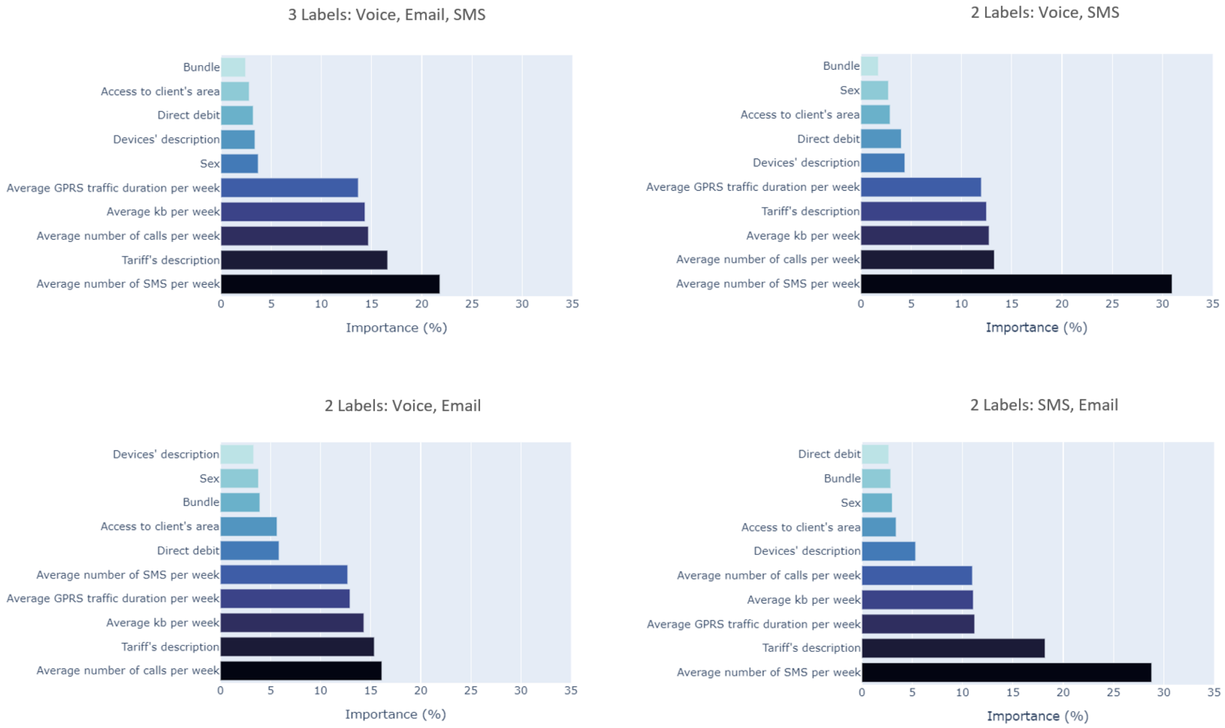


Figure 5.5: Feature's importance bar plot

Overall, looking at the 4 panels, we can see that features related to Voice/SMS communication patterns and data usage have the most influence in the classification process. Surprisingly, the tariff's description has a significant role as well, meaning that customers choose their preferred channel based on their tariffs, or vice versa. If we focus your attention on the two-label classification plots, it is interesting to observe that, depending on the classes involved, the most important ranked features are related with the weekly pattern for that channel. For example, when the model is making a classification between Voice and SMS, it rated both *average number of calls per week* and *average number of SMS per week* as the most relevant features for its decision. The same behavior happens for the pairs Voice-Email and SMS-Email, assuming that the data usage frequency can be considered as an indicator of the Email channel. To conclude, this type of analysis improved our trust in the results since these ratings are capable of being understood within the problem's context.

In addition to the analysis about the features' importance to make the predictions (inter-class analysis), we complemented this information with an analysis based on the features' importance for the characterization of the class, i.e., how much variance they explain inside each channels (intra-class analysis). For this purpose we used a Principal Component Analysis (in the dataset of each channel) to extract the variance values assigned to each feature from the principal component. Then, these values were transformed into importance through the following logic: each feature contributes for the principal component with a variance value, meaning that the lower this value, the more homogeneous the data are and more important this feature is for the characterization of the class. Both intra-class analysis and inter-class analysis are evaluated together as we want to discuss if the selected features are as important to characterize the channel as they are to make their distinction. The Figure 5.6 shows a graphical representation of the ranking of each feature considering the intra-class evaluation and the inter-class evaluation. The slope indicates the agreement level between the two analysis. A slope equal to zero means that the feature has the same importance regarding the two evaluation methods and, ideally, we want a scenario of total agreement.

Regarding the left panel of the Figure 5.6, we can see that the features *access to clients' area*, *debit direct* and *bundle* do not cross others and they show similar agreement level on both evaluations. This result is interesting as both analysis consider that they are not important to neither characterize nor distinguish the class Email (due to a large quantity of missing values detected in the exploratory data analysis). In turn, the features *tariff's description* and *average number of calls per week* are good features to characterize and distinguish the class Email since they obtained high importance values in both evaluations. The feature *Sex* is an example of a disagreement between the two analysis. It does not show much importance when considered to make a distinction between classes. The rest of the features show acceptable levels of agreement and their differences should be interpreted as: whenever the slope value is negative it indicates that the data carried by the feature are similar between each class (inter-class); whenever the slope value is positive it indicates that the data carried by that feature are different between each class (inter-class).

Considering the right panel of the Figure 5.6, we notice that, once more, the features *access to clients' area*, *debit direct* and *bundle* can be isolated from the rest. Their agreement level show that they are not important features for the SMS channel. On the other hand, the feature *average number of SMS per week* is the most important feature to characterize and distinguish the SMS channel. The rest of the features, except *average kb per week*, do not show high agreement level.

Finally, the middle panel of the Figure 5.6 is related to the voice channel. The features access to *clients' area*, *debit direct* and *bundle* are not important for the Voice channel. We consider that the features *average number of calls per week* and *average kb per week* are the most important to both characterize and distinguish the Voice channel (they only disagree in 1 position). Once more the feature *Sex* shows the higher disagreement level. Regarding the intra-class rank of the feature *tariff*, its position can be explained based on the contracted services of each tariff: usually each tariff includes free call minutes which allow almost every client to make as many calls they want regardless of the contracted tariff, so this feature has not importance to characterize the Voice channel. The rest of the features, show a good agreement level which indicates that they are as important as their rank.

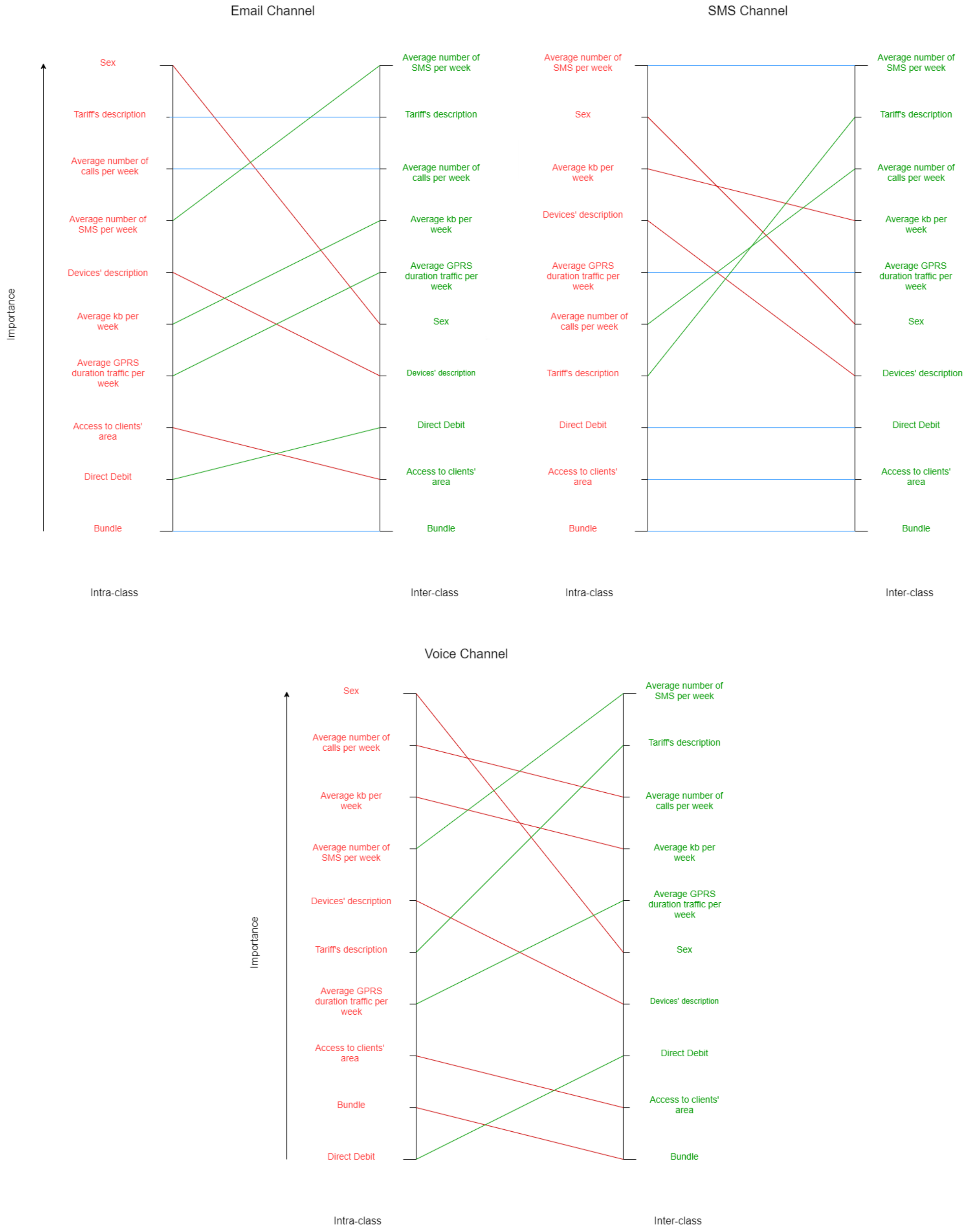


Figure 5.6: Left panel: intra-class analysis and inter-class analysis of the Email channel; Middle panel: intra-class analysis and inter-class analysis of the Voice channel; Right panel: intra-class analysis and inter-class analysis of the SMS channel

To conclude, we can summarize the results of the intra-class analysis and the inter-class analysis in the following points:

- When considering just the inter-class analysis, we can conclude that the three most important features have a clear correspondence with the most important features appointed in each class: *Average Number of SMS Per Week* is related to the SMS channel; *Tariff's Description* can be associated to the Email channel; and *Average Number of Calls Per Week* is a feature related to the Voice channel.
- The 3 least important features to the inter-class analysis correspond to the 3 least important features of the intra-class analysis of each channel.
- The feature *Sex* and *Device's description* show negative slope in all the three panels of the Figure 5.6 meaning that they show homogeneity data within classes and between classes, which make these features not important for the classification.

These results are in agreement with the set of best features selected by the classifier to achieve the best success rate values. The number of selected features were 6. All features *Average Number of SMS Per Week*, *Tariff's Description*, *Average Number of Calls Per Week*, *Average kb per Week*, *Average GPRS traffic duration per Week* and *Sex* can be interpreted regarding their value for each class. In addition, almost none of them have a negative slope regarding the intra and inter class analysis, meaning that although they show heterogeneous data when compared with other features within the same class, this heterogeneity is not significant to affect their contribution for the distinction between classes.

5.6 Results and Discussion

In this section we want to discuss two points: the results achieved, in general, by the proposed approach and the results obtained, in particular, by the channel SMS.

Globally, the model's success rate is positive as it surpass the threshold of 33% per label reached by randomly selecting a channel to contact a client. The results are even more robust when we add both conclusions from the features' importance evaluation and the clients' pattern calls evaluation. Knowing that we can interpret the features rank regardless the channels involved in the classification and clients classified as voice users make in fact a lot of calls on a regular daily basis (when compared to clients marked as email and SMS users) give us confidence in the results.

Regarding the success rate obtained by the channel SMS, we consider that its value can be influenced by the success criteria defined at the labeling process. This channel is the only one whose criteria depends on the clients' communication behavior. We used the number of weekly SMS sent to first mark our samples, then, we included them into our predictive model and finally they were used to determinate the SMS classification performance. Nevertheless, both complementary analysis (features' rank and voice call patterns) proved that the model learned in fact how to differentiate the three labels.

Chapter 6

Conclusion and Future Work

Marketing campaigns are used by companies to reach their clients to both let them know the services they have to offer and to maintain a good relationship with them. The selection of the timing and the channel are two important factors to take into consideration for the success of marketing campaigns. In this work our goal was to determine the best time to contact a customer and to discover which channel is the most appropriate to establish the contact. The output of both problems is a ranking system to help companies to better manage their resources.

Regarding the identification of the best contact hour problem we found interesting results throughout all the development process. In the Exploratory Data Analysis we discovered that clients have preferred day periods to contact the CRM through phone calls. Also, we detected that working days have an higher concentration of contacts than weekends. Taking that into consideration we used historical data to make different approaches (weekday-based and non weekday-based) mostly relying on the clients' patterns calls. The analysis of our models allow us to find important results that can be interpreted considering both temporal data analysis and its context. Particularly, depending on the algorithm used to make the clusters, we identified that the group of customers who are available at 1pm, 5pm and 10pm (Kmodes algorithm) and the group of customers who are available at 2pm, 3pm and 4pm (hierarchical agglomerative cluster) represent both strong clusters which indicates they share similar characteristics. In what concerns the estimation of the best time to contact a client, 3 of our designed models obtained positive results as they showed gains on average between 17% and 24% comparatively to the success rate currently obtained by the organization. Although the validation stage was limited and needs to be improved with a validation made in a real world campaign context, we believe that these models could support the decision-making process when it comes to making a calls' schedule. Finally, we discussed the trade-off between performance and speed of each approach, giving more information to companies select the best option for them.

Considering the results obtained in the identification of the best contact channel problem we were able to achieve gains of about 34% (taking into account just the best channel of each client's rank) when compared to randomly selecting a channel to make the contact. Although the approach shows positive results, our validation phase would benefit from a real business context validation. In addition, we found interesting results through the features' importance analysis since we could conclude that the three most important features to make the predictions have a clear correspondence with the most important features appointed in each class: *Average Number of SMS Per Week* is related to the SMS channel; *Tariff's Description* can be associated to the Email channel; and *Average Number of Calls*

Per Week is a feature related to the Voice channel. Also, we discovered that a customer division into pre-paid and post-paid clients decreases our trust level on the results as the values of the features importance could not be easily interpreted and the pattern calls of clients classified as Voice users were similar to the ones classified as Email or SMS users.

To conclude, throughout this process we achieved our goals and most of all we demonstrate the practical applicability of this study since, currently, the models presented in this work are being integrated into the workflow of the organization.

6.1 Future Work

As it was discussed in the *Context* of this document, the literature is rather scarce in the area of our study. As such, the work conducted in this dissertation presents several opportunities for the development of further research.

One problem pointed out regarding the identification of the best contact hour was the determination of the best weekday to contact a client. We believe that this problematic requires a detailed study to verify the importance of the dimension weekday when the clients' availability is discussed. From a practical perspective, organizations could benefit from an improved implementation of the weekday-based approach presented in the Chapter 4 as they can take advantage of these discoveries to better manage their resources. Knowing both weekday and hour to contact a client could decrease the amount of required tries to contact a costumer.

The approaches based on a clients' characterization showed overall positive results when predicting the hours ranking per client. However, this characterization was almost exclusively based on the customers daily life pattern calls, and it did not contain personal features that could help the grouping process. We think that knowing the customers' career, for example, can add information about unavailable periods of time of a each formed group, enriching their characterization.

Finally, considering the results of the identification of the best channel problem, we concluded that the channel SMS obtained higher success rate values compared to voice and e-mail. In addition, we discussed that its values could be influenced by the success criteria defined at the labeling process. So, we consider that it could be an interesting opportunity to investigate other ways of defining what is a potential SMS user.

References

[1] Tuve Löfström. *Utilizing Diversity and Performance Measures for Ensemble Creation.* 2009.

[2] Beatriz Nery Rodrigues Chagas, Julio Viana, Olaf Reinhold, Fabio Lobato, Antonio F. L. Jacob Jr., and Rainer Alt. Current Applications of Machine Learning Techniques in CRM: A Literature Review and Practical Implications. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018.

[3] Qeethara Kadhim Al-Shayea. *Evaluating Marketing Campaigns of Banking Using Neural Networks.* 2013.

[4] Konstantinos Tsiptsis and Antonios Chorianopoulos. *Data Mining Techniques in CRM.* 2009.

[5] Femina Bahari T and Sudheep Elayidom M. *An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour.* 2015.

[6] John A. Bunge and Dean H. Judson. *Encyclopedia of Social Measurement.* 2005.

[7] R.S.J.d. Baker. *Data Mining.* 2010.

[8] Osmar R. Zaïane. *Principles of Knowledge Discovery in Databases.* 1999.

[9] Chris Rygielski, Jyun-Cheng Wang, and David C.Yen. *Data mining techniques for customer relationship management*, volume 24. 2002.

[10] Edelstein H. Data mining: exploiting the hidden trends in your data. *DB2 Online Magazine.*

[11] Mark Ryan M. Talabis, Robert McPherson, and D. Kaye I. Miyamoto, Jason L. Martin. *Finding Security Insights, Patterns and Anomalies in Big Data.* 2014.

[12] Sanatan Mishra. Unsupervised learning and data clustering, 2017.

[13] Radhwan H. A. Alsagheer, Abbas F. H. Alharan, and Ali S. A. Al-Haboobi. Popular Decision Tree Algorithms of Data Mining Techniques: A Review. *ResearchGate*, 1017.

[14] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to data mining.*

[15] Niklas Donges. A complete guide to the random forest algorithm, 2021.

[16] Gaurav. An introduction to gradient boosting decision trees, 2021.

[17] Vishal Morde and Venkat Anurag Setty. Xgboost algorithm: Long may she reign, 2019.

[18] JAYITA BHATTACHARYYA. Understanding xgboost algorithm in detail, 2020.

[19] Swati Gupta. A Regression Modeling Technique on Data Mining. *International Journal of Computer Applications (0975 – 8887) Volume 116 - No.9*, 2015.

[20] Neha Gupta. Artificial Neural Network. *Vol.3, No.1, 2013-Selected from Inter national Conference on Recent Trends in Applied Sciences with Engineering Applications*, 2013.

[21] Ajith Abraham. Artificial Neural Networks. 2005.

[22] Harika Bonthu. Kmodes clustering algorithm for categorical data, 2021.

[23] Bhanwar Saini. K-means and k-modes clustering algorithm, 2021.

[24] DATA NOVIA. Hierarchical clustering in r: The essentials, 2021.

[25] Khalid Rababah, Dr. Haslina Mohd, and Assoc. Prof. Dr. Huda Ibrahim. A unified definition of CRM towards the successful adoption and implementation. 1, 2011.

[26] Buttle F. Customer relationship Management: Concepts and Tools. 2004.

[27] M. Geib, L. M. Kolbe, and Brenner. CRM collaboration in financial services networks: a multi-case analysis. *Journal of Enterprise Information Management, 19(6), 591 - 607.*, 2006.

[28] Liou and J. J. H. A novel decision rules approach for customer relationship management of the airline market. *Expert Systems with Applications, 36(3, Part 1), 4374-4381.*, 2009.

[29] Schierholz, R., Kolbe, L. M., and Brenner. Mobilizing Customer Relationship Management; A Journey from Strategy to System Design. *In proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006. (HICSS '06).*, 2006.

[30] M. F. Moller. *A scaled conjugate gradient algorithm for fast supervised learning.* 1993.

[31] Sérgio Moro, Paulo Cortez, and Paulo Rita. *A data-driven approach to predict the success of bank telemarketing.* 2014.

[32] Fabrice Talla Nobibon, Roel Leus, and Frits C.R. Spieksma. Optimization models for targeted offers in direct marketing: Exact and heuristic algorithms. *European Journal of Operational Research, Volume 210, Issue 3*, 2011.

[33] Corinna Cortes and Mehryar Mohri. AUC Optimization vs. Error Rate Minimization. 2003.

[34] David S. Coppock. Data modeling and mining: Why lift?, 2002.

[35] Kashyap Trivedi. The best day and time to make sales call, 2019.

[36] Alex Lamascus. Best time to cold call, 2020.

[37] James D. McCaffrey. Example of calculating the gower distance, 2020.

[38] Fatih Karabiber. Jaccard similarity.

Appendices

This page is intentionally left blank.

Appendix A

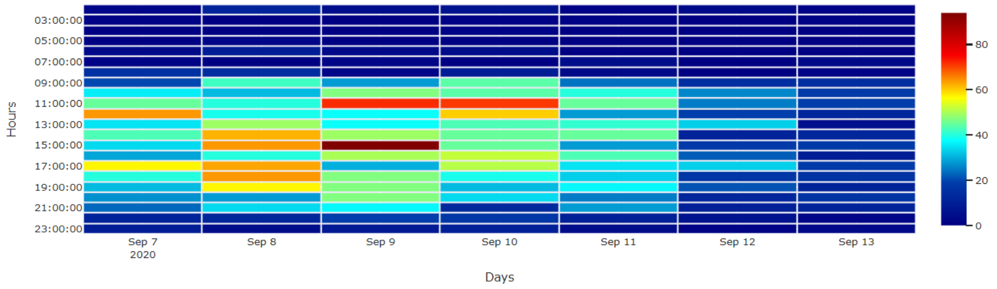


Figure 1: Heatmap Sep 9 2020 - Sep 13 2020 for inbound communication and duration between 0 and 3000 seconds

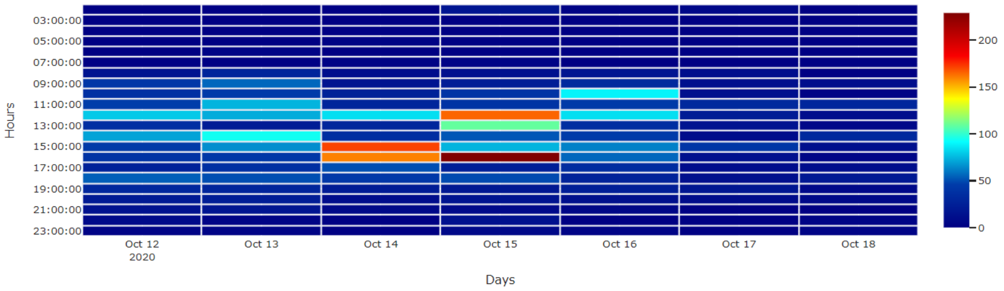


Figure 2: Heatmap Oct 12 2020 - Oct 28 2020 for outbound communication and duration between 0 and 3000 seconds

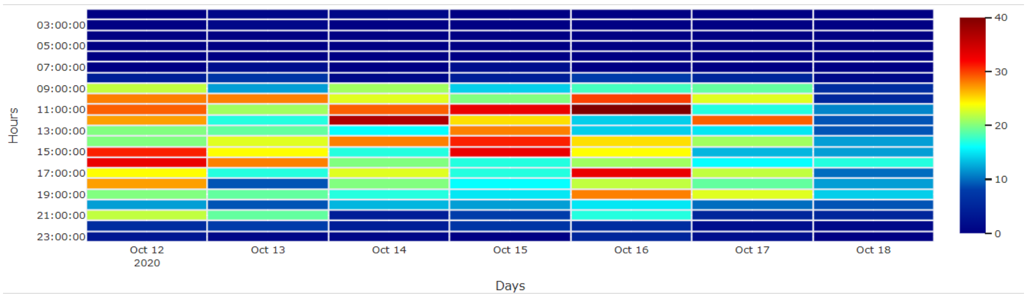


Figure 3: Heatmap Oct 12 2020 - Oct 18 2020 for inbound communication and duration between 180 and 3000 seconds

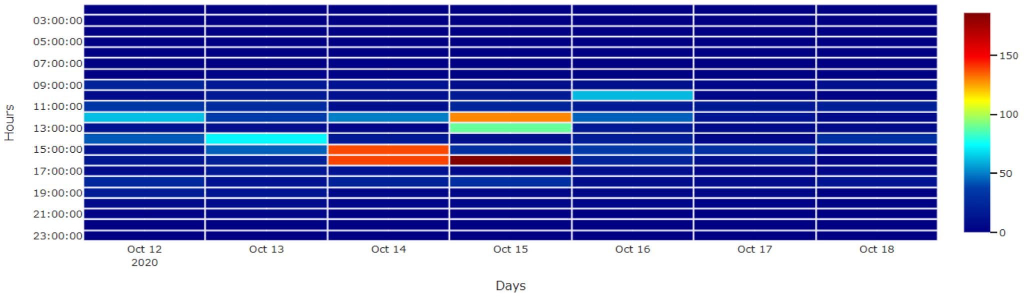


Figure 4: Heatmap Oct 12 2020 - Oct 18 2020 for outbound communication and duration between 180 and 3000 seconds

Appendix B

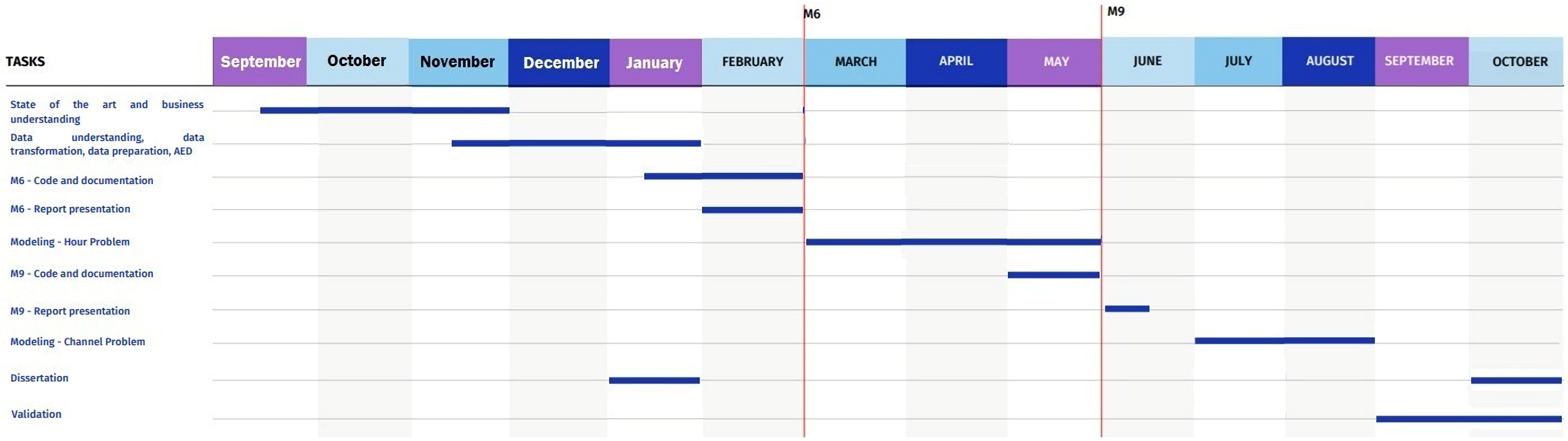


Figure 5: Gantt's Diagram with planned tasks

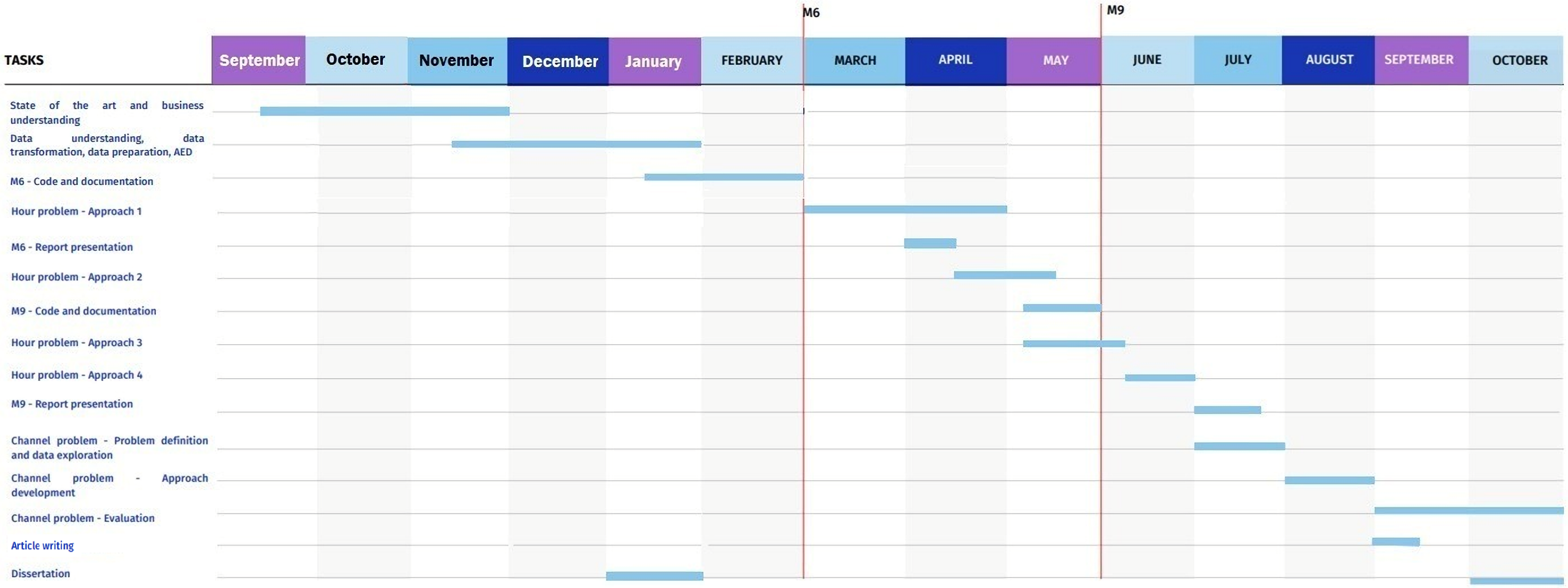


Figure 6: Gantt's Diagram with the real time allocation