NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Customer Analytics: Data Integration and Visit to Stores Study

at Altice Portugal

Mariana Nunes Domingues

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

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

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

CUSTOMER ANALYTICS: DATA INTEGRATION AND VISIT TO STORES STUDY

by

Mariana Domingues

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Supervisor: Professor Pedro Cabral

External Advisor*:* Marco Viana

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Mariana Domingues

Lisboa, 30 novembro 2022

ACKNOWLEDGEMENTS

To my family and friends, a thank you is not enough for all the support during this journey, which was the master's degree, which proved to be more challenging than expected. A deep thank you to Gui and André for the understanding and motivation they gave me.

I also have to thank Marco Viana for all the support, suggestions, and sharing of knowledge; for his patience, guidance, and trust.

To Inês Nunes, Luís Mestre, and the entire team that welcomed me with open arms, thank you for integrating and transmitting a sense of teamwork.

To Professor Pedro Cabral, thank you for your availability, monitoring, and help as a guide in the preparation of this report.

ABSTRACT

As in other industries, in telecommunications it's increasingly important to know our customers, analyzing the data that we manage to obtain and those that customers provide us. This is also important for us to have a competitive advantage over our competitors. This document describes the work on some projects developed during an internship at Altice Portugal, in the Consumer Segment department, describing the implementation and methodologies used in data analysis. Concepts that were necessary for the elaboration of the projects are presented, as well as the tools used to develop them. One of the projects was the development of a predictive model in which the objective was to assign a customer to one of our active stores, trying to understand which variables, in addition to the distance to the store, can influence this choice. The work carried out allowed us to have a more unified structure of information, without redundancies and several sources of information; it also allowed reviewing obsolete processes and information that is no longer useful today, cleaning up necessary resources; and finally, it allowed us to get to know the customer better, understanding better their visits to the physical store.

KEYWORDS

Data Science; CRISP-DM; Machine Learning; Supervised Learning; Multiclassification.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CTT	Correios, Telégrafos e Telefones – commonly known Correios de Portugal, S.A
CP7	7-digit postal code
CRISP-DM	Cross-Industry Standard Process for Data Mining
ESRI	Environmental Systems Research Institute
GIS	Geographic Information Systems
INE	Instituto Nacional de Estatística
KDD	Knowledge Discovery Databases
PT	Portugal Telecom
SEMMA	Sample, Explore, Modify, Model, Assess
TDP	Teledifusora de Portugal, S.A.
TLP	Telefones de Lisboa e Porto
TMN	Telecomunicações Móveis Nacionais, S.A

1. INTRODUCTION

1.1. INTERNSHIP OVERVIEW

The internship began in September 2021 and ended in September 2022, at Altice Portugal. There, I had the opportunity to experience considerable integration into the company, since I joined a large group of new interns at Altice, which allowed me to carry out activities whose purpose was to understand the operation and interests of the company, where it operates and what are their concerns.

I joined the Consumer Segment Management, more precisely the team that supports and helps in planning decisions in this department. My duties went through that, preparing reports and developing files necessary for the proper functioning of the company.

1.2. COMPANY

Altice Portugal is currently the largest telecommunications company in Portugal. We can say that it has been through a lot since its history dates back to 1877.
In the beginning, many concessions were made and in 1968 the Public Company Telephones of Lisbon and Porto (TLP) was created and CTT (Post, Telegraph, and Telephone) provided services in the remaining areas of Portugal. In 1991, TMN (Telecomunicações Móveis S.A) and Teledifusora de Portugal (TDP) were created, responsible for the infrastructure business in this area; there was also the first Lisbon/Macau videoconference. One year later, Telecom Portugal, SA is created, thus creating a separation between Telecommunications and CTT. In 1994, Portugal Telecom (PT) came into existence, joining the companies Telecom Portugal, TLP, and TDP. To expand its horizons, Portugal Telecom then bought Telesp Celular in Brazil. Later, in 2000, the company, which until then was not private, carried out its privatization almost entirely, leading the telecommunications sector in Portugal. Since then, its innovations and willingness to do more for the customer are visible in what has been achieved (Altice, 2022).

Subsequently, the Altice group bought the previous PT, changing its name to Altice Portugal.

Currently, the company Altice Portugal operates in several areas, including the development of technology and its support, support and integration of products in the various businesses, and encouragement of new ideas and projects.

1.3. OBJECTIVES

Since the start of the internship, I was introduced to the systems that my role implies and received a proposition for a very interesting project to develop a reference table, at the database level, with the objective of unifying the information that the company have, that could be consulted, and which could be enjoyed by all who needed it. This table is a reference table at the CP7 level which has geographic, internal and company information. Within the information that this table has, there is the information of the store closest to each CP7, in terms of time traveled from the centroid of CP7 to the store associated with that same CP7.

It was then proposed that I developed an analysis that would allow us to make this store association in a more granular way, that is, to try to understand the store for each customer, not only resorting to the location of CP7. So, the goal of this project was to understand which client variables influenced their choice of physical store and be able to assign the more accurate store to each customer.

The database migration project aimed at migrating processes and useful tables without loss of information.

2. LITERATURE REVIEW

In this chapter, concepts that were necessary for the development of the various projects developed during the internship are discussed. Its reading is important to understand the techniques used in the rest of the work.

2.1. DATA SCIENCE

Nowadays it is increasingly important to transform data into added value for organizations, using the mottos of our institution “From Data to Value” and “Data with Purpose”. For this, we must be able to distinguish what are valuable data (interesting data) from data that only cause turbulence (data that come from statistical fluctuations) and through the former, be able to draw inferences (Shamoo and Resnik, 2009), using data analytics, where we resort to various tools to describe and analyze the data.

Data Science encompasses several areas, such as statistics, data analytics, modeling, and algorithm development, to improve processes (Zhu & Xiong, 2015). It also uses data visualization when, but not only, explores the data, through graphs that help us better understand what the data tell us; and, in addition, it also makes use of data transformations depending on the objective, as well as various tools to ensure the success of the project it intends to respond to. Zhu & Xiong (2015) argue that a data science project includes certain steps, such as identifying the objective of the project, importing data to respond to that same objective, processing data including exploration and cleaning, data modeling, and application of the developed model.

This interesting idea of being able to analyze the data we have at our disposal, and with the great increase in interest in this field that is Data Science, in the 21st century, the data scientist profession is considered the sexiest (Patil & Davenport, 2012).

2.2. DATA MINING PROCESS

Data mining is defined as the extraction of knowledge from data (Witten, Frank, Hall, 2011), according to Hand and Adams (2015), is the analysis of observational data sets to find relationships and summarize the data in different ways that are useful and understandable, and is an application of machine learning (Kotsiantis, 2007). The purpose of these techniques can be, for example, customer segmentation, churn forecast, and better application of marketing strategies, among others. Data Mining is an interdisciplinary field of computer science that combines several tools, including machine learning (Hand, Mannila, & Smyth, 2001).

One of the consequences of the increase in the use of technology and information systems is the increase in the amount of data, so, more data to analyze. Therefore, some methodologies or processes were created that allow organizations to implement projects to get more value from them. Examples of these same technologies are KDD (Knowledge Discovery Databases), CRISP-DM (Cross-Industry Standard Process for Data Mining), and SEMMA (Sample, Explore, Modify, Model, Assess). These methodologies were also guides in the elaboration of this project, being the CRISP-DM the one most followed.

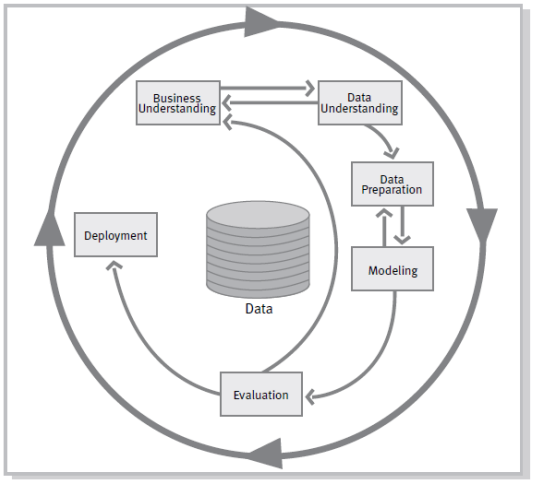


Figure 2.1 – Phases of the CRISP-DM reference model (Chapman et al., 2000)

The process consists of 6 phases that form a cycle, which encourages you to go back to the starting point and evaluate the conclusions. The first step, Business Understanding, aims to understand what is intended with the project at the business level so that it can later be translated into a data mining objective, as well as the choice of an evaluation metric that allows quantifying the success of the project. In the Data Understanding phase, we explore the data that is provided to get to know them better so that we can make better decisions in the future, at which stage we can also find problems related to data quality. When we get to the Data Preparation step, we are ready to make changes to the data, such as cleaning the data, simplifying the values of some variables, creating more interesting variables, and selecting variables, among others. In Modeling, models are usually tested that meet the questions raised in the Business Understanding part, and in Evaluation, we see what values the test results of the previous step obtained to choose the most appropriate model. In the Deployment phase, the objective is to deliver the results obtained with the model developed to the client.

As can be seen in Figure 2.1, this process goes backward and forwards, not following a line, a cascade, which allows the improvement of the results, as well as reviewing the initial question itself.

2.3. MACHINE LEARNING

Nowadays, one of the most attractive topics is the concept of machine learning. Through it we were able to develop algorithms that allow us to describe, predict, and understand various actions, among others, using a computer and data. A model is trained, using a subset of the available and worked data, giving different weights and importance to the different variables available; later it is evaluated with the remaining data that did not enter for the training. As Tom M. Mitchell said, machine learning is “the study of computer algorithms that automatically improve through experience and the use of data” (Mitchell, 1997). This method aims to recognize and extrapolate patterns and learn from them through data, adapting to new circumstances so that in the future it is possible to make predictions without human intervention (Zhang, 2020 and S. J. Russell and P. Norvig, 2016).

The core of this concept is learning from the data we have and that we know so that we can extrapolate to unseen and unknown data. For this, there is a process that makes use of models that are adjusted according to what the data tell us, changing according to what is intended, which is often the decrease in the loss function. This process helps us solve and understand several issues, the main ones being regression problems, classification problems, classification of continuous and non-continuous categories, and recommendation systems. These models work basically by analyzing the inputs and outputs from previous events, that is, we have data for which we already know the conclusion for them. For example, we have data from customers that we know have already churned and we try to understand which variables or indications that may make us better understand that our current customers are at risk of taking this action; or, for example if a customer has purchased a certain selection of products, better understand the relationship between the purchase of certain products, so that we can recommend product B to those who purchased product A.

In recent years, machine learning has been increasingly relevant due to its computational capacity as well as the increase in the amount of data that was created, and the reduction of the necessary investment.

Machine learning is divided, accordingly to Peter Norvig and Bishop, into 3 types presented below (S. J. Russell and P. Norvig, 1995 and Bishop, 2006).

Supervised Learning

Supervised learning is a type of machine learning in which the data is previously classified, and our goal is to be able to predict the classification of new data, that is, it uses the data that is already classified as a basis to predict the classification of unclassified data (Talabis et al., 2015). These base data are an input-output pair, we have the variables considered independent or input, which characterizes the observation, and the dependent variable, which will be the classification or output of the observation with certain independent variables.

There are also two types of algorithms in this type of machine learning: classification algorithms and regression algorithms. The former is used when the variable to be predicted is a categorical variable, which is finite and therefore limited, while regression ones respond to forecasting problems in which the dependent variable is a continuous numerical variable.

Unsupervised Learning

In this concept of machine learning, the objective is to find patterns, since we do not have the data labeled or classified (James et al., 2014). These patterns are found through the variables we have, using algorithms that allow us to group data into segments based on similar features, and non-evident relationships, among others (McCue, 2015). Types of unsupervised learning are, for example, customer segmentation or customer clustering.

Reinforcement Learning

In this training method, the process is carried out through interaction with an environment oriented towards decision-making, which interacts with the algorithm, which then stores information and receives feedback (positive or negative), i.e. trial and error, based on actions passed (Sodhi et al., 2019). Included in this type of learning are genetic algorithms, and swarm intelligence, among others.

2.4. IMPORTANCE OF KNOWING OUR CUSTOMERS

With the increase in data, and more specifically, customer data, companies began to pay more attention to them. The company where I did my internship, one of the main telecommunications companies nationwide, is no exception to this rule. With the amount of data that the company holds, it is often difficult to know where to start. However, if we want to continue to prosper, we must know to whom we want to sell our products, that is, we must know our customers. Kotler & Keller (2006) argue that companies should focus on the customer and know more. And what better way to get to know our customers than to explore the data they provide us?

The fact that we know our customers better leads us to make more personalized offers, intending to increase satisfaction and, consequently, customer loyalty. This, for example, can lead to a reduction in churn, which should be something companies should invest in to increase competitiveness, and try to eliminate defections as much as possible, preparing to identify and act according to each customer they want. defect (Reichheld & Sasser, 1990).

We must also improve the customer experience, which goes beyond just marketing (Schmitt, 2019); that also incorporates, but not only, the customer experience when in contact with the brand/company, for example in its stores.

When there is contact with the customer in a store, this allows us to direct the customer to the offers that we find most suitable with greater ease, since there is human contact, physical contact. It is also easier to ask questions and clarify them. Another reason is the offers that the brand/company makes available are different depending on the region, this is because, in the specific case of Altice, we have fiber in many points but not everywhere, and this means that each customer needs to have an even more personalized experience.

However, companies should be cautious with the personalization of offers, since customers may also find that the proposals are too personalized (Bleier & Eisenbeiss, 2015a; White et al., 2008), realizing that the company has too much information about them.

2.5. GEOGRAPHIC INFORMATION SYSTEMS

With the growing amount of customer data, GIS (Geographic Information Systems) have come to help better understand the customer, allowing the use of geographic data for internal reports, marketing intelligence, and research, helping decision-making (Hess et al. 2004), being increasingly present in strategic decision-making (Nasirin, 2003).

GIS are systems that allow us to combine geographic information with the information we have in databases, which makes their use an asset to any company. Geography can respond to the different needs of both the company and the customer. According to the ESRI (Environmental System Research Institute) GIS "is a system that creates, manages, and maps all types of data. GIS connects data to a map, integrating location data (where things are) with all types of descriptive information (what things are like there)", allowing you to identify patterns, relationships, and geographic contexts.

With the perception that, nowadays, the information that data gives us is very important, being a source of knowledge and information that allows more reasoned decision-making, it is necessary not only to collect data but also to analyze it. Only in this way we can have data-driven organizations.

3. METHODOLOGY

During the internship at Altice Portugal, which took place between September 15, 2021, and September 14, 2022, I was allowed to have contact with different areas as well as different platforms.

Being part of a company created in me an even greater sense of responsibility as my work started to have an impact on processes and the work of other people.

One of the first tasks I was assigned was the creation of a table that would provide all the information available to date on the CP7s in Portugal. This table is currently used by our management teams and is updated daily. Two larger projects also emerged, which later ended up overlapping: when we were asked to analyze which store is closest to each CP7, information that was later integrated into the aforementioned table of the first task, the interest arose in studying a little better what influences our customers' choice of the store; the other project was the oracle database migration, which took a lot of time and dedication from the team.

3.1. TECHNOLOGIES AND TOOLS

ORACLE DATABASE

Oracle Database is a relational database management system that was developed by Oracle Corporation in the 70s (Oracle Corporation, 2017). It is a database that stores database files, which works based on rows and columns in tables, which we can cross with other tables, through the attributes of the tables, we can manage observations, insert, and delete observations, etc. It uses the language of PL/SQL (Procedural Language/Structured Query Language). The PL/SQL language allows for greater ease of data manipulation and is a language that combines the SQL language with procedural programming language.

Some of the advantages of this database are the easy scalability of the data, that is, being able to respond to the user's needs by dealing with the increase or decrease of resources; the guarantee of data isolation, that when we access the data in a reading operation we see a constant instance of the data, this type of database does not allow performing other operations that conflict with the action performed by the user, such as an update of data, alerting, for when this happens for a write-write conflict.

TOAD FOR ORACLE

Toad for Oracle (Tool for Oracle Application Developers), currently a product of Quest Software (Quest, n.d.), was developed in the mid-1990s by Jim McDaniel. This tool allows users to access oracle database data, as well as run scripts, query data, and edit data, among others. As a TOAD user, compared to other tools such as SQL Developer or Navigator, I think this tool is very user-oriented and for quick and practical actions, such as sorting a certain column, after running the query.

ARCGIS

ArcGIS is a software system that provides various tools and applications to the user, which was developed by ESRI (ESRI, 2020). It is a platform that allows the development of geographic information systems, working both the data and adding value with their location, in different areas of business and people management. This software allows you to combine your information with existing information systems in companies.

PYTHON

Python is a very flexible high-level language, which allows you to do various actions on data. This type of language can be used in several libraries, which have built-in modules related to each other; these libraries make it possible to carry out certain actions in a simpler way, using already tested and frequently used codes that define functions that are used repeatedly. Numpy, Pandas, Matplotlib, Seaborn, SciPy are some of the most popular Python libraries.

To make use of this language, we can resort to Jupyter Notebooks, which are interactive platforms, in which we can run a piece of code and visualize its result; we can also write text on them, as well as, in addition to data tables and the like, we can see the graphs and visualizations built. We were able to access this tool through Anaconda (Anaconda, 2020), which is a distributor of the Python language that also allows you to create several work environments, with different libraries.

3.2. TECHNIQUES

This report presents, among other projects, a project that was a multiclassification problem, and therefore, in this part we will focus on methods and techniques used when we talk about supervised learning.

3.2.1 Feature Selection

One of the main phases of a machine learning and data science project is to understand the best variables that can justify and help solve the problem at hand. After extensive exploration of the data, as well as the creation of variables based on the ones provided to us, it's time to select which ones we want to use in our model. This is because an increase in the dimensionality of the data makes the data more dispersed (N. Venkat, 2018) and for that, normally, we should focus only on a subset of variables of the initial variables. The consequence is the elimination of factors that are not interesting, that is, too divergent, or redundant, with high correlations; and thereby improve the performance of the model by making it simpler, as we try to circumvent the risk of overfitting (Yao et al., 2018), which happens when the model justifies quite well the data it has trained with, but when new values are shown, their performance decreases. It also makes it possible to obtain faster algorithms and improve predictive capacity and understanding (Kumar & Minz, 2014).

This process can be done through filter methods that rank the available variables regardless of the algorithm used in the modeling; wrapper methods, which is a method that goes back and forth, once it chooses certain features, trains the model, tries again with new features (adding or removing features) and sees if the model's performance has increased; or through embedded methods, which tries to combine the best of the two previous methods, which are, for example, Lasso and Ridge Regressions.

3.2.2 Classification Algorithms

When we have a problem where we consider applying machine learning, we need to choose what kind of algorithms we are going to use. One of the first questions to ask is whether we want an algorithm that allows us, humans, to understand what happens behind it, that is, an algorithm that can be interpreted and explained, where it is noticeable what weight has been assigned to the different variables. and how they have an impact on the variable to be predicted, called the White-Box models; or if we prefer an algorithm that has higher performance but that cannot be explained or understood so easily, the Black-Box models.

In this project, we decided to follow 2 white-box models, Logistic Regression (Multinomial) and Decision Trees, and 3 Black-Box models, Random Forest, Gradient Boost, and Extreme Gradient Boost Classifier.

Logistic Regression

In this multiclassification problem, we cannot use logistic regression with the base values, since this algorithm can respond to problems in which the dependent variable has only two possible values, and therefore we must resort to multinomial logistic regression. This modification uses the logistic regression model as a basis, causing the multiclassification problem to become several binary classifications, that is, partitioning the classification into several binary classifications with logistic regression applied to each subclassification.

Logistic regression has a continuous variable as its dependent variable, which is the probability of a certain outcome or event happening. This algorithm finds the relationships between the independent variables, or variable, and the event we want to predict the dependent variable; this helps the algorithm to understand what weights to assign to the different variables, which are later transformed into logarithms and combined linearly. Each of the variables is then multiplied by its weight, which is obtained through the maximum-likelihood estimation probabilistic framework (Kuhn and Johnson 2013), and added to the function, which is then adjusted using a logistic or sigmoid function, which makes the values obtained are between 0 and 1, these results being the probability of a certain event occurring (Andrew Ng., 2012). Logistic regression is then a discriminative classifier (Nasteski, 2017).

Decision Trees

Decision trees are models that can be used both in classification problems, in which the model's output is discrete and in regression problems, in which the output is a continuous variable.

One of the great advantages of this algorithm is the fact that it elaborates rules, which are, for the most part, easy to interpret. The idea behind this algorithm is the decomposition of a large and complex decision into small decisions, decisions that are reflected in rules that the model elaborates and that can be easily interpreted and expressed in words (Berry & Linoff, 2009).

Decision trees are built using recursive partitioning and adjust the models to each partition (Loh, 2011), consisting of nodes, branches, and leaves. It starts with the existence of a base node that connects to other nodes, and so on; the nodes represent the rules, that is, an observation arrives that checks the node rule or not, if it checks this condition, it follows one path, otherwise it follows the other path. These nodes leave branches, which are the paths that the observation will follow depending on its characteristics, and the independent variables. The leaves are the representation of the output, that is, in our case, the classification of the observation. The algorithm chooses the best division as well as the best variable to start at the base node and the following nodes and branches (Sammut & Webb, 2017). In Figure 3.1 we can see a brief example how decision trees work.

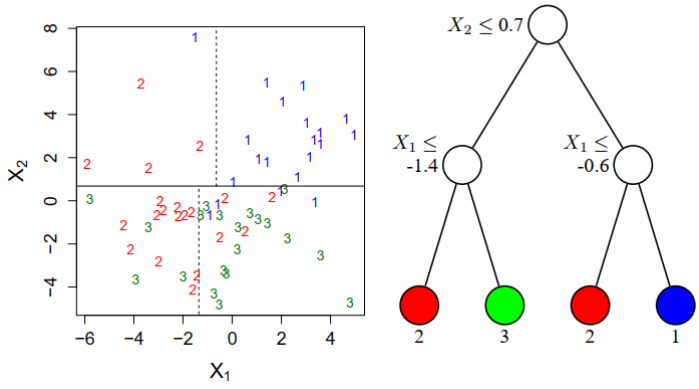


Figure 3.1 – Decision Tree Classifier Example (Loh, 2011)

Random Forest

Like decision trees, random forests are also capable of answering classification and regression problems.

This model is an ensemble method based on decision trees, and it is a meta-estimator. It includes several decision trees that elaborate their rules in parallel, bagging, based on samples of data and variables selected at random from those available (Nagpal, 2017), defining their output. In our case of multiclassification, in the end, the algorithm sees which class had the most votes in all weak models and that is the one that wins, this means, that class is the output of the random forest.

The fact that it is an ensemble method that works with independent decision trees makes overfitting more unlikely to happen and the model becomes more accurate.

Two of the disadvantages regarding decision trees are the non-interpretability of the model developed and the fact that it is a model that is more expensive computationally.

Gradient Boosting

Gradient boosting is a technique that belongs to boosting methods, where several "weak" predictive models are trained, which are usually decision trees (Dorogush et al., 2018. This technique, which can be used both in regression and classification, was presented by Friedman (2002) and aims to improve models sequentially, allowing a "weak" model to learn from the mistakes of the previous one, that is, a model will give more weight to observations that did not have the correct classification (Bishop, 2006) in the previous model. In the end, we will have an algorithm that is a linear combination of several base learners (G. Wang et al., 2011). The Extreme Gradient Boosting and the Light Gradient Boosting Machine were the chosen frameworks to train our model.

Extreme Gradient Boosting

This technique is also an ensemble of decision trees, which was introduced by Chen and Guestrin (2016), with the aim of optimizing the previous model, Gradient Boosting. Extreme Gradient Boosting, XGBoost, is a model that intends to circumvent overfitting, using Gradient Boosting "weak" learners and trying to arrive at the best decision tree model. It also has the advantage of being more computationally efficient.

Light Gradient Boosting Machine

This grandient boosting framework, Light Gradient Boosting Machine, LightGBM, was developed by Microsoft, and uses decision trees as weak learners. It manages to work with large-scale datasets, both in terms of observations and variables; not needing so much memory to be trained and being faster than the previous ones speeding up to 20 times comparing to the traditional gradient boosting (Ke et al., 2017), maintaining a good performance.

3.2.3 Evaluation Metrics

After training a model, we must try to understand if it applies to other datasets, that is, we have to understand if the model is performing well or not.

In the case of classification problems in machine learning, we have several metrics that can be useful to us, depending on the consequences of errors in our predictions. Let's see the following image (Figure 3.2), representative of the confusion matrix of a problem with binary classification:

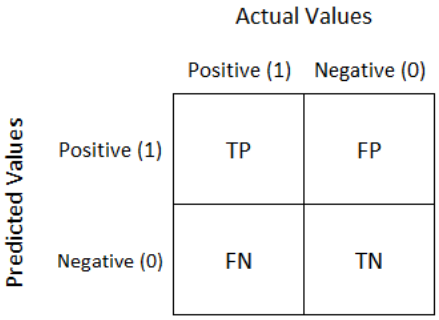


Figure 3.2 – Confusion Matrix (Narkhede, 2018)

TP – True Positives: Observations that were predicted to be positive and that are actually positive

TN – Observations that were predicted to be negative and that actually are negative

FP – Observations that were predicted to be positive but actually are negative (we make an error)

FN – Observations that were predicted to be negative but actually are positive (we make an error)

This confusion matrix alone helps us to understand a little how our model is performing, but it is also helpful in creating other metrics that are also useful for our evaluations, such as the following:

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

It is the proportion of observations that were correctly classified, considering all observations.

$$Error\ Rate = \frac{FP + FN}{TP + FP + FN + TN}$$

It is the proportion of observations that were incorrectly classified, considering all observations.

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

Precision is the proportion of observations that were correctly classified as positive, among all observations that were classified as positive.

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

It is the proportion of observations that were correctly classified as positive, considering all observations that are actually positive (including those that were misclassified as negative).

$$F_1 Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The F1 score is defined as the harmonic mean between precision and recall, the closer to 1 the better is the performance of the model.

We can also have these metrics associated with multiclassification problems by averaging the metrics across the classes. We are able to apply the above formulas also from a perspective of one class vs the other classes.

We can use the F1 score in cases of multi-classification as an evaluation metric by averaging it, however, we must do the Weighted F1-Score when we are faced with data that is not balanced.

3.3. DATA

For all the projects developed in this internship, I used the company's internal information. This information came from tables referring to customers, or geographic units. We regularly receive information from outside sources, such as CTT and INE (Instituto Nacional de Estatítica). However, this information had already been integrated into the company's information systems.

For the design of the reference table, I used data at the CP7 level, as well as information on the parish, county, and district where the same was located. I also made use of data related to variables and characteristics of the business

In the database migration project, no concrete data was needed, just confirm the need of them and migrate them.

In the analysis made of the visits to the stores, I resorted to data collected during the interaction, in the store, or even after it, which are integrated into our systems for consultation and development of other projects. I also used information regarding our active stores and also information that is in the reference table mentioned above.

4. RESULTS

I will first talk about the creation of the CP7 reference table, then the database migration challenge, and finally the analysis made regarding the customer's choice of store.

4.1. GEOMASTER

GEOMASTER, as it is commonly called within the administration, is a table that contains the characterization of the various postal codes in the country, characterized with business and external information. Currently, we have registered 202790 postal codes with an average of 31 accommodation units per CP7.

The interest arose in having a unique geographic segmentation information base within our department, to facilitate access to it, and to have better organization and coherence. There was a need for unification.

At the beginning of this challenge, it was necessary to gather all the information that customers, in this case, the teams in our consumer segment department, could need. For this, a survey of variables that each team would like to have in this table was made. A .xlsx file was made available and all teams were asked to indicate the variables they would like the table to be prepared to have.

When the teams finished filling in the file, we started working on what would become Geomaster. Several variables were disregarded at the beginning since they were not related to CP7s but to customers or services. Subsequently, within the remaining variables, filtering was carried out to determine whether or not it made sense to keep some variables because they were already obsolete or were fed by obsolete processes. After all the filtering together with the various teams, we arrived at the final variables, those that would appear in this table.

I then moved on to the table development part, where I would get the values of variables from other tables, helping me at the time to understand how the processes worked, the commonly used techniques and even to get to know our business a little better. At this stage I created the code in Notepad++ (a text editor) and tested it using the TOAD tool, running the queries necessary for my tests. During development, more variables, that had initially been approved, were eliminated, for several reasons.

The construction of the table is based on starting with the characterization of each CP7 and adding the variables that were within the scope of the project. After a lot of testing, since it would be a reference table, we moved on to the delivery part.

We created documentation with the names of the variables in the table, their types, and their description, as shown in Figure 4.1.

Variável	Tipo	Descrição
COD_DIA	VARCHAR2 (8 Byte)	Código do dia de processamento YYYYMMDD
CP7	VARCHAR2 (255 Byte)	CP7
DTCCFR	VARCHAR2 (255 Byte)	Código de freguesia anterior a 2013 (4260 freguesias) no formato numérico. Ex: o código 034523 fica apenas 34523
DICOFRE	VARCHAR2 (6 Byte)	Código de freguesia anterior a 2013 (4260 freguesias). Ex: o código 034523 fica 034523
FREGUESIA	VARCHAR2 (100 Byte)	Nome da Freguesia anterior a 2013 (4260 freguesias)
DTCC	VARCHAR2 (255 Byte)	Código do Município anterior a 2013 (4260 freguesias), com o 0 inicial
MUNICIPIO	VARCHAR2 (100 Byte)	Nome do Município anterior a 2013 (4260 freguesias)
DT	VARCHAR2 (255 Byte)	Código do Distrito anterior a 2013 (4260 freguesias), com o 0 inicial
DISTRITO	VARCHAR2 (100 Byte)	Nome do Distrito anterior a 2013 (4260 freguesias)

Figure 4.1 – Documentation's example of Geomaster

Sheets were also created, in a .xlsx file, with the information that each team should follow, since they asked for variables with a certain code/name and in the Geomaster table, they had that different code/name.

Initially, the table had 100 variables and currently contains 122 variables. The added variables also underwent a validation process, both for utility and for confirmation that they were associated with CP7. New variables were also added to the documentation.

Currently, the table is updated daily and feeds analysis processes, and strategic marketing processes, among others, supporting operation and decision making. The update is done using a SQL script that runs through a trigger in the Task Scheduler application which then triggers the start of a Batch file which in turn initiates the table update process.

This table feeds various processes from various teams within our process, which in itself is already a positive point. Being functional and people using the information created is a very positive point of this project.

The file created for the various teams to be able to change their processes, in which we have the information, and the description of the variables is also used today as a table dictionary. All elements of our department have access to this document and can consult their information, to better understand what variables, exist to use them in their developments

The Geomaster table was an asset for my personal development, it was a project that I developed and that I managed to get to a good port, as it is also an asset for our management of the consumer segment since the teams start to have this information in a single place, facilitating access to it, as well as having a great impact on the standardization of information. As I mentioned before, it helped me to learn about the normal layout and development of a project in our department.

4.2. Database Migration

This project arose even before it started working at Altice Portugal, in the direction of the consumer segment. The analytics database migration was proposed by the IT services department because the machine where the data was stored at the time was going to be turned off. So, we passed the data we had in an oracle database to another, also an oracle database, more recent and with greater capabilities.

This project had the participation of the entire management since it was in this machine that DSC kept its information and data necessary for the continuity of the services provided.

The project started in 2021 but was only completed in October 2022.

This challenge began by requiring a survey of the processes, both automatic and manual, and the existing tables. Furthermore, it was important to monitor in which processes a particular table was needed, either as input or output, and to know whether or not it would already be in the new database. This survey was also done for the auxiliary tables, which are tables that help us reach the final tables, the ones that both we and other teams use.

This survey was done starting by documenting which tables were needed in each process and identifying them at the beginning of each process. Normally, as the processes run, many of them at least, through a batch that starts a SQL script, this documentation was made at the beginning of each script. Before the start of this project, this documentation did not exist.

Then I used a .xlsx file again to be able to better visualize the work to be done and monitor the project's progress. I made a matrix where I could see which table and in which processes it was used, as well as how many processes resorted to it. Also, through some color codes, it was possible to monitor if the table was present as a synonym in the new database, if it was not yet, or if it was fully migrated.

The most challenging part of this migration project was the constant verification to check if the regular processes were running as intended, as well as the confirmation that the tables were up to date and with the correct data.

One of the positive points that this project brought to the team was the questioning of processes that perhaps were no longer useful and the discontinuation of the same and outdated processes. In addition to processes, it was possible to understand that tables were no longer necessary and that they no longer made sense to have.

As the machine that contained the data ended up being permanently shut down, we decided to take all the tables with us to the new database, and the ones that we considered expendable changed their name using a code, so that in the future, and if no one demonstrates the need for them, we can delete them, thus recovering useful and necessary space in the new database.

Another positive result of this project was the implementation of documentation at the beginning of each script used in the different processes, with their input and output tables (see Figure 4.2). There was already documentation with the changes made to each script, so as not to lose logic and rules that were implemented in the past, however, this insertion of more information is an asset, which saves us time and facilitates a quick understanding of what tables are used in a given process and also the origin of a table.

```
/*****************************************

    TABELAS INPUT:
        TABELA_INPUT_1
        TABELA_INPUT_2
        TABELA_INPUT_3

    TABELAS OUTPUT:
        TABELA_OUTPUT_1
        TABELA_OUTPUT_2
        TABELA_OUTPUT_3

*****************************************/
```

Figure 4.2 – Documentation's example of a script

My balance of the project was quite positive as I felt that we managed to clean up things that no longer made sense to make room for new challenges.

4.3. CUSTOMER AND STORE ANALYSIS

The project developed to understand the customer's choice came about when we were asked to find the closest store to each CP7 for a pilot project that would take place in the company. At the time we used the ArcGIS tool to make this calculation. Instead of calculating the distance in kilometers or meters, we calculated the temporal distance, since we considered that the temporal distance variable was more interesting for the client than the distance variable alone.

This development ended up being also included in the aforementioned table, the geomaster, for future consultation of teams that needed these new variables.

The pilot project for which this variable was necessary consisted in the possibility of exchanging equipment in store, by the customer, and therefore the interest arose in realizing if we could reach a lower level of granularity, that is, if instead of we were referring the customer to the store closest to their CP7 if we were able to define a store for each customer based on their characteristics.

This project was perhaps the most oriented project to the contents taught in the master's degree since the analysis was independent, and freedom was given to choose how to do it. This project was based on the methodology used in some machine learning projects, CRISP-DM, and therefore I will now divide it into the different phases of this type of methodology.

BUSINESS UNDERSTANDING

At this stage, it is intended that we objectively arrive at the question or questions that our client wants to be answered. In this case, the business objective is to associate a store with each customer, trying to understand if there are, in addition to distance, other factors that also contribute to this choice.

For this, it was necessary to work with data about customers and stores, as well as customer interaction in stores. Tools like Toad for Oracle and Jupyter Notebooks were used to develop this project.

As a data mining goal, we understand that it is the creation of a predictive model that allows us to associate a specific customer with a store in order to make their experience more personalized.

As data mining success criteria, we chose a combination of metrics used in classification problems, such as precision, recall, accuracy, and f1-score, which the closer to 1 the better. However, you need to be aware of overfitting.

It will then be a machine learning classification problem, more precisely multiclassification, where the predictable variable will be the store and the input variables will be customer variables.

The steps to take will be:

- Gather and organize data
- Conduct an exploratory analysis of the data
- Insights from exploratory analysis
- Prepare the data to be able to train the models, such as feature importance
- Apply classification models
- Understand the results obtained

One of the things to consider, as a possible challenge, is the fact that the models may not have a great meaning.

DATA UNDERSTANDING

We decided to work with customers who subscribe to a fixed service from the company, and visited stores in the year of 2021. These data are in the oracle database, and I used Toad for Oracle to develop the queries, as well as the construction of auxiliary tables.

When I started exploring the data, I realized that there are customers who have more than one service in different locations and therefore different coordinates, which raised the question of "which is the customer's permanent (or the one where he spends more time) home?". To also work with these customers, we decided to develop a rule so that, in terms of business, the location that most likely would be the customer's first home was chosen. The rule was to first compare the average billing values, choosing the service with the highest billing as the first home service; the next comparison was related to the pricing of the package in which the customer was included; in the case of ties, we moved on to the next criterion, which was the type of service that the customer had, giving priority to the Fiber service, etc., according to a scale that we consider to be from the best service to the worst (on the premise that we will have better service than in the house where we spend more time); Finally, we

confirmed that the service's postal code was the same as the account's postal code, and if so, that service would be the customer's main service.

We started with different records associated with our customers; and we managed, with the above criteria, to have one record per customer for the majority of the customers, that is, we were unable to find the service associated with the first home for less than 0.001% of our customers.

Regarding store records, we started with records of interactions in store in 2021, of which almost 85% are from the consumer segment, and of these 85%, only 67% of interactions contain information about customers in our database (i.e. 33% of these are records that we cannot associate with customer information.)

We continued the exploration and realized that there were records of interactions that had not occurred in the physical store but in the online store, we also discovered that there were customers for whom there was no location, despite having an associated fixed service, and therefore we did not consider them.

We also have information on the company's active stores, spread throughout Portugal, both mainland and islands.

We cross-referenced customer information with information from their interactions in stores and with information from stores, and from there we started to explore the data we had.

We started with 85 variables, related to the customer, their interaction in the store, and finally also related to the store where this interaction took place.

We can see, through the Figure 4.3, which represents the TOP5 of stores with more interactions from customers in the consumer segment and with fixed service or fixed internet service, that the store with the most interactions was the store in Aveiro in the Glicínias forum, with more than 26000 interactions.

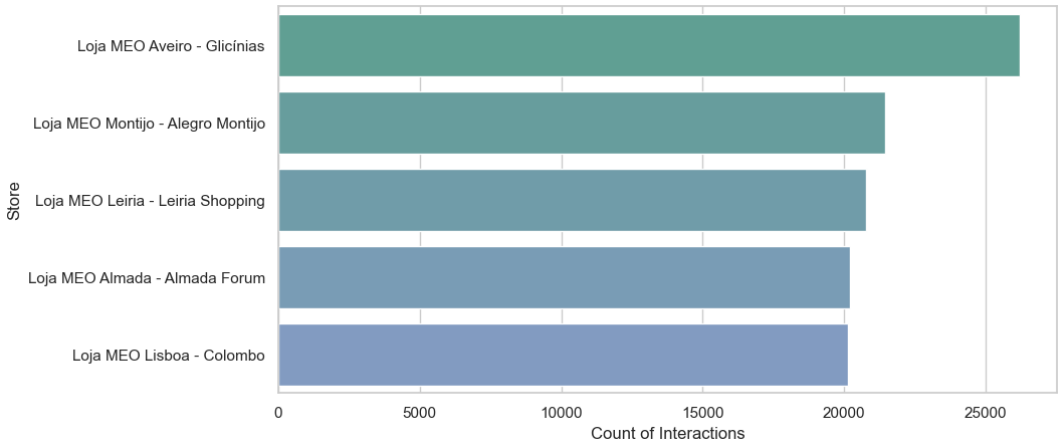


Figure 4.3 – TOP 5 Stores with the most interactions

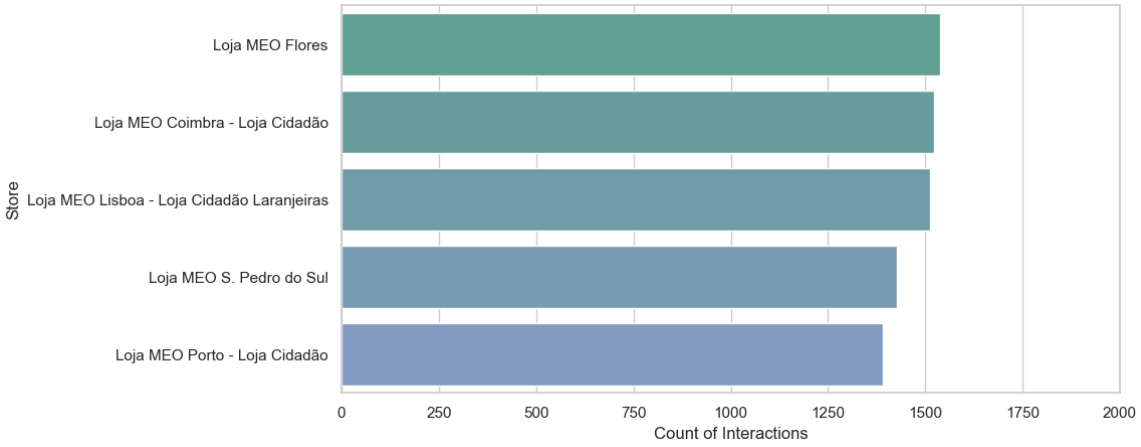


Figure 4.4 – LOWER 5 Stores with the least interactions

However, stores with fewer interactions, as we see in the graph in the Figure 4.4 (it has a shorter scale) have less than 1600 interactions, which makes us have a very unbalanced dataset since this will be the variable to consider as the target in our prediction model.

Still, within the theme of stores, we can see that there are 3 types of stores, those located in shopping centers, those that integrate spaces with Citizen's Store spaces, and street stores. We can see their distribution through the following images (Figure 4.5 and Figure 4.6).

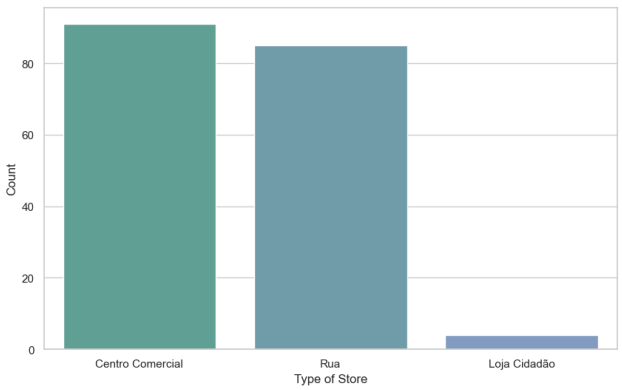


Figure 4.5 – Distribution of the Types of Stores

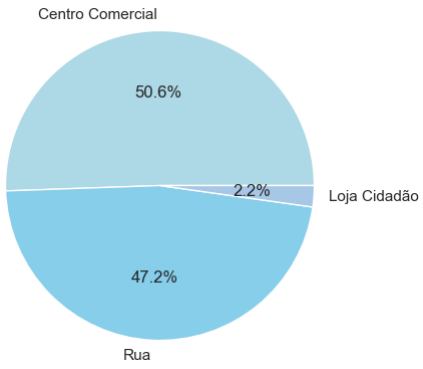


Figure 4.6 – Types of Stores in Relative Values

As for the districts (see Figure 4.7), Lisbon is the district with the most interactions, with almost 300000 interactions, followed by the Porto district, with almost 200000 interactions, values that we already expected.

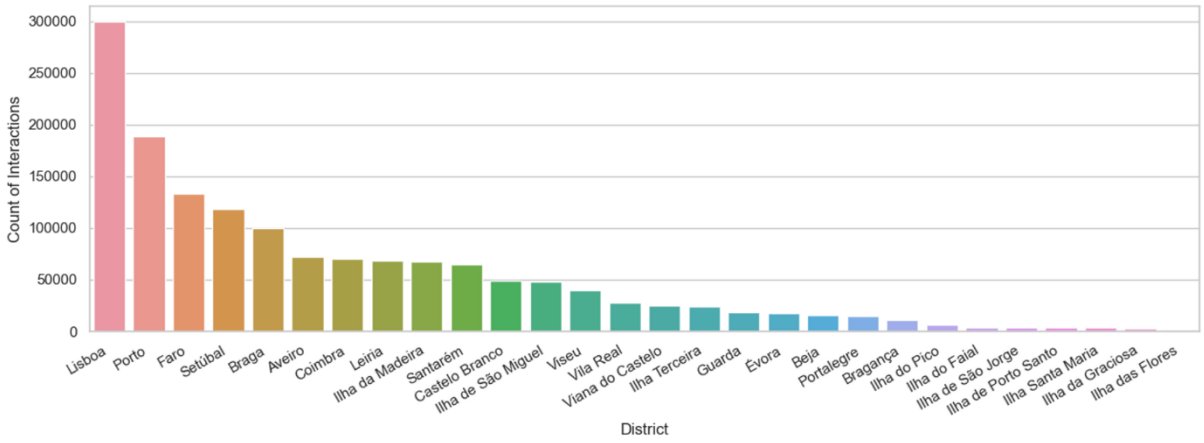


Figure 4.7 – Distribution of the interactions by District

In the variables related to customers, we found missing values in the variable year of birth in some observations; other observations did not have a postal code associated with the service, and others did not have basic service package information.

Regarding the type of service that our customers have, regarding customers who had interactions in the store, we can see (Figure 4.8) that the vast majority benefit from a Fiber service, and the rest are distributed among the other 3 types of service, ADSL (copper), Satellite, and Modular fiber.

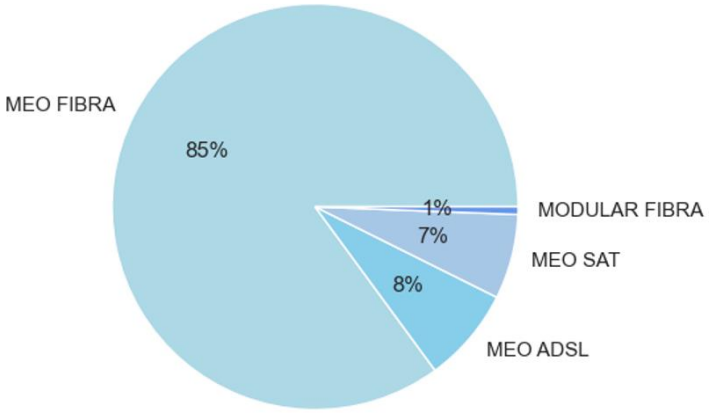


Figure 4.8 – Types of Service from customers that interacted in stores in Relative Values

DATA PREPARATION

In this phase is when we do some data cleaning, create new variables, see the correlations between variables and select the ones that will possibly give us the best results.

Regarding the missing values, the approach was, when considering the year of birth, we chose a year that was close to the mean of our data, but which is not the mean; therefore, for customers who did not have their year of birth, we associated the value 1972 in this variable. As for customers who did

not have CP7 associated with the service, we left them out, as the postal code is a variable with high cardinality and will not be used in the development of the models. However, it will be important for the comparison with the criterion currently used, which is the temporal distance between the CP7 centroid and the nearest store.

As for the variable related to the client service base package, we decided to assign the value 'Other' which will make sense when we work on this variable further.

We created new variables, such as the variable that indicates how many days have passed between the customer started to use our service, the initial variable data_ini_cli, and the first day of the year 2022, which is the variable permanence_days; another variable created was the variable that represents how many days passed between the customer has started to be our customer and the day of the interaction, this being the days_between_ini_cli variable; the age variable also appeared, calculated based on the year of birth. Variables were also built for a deeper exploratory analysis, but that cannot be used in the models, since they depend on the store and the interaction that occurs, they were the distance between the customer's coordinates and the store where the interaction took place, and the month of interaction.

An interesting finding was made when we analyzed the trip to the store of a customer who has been associated to the company for less than 6 months since their entry, where we can see that there are spikes in interactions when customers have multiples of 30 days between the beginning of being a customer and the interaction in the store (see appendix A); we removed customers who had an interaction in a store on the day they became customers, because, in terms of data visualization, it made the analysis of the remaining days almost imperceptible, since these customers account for a small portion of the dataset. There is also a decrease in frequency as this difference in days increases.

Regarding its values, we can easily verify the presence of outliers through the following boxplot (Figure 4.9) and even non-coherent values, since we have negative values for this variable, this is because we can have a customer who, before being our customer, had an interaction in one of our stores, but this information is not updated in the record of previous interactions from the moment the person becomes our customer.

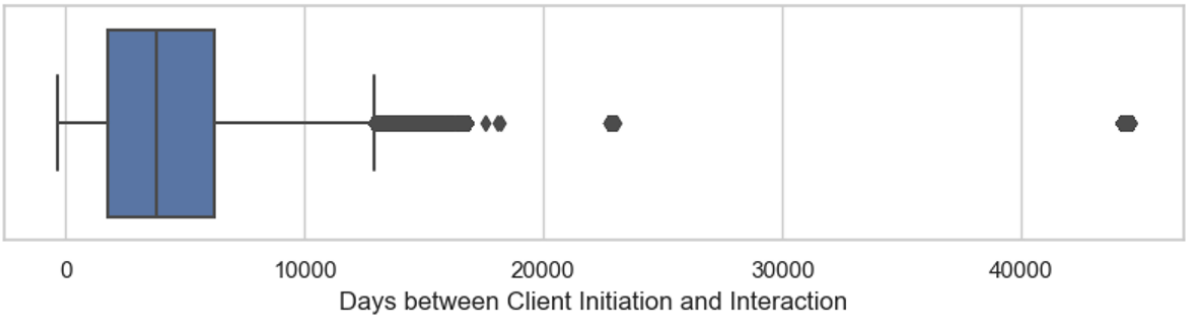


Figure 4.9 – Boxplot of the variable Days between Client Initiation and Interaction

As for the variable created to define the age of the client, we also obtained strange values, since it is not possible to have negative ages, which in this case are a consequence of values in the variable year of birth greater than 2022. We can have an idea of this variable through the next boxplot (Figure 4.10).

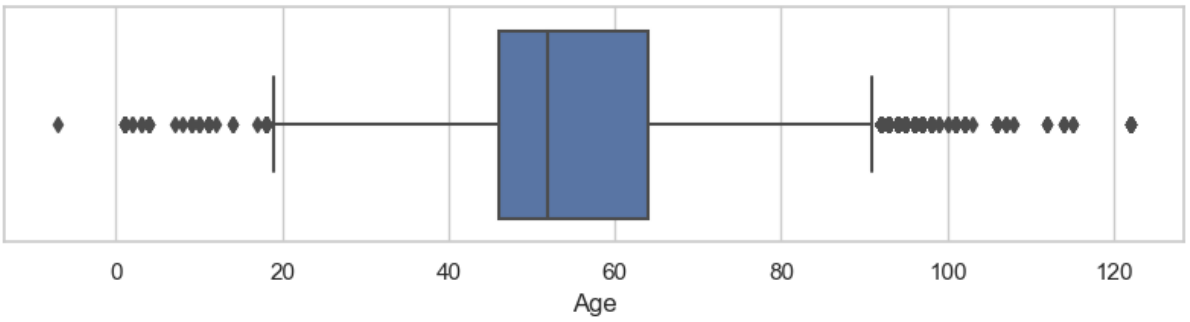


Figure 4.10 – Boxplot of the variable Age

As for the distance variable, using both the longitude and latitude of the customer's home and the store, and through the Haversine calculation, we can determine the distance between the customer and the store where they interacted. This distance calculation considers the spherical shape of the earth. In this variable, we also found values that we considered abnormal. To be able to see the distribution of the variable in a more meaningful way (Figure 4.11), we decided to apply a filter, just for visualization, that the distance between the store and the customer would be less than 50 kilometers, where more than 95% of our dataset is contained.

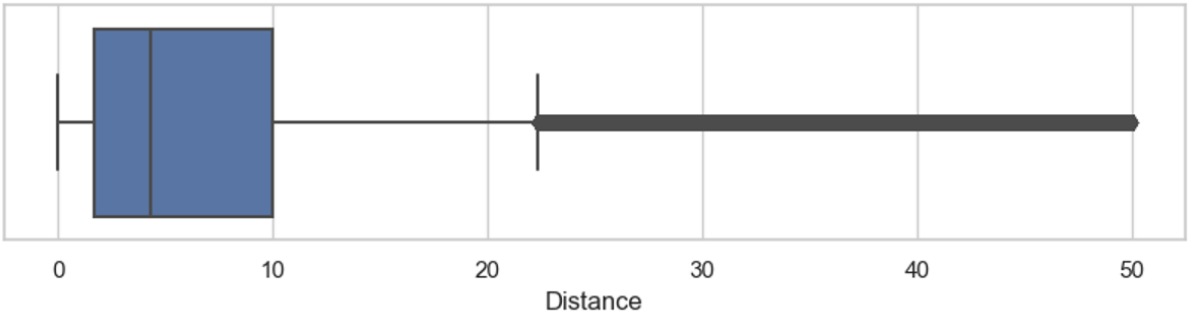


Figure 4.11 – Boxplot of the variable Distance (for values < 50 Km)

We created an interesting variable, which with the distance variable brings us funny insights. The MONTH_INTERACTION variable was created based on the interaction date and allowed us to see, for example, which months had the highest number of interactions. In conjunction with the information on the distance variable, we decided to try to understand in which months the highest number of interactions occurred, when we knew that the distance between the customer and the store was greater than 100 kilometers. The results can be seen in the following image in Figure 4.12.

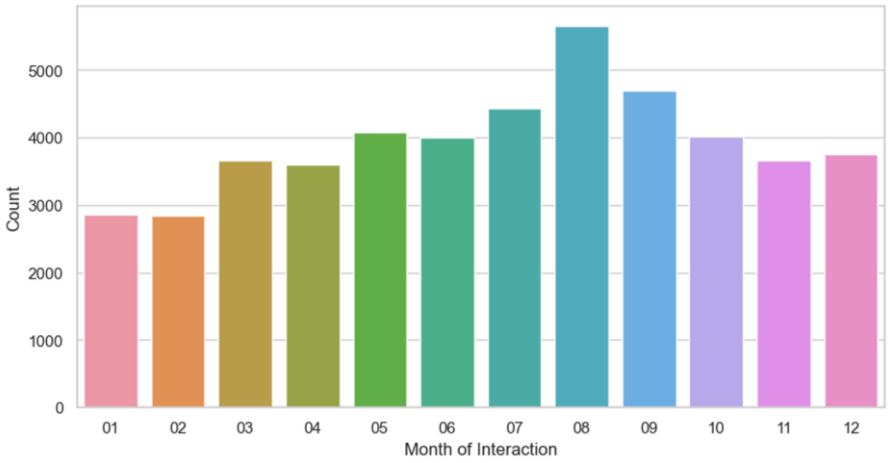


Figure 4.12 – Distribution of Customer Interactions that are more than 50 Km away from the Store where they interacted, per Month

We can therefore observe that the month with the highest number of interactions when the distance from the customer to the store is greater than 100 km is the month of August, which is the month of choice, in terms of the Portuguese national panorama, for the inhabitants to enjoy their holidays. The fact that the customer is on vacation often means being away from their home, where they spend more time, and hence this value, for the month of August, makes sense.

It should be noted that some of these variables will not be considered in the construction of the model since they are variables that we do not know their value until the customer has interaction in the store. The variables month of interaction, days between customer initiation and interaction in the store, and distance from the customer to the store was created with the aim of exploring the data, to better understand them and find anomalous observations so that we can treat or eliminate them to reduce the noise that enters our model.

We also created a variable that allows us to count how long the client has been with us, in days, which has the following distribution (Figure 4.13). It should be noted that what we might think is just one outlier with over 40,000 days, which is reflected in around 109 years, seemed strange to us and therefore we decided not to consider customers who, according to this variable, were enjoying of company services for more than 30,000 days, which even so amounts to more or less 82 years.

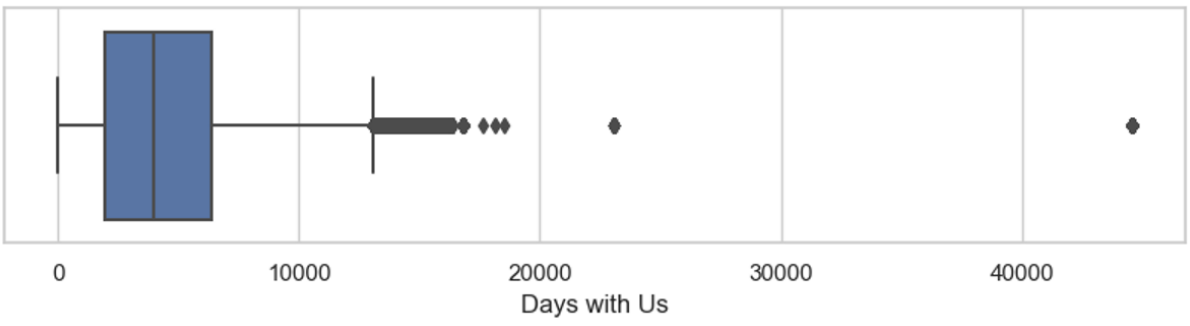
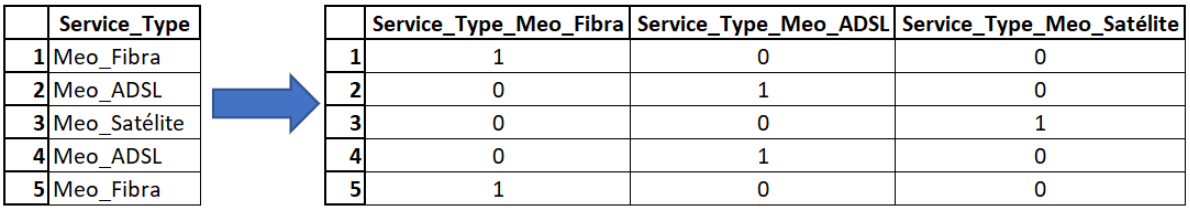


Figure 4.13 – Boxplot of the variable Days with Us

According to our data exploration, we decided to eliminate observations that seemed to us to be outliers, that is, values outside the normal range. However, we did not apply any technique, such as the interquartile technique, deciding only to eliminate observations based on the distribution and what we learned about the data. For this reason, and in accordance with the company's business, we did not consider interactions with customers who were under the age of 18; we excluded interactions that demonstrated a distance between the customer and the store greater than 900 km, since that seemed to be an acceptable distance for someone residing in Portugal and frequenting one of our stores; and we also did not consider interactions with customers who have been with us for more than 30,000 days.

With these filters, we eliminated less than 1% of our dataset.

We decided to treat some variables, transforming values of the same ones that were not so frequent to a value common to all and we also encoded some variables, passing them from being in a column to starting to be included in several, so that the models and other techniques can have the best performance. The type of encoding used was the One-Hot-Encoder which transforms a variable present in a column with, for example, 5 possible values, into 5 different columns with a binary value as shown in the following image (Figure 4.14).



	Service_Type
1	Meo_Fibra
2	Meo_ADSL
3	Meo_Satélite
4	Meo_ADSL
5	Meo_Fibra

	Service_Type_Meo_Fibra	Service_Type_Meo_ADSL	Service_Type_Meo_Satélite
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0
5	1	0	0

Figure 4.14 – Example of One-Hot-Encoder technique

Through this technique, we were able to advance to the variable selection phase.

We can make the selection of variables, trying to understand what impact they can have on the target variable, as we can also see if certain variables are giving us similar information.

We started by looking at the correlations that existed between the variables through the Pearson and Spearman correlation coefficients, to eliminate redundancies, that is, not having two variables giving almost the same information to the model. With this method, we were aware of these variables, but still without eliminating them. We took notes and let's see how they behave in other types of techniques.

One of the techniques used was the Mutual information method, where we were able to perceive which variables are more important for the model when compared to the target variable. We can highlight that in our case, the variables linked to the location of our client have a great weight, a fact that we were already expecting. In the following image (Figure 4.15), we can see the importance of the different metric variables in the target variable, which is the store.

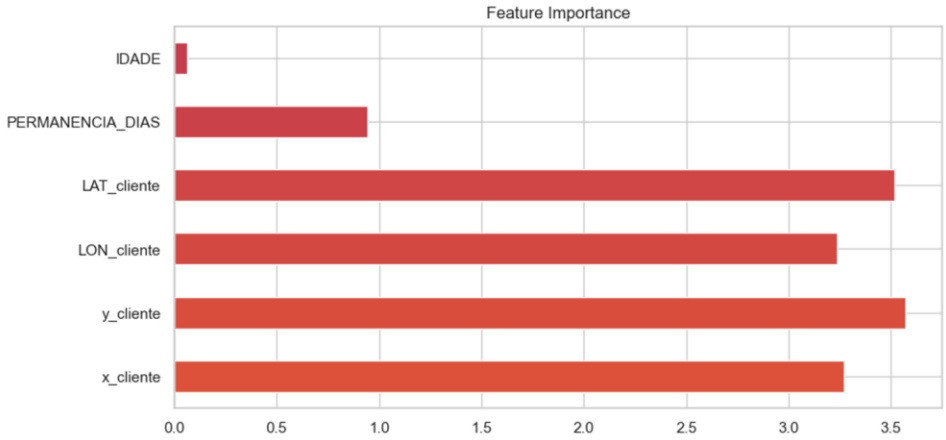


Figure 4.15 – Feature Importance of Metric Variables in Target Variable

Another method used was the use of the attribute of the importance of features that the Random Forest classification model offers, for this, a model was applied using 100 estimators of decision trees, each of which could grow in a maximum of 5 measures, that is, only 5 columns at most are used in each of the decision trees and with a separation criterion of 75, that is, to advance to the next variable analysis the node must host at least 75 observations. The results of this method are in the image below (Figure 4.16).

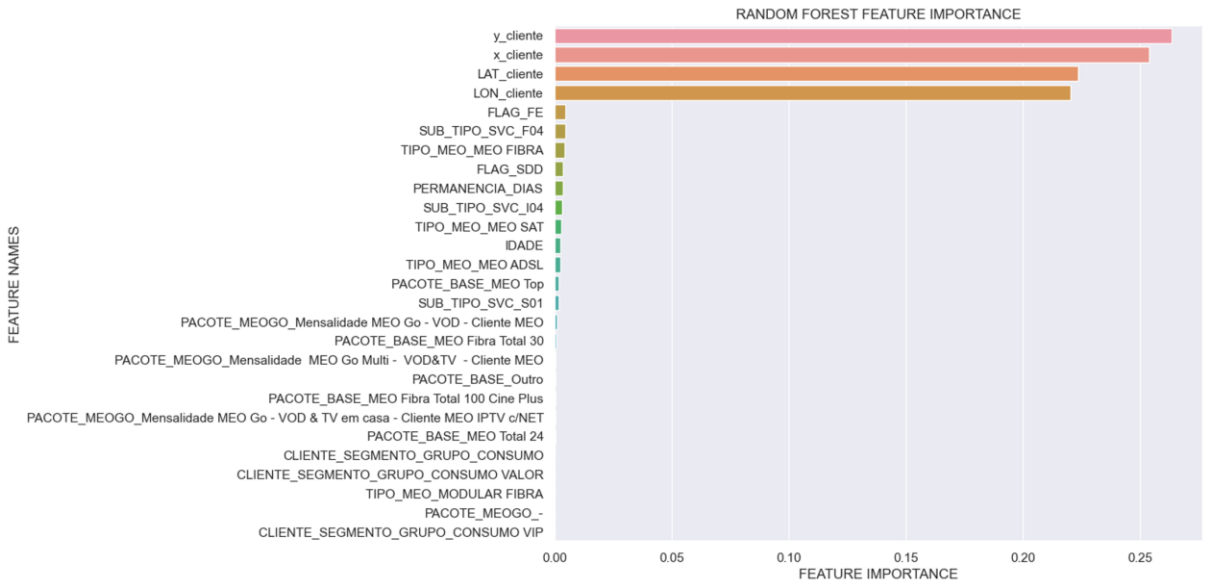


Figure 4.16 – Feature Importance using Random Forest model

Once again, the customer location variables occupy a place on the podium of importance in the target variable.

We also used the same attribute as in the Random Forest, but for the Gradient Boosting model, which we combined with the previous analyses.

So, we decided to have several variable options to train our models. Below, in Table 4.1, we have one of the selections of variables used, as well as its description.

Variable	Description	Type
FLAG_SDD	variable that indicates whether the customer uses direct debit	Binary (1 -True / 0 - False
FLAG_FE	variable that indicates whether the customer uses electronic invoice	Binary (1 -True / 0 - False
LON_CLIENTE	longitude of customer's first home	Float
LAT_CLIENTE	latitude of the client's first home	Float
PERMANENCIA_DIAS	number of days that the customer has service from the company	Int
IDADE	customer's age	Int
TIPO_MEO_MEO_FIBRA	variable that indicates whether the customer has fiber service	Binary (1 -True / 0 - False
TIPO_MEO_MEO_SAT	variable that indicates whether the customer has satellite service	Binary (1 -True / 0 - False
TIPO_MEO_MEO_ADSL	variable that indicates whether the customer has ADSL service	Binary (1 -True / 0 - False
PACOTE_BASE_MEO Top	variable that indicates whether the customer has a Meo Top base package	Binary (1 -True / 0 - False
PACOTE_MEOGO_-	variable that indicates whether the customer has a Meo Go package named	Binary (1 -True / 0 - False

Table 4.1 – Some of the selected Variables

With the study of the variables and their selection, we were able to move on to the part of training the models.

MODELING

At this stage, we chose to divide the dataset in a proportion of 70% to train the model and 30% to evaluate its performance since our objective is that the model comes as close as possible to reality, but we do not want the model to justify too much. our data well and for unseen data, it performs poorly. The 30% of data that is left out is for us to have that safeguard and to evaluate our model more objectively.

Several models were applied, such as:

- Multinomial Logistic Regression with all variables, using the 'saga' solver (A)
- Multinomial Logistic Regression with a subset of variables, using the 'saga' solver (B)
- Multinomial Logistic Regression with only location variables, using the 'saga' solver (C)
- Decision Tree with default values with a subset of variables (D)
- Decision Tree with default values with all variables (E)
- Decision Tree with a subset of variables, using the ‘Gini’ criterion, the minimum number of observations per leaf and minimum value for separation of 50 observations, and choosing the best splitter, which allows the decision tree to choose the variable with greater importance to do the split (F)
- Random Forest with all the variables, using 50 decision trees, the 'entropy' split criterion, which chooses the variables that have less entropy as the variable to split, and with the minimum value for separation of 75 observations (G)
- Random Forest with a subset of the variables, and with the same characteristics as the model above (H)
- Extreme Gradient Boosting with a subset of variables, using 50 decision trees and a learning rate of 0.1 (I)

- Extreme Gradient Boosting with all variables, using 50 decision trees and a learning rate of 0.1 (J)
- Light Gradient Boosting Machine with all variables, with 50 decision trees, with the multiclass objective that allows you to make one class vs the others, a L2 regularization of 0.1, to prevent overfitting (K)
- Light Gradient Boosting Machine with all the variables and the previous parameters, with the addition of the boosting type based on DART (Dropouts meet Multiple Additive Regression Trees) to increase the accuracy and even with a number of leaves, final nodes, set to 100 (L)
- Light Gradient Boosting Machine with the same parameters as the model above, but using only a subset of variables (M)
- Light Gradient Boosting Machine same as above model, however without L2 regularization of 0.1 to avoid overfitting (N)

EVALUATION

Arriving at the model evaluation phase, we must evaluate the metrics obtained for each of the trained models, to know which one best approximates reality. As we are facing a multiclassification problem in which we are facing an imbalanced dataset, we must look at the previously mentioned metrics, such as recall, precision, and f1-score, but for the weighted values.

These metrics are obtained by comparing the output of each model when applied to the subset of observations (the 30% of data) that were left out when we trained the models. This allows us to ensure that we are not overfitting since the model is being applied to data that it has not yet seen.

To have an easier way of comparing the results of the different models, we built the following graph, in Figure 4.17, with the F1-Score weighted metric obtained in each one.

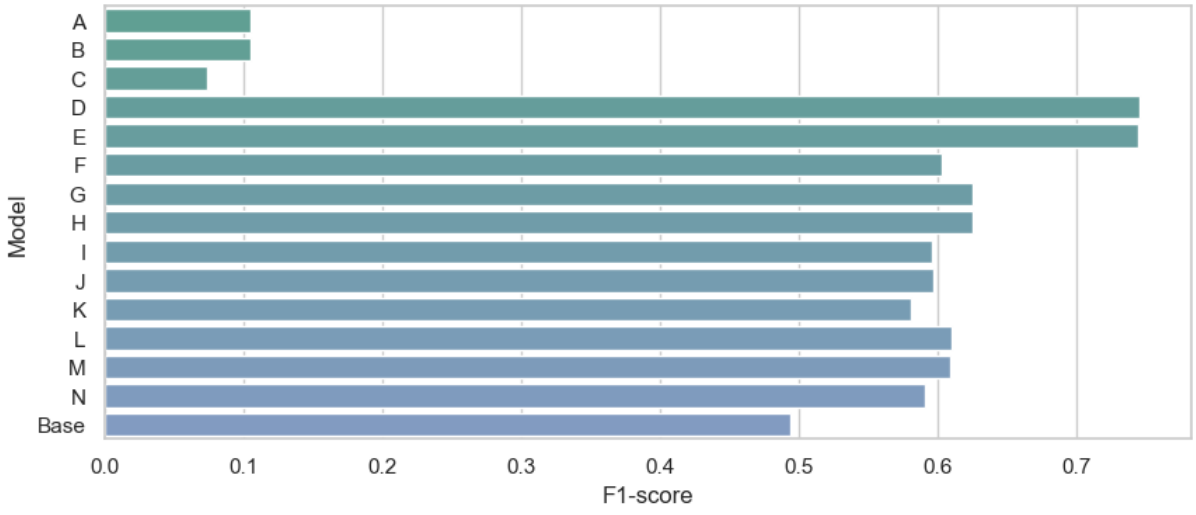


Figure 4.17 – Performance of the models based on the F1-score metric

As we can see, all trained models, except the logistic regression models, performed better in terms of predicting our client's store, compared to the base model, which is currently used. This base model only considered the postal code where the customer is located and was based only on minimizing the temporal distance between the customer and the various store options. The models with the best performances were the Decision Trees with default values, and both the model that used all the variables and the model that used only a subset of them, obtained good results.

Comparing the models with the best performance with our base model, using other weighted metrics (Table 4.2), we can see that the results are better in the developed models.

Model	Weighted Precision	Weighted Recall	Weighted F1-score
D	0.745	0.745	0.745
E	0.745	0.745	0.744
Base	0.551	0.511	0.494

Table 4.2 – Comparison between the models with the best performance and base model

To better try to understand one of the decision trees regarding the weights assigned to each variable when building it, I used an attribute of it. When we train a model of decision trees in the python environment, this one in an attribute, as well as, for example, the Random Forest and the Gradient Boosting, which is called feature_importances_ that allows the user to have a notion of which variables were most important in the construction and training of the model in question. This is what I did when I arrived at the results above (Figure 4.17 and Table 4.2). For a better understanding, we have below, in Figure 4.18, a representation of these amounts.

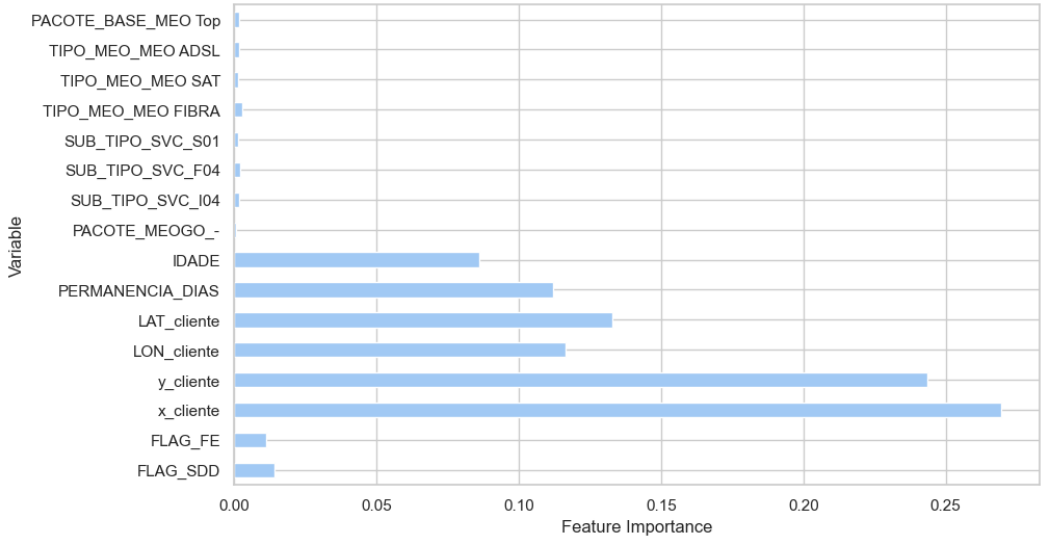


Figure 4.18 – Feature Importance of some Variables in Model E

As before, in the exploration phase (Figure 3.17 and Figure 3.18), also in the training of this model, the variables related to location are the ones that have the greatest weight. However, we see that other

variables, such as the number of days since the customer has had a service from the company, also play a role in the choices made in this model.

DEPLOYMENT

According to the methodology used to approach this project, the next step will be the deployment, in which we deliver the final product to the client and develop tools that help in the functioning and maintenance of the created project.

This project has not reached this stage yet, as we are still analyzing how best we can integrate this new information into the information we already have. It will perhaps be useful information to have when we analyze our clients, what are their preferences, among others.

I think that the developed project brought a new opportunity to integrate useful information into the remaining projects and also in the processes that we have already implemented.

The results obtained, as we can see in Figure 4.2, were positive, since the performance of most of the models created was superior to the performance of the current model. I feel that it was a project in which I was able to apply methodologies and concepts learned, in a real situation and with a possible future positive impact.

It was possible to understand our customers a little better, not only in terms of the case under study, and the visits to stores, but I also had the opportunity to extract some curious insights that previously would have been just my suspicions but which the data proved.

5. CONCLUSIONS

This report describes some of the tasks assigned to me during my internship, which lasted one year, at Altice Portugal. Several projects were proposed, of which I decided to present 3 in this internship report since they were challenging and contributed to my development as a data scientist. In addition to the projects, I was assigned recurring tasks, some of which are described in the project's chapter.

Regarding the Geomaster project, the department was satisfied with the results and the ease of access to information. There is a continuous improvement of the information that exists in this table, and everyone's desire is to make it a table rich in useful information and transversal to all teams. As mentioned before, this table is updated daily, and sometimes it is necessary to review it when the process does not run completely due mainly to the unavailability of other tables or code errors when developing a new version of the script, which is bad. It is a project in constant improvement.

As for the database migration project, we concluded the first phase of data integration from the old database to the new one. However, one day awaits us, a review and subsequent deletion of tables and information that is no longer useful or updated.

In the analysis of the customers' choice of store, the results were interesting, since we were able to achieve a better performance than the simple model that we have now applied for assigning a store to each customer. I confess that I was not expecting the results obtained, but I am happy to have managed to carry out an analysis that takes our analysis further. However, I think that, in the future, it would be interesting to carry out this analysis for districts or localities with a larger population and a greater choice of stores, such as Porto and Lisbon, which have considerably more stores available to their customers. And it may be interesting to see, for example, the age distribution of those who frequent stores located in shopping centers.

5.1. CONNECTION TO THE MASTER PROGRAM

The Master in Data Science and Advanced Analytics at Nova IMS gives us a lot of tools to have a critical analysis of the different projects to which we contribute. It is a master's degree where we work with data, in different ways and with different tools, which allows us to be well-prepared for future challenges. One of the main competencies that I acquired, and that the master's degree encourages us to develop is knowing how to look and what to look for.

Being part of a team that is in contact with various topics, which supports the department, combined with my specialization in Business Analytics, allowed me to develop my business orientation.

Regarding the curricular units that I considered to have better prepared me for this internship, they were Storing and Retrieving Data, Data Visualization, Data Mining, Machine Learning, and Business Cases. This last curricular unit helped a lot in understanding the need to comprehend the business for which we develop solutions and analyses.

5.2. INTERNSHIP EVALUATION

The fact that the internship took place in one of the most renowned companies in Portugal allowed me to get in touch with different realities and different people; to know the market and what position the company occupies, not only in terms of services but also in terms of social awareness, among others.

It was a challenging internship as it was my first contact with the world of work, where I discovered and continue to discover what kind of professional I am and want to be. It allowed me to learn new work methods and realize that there is no one with the right answers.

The team I integrated is multidisciplinary and, therefore, is in contact with different topics, which was sometimes a little overwhelming. This is because, anyone who joins a company, they are not used to the type of language, the acronyms used, or the keywords that have a specific concept within the company.

However, the whole experience was very enriching both professionally and personally.

It allowed me to get in touch with different tools than the ones I learned and made me more responsible, without losing the sense that I also make mistakes and that mistakes are reversible or can be solved.

5.3. LIMITATIONS AND FUTURE WORK

Regarding the difficulties found in the work carried out, the main one was time, especially in the analysis project of the stores and customers, this because the gathering of data and its exploitation consumes a lot of time, and when I started to understand the problem and stay on top of the project, the internship time was ending and with it the need to deliver this internship report work. This does not preclude the continuation of this particular project, as the insights gleaned from it were interesting. As previously mentioned, it is in my interest to carry out a more concrete analysis, using, for example, the Lisbon and Porto districts. The time variable also led to the fact that more models were not tried, or even the same models, but with different chosen variables; this being a possible future improvement; and also, the use of the grid-search method for each model and with different variables, is a future step in this project. Another challenge in this project was having an imbalanced dataset; At the time, I questioned within the company if they usually considered any technique that would allow this characteristic to be circumvented, and I realized that they prefer to leave the data as they come regarding this imbalance. So, another possible improvement is the application of methods such as under-sampling or oversampling to see what results we can get.

In the project of the reference table, I will continue to monitor the same table, adding whenever necessary and justifiable, the variables useful to the different teams.

For the database migration project, the team's objective is to abandon in the future some data structures that are no longer used by our team or by other teams, both within the department and

outside it. To this end, over time we will review and assess the need to maintain these structures or not.

As for recurring tasks, it is in my interest to invest in automating them, to gain time to embrace other challenges within the company.

REFERENCES

Anaconda distribution. (n.d.). Anaconda. https://www.anaconda.com/products/distribution

Ng, A. (2000). CS229 Lecture notes. *CS229 Lecture notes*, *1*(1), 1-3.

Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons.

Best Oracle developer and administrator database tools | Free trial. (n.d.). Quest | IT Management | Mitigate Risk | Accelerate Results. https://www.quest.com/products/toad-for-oracle/

Bhardwaj, A. (2020, June 14). *Calculating distance between two Geolocations in Python*. Medium. https://towardsdatascience.com/calculating-distance-between-two-geolocations-in-python-26ad3afe287b

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

Bleier, A., & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, *34*(5), 669-688. https://doi.org/10.1287/mksc.2015.0930

Brownlee, J. (2020, August 31). *Multinomial logistic regression with Python*. MachineLearningMastery.com. https://machinelearningmastery.com/multinomial-logistic-regression-with-python/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, *9*(13), 1-73.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Database. (n.d.). Oracle | Cloud Applications and Cloud Platform. https://www.oracle.com/database/index.html

Patil, T. H. D. J., & Davenport, T. (2012). Data scientist: The sexiest job of the 21st century. *Harvard business review*, *90*(10), 70-76. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*. https://doi.org/10.48550/arXiv.1810.11363

ESRI. (2022, May 11). *What is GIS?*. GIS Mapping Software, Location Intelligence & Spatial Analytics | Esri. https://www.esri.com/en-us/what-is-gis/overview

ESRI. (2022, July 25). *Download ArcGIS pro*. https://pro.arcgis.com/en/pro-app/latest/get-started/download-arcgis-pro.htm

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232. https://doi.org/10.1214/aos/1013203451

Hand, D. J., & Adams, N. M. (2015*). Data mining in Wiley StatsRef: Statistics Reference Online* (pp. 1-7). Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781118445112.stat06466.pub2

Hess, R. L., Rubin, R. S., & West, L. A. (2004). Geographic information systems as a marketing information system technology. *Decision Support Systems*, *38*(2), 197-212. https://doi.org/10.1016/S0167-9236(03)00102-7

James, G. M., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in R*. Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30.*

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324. https://doi.org/10.1016/s0004-3702(97)00043-x

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

Kumar, V. (2014). Feature selection: A literature review. The Smart Computing Review 4 (3). 10.6029/smartcr.2014.03.007.

Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14-23. https://doi.org/10.1002/widm.8

McCue, C. (2015). Identification, characterization, and modeling. *Data Mining and Predictive Analysis*, 137-155. Butterworth-Heinemann.https://doi.org/10.1016/b978-0-12-800229-2.00007-9

Mitchell, T. M. (1997). *Machine learning* (Vol. 1, No. 9). New York: McGraw-hill.

Nagpal, A. (2017). Decision tree ensembles-bagging and boosting. *Towards Data Science*. https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9#:~:text=The%20main%20principle%20behind%20the,to%20form%20a%20strong%20learner.&text=Bagging%20(Bootstrap%20Aggregation)%20is%20used,sample%20chosen%20randomly%20with%20replacement

Narkhede, S. (2018). Understanding confusion matrix. *Towards Data Science*, *180*(1), 1-12. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Nasirin, S., & Birks, D. F. (2003). DSS implementation in the UK retail organisations: A GIS perspective. *Information & Management*, *40*(4), 325-336. https://doi.org/10.1016/s0378-7206(02)00015-0

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS B*, *4*, 51-62. https://doi.org/10.20544/horizons.b.04.1.17.p05

Pedregosa (et al. 2011). Scikit-Learn: Machine Learning in Python. the Journal of machine Learning research, 12, 2825–2830.

Portugal, P. (n.d.). *Altice Portugal*. https://www.telecom.pt/en-us/a-pt/Pages/historia.aspx

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106. https://doi.org/10.1007/bf00116251

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach*.

Reichheld, F. F., & Sasser, W. E. (1990). Zero Defections: Quality Comes to Services. *Harvard Business Review, 68*(5), 105-111, Sept-Oct, 1990. Retrieved from https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Shamoo, A. E., & Resnik, D. B. (2009). Responsible conduct of research: Oxford University Press. *New York*.

Sodhi, P., Awasthi, N., & Sharma, V. (2019). Introduction to machine learning and its basic application in python. In *Proceedings of 10th International Conference on Digital Strategies for Organizational Success*. https://doi.org/10.2139/ssrn.3323796

Talabis, M., McPherson, R., Miyamoto, I., & Martin, J. (2014). *Information security analytics: Finding security insights, patterns, and anomalies in big data*. Syngress.

Venkat, N. (2018). The curse of dimensionality: Inside out. *Pilani (IN): Birla Institute of Technology and Science, Pilani, Department of Computer Science and Information Systems*. http://doi.org/10.13140/RG.2.2.29631.36006

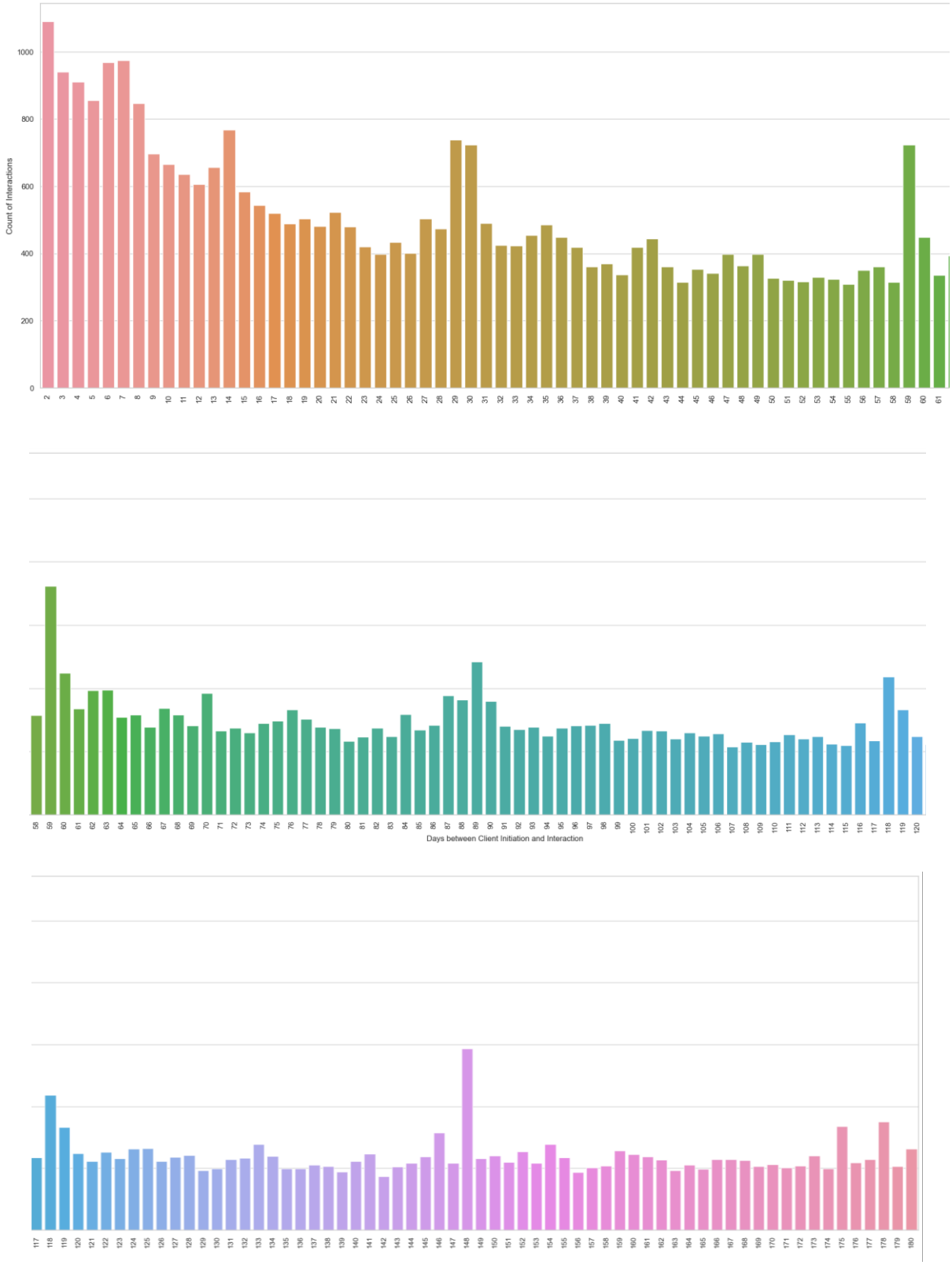
Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223-230. https://doi.org/10.1016/j.eswa.2010.06.048

Witten, I. H., Frank, E., & Hall, M. A. (2011*). Data mining: Practical machine learning tools and techniques.* Elsevier.

Zhu, Y., & Xiong, Y. (2015). Towards data science. *Data Science Journal*, *14*. https://doi.org/10.5334/dsj-2015-008

APPENDIX

A. Days between Client Initiation and Interaction



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

Universidade Nova de Lisboa