NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

CUSTOMER CHURN PREDICTION IN PORTUGUESE BANKING SECTOR

Using a Machine Learning Approach

Inês Tomás Pires

Master Thesis

presented as partial requirement for obtaining the Master Degree Program in Information Management

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 CHURN PREDICTION IN PORTUGUESE BANKING SECTOR

Using a Machine Learning Approach

by

Inês Tomás Pires

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervised by

Miguel de Castro Simões Ferreira Neto, PhD, NOVA Information Management School

João Bruno Morais de Sousa Jardim, PhD, NOVA Information Management School

December, 2023

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 acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Inês Tomás Pires

Lisbon, 4th December 2023

ACKNOWLEDGEMENTS

First, I would like to show my gratitude to my supervisors, professors Bruno Jardim and Miguel Neto, for all the support and comprehension in the development of this work and for believing in the potential of this study.

Second, I want to thank José Coelho and João Gomes, the JJ's, for being my life saviors and helping me with the dilemmas throughout this dissertation. It wouldn't have been possible without their help and a little bit of their fun.

I also want to give special thanks to Leandro Gonçalves for showing me his passion for data and for all the knowledge I gather with him every day. His passion became mine too.

Last but (definitely) not least, to Patrícia Ferreira, for dealing with all my doubts and insecurities and for believing in me. Thank you for always being there and for being my shelter.

ABSTRACT

This study focuses on developing a predictive model for customer churn in a Portuguese bank, using machine learning techniques. Following the CRISP-DM methodology, the analysis encompasses comprehensive EDA, data preparation and visualizations, laying the foundation for model selection. Whitin the subset of evaluated models, such as tree-based and ensembled models, Gradient Boosting emerges as a standout performer, demonstrating notable predictive capabilities. Beyond the identification of customers at risk to churn, this model provides valuable insights, crafting proactive retention strategies. The precision in identifying customers with a high probability of churn enhances informed decision-making. For that reason, an interactive dashboard is developed to empower stakeholders in addressing potential churn risks. These findings underscore the importance of leveraging machine learning in banking scenarios, emphasizing the potential for predictive analytics to enhance customer retention strategies and overall business outcomes.

KEYWORDS

Banking Sector; Business Intelligence; Customer Churn; Machine Learning; Power BI

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

Statement of Integrity i
Acknowledgements ii
Abstract iii
List of Figures vi
List of Tables viii
List of Abbreviations and Acronyms ix
1. Introduction 1
1.1. Background and Problem Identification 1
1.2. Study Objectives 1
1.3. Study Relevance and Importance 2
2. Literature review 3
2.1. Customer Churn Definition 3
2.1.1. Churn in banking sector 4
2.1.2. Case studies in banking sector 4
2.2. Predictive Models 5
2.2.1. Machine Learning Models in Customer Churn 6
3. Methodology 8
3.1. CRISP-DM 8
3.1.1. Business Understanding 9
3.1.2. Data Understanding 10
3.1.3. Data Preparation 18
3.1.3.1. Duplicated values 19
3.1.3.2. Missing values 19
3.1.3.3. Outliers treatment 19
3.1.3.3.1. Z-Score 20
3.1.3.3.2. IQR 20
3.1.3.4. Feature selection 22
3.1.3.5. Encoding 23
3.1.3.6. Data normalization 24
3.1.4. Modelling 25
3.1.5. Evaluation 26
3.1.6. Deployment 27

4. Results and Discussion 28
4.1. Machine Learning Models 28
4.1.1. Decision Trees 28
4.1.2. Extra Trees 29
4.1.3. Random Forest 30
4.1.4. XGBoost 30
4.1.5. Gradient Boosting 31
4.2. Discussion 32
4.3. Power BI Dashboards 33
5. Conclusions and Future Works 36
5.1. Conclusions 36
5.2. Limitations 36
5.3. Future Works 37
Bibliographical References 38

LIST OF FIGURES

Figure 2.1 – Graph representing the word “Churn” between Dec. 2013 and Dec. 2023 (source: Google Trends) 3
Figure 3.1 – Phases of the CRISP-DM reference model (Chapman et al., 2000) 9
Figure 3.2 – Timeline regarding train and test datasets 10
Figure 3.3 – Churn rate 12
Figure 3.4 – Churn by Gender 13
Figure 3.5 – Churn by Opening Channel 13
Figure 3.6 – Churn by Plan 13
Figure 3.7 – Churn by Blocked Account 13
Figure 3.8 – Churn by Blocked Client 14
Figure 3.9 – Churn by Cross-Selling 14
Figure 3.10 – Churn by Debit Card 14
Figure 3.11 – Churn by Credit Card 14
Figure 3.12 – Churn by Warranty 15
Figure 3.13 – Boxplot and Histograms’ Age variable 16
Figure 3.14 – Boxplot and Histograms’ Tenure variable 16
Figure 3.15 – Boxplot and Histograms’ Blocked variable 16
Figure 3.16 – Boxplot and Histograms’ Cross_Selling variable 16
Figure 3.17 – Boxplot and Histograms’ Last_Transaction variable 17
Figure 3.18 – Boxplot and Histograms’ Balance_DO variable 17
Figure 3.19 – Boxplot and Histograms’ Balance_DP variable 17
Figure 3.20 – Boxplot and Histograms’ Balance_Credit variable 17
Figure 3.21 – Boxplot and Histograms’ Outstanding_Balance variable 18
Figure 3.22 – Boxplot and Histograms’ Average_Balance_6M variable 18
Figure 3.23 – Boxplot and Histograms’ Average_Transactions_6M variable 18
Figure 3.24 – Boxplot and Histograms’ Age variable after data preparation 21
Figure 3.25 – Boxplot and Histograms’ Tenure variable after data preparation 21
Figure 3.26 – Boxplot and Histograms’ Cross_Selling variable after data preparation 21
Figure 3.27 – Boxplot and Histograms’ Balance_DP variable after data preparation 21
Figure 3.28 – Boxplot and Histograms’ Balance_Credit variable after data preparation 22
Figure 3.29 – Boxplot and Histograms’ Outstanding_Balance variable after data preparation 22
Figure 3.30 – Boxplot and Histograms’ Average_Transactions_6M variable after data preparation 22
Figure 3.31 – Spearman heatmap between numerical variables 23

Figure 4.1 – Decision Tree’s AUC and Confusion Matrix 29
Figure 4.2 – Extra Trees’ AUC and Confusion Matrix 29
Figure 4.3 – Random Forest AUC and Confusion Matrix 30
Figure 4.4 – XGBoost’s AUC and Confusion Matrix 31
Figure 4.5 – Gradient Boosting’s AUC and Confusion Matrix 32
Figure 4.6 – Overview’s page of the Power BI dashboard 34
Figure 4.7 – Churn Analysis' page of the Power BI dashboard 35

LIST OF TABLES

Table 3.1 – Dataset's features summary..... 10
Table 3.2 – Descriptive statistics of numerical variables 24
Table 3.3 – Descriptive statistics of categorical variables..... 24
Table 4.1 – Model's scores summary table..... 28

LIST OF ABBREVIATIONS AND ACRONYMS

AUC	Area Under Curve
BI	Business Intelligence
CRISP-DM	Cross Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
FN	False Negatives
FP	False Positives
IQR	InterQuartile Range
SMOTE	Synthetic Minority Over-sampling Technique
TN	True Negatives
TP	True Positives
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

The way people communicate, and access information nowadays is quite different from what it was 20 years ago, and this is changing every day at a rapid pace. The digital era changed how fast we receive information daily, which has allowed companies to take advantage of it for the benefit of their business and has brought with it more and more competitiveness and competition between industries.

With world globalization, it is currently possible to do practically everything through a computer or a smartphone, and banking sector is an example of this. The digital transformation brought industries new ways of doing business. It is increasingly common to find banks that adapt to this transformation, and this goes through home banking: it is possible to check balances and movements; manage credit and debit cards; create time deposits; make payments and credit simulations and so on. This is making this industry more and more digital.

However, it becomes more difficult for banks to retain their clients. Although it is related to a lot of factors – lack of personalization, poor customer service, and fees and charges, it is also related to the fact that digital banking made it easier for customers to switch bank. And more important than gaining new customers, is it more importantly to retain the old ones.

Customer churn is a concept that is well known between businesses and is related to the loss of customers. In align with this, it is understandable that a high percentage of churn rate is bad for business, and it is important to understand what is causing that effect and what can be done to avoid this.

Taking this into consideration, banks need to find new ways to make their business grow, and this study will focus on that.

1.2. STUDY OBJECTIVES

This study aims to understand the client's behaviour in the Portuguese banking sector, to build a model that can predict the loss of clients by using machine learning tools, with a Power BI approach.

Throughout this dissertation, it will be studying the customer's churn behaviour, focusing only on individual clients that already have closed all their bank accounts.

Not all clients have the same behaviour and for that reason it is important to distinguish those ones, taking into consideration some criteria, such as: how many transactions and how many different financial products each client has; if the client is blocked and/or if the client has a blocked account; if the client has a credit/debit card; and how long the client has a relationship with the Bank.

At last, the goal is to consider all these aspects to predict which clients are going to close their bank accounts, so that banks can maintain those ones and generate more value.

1.3. Study Relevance and Importance

Banks play an important role in the economy as they provide economic stability for the financial system and provide services in a way to facilitate transactions, loans, product sales and even promote job creation. It is therefore a highly regulated sector by several entities as it plays an active role in the economy of a country.

That is why it is essential that people who trust banks with their money and assets feel safe with the service provided. The importance of customer satisfaction in the banking sector is crucial for generating value and retaining clients. Understanding what went wrong will make the bank's business generate more money.

This study will be relevant for the business of the bank under study since the future goal is to implement this model in this banking institution. More than developing a model, is using the framework later, by monitoring target customers to help understand what actions to take with customers who eventually will close all their bank accounts.

2. LITERATURE REVIEW

This chapter is dedicated to the literature based on dissertations and papers like the case of this study. This will provide an overview of what already has been studied, by analysing the methods that were used and searching what is still missing about the subject.

2.1. CUSTOMER CHURN DEFINITION

Customer churn can be defined as the loss of customers who cease to do business with a certain company. It is a term that has been growing in popularity, as the chart from Figure 2.1 shows, which represents the word "churn" searched on google engine between 1st December 2013 and 1st December 2023 (the vertical axis represents the popularity peak).

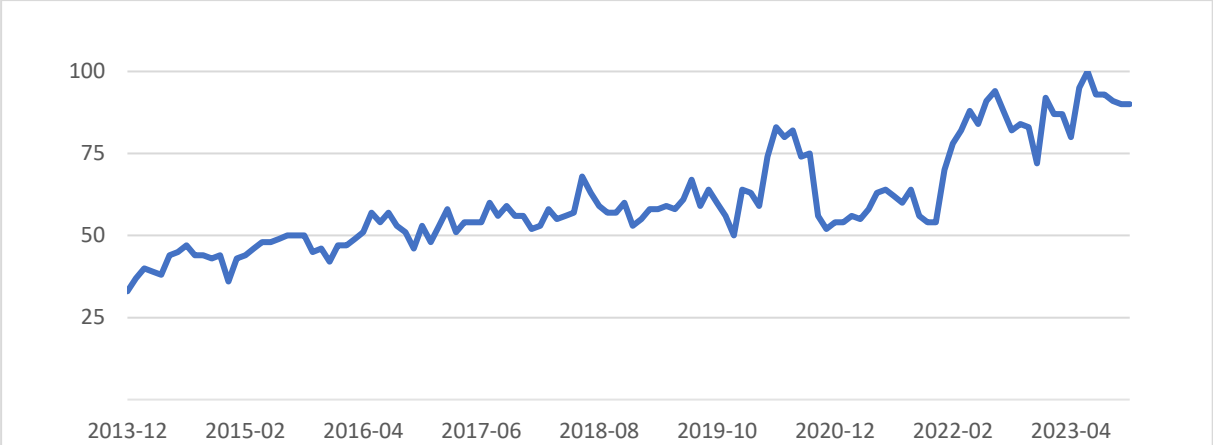


Figure 2.1 – Graph representing the word "Churn" between Dec. 2013 and Dec. 2023 (source: Google Trends)

This phenomenon is a major concern for businesses as it impacts their revenue and profitability (Santiago da Cunha, 2023). Understanding the factors that contributes to customer churn will lead companies to take actions to retain customers and minimize churn.

Businesses need to "shift their attention from customer acquisition to customer retention" (Oyeniyi & Adeyemo, 2015 since it becomes "very difficult for a company to attempt to win new business" Hadden, 2008. Hadden also states that the solution "should not target the entire customer base because (i) not all customers are worth retaining, and (ii) customer retention costs money; attempting to retain customers that have no intention of churning is a waste of resources".

Identifying customers who are likely to churn can be a daunting task for businesses. However, it is essential to have a proactive approach in retaining customers, which involves predicting and preventing churn. Predictive models can assist companies in determining which customers are most likely to leave and provide them with opportunities to take preventive measures. "Practically, the probability to end the relationship with the company is then used to rank the customers from most to least likely to churn, and customers with the highest propensity to churn receive marketing retention campaigns" (Coussement et al., 2017).

2.1.1. Churn in banking sector

The banking sector is particularly vulnerable to customer churn due to the high level of competition and the ease of switching to other banks. It can be costly for banks as it not only results in lost revenue but also increases customer acquisition costs.

Predicting customer churn is crucial for banks to improve customer retention and maintain a competitive edge. According to many sources, it says that it can costs five (5) times more to acquire a new customer than retaining an existing one. That's why businesses rather spend money on customer churn prediction than gaining new customers.

In light of this, it is important for banks to prioritize their attention towards their clients. By doing so, they can gain valuable insights into customer behaviour and preferences. Moreover, banks can identity patterns that can help them anticipate and predict their customers' future actions (Rosa, 2018).

By closely studying and analysing customer interactions, transactions, and historical data, banks can uncover trends and tendencies that offer significant value in understanding customer needs and expectations (Ejeu, 2021). These patterns can reveal valuable information, such as purchasing habits, financial goals, and risk tolerance levels. Armed with this knowledge, banks can tailor their offerings and services to align with each customer's unique requirements and preferences.

The ability to accurately predict a customer's next move provides several advantages for banks (Warikoo, 2022). Firstly, it allows them to proactively offer personalized recommendations, products, and services that cater to the specific needs and aspirations of individual customers. This proactive approach enhances customer satisfaction and loyalty, as customers feel understood and valued by their financial institution (Verbeke et al., 2012).

2.1.2. Case studies in banking sector

Customer churn in banking sector has been studied by several authors in several countries. In this section, it will be analysed studies with different approaches. This will give a wide understanding of what is causing customers to churn in order to create new strategies to mitigate churn and improving customer retention in the banking industry.

A research conducted (Sashikala, 2015) in India, with a sample size of 200 respondents, concluded that "customer service plays a very important role, followed by customer convenience, reliability of the bank and the product benefits". Furthermore, "on the basis of the factors identified and the contribution made by each to reduce customer churn, the banks can develop differentiation strategies and ensure customer retention".

In Lima's study (de Lima Lemos et al., 2022) it was used data from a large Brazilian financial institution. In this paper, it was concluded that the volume of commercial and housing loans active contributed as the most powerful predictor of customer churn, followed by client's

profitability and by the number of spontaneous movements carried out in the current account.

Another study made regarding customer retention by banks in New Zealand was taking into consideration (among others) the following elements: durability of relationship; customer satisfaction; corporate image; customer loyalty; customer's demographic and customer retention rate (Cohen et al., 2007). Regarding this study, the banks "has a positive impact on customers' likelihood of staying with their bank" which give an insight on the durability of the relationship between customer and bank. This is also related to the customer's age and level of education factors. On the other hand, the New Zealand banks "cannot rely upon price competition alone in order to be competitive" since there are other elements that can influence on the customer retaining, like the customer satisfaction, which should be an ongoing relationship between the customer and the bank.

Beside identifying patterns, it is also important to understand that there are external factors that can have impact on the customer's behaviour, such as the economic crisis that began in 2008. According to Akshay and António (Cabeças, 2014), this crisis influenced the customer's behaviour in the Portuguese banking sector in the reduction of savings and of the volume of credit obtained from the banks. It also conditioned customer's satisfaction and affected the level of loyalty.

By immersing in these studies, it is possible to gain practical knowledge since it offers different perspectives that can contribute to the development of best practices for customer churn management within the banking sector.

2.2. Predictive Models

The process of prediction involves collecting and analysing data, identifying patterns and trends in data, and using statistical or machine learning algorithms to develop models that can forecast future outcomes. Customer churn prediction is one of these predictive modelling techniques that can use machine learning algorithms to identify customers who are likely to churn.

By analysing customer data such as purchase history, demographics, and behavioural patterns, predictive models can forecast which customers are most likely to leave the company. That's why (Hadden, 2008) state that companies need to understand their customers. This helps them to take proactive measures to prevent customer churn by providing targeted incentives, personalized offers or improving customer service.

According to (Verbeke et al., 2012) "based on historical data a model can be trained to classify customers as future churners or non-churners". Churn can be caused by a variety of factors, and customers may leave a company for reasons that are difficult to predict (Rosa, 2018). Additionally, customer behaviour may change over time, which can make it challenging to maintain accurate predictive models due to the complex and dynamic nature of customer

behaviour (de Lima Lemos et al., 2022). The effectiveness of churn prediction relies on evaluating how accurately prediction models can identify whether a customer is likely to churn or not. Hence, evaluating the accuracy of prediction models in identifying customers likely to churn is crucial for determining the success of churn prediction, since companies can take proactive measures to prevent it from happening (Ismail et al., 2015).

By leveraging machine learning algorithms and predictive modelling techniques, companies can identify potential churners and take proactive measures to prevent customer churn. Therefore, companies must continually evaluate the accuracy of their predictive models to ensure they remain effective and relevant.

2.2.1. Machine Learning Models in Customer Churn

There are several approaches used in customer churn prediction, including statistical modelling, machine learning algorithms, and data mining techniques. Each approach has its strengths and weaknesses, and the choice of each depends on the specific requirements and constraints of the business problem. In recent years, machine learning models are beginning to gain relevance, especially in churn prediction (Oluwatoyin et al., 2023).

As one reads about the literature, it becomes clearer that some studies use different machine learning methods from each other's. Based on a 40 publications' research, from 2010 and June 2020, most of these customer churn's studies found that decision trees, support vector machines and logistic regression were the 3 most popular and useful methods (Tianyuan & Moro, 2021).

Machine learning algorithms are programmed algorithms that use statistical methods and artificial intelligence techniques to automatically learn patterns in data in order to make predictions or take actions. They can be used in a wide range of applications, including image and speech recognition, natural language processing, fraud detection and customer churn.

There are several reasons why machine learning algorithms are often chosen for customer churn prediction. For Lemo's (de Lima Lemos et al., 2022) there are characteristics that motivate the use of machine learning techniques:

1. Scalability: machine learning algorithms can be trained on large data sets and scaled to handle high volume data processing in real time.
2. Automation: machine learning algorithms can be automated to make predictions and take actions in real-time, enabling businesses to respond quickly to changes in customer behaviour and take proactive measures to prevent churn.
3. Ability to handle complex data: machine learning algorithms can effectively handle large and complex data sets as also identify patterns.
4. Flexibility: machine learning algorithms can be applied to a wide range of data types, including structured and unstructured data, and can be used to solve a variety of business problems beyond customer churn prediction.

Machine learning algorithms offer a powerful and flexible approach to customer churn prediction, enabling businesses to gain insights into customer behaviour, identify at-risk customers and take proactive measures to prevent churn.

This work will use machine learning algorithms as the literature highlights their growing relevance in customer churn prediction. With this understanding, this study aims to delve deeper into the application and effectiveness of specific machine learning algorithms in the banking sector for customer churn prediction.

3. METHODOLOGY

The adopted methodology in this study is align with the research objectives of generating value and maintaining customers. This section provides an overview of the research approach, encompassing the data collection process, data analysis techniques and the specific methods employed to address the research questions surrounding why banks aim to predict their client's behaviour and achieve the research objectives.

A clear and transparent account of the steps undertaken to ensure the validity and reliability of the findings will be present. In this section, it will be demonstrated the systematic and rigorous nature of the research process, highlighting the adherence to established frameworks and principles.

In this study, a mixed-methods approach is employed, combining quantitative and qualitative data collection and analysis techniques. This approach enables a comprehensive exploration of the research topic, allowing for a deeper understanding of the complexities and nuances associated with the phenomenon under investigation.

Furthermore, this section describes the research design, sample selection, data sources, and the tools and software utilized for data analysis. It also outlines the ethical considerations considered during the research process, ensuring the protection of clients' rights and privacy.

By providing a clear methodology, this study aims to enhance the credibility and robustness of the findings, enabling readers to assess the reliability of the results and replicate the research process if desired. The methodology section serves as a roadmap for the research journey, guiding the reader through the various stages undertaken to address the research questions and contribute to the existing knowledge in the field.

3.1. CRISP-DM

The CRISP-DM presented in Figure 3.1 is a well-established and widely adopted framework for guiding the process of data mining and analytics (Chapman et al., 2000). This methodology provides a structured approach to solving data-driven problems, making it particularly valuable in the field of data science and research. In the context of this project, the CRISP-DM will serve as the cornerstone of the research methodology, offering a systematic and organized process for tackling the research questions and objectives outlined.

The significance of CRISP-DM lies in its ability to break down the data analysis process into distinct phases, ensuring that each step is carefully considered and executed. These phases include Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. By adhering to this framework, it is aimed to enhance the efficiency of the research, ultimately leading to more meaningful and reliable results.

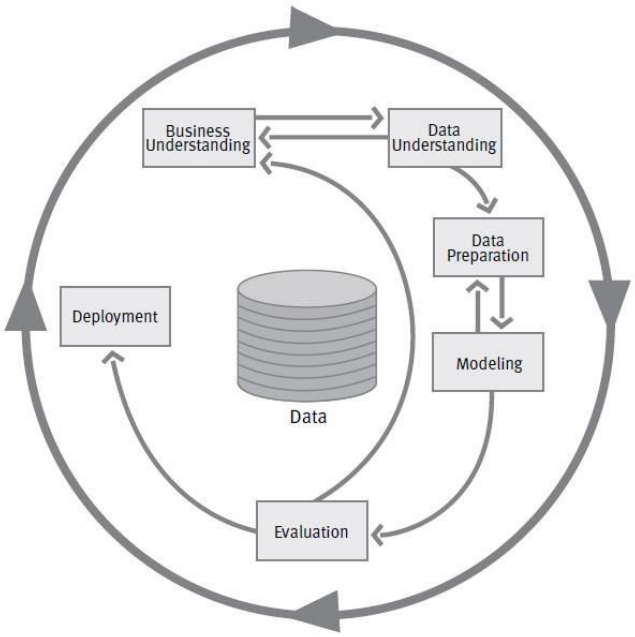


Figure 3.1 – Phases of the CRISP-DM reference model (Chapman et al., 2000)

3.1.1. Business Understanding

Within the context of this research project, the Business Understanding phase involves a comprehensive examination of the reasons why a bank would want to predict customer behaviour, focusing on generating value and maintaining clients. The research question, "Why does a bank want to predict their client's behaviour?" serves as the foundation for understanding the business problem and aligning the analysis with the bank's objectives.

In this context, the data used in this project originates from operational sources and was carefully acquired, following all necessary authorizations from the Bank. This operational data provides a rich source of information, encompassing customer transaction records. The data was meticulously treated to meet the standards of data quality and security (by masking the client's personal data), ensuring that it serves as a reliable foundation for our research.

The computational resources employed for this project include a Windows 11 computer as the primary platform for analysis. This analysis was conducted within the Anaconda 2.5.0 environment, leveraging the versatility and capabilities of Jupyter Notebook. Within this environment, a variety of libraries, such as Pandas, NumPy, Matplotlib, Seaborn, and others, were harnessed to facilitate data preprocessing, modelling, and data visualization.

Furthermore, the requirements of the project are oriented towards producing actionable insights that allow the prediction of clients who are at risk of churning. By focusing on this requirement, it will empower this Portuguese Bank to take proactive measures that generate value and retain clients. It is also important to mention that the dataset for this analysis is available in a csv file format, extracted from the bank's BI server.

3.1.2. Data Understanding

The Data Understanding phase encompasses the insights and information that was gathered about the dataset, aiming to identify and mitigate potential issues that may arise in subsequent project phases. This is achieved with the careful examination of data tables and graphical representations, allowing to assess the data quality and characteristics.

The initial step entails the collection of data from the csv file, which includes the selection of software libraries for the project's development within the Jupyter environment.

This dataset pertains to individuals and is structured in a tabular format contained within a single csv file. It comprises a total of 25.249 observations (rows) and 21 variables (columns), providing a source of information for analysing customer interactions and predicting churn.

The data spans from 1st April 2022 to 30th September 2022, capturing a diverse range of transactional data. The snapshot date, set as 30th September 2022, acts as the focal point for assessing behaviour leading up to subsequent analysis period. To allow for customers to make decisions about their accounts, a quarantine period is designated from 1st October 2022 to 31st December 2022. The target variable – *Churn*, indicates whether clients, as of the snapshot date, churned or not. Churn determination is based on events occurring between 1st January 2023 and 30th June 2023. A more visual timeline is represented on Figure 3.2.

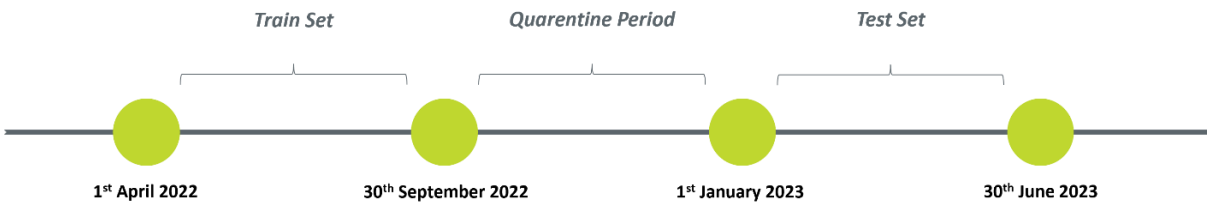


Figure 3.2 – Timeline regarding train and test datasets

With EDA it is possible to understand the dataset since it is a fundamental step in this phase. It involves the comprehensive description of key characteristics within the dataset, providing insights through the application of numerical and data visualization techniques. Thus, EDA helps to gain a profound understanding of the observations, unveiling patterns, anomalies and confirming assumptions.

In Table 3.1, it is presented the variables, its data types, description, and missing values.

Table 3.1 – Dataset's features summary

Variable	Type	Obs.	Missing Values	Description
Client_Number	int64	25.249	0	Client ID
Gender	object	25.249	0	Client gender
Age	float64	25.249	0	The age (in years) of the client

Tenure	int64	25.249	0	Total amount of months that the client has been with the Bank
Online	int64	25.249	0	Flag that indicates if the client was opened online or presential
Plan	Object	25.249	0	Client's plan
Account_Blocked	int64	25.249	0	Flag that indicates if the client has at least one blocked account
Client_Blocked	int64	25.249	0	Flag that indicates if the client is blocked
Blocked	float64	25.249	0	Total amount of months that the client has (at least) one account that has been blocked
Cross_Selling	Int64	25.249	0	Indicates how many different products a client has
Debit_Card	int64	25.249	0	Flag that indicates if the client has an active debit card
Credit_Card	int64	25.249	0	Flag that indicates if the client has an active credit card
Warranty	int64	25.249	0	Flag that indicates if the client has warranties associated
Last_Transaction	int64	25.249	0	Total amount of months of the client's last transaction regarding current accounts
Balance_DO	float64	25.249	0	Balance amount (converted to euro) regarding current accounts
Balance_DP	float64	25.249	0	Balance amount (converted to euro) regarding term accounts
Balance_Credit	float64	25.249	0	Balance amount (converted to euro) regarding credit accounts
Outstanding_Balance	float64	25.249	0	Balance amount (converted to euro) regarding outstanding commissions
Average_Balance_6M	float64	25.249	0	Average balance amount (converted to euro) in the last 6 months
Average_Transactions_6M	int64	25.249	0	Average transactions in the last 6 months
Churn	int64	25.249	0	If the client has been closed the 6 months after

Within the dataset, it is possible to encounter various variables, each with unique roles and significance. To provide a clearer understanding, it will be given some more detail regarding specific features:

- *Account_Blocked*: A categorical variable related to whether the client has at least one blocked account. Depending on the nature of the block, certain transactions or account activities may be restricted.
- *Client_Blocked*: This categorical variable indicates whether the client is completely blocked, which means they are prevented from accessing their bank accounts entirely.
- *Blocked*: This variable indicates how long (in months) a client has been blocked by the Bank, where the ones that aren't blocked are marked as "999" (to distinguish the ones that have been blocked for less than 1 month).
- *Plan*: The Plan variable is regarding the client's plan and there are 3 different plans (in ascending order by the benefits that each plan provides): BASIC, PLUS and PREMIUM. For those who doesn't have any plan associated are marked as "SEM PLANO".

- *Warranty*: This categorical variable indicates the presence of any warranties associated with the client's accounts. These warranties may encompass guarantees, mortgages on the client's property and other similar commitments.
- *Last_Transaction*: Like the *Blocked* variable, this feature represents how long (in months) a client hasn't made a transaction, where the ones that make haven't made any transactions at all are marked as "999" (to distinguish the ones whose last transaction was made less than 1 month ago).
- *Outstanding_Balance*: This numerical variable represents the balance of outstanding commissions owed by the client to the bank. Notably, these outstanding commissions are based on data as of 30th September 2022.
- *Average_Transactions_6M*: This numerical variable represents the average transactions that a client made regarding the reference date from the past 6 months, where the ones that make haven't made any transactions at all are marked as "999".

Besides understanding what kind of variables the dataset has and the meaning of them, it is also important to understand the significance of the target variable among the other attributes. The target variable – *Churn*, is the focal point, as it represents the outcome of interest. A better comprehensive of this variable is crucial as it guides the entire data mining process, shaping the selection of relevant features, and influencing the model's predictive capabilities.

In this context, a valuable initial step involves understanding the churn rate within the dataset. As it shows in the below pie chart, about 2,70% of customers have churned. Given that this percentage is a small number (which leads to an unbalanced dataset that will be dealt later), it needs to be ensured that the chosen model will predict with accuracy these churned customers. Figure 3.3 displays this pattern.

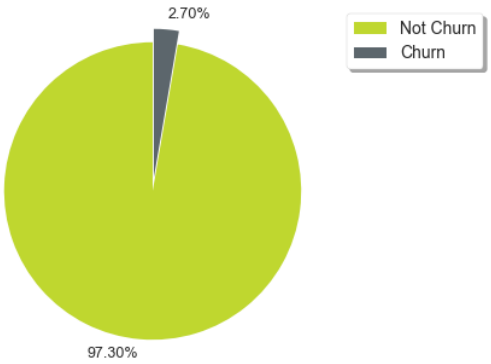


Figure 3.3 – Churn rate

Regarding the relation between the target variable and the other ones within the dataset, there are insights that should be considered. The clustered bar charts regarding this relation are shown below.

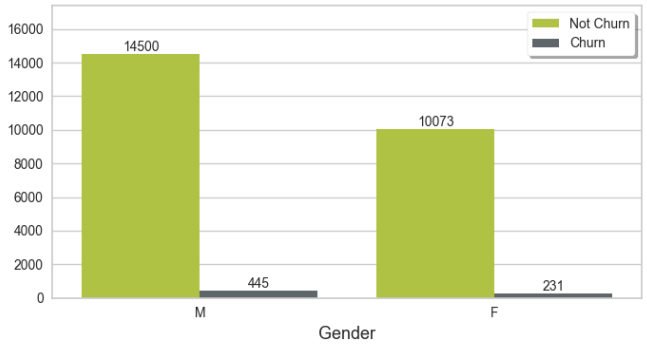


Figure 3.4 – Churn by Gender

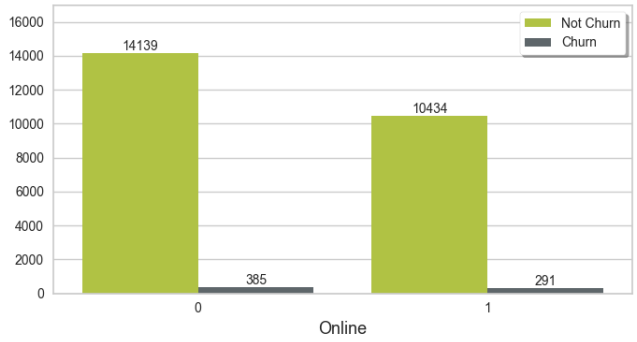


Figure 3.5 – Churn by Opening Channel

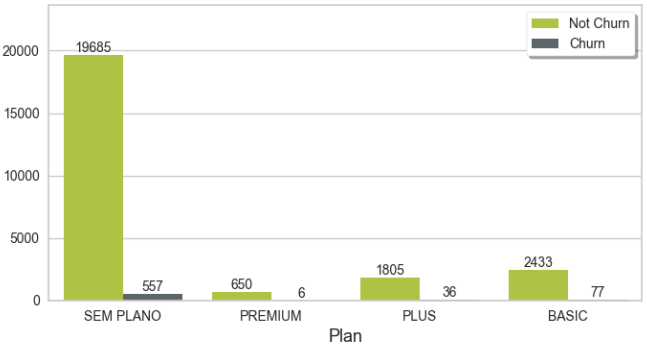


Figure 3.6 – Churn by Plan

Figure 3.7 – Churn by Blocked Account

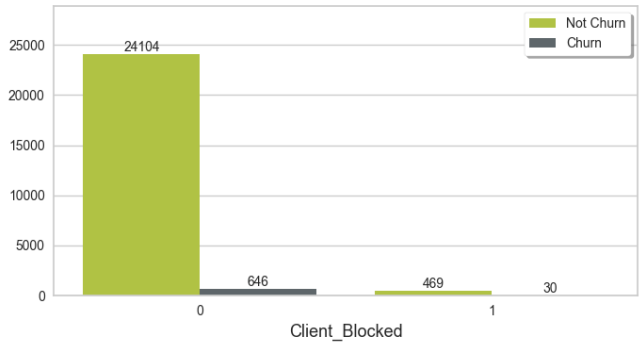


Figure 3.8 – Churn by Blocked Client

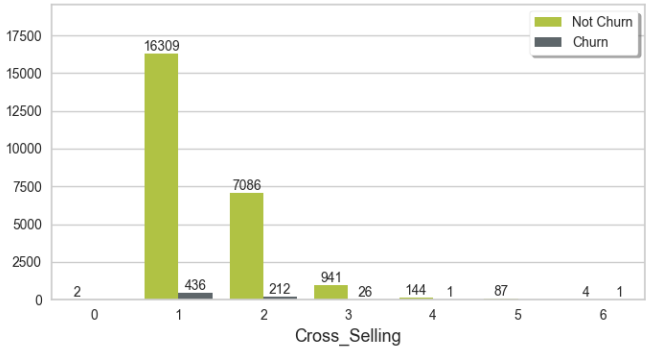


Figure 3.9 – Churn by Cross-Selling

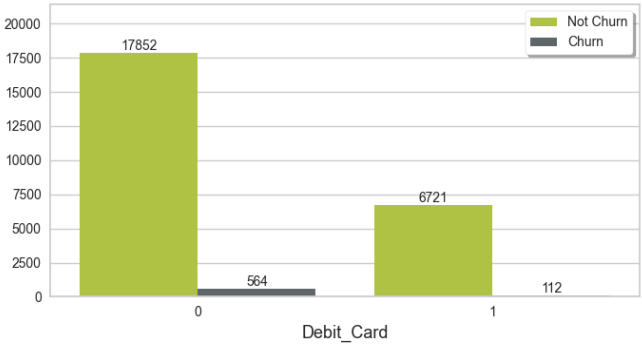


Figure 3.10 – Churn by Debit Card

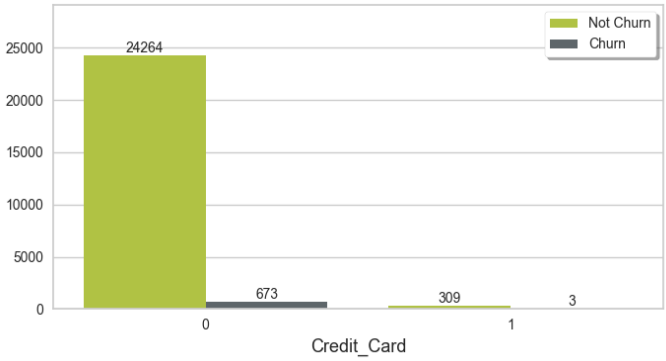


Figure 3.11 – Churn by Credit Card

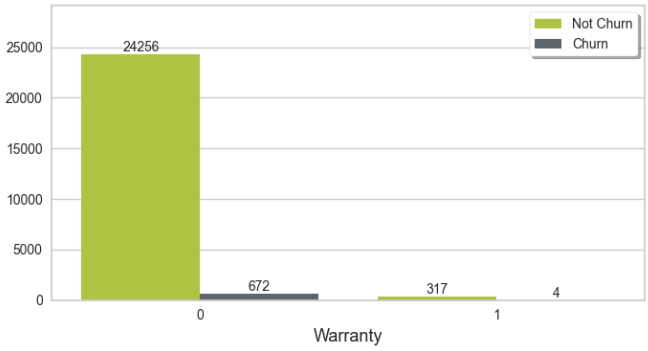


Figure 3.12 – Churn by Warranty

From these clustered bar charts, it is possible to get a few insights, such as:

- Most customers are males, and likewise, the predominant group of churn customers is also male (Figure 3.4).
- The predominance group regarding the opening channel is presential, nevertheless, the bar chart indicates that there is no relation between the churned customers and the opening method, since it there is no practical difference (Figure 3.5).
- According to Figure 3.6 there's seems to be a relation between the client's plan and churned customers, since the higher the plan, the lower the churn, which means that there are more churned clients without a plan than with the PREMIUM plan.
- Regarding the client and accounts status, there's seems to be no relation between clients / accounts blocked with churned customers (Figure 3.7 and Figure 3.8). Most of the clients don't have warranty products (Figure 3.12) and the majority has just one different kind of products, which also correspond to the majority of the churn client's (Figure 3.9).
- There is a significant number of customers that have a debit card, however most of the customers that churned didn't have that product (Figure 3.10).
- On the other hand, there are a few number of customers that have a credit card, and the majority of the customers that churned didn't have that product (Figure 3.11).

Besides the relation between the target variable and the other features, it could be interesting to analyse and observe the numerical features, using boxplots and histograms, as it will help to give a better understanding of these features within the dataset. Below is shown, for each numerical variable, the boxplots and the histogram charts, respectively.

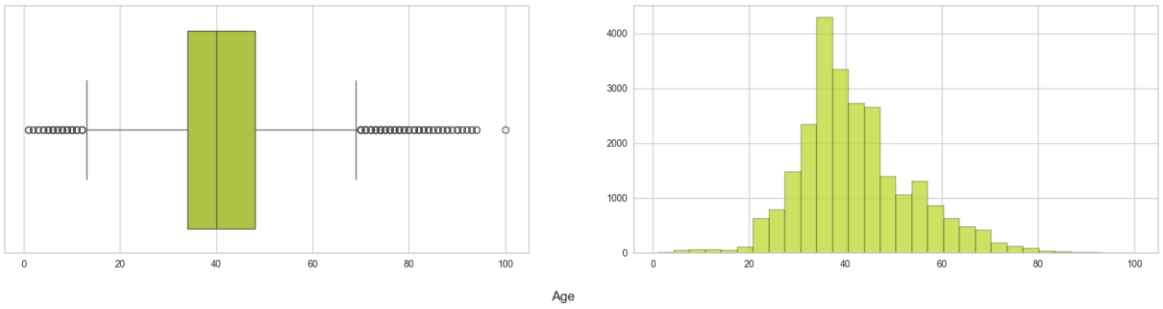


Figure 3.13 – Boxplot and Histograms' Age variable

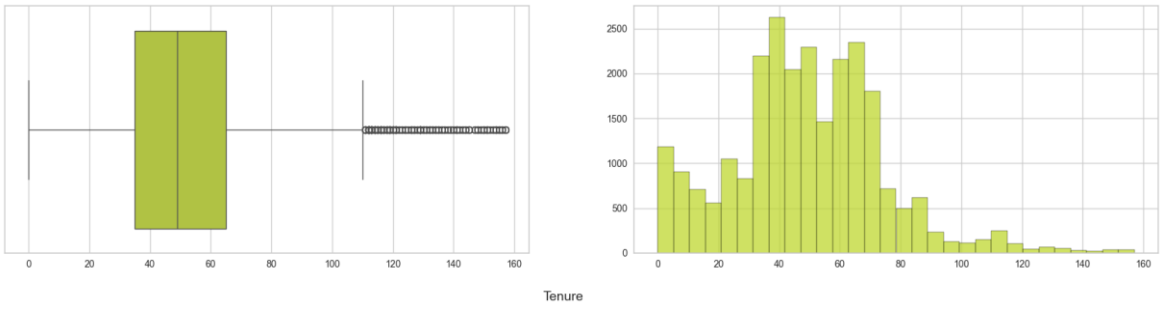


Figure 3.14 – Boxplot and Histograms' Tenure variable

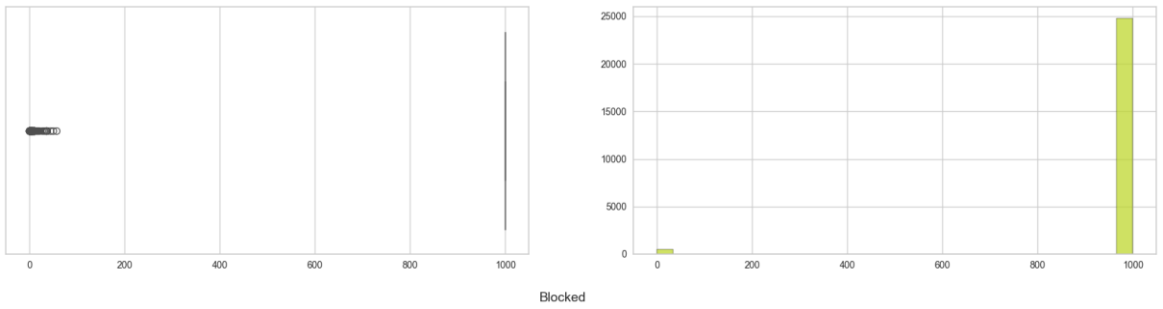


Figure 3.15 – Boxplot and Histograms' Blocked variable

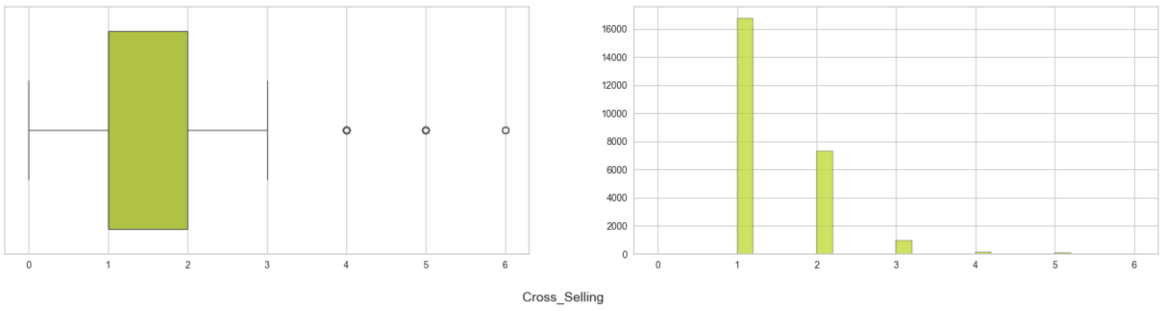


Figure 3.16 – Boxplot and Histograms' Cross_Selling variable

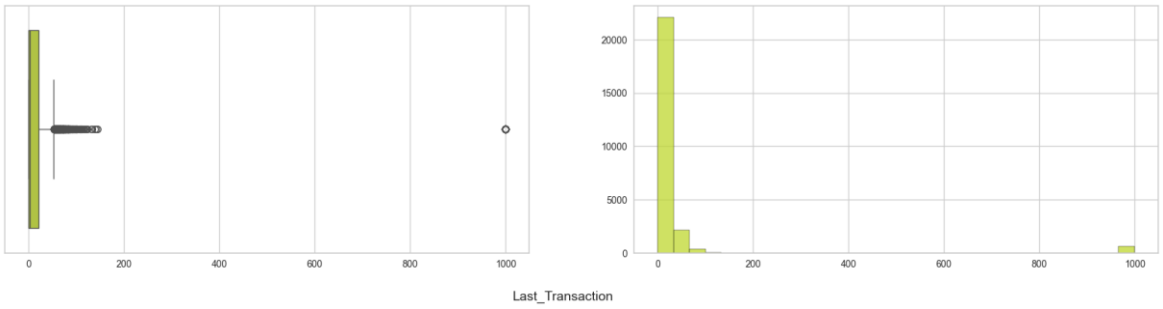


Figure 3.17 – Boxplot and Histograms' Last_Transaction variable

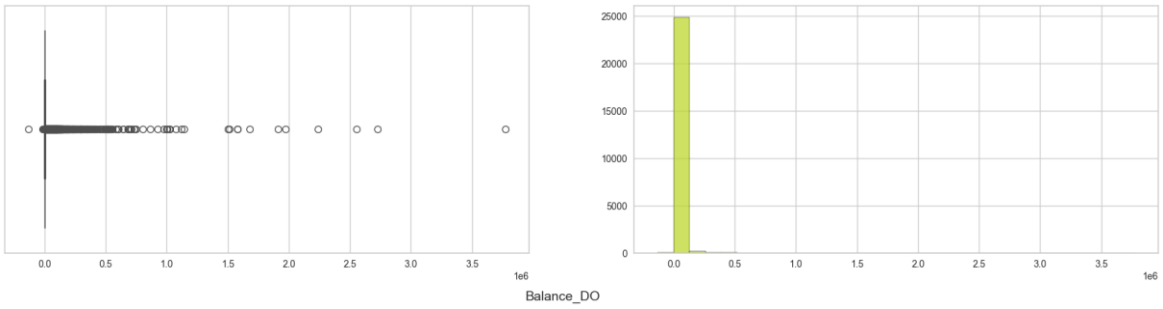


Figure 3.18 – Boxplot and Histograms' Balance_DO variable

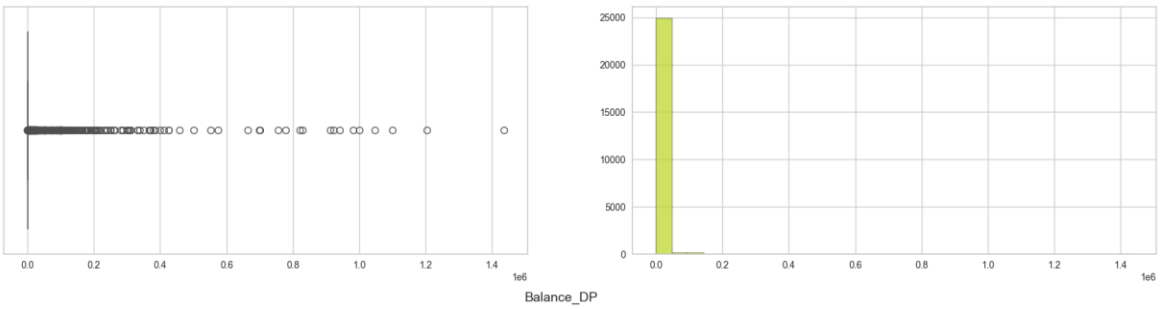


Figure 3.19 – Boxplot and Histograms' Balance_DP variable

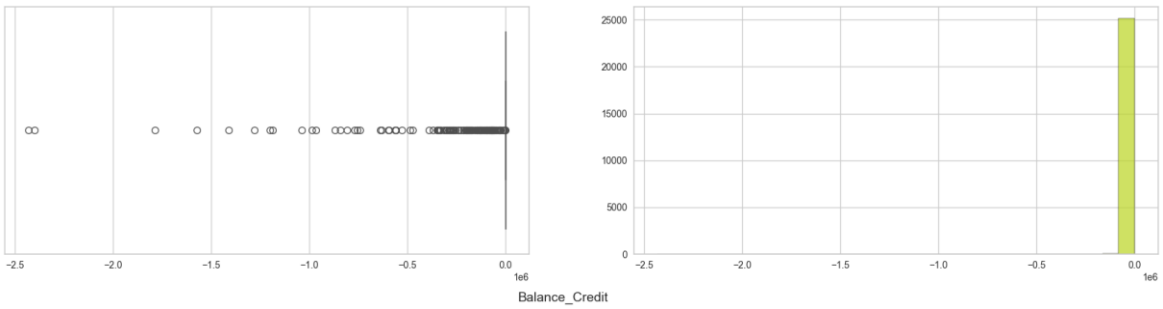


Figure 3.20 – Boxplot and Histograms' Balance_Credit variable

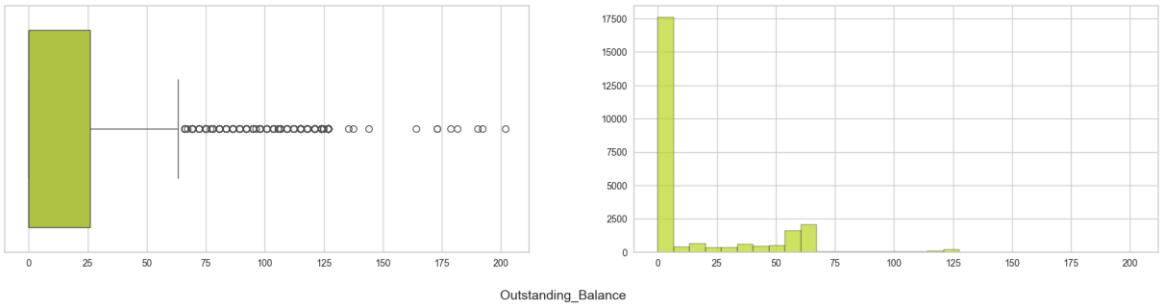


Figure 3.21 – Boxplot and Histograms' Outstanding_Balance variable

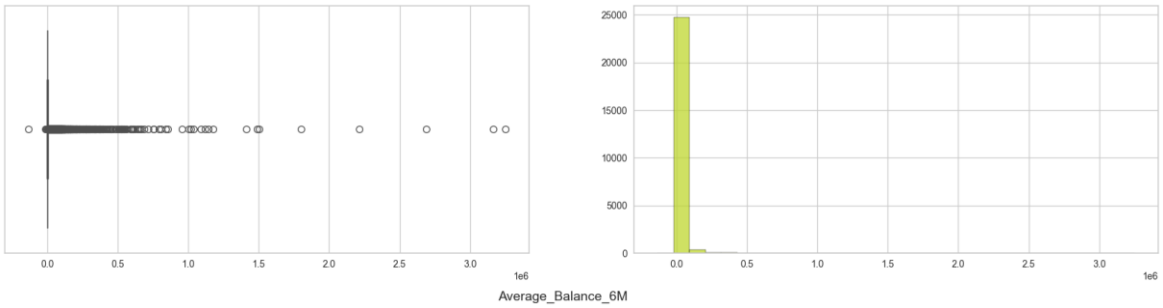


Figure 3.22 – Boxplot and Histograms' Average_Balance_6M variable

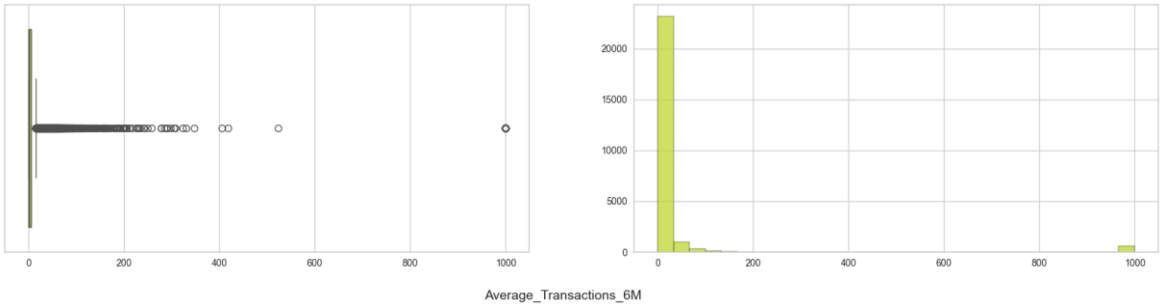


Figure 3.23 – Boxplot and Histograms' Average_Transactions_6M variable

For *Age* (Figure 3.13) and *Cross_Selling* (Figure 3.16) variables, they seem to have a right skewed distribution, whereas the other variables – *Tenure* (Figure 3.14), *Blocked* (Figure 3.15), *Last_Transaction* (Figure 3.17), *Balance_DO* (Figure 3.18), *Balance_DP* (Figure 3.20), *Balance_Credit* (Figure 3.19), *Outstanding_Balance* (Figure 3.21), *Average_Balance_6M* (Figure 3.22) and *Average_Transactions_6M* (Figure 3.23), don't have any kind of distribution. Overall, in a first glance, this dataset is imbalanced, which will be handle later in the next chapter.

3.1.3. Data Preparation

Data Preparation is essential for ensuring the data's quality and readiness for analysis. In this phase, data will be cleaned and preprocess, addressing issues such as data format standardization.

Data cleaning is the process of identifying and rectifying incorrect or irrelevant data within a dataset. The primary goal of data cleaning is to improve data quality and enhance the productivity of the analysis. It's also important to note that it will allow to retain the highest quality of information within the dataset, ultimately enabling to draw more accurate conclusions and achieve better results in subsequent phases. Additionally, it's important not to delete a substantial amount of data, generally avoiding the removal of more than 3-4% of observations. This strategy helps maintain a robust dataset with ample observations and ensures that valuable data will be retained for later phase of the model.

3.1.3.1. Duplicated values

Regarding duplicated values, when not handled, it can lead to inconsistencies and skew results, making it crucial to eliminate them to maintain the integrity of the analysis. Nevertheless, this dataset did not have duplicated values.

3.1.3.2. Missing values

Dealing with incomplete information in a dataset, where certain observations in a dataset have lack information, is a vital component of data preparation. Handling missing values is a significant aspect of data preparation as it allows for more effective data analysis and the extraction of valuable insights. Nevertheless, this dataset doesn't have missing values.

3.1.3.3. Outliers treatment

In addition to the methods previously employed during Data Preparation, addressing outliers is also crucial. An outlier represents an observation that deviates significantly from most values within the dataset. This deviation, whether exceptionally high or low, has the potential to introduce complications in statistical analysis or, conversely, provide valuable insights into the dataset. Understanding how to identify and address outliers is very important for gaining a more profound comprehension of the dataset.

While addressing outliers is an important step for maintaining statistical robustness, it's equally important to strike a balance and consider the potential impact on the dataset's overall information content.

In this scenario, the treatment of the outliers is exclusively applied to the numerical variables, and their identification is accomplished through data visualization techniques, namely through boxplots and histograms. The graphs from Figure 3.13 to Figure 3.23 represent how each numerical variable is distributed, enabling a comprehensive analysis of the distribution of observations.

By analysing these graphs, one can come to conclusion that all numerical variables have an irregular statistical distribution, where those anomalies could be considered as outliers. Two commonly employed techniques, namely Z-Score and IQR were evaluated to identify and treat

outliers. The choice between these two techniques was guided by a careful consideration of the impact on the dataset and subsequent model performance.

3.1.3.3.1. Z-Score

The Z-Score technique was selected for outlier treatment based on several considerations. This method calculates the number of standard deviations a data point is from the mean, providing standardized measure of deviation. Through a comprehensive analysis of the dataset, it was observed that using the Z-Score resulted in the removal of approximately 8% of the dataset. This event was deemed beneficial, particularly after subsequence analyses revealed that the predictive model exhibited superior performance when outliers were excluded. The decision to remove outliers using this technique was influenced by a desire to improve model generalization without excessively distorting the underlying distribution of the data.

3.1.3.3.2. IQR

While the IQR technique also offers a robust method for outlier identification, it was decided against its extensive application in this study. Analysis using the IQR technique revealed the potential removal of around 21% of the dataset. However, this level of data removal raised concerns about the risk of overfitting the model to the training data, potentially compromising its ability to generalize to unseen instances. Given the inherent imbalance in the dataset, preserving a representative sample of the minority class – churn, was deemed paramount to avoid introducing biases and achieving a balanced compromise between outlier treatment and model performance.

In summary, the decision to employ Z-Score technique for outlier treatment was driven by its more conservative impact on the dataset, ensuring a balanced approach to model training and avoiding the potential pitfalls associated with overtraining. This choice aligns with the overarching goal of developing a predictive model for customer churn that not only performs optimally but also maintains the integrity of the underlying data distribution.

Still the decision of using Z-Score, the dataset seems to still have some outliers that can be shown from Figure 3.24 to Figure 3.30. Due to potential loss of data, it was decided to keep those outliers. Despite that, while keeping outliers could become an issue later, there are machine learning models that can handle outliers effectively, which will be approached later in the next chapter.

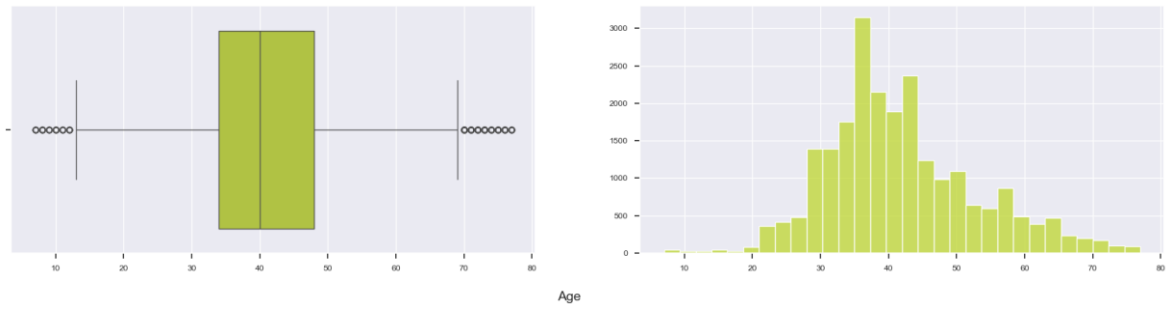


Figure 3.24 – Boxplot and Histograms' Age variable after data preparation

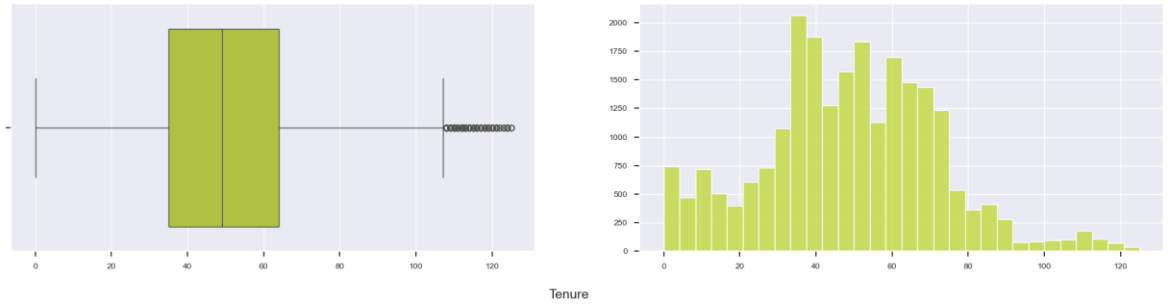


Figure 3.25 – Boxplot and Histograms' Tenure variable after data preparation

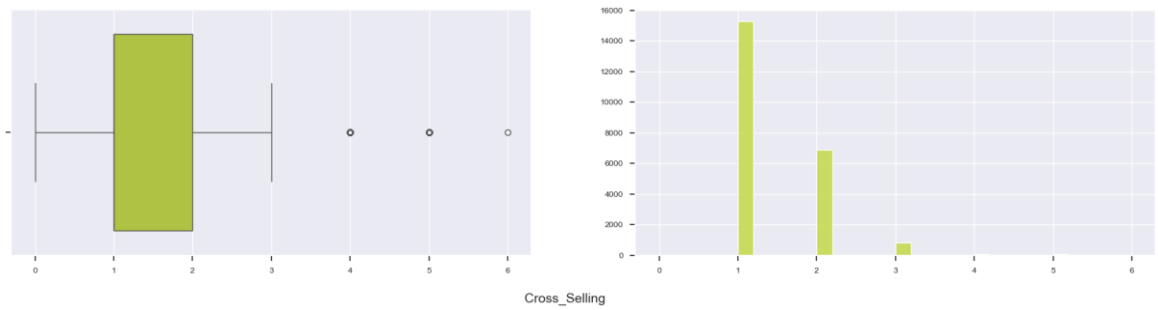


Figure 3.26 – Boxplot and Histograms' Cross_Selling variable after data preparation

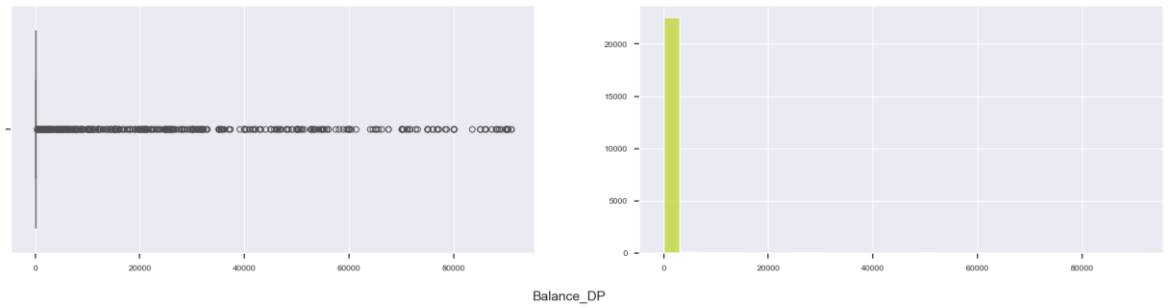


Figure 3.27 – Boxplot and Histograms' Balance_DP variable after data preparation

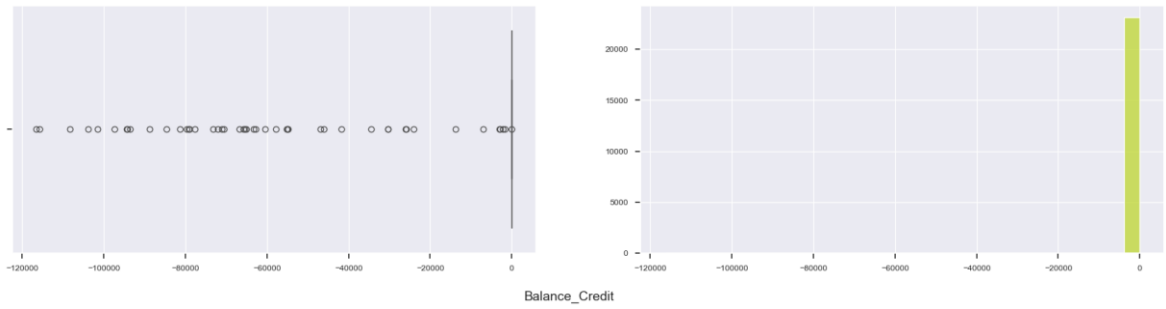


Figure 3.28 – Boxplot and Histograms' Balance_Credit variable after data preparation

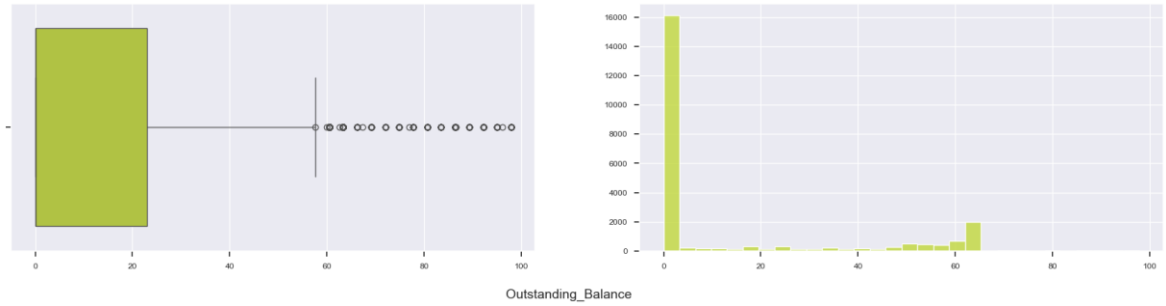


Figure 3.29 – Boxplot and Histograms' Outstanding_Balance variable after data preparation

Figure 3.30 – Boxplot and Histograms' Average_Transactions_6M variable after data preparation

3.1.3.4. Feature selection

Examining the correlation between variables is another crucial step in the Data Preparation process. This analysis is exclusively applied to numerical features, enabling to identify potential redundancy among them. To accomplish this, it was employed the Spearman correlation coefficient, since this one access non-linear relationships and is more robust to the presence of outliers and non-normally distributed data. In this context, the heatmap chart is presented in Figure 3.31 for visual reference.

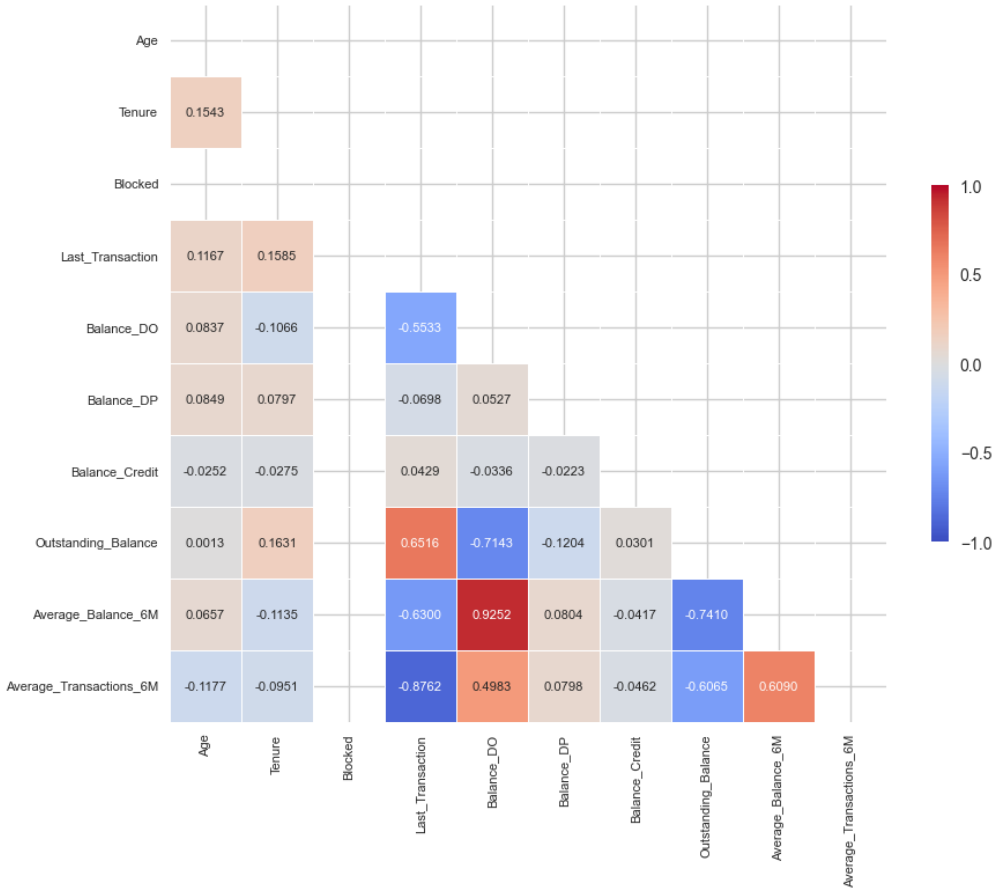


Figure 3.31 – Spearman heatmap between numerical variables

Based on the heatmap, it becomes evident that the variables *Last_Transaction*, *Average_Balance_6M* and *Balance_DO* exhibit an exceptionally strong correlation. Besides, the feature *Blocked* doesn't show correlation values due to the data preparation done previously.

Consequently, it was removed the features *Last_Transaction*, *Average_Balance_6M*, *Balance_DO* and *Blocked*.

3.1.3.5. Encoding

Data needs to be specifically prepared before fitting into a modelling process. In this regard, there are many modelling techniques that needs numerical inputs, requiring the conversion of these categorical variables into numerical.

To achieve this conversion, it was adopted the Label Encoder method to transform the original categorical labels into numerical. The resulting can then be used as input for the machine learning models further in this study. In the context of this project, 2 variables were transformed, namely *Gender* and *Plan*.

3.1.3.6. Data normalization

Having established the foundations of outliers' treatment, correlation analysis and explored the utility of the encoding method, the significance of data normalization becomes increasingly apparent. In the scope of data science, where diverse datasets often exhibit varying scales and structures, normalization acts as a crucial bridge. A normalized dataset facilitates more reliable model training, reduces the impact of outliers, and contributes to the production of insightful and actionable results.

One widely used approach for normalization in case the dataset has outliers is the Robust Scaler. Unlike other scalers, such as the Standard Scaler, the Robust method is less influenced by outliers, which is the case of this dataset, since it scales features based on the IQR rather than the mean and the standard deviation.

Afterall, the descriptive statistics for the numerical variables after the data preparation sums into Table 3.1Table 3.2 and for the categorical variables In Table 3.3.

Table 3.2 – Descriptive statistics of numerical variables

Variable	Count	Mean	Std	Min	25%	50%	75%	Max	Missing Values
Age	23.175	0,12	0,79	-2,36	-0,43	0,00	0,57	2,64	0
Tenure	23.175	-0,01	0,80	-1,69	-0,48	0,00	0,52	2,62	0
Balance_DP	23.175	1,29	3,68	-0,09	-0,09	0,00	0,91	78,33	0
Balance_Credit	23.175	0,12	0,79	-2,36	-0,43	0,00	0,57	2,64	0
Outstanding_Balance	23.175	-0,01	0,80	-1,69	-0,48	0,00	0,52	2,62	0
Average_Transactions_6M	23.175	0,63	1,06	0,00	0,00	0,00	1,00	4,25	0
Churn	23.175	0.03	0.16	0,00	0,00	0,00	0,00	1,00	0

Table 3.3 – Descriptive statistics of categorical variables

Variable	Count	Unique	Top	Freq	Missing Values
Gender	25.249	2	1	13.598	0
Online	25.249	2	0	13.169	0
Plan	25.249	4	3	18.554	0
Account_Blocked	25.249	1	0	23.175	0
Client_Blocked	25.249	1	0	23.175	0
Cross_Selling	25.249	7	0	15.283	0
Debit_Card	25.249	2	0	16.647	0
Credit_Card	25.249	2	0	22.971	0
Warranty	25.249	2	0	22.969	0

3.1.4. Modelling

The Modelling section is regarding the application of data mining techniques and algorithms to build the predictive model. This is where the core analysis of the data will take place, and various modelling methods will be employed to make predictions.

In the context of the customer churn prediction problem, the robustness of machine learning models to outliers emerged as a consideration during the analysis phase. Even though outliers were treated previously, the dataset still present some of them. The models under consideration were evaluated for their ability to handle outliers, with a focus on preserving predictive accuracy and model generalization.

In the pursuit of developing a predictive model for customer churn, the selection of appropriate machine learning models played a crucial role in shaping the analytical framework. A key consideration in this process was the anticipated presence of outliers within the dataset and the necessity to deploy models inherently robust to such anomalies. The choice of models was made based on the literature.

The selection of ensemble and tree-based models was motivated by their inherent characteristics that render them robust to outliers. These models operate on principles of aggregating predictions from multiple decision trees or weak learners. The ensemble nature of these models is expected to distribute the influence of outliers, making them promising candidates for scenarios where data may exhibit anomalies.

Consequently, the machine learning models used were the following:

- Decision Trees;
- Extra Trees;
- Random Forest;
- XGBoost; and
- Gradient Boosting.

When building a predictive model using machine learning, it is important to partition the dataset into two distinct subsets: one dedicated to training the model, which corresponds to 80% of the dataset; and another reserved for testing its performance, which corresponds to 20% of the remaining dataset. This splitting ensures that the model, once trained, can assessed on unseen data, providing a robust evaluation of its generalization capabilities, and preventing overfitting to the training set.

In the context of churn prediction, where the consequences of misclassifying potential churners can be significant, leveraging resampling method is instrumental. These resampling methods are important since they address class imbalance, which occurs when the number of instances of one class (e.g., customer who churn) significantly outweighs the other in the dataset. To mitigate potential biases and ensure the model learns equally from both classes, oversampling the minority class was applied, using the SMOTE technique. This will allow the

model to glean insights from both scenarios, improving its capacity to identify and predict customer churn effectively. By striking a balance between the representation of both churn or non-churn instances, this technique contributes to a more robust and accurate predictive model.

Furthermore, the *GridSearchCV* technique was employed for each model to identify the optimal values for hyperparameters. However, the *RandomizedSearchCV* methodology was needed by computational considerations. While *GridSearchCV* systematically explores a predefined hyperparameter grid through an exhaustive search, its time-consuming nature prompted the adoption of *RandomizedSearchCV*. The last one employs a randomized sampling approach within the hyperparameter space, offering a more computationally efficient alternative. This strategy not only accelerates the hyperparameter tuning process but also upholds the commitment to optimizing model performance for effective customer churn prediction.

3.1.5. Evaluation

The Evaluation phase will be focusing on assessing the performance of the models. It will also be employed evaluation metrics such as accuracy, precision, recall and the F1-Score to measure the quality of the models. When dealing with imbalanced datasets and outliers it is important to consider a comprehensive set of evaluation metrics. Metrics such as accuracy provide an overall view of model performance, but in scenarios where class distribution is uneven, precision, recall and the F1-Score become vital. Precision measures the accuracy of positive predictions, recall gauges the ability to capture all positive instances, and the F1-Score harmonizes these two metrics, offering a balanced assessment.

Besides the metrics previously referred, there are other metrics that will be present with some visualizations, namely:

- AUC: is a metric to evaluate the performance of binary classification models with the help of graphical representation of the trade-off between the true-positives (sensitivity or recall) and the false-negatives (1 - specificity). The AUC is particularly useful when dealing with imbalanced datasets since it evaluates a models' performance across various threshold, providing insights into how well the model handles different class distributions.
- Confusion Matrix: is a table that is also used to evaluate the performance of binary classification models which provides a comprehensive view of how well a model is doing in terms of making correct and incorrect predictions. This table is a four-square table that is divided into:
    a. TP: Instances where the model correctly predicts the positive class.
    b. TN: Instances where the model correctly predicts the negative class.
    c. FP: Instances where the model incorrectly predicts the positive class.
    d. FN: Instances where the model incorrectly predicts the negative class.

To robustly assess the performance of the models, cross-validation was employed as a rigorous evaluation technique. This method involves systematically partitioning the dataset into multiple subsets, training the models on various combinations of these subsets, and assessing their performance on held-out performance.

3.1.6. Deployment

The last step of the CRISP-DM methodology is the Deployment phase, where the insights and models derived from the data analysis are put into practice.

In addressing the challenge of customer churn in this Portuguese bank, the Business Intelligence team will integrate the predictive model into the bank's operational framework. The developed model, that will previously be tuned based on the business objectives, will be ready to be deployed as a strategic tool to mitigate customer churn.

Given the business objectives of the customer churn prediction, the focus is on obtaining probabilistic predictions rather than deterministic outcomes, aligning with the need for nuanced risk assessment and decision-making.

The model was designed to be dynamically feed with the latest customer interactions, allowing it to adapt and refine predictions at the end of the day. Based on that, stakeholders – that could be from the Marketing Department and the Wealth Management, will be equipped with a predictive tool that will assess the probability of churn.

This phase is a bridge between analysis and action, where the insights gained from data analysis are translated into actionable solutions. It's a dynamic and interactive process that demands ongoing attention, adaption, and collaboration to drive meaningful business value.

4. RESULTS AND DISCUSSION

4.1. MACHINE LEARNING MODELS

An extensive exploration of various machine learning models for predicting customer churn is present in this chapter. It will be provided a comprehensive analysis of the 5 different algorithms, evaluating their performance based on evaluation metrics mentioned on the previous chapter. In Table 4.1 it is present a summary table with all the scores for each different model.

Table 4.1 – Model's scores summary table

Model	AUC	Accuracy	Recall	Precision	Specificity	F1-Score
Decision Tree	0,948291	0,947741	0,951764	0,943338	0,943786	0,947532
Extra Trees	0,994226	0,963463	0,967843	0,958850	0,959157	0,963325
Random Forest	0,990110	0,953167	0,965610	0,941433	0,940931	0,953368
XGBoost	0,995968	0,974646	0,973202	0,975599	0,976065	0,974399
Gradient Boosting	0,997081	0,983503	0,979008	0,987610	0,987923	0,983290

For each model it will be present visualizations from the evaluation metrics, such as the AUC and the Confusion Matrix.

4.1.1. Decision Trees

Decision Trees are a non-linear model that makes decisions by recursively splitting the data based on the most significance features. They are interpretable and can handle both numerical and categorical variables.

Using the *GridSearchCV*, the tuned hyperparameters are the following:

- Criterion: entropy
- Maximum Depth: 47

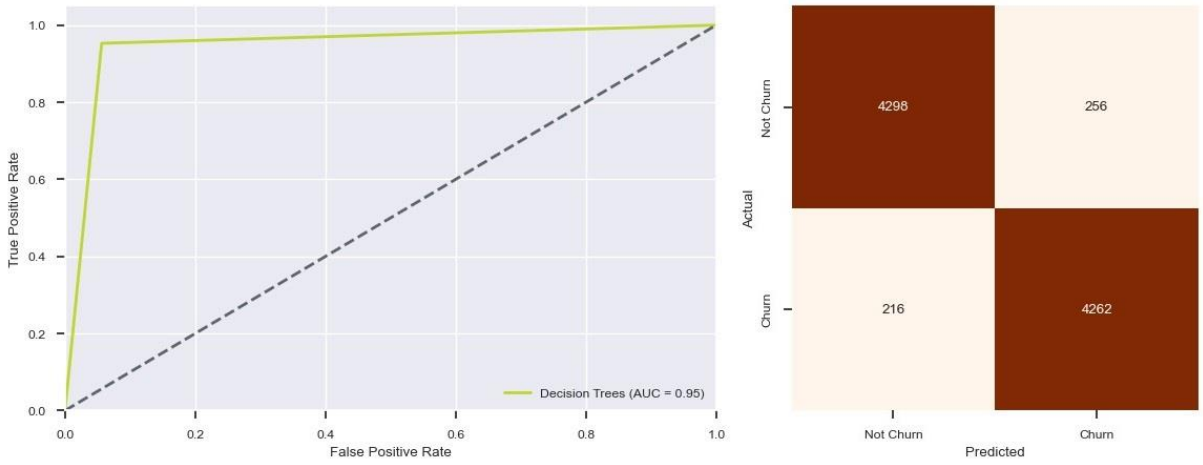


Figure 4.1 – Decision Tree's AUC and Confusion Matrix

Cross-validation scores (train): [0.94726644 0.93785467 0.94269896 0.95017301 0.9449058]
Cross-validation scores (test): [0.89042612 0.86607637 0.86212625 0.88095238 0.8709856]

Decision Trees display robust performance with a ROC of 95% and an accuracy of 95%. Notably, the model achieves high recall (95%) and specificity (95%). (Figure 4.1). Cross-validation results suggest consistent performance, though there is a slight drop in test accuracy.

4.1.2. Extra Trees

Extra Trees is an ensemble learning method that builds multiple decision trees and combines their predictions, introducing additional randomness in the tree-budling process.

Using the *RandomizedSearchCV*, the tuned hyperparameters are the following:

- Maximum depth: 30
- Minimum split samples: 4

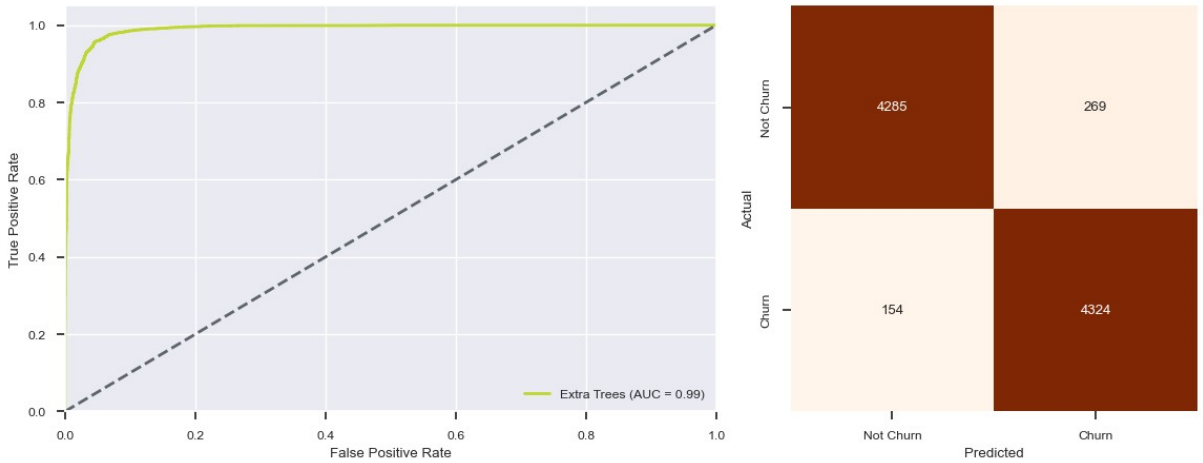


Figure 4.2 – Extra Trees' AUC and Confusion Matrix

Cross-validation scores (train): [0.95017301 0.95072664 0.94712803 0.95114187 0.9529346]
Cross-validation scores (test): [0.91311566 0.90149419 0.91306755 0.89922481 0.89368771]

Extra Trees stands out with exceptional performance, boasting a ROC of 99% and an accuracy of 98%. The model achieves high recall (98%) and precision (98%) (Figure 4.2). Cross-validation results affirm good generalization though there Is a slight decrease in test scores.

4.1.3. Random Forest

The Random Forest is an also an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the class for classification problems. It's robust, handles outliers well, and is less prone to overfitting.

Using the *GridSearchCV*, the tuned hyperparameters are the following:

- Maximum depth: 2
- Minimum split samples: 26

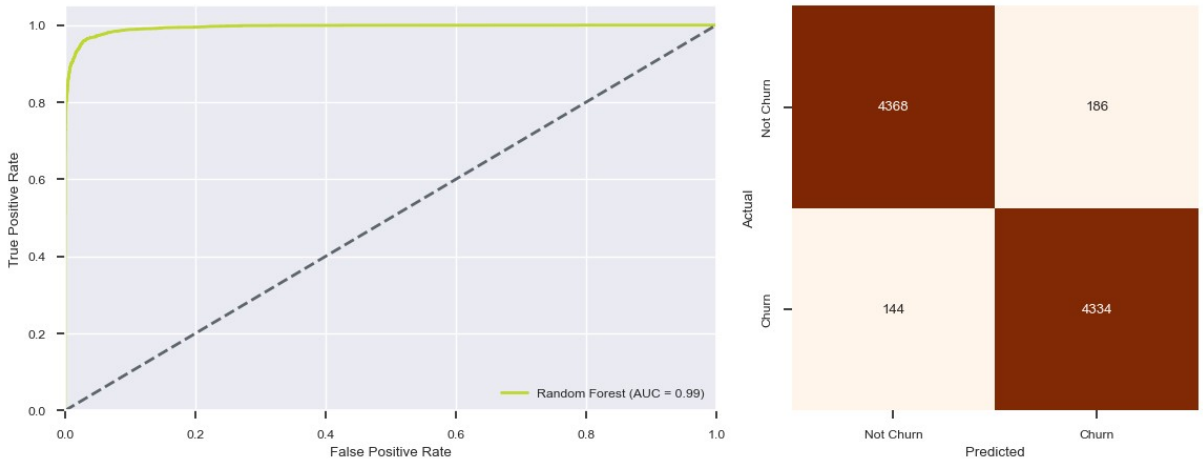


Figure 4.3 – Random Forest AUC and Confusion Matrix

Cross-validation scores (train): [0.96110727 0.96525952 0.96138408 0.96373702 0.9631782]
Cross-validation scores (test): [0.91145545 0.90924184 0.91638981 0.92192691 0.91638981]

Random Forest has a ROC of 99% and an accuracy of 96% The model achieves high recall (97%) and precision (96%), indicating strong predictive capabilities. Specificity is notable at 96% (Figure 4.3). Cross-validation underscores its robustness, maintaining high scores on both training and test sets.

4.1.4. XGBoost

The XGBoost is an ensemble learning method known for its speed and performance. It builds trees correcting the errors of the previous ones.

Using the *RandomizedSearchCV*, the tuned hyperparameters are the following:

- Maximum depth: 14
- Learning rate: 0,3

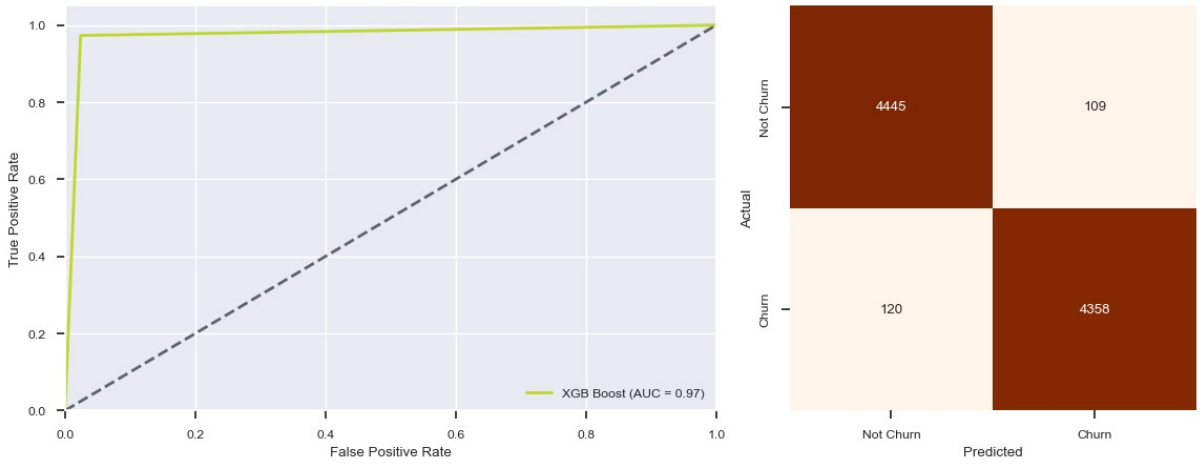


Figure 4.4 – XGBoost's AUC and Confusion Matrix

Cross-validation scores (train): [0.97121107 0.97411765 0.97453287 0.97273356 0.9725913]
Cross-validation scores (test): [0.9513005 0.93912562 0.94241417 0.95902547 0.94739756]

XGBoost is another model with great score, with an AUC of 99% and with accuracy of 98%, with balanced metrics in recall (98%) and precision (99%) (Figure 4.4). Cross-validation consistently supports its robustness and generalization.

4.1.5. Gradient Boosting

The Gradient Boosting is another ensemble method that builds trees sequentially, with each tree correcting the errors of the previous ones. It's effective but can be computationally expensive.

Using the *RandomizedSearchCV*, the tuned hyperparameters are the following:

- Maximum depth: 18
- Learning rate: 0,8

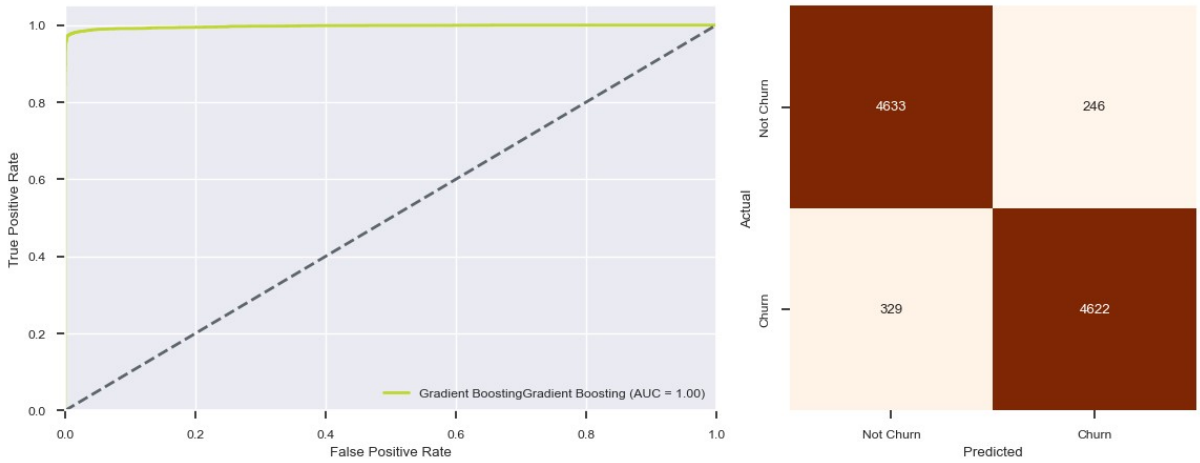


Figure 4.5 – Gradient Boosting's AUC and Confusion Matrix

Cross-validation scores (train): [0.98435986 0.98588235 0.98519031 0.98477509 0.9842192]
Cross-validation scores (test): [0.97122302 0.96513558 0.95791805 0.97674419 0.96179402]

Gradient Boosting emerge as the top-performing model with great scores. The model achieves a ROC of 99% and accuracy of 98%. High recall (98%) and precision (99%) emphasize its strong predictive capability (Figure 4.5). Cross-validation consistently supports its robustness and generalization.

4.2. DISCUSSION

In evaluating the predictive performance of the models, a set of metrics was employed. The models consistently demonstrated strong capabilities in capturing patterns within the data, as evidenced by high scores across these metrics. Notably, the cross-validation results further reinforced the robustness of the models, showing their ability to generalize well to diverse subsets of the dataset. Ensemble methods were strategically leveraged, contributing to enhanced model stability and predictive accuracy. These findings collectively underscore the efficacy of the models in addressing the complexities of the customer churn prediction problem.

Overall, based on the results, the conclusions for each model are the following:

- Decision Trees: Performs well with good balance between precision and recall, indicating a robust ability to identify both churned and not churned customers.
- Extra Trees: Demonstrates high accuracy and recall, making it a strong candidate for identifying churned customers.
- Random Forest: Shows robust performance with high accuracy and good balance between precision and recall.
- XGBoost: Performs like Random Forest, demonstrating high accuracy and good balance between precision and recall.

- Gradient Boosting: Shows good performance, specialty in precision, indicating a reliable ability to correctly identify churned customers.

The results suggest that the choice of the optimal model depends on the specific priorities and constrains of the business, emphasizing the importance of considering both false positives and false negatives in the context of customer churn prediction.

In the pursuit of an optimal model for predicting customer churn, the selection process was anchored in aligning with the critical priorities of the business. Recognizing the significance of both precision and recall in the context of customer churn, the decision was made to Favor ensemble methods, specifically Random Forest, as it exhibited a balance between these two metrics.

The emphasis on achieving high accuracy wile effectively identifying both churned and not churned customers drove this decision. Additionally, the consideration of business priorities underscored the suitability of ensemble methods like Random Forest, which provides a robust framework for making predictions that align closely with the nuanced requirements of churn prediction problem.

This strategy choice reflects a commitment to delivering a model that not only meets the quantitative metrics but also aligns with the overarching and strategic vision of the business.

4.3. POWER BI DASHBOARDS

Business Intelligence (BI) is a set of techniques and tools that enable organizations to collect, analyse, and interpret data to make informed decisions. BI plays a vital role in customer churn prediction, as it provides insights into customer behaviour and helps identify the factors that lead to churn. By leveraging BI tools, companies can build predictive models that accurately forecast customer churn and take preventive measures to retain customers.

Power BI is a business analytics service provided by Microsoft that offers users to visualize and analyse data from different sources in a user-friendly format, facilitating decision-making. With its powerful features, Power BI is an invaluable tool for analysing and presenting data, particularly in the domain of customer churn prediction.

One of the key advantages of Power BI is its inclusion of built-in machine learning models, which can be used to forecast customer churn with accuracy. By leveraging Power BI's capabilities, businesses can gain actionable insights from their data, empowering them to make data-driven decisions and devise effective churn prevention strategies.

Aligning with this, a Power BI dashboard was created that will leverage the insights derived from the deployment of the machine learning model, developed, and trained to predict the probability of a customer to churn.

The first page of the dashboard (Figure 4.6) serves as a holistic overview of the customer, providing stakeholders with key metrics and visualizations essential for understanding the status of customer relationships. These visualizations empower decision-makers to understand the trends in customer behaviour, offering a quick snapshot of the overall health of the customer base.

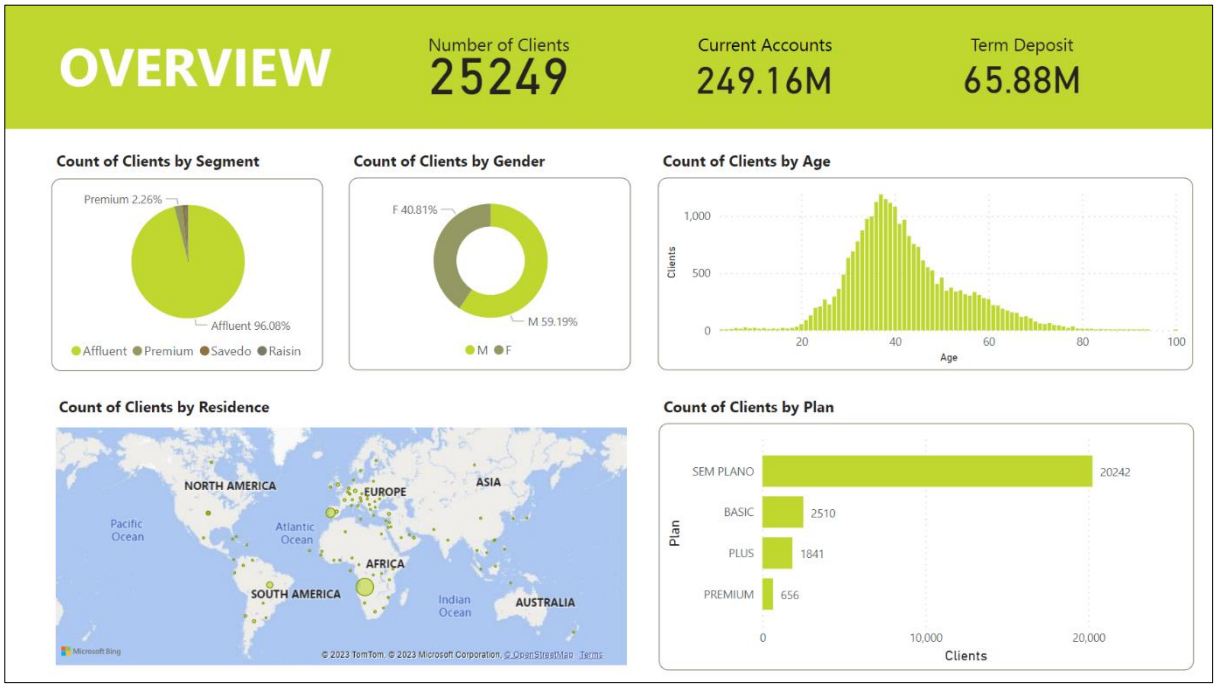
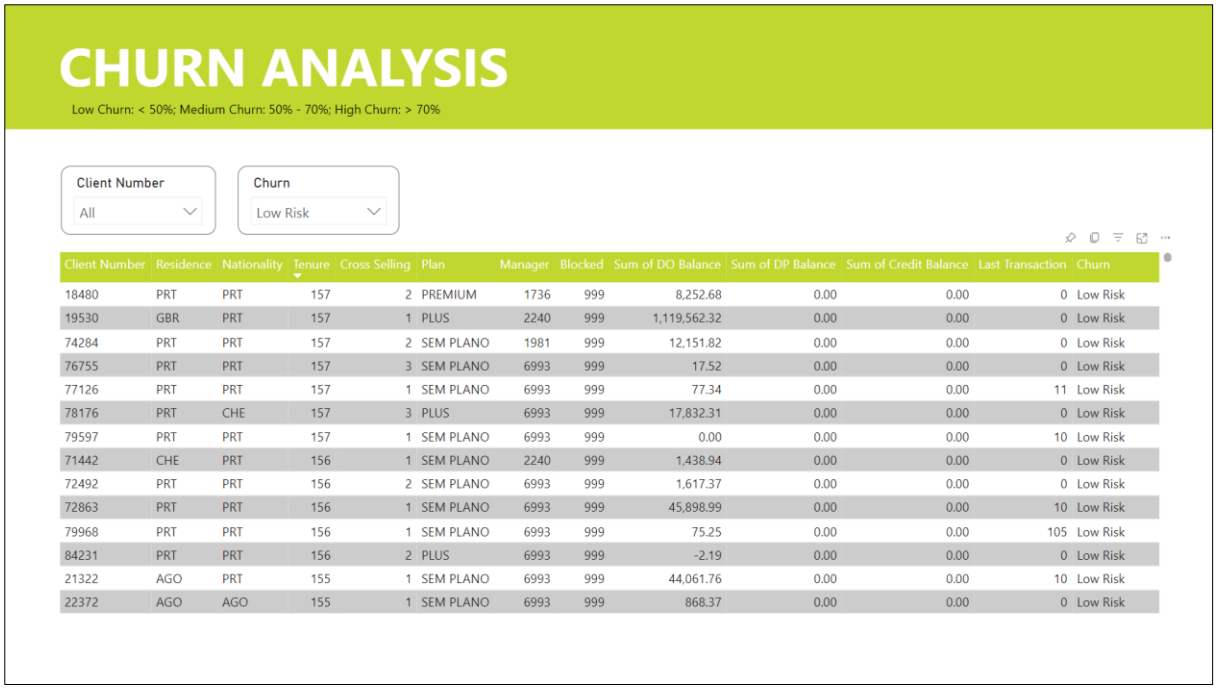


Figure 4.6 – Overview's page of the Power BI dashboard

The second page (Figure 4.7) focuses into a granular detail of individual clients, presenting a comprehensive table with essential information. However, this page focuses on different kind of churn labels. These labels will serve as actionable insights derived from the machine learning model. The model calculates the likelihood of a customer churning and categorize them into 3 distinct labels: Low Risk (lower than 50%), Intermediate Risk (between 50% and 70%), and High Risk (higher than 70%). The incorporation of these labels not only helps stakeholders but also facilitates a proactive approach to customer retention in order to prioritize their efforts based on the risk levels.



Client Number	Residence	Nationality	Tenure	Cross Selling	Plan	Manager	Blocked	Sum of DO Balance	Sum of DP Balance	Sum of Credit Balance	Last Transaction	Churn
18480	PRT	PRT	157	2	PREMIUM	1736	999	8,252.68	0.00	0.00	0	Low Risk
19530	GBR	PRT	157	1	PLUS	2240	999	1,119,562.32	0.00	0.00	0	Low Risk
74284	PRT	PRT	157	2	SEM PLANO	1981	999	12,151.82	0.00	0.00	0	Low Risk
76755	PRT	PRT	157	3	SEM PLANO	6993	999	17.52	0.00	0.00	0	Low Risk
77126	PRT	PRT	157	1	SEM PLANO	6993	999	77.34	0.00	0.00	11	Low Risk
78176	PRT	CHE	157	3	PLUS	6993	999	17,832.31	0.00	0.00	0	Low Risk
79597	PRT	PRT	157	1	SEM PLANO	6993	999	0.00	0.00	0.00	10	Low Risk
71442	CHE	PRT	156	1	SEM PLANO	2240	999	1,438.94	0.00	0.00	0	Low Risk
72492	PRT	PRT	156	2	SEM PLANO	6993	999	1,617.37	0.00	0.00	0	Low Risk
72863	PRT	PRT	156	1	SEM PLANO	6993	999	45,898.99	0.00	0.00	10	Low Risk
79968	PRT	PRT	156	1	SEM PLANO	6993	999	75.25	0.00	0.00	105	Low Risk
84231	PRT	PRT	156	2	PLUS	6993	999	-2.19	0.00	0.00	0	Low Risk
21322	AGO	PRT	155	1	SEM PLANO	6993	999	44,061.76	0.00	0.00	10	Low Risk
22372	AGO	AGO	155	1	SEM PLANO	6993	999	868.37	0.00	0.00	0	Low Risk

Figure 4.7 – Churn Analysis' page of the Power BI dashboard

Overall, customer churn prediction is a complex problem that requires the use of machine learning algorithms and data visualization tools such as Power BI. Understanding the factors that contribute to customer churn and using appropriate evaluation metrics can help businesses predict and minimize customer churn, thereby improving their revenue and profitability.

5. CONCLUSIONS AND FUTURE WORKS

5.1. CONCLUSIONS

In the pursuit of developing an effective customer churn prediction model for a Portuguese bank, this study navigated through various challenges and insights. The initial dataset, characterized by a notable class imbalance, prompted the use of the SMOTE technique to enhance the model's sensitivity to instances of churn. Furthermore, correlation and data normalization were also performed to avoid overfitting. While the conventional accuracy metric provided a fundamental evaluation, the emphasis on recall, precision, F-1 score and specificity brought a better understanding on the performance evaluation.

Among the array of models explored, including Decision Trees, Extra Trees, Random Forest, the algorithms showed robust capabilities in identifying potential churn cases. Not only the decision tree-based models showcased a good balance between interpretability and predictive accuracy, such as Random Forest and XGBoost, but also the gradient boost models showed the capability of getting good results. The models not only provide a mechanism for identifying customers at risk of churn but also offer valuable insights for crafting proactive retention strategies. The interpretability versus performance trade-off in model selection highlights the importance of aligning the model's complexity with the need for comprehensive insights.

5.2. LIMITATIONS

However, it is crucial to acknowledge the limitations of the models. While the current features set has proven valuable, there remains untapped potential in leveraging additional features. For instance, exploring the destinations of transfers could provide insights into whether funds are moving to accounts within the same person. This distinction could be fundamental in understanding a client's intention to exit the bank, offering a nuanced perspective on potential churn.

Understanding the relationship between customers associated with the same account presents another avenue for improvement. In cases where a primary account holder closes their account, investigating the influence on accounts held by secondary holders could provide a more comprehensive view of potential influencing behaviors. This insight might be crucial in predicting the likelihood of a client closing multiple accounts following the primary account closure.

A notable challenge encountered during this study was the presence of a significant number of client accounts that exhibited a long period of inactivity. While these clients may not be immediate candidates for account closure, their lack of financial movement poses a different dilemma. Such inactive accounts may not be generating value for the Bank, raising questions

about strategies to re-engage these clients or whether the resources allocated to maintaining these accounts align with the overall business objectives.

Addressing these challenges involves a delicate balance between encouraging client activity and avoiding undue pressure, which could lead to unwanted consequences. Future research should explore strategies for reactivating dormant accounts or optimizing resource allocation to ensure that efforts align with both customer satisfaction and the bank's financial objectives.

5.3. FUTURE WORKS

For future developments, there are a spectrum of opportunities for refinement and expansion:

- Model Refinement: fine-tuning the hyperparameters of the ensemble models can further elevate the predictive performance of the models.
- Model Deployment and Monitoring: implementing real-time monitoring systems and establishing a feedback loop from implemented retention strategies to the model could also be a future work, since it would ensure the adaptability of the model when finding patterns.
- Timeseries model: instead of using a specific reference date, it would be interesting to adapt this binary problem into a timeseries forecast problem, since it would be looking at patterns over time which could lead to more personalized strategies and specific action to retain customers.

In conclusion, while celebrating the accomplishment of the current model, the journey of refinement and responsible deployment continues. The outlined future works aim to address current limitations and contribute to continuous improvement of customer churn prediction for the banking sector.

BIBLIOGRAPHICAL REFERENCES

Cabeças, A. (2014). *Comportamento dos clientes da banca portuguesa em período de crise económica* [Universidade Autónoma de Lisboa]. http://hdl.handle.net/11144/472

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*. DaimlerChrysler.

Cohen, D., Gan, C., Yong, H. H. A., & Chong, E. (2007). Customer Retention by Banks in New Zealand. *Banks and Bank Systems*, *2*(1), 40–55. Customer Retention by Banks in New Zealand

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36. https://doi.org/10.1016/j.dss.2016.11.007

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications*, *34*(14), 11751–11768. https://doi.org/10.1007/s00521-022-07067-x

Ejeu, E. (2021). *Analysis of customer churn in Ugandan commercial banks: a case study of Stanbic Bank* [College of Business and Management Sciences]. http://makir.mak.ac.ug/handle/10570/8350

Hadden, J. (2008). *A Customer Profiling Methodology for Churn Prediction* [School of Applied Sciences]. https://dspace.lib.cranfield.ac.uk/handle/1826/3508

Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). A Multi-Layer Perceptron Approach for Customer Churn Prediction. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(7), 213–222. https://doi.org/10.14257/ijmue.2015.10.7.22

Oluwatoyin, A. M., Misra, S., Wejin, J., Gautam, A., Behera, R. K., & Ahuja, R. (2023). *Customer Churn Prediction in Banking Industry Using Power Bi* (pp. 767–774). https://doi.org/10.1007/978-981-19-1142-2_60

Oyeniyi, A. O., & Adeyemo, A. B. (2015). Customer Churn Analysis In Banking Sector Using Data Mining Techniques. *African Journal of Computing & ICT African Journal of Computing & ICT Reference Format: A. O. Oyeniyi & A.B. Adeyemo*, *8*(3), 165–174. https://www.semanticscholar.org/paper/Customer-Churn-Analysis-In-Banking-Sector-Using-Oyeniyi-Adeyemo/0fd0067775eef2aa52547229c17e06128fdc0633

Rosa, N. B. da C. (2018). *Gauging and Foreseeing Customer Churn in the Banking Industry: A Neural Network Approach* [NOVA IMS]. https://run.unl.pt/handle/10362/71584

Santiago da Cunha, S. S. (2023). *Customer Churn Prediction in the Banking Industry* [NOVA Information Management School]. http://hdl.handle.net/10362/149182

Sashikala, P. (2015). Customer Retention in the Indian Banking Sector -Managing Customer Churn. *International Journal of Applied Engineering Research*, *10*, 35567–35584. https://doi.org/10.13140/RG.2.2.26244.27522

Tianyuan, Z., & Moro, S. (2021). Research Trends in Customer Churn Prediction: A Data Mining Approach. In *Advances in Intelligent Systems and Computing: Vol. 1365 AIST* (pp. 227–237). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-72657-7_22

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Warikoo, A. (2022, March 29). *Banking Consumer Behavior: The tale of fragmented customers*. https://www.linkedin.com/pulse/banking-consumer-behavior-tale-fragmented-customers-akshay-warikoo?trk=articles_directory