UNIVERSIDADE CATÓLICA PORTUGUESA

The applicability of Business Analytics and Business Intelligence techniques in the Wealth Management Sector

Case Study at Banco Carregosa

Diogo Tavares Castro Silva

Universidade Católica Portuguesa

2023

UNIVERSIDADE CATÓLICA PORTUGUESA

The applicability of Business Analytics and Business Intelligence techniques in the Wealth Management Sector

Case Study at Banco Carregosa

Final Project submitted as an Internship Report to

Universidade Católica Portuguesa

for the Master degree of Management with specialization in Business Analytics

by

Diogo Tavares Castro Silva

Under guidance of

Maria Alice Trindade

Católica Porto Business School, Universidade Católica Portuguesa

April, 2023

Acknowledgements

I would like to express my gratitude and appreciation towards some people, without whom I would not be in this position and with this opportunity.

First of all, I am extremely grateful to my supervisor, professor Maria Alice Trindade, for the cooperation and support demonstrated during the internship and for always being available in providing guidance.

I also want to express my gratitude to Banco Carregosa, especially to the Marketing Department, for the friendly welcome and assistance throughout the internship.

I am deeply indebted to my parents and my sister, for their unwavering confidence in me and for providing me with the opportunity to pursue this Master’s degree.

I am also extremely grateful to Diana for all the moments and the support given since the beginning.

Finally, I also want to thank all my close friends, for all the companionship demonstrated throughout the years.

Resumo

O Banco Carregosa é uma instituição bancária portuguesa especializada em investimentos e gestão de património. O Banco Carregosa prospera no relacionamento com os seus clientes, oferecendo um serviço personalizado, com base no princípio de que cada cliente é único, ou seja, cada cliente tem metas financeiras únicas, preferências de serviço e níveis de tolerância ao risco únicos.

No entanto, uma vez que o setor bancário é altamente competitivo, impulsionado pelas mudanças nas necessidades e preferências dos clientes, avanços na tecnologia e mudanças no quadro regulatório, o Banco Carregosa procura uma estratégia para superar essa feroz rivalidade. A análise de dados surgiu como uma ferramenta crucial para as instituições obterem uma vantagem competitiva, adquirindo conhecimentos sobre o comportamento do cliente, o que lhes permite fornecer um serviço mais personalizado e direcionado.

Esta dissertação demonstra os benefícios da aplicação de tecnologias de Business Intelligence (BI) e Business Analytics (BA) no setor de Gestão de Património e, especificamente, no Banco Carregosa. Esta dissertação foca-se em fornecer ao Banco Carregosa tecnologias fundamentais para superar a rivalidade, em vez de se concentrar nos resultados dos modelos empregados. As técnicas utilizadas foram Market Basket Analysis e Churn Analysis, juntamente com o desenvolvimento de dashboards com informações relevantes para o processo de tomada de decisão.

No final, concluímos que o BI e o BA forneceram ao Banco Carregosa vários benefícios, permitindo uma compreensão mais profunda dos seus dados e melhorias na eficiência das suas operações.

Palavras-chave: *Business Intelligence, Business Analytics, Machine Learning,* Visualização de Dados, *Private Banking, Wealth Management*

Palavras: 9541

Abstract

Banco Carregosa is a Portuguese banking institution, specialized in investments and wealth management. Banco Carregosa thrives in the relationships with customers, by offering a personalized service, based on the principle that each client is unique, i. e., each client has unique financial goals, service preferences and risk tolerance levels.

However, since the banking sector is highly competitive, driven by changing customer needs and preferences, advances in technology and changes in the regulatory framework, Banco Carregosa is looking for a strategy to overcome this fierce rivalry. Data analytics has emerged as a crucial tool for institutions to gain a competitive edge, by acquiring insights into customer behaviour, which allows them to provide a more personalized and targeted service.

This dissertation demonstrates the benefits of applying Business Intelligence (BI) and Business Analytics (BA) technologies in the Wealth Management sector and, specifically, in Banco Carregosa. This dissertation is more concerned with providing Banco Carregosa with fundamental technologies to overcome rivalry, instead of focusing on the outcomes of the models employed. The techniques applied were Market Basket Analysis and Churn Analysis, alongside the development of dashboards with relevant information for the decision-making process.

In the end, we concluded that BI and BA provided Banco Carregosa with several benefits by enabling a deeper understanding of their data and improvements in the efficiency of their operations.

Keywords: Business Intelligence, Business Analytics, Machine Learning, Data Visualization, Private Banking, Wealth Management

Index

Table Index

Figure Index

First Chapter

1.1 Introduction

This dissertation was developed under an internship at Banco Carregosa, which proposed an improvement on the treatment of its data, by implementing BI and BA techniques in the organizational context to assess the applicability and results of these tools in a Wealth Management institution.

Companies are being flooded with a rising amount of data and information and, for businesses to succeed, knowledge extraction from this data is crucial. Data is already being collected at a rate that has never been seen before, and this tendency will only continue as technology to capture, store and analyze data is developed (Gantz et al., 2011). All sectors of activity must adapt to these contemporary and ongoing developments, and the banking industry is not an exception. Banks have at their disposal large datasets from internal and external sources and, to thrive in a highly competitive market, they must become closer to their customers, understand and anticipate their demands, and proactively position their services (Lacković et al., 2020). The importance of applying BI and BA tools in banking institutions is well known, which is shown in the vast amount of material available on the subject. There is, however, a gap in the literature regarding the significance of these technologies in the Wealth Management sector, which accounts for the lack of information about this subject.

Banco Carregosa is a banking institution, specialized in investments and wealth management. The services of Banco Carregosa are divided into two business units: Private Banking and Affluent Banking. The first is related to clients with larger wealth portfolios, while the latter pertains to the commercial segment, focused on savings and investments. Banco Carregosa is promoting the expansion of its data analytics field and has thus proposed this internship to comprehend the applicability and benefits of BI and BA tools.

The purpose of this report is to illustrate the effects of applying BI and BA tools to the performance of Banco Carregosa and to evaluate the effectiveness of using such tools in the Wealth Management sector. The main goal of this dissertation is to supply Banco Carregosa with models and procedures to assist in the decision-making process, as opposed to demonstrating the results found.

The data in the subsequent analyses were provided by Banco Carregosa. However, to maintain confidentiality, the numbers displayed in this paper do not correspond to reality.

1.1.1 Report Restructuring Objectives

The main objective of this report is to demonstrate the benefits of using BI and BA technologies in a Wealth Management institution. However, instead of concentrating on the outcomes of the analyses, this dissertation focuses on providing Banco Carregosa with better analytical tools to enhance the analytical process and improve decision-making. This procedure is important for Banco Carregosa since it enables the use of these instruments with the appropriate adjustments in terms of inputs and outputs, following the desired outcomes.

Consequently, to respond to the research question "what is the applicability of Business Intelligence and Business Analytics tools in a Wealth Management Sector?", we created models and algorithms to assist Banco Carregosa's activities.

1.1.2 Research Methodology

The research methodology followed is Action Research, which can be described as a collaborative problem-solving relationship where the action researcher and the participants actively work together in diagnosing a problem and developing solutions (Coghlan, 2003). The basic idea is that the action researcher collaborates directly with people who are affected by an issue to solve it using a scientific method (Coghlan, 2003).

The Action Research methodology can be described as an action-reflection cycle (Dick, 2015). This methodology approaches real-world problems in a participatory and collaborative way to improve knowledge and action. The participants begin by observing and reflecting on a question/problem and creating an action plan (McNiff et al., 2011). Then, they will observe and evaluate their findings and modify their study according to their objectives. This can lead to a new direction of analysis, which restarts the improvement process (McNiff et al., 2011).

In the context of Banco Carregosa, before each investigation, the team discussed a crucial question, for example, "is there any particular set of products/services that customers buy jointly?". Considering these relationships found, in case there was a client that only acquired one of the products, Banco Carregosa could present the remaining products, which offers great potential for cross-selling. Following that, the data was collected and cleaned in several iterations of experimentation, until reaching analysis with valuable outcomes, at which point it underwent further analysis to gather new insights and reflect on new questions.

1.1.3 Structure

The dissertation is divided into six chapters. The first chapter provides context for the dissertation, a brief description of Banco Carregosa, the objectives of the internship, the literature gap addressed and the methodology used.

Chapter number two provides a deeper characterization of Banco Carregosa, its evolution since being founded and a brief description of the products/services offered.

The third chapter provides a theoretical framework for the dissertation. This chapter covers the main concepts adjacent to the Data Analytics area, such as, BI and BA, Big Data and Data Visualization.

The fourth chapter refers to the analytical techniques employed during the internship that will serve as a base for the analytical process at Banco Carregosa.

The fifth chapter is divided into six sub-chapters, which refer to the different stages and activities performed during the internship. This chapter includes the results obtained from the analysis.

The last chapter discriminates the conclusions drawn from the project and provides recommendations for the future work of Banco Carregosa, based on the limitations found during the internship.

Second Chapter

2.1 Banco Carregosa

Banco Carregosa is a Portuguese banking institution established in 1833, which specializes in investments and wealth and portfolio management. Initially, Banco Carregosa was founded as a brokerage firm called L. J. Carregosa, focused on trading and offering currency and exchange services. However, in 1994, L.J. Carregosa was established as a business with the designation L. J. Carregosa – Sociedade Corretora S.A.. In 2008, with the approval of Banco de Portugal, this firm was transformed into a banking institution.

Banco Carregosa is divided into two different activity segments: Private and Affluent Banking. As referred in the previous chapter, Private Banking is tied with clients with greater wealth portfolios, while Affluent Banking is related to the commercial segment. For all clients and their different investor profiles, Banco Carregosa offers a wide variety of financial products/services and wealth management solutions, as well as its own trading platform, GoBulling, created in 2007, which allows clients to manage their assets online.

Banco Carregosa is distinguished by its thriving relationship with their customers, by offering a specialized and unique service to customers, which is based on their preferences, such as risk tolerance level and financial objectives.

Third Chapter

3.1 Literature Review

People frequently believe that what they observe is a complete truth (Harford, 2020). However, we cannot understand the world if we do not completely comprehend the statistics behind it (Harford, 2020). We must dive deeper to try to reach this truth (Harford, 2020). Statistics is the perfect tool in assisting individuals in better grasping the complex world around them (Harford, 2020).

3.1.1 Business Intelligence and Business Analytics

Hal Varian, Chief Economist at Google (Chen et al., 2012, p.1166), once said that: *"So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on."*. At a period when businesses are dealing with fierce competition, producing comparable products and employing similar technologies, BI and BA tools are playing a differentiation role (Davenport, 2006).

BI and BA systems are analytical tools that combine the gathering, analysis and storage of data, to acquire useful and complex information to better comprehend the firm's capabilities, the current state of the market and the future directions to pursue, to provide a better understanding of the surrounding environment to decision makers (Negash, 2004).

There is not a clear distinction between BI and BA definitions and many authors have only defined them slightly differently (Turban, 2014), since they are related towards achieving similar goals, offering insights into the data available and defining future actions. However, whereas BI is more focused on descriptive analysis, by identifying patterns and relationships that occurred with historical data, BA is more focused on a predictive analysis by using sophisticated statistical techniques coming from data mining and machine learning tools to analyze current data and generate predictions (Calzon, 2022).

Still, BI and BA tools are not beneficial by themselves (Harford, 2020), it is essential to interpret the results, have a critical mindset and extensive knowledge of the subject, to transform this information into valuable knowledge, which can enhance the decision-making process and improve the company's performance (Ćurko et al., 2007).

Since the 1950s, scientists studying artificial intelligence had been using the term "intelligence", however, only during the 1990s, did the term "Business Intelligence" start to emerge in IT and business industries (Davenport, 2006). The concept of "Business Analytics", which is considered the primary analytical element in BI, was introduced in the late 2000s (Davenport, 2006). Currently, these technological tools are vital for the success of any institution. According to a survey conducted by Bloomberg Businessweek (2011), it was discovered that 97% of companies whose annual revenues exceeded 100$ million employed some form of BI and BA techniques (Chen et al., 2012).

Since the development of the Internet and the large-scale generation of data, the opportunities for data collection and analytical research have exponentially grown since web systems enable businesses to interact directly with customers online (Chen et al., 2012). By tracking the client's online actions and surfing through specific IP addresses, cookies and server logs, institutions can compile data about the client's behaviours, interests and purchase habits (Chen et al., 2012). Using this data and the appropriate BI tools, businesses may improve the placement of their products on their websites and their recommendations to customers (Chen et al., 2012). Thus, rather than being a classic business-to-customer relationship, the market is increasingly becoming a "*conversation*" between the company and the client (Lusch et al., 2010).

This global information revolution has impacted all industries, and the financial sector is no exception: banking institutions need to fully utilize the vast amount of information about clients and transactions held on their information systems, to obtain a competitive advantage over their rivals. However, if this analysis is performed without adequate tools, this process can quickly become expensive and yield insufficient results (Ćurko et al., 2007).

For instance, the success of banking institutions is closely tied to their Customer Relationship Management (CRM) (Kaur et al., 2013) and the efficiency of their operations, thus exploiting BI technologies is crucial in an endless race towards achieving a competitive advantage in the market (Ćurko et al., 2007). Banks have a vast amount of data gathered and they must understand the customers' behaviours, to efficiently satisfy their needs and expectations and prevent churn, this is, to prevent clients from ending their relationships, which can lead to a significant decrease in revenues (Coşer et al., 2020a). These technologies, for example, can also effectively help banks maintain constant control over activities such as fraud detection and risk, customer and product management (Ćurko et al., 2007).

3.1.2. Big Data

The concept of "Big Data" has brought little consensus regarding its origin, size and characteristics (Gandomi et al., 2015). It is argued that Big Data is a concept which originated in the mid-1990s (Diebold et al., 2012), however, the term has only become widespread recently. At the present, data is being gathered at an unprecedented rate: before 2011, 1.8 Zettabytes (ZB) of data were created globally, according to research by the International Data Corporation (IDC).

However, over the subsequent five years, this number has increased by nine times, and this tendency will only continue as storage capacities rise, making it possible to store more data (Gantz et al., 2011). Consequently, there has been little agreement regarding the designation of how large a data set must be to be considered “Big Data”, since the developments of technologies are facilitating the generation of data and since two similar-sized sets of data may require different management approaches based on their characteristics (Gandomi et al., 2015).

Originally, Big Data was a term used to describe enormous data sets that conventional databases were unable to store and interpret (Lacković et al., 2020). However, several authors have provided their perspectives on the features of Big Data: Laney (2001) suggests that Volume, Variety and Velocity (Three V’s) are the dimensions of these data sets. Furthermore, according to Gantz et al. (2011), “big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis”, which summarizes Big Data as Four V’s, being Volume, Variety, Velocity and Value. Özköse et al. (2015) defend the existence of five V’s, adding the term Veracity. Despite all the ambiguity regarding the characteristics of Big Data, the Three V’s appear to have emerged as a consistent foundation for most authors (Chen et al., 2012; Kwon et al., 2014).

Collecting and managing a large volume of data is insignificant unless we address the challenge of extracting meaningful value from the sets of data (Bakshi, 2012). Therefore, “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation” (Gartner, 2014). As an example of the positive impact of Big Data, in a study conducted by McKinsey, it was concluded that its analysis has the potential value of reducing the U.S. expenditure on healthcare by over 8% (Manyika et al., 2011).

However, Big Data encounters significant challenges in extracting knowledge from large datasets, since these are rapidly being generated and can be of various types (Lacković et al., 2020). Traditional management techniques and systems are based on the relational database management system (RDBMS), which only applies to structured data, which refers to the tabular data found in spreadsheets or relational databases (Chen et al., 2014). This type of data only constitutes 5% of all existing data (Cukier, 2010), with the other 95% being composed of unstructured data, which can be in the form of audio, video, images or unstructured text.

3.1.3. Data Visualization

In today's data-driven corporate environment, having ongoing access to analytical information on business activities is crucial for both small and large corporations (Orlovskyi et al., 2020). Industries have an abundant quantity of information regarding their activities, generated from various heterogeneous sources. Thus, organizing and presenting this information succinctly becomes a difficult and complex process (Dabbebi et al., 2017).

Data visualization tools have been widely used to address this situation and enhance the decision-making process. By providing a thorough overview of large volumes of data in a visually appealing format, data visualization technologies play a role in connecting large databases with a human vision system (Qin et al., 2020), to produce novel insights for the company's performance (Yu-Wen et al., 2022), allowing organizations to bridge the gap between data and decision making.

Dashboards are highly visual interfaces, which are built in an approachable, comprehensible and simple format, to help users in understanding the underlying data behind the graphs (Muskan et al., 2022). It is unlikely that a pre-built dashboard could satisfy all of the requirements of a complex organization (Brath et al, 2004), thus a deep comprehension of the raw data and the user's context is required to ensure that the information is effectively communicated and that noise is minimized. So, it is essential to gather knowledge of the raw data, formulate a research question, clean the data and then choose the appropriate visualization (Muskan et al., 2022).

Microsoft's Power BI tool is a useful resource for creating visualizations and thus improving business decision-making. Power BI is an integrated set of software that turns unrelated data sources into cohesive, visually immersive and interactive insights (Microsoft, 2023). Power BI provides a tool to establish connections between different data sources, enables data cleaning through edit queries, new metrics or the addition of new columns to databases and integrates filtering options for deeper user analysis (Bhargava et al., 2018). Power BI provides a wide variety of visuals, the official Microsoft visuals which are built-in in the installation package and the custom visuals developed by third-party users (Bhargava et al., 2018).

Fourth Chapter

4.1 Analytical Techniques

The present chapter focuses on the BI and BA techniques that were applied during the internship at Banco Carregosa. The main focus is to provide Banco Carregosa with effective analytical methods, to enhance their processes by improving productivity and improving the veracity of their insights.

4.1.1 Association

Association techniques are useful for finding relationships in large databases, by discovering patterns and correlations between items, which enables the identification of cross-selling opportunities (Cavique, 2007).

Market Basket Analysis (MBA), also known as association rule mining, is a data mining technique that originated in the marketing field, that aims to provide information regarding the purchase behaviour of the buyer (Kaur et al., 2016). Initially, this technique was used to understand the client's buying patterns in supermarkets, and to get a deeper knowledge about which products were bought together, i. e., which goods would be placed together in the supermarket "basket" (Aguinis et al., 2013). MBA allowed the supermarkets to make decisions regarding the placement of the products on their shelves, thereby increasing the likelihood of customers finding both products and buying them (Chen et al., 2005; Russell et al., 2000).

More recently, MBA has been employed in other industries, including the banking sector, by identifying association rules between groups of products/services bought frequently together and assessing the extent to which they co-occur. The fact that an MBA enables researchers to evaluate the presence of relationships by adopting an inductive method of theorizing is one of the main reasons for its growing acceptance across scientific domains (Locke, 2007). However, some argue that this approach is currently being underutilized in management research (Shepherd et al., 2011).

No method can overcome the challenges caused by errors in the data collection and data entry phases (Aguinis et al., 2013). Despite this, MBA provides a level of flexibility while exploiting an existing dataset, which allows the analysis of data considered "unusable" (Mcdonald et al., 2000; Roth et al., 1999). Regarding the problem of missing values, MBA considers that missing data indicates that no option was selected, as opposed to being caused by errors, which means that missing responses are considerably significant since the association rule created is the following: the presence of product X predicts the absence of product Y (Aguinis et al., 2013). Concerning the presence of outliers, MBA association rules suffer less influence compared to traditional data-analytic approaches, since outliers are infrequent occurrences (He et al., 2004).

In an MBA, three indexes are used to comprehend the relevance of the association rule found – lift, support and confidence (Berry et al., 2004). The lift measurement is used to characterize the correlation between variables (Aguinis et al., 2013): if the lift value is greater than 1, this means that the variables are positively correlated, being the presence of X associated with the presence of Y. The support measurement is related to the probability of both variables co-occurring (Aguinis et al., 2013). However, in the case of having an extremely large data set, the usefulness of this measure decreases because there are numerous alternative combinations (Cohen et al., 2001). Finally, the confidence measure explains the probability of the consequent being selected if the antecedent has already been chosen (Aguinis et al., 2013). This measurement is more capable of identifying the nature of the association and it even outperforms the support measure because it remains effective with large datasets (Aguinis et al., 2013).

Even though these indexes are fundamental for the analysis of association rules, the interpretation of their values is context-dependent (Aguinis et al., 2010).

The rules presented are composed of an antecedent or a left-hand side (LHS) which leads to a consequent or a right-hand side (RHS). The example in table 1 has one rule, which refers that 80% of clients that buy product X also buy product Y, which indicates that both products are correlated and present an opportunity for cross-selling.

LHS	RHS	Support	Confidence	Lift
X	Y	0,2	0,8	1,3

Table 1 - Association rule example

4.1.2 Classification

Statistical and data mining techniques are useful to create prediction models, by identifying relationships in the data and forecasting behaviour by fitting a model based on the existing data (Sharma et al., 2011).

According to Han et al. (2012), predictive analysis can be undertaken via regression and classification techniques. While the Regression algorithm is responsible for predicting the output in continuous values, the Classification algorithm is a supervised machine-learning technique, responsible for categorical values.

Logistic Regression is a supervised machine-learning technique, used for solving classification problems. This tool predicts a categorical dependent variable, which is binary (can only assume the value 1 or 0), using a given set of independent variables.

Coşer et al. (2020) and Sharma et al. (2011) applied Logistic Regression models when predicting customer churn in banking and cellular network services, respectively, given that this model produced positive performance values. In both types of research mentioned, the authors used socio-demographic independent variables, such as age, gender, geography, and other activity-related information.

Creating models to explain and predict customer churn is one of the most useful applications of these techniques (Keramati et al., 2016). Customer churn is a term used in the cellular mobile telecom service industry to refer to a customer that switches from service providers (Sharma et al., 2011). In the banking industry, customer churn refers to the clients that ended their activities with the bank, i. e., clients that closed their respective accounts.

Retaining current customers is a crucial strategy for survival within competitive markets, given the significance of customers as the most valuable assets of companies (Keramati et al., 2016). Customer churn represents a fundamental problem in the competitive environment of banking institutions since the costs associated with recruiting new customers are higher than the costs of sustaining existing clients (Athanassopoulos, 2000). Moreover, existing clients typically provide higher long-term profits (Verbeke et al., 2011). Nie et al. (2006) claim that banks can boost profitability by about 85% by increasing customer retention by 5%. Thus, ensuring customer loyalty is crucial in banking institutions.

Fifth Chapter

5.1 Activities Performed

Banco Carregosa is promoting the development of its data analytics field, to thrive in a competitive banking sector. The techniques described in the previous chapter were applied, to analyse the positive benefits of BI and BA in the Wealth Management sector. The development of financial and analytical dashboards will improve how Banco Carregosa's information is displayed, to expedite the decision-making process.

Association algorithms were used to discover patterns and cross-selling opportunities. Information about which products are often purchased together is vital for Banco Carregosa to improve targeting and provide better offers to specific clients. Following that, to analyse customer loyalty and the probability of churn, we used Classification techniques, more specifically a Logistic Regression model, to extract this information from the dataset. For both these analyses, the appropriate codes were created using RStudio.

The techniques and models developed can be seen in figure 1.

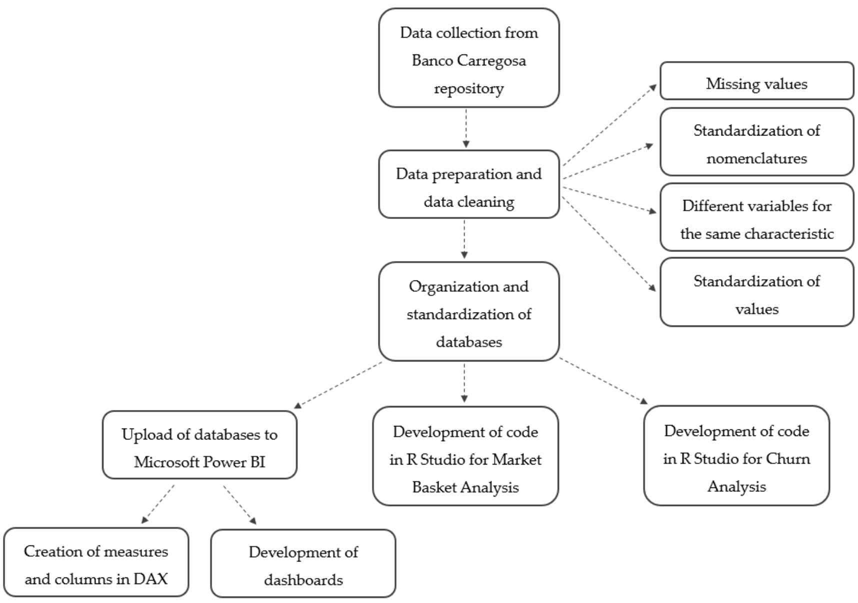

Figure 1 - Activities Performed

5.2 Data Preparation

As stated in chapter one, the data subject to analysis was provided by Banco Carregosa, via their internal software.

For the Data Preparation and Cleaning process (see in figure 2), Office Excel was the main software used, to make the data suitable for the subsequent analysis.

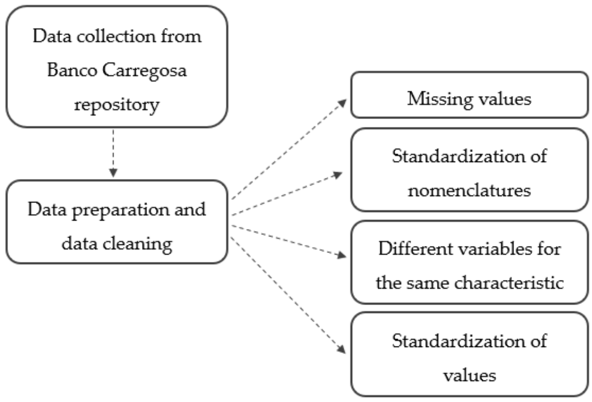

Figure 2 - Data Preparation Stages

The largest setbacks during this phase were related to four major problems. First, the setback related to missing data, which increased when analyzing data from past years. The missing values were replaced by the nomenclature "Unknown", to highlight the problem of missing values when presenting the findings to the team. Second, since a portion of data is manually collected, some variables were incorrectly stored, thus some nomenclatures needed to be updated and standardized, to ensure consistent data. Third, different variables which related to the same characteristic needed to be merged, for example, for the opening of accounts, the dataset provided two variables related to the main objective of the account, one filled by the customer and another by the account manager, which was merged and standardized. Finally, the values presented in the databases were not uniform, i. e., some databases contained values with commas while others contained values with dots. Consequently, data sets were standardized before starting the analyses. After all these steps, the databases were ready to be analyzed. To connect the databases, the tables were connected by the variable “Code”, which is a unique numerical code that represents the number of each account.

5.3 Data Visualization Results

Banco Carregosa had established the necessity of upgrading the display of information, with the creation of interactive dashboards that allowed for a simple and effective analysis.

We used Power BI to accomplish this task, shown in figure 3, which proved its relevance for the daily operations of Banco Carregosa. Power BI allowed for the creation of interactive financial and analytical dashboards, with the capacity of filtering data. Furthermore, the DAX tool (Data Analysis Expressions) in Power BI allows for further analysis, since it is capable of creating new measures and columns based on the existing databases. Thus, with this tool, employees can extract data quickly, accurately and in real-time, due to Power BI’s user-friendly interface and customization capabilities.

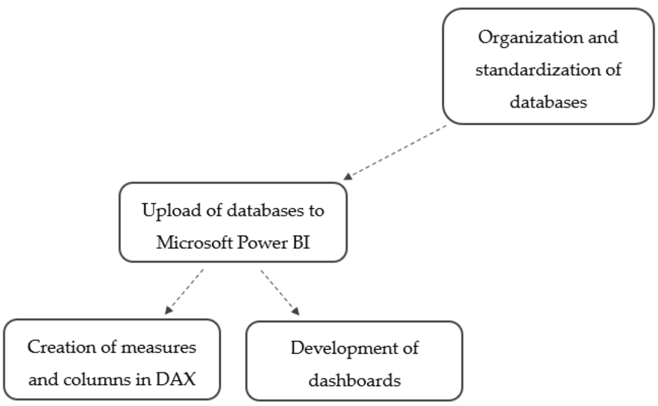

Figure 3 - Data Visualization Stages

We designed 11 interactive dashboard pages, intending to enhance the display of information and the decision-making process. The design of the dashboards is based on the principles highlighted by the authors Suk et al. (2010) and Yu-Wen et al. (2022). Suk et al. (2010) refer to the importance of using background colors, such as white or black, since these are emotionally neutral. Yu-Wen et al. (2022) argue that line, column and map charts are more likely to facilitate the extraction of useful information, and thus should be frequently used. In addition, Yu-Wen et al. (2022) also defend that headings and titles should clearly explain the contents of the visualizations, to reduce misleading conclusions, since their eye-tracking experiment concluded that participants initially concentrated their attention on the headings, rather than the visualizations. The filtering of these pages developed, related to date and business units (Private and Affluent Banking), varies depending on the information presented.

For privacy and data security reasons, the sensitive data present in the figures was blurred, to protect the integrity of Banco Carregosa.

5.3.1 Evolution of Patrimony and Commissions Dashboard

The first dashboard is related to the overall evolution of the performance of the existing accounts, between 2016 and 2022 (figure 4). On this page, we can analyse the evolution of the patrimony and commissions generated by accounts over those years, as well as the evolution of the average values of patrimony and commissions.

For this analysis, the visualizations used were Clustered column charts and Line charts, for a better understanding of the evolution through the period. The Card visualization also was used, to provide an easier analysis of the specific values.

The data provided by Banco Carregosa only had information about the total patrimony and commission values for each account, so we created two new columns (see Attachments A1 and A2), to calculate the average values of patrimony and commissions. We also created some measures (see Attachment B1) to facilitate the analysis of the total values and avoid misunderstandings.

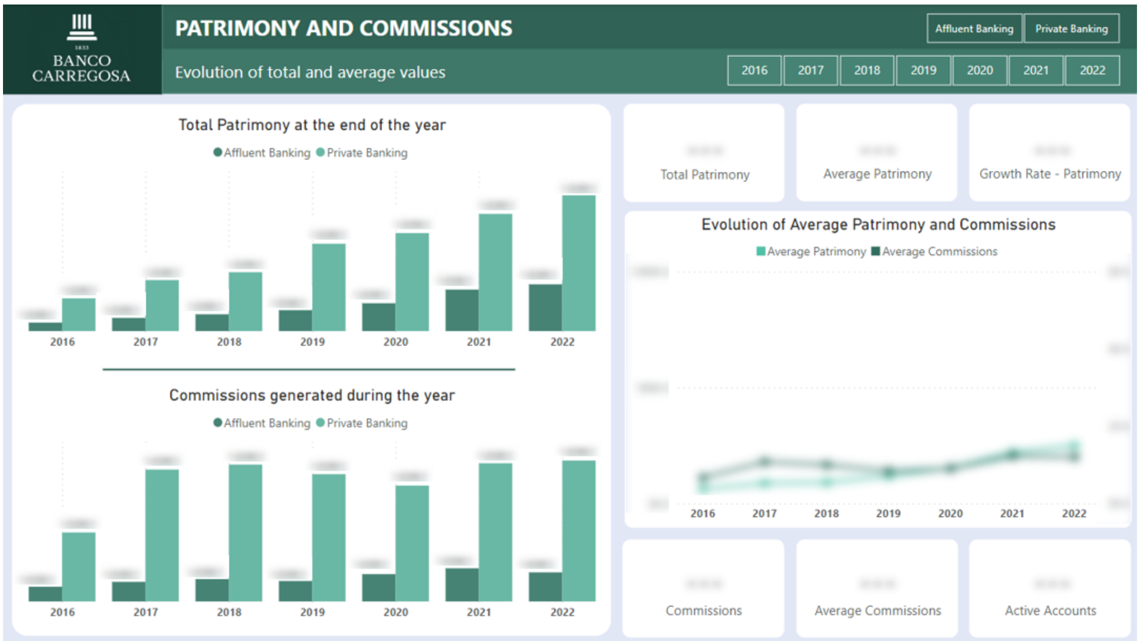

Figure 4 - Patrimony and Commissions Dashboard

5.3.2 Accounts Dashboard

The second dashboard elaborated is related to the characteristics of the accounts. The first page (figure 5) presents the characteristics of the active accounts in 2022: the number of accounts and the commissions' values by patrimony range, the patrimony by product category, the ratio between accounts from companies and individuals and the respective patrimony, the ratio between financial and non-financial accounts and the respective patrimony and, finally, the percentage of accounts that use GoBulling, Banco Carregosa's trading platform.

The visualizations used were Stacked column charts, to easily observe the contrast between the Business units and the ratio of accounts that used the trading platform. The Pie charts were also used to highlight the distinction between accounts from companies and individuals, as well as financial and non-financial companies. The Stacked bar chart visualization represents the patrimony by product category.

We calculated two measures on this page. As stated before, the dataset only provided information about the total values of patrimony, so one measure calculated the average patrimony value (see Attachment B2), while the other measure calculated the percentage of accounts that use GoBullying (see Attachment B3).

Figure 5 - Accounts in 2022 Dashboard

The following page (figure 6) provides information about the new accounts opened between 2016 and 2022 and the monthly evolution. This dashboard page also offers details about the total and average values of patrimony and commissions generated by the opened accounts, until the end of the opening year. For those accounts, we also have information about the evolution of commissions generated by service groups.

Line and Stacked column charts were used to facilitate the evolution analysis, and Cards were used to present the specific values.

For the Line chart "Accounts Opened by Month" we created a new column, to display the x-axis with the format "mmm" (see Attachment A3).

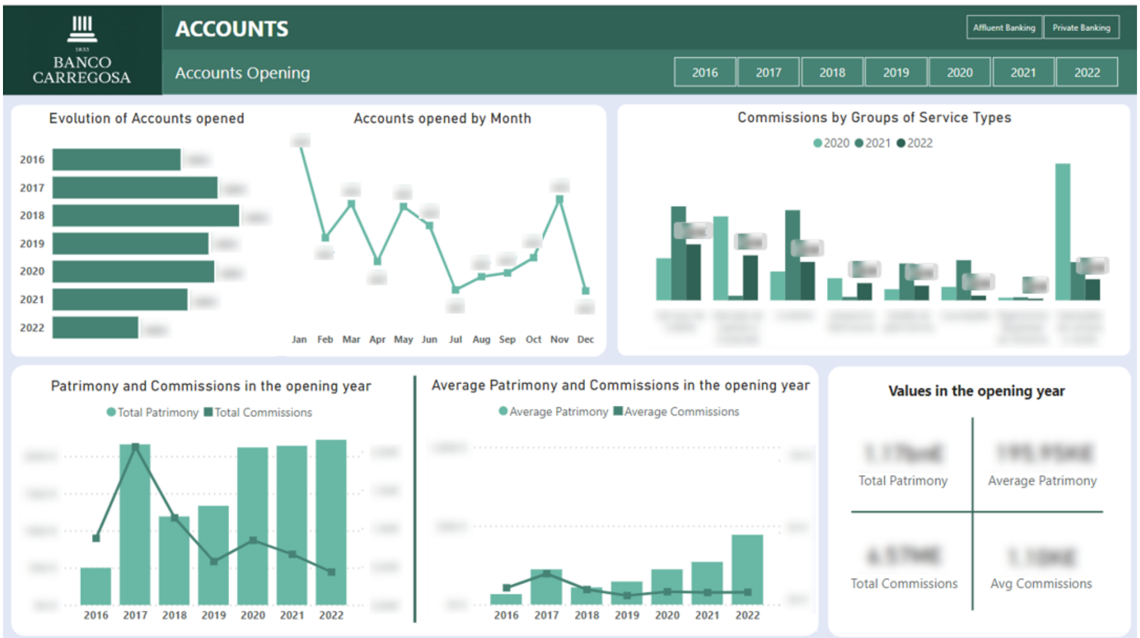

Figure 6 - Accounts Opened Dashboard

The next two pages (figures 7 and 8) provide an overall view regarding the reasons indicated by clients when opening the accounts. Figure 7 supplies information about the main purposes chosen by clients, between 2020 and 2022, while figure 8 furnishes details regarding the total and average values of patrimony and commissions by purpose. On both pages, it is possible to filter by the main purposes chosen.

In figure 7, the Matrix visualization appears as a complement to the Clustered column chart, by representing the least chosen purposes. The three Donut charts present the frequency of the reason chosen during the process of opening the account. Finally, the Card visualization is used to emphasize which individual purpose is being analysed or to highlight that the dashboard is showing the totality of purposes. In figure 8, the Line and Clustered chart supports a clear view of the total and average values for each purpose chosen.

For the Card visualization, we created a measure to emphasize the stated above (see Attachment B4). Similar to previous dashboards, the dataset related to the purpose of new accounts only included the total values of patrimony and commissions, so we created two columns to calculate the respective averages (see Attachments A1 and A2).

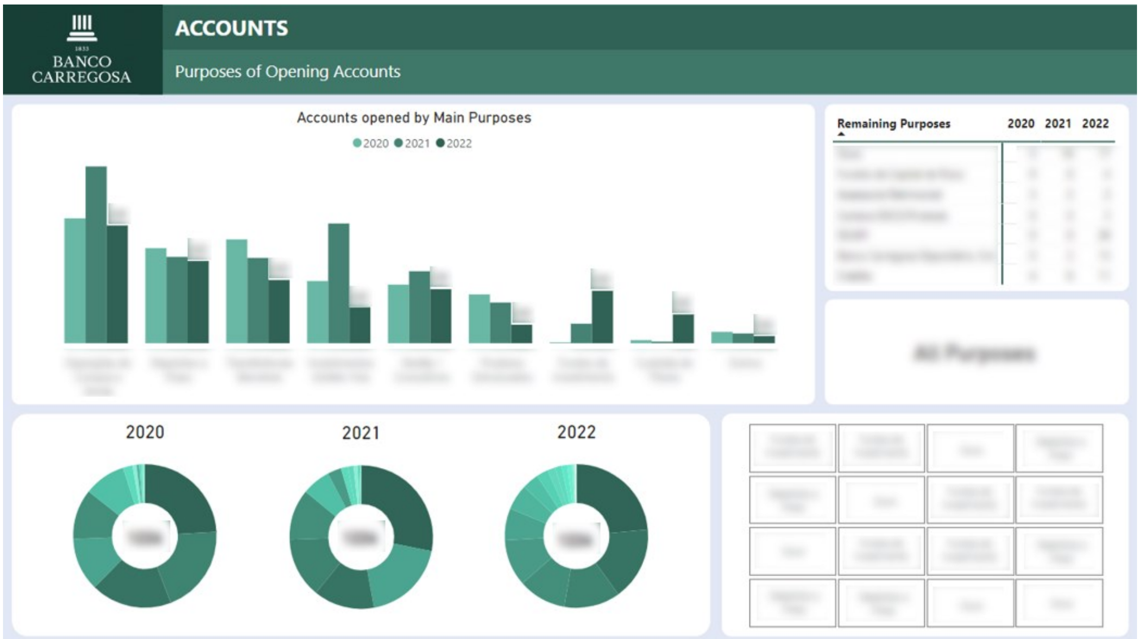

Figure 7 - Purposes of Opening Accounts Dashboard

Figure 8 - Patrimony and Commissions Dashboard

Finally, the last page of the accounts dashboard concerns the closure of accounts in 2022 (figure 9). Since Banco Carregosa had a massive internal process of closing accounts during 2020 and 2021, it was established that a comparison of the current year would not be suitable, thus the analysis only focuses on 2022. This page delivers an evolution of accounts closed by month, with the respective motives of closure. In addition, it was also requested to pay close attention to the motive "Compliance" since it is related to the internal procedure previously mentioned, which proved to deviate the evolution of closed accounts to September. To understand the impact of the closure of accounts, this page also supplies an analysis regarding the commissions generated from 2020 until 2022 from the accounts closed in 2022, regarding the motive of closure and service type.

On this page, four different types of visualizations were used. Clustered column charts were used to demonstrate the evolution by month and the commission values regarding service type. The Treemap was useful to understand the most frequent reasons for closure and the Matrix was useful to comprehend the impact in terms of total and average commissions. Finally, Card visualizations were used to present the number and the percentage of accounts closed, which can be filtered by both business unit and motive.

Similar to previous pages, we created a measure related to the average commission (see Attachment B5) and a new column, to display the x-axis with the format "mmm" (see Attachment A3).

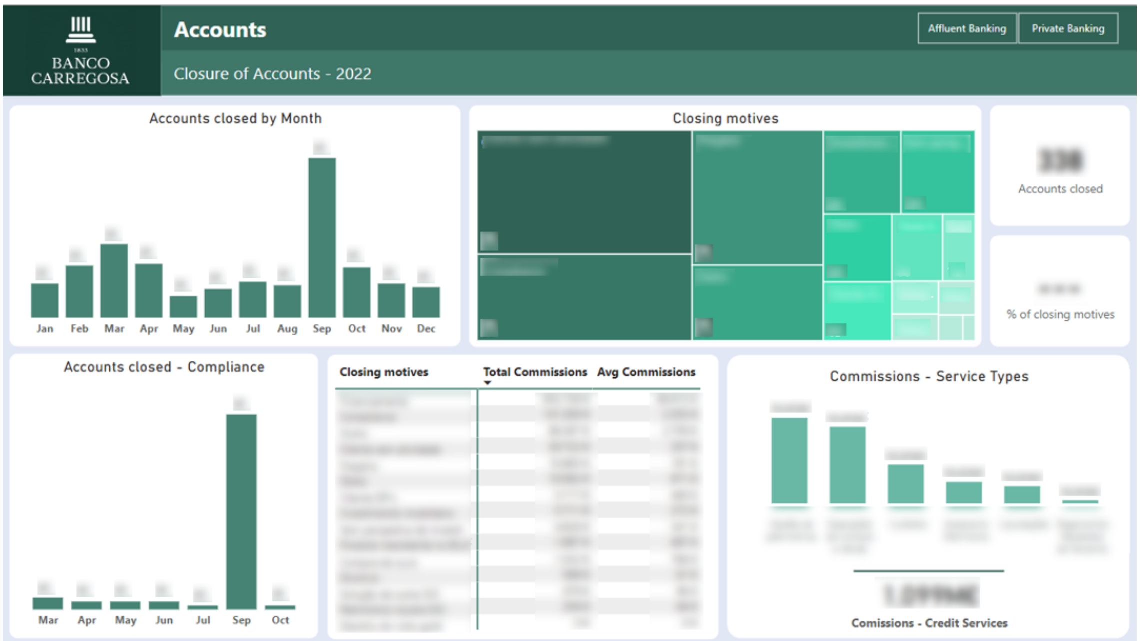

Figure 9 - Accounts Closed in 2022 Dashboard

5.3.3 Leads Dashboard

The third dashboard elaborated provides information about leads, i. e., a contact from a possible client that demonstrated interest in a service from Banco Carregosa. Leads from 2019 until 2022 were analysed.

The first page (figure 10) displays the evolution of leads received and leads contacted by commercials. It is fundamental for banking institutions to respond to all leads received, so it was fundamental to understand not only the number of leads received, but also the percentage of contacts. The page also offers a view regarding the type of lead received: “main lead” is related to first contact from a possible new client, “repeated lead” refers to clients that had already contacted Banco Carregosa in the past and “client lead“ refers to a contact from a client with an already active account. Furthermore, the page delivers information about the number of leads and contacts per type of lead.

The Line and Clustered column charts were useful to provide information regarding the absolute values and also the percentage of leads contacted, while the Treemap allowed for an easy overview regarding the most frequent lead type.

On this page, we created a measure (see Attachment B6) to determine the percentage of leads contacted compared to all the leads received.

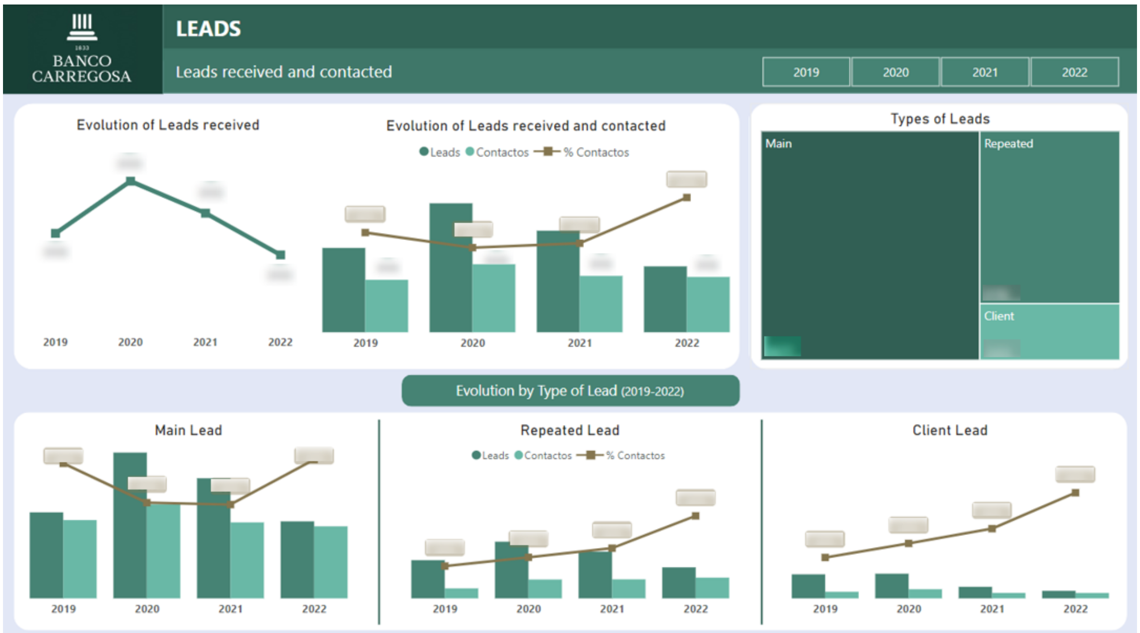

Figure 10 - Leads Received and Contacted Dashboard

The next page (figure 11) characterizes the leads received from 2019 and 2022, regarding their origin and the ratio of leads which were considered null. The Line and Clustered column chart provides information regarding campaigns launched by Banco Carregosa and the number of leads generated: the columns indicate the number of leads received by each campaign in 2022, while the line represents the evolution compared to the average of leads received between 2015 and 2021 for those campaigns.

A Treemap visualization, Pie Chart and Cards were used to present the characteristics of the leads.

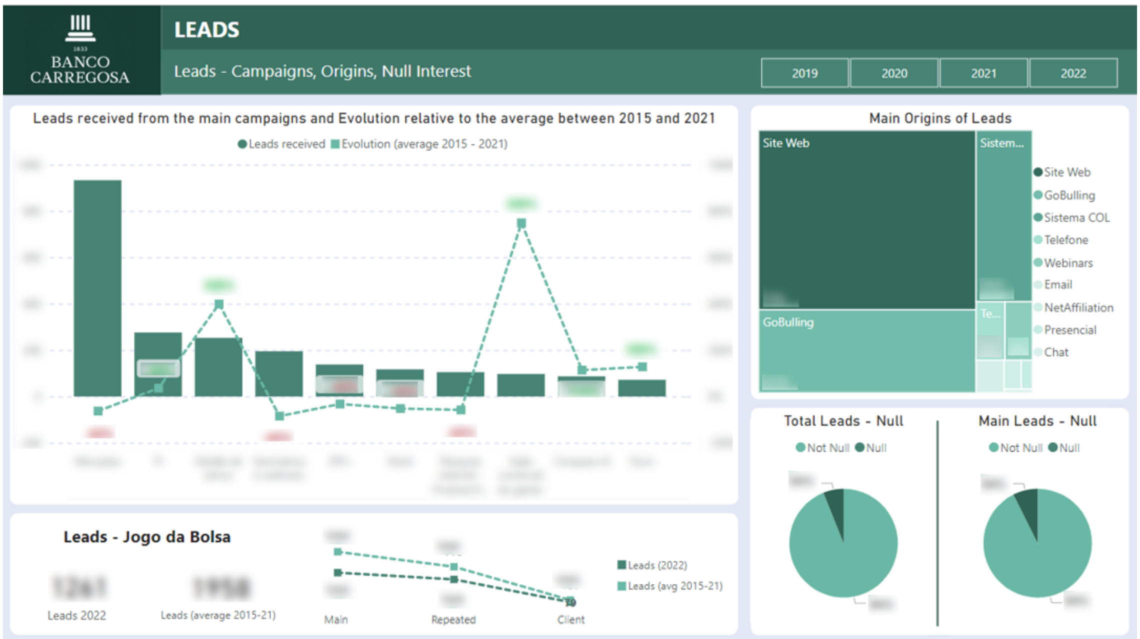

Figure 11 - Leads Characteristics Dashboard

The final page of the Leads dashboard assesses the data relative to the main leads and their conversion into accounts (figure 12). The Line and Clustered column charts provide information about the number of main leads received and the ratio of those converted. In this analysis, it was requested to separate the campaign "Markets" from the remaining campaigns. This display also supplies data regarding the campaign and origins of the converted leads, through the use of a Treemap and a Pie chart. Finally, one of the Cards presents the percentage of leads converted from the total number of leads received, while the other card provides knowledge regarding the average number of days between the reception of the lead and its conversion.

Two new measures were created: one represents the percentage of main leads converted (see Attachment B7), while the other measure calculates the average number of days of conversion (see Attachment B8).

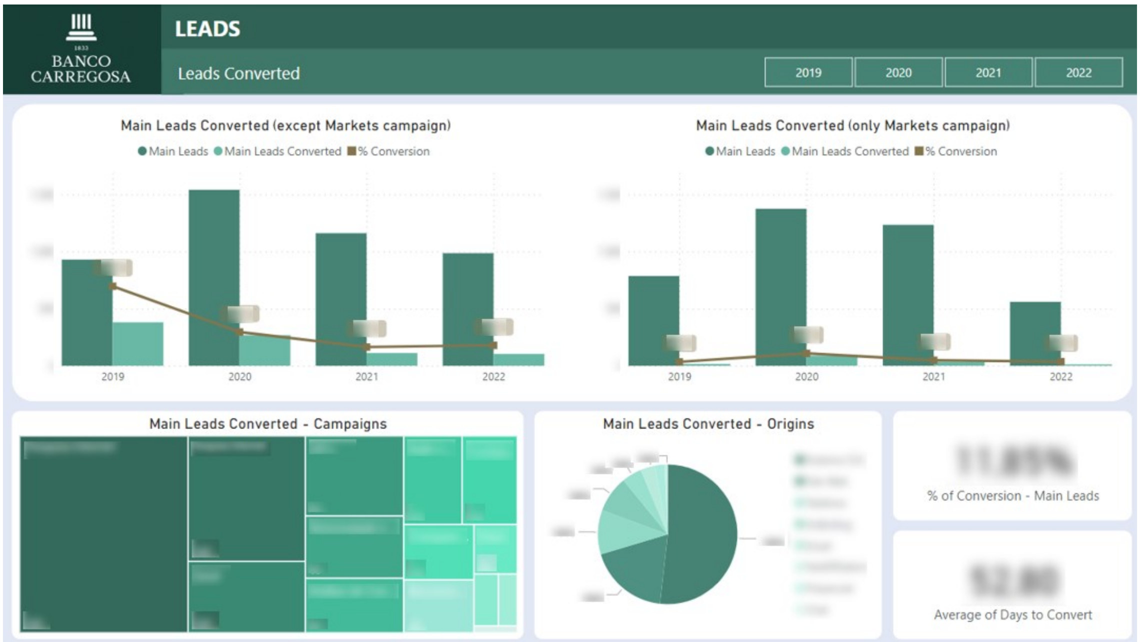

Figure 12 - Leads Converted Dashboard

5.3.4 Term Deposit "Bem-Vindo" Dashboard

The last dashboard offers insights on the topic of Term Deposits "Bem-Vindo", namely the evolution of Term Deposits "Bem-Vindo" and the values in case of conversion into different products/services of Banco Carregosa (figure 13).

The Clustered column chart provides an evolution of the number of Term Deposits "Bem-Vindo" opened from 2014 (the year when Banco Carregosa created this Term Deposit) until the present year. The Treemaps are related to the incorporation value of the deposit and, in case of conversion into other service of Banco Carregosa, the most frequent products/services. The Clustered column chart "Commissions after Term Deposits "Bem-Vindo"" offers details about the value of the commissions generated after the Term Deposit "Bem-Vindo". This visualization does not include the year 2014, since this deposit was only created in September 2014, so the commissions received during that period are negligible for this comparison. The "% Conversion" card refers to the percentage of clients that continued their activity with Banco Carregosa after the Term Deposit "Bem-

Vindo", while the remaining cards have statistics comparing the interests paid by Banco Carregosa from the Term Deposits and the commissions received from the new clients that continued their activities with other products/services. Finally, this page has two Line charts. The first one portrays an evolution and comparison between the interest rates executed and the number of Term Deposits "Bem-Vindo" opened. The second Line chart gives information about the average number of Term Deposits opened, by interest rates. As an illustration, the interest rate of 2% was used for 35 months, during which 997 Term Deposits were created, which is equivalent to having 28,5 Term Deposit per month of activity.

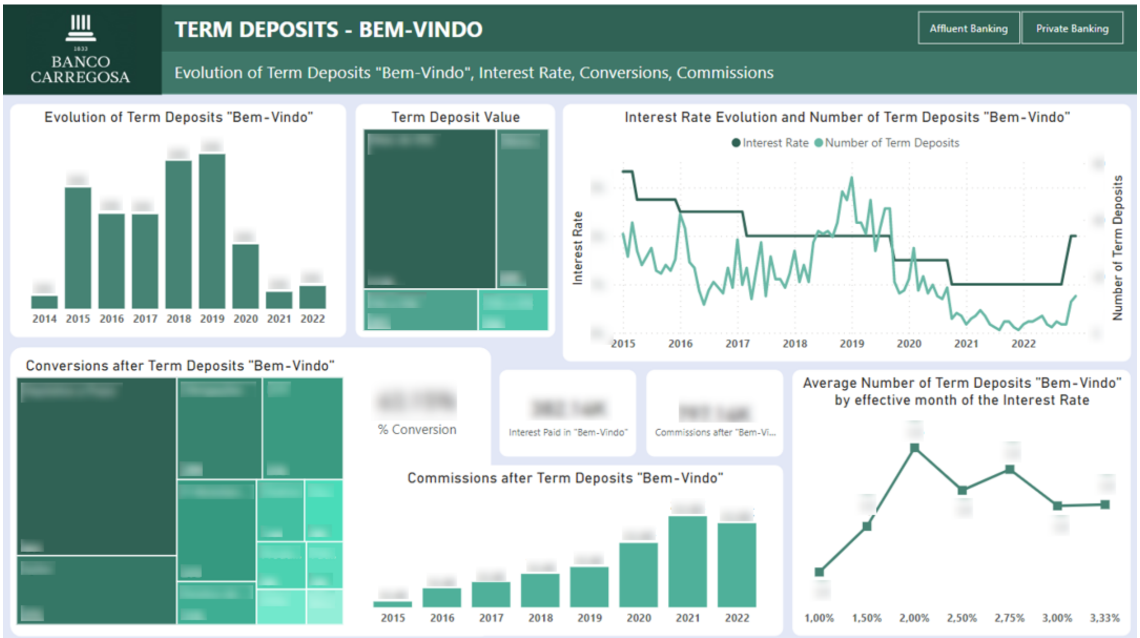

Figure 13 - Term Deposits "Bem-Vindo" Dashboard

5.4 Market Basket Analysis

As referred to in the literature, MBA is an association technique that is suitable for detecting patterns and relationships in large-scale datasets, which can be vital for the identification of cross-selling opportunities (Cavique, 2007).

In the context of Banco Carregosa, as seen in figure 14, we applied an MBA to investigate the relationships between products/services purchased by customers, to discover if any patterns would indicate that customers who had acquired a particular product/service were more likely, or not, to acquire another product/service.

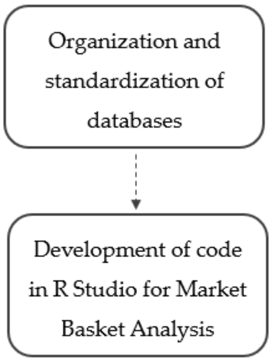

Figure 14 - Market Basket Analysis Stages

5.4.1 Data collection

To proceed with the MBA, we retrieved data regarding clients and the products/services purchased from 2019 until 2022. The data collected included details regarding the number of products traded by clients.

5.4.2 Data preparation

After collecting the data, which included details about the number of products bought, it was necessary to clean and prepare this data. Products with similar characteristics were categorized into groups. For example, "Treasury bonds" and "Senior bonds" were grouped into the category "Bonds". To finish, with the help of the Excel function "TEXTJOIN", the names were combined and delimited by a semicolon.

5.4.3 Analysis

Due to the magnitude of the data collected, we used RStudio, version 2022.12.0.353. R is a statistical and graphical program that suits advanced analysis, so it was appropriate for the MBA (Posit Team, 2022).

The main package used was *Arules*[1], which provides the infrastructure for representing, manipulating and evaluating transaction data and patterns, to form association rules. The algorithm *Apriori* searches for common itemsets level-wise.

The *Apriori* algorithm default behaviour searches for rules with a minimum support of 0,1, minimum confidence of 0,8 and a maximum of 10 item sets. However, with these default values, the algorithm only found 2 rules, so we altered the values, to obtain more rules. The code developed in RStudio can be seen in Attachment C1. Table 2 presents the leading rules for a minimum support of 0,05 and a minimum confidence of 0,8. As stated before, for confidentiality reasons, the values presented do not correspond to reality.

Antecedent	Consequent	Support	Confidence	Lift
Futures, Bonds	Shares	0,165	0,962	1,667
Futures	Bonds	0,145	0,870	1,534
CFDs, ETF	Shares	0,084	0,837	1,821
Bonds, ETF	Incorporation rights	0,075	0,804	1,120
Futures	ETF	0,050	0,778	1,248

Table 2 - MBA Structured Products results

[1] https://cran.r-project.org/web/packages/arules/index.html

Looking at the first rule of table 2, we can assess that clients that purchased Futures and Bonds also tend to buy Shares, with a confidence level of 96,2%. This metric, as explained in the literature review, explains the probability of the consequent being selected if the antecedent has already been chosen. Thus, we can conclude that 96,2% of the accounts that have transacted Futures and Bonds, also have Shares.

When presenting this analysis to the team, it was agreed that the MBA would be helpful to study one product in particular, "Structured Products". The objective was to analyse which clients are more likely to buy this product. We conducted another MBA, but with the consequent locked as "Structured Products". The minimum support value changed to 0,03 for this analysis. The code can be seen in Attachment C2. Table 3 presents the findings.

Antecedent	**Consequent**	**Support**	**Confidence**	**Lift**
Bonds, Futures, Investment Funds	Structured Products	0,035	0,875	8,666
Futures, Shares, Term Deposits	Structured Products	0,031	0,821	6,487
ETF, Shares	Structured Products	0,034	0,784	4,801
Bonds, ETF	Structured Products	0,037	0,760	4,127
Part Units, Term Deposits	Structured Products	0,031	0,714	4,241

Table 3 - MBA Structured Products results

By analyzing the first row of table 3, we can perceive that 87,5% of the clients who purchased “Bonds”, “Futures” and “Investment Funds” also purchased “Structured Products”.

5.4.4 Results

Both these analyses proved to be beneficial for Banco Carregosa, by equipping the respective teams with valuable knowledge regarding the products transacted by customers and their respective cross-selling opportunities. Separately to these analyses, we also created a database with information regarding accounts and the products/services purchased, which can easily be used to select specific accounts for cross-selling opportunities.

The output of the global analysis (shown in table 2) detected that trading instruments, such as Bonds and Shares, are usually linked with the purchase of other low-risk trading instruments. The dataset’s characteristics can have an impact on this analysis. It is expected that trading instruments will emerge more frequently in the MBA since they are purchased in greater quantities.

The later analysis, displayed in table 3, gave the respective teams insights regarding the product “Structured Products”. This analysis allowed the team to gather insights regarding the products/services more frequently purchased alongside “Structured Products”. For example, in the first row of table 2, the team could retrieve that 87,5% of accounts that bought “Bonds”, “Futures” and “Investment Funds” also bought “Structured Products”. To promote this product to the remaining 12,5% of clients, the team used the database independently built, to filter those specific clients and offer them specific campaigns for this product, since they have a higher likelihood of investing.

The MBA allowed Banco Carregosa to understand its cross-selling opportunities based on clients' behaviour, which is a valuable tool for increasing transactions and broadening the expertise of internal teams.

5.5 Churn Analysis

Customer churn analysis is fundamental for competitive industries, since the costs related to recruiting new clients are higher than the costs of maintaining existing ones, due to expenses such as credit searching and advertising costs. The best strategy for maximizing profits is to retain current clients and avoid their churn (Sharma et al., 2011).

For this dissertation, we applied a Logistic Regression model, which had already been used in other customer churn studies, such as Coşer et al. (2020), given that it showed positive performances.

In the context of Banco Carregosa, we developed a model (see figure 15) to get detailed information about which customers churned and their respective characteristics, to analyze if any trait is related to the churn. Based on the available data, we studied socio-demographic indicators, such as *age*, *gender*, *profession*, *literacy prowess*, etc.

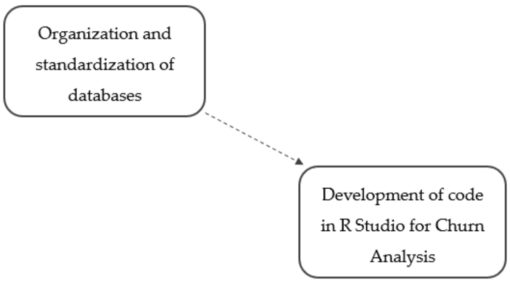

Figure 15 - Churn Analysis Stages

5.5.1 Data collection

For this analysis, we decided to analyse customer churn by year. For example, for 2022, we gathered information regarding clients with active accounts and their characteristics. In this dataset, there were clients which closed their accounts during this year and also clients who didn´t and maintained their activities. We decided to focus on the client's features, such as *age, gender, marital status, profession, literary prowess, investor profile* and *campaign* that led to the account being opened and the *number of accounts holders*. It was also agreed to differentiate between the two business units, Private Banking and Affluent Banking, to compare both outputs.

5.5.2 Data preparation

After collecting the data, we started with its standardization, to prepare it for the model. First, the categorical variables (*campaign, marital status, gender, profession, literary prowess and investor profile*) were transformed into dummy variables. The variable *number of account holders* was kept as a numerical variable. Finally, the variable *age* was also converted into a dummy variable, organized in intervals decided by the team. We additionally divided the dataset into two datasets according to their business units, as suggested by the team.

As mentioned, the variables chosen to be analyzed were the following:

- Age – Below 30, from 30 to 39, from 40 to 49, from 50 to 59, from 60 to 69, from 70 to 79, above 80;
- Gender – Male or Female;
- Marital Status – Single, Married, Divorced, Widowed, Separated, etc;
- Profession – Tourism, Industry, Military, Services, Distribution, etc;
- Literary prowess – Superior, Secondary, Basic, No literacy;
- Investor profile – Dynamic, Aggressive, Moderated, Conservative;

- Campaign – A, B, C, D, E, F.

Finally, to examine the customer churn, we added a binary variable to the database, called "churn". This is the dependent variable, where the value 1 indicates the clients who churned and the value 0 denotes the opposite.

5.5.3 Analysis

The dataset was uploaded into RStudio, version 2022.12.0.353. As stated in the literature review, the Logistic Regression model forecasts a binary categorical dependent variable, using a given set of independent variables. Before applying the model, we converted the dependent variable into a factor. The code created can be seen in Attachment C3. As stated before, the values presented do not correspond to reality.

We ran the following code for the Logistic Regression model, with all the variables previously mentioned, for the business unit Affluent Banking:

LogModel <- glm (Churn ~., family = "binomial", ChurnDPI)

Furthermore, we applied the step function, to remove the variables which are not significant to the model, thus improving it. The variables *gender* and *literacy prowess* were not significant, thus disappearing from the step model:

StepLog <- step (LogModel)

After observing the Akaike Information Criteria (AIC) of both models, i. e., the estimator of prediction error, we concluded that the second is a better-fit model since its AIC is lower than the first model. Table 4 demonstrates the results from the step model above mentioned.

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-3,105	0,294	-10,485	0,127
Campaign=A	0,651	0,248	2,452	0,154

Campaign=C	0,193	0,344	-2,880	0,017*
Campaign=D	-0,998	0,329	-2,874	0,084.
Age=30 to 39	-0,546	0,881	1,456	0,075.
Age=40 to 49	0,724	0,309	2,338	0,045*
Age=50 to 59	0,756	0,324	1,805	0,027*
Age=60 to 69	0,756	0,304	2,204	0,150
Age=70 to 79	-0,991	0,258	2,704	0,005**
Age=Above 80	-1,340	0,414	3,051	0,009**
Investor profile=defensive	-0,954	0,154	1,502	0,008**
Profession=services	-0,351	0,254	-1,340	0,223
Profession=turism	-1,987	0,800	-0,584	0,130
Profession=distribution	0,005	1,008	1,205	0,080.
Marital Status=single	0,364	0,257	-1,345	0,095.
Marital Status=married	0,021	0,831	2,745	0,015*
Marital Status=married with Acquired Communion	0,425	0,151	1,506	0,132

Table 4 - Logistic Regression results

Signif. codes: 0 '*' 0.001 '*' 0.01 '' 0.05 '.' 0.1 ' ' 1

The outcomes of the logistic regression in R enable Banco Carregosa to perform analyses such as individuals between the ages of 40 and 59, married and whose campaign is C have more probability of closing their accounts, while clients with more than 70 years, with defensive investor profiles, have a bigger probability of continuing their activity with Banco Carregosa.

For instance, individuals between 40 and 49 have *exp(0,724)* = 2,06 times more probability of closing their accounts. Clients between 50 to 59 years old have *[exp(0,756) - 1] * 100* = 113% larger probability of churning.

5.5.4 Results

This algorithm proved to be suitable for the activity of Banco Carregosa because it provided the team with detailed information regarding the type of customers who are more likely to close their accounts. Consequently, Banco Carregosa can identify these profiles and create specific campaigns to boost the engagement of these clients, to reduce churn. Furthermore, Banco Carregosa can apply this algorithm to analyse other variables, such as activity-related indicators, for example, the time of inactivity of the client.

In this context, and by analyzing table 4, Banco Carregosa could create campaigns or focus on contacting married clients with less than 59 years.

5.6 Market Basket Analysis and Churn Analysis

An additional analysis was performed to identify any connections between the products/services purchased and the propensity of customer churn. In the present context and with the databases previously withdrawn, we decided to proceed with an MBA, where the consequent would be locked with the variable corresponding to the closure of accounts.

5.6.1 Analysis

Similar to the prior MBA, we employed RStudio, version 2022.12.0.353. The main package was *Arules* and the selected algorithm was *Apriori.* The code developed can be seen in Attachment C4.

For this analysis, we used a minimum support of 0,05 and a minimum confidence of 0,1.

Antecedent	Consequent	Support	Confidence	Lift
Term Deposit	Account closed	0,105	0,311	2,018
Bonds, Futures	Account closed	0,087	0,238	1,356
Futures	Account closed	0,089	0,224	1,545
Bonds, Shares	Account closed	0,097	0,203	1,004

Table 5 - MBA accounts closed results

Through this MBA, Banco Carregosa can perform analyses such as 31,1% of clients that purchased Term Deposits closed their accounts, while from the clients that bought both Bonds and Futures, 23,8% of those accounts were closed.

5.6.2 Results

This analysis, similar to the previous MBA conducted, provides the team with knowledge regarding relationships between the products/services and the propensity of customer churn. By applying this algorithm, Banco Carregosa can understand the risk of churn based on products transacted and analyse its impact on the activity of the bank.

After discussing these particular findings with the team, it was agreed that the first rule, regarding the Term Deposits, is consistent with reality since customers that transact Term Deposits are frequently switching banks, in search of the most profitable interest taxes, thus resulting in a higher churn rate.

5.7 Discussion of Results

Data Visualization tools are gaining importance with the volume of information and data produced in recent years. Due to the complexity of organizing and presenting this information (Dabbebi et al., 2017), dashboard creation has grown in importance by providing an overview of large data sets in a visually appealing format and improving the decision-making process (Qin et al., 2020; Yu-Wen et al., 2022). The dashboards created in Banco Carregosa reinforced the idea mentioned in the literature, Banco Carregosa gained new insights of its data which assisted collaborators in making decisions to improve their activity.

For the MBA, the literature refers the importance of this analysis in comprehending the buying pattern of clients, which allows for the discovery of association rules between products that are frequently purchased together (Cavique, 2007; Kaur et al., 2016). In Banco Carregosa, the execution of this model reinforced the significance mentioned in the literature, by allowing the institution to identify association rules between its products/services and understanding cross-selling opportunities based on clients' behaviours, which is a valuable tool for boosting transactions and broadening the expertise of internal teams.

The literature also illustrates the significance of Churn Analysis, by employing statistical models. Since it is more expensive to acquire new clients than to maintain the current ones (Athanassopoulos, 2000; Sharma et al., 2011), understanding the traits of clients which tend to discontinue their activity with the institution is a vital strategy. Applying this technology in Banco Carregosa proved the points referred in the literature, since it gave internal teams important knowledge about specific clients' characteristics. By focusing on those clients who are more likely to close their accounts and developing preventive measures, Banco Carregosa could reduce expenditures.

Sixth Chapter

6. Conclusion

The main objective of this report was to demonstrate the benefits and the applicability of using BI and BA technologies in Banco Carregosa and the Wealth Management sector. To measure this, we supplied Banco Carregosa with models and techniques to improve the decision-making process and to facilitate the visualization of data.

Since Banco Carregosa is just beginning its journey in the matter of data analytics, we encountered some limitations during the internship, particularly during the data preparation stage. Banco Carregosa's repository is not fully optimized and connected, so the data extraction procedure was long. Furthermore, as previously stated, we ran into several setbacks while cleaning the data. However, despite the limitations encountered, the main objective of this report was successfully achieved. The activities performed and the results were presented to some elements of Banco Carregosa's Board and the Marketing team and perceived as useful to support decision-making and combat the fiercely competitive banking sector. The techniques developed allowed the institution to have a deeper awareness of its environment, which creates opportunities to deliver an enhanced and more adequate service to its clients.

Based on the limitations encountered, we can set some suggestions in the matter of analytics in Banco Carregosa, to improve the resources and mechanisms used and make this process more agile. The missing data and inconsistent terminology problems must be addressed, since they are creating biases in the analysis. Therefore, the data storage process needs to be more strict and, if possible, automated. To overcome the obstacle of number standardization, Banco Carregosa should perform a thorough cleaning of its system, to standardize commas and dots in their values, and keep values coherent. Finally, Banco Carregosa already has the technical foundations of the models developed, however, this is a process that always needs to be optimized and adjusted to concur with the pertinent analysis. In terms of data visualization, Banco Carregosa should continue to develop interactive dashboards and distribute them among key collaborators of the institution, since these dashboards deliver accurate and updated information, which can be crucial to improve Banco Carregosa's strategies.

References

Aguinis, H., Forcum, L. E., & Joo, H. (2013). Using Market Basket Analysis in Management Research. *Journal of Management*, *39*(7), 1799–1824. https://doi.org/10.1177/0149206312466147

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, H. J., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*(3), 515–539. https://doi.org/10.1177/1094428109333339

Athanassopoulos, A. D. (2000). *Customer Satisfaction Cues To Support Market Segmentation and Explain Switching Behavior* (Vol. 47).

Bakshi, K. (2012). Considerations for big data: Architecture and approach. *IEEE Aerospace Conference Proceedings*. https://doi.org/10.1109/AERO.2012.6187357

Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques for marketing, sales, and customer relationship management* (2nd ed.).

Bhargava, M. G., Kiran, K. T. P. S., & Rao, D. R. (2018). Analysis and design of visualization of educational institution database using power bi tool. *Global Journal of Computer Science and Technology*, *18*(C4), 1-8.

Brath, R., & Peters, M. (2004). *Dashboard Design why design is importat*.

Calzon, B. (2022). *What Is The Difference Between Business Intelligence And Analytics?*

Cavique, L. (2007). *A Scalable Algorithm for the Market Basket Analysis*.

Chen, H., Chiang, R. H. L., Storey, V. C., Lindner, C. H., & Robinson, J. M. (2012). *Business Intelligence and Analytics: From Big Data to Big Impact Quarterly-Business Intelligence and Analytics: From Big Data to Big Impact* (Vol. 36, Issue 4).

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, *40*(2), 339–354. https://doi.org/10.1016/j.dss.2004.04.009

Coghlan, D. (2003). Practitioner Research for Organizational Knowledge: Mechanistic- and Organistic-oriented Approaches to Insider Action Research. *Management Learning*, *34*(4), 451–463. https://doi.org/10.1177/1350507603039068

Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J. D., & Yang, C. (2001). Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, *13*(1), 64–78. https://doi.org/10.1109/69.908981

Coşer, A., Aldea, A., Maer-Matei, M. M., & Beşir, L. (2020). Propensity to churn in banking: What makes customers close the relationship with a bank? *Economic Computation and Economic Cybernetics Studies and Research*, *54*(2), 77–94. https://doi.org/10.24818/18423264/54.2.20.05

Cukier. (2010). *Data, data everywhere*.

Ćurko, K., Pejic Bach, M., & Radoniü, G. (2007). *Business Intelligence and Business Process Management in Banking Operations*.

Dabbebi, I., Iksal, S., Gilliot, J.-M., May, M., & Garlatti, S. (2017). *Towards Adaptive Dashboards for Learning Analytic: An Approach for Conceptual Design and implementation*. 120–131. https://doi.org/10.5220/0006325601200131ï

Davenport, T. H. (2006). *Competing on Analytics Compliments of…*. www.hbr.orgorcall800-988-0886.www.hbr.org

Dick, B. (2015). Reflections on the SAGE Encyclopedia of Action Research and what it says about action research and its methodologies. *Action Research*, *13*(4), 431–444. https://doi.org/10.1177/1476750315573593

Diebold, F. X., Brown, L., Cheng, X., Cunha, F., Diebold, S., Foster, D., Halperin, M., Lohr, S., Mashey, J., Nickolas, T., Olson, L., Pai, M., Pospiech, M., Schorfheide, F., Shin, M., & Steele, M. (2012). *"A Personal Perspective on the Origin(s) and*

Development of 'Big Data': The Phenomenon, the Term, and the Discipline", Second Version A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline *. http://ssrn.com/abstract=2202843http://economics.sas.upenn.edu/pier

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gantz, J., & Reinsel, D. (2011). *IDC I V I E W E x t r a c t i n g V a l u e f r o m C h a o s Sponsored by EMC Corporation*. http://www.emc.com/digital_universe.

Gartner. (2014). *Gartner Glossary - Big Data*.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (Third Edition).

He, Z., Xu, X., Huang, J. Z., & Deng, S. (2004). Mining class outliers: concepts, algorithms and applications in CRM. *Expert Systems with Applications*, *27*(4), 681–697. https://doi.org/10.1016/j.eswa.2004.07.002

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, *85*, 78–85. https://doi.org/10.1016/j.procs.2016.05.180

Kaur, M., Singh, K., & Sharma, N. (2013). *International Journal on Recent and Innovation Trends in Computing and Communication Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers*. http://www.ijritcc.org

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*(1). https://doi.org/10.1186/s40854-016-0029-6

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, *34*(3), 387–394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

Lacković, I. D., Kovšca, V., & Vincek, Z. L. (2020). A review of selected aspects of big data usage in banks' risk management. In *Journal of Information and Organizational Sciences* (Vol. 44, Issue 2, pp. 317–330). University of Zagreb, Faculty of Organization and Informatics. https://doi.org/10.31341/jios.44.2.7

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1

Locke, E. A. (2007). The case for inductive theory building. *Journal of Management*, *33*(6), 867–890. https://doi.org/10.1177/0149206307307636

Lusch, R., Liu, Y., & Chen, Y. (2010). *The Phase Transition of Markets and Organizations The New Intelligence and Entrepreneurial Frontier*.

Manyika, J., McKinsey Global Institute, Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big Data : The next frontier for innovation, competition, and Productivity*.

Mcdonald, R. A., Thurston, P. W., & Nelson, M. R. (2000). *A Monte Carlo Study of Missing Item Methods*.

McNiff, J., & Whitehead, J. (2011). *ONE What Is Action Research?*

Microsoft. (2023). What is Power BI? [Web Page]. Retrieved from https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

Muskan, Singh, G., Singh, J., & Prabha, C. (2022). Data Visualization and its Key Fundamentals: A Comprehensive Survey. *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, 1710–1714. https://doi.org/10.1109/ICCES54183.2022.9835803

Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems*, *13*. https://doi.org/10.17705/1CAIS.01315

Nie, G., Zhang, L., Li, X., & Shi, Y. (2006). The analysis on the customers churn of charge email based on data mining - Take one internet company for example. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 843–847. https://doi.org/10.1109/icdmw.2006.154

Orlovskyi, D., & Kopp, A. (2020). *Information Technology and Interactions (IT&I-2020) 163*.

Özköse, H., Arı, E. S., & Gencer, C. (2015). Yesterday, Today and Tomorrow of Big Data. *Procedia - Social and Behavioral Sciences*, *195*, 1042–1050. https://doi.org/10.1016/j.sbspro.2015.06.147

Posit team (2022). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: a survey. *VLDB Journal*, *29*(1), 93–117. https://doi.org/10.1007/s00778-019-00588-3

Roth, P. L., Switzer Iii, F. S., & Switzer, D. M. (1999). *Missing Data in Multiple Item Scales: A Monte Carlo Analysis of Missing Data Techniques*.

Russell, G. J., & Petersen, A. (2000). *Analysis of Cross Category Dependence in Market Basket Selection*.

Sharma, A., & Kumar Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. In *International Journal of Computer Applications* (Vol. 27, Issue 11).

Shepherd, D. A., & Sutcliffe, K. M. (2011). INDUCTIVE TOP-DOWN THEORIZING: A SOURCE OF NEW THEORIES OF ORGANIZATION. *Academy of Management Review*, *36*(2), 361–380. https://doi.org/10.5465/amr.2011.59330952

Suk, H. J., & Irtel, H. (2010). Emotional response to color across media. *Color Research and Application*, *35*(1), 64–77. https://doi.org/10.1002/col.20554

Harford, T. (2020). *How to Make the World Add Up*.

Turban, Efraim. (2014). *Business intelligence and analytics : systems for decision support*.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

Yu-Wen, H., Yu-Ju, Y., & Wei, J. (2022). User Perception and Eye Movement on A Pandemic Data Visualization Dashboard. *Proceedings of the Association for Information Science and Technology*, *59*(1), 121–131. https://doi.org/10.1002/pra2.610

Attachments

Attachment A

A1:

```
Average Patrimony = ‘Total Accounts'[Total Patrimony at the end of the year] / 'Total
Accounts’[Number of accounts at the end of the year]
```

A2:

```
Average Commissions = ‘Total Accounts’[Commissions of the year] / 'Total
accounts'[Number of accounts at the end of the year]
```

A3:

```
Small Month = FORMAT('Opened Accounts[Approval Date], "mmm")
```

Attachment B

B1:

```
Total Patrimony = IF(SUM('Total Accounts'[Total Patrimony at the end of the
year])>7000000000,
    "---",
    SUM('Total Accounts'[Total Patrimony at the end of the year]))
```

Note: This measure was created to facilitate the analysis of the total values and avoid misunderstandings. For instance, when filtering by year, the "Total Patrimony " card reveals the sum of the patrimony of the accounts in the selected year. However, if the page is not filtered by year, the value displayed on the card would not represent any legitimate value (in this case, the value would represent the sum of all patrimony registered at the end of all years, which would deviate the analysis, because if an account had patrimony in 2020 and 2021, the Card visualization would sum both values). For those cases, to avoid this mistake, some measures were created, to display the symbol “---“.

B2:

```
Average Patrimony = SUM(Companies[Total Patrimony])/SUM(Companies[Number of Accounts])
```

B3:

```
GoBulling usage percentage = SUM(GoBulling[Accounts using GoBulling]) / (SUM(GoBulling[Accounts using GoBulling]) + SUM(GoBulling[Accounts not using GoBulling]))
```

B4:

```
Name of Purpose = SELECTEDVALUE(Purposes[Purposes],"Totality of Purposes”)
```

B5:

```
Average Commissions = SUM('Opened Accounts'[Commissions of the year]) / COUNT('Opened Accounts'[Approval Date])
```

B6:

```
Leads Contacted = SUM(Leads[Contacted]) / COUNT(Leads[Registered])
```

B7:

```
Percentage of Leads Converted = SUM('Leads Converted'[Dummy where 1 equals converted]) / COUNT('Leads Converted'[Registered])
```

B8:

```
Average days to convert = AVERAGE('Leads Converted'[Days to Convert])
```

Attachment C

C1:

```
Rules <- apriori(data, parameter = list(supp=0.05, conf= 0.8))
sort <-sort(Rules, decreasing = TRUE, by="confidence")
inspect(sort)
```

C2:

```
RulesSP <- apriori(data, parameter = list(supp=0.03, conf=0.8),
                   appearance = list(rhs=c("Structured Products"), default="lhs"))

SortSP <- sort(RulesSP, decreasing = TRUE, by="confidence")

inspect(SortSP)
```

C3:

```
str(data)
data$Churn <- as.factor(data$Churn)
data$Code <- NULL

Model <- glm(Churn ~., family = "binomial", data)
summary(Model)

Step <- step(Model)
summary(Step)

AIC(Model)
AIC(Step)
```

C4:

```
data[] <- lapply(data, as.factor)

Model <- apriori(data, parameter = list(supp=0.05, conf=0.1),
                 appearance = list(rhs=c("Account Closed=1"),
                                   lhs=c("Shares=1","Gold=1","CFDs=1",
                                         "Term Deposits=1","ETF=1",
                                         "Forex=1","Forex on Crypto=1",
                                         "Futures=1","Bonds=1","Options=1",
                                         "Comercial=1","Structures Products"=1,
                                         "Investment Funds=1""Warrants=1"),
                                   default="none"))

Sort <- sort(Model, decreasing = TRUE, by="confidence")

inspect(Sort[1:15])
```