UNIVERSIDADE CATÓLICA PORTUGUESA

Analytical CRM in a management consulting firm

An application of data driven techniques

Inês Oliveira Amorim

Católica Porto Business School

2021

UNIVERSIDADE CATÓLICA PORTUGUESA

Analytical CRM in a management consulting firm

An application of data driven techniques

Final Work in Organisational Context
presented to Universidade Católica Portuguesa
in order to obtain the master's degree in
Management with specialisation in Business Analytics

by

Inês Oliveira Amorim

under the guidance of
Prof. Dr. Vera Lúcia Miguéis Oliveira e Silva

Católica Porto Business School, Universidade Católica Portuguesa
May 2021

Acknowledgments

I would like to express my sincere gratitude to all the people without whom it would not have been possible to carry out this work.

First and foremost, I would like to express my deepest appreciation to my thesis advisor, Professor Vera Miguéis, for her continuous support, availability, motivation, and guidance throughout this journey. Thank you for all the suggestions, for the constant concern to follow my work during my internship at Inova+, for the sharing of knowledge and for the valuable comments on this thesis.

Furthermore, I would also like to express my gratitude to the Professors from the specialisation in Business Analytics for their excellent teaching capacities, valuable insights, and patience, especially in these difficult times that we all face, and to my colleagues for their collaboration and friendship. I am very glad I chose this path of analytics.

In addition, I would also like to thank the company of my internship, Inova+, for welcoming me with open arms and for the availability to provide me with the necessary data. A special thanks to Dr. Nuno Soares, my internship mentor, and to my internship colleagues for the joyful moments.

I could not have completed this thesis without the support of my friends, particularly those who went through this process with me, contributing with discussions and advices for the research, while also being happy distractions.

Lastly, I would like to express my gratitude to my family, especially to my parents, for their support and encouragement, and for all the efforts they made to provide me with an excellent education. I also thank my boyfriend for all his help, patience, and love.

Resumo

Considerando o ambiente competitivo em que as empresas operam atualmente e a importância do *customer relationship management* (CRM), é crucial analisar os dados relacionados com clientes para adquirir mais conhecimento e obter importantes *insights* sobre os mesmos, a fim de aumentar a sua retenção e o desempenho da empresa.

A investigação apresentada resultou de um estágio curricular realizado na empresa Inova+, uma consultora especializada no apoio ao crescimento de organizações. Neste sentido, o objetivo desta investigação visa apoiar o sistema CRM e as estratégias de gestão de clientes da Inova+, contribuindo para a melhoria e fortalecimento das relações entre a empresa e os seus clientes. Para esse efeito, uma metodologia quantitativa utilizando ferramentas analíticas, nomeadamente ferramentas de *data mining*, foi adotada para estudar várias dimensões do CRM.

Neste contexto, esta investigação focou-se em quatro aspetos principais em análise, que permitiram obter um conhecimento mais detalhado sobre os clientes da empresa. Inicialmente, a observação de *KPIs* relativos ao CRM e ao desempenho da empresa através da construção de dashboards. Em segundo lugar, foi aplicado um modelo de previsão de séries temporais relativo ao volume de negócios potencial. Adicionalmente, foram identificados segmentos de clientes de acordo com o seu comportamento de compra através da aplicação de um modelo RFM e foi desenvolvida uma análise de *clustering*. Por fim, foram identificados fatores significativos que influenciam a probabilidade de adjudicação de uma proposta comercial, tais como o país, tipo de organização e setor económico da empresa cliente, bem como o serviço associado.

Palavras-chave: B2B; Consultoria; *Customer Relationship Management*; *Data Mining*; Previsão

Abstract

Considering the competitive environment in which companies operate nowadays and the importance of customer relationship management (CRM), it is crucial to analyse customer-related data to gain knowledge and insights about them in order to increase their retention and company's performance.

The presented investigation resulted from a curricular internship carried out at Inova+, a management consulting firm specialised in supporting the growth of organizations. In this sense, the aim of this investigation is to support the CRM system and the customer's management strategies of Inova+, contributing to the improvement and strengthening of relations between the company and its customers. For this purpose, a quantitative methodology using analytical tools, namely data mining tools, was adopted to study various dimensions of CRM.

In this context, this investigation focused on four main aspects under analysis, which allowed to obtain a more detailed knowledge about the company's customers. Initially, the observation of KPIs regarding the CRM and the company's performance through the construction of dashboards. Secondly, a time-series forecasting model for prospective revenues was applied. Additionally, an identification of customer segments according to their purchasing behaviour through the application of a RFM model and a clustering analysis was carried out. Finally, significant factors that influence the probability of adjudication of a commercial proposal were identified, such as the country, type of organisation and economic sector of the client company, as well as the service associated.

Keywords: B2B; Customer Relationship Management; Data Mining; Forecasting; Management consulting

Table of Contents

List of Figures

List of Tables

Acronyms

aCRM – Analytical CRM

AI – Artificial Intelligence

AIC – Akaike Information Criterion

AUC – Area under the receiver operating characteristic (ROC) curve

BA – Business Analytics

BI – Business Intelligence

BIC – Bayesian Information Criterion

B2B – Business to business

B2C – Business to consumer

CART - Classification And Regression Tree

CRM – Customer Relationship Management

DM – Data Mining

ERP – Enterprise Resource Planning

EU – European Union

FN – False Negative

FP – False Positive

HR – Human Resources

ID3 - Iterative Dichotomiser 3

IT – Information Technology

KPI – Key Performance Indicator

MAPE – Mean Absolute Percentage Error

OR – Odds Ratio

R&D – Research and Development

ROC (Curve) - Receiver operating characteristic (curve)

RFM – Recency, Frequency and Monetary

RFI – Recency, Frequency and Intensity

RM – Relationship Marketing

ROI – Return on Investment

ROS – Return on Sales

SME – Small and Medium Enterprise

TN – True Negative

TP – True Positive

Introduction

Nowadays, it is evident the importance of technologies and data in the growth of organisations (T. H. Davenport & Harris, 2007). Companies build relationships with their customers over time through a variety of interactions (sales, calls and emails, social media connections, etc.) (Sharda et al., 2018) and they accumulate valuable data regarding customer behaviour from all these customer touch points (M. Berry & Linoff, 2004). In parallel, companies are gaining competitive advantage by changing the focus from the product/service to the customer, adopting customer-centric marketing strategies (Peppers & Rogers, 2016). It is in this context that many companies have adopted customer relationship management (CRM) systems, that empower firms by facilitating their relationships with customers, storing and managing relevant customer data (Kotler & Keller, 2012; Payne & Frow, 2005). Data analysis is a great help in the decision-making process of organisations, especially those focused on their relationships with customers (Waller & Fawcett, 2013).

To adopt customer centred approaches, it is important that companies understand their customers wants and needs (Verhoef et al., 2002), and one important aspect to fortify and reinforce customer relationships is the identification of groups of customers who can be treated similarly. Customer segmentation allows to define the target market for personalised marketing actions (Miguéis, 2012), thus increasing customer retention.

The present investigation resulted from a curricular internship held at the Marketing and Business Development department of Inova+, a management consulting firm specialised in supporting the growth of organisations, and it aims to support the CRM system of the company and its customer's management strategies by analysing and interpreting data retrieved from both the accounting and the CRM system of the company. The objective is to gain customer

knowledge and retrieve meaningful business insights in order to help the company fortify the relationship with its customers, improving customer retention and sales performance.

In this context, this research focuses on four main aspects in analysis: the current CRM performance of the company, the future prospect of the volume of business, the customer's segmentation, and the factors that influence a commercial proposal adjudication. Therefore, the following research questions are proposed:

- Is it possible to gather some insights regarding CRM performance from the CRM system of Inova+?
- Is it possible to segment Inova+ customers based on their value to the company? What are the Inova+ customer's segments?
- What is the probability of a commercial proposal being adjudicated and which factors influence it?

In this research the company Inova+ is used as a case study, where a quantitative approach was followed with the application of descriptive and predictive analytics tools to the firm's data through the construction of performance dashboards, the development of a time-series forecasting model and the application of the data mining techniques of clustering to segment the customers. Classification was also used to predict the commercial proposals probability of adjudication.

The structure of this thesis is divided into 5 main chapters. Chapter 1 presents a literature review addressing the concepts of marketing, customer relationship management and market segmentation. Additionally, it introduces the B2B context in which management consulting firms operate and respective performance measures. The concepts of business analytics, data mining and its main techniques used to support analytical CRM are also explored, emphasising

clustering and classification. Chapter 2 concerns the presentation of the Inova+ company, exploring its areas of expertise and CRM system. Chapter 3 outlines the chosen methodology and the data employed in this study. Chapter 4 refers to the construction and analysis of the four main aspects of this thesis, from the performance dashboards to the proposals predictive (classification) models. In this chapter the principal results are presented and discussed according to the objectives of the study. Lastly, in the final chapter the main conclusions and managerial implications are presented, as well as the limitations of the study along with some suggestions for future research.

Chapter 1: Literature Review

This chapter is divided into five main parts and has as main objective to present a summary of the literature regarding marketing and customer relationship management concepts, market segmentation (also referred to as customer's segmentation) and the B2B context in which management consulting firms operate. The final section of chapter 1 aims to introduce the era of data organisations live nowadays, giving emphasis to business analytics, data mining and its main techniques used to support analytical CRM.

1.1 Marketing and Relationship Marketing

Marketing started by being centred in the product and the transactions, with the concept of marketing mix (Borden, 1964) and the four P's of marketing (Product, Price, Place and Promotion) emerging in literature by McCarthy in 1960 (McCarthy & Perreault, 1993). Over time, the definition and scope of marketing has evolved.

Currently, marketing is about identifying and fulfilling human needs, and can be defined as the art and science of identifying the target audience and attract, maintain and develop the customers base through the creation of superior customer value by satisfying the needs and requirements of the target market (American Marketing Association, 2017; Chartered Institute of Marketing (CIM), 2015; Kotler & Keller, 2012).

With the growth of competition and the maturity of the industries, there has been a changeover from transactional marketing to relationship marketing (Gummesson, 1987; Payne, 2005; Wilson & Gilligan, 2005). Relationship marketing (RM) appeared in the 1980s (it was first mentioned in the literature by Berry in 1983) due to the shift of focus from customer acquisition to customer

retention (L. L. Berry, 1995, 2002; Kotler & Keller, 2012; Payne, 2005; Sheth, 2002). Relationship marketing is defined as "attracting, maintaining and enhancing customer relationships" (L. L. Berry, 1983, 2002). Coviello et al. (1997) defines RM as an integrative marketing that is cross-functional and includes database marketing, interaction marketing and network marketing. According to Gummesson (1987), RM (or interactive marketing as the author named it) refers to all the activities a company pursues to build, maintain and develop customer relationships, where interaction is key, especially for services. Kotler & Keller (2012) claim that the RM goal is to establish enduring relationships with people (namely customers, employees, marketing partners and the financial community) and organisations in order to earn and retain their business. According to Gronroos (1990), customer relationships do not necessarily need to be long-term relationships, as long as it is beneficial for both parties.

In short, relationship marketing is concerned with the management and improvement of organisation's relationships over time, both with their customers but also with other stakeholders (Coviello et al., 1997; Hollensen, 2015; Kahan, 1998; Payne, 2005), as Figure 1 illustrates.

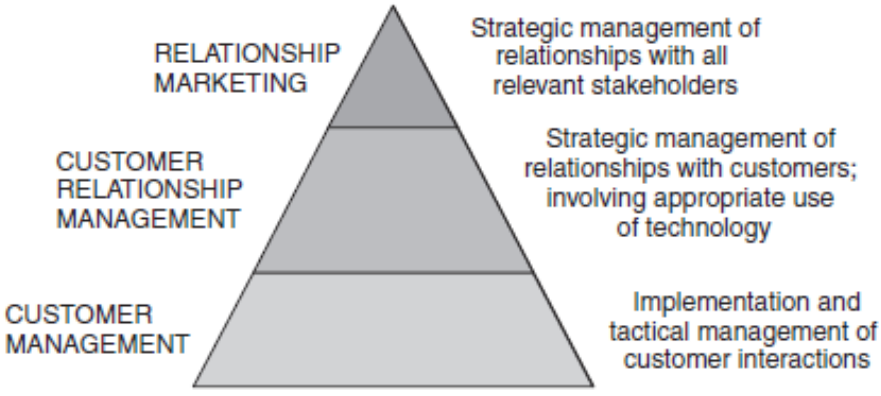

Figure 1: Relationship Marketing (RM), Customer Relationship Management (CRM) and Customer Management (Payne, 2005)

1.2 Customer Relationship Management (CRM)

CRM is based on the concept of relationship marketing (Payne, 2005; Payne & Frow, 2005), and enables companies to collect and analyse customers real-time data, allowing them to develop relationships with their customers (Gummesson, 2004; Kotler & Keller, 2012). In the following sections regarding CRM, a general definition and main categories will be presented, alongside with the challenges companies face when adopting and implementing an CRM system. CRM importance in providing customer knowledge and the relation with firm performance is also explored in this section.

1.2.1 General Definition

Globalisation has made competition between firms fiercer (T. H. Davenport & Harris, 2007; Sheth, 2002), as the market (including competitors, suppliers and customers) is now at a worldwide level (Ling & Yen, 2001). Furthermore, in the last decades the world has witnessed a massive transformation and evolution of information technology (IT) (H. Chen et al., 2012). This progress was not exclusive at the individual level. Enterprises are now more technological than ever and can count with the support of intelligent systems to control more efficiently their processes (Kusiak, 2018), both regarding customers, suppliers, employees, budgets, warehouses, logistics and stores.

One of these intelligent systems is the Customer Relationship Management (CRM) system that helps companies manage their relationships with customers more efficiently (Sharda et al., 2018; Sheth, 2002), and there are a myriad of perspectives and definitions of it.

According to Payne & Frow (2005), CRM provides companies the opportunity to use customers data and information to understand them better, leading enterprises to customise market offerings, services, programs,

messages, and media according to their customers characteristics and requirements (Kotler & Keller, 2012).

One main aspect of CRM is that it needs to be a cross-functional system focused on relationship development (Becker et al., 2009; I. J. Chen & Popovich, 2003; Payne, 2005), contemplating elements such as people, processes, operations and marketing (Payne & Frow, 2005), technology and content (I. J. Chen & Popovich, 2003; Krizanova et al., 2018).

According to Ling & Yen (2001), CRM consists of a set of processes and interactive systems that support a business strategy to build long-term valuable customer relationships (Krizanova et al., 2018; Ling & Yen, 2001; Payne & Frow, 2005). This definition puts emphasis on the importance that managing customers' relationships has in maximising companies' value, and is in line with the CRM definitions of many other authors such as Chen & Popovich (2003), Gummesson (2004), Payne (2005), and Reimann et al. (2010).

For Ngai et al. (2009), it is important that we view CRM "as a comprehensive process of acquiring and retaining customers, with the help of business intelligence, to maximize the customer value to the organization". According to Reinartz et al. (2004), CRM process entails three distinct stages: initiation (with the identification of potential customers and acquiring them, or the regain of customers); maintenance (retaining customers is the goal, and this phase includes up-selling and cross-selling) and termination (customers exit). In turn, Sin et al. (2005) settles that CRM is a multi-dimensional concept, and it has four components: key customer focus (customer-centric focus), CRM organisation, knowledge management and technology-based CRM.

For further details on CRM and its definitions, please refer to Ngai (2005) and Payne & Frow (2005).

1.2.2 Main Categories

Referring to the viewpoints of Kim & Mukhopadhyay (2011), there are two broad categories of CRM technologies: the targeting-related CRM (similar to analytical CRM), that intends to increase the knowledge of the company regarding customers (it includes marketing automation, analytics and business intelligence), and the support-related CRM (similar to operational CRM) which is associated with the relationship building with customers.

In contrast, Dyché (2001) and Payne (2005) argue that CRM can be divided into three types or categories: Operational CRM, Collaborative CRM, and Analytical CRM. Operational CRM refers to the automation of business processes (Ngai et al., 2009) and the day-to-day operations of a firm (Dyché, 2001; Payne, 2005; Peppers & Rogers, 2016). It involves collecting customers data through different customer contact-points (Kimball & Ross, 2013) such as contact centre, contact management system, mail, fax, sales force and web (Mirzaei & Iyer, 2014). With the Operational CRM, the operational efficiency of the firm is improved (Xu & Walton, 2005). Collaborative CRM enables a direct interaction between customers, the firm and its employees (Payne, 2005), through a variety of channels (although typically via web). It enables to improve the quality of customer interactions (Dyché, 2001). This type of CRM is based on the individualisation of the marketing activities and value proposals towards the customers, developing and enhancing customer loyalty (Kracklauer et al., 2004). Usually these CRM systems are integrated with enterprise-wide systems such as ERPs (Xu & Walton, 2005). Analytical CRM (aCRM) refers to the use and analysis of data originated from the customers touch-points and the operational CRM (Dyché, 2001; Payne, 2005) that is usually stored in databases and data warehouses (Kimball & Ross, 2013). A combination of analytical tools are adopted to analyse and interpret the data (Mirzaei & Iyer, 2014). Data mining is one of these tools and is used to identify

patterns and customers behaviours (Xu & Walton, 2005), to help firms in their decision-making processes (Kimball & Ross, 2013). Quoting Ngai et al. (2009), analytical CRM refers to "the analysis of customer characteristics and behaviours so as to support the organisation's customer management strategies". Analytical CRM contributes to a firm's competitive advantage by enabling it to build and augment customer value (Peppers & Rogers, 2016), while also empowering it with the ability to improve marketing analytics, customer service and sales performance (Ranjan & Bhatnagar, 2011).

According to Kracklauer et al. (2004), there are four dimensions of aCRM: Customer Identification (this entails the definition of target clients and customer segmentation); Customer Attraction (marketing and sales efforts, including promotions or free samples for example); Customer Retention (company needs to ensure costumers satisfaction and loyalty); Customer Development (Creating opportunities for cross-selling or product and service bundling in order to increase customer lifetime value). Figure 2 shows the four CRM dimensions and the selected tactical tools for achieving the respective core tasks. Some measures like benchmarking and one-to-one marketing are useful in other phases of customer relationship management.

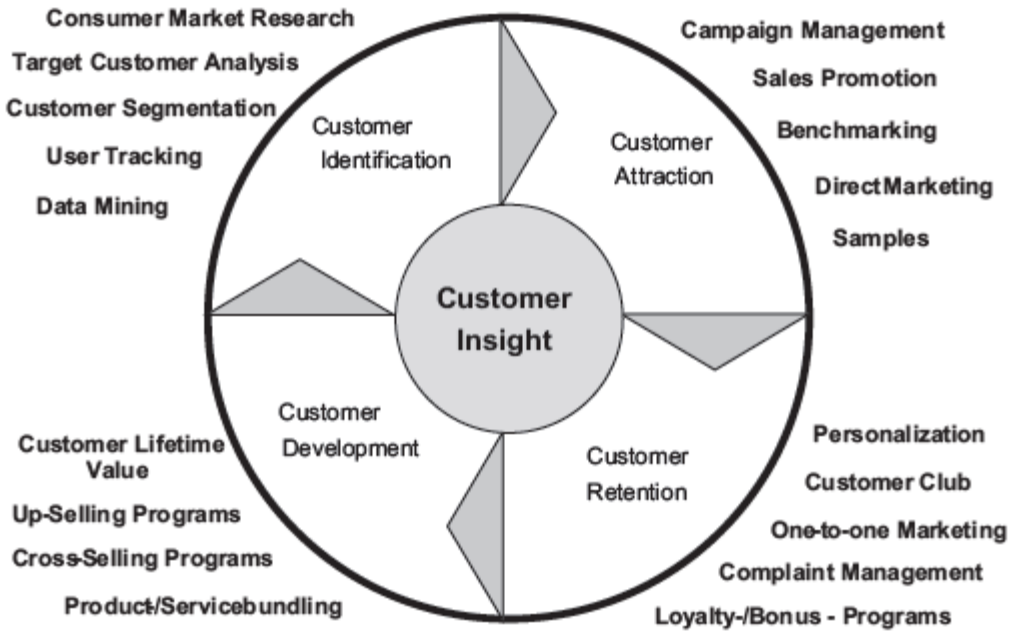

Figure 2: CRM dimensions and respective tactical tools (Kracklauer et al., 2004)

1.2.3 Adoption and Implementation

The increase in market competition (Bose, 2009), the advances of technology, and the fact that it costs less to retain a customer than to attract a new one (Park & Kim, 2003) are the main business drivers for the implementation of CRM systems (Ling & Yen, 2001).

As to achieve the potential CRM systems can provide to firms that implement it, it is crucial that it exists a commitment in the whole company to the system (Stringfellow et al., 2004), and it should be well integrated and in line with the company goals and strategy (Bohling et al., 2006). Therefore, there is a need that the system itself is cross-functional (I. J. Chen & Popovich, 2003; Kimball & Ross, 2013; V. Kumar et al., 2006) and process-oriented (Payne & Frow, 2005), providing organisational learning and a culture of data sharing across the entire firm (I. J. Chen & Popovich, 2003; Peltier et al., 2013). This alone can be a factor of disruption or success regarding CRM systems. According to Cooper et al. (2008), it is practically mandatory that CRM projects

involve the whole firm, from the general management (more strategic) to the IT and Sales and Marketing departments. The customer focus should not be isolated in the marketing departments (Reinartz et al., 2004). Usually, it is the sales personnel that input customers' information on CRM, but not always in a consistently way (Stein et al., 2013), and this can be a factor for CRM not achieving its full potential.

Although companies are investing a great amount of resources in the adoption and implementation of CRM systems, many of them do not see satisfactory results (S. H. Kim & Mukhopadhyay, 2011) or the expected return in terms of performance (Jayachandran et al., 2005). In order to achieve a successful implementation of CRM in a firm, technology is not enough (I. J. Chen & Popovich, 2003; Parvatiyar & Sheth, 2001), and the personnel is key (Becker et al., 2009; Cambra-Fierro et al., 2017; Gummesson, 2004; Reinartz et al., 2004). Business and technology need to work together to ensure the success of CRM (Kiron & Bean, 2013; V. Kumar et al., 2006; Ling & Yen, 2001). The organisational structure and commitment towards the CRM, alongside with the management support and the employees motivation to use it is vital for its success (Becker et al., 2009).

1.2.4 CRM and the Importance of Customer Knowledge

Ranjan & Bhatnagar (2011) assert that an organization that not only implements CRM effectively but also does an analysis of the customer data is on the right path to an effective management, aligning the strategy of the company with the actions of marketing towards their clients, leading to an increase in their satisfaction. CRM, along with the posterior data analysis, empowers firms to carry out data-driven decision making strategies (Stein et al., 2013).

It is crucial for companies to know their customers behaviours and patterns in order to be able to satisfy their needs and requirements, and CRM helps optimize this process of getting to know the customers (Verhoef et al., 2002), and it effects positively the organizational performance (M. Kumar & Misra, 2020; Park & Kim, 2003). An effective CRM enables firms to generate customer insights, and they should use these insights to tailor the services itself (Bailey et al., 2009). As Tseng & Wu (2014) stated, "customer knowledge has a positive influence on service quality and CRM is the partial intervening variable between customer knowledge and service quality". This is in line with Mithas et al. (2005) research that establishes that firms with CRM systems gain customer knowledge and improve customer satisfaction, and with Hallencreutz & Parmler (2019) and Durdyev et al. (2018) researches that determine that service quality is the main driver and influencer of customer satisfaction.

According to Zineldin (2006), by creating and strengthening these customer relationships with the establishment of an CRM system, a company can outperform its competitors by being one step ahead in terms of understanding what their customers truly want and when they want it, increasing their satisfaction and latter loyalty and retention (Lages et al., 2008; Lam et al., 2004).

1.2.5 CRM and Firm Performance

Customer Relationship Management (CRM) systems make it easier for companies to manage more effectively their customers relationships across the different stages and touch-points (Mithas et al., 2005) and this information systems create value to the businesses that adopt them successfully (Kimball & Ross, 2013), leading to a potential escalation in firm performance and competitive advantage (Krizanova et al., 2018; Zahay, 2008).

Nowadays, it is known that data analysis is a great support to decision-making in firms (Hallikainen et al., 2020; Waller & Fawcett, 2013). The use of business analytics in CRM, namely the firm's data management capability and customer response capability, accompanied with IT competence, can lead to an improvement in the performance of the CRM system (Nam et al., 2019). Analytics bring knowledge to the organisations, with the ability to discover patterns in data and integrate it, gaining meaningful insights, and leading organisations to success and increased value (Kiron & Bean, 2013). Customer analytics allows a more accurate targeting of marketing actions (Stein et al., 2013), improving firm performance, namely customer satisfaction (Anderson et al., 1994) and sales growth (Hallikainen et al., 2020). According to Kitchens et al. (2018), companies that embrace customer analytics create potential to achieve competitive advantage.

Mithas et al. (2011) states that the information management capability has a great importance in influencing firm performance and success, and the CRM system is an enabler of this, entailing three organisational capabilities: customer management capability, process management capability, and performance management capability.

Contrary to these views, Reimann et al. (2010) argues that CRM does not impact directly firm performance. Instead, the business strategies of cost leadership and differentiation are the link between the CRM system and the company performance. Because CRM enables firms to better known their customers, it gives them the opportunity to adapt their offerings to the customers' requirements (differentiation), and it facilitates the demand forecasting, allowing the firms to choose the clients they want to retain, increasing efficiency and lowering costs (cost leadership) (Reimann et al., 2010).

1.3 Market Segmentation

As a component of marketing strategy, market segmentation was originally introduced by Smith in 1956, and is based on the economic theory of imperfect concurrency developed by Robinson (1969) and Chamberlin (1949) in the 1930s. Segmentation represents a firm's necessity of rational and precise adjustment of product(s) supply and marketing efforts to meet customers' requirements and demand.

According to Kotler & Keller (2012), market segmentation is the identification of distinct groups of customers who differ in their needs and wants. Market segmentation is vital for companies, enabling them to promote and enhance their relationship with customers, improving sales and profits (Miguéis et al., 2012). This ability to identify profitable customers and build long-term relationships with them, increasing loyalty and customer retention, is a key competitive factor to a firm in today's business world (Lee & Park, 2005; Yankelovich & Meer, 2006).

In a B2B environment, it is important for companies to understand their customers' needs and characteristics (Hosseini et al., 2010). Customer segmentation, included in the customer identification phase of CRM (Kracklauer et al., 2004; Ngai et al., 2009), refers to the division of the database of all customers into smaller segments, where customers who belong to the same group have similar characteristics (Jing Wu & Lin, 2005).

Customer segmentation is beneficial for firms, allowing them to identify potential target groups (Hollensen, 2015; Jing Wu & Lin, 2005) and develop value propositions for them (Bailey et al., 2009) according to their specific needs (Bose, 2009; Xu & Walton, 2005). Moreover, it is known that it is less costly for companies to invest money in retaining customers than to attract new ones (Cheng & Chen, 2009; Park & Kim, 2003), especially in services (Ennew & Binks, 1996).

The application of data mining techniques to identify the customers segments, and posteriorly develop efficient retention campaigns and incentives for those clients who are at risk, will increase customers retention and the profitability of the firm (Jahromi et al., 2014). Because customer retention is more profitable than customer attraction and acquisition, it is crucial for companies to identify the strategically important customers and their segments (Xu & Walton, 2005). When customers are satisfied with the supplier they will stay in that relationship, increasing customer retention (Eriksson & Vaghult, 2000).

According to Dagger & Sweeney (2007), for service firms, data regarding customers and their experience can be useful in segmenting the customers as novice and longer-term customers, allowing the company to address each group according to their different needs, customising strategies and subsequently enhancing profitability. These customers segmentation is important because "perceptions of service quality are not stable over time" (O'Neill & Palmer, 2001).

Customer segmentation can be based on different aspects, namely geographic (countries, regions, etc.), demographic (age, income, gender, etc.), psychographic (personal traits, values, lifestyle) or behavioural factors (customers are grouped on the basis of their use or response to a product or service) (Kotler & Keller, 2012). One of the most popular behavioural segmentation models regarding customers is the RFM model. This model is explained in the following paragraphs.

RFM Model

Introduced by Alden's catalogue company in the 1920s (McCarty & Hastak, 2007; Roel, 1988), the RFM (Recency, Frequency, Monetary) model has been extensively applied as a behavioural analysis technique for customer segmentation (Kahan, 1998), allowing companies to know who are their most

recent buyers, the most frequent customers, and the largest spenders (Hughes, 1994). For some authors, it is also known as RFI (Recency, Frequency, Intensity) model (Kimball & Ross, 2013). Furthermore, it is important to note that it is not applicable to the prospecting for new customers (McCarty & Hastak, 2007), since we would not have the data required about the clients' purchase historical (transactional data).

RFM segmentation analysis allows to differentiate important customers from large datasets by three variables: recency, frequency and monetary (value) (Cheng & Chen, 2009; Hu & Yeh, 2014; Jing Wu & Lin, 2005). Recency refers to the length of a time period between the last purchase and present (Hosseini et al., 2010; Wei et al., 2010), so lower recency implies it is a more recent customer. It can be assessed in days, and it allows companies to answer the following question: how many days has it been since the customer's last purchase? (Kimball & Ross, 2013). Frequency measures the number of transactions (purchases) made in a particular time period (Wei et al., 2013). Monetary (value) represents the total amount of money spent during a specified time frame (Hu & Yeh, 2014; Jing Wu & Lin, 2005).

It is important to define the scaling for RFM (Cheng & Chen, 2009), and one of the most used approaches was developed by Hughes (1994), that considers that the three categories have identical weights and are equal in importance. In practice, we start by sorting the data in a descendant order from newest to oldest date, and then divide it into five equal parts (quintiles), assigning a score of 5 to the top group (more recent customers), 4 to the next, and so on (Jing Wu & Lin, 2005). Then the variables frequency and monetary are sorted in the same way, divided in quintiles and subsequent scored from 5 to 1 (Wei et al., 2010). Therefore, the database is divided into 125 roughly equal groups according to the categories recency, frequency, and monetary value.

Customers with higher scores (of 5) are typically the most profitable (Hosseini et al., 2010; Hughes, 1994).

Once we have the three attributes (RFM), another approach we can use is the k-means algorithm for clustering (Cheng & Chen, 2009). Customers can be segmented into different groups that exhibit similar RFM values, and then based on this segmentation, companies develop customised marketing actions and strategies (Hu & Yeh, 2014). This approach has been applied to different industries and contexts such as the online retail industry (D. Chen et al., 2012) and e-commerce (Jun Wu et al., 2020), the hairdressing business (Wei et al., 2013), the automotive supply chain industry (Hosseini et al., 2010) and the fintech industry (Sheikh et al., 2019).

1.4 Business to Business (B2B) Context

A context in which CRM gains particular relevance is the business to business (B2B) context. In a B2B market (companies who sell to enterprise customers (Peppers & Rogers, 2016)), it is important that firms personalise and tailor their offers according to the customer needs, and CRM helps and optimises this process (Verhoef et al., 2002).

Over time, the business to consumer (B2C) context has received more attention than B2B markets, both from academics and managers, hence there is a "B2B knowledge gap" (Lilien, 2016). The B2B domain is more heterogeneous than the B2C, not only in terms of customers size but also on the performance needs of the companies (Lilien, 2016).

Nowadays, companies deal with huge amounts and variety of data (Erevelles et al., 2016), and this is often unstructured data (emails, phone calls, social media, etc) (Kiron & Bean, 2013). Marr (2017) suggests that "artificial intelligence (AI), machine learning, and big data are facilitating operations for B2B companies". AI

can analyse this unstructured data and assist in the lead generation, and machine learning supports the monitoring and analysis of customer behaviour and patterns in the B2B market, helping predict sales and improving content marketing (T. Davenport, 2018; T. Davenport et al., 2020; Marr, 2017).

In the B2B context, CRM utilisation impacts sales performance and its process effectiveness (Rodriguez & Honeycutt, 2011). In the following section more details regarding performance measures and indicators are provided.

1.4.1 Performance Measures and Indicators

Business performance management and measurement is relevant, especially nowadays that the competition is global and the exclusive use of the financial dimensions to measure performance is obsolete (Kaplan & Norton, 1992; Neely, 1999). Particularly in services, non-financial measures are important to measure performance, and flexibility stands out as one of them, namely personnel flexibility, customisation and the ability to adjust capacity promptly (Arias Aranda, 2003). Additional examples of non-financial performance measures are customer satisfaction and retention, brand image, trust, reputation, and customers loyalty (Hallencreutz & Parmler, 2019). Successful firms measure customer satisfaction as this can be a factor that leads to repeat business and higher profitability (Kotler & Keller, 2012). Customer satisfaction has long been used as a source of competitive advantage for many companies, and it is crucial in evaluating the service quality (Grigoroudis et al., 2013; Parasuraman et al., 1988).

Although there is a lack of academic works addressing the topic of CRM performance measurement, H.-S. Kim & Kim (2009) propose a very complete CRM scorecard to diagnose and appraise a firm's CRM practice. This CRM scorecard contemplates four perspectives: organisational performance

perspective, customer perspective, process perspective, and infrastructure perspective. The first one is related with the company's profitability and performance (Reinartz et al., 2004). In the customer perspective, it is important to assess customer satisfaction (Fornell, 1992) and subsequent loyalty (Agustin & Singh, 2005). In terms of process perspective, since CRM is aligned with the corporate strategy to maintain and develop customer relationships (Lindgreen et al., 2006; Park & Kim, 2003), it is crucial to evaluate customer acquisition and retention and cross/up sell rates (Reinartz et al., 2004). The infrastructure perspective contemplates aspects related with the CRM technology, the human capital of the company (employees behaviour (Donavan et al., 2004) and satisfaction (Maxham & Netemeyer, 2003)), the organisational alignment towards the CRM system (with the presence of training and reward systems for example) and the organisational culture and market orientation (Jaworski & Kohli, 1993).

Regarding business performance measurement, there is no strict method to consider, as it depends on the industry and the company itself, however the majority of authors agree that it needs to combine financial and non-financial measures (Neely, 1999; Wilson & Gilligan, 2005). According to Mithas et al. (2011), business performance measurement needs to consider four different dimensions, including customer-focused performance (customer satisfaction is comprised in this dimension), financial and market performance (revenues, profits and financial ratios are analysed), human resource performance (employee satisfaction is relevant here), and finally organisational effectiveness (comprising the level of innovation, production and supply chain flexibility of the firm). For Sin et al. (2005), business performance is a mix of marketing performance (where trust and customer satisfaction are covered) and financial performance (with the typical financial ratios of return on investment (ROI) and return on sales (ROS)). Emphasising customer

relationship performance, customer satisfaction and customer retention are good indicators of performance (Jayachandran et al., 2005).

Performance targets need to be consistent with long-term goals and with the strategy of the company (Grant, 2016). Developed by Kaplan & Norton (1992), the balanced scorecard provides a framework that is aligned with the strategy and vision of a company, and it incorporates four perspectives (customer perspective, financial perspective, internal business perspective and innovation and learning perspective), giving managers an overview of the performance of the business. Enterprises can count with the help of scorecards and dashboards in their business performance management to analyse and visualise a variety of metrics and indicators (H. Chen et al., 2012; Parmenter, 2020; Sharda et al., 2018).

Data visualisation is part of descriptive analytics and is vital for making good business decisions as it facilitates the contextualisation and communication of data (Evans, 2014; Sharda et al., 2018). Dashboards are a data visualisation tool that display the most important information and enable the analysis and monitoring of KPIs and other metrics, usually in a single screen (Eckerson, 2010; Few, 2005). KPIs can be defined as indicators that focus on aspects of the company's performance that are critical for its success (Parmenter, 2020). Performance dashboards have, as the name indicates, the purpose of measure performance and can be divided in three types: operational dashboards (monitor operational processes and activities as they occur), tactical dashboards (measure and analyse the performance of departmental activities, processes, and goals), and strategic dashboards (track the progress and manage the execution of strategic objectives on an enterprise scale, often implemented using a balanced scorecard approach) (Eckerson, 2010; Kerzner, 2017).

1.4.2 Management Consulting Organisations

Management consulting organisations are part of the B2B market and are the archetype of knowledge-intensive firms (Starbuck, 1992; Werr & Stjernberg, 2003) since their main asset is the expertise and experience of the personnel (Kipping & Engwall, 2002). In management consulting companies people, processes and systems work as a knowledge creator (Anand et al., 2007). It is important that this type of organisations adopt continuous learning policies in order to keep the knowledge up to date (Starbuck, 1992). Management consulting firms are typically structured as a pyramid, with partners and seniors at the top level, managers at the medium level, and junior consultants and trainees at the bottom (Jacobs & Chase, 2018).

For successful management consulting firms, the client-consultant communication is fundamental to provide a service with quality , and the top three performance indicators of consulting firms success are customer satisfaction, profitability and repeat business (Kumar et al., 2000). Management consulting organisations that use CRM systems collect and store data from customers more easily, and they can assess performance measures such as the customer retention rate, projects margins and profits, customer lifetime value and the net promoter score (Eccles, 1991; Ibatova et al., 2018). In the new era of data, it is also important to assess the customers engagement in social media, the leads generated and the email (mailings) response rate (Marr, 2012).

1.5 The New Era of Data

1.5.1 Business Analytics

Nowadays organisations are loaded with information and data from all types of varieties and sources (T. H. Davenport & Patil, 2012). In order to stay

above competition, that is getting more intense (Dawar & Bendle, 2018), it is crucial that they know how to leverage value from all this data, and business analytics (BA) is exactly about that (Acito & Khatri, 2014; Evans, 2014). Companies are more productive and profitable when data-driven decisions are made (McAfee & Brynjolfsson, 2012; Sharma et al., 2014). Hence, top performing firms apply business analytics five times more than lower performers (LaValle et al., 2011).

Data is valuable (it was even considered the new oil by Acito & Khatri in 2014), and the analysis of this important asset can provide insights that are then converted into decisions that create value for the company (Sharma et al., 2014). This requires an aligning of strategy and business performance management with analytics (Acito & Khatri, 2014), where strategy can be defined as "a purposeful plan of action that requires making choices regarding the deployment of resources" (Mintzberg, 1987).

The term business intelligence (BI) appeared first than business analytics (H. Chen et al., 2012), and can be defined as the "use of computers to collect, manage, analyse and report data" (Luhn, 1958). Business analytics is a subset of BI (T. H. Davenport & Harris, 2007).

Business analytics (or simply analytics) can be seen as the process of making decisions and recommendations for activities based on insights extracted from historical data (Sharda et al., 2018; Turban et al., 2011). It strengthens an organisation's decision making process (H. Chen et al., 2012) and problem solving skills (Liberatore & Luo, 2010), improving organisational performance (Holsapple et al., 2014). According to Evans (2014) and Davenport & Harris (2007), business analytics is the use of data, IT, statistics, quantitative methods and mathematical or computer-based models in business contexts with the purpose of gaining improved insights about business operations, and make

better data-based decisions. For an overview regarding the definitions of BA see Holsapple et al. (2014).

Business Analytics is divided into three scopes, as described in Figure 3, that are the following: descriptive analytics, predictive analytics, and prescriptive analytics. According to Evans (2014), descriptive analytics refers to the analysis of data to understand past and current business performance, predictive analytics is about analysing historical data to predict future behaviours and patterns., and prescriptive analytics uses optimisation to identify the best business alternatives and actions (for example, to minimise or maximise some objective).

Figure 3: The three scopes of Business Analytics (Sharda et al., 2018)

Predictive analytics is about making predictions, inform decisions and forecast future movements of the customers and industries (Mirzaei & Iyer,

2014), and has been used in the past by companies in order to increase performance and get competitive advantage (Calixto & Ferreira, 2020).

In the context of aCRM, predictive analytics tools (including the application of data mining techniques (explored in detail in sections 1.5.2 and 1.5.3)) are common methods of exploring customer data (Mirzaei & Iyer, 2014).

1.5.2 Data Mining

Data Mining (DM) can be defined as the process of discovering useful information in large repositories of data (Tan et al., 2005). With the preparation and analysis of data, DM aims to obtain knowledge and insights from it (Tsiptsis & Chorianopoulos, 2009). Therefore, data mining is a term that refers to the discovery or mining of knowledge from large amounts of data (Han et al., 2011; Sharda et al., 2018; Turban et al., 2011). DM is also characterised as the process of discover unknown patterns in data (Bose, 2009; Witten et al., 2017). In order to do so, DM combines advanced statistics methods, mathematics and artificial intelligence techniques (Sharda et al., 2018).

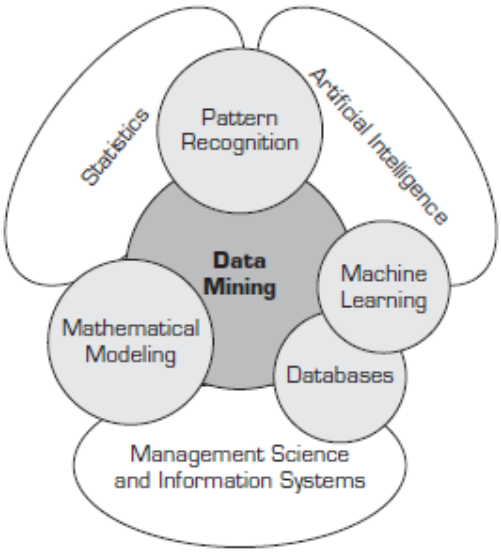

Figure 4: Data Mining as a combination of multiple disciplines (Turban et al., 2011)

Data Mining is a step in the Knowledge Discovery in Databases (KDD) process, which is the process of transforming raw data from the data warehouse into useful information and knowledge (Fayyad et al., 1996; Tan et al., 2005). The KDD process involves several stages: data selection, data preprocessing, data transformation, data mining, and finally interpretation/evaluation (Sharda et al., 2018), as seen in Figure 5.

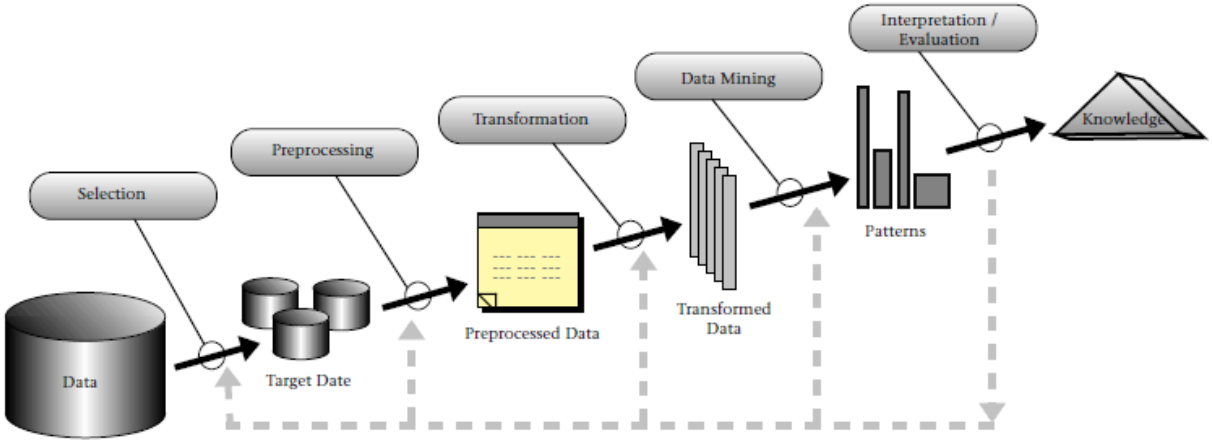

Figure 5: The KDD Process (Fayyad et al., 1996)

Data selection initiates with what will be the application domain and also includes choosing the relevant data to the analysis (Fayyad et al., 1996). Data preprocessing is the step regarding cleaning the data, by dealing with missing values, noisy data and inconsistent data, and is one of the most time-consuming stages of the KDD process (Tan et al., 2005). The next step is data transformation, and it consists of converting the data to the desired formats, by aggregating or normalise it, to then apply data mining algorithms (Han et al., 2011). Data mining is the subsequent step, and consists in discover patterns in the prepared dataset using intelligent methods (Han et al., 2011). With the results of the data mining stage, it is time to evaluate and interpret the patterns discovered. Data visualisation and representation techniques can then be applied to present the knowledge acquired from the data (Sharda et al., 2018).

1.5.3 Data Mining Applications on CRM

The analysis of customer characteristics and behaviours is the foundation for the development of a competitive CRM strategy (Erevelles et al., 2016), and for this the application of data mining techniques such as classification, association, clustering or forecasting can be extremely helpful (Bose, 2009; Cheng & Chen, 2009; Ngai et al., 2009). This will allow firms to be more sensitive and aware of their customers' needs (Bose, 2009; D. Chen et al., 2012), leading to a more efficient and effective marketing effort (Gummesson, 2004; Lee & Park, 2005). Mining customers' data will uncover hidden knowledge (Ngai et al., 2009; Ranjan & Bhatnagar, 2011) and bring insights to the companies (Tsiptsis & Chorianopoulos, 2009), and CRM facilitates it (I. J. Chen & Popovich, 2003). Data mining enables firms to automate the analysis of complex KPIs (Calixto & Ferreira, 2020).

In line with the characterisation of the four dimensions of analytical CRM by Kracklauer et al. (2004), Ngai et al. (2009) illustrates that data mining techniques can be applied to all of these components (Figure 6).

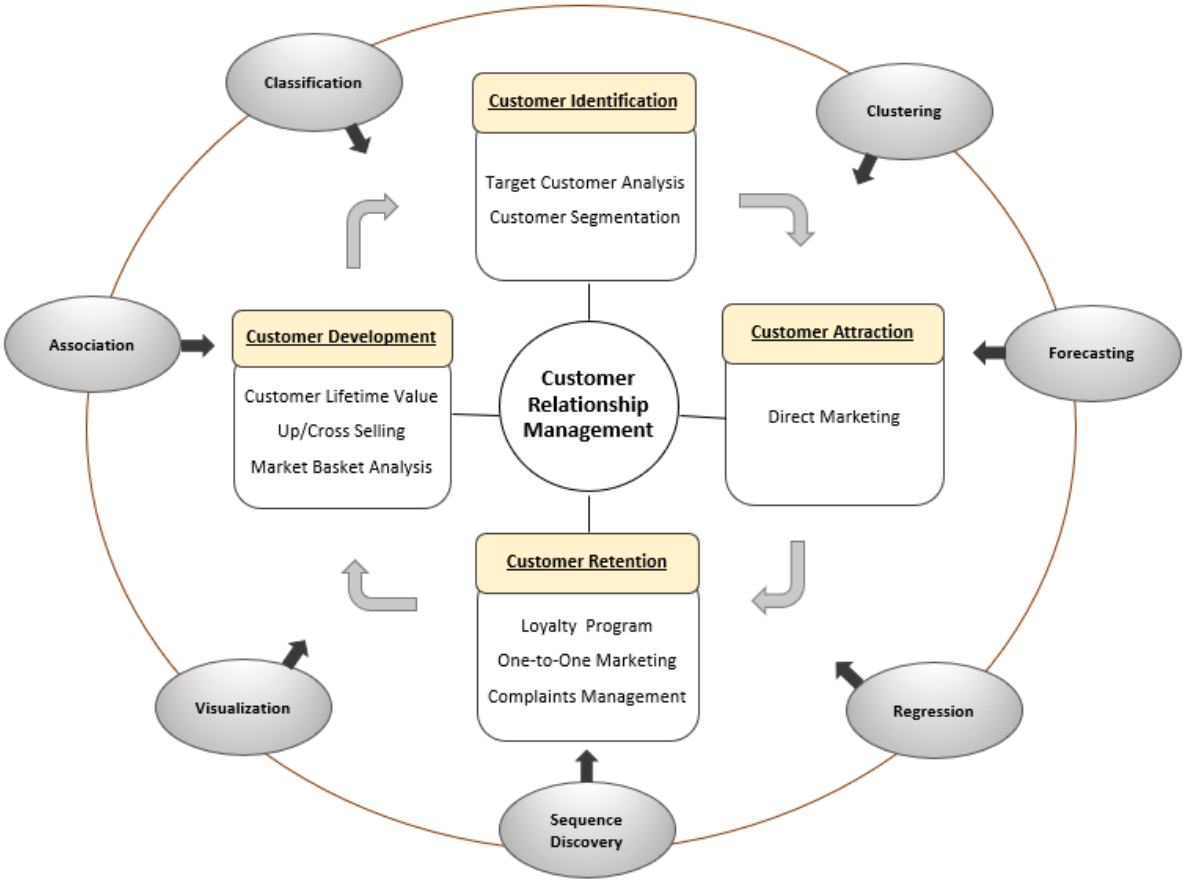

Figure 6: Classification framework for data mining techniques in CRM (Ngai et al., 2009)

Classic examples of data mining techniques are clustering, classification and association rules. The next sections present a brief explanation and summary of the two data mining techniques applied in this research: clustering and classification.

1.5.3.1 Clustering

Clustering is the process of grouping a set of data objects into multiple groups (clusters), with no predefined classes (Han et al., 2011; Shalev-Shwartz & Ben-David, 2014). The goal is that objects within a cluster be similar to one another and different from the objects in other groups (Tan et al., 2005). Most of the clustering algorithms require apriori specification of the number of clusters to find (Sharda et al., 2018; Tibshirani et al., 2001). Common

approaches to determine the number of clusters are the davies-bouldin index and the elbow curve. The davies-bouldin index (Davies & Bouldin, 1979) is a function of the ratio of the sum of intra-cluster scatter to inter-cluster scatter. A good value for the number of clusters is associated to lower values of this index. The elbow curve is a heuristic method (Thorndike, 1953) based on a plot of an error measure (the within cluster dispersion defined typically as the sum of squares of the distances between all items and the centroid of the correspondent cluster divided by the number of clusters) versus the number of clusters (k). As the number of clusters increases, the error measure decreases. The assumed appropriate number of clusters is in the "elbow" point (when the decrease of k smooths significantly). The most popular types of clustering techniques are the partitioning methods and the hierarchical methods (Tan et al., 2005).

Partitioning methods create clusters by optimising an objective partitioning criterion, such as the distance dissimilarity function. Given a database of n objects, it constructs k partitions of the data, where each partition represents a cluster. Each data object is in exactly one cluster. The most commonly used partitioning method is the k-means algorithm (Witten et al., 2017). The k-means algorithm developed by MacQueen (1967) assigns a set of n items to k clusters. The number of clusters is pre-defined by the analyst and each item is assigned to the nearest cluster based on the minimum distance between the item and the cluster mean (M. Berry & Linoff, 2004). This is an efficient and high-speed clustering algorithm, applicable for large datasets (Fred & Jain, 2002; Tsiptsis & Chorianopoulos, 2009).

Hierarchical clustering methods create a tiered decomposition of the given set of data objects. Based on how the hierarchical decomposition is formed, can be classified as agglomerative ("bottom-up") or divisive ("top-down"). In the agglomerative approach, starts with the points as individual clusters and,

at each step, merge the closest pair of clusters until all the groups are merged into one. In the divisive approach, it starts with all the objects in the same cluster (all-inclusive cluster) and, at each iteration, a cluster is split until each object is in one cluster, or a termination condition holds (Han et al., 2011).

1.5.3.2 Classification

The goal of classification techniques is the prediction of the class attribute value of the items. Unlike clustering, classification considers predefined classes. Classification is a supervised learning model in data mining (M. Berry & Linoff, 2004; Shalev-Shwartz & Ben-David, 2014). It is a two-step process, that consists in a training (learning) step where the predictive models are constructed (based on the analysis of a set of training data), and in a testing step (where the models are used to predict the class labels of new data instances where the class label is unknown) (Han et al., 2011; Mitra et al., 2002). Examples of common classification techniques are logistic regression, decision trees, random forests, nearest neighbours' classifiers (kNN) and the Naïve Bayes classifier.

Logistic regression is a classical statistical technique used for classification. It considers a binary dependent variable (the probability of an outcome occurring) constrained between 0 and 1, and assumes a linear relationship between the dependent variable expressed in the logit scale (applying a log-odds transformation) and the independent variables (Afifi et al., 2012, 2020). Besides its ease of use and robust results, one of the advantages of logistic regression is that it does not rely on assumptions of normality for the independent variables and may handle non-linear effects (Fahrmeir & Kaufmann, 1985; Janzen & Stern, 1998).

A decision tree predicts the class label associated with an instance by moving from a root node of a tree to a leaf (node that include the final classifications) (Shalev-Shwartz & Ben-David, 2014) and is based on recursive partitioning. Each non-leaf node (labelled with the attributes used to split the data) of the tree contains a split point that determines how the data will be further divided (Sharda et al., 2018). The decision trees generation process consists of two phases: tree construction and tree pruning (to avoid overfitting). The attributes to test and select during the tree construction phase vary according to the split criterion chosen. The information gain is the splitting mechanism used in the ID3 and C4.5 algorithms (Quinlan, 1993) and is based on the decrease in entropy after a dataset is split on an attribute, leading to more homogeneous branches. The gain ratio is a modification of the information gain criterion that considers the intrinsic information of a split (trade-off between the information gain we get and the diversity of branches we are creating). The gini index is used in CART algorithm (for more details see Breiman et al. (1984)) and uses a gini index instead of an entropy measure, where the attribute providing the smallest gini is chosen to split the node. The main advantages of the decision trees include the capacity to handle data measured on different scales (it is not necessary to normalise the data before proceeding with the classification) and the lack of assumptions about the frequency distributions of the data in each class (Friedl & Brodley, 1997). However, decision trees may suffer from overfitting, leading to suboptimal performances (Dudoit et al., 2002).

Random forests, a technique developed by Breiman (2001), is an ensemble method based on multiple decision trees (a forest) (Han et al., 2011). Each decision tree is generated based on different training sets, which are drawn independently. Moreover, in each partition of the data in a tree of the forest, a random sample of attributes to split the data is randomly chosen among those

available. Once the trees have been constructed, the algorithm stores the class assigned to each test item in all trees, and the final class, defined by majority voting, is an aggregation across the trees (Tan et al., 2005, 2013). Random forests have several advantages, such as the ability to handle categorical and continuous independent variables, the computational efficiency, and it avoids overfitting (Breiman, 2001).

The K-nearest neighbours (kNN) is an instance-based learning classifier based on learning by analogy (Han et al., 2011; Witten et al., 2017) and it works by finding the k training instances that are closest to the test instance and classifies the test instance as the predominant class of the k nearest training instances (X. Wu et al., 2008). It requires an appropriate preprocessing of the data, namely the normalisation of the variables since different attributes can be measured on different scales and this can lead to problems in the predictions (Shalev-Shwartz & Ben-David, 2014).

Naïve Bayes is a bayesian classifier that uses probability theory (Bayes theorem) to build classification models based on the past occurrences that can place a new instance into a most probable class (Sharda et al., 2018). The goal is to identify the hypothesis with the maximum posteriori probability. In this approach it is assumed that the attributes are statistically independent from each other (Shalev-Shwartz & Ben-David, 2014; Witten et al., 2017). Despite this unrealistic independence assumption, the Naïve Bayes classifier is remarkably successful in practice (Rish, 2001), and is particularly suited for when the dimensionality of the dataset is high (Islam et al., 2007), although computationally costly.

Chapter 2: The Company Inova+

This chapter contains a description of the Inova+ company, its organisational structure and areas of expertise. In addition, it includes a brief presentation of the CRM system and its respective limitations.

2.1 Description of the Company

Founded in 1997, Inova+ is a Portuguese specialised consulting firm that supports the growth of organisations through innovation, international cooperation, digital transformation, and access to funding. It is an European leader supporting R&D (Research and Development) and innovation projects, including Portuguese and European funded projects (Horizon Europe, for example) across multiple fields and areas of expertise, with a portfolio comprising corporations, SMEs (Small and Medium Enterprises), start-ups, scientific and R&D organisations as well as public authorities. Inova+ is present in four countries, namely Portugal (Porto and Lisbon), Belgium (Brussels), Germany (Heidelberg) and Poland (Warsaw), counting with over 80 employees and an annual turnover exceeding €5M in 2017 (Inova+, 2020).

The company performs and is distributed amongst three core business units: International, Digital, and Consulting. The International unit, operating in numerous fields, supports the development and promotion of innovative initiatives at an international level, collaborating with an extensive network of partners in several countries. According to the specifications of the project and the requirements of the clients, the firm can assume the role of partner, coordinator, or service provider/consultant. Working side by side in various projects with the European Commission and other entities, the sub-divisions of the International unit include international funding (over 15 European funding programmes, such as Horizon 2020 and Horizon Europe), Science and

Technology, capacity building and cooperation (Erasmus+ programme), and support to European Institutions.

The Digital unit is subdivided in R&D insourcing and project management, applied R&D, and digital transformation. Inova+ has its own R&D department, "which is able to design and implement projects in different technical and scientific areas" (Inova+, 2020). They support organizations with the digital transformation process and the adoption of solutions according to Industry 4.0., and they develop customised solutions in different fields such as Health, Smart Buildings and Industrial Automation, and Sustainable Cities.

Lastly, the Consulting unit is one of the biggest of the firm, and in terms of clients it operates in three organisation segments: corporate, scientific (universities and R&D centres), and institutional (public entities, municipalities, etc.). The areas of intervention include public funding, project management, business innovation, support to start-ups and SIFIDE II (tax incentives).

Alongside the business units, the company also holds shared services departments: IT, HR (Human Resources), Financial, Administration, Marketing, Communication and Business Development. Although these shared services have a more internal focus, they also work with the business units' side by side and are in close contact with stakeholders, namely with clients, partners, and suppliers.

2.2 CRM System

Taking into consideration the rapid technological developments and the need to fortify customer relationships, Inova+ implemented in February of 2019 the Microsoft Dynamics 365 CRM system, in order to manage more efficiently their company's interactions with customers, improving the accounts and

organizations management and the business process operational efficiency and management.

The CRM system helps in the optimisation and automation of sales, and it includes a customer database component, enabling the company to keep track of their accounts and contacts, expand sales from leads to opportunities, and create sales collateral such as proposals, contracts, orders, and invoices. The layout and organisation of the CRM system is simple, allowing a straightforward navigation, as it includes a side panel containing the main entities of the system such as Clients and Partners (Organizations and Contacts), Leads (Leads, Opportunities and Competitors), Sales (Proposals, Contracts, Orders and Invoices), Working (Services, Candidatures and Projects), Tools (includes reports) and Marketing (Lists and Campaigns).

It was implemented with four main goals, namely the management of customers/partners and contacts (creating a centralised registration and consultation of information regarding customers, partners and contacts), sales management (with the registration and consultation of commercial proposals and two business process flows), candidatures and projects management (allowing the registration of submitted applications, approved internal projects and of services resulting from customer applications), and finally communication management (fostering mass communication with customers and contacts via email and the consultation of all information related to the business).

The two business process flows present in Inova+ are distinguished by their genesis and are the proposal flow and the candidature flow. The proposal flow starts when a new proposal is created, and it comprises the entities proposal, services portfolio (where the service associated with the proposal is introduced), purchase order (if the proposal is adjudicated) and invoice. The candidature flow is related with one of the main services of Inova+ which are the national and international funding candidatures, and it initiates when a new candidature is

made, including the entities candidature, services portfolio, project (where details of the candidature are introduced), services, purchase order and invoice.

It is important to mention that the Inova+ CRM system has several limitations, mainly due to the lack of commitment and motivation of the employees towards the system, leading to a sub utilisation of the system and making it not reach its full potential. Although it is a firm's goal since its implementation in 2019, the CRM system is not yet integrated with the ERP and accounting systems of the company. Consequently, there is no crossing of data between the systems, namely the information regarding customers identification (for example, the same customer has an id number in CRM different than the id number in the ERP system), and the invoices information, that is lacking in the CRM system. Another significant limitation of the CRM system is that currently only two entities are used, which are the Clients and Partners entity (with information regarding organisations and contacts) and the entity of Sales (where proposals are introduced, leading to contracts if adjudicated). Although these entities are used, they are not updated over time, meaning that the system can have proposals as drafts that do not include the value associated, their status update or the service associated. Additionally, no more than 13 (out of over 80) employees have credentials to enter and operate in the system, resulting in a substantial amount of missing information (specifically involving proposals and contracts associated services and values). These limitations lead to a deficient utilisation of the CRM system, where the data does not fully reflect the reality of the company.

Chapter 3: Methodology

The purpose of this investigation is to enhance analytical CRM of Inova+ by analysing the company's and its customer's data with the aim of gaining customer knowledge and support their customer's management strategies, helping them apply data driven decisions in order to increase customer retention and sales performance. The research methodology applied in this investigation is a quantitative approach centred on a case study.

In this investigation the quantitative approach followed contemplated four different procedures according to the data and objectives in mind: dashboards, time-series forecasting model, clustering and classification techniques. The data analysed in this research was extracted from the CRM and the accounting system of the company. To perform the analysis, software such as Excel, Power BI and R Studio were used. In the following sections, additional details regarding the methodology and data utilised for each of the four constructs are presented.

3.1 Performance Dashboards

The objective of this section of the study was to analyse important indicators regarding accounts and customers, proposals, contracts, and revenues, that are relevant to the commercial department of Inova+, helping in its decision-making process. In order to do so, performance dashboards were constructed using Power BI.

The CRM system of the company includes information regarding accounts (client's and other organisations that had contact with the company), proposals and contracts. These data were extracted from the CRM system in the form of Excel files, and later uploaded into Power BI to create the dashboards.

The data model with the three tables (accounts, proposals, and clients) and the relationships between them can be observed in Figure 7. The attribute "company", common in the three tables, was used to connect them.

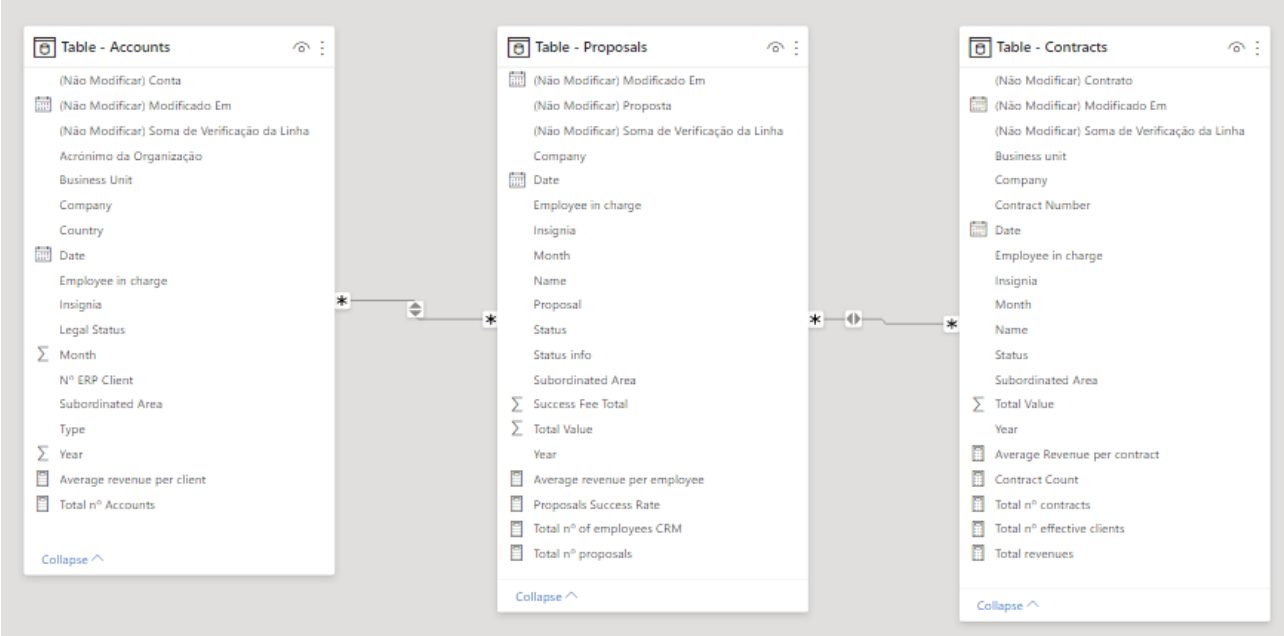

Figure 7: Performance dashboards data model

Three dashboards were created with the information from the tables (retrieved from the CRM system) that comprised the period between 31 of January of 2019 and 31 of December of 2020. So as to achieve the construction and analysis of the dashboards and respective KPIs, important measures were created in Power BI. Examples of these measures are the total number of accounts, proposals, clients, and employees, and the total revenues and total number of returning clients.

3.2 Revenues Forecast

In this part of the investigation the purpose is to apply predictive analytics to the data retrieved from the accounting system of Inova+ by employing a time-series forecasting model to predict the volume of business for 2021. Additionally, it was predicted the year in which the revenues would surpass the 10 million Euros mark, which is a company's goal. As a side note, the words volume of

business, revenues and turnover are used in this section with the same meaning: the income that the company generates through its business and services provided (sales) and other supplementary income sources.

The data regarding the volume of business was extracted to an Excel file from the accounting system of the company. It can be seen on Table 1 and it comprises the period between 2010 and 2020, inclusive.

Year	Annual Turnover
2010	2,356,102.99 €
2011	2,152,249.27 €
2012	2,566,360.90 €
2013	3,127,440.14 €
2014	2,921,146.54 €
2015	3,382,687.86 €
2016	3,367,842.32 €
2017	4,645,606.32 €
2018	5,148,172.77 €
2019	5,319,653.16 €
2020	5,065,232.46 €

Table 1: Annual turnover from 2010 to 2020

This data is a time series and it can be defined as a set of observations for a given variable of interest collected over time (Sharda et al., 2018). There are several methods for time series forecasting (for an overview see De Gooijer & Hyndman (2006) and Gardner (1985, 2006)). In order to choose the right forecasting model for the data, a graphic (Figure 8) was made to check if there was a trend.

Figure 8: Graphic of annual turnover from 2010 to 2020 and trendline

It can be observed that the turnover is steadily increasing over time, indicating a positive (additive) trend. A trendline was constructed and included in the graphic (Figure 8). Therefore, the chosen method to perform the turnover forecast analysis was the Holt's linear method (Holt, 2004).

Holt Model

Exponential smoothing can be described as a time series forecasting method for univariate data, where the only input to the forecasting system is the historical data of revenues, in this case (Winters, 1960). The Holt Model is a time series smoothing model with trend adjustments where two separate equations work together to generate the final forecast: the smoothing equation and the trend equation (Holt, 2004). These equations can be seen in Figure 9.

$$S_t = \alpha A_t + (1-\alpha)(S_{t-1} + T_{t-1})$$
$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$
$$\hat{A}_{t+k} = S_t + kT_t$$

Figure 9: Holt Model smoothing, trend and forecast equations

The smoothing equation includes the observed turnover for period t (A_t) and directly adjusts the last smoothed value (S_{t-1}) for last period's trend (T_{t-1}). The trend is expressed as the difference between the last two smoothed values and the last period's trend. The exponential smoothing approach requires that alpha (coefficient for the level smoothing) and beta (coefficient for the trend smoothing) be between the values of 1 and 0, inclusive (Carlberg, 2012).

The forecasting of revenues using the Holt model was made in Excel. For initialising the smoothing equation, S_{t-1} was calculated as the average of the revenues from 2010 to 2015. To initialise the trend (T_{t-1}), the slope of the trendline was used (340,398) as this value represents the average growth in revenues per year. Initially the alpha and beta were set to 0.5 each. This was followed by the application of the equations to forecast revenues from year 2016 until 2021. With the goal of choosing the ideal alpha and beta values and to assess the accuracy and quality of the model, a minimisation of a measure of forecast error was utilised using solver in Excel. The forecast error measure used was the mean absolute percentage error (MAPE), as it is scale independent (Chatfield, 2000). The MAPE formula can be seen in Figure 10 and it measures the size of the error in percentage terms, where the error is defined as the difference between the actual value and the forecast.

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{e_t}{A_t} \right|}{n}$$

Figure 10: Mean Absolute Percentage Error (MAPE) formula

3.3 Market Segmentation

Customer's segmentation (also known as market segmentation) allows companies to efficiently select and target their customers, achieving a more

personalised marketing (Bailey et al., 2009). It enables organisations to understand their customers and build differentiated strategies, tailored to their characteristics and needs, retrieving customers insights (Tsiptsis & Chorianopoulos, 2009) important for increasing customer's retention (Payne & Frow, 1999).

In this chapter of the thesis, the RFM model was applied to segment the clients' companies according to their historical purchase behaviour, reflected by the recency, frequency, and monetary value of the transactions. The data used for this analysis was extracted from the accounting system of Inova+ and it includes the client's company designation, the value received and the date of the transaction. It includes 549 different organisations and comprises the period between 28 of January of 2010 until 27 of January of 2021. The recency was computed in days and it refers to the number of days between the most recent transaction and the date of reference (27/01/2021). The frequency was modelled as the total number of transactions made in the period of analysis (2010 to 2021), representing a long-term portrait of the company sales historic. The monetary value was computed as the average amount of money spent per transaction, in Euros.

To conduct the customers segmentation using the RFM model, two approaches were followed. The first one was to divide the RFM data into quintiles and apply the RFM scoring methodology followed by Hughes (1994). In the RFM scoring method, the data was organised in a descendent order from the most recent to the oldest transactions. Then, for each observation (customer) a score from 1 to 5 was assigned regarding the three variables of RFM. A recency score of 5 refers to the most recent transactions, and a frequency score of 5 refers to customers with higher purchasing frequencies. The higher the average monetary value spent by the client company in a transaction, the higher the M score (equalling 5), for instance.

Additionally, the K-means clustering algorithm was utilised to conduct customer segmentation on the RFM data. The algorithm used to segment customers was the k-means algorithm due to the speed, efficiency, and facility of application to the client's database under study. The data was normalized using the z-score normalization, as it handles outliers (Han et al., 2011).

3.4 Proposals Adjudication Predictive Models

The purpose of this part of the investigation is to develop predictive models to assess which variables are statistically significant and influence the probability of a proposal being adjudicated, along with knowing the probability of a proposal being adjudicated or not according to such attributes. This is important for the Marketing & Business Development department of the company as it allows them to better target their current and prospective customers and respective proposals. For this, data extracted from the CRM system regarding proposals was used, alongside with other additional data that characterises the client's organisations.

The data, extracted from the CRM system into an excel file, regarding the commercial proposals of the Inova+ company, in the period from February 2019 to December 2020, includes information concerning the client's company and the proposing company (Inova+). The clients and proposals data include the enterprise's designation, id number of the proposal, designation of the proposal, service associated, total value of the proposal, total success fee value associated with the proposal (that is only received if the proposal is adjudicated), status of the proposal (adjudicated, lost, or ongoing) and date of the proposal (usually it is the date of the proposal's insertion into the CRM system). Regarding Inova+, the proposals excel file include the variables business unit, subordinated area, and employee. The latter refers to the person who entered the proposal

information in the CRM and may very possibly not have followed the proposal and the customer closely. Therefore, it was excluded from the final database.

3.4.1 Data Preprocessing

During the phase of selecting and preparing the data (data pre-processing step of the KDD process), several alterations were made to the variables to create the final database to support the predictive models.

The implemented transformations to the variables from the excel database extracted from the CRM system included the elimination of the variable date, that lead to the creation of two new variables: year and month. The variable total success fee also was excluded, since only a small percentage of the proposals had a success fee associated. In order to replace it, a binary variable named success fee was created, that is equal to 1 when a proposal has a success fee associated and is equal to 0 otherwise. The variables id proposal number and the designation of the proposal were also eliminated, as they only represent an internal registration for the company. From the variable status of the proposal, only two options were selected for the final database. The selected proposals were the adjudicated and the lost ones, as the ongoing proposals don't have a closed status and are still being negotiated with clients. Regarding proposals in draft, their status was analysed through the cross check of information in the accounting system. If there was an invoice regarding that client, the proposal was labelled as adjudicated, and if not, it was considered lost (not adjudicated).

Additionally, with the goal of increasing the information concerning clients and their characteristics in the database, four variables were included in the final database: type, country, economic sector, and covid-19. The variable type indicates the nature of the client's company and it comprises 5 different levels.

The variable country refers to the country of the client's company, and it is a binary variable, being equal to "Portugal" when the proposal is for a Portuguese company and "Other" when the proposal is for a foreign company. This information about the type and country of the companies was present in the organisations section of the CRM system and it was extracted from there. The binary variable covid-19 was added to the final database as it represents a pandemic period that can affect the adjudication of proposals of Inova+. It assumes the value of 1 for proposals from March until December 2020, and it is 0 before that period.

Regarding the variable economic sector, it represents the sector in which the client companies operate and the data was consulted and extracted from Orbis, a Bureau Van Dijk online database for worldwide companies (Bureau van Dijk, 2021). The classification used for the economic sector variable was the NACE Rev. 2, which is the European standard classification of economic activities (Eurostat, 2008). It is important to mention that this classification is part of an integrated system of statistical classifications regarding economic activities (the ISIC – International Standard Industrial Classification - at a world level) and therefore can be used for European and non-European companies (Eurostat, 2008; United Nations, 2008). The NACE Rev.2 framework is the EU level of the CAE Rev. 3 framework in Portugal, and therefore are equivalent. The information regarding the main section classification of NACE Rev. 2 was introduced in the database for each company. Some sectors were grouped (and hence don't have the exact same nomenclature as the NACE Rev. 2 main section) because of the low number of observations present in the database. It was the case of the sectors Public Administration, Education, Health and Social Services, Service Activities and Other Activities.

In order to finalise the stage of the preparation of the data, the variable company was eliminated from the dataset as it was the designation of each company, thus not relevant for the construction of the predictive models, and it was confirmed that there were no missing values in the final database. Table 2 lists the predictor variables and the outcome (target) variable, as well as their respective typology.

Predictive Variables		
Variable	Typology	Categories
Type	Categorical - Nominal	▪ SME ▪ Education and research organization ▪ Public body ▪ Large enterprise ▪ Non profit
Country	Categorical - Binary variable	▪ Portugal ▪ Other
Economic Sector	Categorical - Nominal	▪ Public Administration, Education, Health and Social Services ▪ Professional, scientific, and technical activities ▪ Manufacturing ▪ Service activities (administrative, financial and insurance, tourism, transportation, arts, and other activities) ▪ Information and communication ▪ Other activities (Agriculture, Construction, Energy, Water, Wholesale and retail trade)
Business Unit	Categorical - Nominal	▪ Consulting ▪ International ▪ Digital ▪ Administration
Subordinated Area	Categorical - Nominal	▪ Empresas Norte ▪ Empresas Sul

		▪ Institucional ▪ Científico ▪ Horizon2020 ▪ Transformação Digital ▪ Administration ▪ CBC - Capacity Building and Cooperation ▪ Support to European Institutions ▪ Inovação Social ▪ RD - Regional Development
Service	Categorical – Nominal	▪ National candidature ▪ European candidature ▪ Specialized services ▪ Tax incentives candidature ▪ National project management ▪ European project management ▪ Organizational consultancy ▪ Funding advising ▪ Representation in Brussels ▪ Business strategy consultancy ▪ Customized ICT development ▪ Implementation of standards ▪ Support consultancy for entrepreneurship ▪ Outsourcing of R&D ▪ Other
Proposal Value	Numeric – Continuous – Ratio Scale	
Success Fee	Categorical - Binary variable	▪ No ▪ Yes
Year	Numeric - Discrete	
Month	Categorical - Nominal	▪ January ▪ February ▪ March ▪ April ▪ May

		▪ June ▪ July ▪ August ▪ September ▪ October ▪ November ▪ December
Covid-19	Categorical - Binary variable	▪ No ▪ Yes
Target Variable		
Variable	Typology	Categories
Adjudicated Proposal	Categorical - Binary variable	▪ No ▪ Yes

Table 2: Variables (predictor and target) list and typology

3.4.2 Prediction Algorithms

The data mining classification techniques used in this study were logistic regression, decision trees, random forests, K-nearest neighbors (kNN) and Naïve Bayes. These techniques were chosen due to their robustness and ease of application. For all the classification models, the 70:30 ratio for splitting the data was used, that is, 70% of the data was used for the training set and 30% for the test set.

When variables are measured at different scales, it can affect the data analysis. This is particularly important for distance-based classification methods such as the kNN. To help avoid this problem, the data was normalized in the kNN model. The normalization applied was the z-score normalization, as it handles outliers (better for skewness data). In z-score

normalization, the values for an attribute are normalized based on its mean and standard deviation (Han et al., 2011).

Regarding the variables in analysis during the model's development stage, the first model to be constructed considered all the explanatory variables, i.e., the variables type, country, economic sector, business unit, subordinated area, service, proposal value, success fee, year, month and covid-19. Logistic regression requires there to be little or no multicollinearity among the independent variables, signifying that the independent variables should not be too highly correlated with each other (Afifi et al., 2020). Therefore, a second model was carried out considering the problems of multicollinearity and disregarding a variable (see section 4.4.1 - Explanatory Variables).

After the application of the full model, it was necessary to proceed with a feature selection process to include in the classification models only the most significant predictor variables, hence creating more simplest models and increasing model performance (Tan et al., 2005). Two approaches were followed, the filter and the wrapper method. In the filter method, an analysis of the correlation between the dependent (adjudicated proposal) and the dependent variables was performed, using the Pearson's Chi-squared test of independence, and also an analysis of how much information gain is given by each independent variable about the dependent variable (adjudicated proposal).

In the wrapper method, a combination of stepwise forward selection and backward elimination methods was applied to the logistic regression full model through performing a stepwise model selection by Akaike's information criterion (AIC), in order to find an optimal and parsimonious model (that does not include unnecessary variables) (Afifi et al., 2020). In this combined stepwise process, at each step the procedure selects the best

attribute and removes the worst from among the remaining attributes (Han et al., 2011).

After the model development and training stage, it was necessary to test and measure the performance of the models to compare and assess them. The measures adopted were based on the confusion matrix (Table 3) and are: accuracy, sensitivity (or recall), specificity and precision. The formulas for these performance measures can be seen in Table 4. The accuracy is the ratio of correctly classified instances (true negatives and true positives) divided by the total number of instances (Sharda et al., 2018). The sensitivity (also named recall or true positive rate) is the proportion of actual positive cases (TP + FN) that are correctly identified as such by the models. The specificity (or true negative rate) is the ratio of correctly classified negatives divided by the actual negative cases (TN + FP). Precision is the proportion of instances classified as positive (TP + FP) that were correctly identified.

Confusion Matrix		**Observed (Adjudicated Proposal)**	
		No	**Yes**
Predicted (Adjudicated Proposal)	**No**	TN	FN
	Yes	FP	TP

Table 3: Confusion matrix

Performance Measure	**Formula**
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Sensitivity (or Recall or True Positive Rate)	$\frac{TP}{TP + FN}$
Specificity (or True Negative Rate)	$\frac{TN}{TN + FP}$
Precision	$\frac{TP}{TP + FP}$

Table 4: Performance measures formulas

Additionally, in the logistic regression method four more performance measures were used to compare the models: the AIC, the BIC, the analysis of variance (ANOVA) and the AUC (area under the ROC curve). The AIC and the BIC (Bayesian Information Criterion) were used to assess the models fit to the data, and the AUC of the ROC Curve to measure how much the models are capable of distinguish between classes, where the higher the AUC (close to 1) the better. The AIC and the BIC criterions are very similar, but the AIC tries to find the model that predicts better while BIC tries to identify the right model (the one that has exactly the important variables). The smaller the AIC and the BIC values, the better the model fit. The ANOVA method was used to compute an analysis of variance (or deviance) and help compare the 4 models through the likelihood ratio test measuring the goodness of fit to the data.

Chapter 4: Results and Discussion

This chapter presents an analysis and discussion of the main results of the four focal aspects constructed and approached in this research: the performance dashboards to analyse the CRM and company's performance, the revenues time-series forecast model to predict the turnover for 2021 and the year in which the firm will surpass the 10 million Euros mark, the application of the RFM model to segment customers and finally the predictive models created to assess the probability of adjudication of a commercial proposal and influencing factors.

4.1 Performance Dashboards and KPIs

In the following sections of this subchapter the accounts and proposals dashboard, the clients and contracts dashboard and the sales and company overview dashboard are presented and discussed. The three dashboards include a date and year filter to enable the option of choosing the relevant period for analysis. The date goes from 31/01/2019 to 31/12/2020 for the accounts and proposals dashboard, and from 01/02/2019 to 31/12/2020 for the remaining dashboards, as the information regarding contracts and sales only started to be introduced in the CRM system in February of 2019.

4.1.1 Accounts & Proposals Dashboard

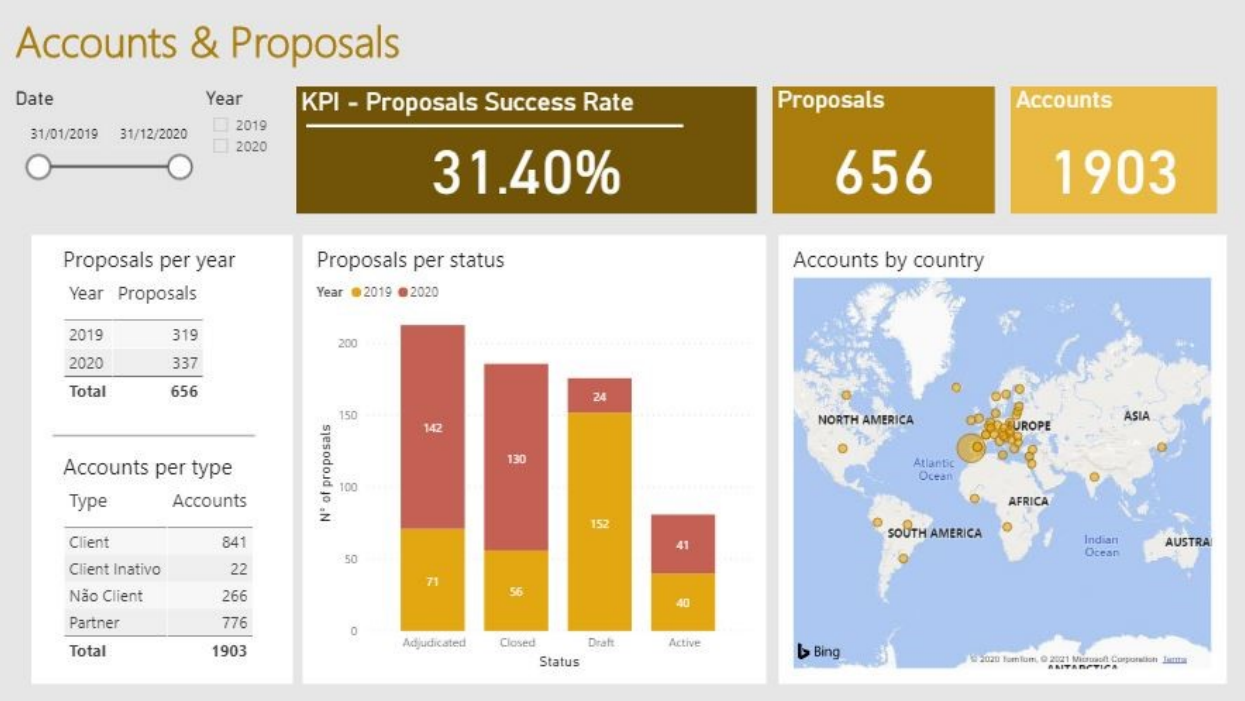

Figure 11: Accounts & Proposals Dashboard

In Figure 11 the dashboard regarding the accounts (organisations) and proposals of the company is displayed. The company has over 1900 accounts in its CRM system, and these accounts represent organisations from all over the world, with a special focus in Europe, the majority of them being from Portugal (826 accounts). These accounts can be clients, partners, or other organisations. Between 2019 and 2020, Inova+ made a total of 656 proposals. The status of the proposals can be adjudicated, closed (non-adjudicated proposals), draft and active (proposals waiting for a decision from the client). It cannot be concluded that more proposals were adjudicated in 2020 than in 2019 because more than 150 proposals from 2019 are with the status "draft" in the CRM system. This confirms the limitations of the CRM of the company, as much information is lacking and not updated in the system such as the proposals status.

The vital KPI in analysis here is the proposals success rate, that represents the percentage of commercial proposals that were adjudicated. It was

calculated by dividing the total number of contracts (equivalent to adjudicated proposals) by the total number of proposals. In the current state of the art of the CRM system, the proposals success rate in the period 2019-2020 is 31.40%. If we look at the year 2020 alone, where we have less proposals in the status of draft, this KPI takes the value of 42.14%. The proposals success rate also characterises the success rate of the commercial team, and it is an important metric to monitor.

4.1.2 Clients & Contracts Dashboard

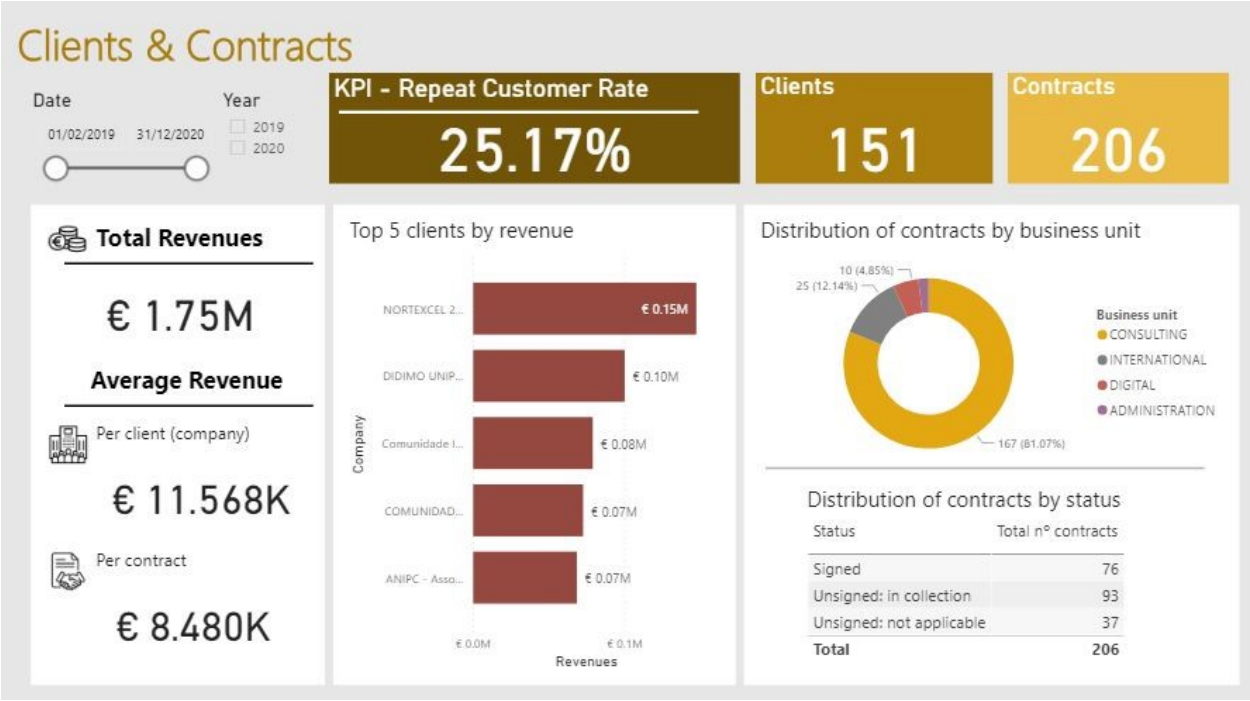

Figure 12: Clients & Contracts Dashboard

The dashboard regarding clients and contracts is present in Figure 12. The total number of clients of Inova+ between 2019 and 2020 is 151, with 206 contracts attributed. The monitoring of the total number of clients and projects in course is important to track because it allows the company to plan their teams size and schedule. Between 2019 and 2020, the company had total revenues above 1.75 million Euros and an average revenue per client and per

contract of 11,568 Euros and 8,480 Euros, respectively. In the bar chart the top 5 clients that most contribute to the revenues and their corresponding contribution are represented. The business unit of consulting represented more than 81% of the total contracts, which makes sense since it is the biggest area of the company.

The repeat customer rate (also known as repeat business rate) is an important KPI that tells us the percentage of clients that are returning clients, i.e., clients organisations that counted with the help of Inova+ for developing their consulting projects more than one time. The repeat customer rate was calculated through the creation of a Power BI measure that divided the total number of returning clients by the total number of clients, and it was 25.17%. This means that circa 25% of the client's organisations worked with Inova+ in more than one project.

4.1.3 Sales & Company Overview Dashboard

Figure 13: Sales and company overview dashboard

In Figure 13 is displayed the sales and company overview dashboard. This dashboard gives an overview of the firm revenues over time. The total revenues reached 1.75 million Euros, 1.21 million Euros of which from 2020. The "Total revenues by year" graphic permits to drill-down and see the evolution of revenues by months (where the months with higher revenues were October and November) and successively by days. Once more it is important to mention that these conclusions are compromised by the limitations of the CRM system, because as we could see in the accounts and proposals dashboard, most of 2019 proposals were in the status "draft", which means they did not have a service or value associated, and we cannot know if they were adjudicated or not.

The distribution of revenues by the business units and corresponding subordinated areas can be seen in the pie chart and in the horizontal columns chart. The consulting unit represents almost 62% of the revenues (it represented 81% of the contracts) and the institutional and "Empresas Norte" are the subordinated areas (included in the consulting business unit) with the higher revenues.

The company has over 80 employees but only 13 of them work and utilise the CRM system. The average revenue per CRM employee is 134.370 Euros, but the real average revenue per employee (considering 80 employees and a total revenue of 1,746,83.36€, which is the exact value for total revenues) is 21,835.45€, which is a much more reasonable number. Additionally, a treemap chart was constructed to visualise the distribution of revenues per CRM employee. The two employees with the highest amount of revenues are part of the commercial department of Inova+. Note that the employee attribute does not mean that they accompanied the client's company project up close, it only represents the employee who introduced the proposal in the system, so it makes sense that the revenues associated with these two employees are

higher because they are the people responsible for introducing the proposals and subsequent contracts information in the CRM system.

4.2 Revenues Forecast

The forecasting of revenues is important for the company as it allows it to better establish business quotas, to plan their capital expenditures and budgeting, and to manage their human resources more efficiently, thus leading to informed data-driven business decisions. In result of the application of the holt model, the revenues forecast for 2021 is 5,997,430.05 Euros, as observed in Figure 14. In order to have an indication of the confidence level of the forecast, a 95% confidence level interval was constructed (Newbold et al., 2013), where the lower value is 5,808,806.67 Euros and the upper value is 6,186,053.43 Euros. It can be concluded, with 95% confidence, that the 2021 revenues forecast is 5,997,430.05 Euros with a margin of error of 188,623.38 Euros.

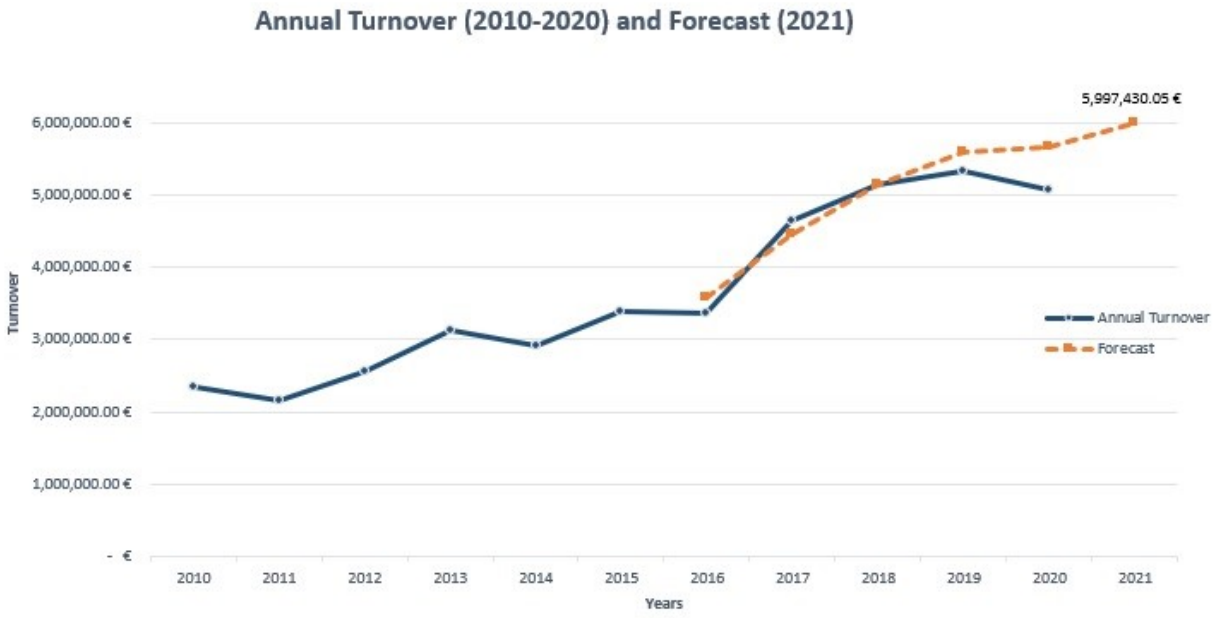

Figure 14: Annual turnover (2010-2020) and turnover forecast for 2021

The optimal values (that minimise MAPE) for alpha and beta are 0.507 and 0, respectively. An alpha value of 0.507 means that the series is smoothed, and a beta of 0 signifies that the trend is constant over time. MAPE was used to measure forecast accuracy, and it is approximately 5.3%, which indicates a highly accurate forecasting (Lewis, 1982; Moreno et al., 2013).

Additionally, a turnover forecast using holt method (and the optimal values of alpha and beta) for the period 2022-2035 was made to discover the year in which the turnover would be above 10 million Euros. The answer is 2033, where the revenues forecast is 10,082,206.05 Euros.

4.3 Market Segmentation

On Table 5 it can be observed a summary of the three RFM variables: recency, frequency, and monetary.

Variables		
Recency (in days)	Minimum - Maximum	0 – 4,017
	Median	651
	Mean (Standard deviation)	1,077 (1,087)
Frequency	Minimum - Maximum	1 – 15
	Median	2
	Mean (Standard deviation)	2.48 (2.11)
Monetary (in Euros)	Minimum - Maximum	209.60 – 116,348.76
	Median	4,750
	Mean (standard deviation)	8,631 (12,265.70)

Table 5: Summary measures of RFM variables (N=549)

The following sections present the two approaches followed for the segmentation of Inova+ customers, RFM scoring method and k-means clustering, and a discussion of the main results.

4.3.1 RFM Scoring Method

Table 6 represents the RFM scores of the dataset of the 549 companies. For example, a recency score (R) of 5 refers to the most recent transaction occurring between 0 and 803 days before the 27 of January of 2021. For the variable frequency (F), a score of 5 means that the company had between 13 to 15 transactions with Inova+ over the time period in analysis (11 years). A monetary score (M) of 5 refers to the transactions with an average monetary value equal or above 93,289 Euros.

RFM Score	Number of clients
511	195
311	70
411	64
521	62
111	38
211	28
531	17
512	15
321	9
421	7
541	7
513	6
412	5
532	5
212	4
522	3
221	2
431	2
551	2
113	1
213	1
223	1
312	1
325	1
331	1
515	1
523	1

Table 6: RFM Scores and respective frequency (N=549)

After the RFM scores assigned, six segments were created and labelled (observed in the treemap of Figure 15) according to its scores. The RFM score of 511 is the most common, which means 195 companies from the total of 549 have made a recent transaction (i.e., in the last 2 years and 2 months), but with a low frequency and monetary value. Therefore, these clients were labelled as new clients, along with the clients with the RFM scores of 512, 513 and 515. They represent almost 40% of total clients, and it is important to reinforce the communication and present them the areas of expertise of Inova+, in order to increase the customer retention.

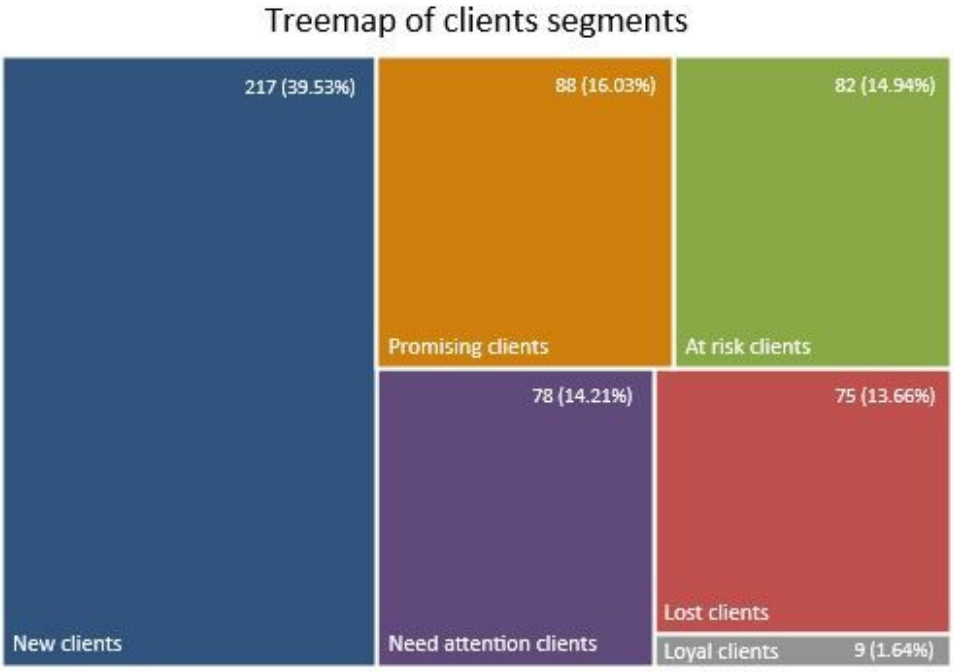

Figure 15: Treemap of the customers segments and respective frequencies

Promising clients represent more than 16% of the total clients and they had the RFM scores of 521, 522, 523, 531 and 532. These are clients that made a recent transaction with Inova+ and had a frequency of past transactions above 4 (in the case of frequency score of 2) and above 7 (frequency score of 3). These clients are returning clients, but it is important to increase the revenues they provide. Suggestions for the marketing team would be to follow-up closely these clients, contact them on a regular basis to create brand awareness and

measure their satisfaction with the projects they assigned to Inova+. At risk clients (RFM scores of 311, 312, 321, 325 and 331) represent almost 15% of total clients and these are companies where the last transaction with Inova+ was more than 4 and a half years ago (in 2017 or before). They are at risk of being churned, and it crucial that Inova+ conquers these companies. The marketing team should send personalised emails to reconnect and present their most recent value proposals and projects. The RFM scores of 411, 412, 421 and 431 represent the "Need attention" clients. These clients were classified as such because it is necessary to reactivate them, since the most recent transaction was more than 2 years ago. It is suggested that Inova+ contacts them with specific recommendations of projects and candidatures based on their past purchases. Lost clients (scores of 111, 113, 211, 212, 213, 221 and 223) represent 13.66% of total clients and loyal clients (541 and 551 scores) represent only 1.64% of total clients. This is a very low percentage, and Inova+ needs to improve this metric. It can also be seen that the average monetary value spent by these clients is very low (score of 1) and it is important to make these frequent clients to spend more. The commercials of Inova+ can suggest in their meetings with the client's projects that could benefit the company and with a higher value associated.

4.3.2 K-means Clustering

The k-means clustering algorithm requires the apriori definition of the number of clusters (k). In order to support the choice of the number of clusters, it was computed an elbow curve (Figure 16) and the davies-bouldin index (Figure 17) for different values of the number of clusters.

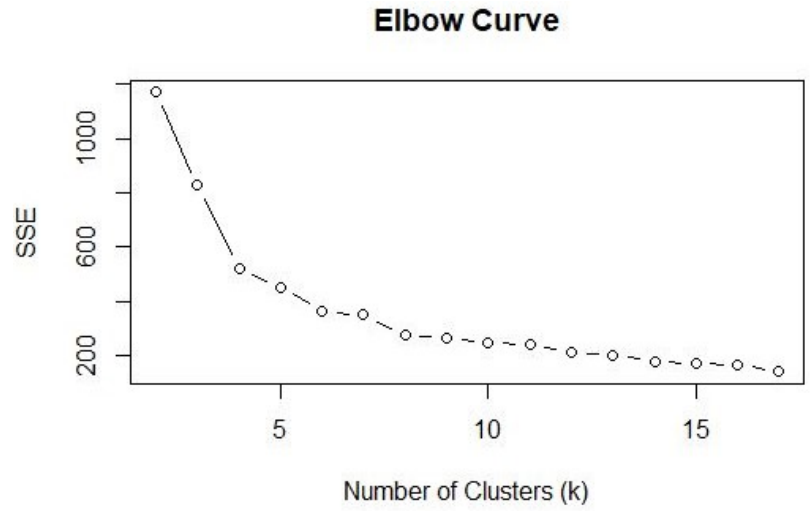

Figure 16: Elbow curve for the k-means algorithm

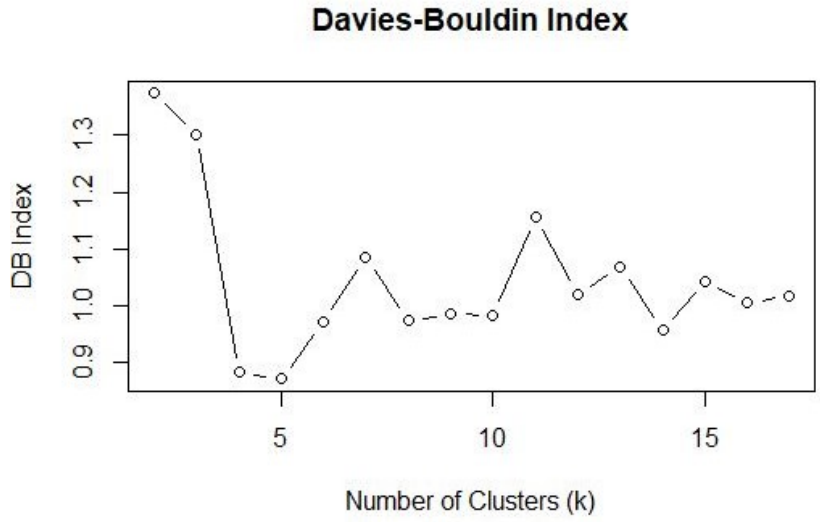

Figure 17: Davies-Bouldin Index Graphic

According to the davies-bouldin index, the most appropriate number of clusters would be four or five, as these correspond to the lowest values of the index. From the elbow curve, it was concluded that five clusters seemed to be the most appropriate option. Therefore, the k-means clustering for customers' segmentation was proceeded with five clusters.

Figure 18 is a 3D graphic representing the clusters and their positioning according to the three variables of the RFM model, and Table 7 represents the five clusters and their respective average recency, frequency and monetary values (centroids), size (number of clients companies) and proportion.

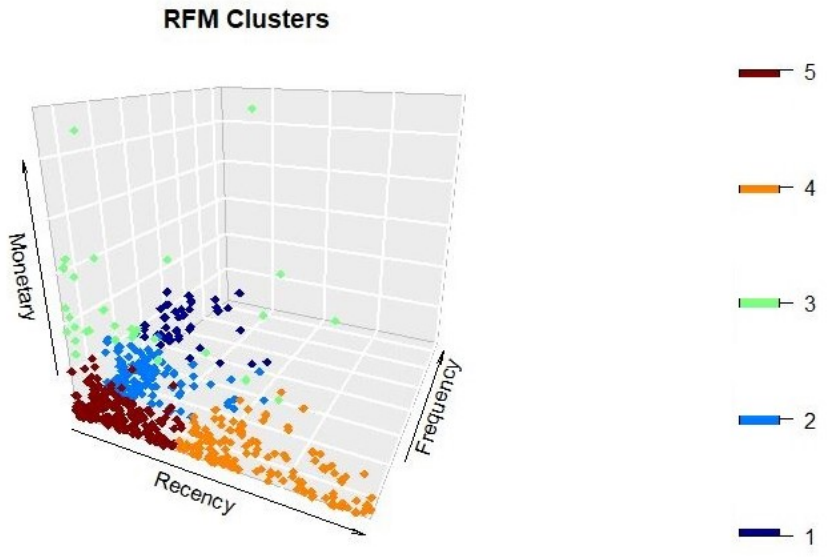

Figure 18: RFM Clusters 3D Graphic

	Recency Centroid	Frequency Centroid	Monetary Centroid	Size (Proportion)
Cluster 1	377.85	8.09	14,700.69	41 (7.47%)
Cluster 2	491.38	3.94	6,965.70	125 (22.77%)
Cluster 3	988.83	1.73	49,098.17	30 (5.46%)
Cluster 4	2,632.66	1.43	5,342.77	141 (25.68%)
Cluster 5	535.79	1.35	4,899.41	212 (38.62%)

Table 7: Cluster's centroids, size and proportion (N=549)

Clusters 5, 4 and 2 have more clients included (representing more than 87% of total clients). The graphics presented in Figure 19 allow to differentiate the clusters considering combinations of variables: recency-frequency, frequency-monetary, and recency-monetary.

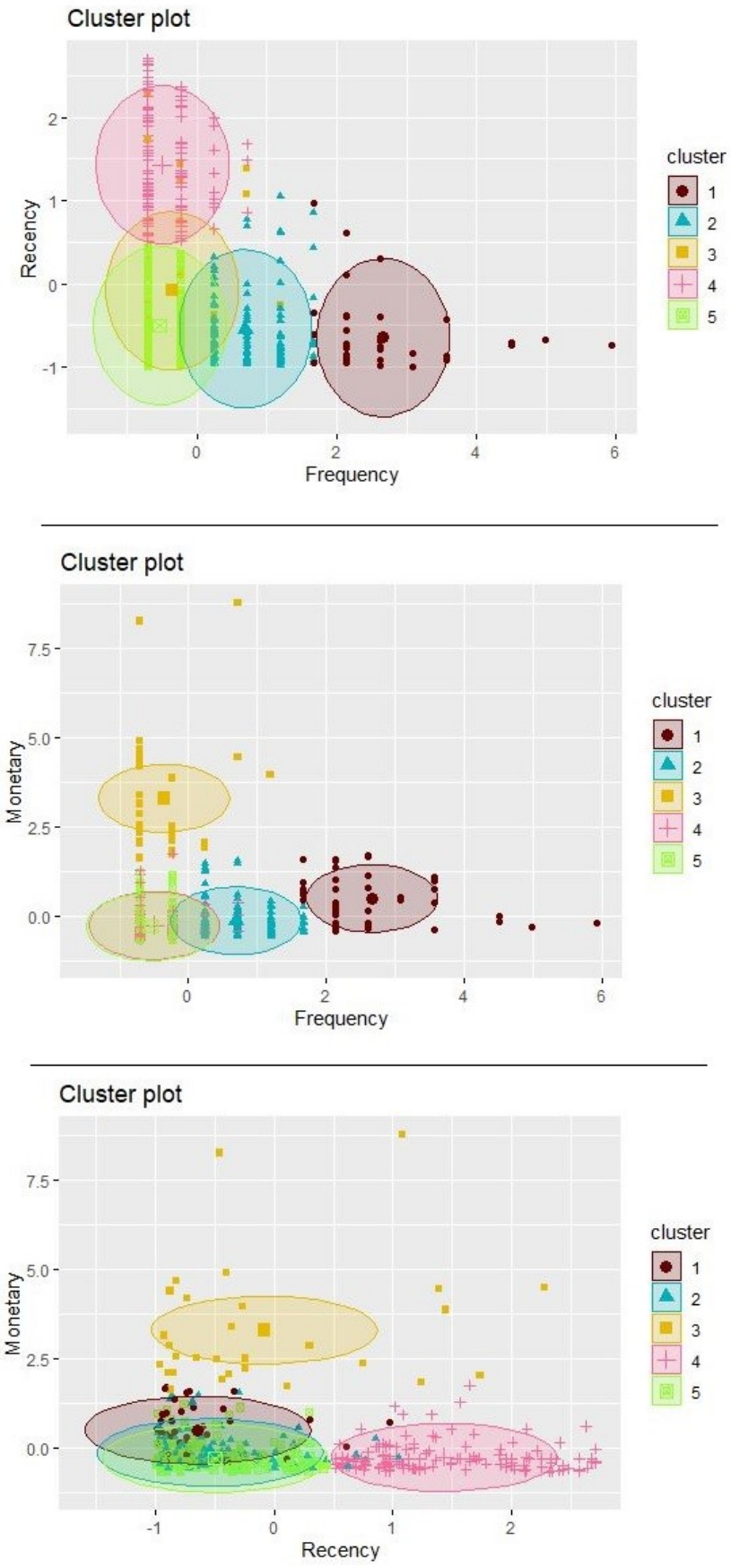

Figure 19: Clusters graphics with recency, frequency and monetary variables

Cluster 3 includes the clients that spend more money in their transactions and projects with Inova+ (this cluster represents the big spenders' segment). Cluster 1 represents the frequent client's segment (clients that interact with Inova+ on a regular basis), with an average frequency of 8 transactions over the period in analysis. Cluster 4 respects to lost client's segment, and has an average recency of 2632 days, representing clients whose last transaction with the company was, on average, more than 7 years ago. As this segment represents almost 26% of total clients, it is important to reconnect with them and reconquer their trust and business providing them the best possible service. Clusters 2 and 5 are very similar in terms of recency and monetary values of the transactions. They only differ more significantly in terms of frequency, where the cluster 2 has an average of almost 4 transactions per client. Cluster 5 represents the segment of new clients and cluster 2 can be classified as the segment of promising clients, that can be retained by Inova+ with the help of personalised marketing actions.

4.3.3 Discussion

The RFM analysis helps the company evaluate their customer's value based on their recency, frequency, and monetary value. Segmenting the customer database, whether through the RFM score approach or with the k-means clustering technique, provides enormous benefits for Inova+, as it helps the marketing department to discover distinct groups of customers and develop targeted marketing programmes and initiatives, boosting sales and revenues.

The main difference in terms of results between the RFM scoring approach and the k-means clustering technique is the number of customer's segments. The first approach, the RFM scoring, differentiates the customers in six segments based on the analyst intuition, and it is visible in Table 8. The

application of the k-means clustering algorithm resulted in five customer's segments (Table 9), labelled according to their RFM values.

Segment	Number of clients	RFM Scores
New clients	217 (39.53%)	511, 512, 513, 515
Promising clients	88 (16.03%)	521, 522, 523, 531, 532
At risk clients	82 (14.94%)	311, 312, 321, 325, 331
Need attention clients	78 (14.21%)	411, 412, 421, 431
Lost clients	75 (13.66%)	111, 113, 211, 212, 213, 221, 223
Loyal clients	9 (1.64%)	541, 551

Table 8: RFM Scoring customer's segments

Cluster and Segment	Number of clients	Average Recency	Average Frequency	Average Monetary
Cluster 1: Frequent clients	41 (7.47%)	377.85	8.09	14,700.69
Cluster 2: Promising clients	125 (22.77%)	491.38	3.94	6,965.70
Cluster 3: Big spenders	30 (5.46%)	988.83	1.73	49,098.17
Cluster 4: Lost clients	141 (25.68%)	2,632.66	1.43	5,342.77
Cluster 5: New clients	212 (38.62%)	535.79	1.35	4,899.41

Table 9: K-means clustering customer's segments

The new clients represent around 39% of the total customers in both analyses. In order to retain them, it is important to reinforce the communication and develop targeted marketing campaigns to appeal to new projects adjudications. In terms of the other segments, the two approaches start to differ more. In the clustering analysis, promising clients, that are returning clients that spend on average 6,965.70 Euros per transaction with Inova+, represent more than 22% of total customers, while in the RFM scoring approach they only represent 16.03% and refer to clients who had 4 or more

transactions in the past. The clustering analysis allows to identify the best customers groups and differentiate between frequent clients and big spenders, while the RFM scoring approach differentiates at risk clients, need attention clients and loyal clients. The segment regarding lost clients is higher in terms of size in the clustering analysis, but this can be due to the fact that it also considers at risk and need attention clients (using the RFM scoring labels).

Calculate RFM scores and divide the customers into segments according to its respective score can be challenging, as it depends on the interpretation of the analyst. In the k-means algorithm, the customers are segmented into different groups according to their similarities regarding the RFM values. One of the disadvantages of the k-means clustering is that it needs to define the number of clusters apriori, and this can be problematic to implement for people not familiarised with analytics and data mining techniques and software. However, although it requires analytic expertise, it is an efficient algorithm.

4.4 Proposals Adjudication Predictive Models

4.4.1 Exploratory Data Analysis

The database for the construction of the predictive models regarding the adjudication of commercial proposals of Inova+ includes 694 observations and 12 variables, with a time frame from February 2019 to December 2020. The dependent variable (also known as outcome or response variable) is named "Adjudicated Proposal" and concerns the status of the commercial proposals. It is a binary variable, being equal to 1 (Yes) when the proposal is adjudicated and equal to 0 (No) when the proposal is not adjudicated. The 11 explanatory (also known as predictor or independent) variables are: type, country (binary variable), economic sector, business unit, subordinated area, service, proposal

value, success fee (binary variable), year, month, and covid-19 (binary variable). For the descriptive statistical analysis, the minimum, maximum, median, mean, and standard deviation were presented for the numeric variable "proposal value", and absolute and relative frequencies (n/%) for the numeric variable year and the categorical variables.

Variables	Categories	n	%
Proposing Company			
Business Unit	Consulting	560	80.7%
	International	80	11.5%
	Digital	36	5.2%
	Administration	18	2.6%
Subordinated Area	Empresas Norte	238	34.3%
	Empresas Sul	133	19.2%
	Institucional	112	16.1%
	Científico	65	9.4%
	Horizon2020	47	6.8%
	Transformação Digital	36	5.2%
	Administration	18	2.6%
	CBC - Capacity Building and Cooperation	18	2.6%
	Support to European Institutions	12	1.7%
	Inovação Social	11	1.6%
	RD - Regional Development	4	0.6%
Client Company			
Type	SME	240	34.6%
	Education and research organization	120	17.3%
	Public body	116	16.7%
	Large enterprise	115	16.6%
	Non profit	103	14.8%
Country	Portugal	638	91.9%
	Other	56	8.1%
Economic Sector	Public Administration, Education, Health and Social Services	264	38.0%
	Professional, scientific and technical activities	123	17.7%
	Manufacturing	118	17.0%
	Service activities (administrative, financial and insurance, tourism, transportation, arts and other activities)	69	9.9%
	Information and communication	63	9.1%
	Other activities (Agriculture, Construction, Energy, Water, Wholesale and retail trade)	57	8.2%

Table 10: Descriptive analysis of the companies – proposing and client company (N = 694)

The database includes for each proposal (each observation) specific information regarding the companies (both Inova+, which is the proposing

company, and the client's companies) and the proposals. The categories of the variables regarding the companies are presented in Table 10.

Most of the commercial proposals are part of the consulting business unit of Inova+ (represents 80.7% of the total proposals), with the frequent subordinated areas being "Empresas Norte" (34.3%), "Empresas Sul" (19.2%), and "Institucional" (16.1%) that are included in the consulting unit. Regarding the client's companies, "SME", which stands for "Small and Medium Enterprise" is the most frequent type of company (34.6%) and most of the client companies were from Portugal (91.9%). The proposals to foreign countries include the countries Germany (with 12 proposals total), Spain (9 proposals total), Lithuania (7 proposals total), Ireland (6 proposals total) and others such as France, the Netherlands or Brazil that count with 3 or less proposals. The client's companies are mainly from the economic sectors "Public Administration, Education, Health and Social Services" (38.0%), "Professional, scientific and technical activities" (17.7%), and "Manufacturing" (17.0%).

The variables regarding the commercial proposals are shown in Table 11.

Variable	Categories	n	%
Service	National candidature	221	31.8%
	European candidature	102	14.7%
	Specialized services	97	14.0%
	Tax incentives candidature	56	8.1%
	National project management	43	6.2%
	European project management	36	5.2%
	Organizational consultancy	36	5.2%
	Funding advising	28	4.0%
	Representation in Brussels	16	2.3%
	Business strategy consultancy	12	1.7%
	Customized ICT development	11	1.6%
	Implementation of standards	8	1.2%
	Support consultancy for entrepreneurship	8	1.2%
	Outsourcing of R&D	2	0.3%
	Other	18	2.6%
Year	2019	339	48.8%
	2020	355	51.2%
Month	January	33	4.8%
	February	54	7.8%

Variable	Categories	n	%
	March	81	11.7%
	April	67	9.7%
	May	56	8.1%
	June	54	7.8%
	July	49	7.1%
	August	44	6.3%
	September	56	8.1%
	October	91	13.1%
	November	62	8.9%
	December	47	6.8%
Covid-19	No	401	57.8%
	Yes	293	42.2%
Success fee	No	584	84.1%
	Yes	110	15.9%
Proposal Value (Euros)	Minimum - Maximum	0 - 1000000	
	Median	5000.0	
	Mean (Standard deviation)	12255.67 (41978.01)	
Adjudicated proposal	No	327	47.1%
	Yes	367	52.9%

Table 11: Descriptive analysis of the proposal's variables (N=694)

Table 11 shows the descriptive analysis of the proposals. National candidatures (31.8%), European candidatures (14.7%), and Specialized services (14.0%) are the most common services associated with the commercial proposals of Inova+. The data is well balanced in terms of the variable year, with about half of the proposals presented in 2019 (48.8%) and half in 2020 (51.2%) - 42.2% in the "Covid-19 period" (after March 2020). Only 15.9% of the proposals include a success fee. The success fee can represent a monetary amount due to Inova+ when the proposal is adjudicated (this represents 7.2% of the total proposals, and ranges from 100 Euros to 105,000 Euros), a success fee percentage (which happens in 8.36% of the proposals, ranging from 1% to 10%, being 5% the most common) or both (there are 2 proposals in the database that include both: success fee amount and success fee percentage).

The proposal value numeric variable is represented in Euros and it ranged between 0 and 1,000,000 Euros, with mean of 12,255.67 Euros, standard deviation of 41,978.01 Euros and median of 5,000 Euros. A descriptive statistics analysis was performed to see the distribution of the variable. The mean of the

variable is higher than the median, indicating that the proposal value variable has a positive symmetry, as observed in Figure 20 with the density plot.

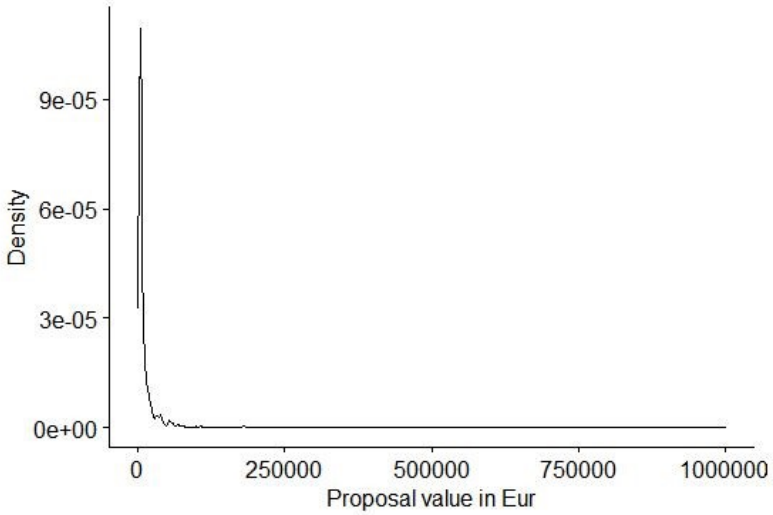

Figure 20: Density plot for the proposal value variable

Regarding the proposal value variable, outliers were identified through a box plot, present in Figure 21, which are observations that seem to be inconsistent with the remainder of the data (Afifi et al., 2012, 2020; Barnett & Lewis, 1978). The observations that are too distant from the rest of the data (displayed as isolated points in the box plot graphic) can be considered outliers.

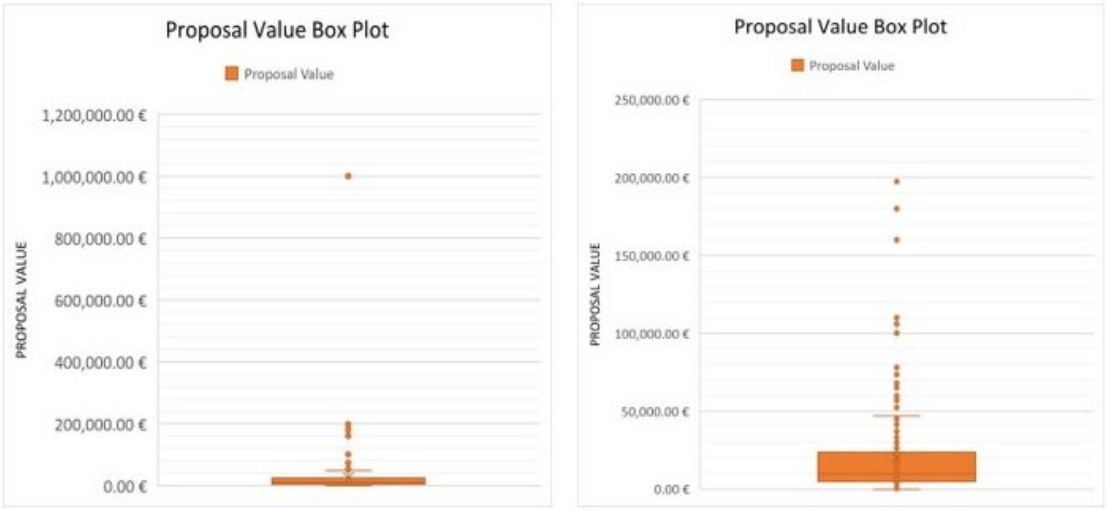

Figure 21: Box Plots of the proposal value variable

These observations were not removed from the dataset because they represent actual values, as they refer to the values associated with the commercial proposals made to clients and prospect clients of Inova+, and do not represent bias or an incorrect data insertion.

It is relevant to mention that only 15 proposals have a proposal value of 0 Euros, and all of them have a high success fee associated, so these are proposals regarding incentives candidatures, where the monetary benefit for Inova+ will be with the success fee associated. Additionally, the outcome variable adjudicated proposal is well balanced. From the universe of 694 commercial proposals, 52.9% were adjudicated.

Explanatory Variables

The predictive models constructed considered different sets of explanatory variables (attributes). Regarding the problem of multicollinearity in the full model, the correlation between the business unit and subordinated area variables was confirmed through a contingency table with the frequencies of the variables (see Figure 22) and through a Pearson's Chi-squared test of independence, where the p-value was essentially 0 (lower than 5% confidence), meaning that the null hypothesis of independence between the variable business unit and subordinated area was rejected.

```
> table(data$Subordinated_Area,data$Business_Unit)

                                  Consulting Administration Digital International
Empresas Norte                           238              0       0             0
Administration                             0             18       0             0
CBC - Capacity Building and Cooperation    0              0       0            18
Científico                                64              0       0             1
Empresas Sul                             133              0       0             0
Horizon2020                                1              0       0            46
Inovação Social                           11              0       0             0
Institucional                            112              0       0             0
RD - Regional Development                  1              0       0             3
Support to European Institutions           0              0       0            12
Transformação Digital                      0              0      36             0
```

Figure 22: Contingency table for variables subordinated area and business unit

The high correlation between the variables business unit and subordinated area lead to problems in the estimation of the coefficients in the logistic regression. Additionally, the performance of the classification models was compared for the full model with and without the business unit variable. Since the performance (accuracy rate) was very similar, as well as equal in some cases, and because the subordinated area comprises almost the same information regarding Inova+ proposals as the variable business unit, the full model applied disregarded the variable business unit.

Regarding the feature selection process, the results of the Pearson's chi-squared test of independence showed that the variables with a significant relationship (where the p-value was lower than 0.05, leading to the rejection of the null hypothesis of independence) to include in the classification models were the variables country, economic sector, business unit, subordinated area, service, proposal value, month and covid-19. The variables with the higher information gain were the variables country, economic sector, subordinated area, service and month. Because of the correlation between the business unit and subordinated area variables, only the latter one was used. Although the proposal value variable has a significant impact on the adjudicated proposal variable, it was not included in the feature selection model as it comprises many levels that include both classes (adjudicated and non-adjudicated proposals). In Figure 23 we can see that proposals with an associated value above 125.000 Euros are not adjudicated.

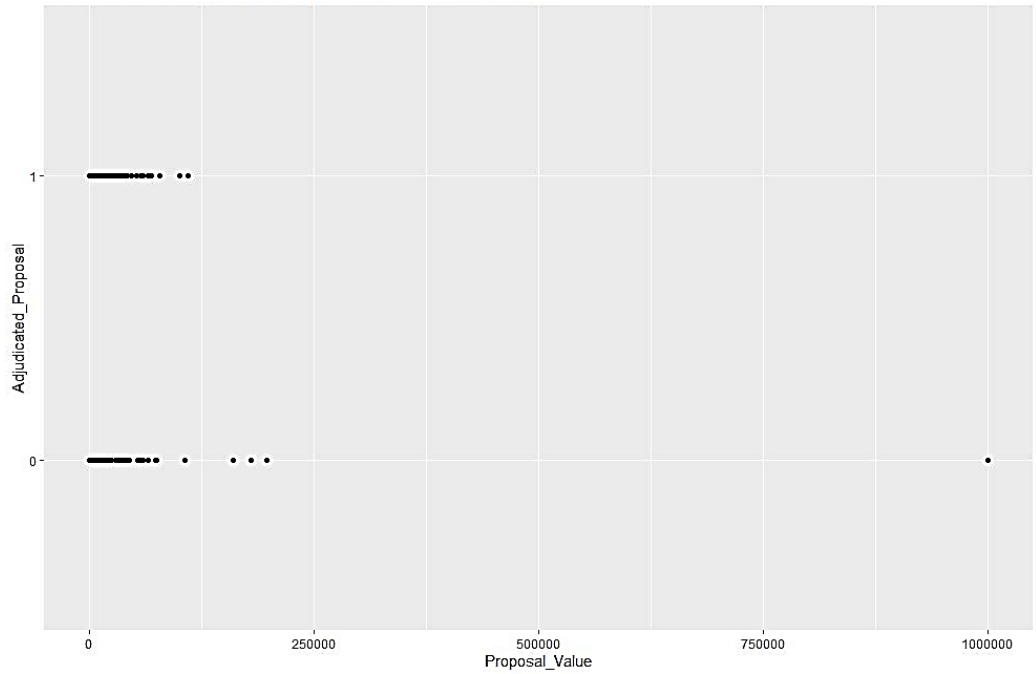

Figure 23: Variable proposal value distribution by proposal adjudication status

Although the variable month is statistically significant to the predictor variable and provides information gain to the models, it was not included in the feature selection model because it is highly correlated with the covid-19 variable and because of the limitation it carries due to the CRM system being implemented in February 2019, meaning that the only data regarding January is exclusive of the year 2020.

Country, proposal value, year and covid-19 were the variables selected from the stepwise selection procedure.

Moreover, a 4th model was constructed and analysed, that includes the attributes type, country, economic sector, service and covid-19. An overview of the explanatory variables included in each model can be seen on Table 12.

Models	Explanatory Variables
Full Model	Type, Country, Economic Sector, Subordinated Area, Service, Proposal Value, Success Fee, Year, Month, Covid-19
Feature Selection (FS) Model	Country, Economic Sector, Subordinated Area, Service, Covid-19

Models	Explanatory Variables
Stepwise Model	Country, Proposal Value, Year, Covid-19
Model 4	Type, Country, Economic Sector, Service, Covid-19

Table 12: Classification models explanatory variables

4.4.2 Logistic Regression

In the logistic regression models, the reference category of the explanatory variables is "SME" for the variable type, "Public Administration, Education, Health and Social Services" for the economic sector attribute, "Empresas Norte" for the subordinated area variable, "National candidature" as the service class of reference, and month "February" for the variable month. For the binary variables' success fee and covid-19 the reference category was "No". The outputs with the coefficients, odds ratios, and p-values of the variables for the four logistic models developed can be seen in the appendix.

The full logistic regression model for the adjudicated proposal allows to conclude that the odds of having a proposal adjudicated is 9 times higher for Portuguese companies when comparing to companies from other countries (OR = 9.382, p-value < 0.001). When comparing to "Public Administration, Education, Health and Social Services" companies (reference category), proposals from the economic sectors of "Manufacturing" (OR = 3.587, p-value = 0.018) and "Professional, scientific and technical activities" (OR = 2.672, p-value = 0.030) have significantly higher odds of being adjudicated. Regarding the service associated to the proposals, "other services" (OR = 0.185, p-value = 0.024) have significantly lower odds than national candidatures (reference category) proposals. Proposals submitted in the covid period (after March 2020) have significantly higher odds of being adjudicated than proposals submitted before March 2020 (OR = 7.953, p-value = 0.013). The type of company, the subordinated area, the success fee, the year, the month and the

proposal value do not affect significantly the proposal adjudication (p-value > 0.05).

Results of the feature selection logistic regression model revealed 2 variables with a statistically significant (p-value < 0.05) effect: country and covid-19. The odds of proposal adjudication are 12.1 times higher for Portuguese companies than for other countries (OR = 12.171, p-value < 0.001). Proposals submitted in the covid period have significantly higher odds of being adjudicated than proposals submitted before March 2020 (OR = 2.087, p-value = 0.001). Alongside that, there are 3 variables with a marginal significant effect (0.05 < p-value < 0.10) in the odds of proposal adjudication: economic sector, subordinated area, and service.

The stepwise logistic regression model has 3 significant variables. The odds of proposal adjudication are 5.9 times higher for Portuguese companies than for other countries companies (OR = 5.909, p-value < 0.001). For a one-year increase in the variable year, the odds of proposal adjudication are 0.493 times lower (Coefficient = -0.707, OR=0.493, p-value=0.037), i.e., for every one-year increase in year, the odds of proposal adjudication are 50.7% lower, ceteris paribus (the other explanatory variables remaining constant). It indicates that proposals from 2020 (the recent year in analysis) are less likely to be adjudicated comparing with 2019. This may be due to the impact of the first months of the year 2020 when covid-19 affected the world, although proposals submitted in the covid period have significantly higher odds of being adjudicated than proposals submitted before March 2020 (OR = 3.590, p-value < 0.001). These contradictory results may be due to the correlation between the variables year and covid-19, which was confirmed through a Pearson's Chi-squared test of independence, where the p-value was essentially 0 (lower than 5% confidence), meaning that the null hypothesis of independence between

the variable year and covid-19 was rejected. Proposal value is not a significant variable.

In the final logistic regression model (model 4), the results revealed 5 variables with a significant effect (p-value < 0.05) in the odds of proposal adjudication: type, country, economic sector, service and covid-19. Regarding the variable type, when comparing to "SME" companies (reference category), "Education and research organisation" have significantly higher odds of proposal adjudication (OR=2.787, p-value = 0.022). When comparing to "National candidatures" proposals (reference category), "Customized ICT development" proposals (OR = 0.081, p = 0.023) have significantly lower odds of proposal adjudication.

	Accuracy	Sensitivity	Specificity	Precision	AIC	BIC	AUC
Full Model	59.52%	71.17%	**46.46%**	59.85%	660.592	869.696	0.5864046
Feature Selection Model	**61.43%**	79.28%	41.41%	**60.27%**	653.911	787.737	0.6011011
Stepwise Model	54.29%	**81.98%**	23.23%	54.49%	639.382	660.293	0.5272545
Model 4	59.52%	76.58%	40.40%	59.03%	650.046	758.781	0.6099281

Table 13: Performance measures of the logistic regression models

As shown in Table 13, the model with the highest accuracy is the feature selection model. The AIC is the 2nd lowest, which means that the model has a good fit to the data, and the AUC is the 2nd highest (0.6011), which means there is a 60.11% chance that the model will be able to correctly distinguish between adjudicated and non-adjudicated proposals. Therefore, the best predictive logistic regression model is the feature selection model.

4.4.3 Decision Trees

In this study three split criterions were applied in the construction of the decision trees: information gain, gini index, and gain ratio. It was performed a

post-pruning to avoid overfitting. The complexity parameters (cp) according to the models performed are shown in Table 14. Furthermore, in Table 14 is exhibited the performance measures of the decision trees for the four models with the information gain, gini index and gain ratio splitting criterions. There are no major differences in terms of performance between the three criterions, and the model with the variables type, country, economic sector, service and covid-19 has the higher accuracy and precision (confidence regarding the predictions of adjudicated proposals) from all the models.

	Accuracy	Sensitivity	Specificity	Precision
Information Gain				
Full Model (cp=0.02)	54.29%	**83.78%**	21.21%	54.39%
Feature Selection Model (cp=0.02)	54.76%	69.37%	**38.38%**	55.80%
Stepwise Model (cp=0.02)	53.33%	72.97%	31.31%	54.36%
Model 4 (cp=0.02)	58.10%	79.28%	34.34%	57.52%
Gini Index				
Full Model (cp=0.02)	54.29%	**83.78%**	21.21%	54.39%
Feature Selection Model (cp=0.025)	54.76%	69.37%	**38.38%**	55.80%
Stepwise Model (cp=0.02)	52.38%	72.97%	29.29%	53.64%
Model 4 (cp=0.02)	**58.57%**	79.28%	35.35%	**57.90%**
Gain Ratio				
Full Model (cp=0.02)	54.29%	**83.78%**	21.21%	54.39%
Feature Selection Model (cp=0.025)	54.76%	69.37%	**38.38%**	55.80%
Stepwise Model (cp=0.02)	52.38%	72.97%	29.29%	53.64%
Model 4 (cp=0.02)	**58.57%**	79.28%	35.35%	**57.90%**

Table 14: Performance measures of decision trees

In Figure 24 it is presented the gini index and gain ratio decision tree for the model 4 (they are equal), which has the highest accuracy (58.57%). The first split feature is the country, and for commercial proposals of non-Portuguese client's companies the predicted outcome is non-adjudicated proposals, representing 10% of total. The 2nd split feature is the covid-19 variable, and proposals from March 2020 onwards (when the attribute covid-19 is 1) have a higher probability of being adjudicated (representing 37% of total). The other split features are the service, economic sector, and type of organisation, respectively, where proposals for non-profit companies, SME or public body are predicted to be non-adjudicated.

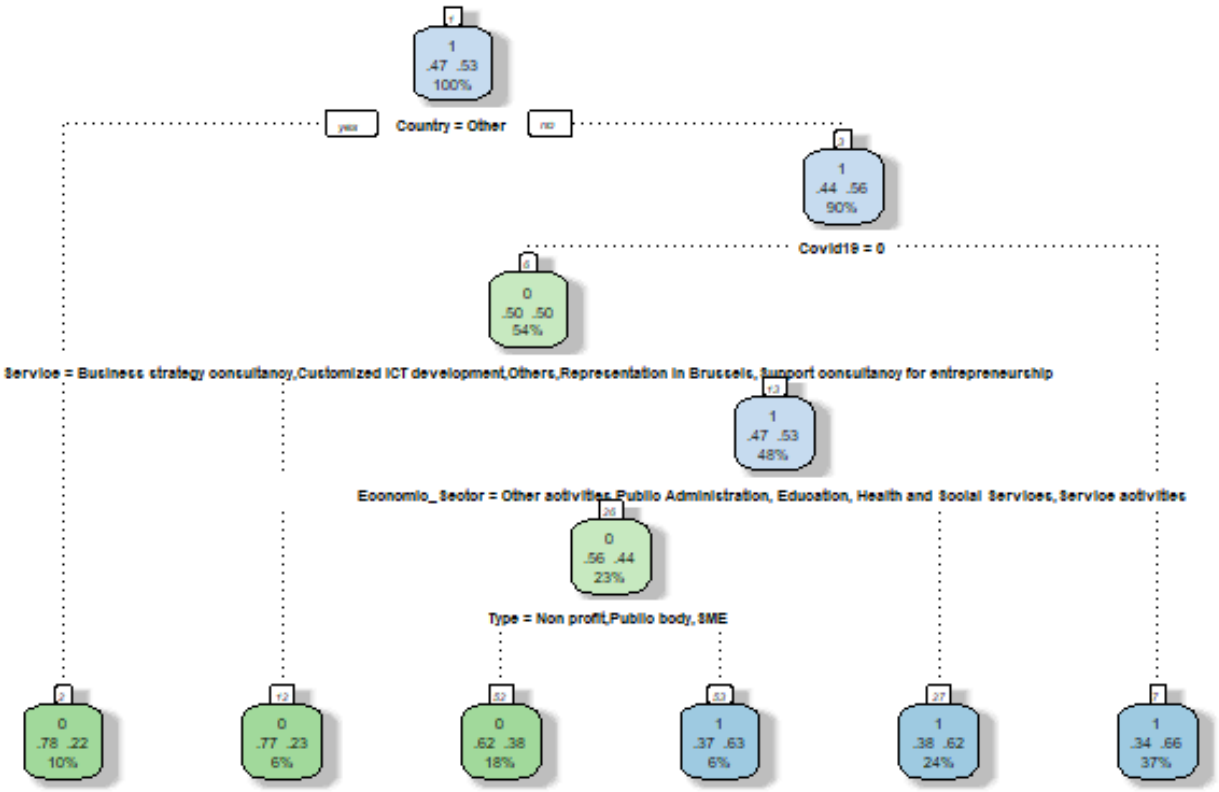

Figure 24: Gini Index and Gain Ratio Decision Tree – Model 4

4.4.4 Random Forests

The two fundamental parameters in the random forest algorithm are the number of trees generated (in the forest) and the number of features to consider in each decision node. In this case, 500 trees were generated for the four models, and the number of variables tried at each split (named "mtry" in Table 15) were 3 and 4 depending on the model, chosen according to the highest accuracy. The performance measures of the random forests' models can be observed on Table 15.

	Accuracy	Sensitivity	Specificity	Precision
Full Model (mtry = 3)	**63.33%**	**79.28%**	45.45%	61.97%
Feature Selection Model (mtry = 4)	58.57%	70.27%	45.45%	59.09%
Stepwise Model (mtry = 4)	62.86%	60.36%	**65.65%**	**66.34%**
Model 4 (mtry = 3)	56.19%	66.67%	44.44%	57.36%

Table 15: Performance measures of random forests (nº of trees = 500)

The highest accuracy achieved in the random forest algorithm is 63.33% with the full model, followed by the stepwise model with 62.86%. The random forest with the highest sensitivity is the full model that considers all the explanatory variables, and the random forest with the highest specificity and precision is the stepwise model that includes the variables country, proposal value, year, and covid-19.

The random forests technique allows a measure of the importance of each variable for the classification to be obtained. In this case, in the full model the most important variables are the country, month, service, and subordinated area. The variable country also stands out from the other variables in terms of importance in the feature selection model. In the stepwise model, the variables that stand out in terms of importance are covid-19, proposal value and country. The year is not so relevant to predict the probability of proposal

adjudication. In model 4, country is the most important variable by far, followed by type of company and covid-19.

4.4.5 K-nearest neighbours (kNN)

The choice of the right k is crucial, as it can lead to misclassifications of the adjudicated proposal variables. In order to tune and optimise this parameter, an accuracy plot was made were the k chosen for each model was the one that maximized the accuracy. In Table 16 is shown the performed models, the number assigned to k in each model and their respective performance measures.

	Accuracy	Sensitivity	Specificity	Precision
Full Model (k=34)	58.10%	79.28%	34.34%	57.52%
Feature Selection Model (k=30)	56.19%	75.68%	34.34%	56.38%
Stepwise Model (k=20)	56.19%	67.57%	**43.43%**	57.25%
Model 4 (k=27)	**60.00%**	**84.68%**	32.32%	**58.39%**

Table 16: Performance measures of kNN models

The kNN model with the highest accuracy (60%), sensitivity, and precision is the model 4, although it has the lowest specificity from all the models in analysis. This model includes the variables type, country, economic sector, service and covid-19, and the k that maximises the accuracy is equal to 27.

4.4.6 Naïve Bayes

In Table 17 we can see the performance measures of the constructed naïve bayes models.

	Accuracy	Sensitivity	Specificity	Precision
Full Model	51.43%	90.99%	7.07%	52.33%
Feature Selection Model	**58.57%**	75.68%	**39.39%**	**58.33%**
Stepwise Model	52.38%	**91.89%**	8.08%	52.85%
Model 4	56.67%	77.48%	33.33%	56.58%

Table 17: Performance measures of Naïve Bayes models

The naïve bayes model with the highest accuracy (58.57%), specificity and precision is the feature selection model, although it has the lowest sensitivity (true positive rate) from all the 4 models in analysis.

4.4.7 Model Selection and Main Conclusions

To study the probability of a commercial proposal being adjudicated (or not) and which factors influence it, predictive models were developed considering different attributes and classification techniques. The results in terms of performance of the best models selected from each classification technique and their respective probabilities of proposal adjudication (named Y in the Table 18) are present in Table 18.

	Accuracy	Sensitivity	Specificity	Precision	P(Y=0)	P(Y=1)
Log. Reg. – Feature Selection Model	61.43%	79.28%	41.41%	60.27%	30.48%	69.52%
DT Gini Index/Gain Ratio – Model 4	58.57%	79.28%	35.35%	57.90%	27.62%	72.38%
Random Forest – Full Model	**63.33%**	79.28%	45.45%	61.97%	32.38%	67.62%
Random Forest – Stepwise Model	62.86%	60.36%	**65.65%**	**66.34%**	51.90%	48.10%
kNN – Model 4	60.00%	**84.68%**	32.32%	58.39%	23.33%	76.67%
Naïve Bayes – Feature Selection Model	58.57%	75.68%	39.39%	58.33%	31.43%	68.57%

Table 18: Predictive models performance measures and probabilities regarding adjudication of proposals

The probability of a proposal being adjudicated varies between 48.10% in the random forest stepwise model and 76.67% with the kNN model. The average probability of a proposal being adjudicated is 67.14%. While this seems a very good number, it is important to also look at the accuracy of the models. The accuracy goes from 58.57% (naïve bayes and decision tree models) to 63.33% in the random forest full model. This signifies that 63.33% of the predictions made by the random forest in the full model were correct. The random forest stepwise model has the second highest accuracy (62.86%), followed by the logistic regression feature selection model (61.43%) and the kNN (model 4) classifier with 60% accuracy.

The models have an average accuracy of 60.79%. Although this may look a bad indicator, in this domain it is not. All the indicators present in the CRM system of the company were utilised in the analysis, as well as the ones added to the database (such as the economic sector or covid-19 attributes). However, there are many influencing factors that can affect proposals adjudication in this B2B domain, and the variables utilised in the database may only represent a small percentage of all the information that could eventually influence the adjudication of a proposal.

In the process of model selection, it is important to look at the other performance measures. In terms of correctly identifying the adjudicated proposals, the kNN model stands out with an 84.68% sensitivity rate. The specificity gives us the proportion of non-adjudicated proposals that are correctly identified, and the random forest (stepwise model) performs better in this aspect (65.65%). Regarding the precision metric, that gives the confidence concerning the prediction of adjudicated proposals, the model that performs better is the random forest stepwise and full model (66.34% and

61.97%, respectively), followed by the logistic regression model with a precision of 60.27%.

Overall, the predictive models with a superior performance are the random forests, having an accuracy of 63.33% and 62.86% (full model and stepwise model, respectively) and a probability of 57.86% for a proposal being adjudicated, on average.

As regards to the variables that influence this probability of proposal adjudication, there are two that stand out as particularly relevant and significant: country and covid-19. Commercial proposals made for Portuguese companies have a higher probability of being adjudicated than for non-Portuguese companies. Additionally, proposals from March 2020 onwards have also a higher probability of adjudication, which makes sense as a majority of these proposals were candidatures to monetary incentives and funding programmes, both national and at an European level. Other relevant variables are the economic sector of the client's company, the subordinated area of the business unit that works in that proposal, the service associated, and the type of organisation. The proposal value, the success fee associated, the year and the month attributes do not significantly affect the proposal adjudication.

Conclusion

In this final chapter, a summary of the research conducted and the main conclusions that were possible to obtain through it are presented, recognising the contributions of this thesis and some managerial implications for the company Inova+. This chapter also identifies some of the limitations of this study and possible suggestions for future research.

The interest of this study arose from a curricular internship carried out at the Marketing and Business Development department of the company Inova+, a management consultant firm specialised in supporting the growth of organisations through innovation, international cooperation, digital transformation, and access to funding. The goal of the company was to support and obtain insights from the analysis of data retrieved from the CRM system, gaining knowledge regarding their business customers and commercial sales processes, in order to increase customer retention and revenues performance.

Throughout this research it was possible to perceive the importance and relevance of the roles of analytics and data mining models in the support of marketing strategies and company's goals. This importance was observed in several phases of this project. The main contributions of this thesis for the literature in this regard are the following:

- The analysis of the state of the art of the CRM system of the company and the measurement of KPIs relevant to the business, such as the proposals success rate and the repeat customer rate;
- The development of a forecast model to predict revenues for 2021 and identify the year in which the company will reach its goal of surpassing 10 Million Euros in annual turnover;

- The identification of market segments based on customer's purchasing behaviour inferred from recency, frequency and monetary value spent on transactions using a clustering technique;
- The development of classification models that addressed the factors influencing proposals adjudication and their respective performance measurement (accuracy, etc.) to identify the variables that have the highest impact in the adjudication of a commercial proposal.

Inova+ has customers from several parts of the world, but the major target is Portuguese companies. They represent the majority of total clients and are a significant factor in determining the adjudication of a proposal, since commercial proposals made for Portuguese companies have a higher probability of being adjudicated than for non-Portuguese companies. The application of the k-means clustering algorithm allowed the definition of segments of customers with similar purchasing behaviours, and thus, facilitated the opportunity for Inova+ to develop differentiated marketing strategies and actions more oriented to each segment to improve its results. Novice customers represent more than 38% of total customers, which indicates that the company can attract and acquire customers easily. However, less than 8% of total customers are frequent (average of more than 8 transactions) and less than 6% are classified as big spenders, which reveals the firm's difficulties in retaining profitable customers. It is necessary to monitor and evaluate the service provided to customers in order to increase their satisfaction and respond more accurately to their needs and expectations.

One of the main limitations of this research consisted in the time period available to carry out this study, since it was part of a curricular internship, the time allotted to it was reduced. Another important limitation was related with the deficient utilisation of the CRM system of the company by the personnel and its short time of existence (since 2019). There were many missing data that constrained the evolution of this research and that would be relevant to study,

namely information regarding the leads generation, marketing campaigns and the national, European and/or funding candidatures success rate, since this is an important indicator that client companies look at when deciding which company to pursue business with. It would also be pertinent to study and evaluate the response of customers to the newsletter and email marketing campaigns and verify if this had an impact in the proposal's adjudication.

The variables included in the predictive models only represent a small percentage of the total factors that influence the probability of adjudication of a commercial proposal. For future research it would be interesting to include and study the impact of additional variables in the proposals adjudication, such as the team size of Inova+ affected to that proposals, the success rate of Inova+ in that specific service and a variable that reflected the competition (for example, if the customer has had proposals from other competitor companies or not). Management consulting firms have typically high personnel turnover rates. Usually, the services provided by Inova+ have several months (and even years) of duration. Therefore, it would be interesting to study if this change in the team during a service would affect the customer satisfaction and the probability of adjudication. A simulation of teams scheduling by project would also be interesting.

Additionally, as suggestions for future work, a classification model (such as decision trees) could be applied to the customer's segments derived from k-means clustering. This would allow Inova+ to allocate its new customers to each segment without the need to develop clustering analysis again. It is essential for the company to increase employee engagement in relation to CRM and improve the monitoring and follow-up of customers, therefore, it would also be relevant as future work to study the impact of CRM training and incentives to embrace it in the company's performance.

Bibliography

Acito, F., & Khatri, V. (2014). Business analytics: Why now and what next? *Business Horizons*, *57*(5), 565–570. https://doi.org/10.1016/j.bushor.2014.06.001

Afifi, A., May, S., & Clark, V. A. (2012). *Practical Multivariate Analysis* (5th ed.). CRC Press - Taylor & Francis Group.

Afifi, A., May, S., Clark, V. A., & Donatello, R. A. (2020). *Practical Multivariate Analysis* (6th ed.). CRC Press - Taylor & Francis Group.

Agustin, C., & Singh, J. (2005). Curvilinear Effects of Consumer Loyalty Determinants in Relational Exchanges. *Journal of Marketing Research*, *42*(1), 96–108. https://doi.org/10.1509/jmkr.42.1.96.56961

American Marketing Association. (2017). *Definitions of Marketing*. https://www.ama.org/the-definition-of-marketing-what-is-marketing/

Anand, N., Gardner, H. K., & Morris, T. (2007). Knowledge-Based Innovation: Emergence and Embedding of New Practice Areas in Management Consulting Firms. *The Academy of Management Journal*, *50*(2), 406–428.

Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: Findings from Sweden. *Journal of Marketing*, *58*(3), 53–66. https://doi.org/10.1177%2F002224299405800304

Arias Aranda, D. (2003). Service operations strategy, flexibility and performance in engineering consulting firms. *International Journal of Operations & Production Management*, *23*(11), 1401–1421. https://doi.org/10.1108/01443570310501907

Bailey, C., Baines, P. R., Wilson, H., & Clark, M. (2009). Segmentation and customer insight in contemporary services marketing practice: why

grouping customers is no longer enough. *Journal of Marketing Management*, *25*(3–4), 227–252. https://doi.org/10.1362/026725709X429737

Barnett, V., & Lewis, T. (1978). *Outliers in Statistical Data* (1st ed.). John Wiley & Sons Ltd.

Becker, J. U., Greve, G., & Albers, S. (2009). The impact of technological and organizational implementation of CRM on customer acquisition, maintenance, and retention. *International Journal of Research in Marketing*, *26*(3), 207–215. https://doi.org/10.1016/j.ijresmar.2009.03.006

Berry, L. L. (1983). Relationship marketing. *Emerging Perspectives on Services Marketing*, *66*(3), 33–47.

Berry, L. L. (1995). Relationship Marketing of Services - Growing Interest, Emerging Perspectives. *Journal of the Academy of Marketing Science*, *23*(4), 236–245. https://doi.org/10.1177/009207039502300402

Berry, L. L. (2002). Relationship Marketing of Services - Perspectives from 1983 and 2000. *Journal of Relationship Marketing*, *1*(1), 59–77. https://doi.org/10.1300/J366v01n01_05

Berry, M., & Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (2nd ed.). Wiley Publishing, Inc.

Bohling, T., Bowman, D., LaValle, S., Mittal, V., Narayandas, D., Ramani, G., & Varadarajan, R. (2006). CRM Implementation: Effectiveness Issues and Insights. *Journal of Service Research*, *9*(2), 184–194. https://doi.org/10.1177/1094670506293573

Borden, N. H. (1964). The concept of the marketing mix. *Journal of Advertising Research*, *4*(2), 2–7.

Bose, R. (2009). Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*, *109*(2), 155–172.

https://doi.org/10.1108/02635570910930073

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Chapman & Hall/CRC.

Bureau van Dijk. (2021). *Orbis Database*. https://orbis.bvdinfo.com/

Calixto, N., & Ferreira, J. (2020). Salespeople Performance Evaluation with Predictive Analytics in B2B. *Applied Sciences*, *10*(11), 4036. https://doi.org/10.3390/app10114036

Cambra-Fierro, J. J., Centeno, E., Olavarria, A., & Vazquez-Carrasco, R. (2017). Success factors in a CRM strategy: technology is not all. *Journal of Strategic Marketing*, *25*(4), 316–333. https://doi.org/10.1080/0965254X.2016.1148760

Carlberg, C. (2012). *Predictive Analytics: Microsoft Excel* (1st ed.). Pearson Education, Inc.

Chamberlin, E. H. (1949). *The theory of monopolistic competition: A re-orientation of the theory of value* (6th ed.). Oxford University Press, London.

Chartered Institute of Marketing (CIM). (2015). *Marketing and the 7Ps: A brief summary of marketing and how it work*. https://www.cim.co.uk/media/4772/7ps.pdf

Chatfield, C. (2000). *Time-Series Forecasting* (1st ed.). Chapman & Hall/CRC.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and

Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188. https://doi.org/10.2307/41703503

Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM) - People, process and technology. *Business Process Management Journal*, *9*(5), 672–688. https://doi.org/10.1108/14637150310496758

Cheng, C.-H., & Chen, Y.-S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3, Part 1), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

Cooper, M. J., Gwin, C. F., & Wakefield, K. L. (2008). Cross-functional interface and disruption in CRM projects: Is marketing from Venus and information systems from Mars? *Journal of Business Research*, *61*(4), 292–299. https://doi.org/10.1016/j.jbusres.2007.07.028

Coviello, N. E., Brodie, R. J., & Munro, H. J. (1997). Understanding contemporary marketing: Development of a classification scheme. *Journal of Marketing Management*, *13*(6), 501–522. https://doi.org/10.1080/0267257X.1997.9964490

Dagger, T. S., & Sweeney, J. C. (2007). Service Quality Attribute Weights: How Do Novice and Longer-Term Customers Construct Service Quality Perceptions? *Journal of Service Research*, *10*(1), 22–42. https://doi.org/10.1177/1094670507303010

Davenport, T. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, *1*(2), 73–80. https://doi.org/10.1080/2573234X.2018.1543535

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The new science of*

winning (1st ed.). Harvard Business School Press.

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, *90*(5), 70–76.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Dawar, N., & Bendle, N. (2018). Marketing in the age of Alexa. *Harvard Business Review*, *96*(3), 80–86.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*(3), 443–473. https://doi.org/10.1016/j.ijforecast.2006.01.001

Donavan, D. T., Brown, T. J., & Mowen, J. C. (2004). Internal Benefits of Service-Worker Customer Orientation: Job Satisfaction, Commitment, and Organizational Citizenship Behaviors. *Journal of Marketing*, *68*(1), 128–146. https://doi.org/10.1509/jmkg.68.1.128.24034

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, *97*(457), 77–87. https://doi.org/10.1198/016214502753479248

Durdyev, S., Ihtiyar, A., Banaitis, A., & Thurnell, D. (2018). The construction client satisfaction model: a PLS-SEM approach. *Journal of Civil Engineering and Management*, *24*(1), 31–42. https://doi.org/10.3846/jcem.2018.297

Dyché, J. (2001). *The CRM Handbook: A Business Guide to Customer Relationship Management* (1st ed.). Addison-Wesley Professional.

Eccles, R. G. (1991). The performance measurement manifesto. *Harvard Business Review*, *69*(1), 131–137.

Eckerson, W. W. (2010). *Performance dashboards: measuring, monitoring, and managing your business* (2nd ed.). John Wiley & Sons, Inc.

Ennew, C. T., & Binks, M. R. (1996). The impact of service quality and service characteristics on customer retention: Small businesses and their banks in the UK. *British Journal of Management*, *7*(3), 219–230. https://doi.org/10.1111/j.1467-8551.1996.tb00116.x

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904. https://doi.org/https://doi.org/10.1016/j.jbusres.2015.07.001

Eriksson, K., & Vaghult, A. L. (2000). Customer Retention, Purchasing Behavior and Relationship Substance in Professional Services. *Industrial Marketing Management*, *29*(4), 363–372. https://doi.org/https://doi.org/10.1016/S0019-8501(00)00113-9

Eurostat. (2008). *NACE Rev. 2 - Statistical classification of economic activities in the European Community*. Eurostat. https://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-07-015

Evans, J. (2014). *Business Analytics* (1st ed.). Pearson.

Fahrmeir, L., & Kaufmann, H. (1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics*, *13*(1), 342–368.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Few, S. (2005). Dashboard design: Beyond meters, gauges, and traffic lights. *Business Intelligence Journal*, *10*(1), 18–24.

Fornell, C. (1992). A National Customer Satisfaction Barometer: The Swedish Experience. *Journal of Marketing*, *56*(1), 6–21. https://doi.org/10.1177/002224299205600103

Fred, A., & Jain, A. (2002). Data clustering using evidence accumulation. *Object Recognition Supported by User Interaction for Service Robots*, *4*, 276–280 vol.4. https://doi.org/10.1109/ICPR.2002.1047450

Friedl, M., & Brodley, C. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*(3), 399–409. https://doi.org/10.1016/S0034-4257(97)00049-7

Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*(1), 1–28. https://doi.org/10.1002/for.3980040103

Gardner, E. S. (2006). Exponential smoothing: The state of the art-Part II. *International Journal of Forecasting*, *22*(4), 637–666. https://doi.org/10.1016/j.ijforecast.2006.03.005

Grant, R. M. (2016). *Contemporary Strategy Analysis* (9th ed.). John Wiley & Sons Ltd.

Grigoroudis, E., Tsitsiridi, E., & Zopounidis, C. (2013). Linking customer satisfaction, employee appraisal, and business performance: an evaluation methodology in the banking sector. *Annals of Operations Research*, *205*(1), 5–27.

Gronroos, C. (1990). Relationship approach to marketing in service contexts: The marketing and organizational behavior interface. *Journal of Business Research*, *20*(1), 3–11. https://doi.org/https://doi.org/10.1016/0148-2963(90)90037-E

Gummesson, E. (1987). The new marketing—Developing long-term interactive relationships. *Long Range Planning*, *20*(4), 10–20. https://doi.org/https://doi.org/10.1016/0024-6301(87)90151-8

Gummesson, E. (2004). Return on relationships (ROR): the value of relationship marketing and CRM in business-to-business contexts. *Journal of Business & Industrial Marketing*, *19*(2), 136–148. https://doi.org/10.1108/08858620410524016

Hallencreutz, J., & Parmler, J. (2019). Important drivers for customer satisfaction – from product focus to image and service quality. *Total Quality Management & Business Excellence*, 1–10. https://doi.org/10.1080/14783363.2019.1594756

Hallikainen, H., Savimäki, E., & Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management*, *86*, 90–98. https://doi.org/10.1016/j.indmarman.2019.12.005

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufman.

Hollensen, S. (2015). *Marketing Management: A Relationship Approach* (3rd ed.). Pearson.

Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. *Decision Support Systems*, *64*, 130–141. https://doi.org/10.1016/j.dss.2014.05.013

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070

Hu, Y.-H., & Yeh, T.-W. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowledge-Based*

Systems, *61*, 76–88. https://doi.org/10.1016/j.knosys.2014.02.009

Hughes, A. M. (1994). *Strategic Database Marketing*. Probus Publishing Company.

Ibatova, A. Z., Kuzmenko, V. I., & Klychova, G. S. (2018). Key performance indicators of management consulting. *Management Science Letters*, *8*(5), 475–482. https://doi.org/10.5267/j.msl.2018.3.004

Inova+. (2020). *Inova+ Webpage*. https://inova.business

Islam, M. J., Wu, Q. M. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007). Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 1541–1546.

Jacobs, F. R., & Chase, R. B. (2018). *Operations and Supply Chain Management* (15th ed.). McGraw-Hill.

Jahromi, A. T., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*(7), 1258–1268. https://doi.org/10.1016/j.indmarman.2014.06.016

Janzen, F. J., & Stern, H. S. (1998). Logistic Regression for Empirical Studies of Multivariate Selection. *Evolution*, *52*(6), 1564–1571. https://doi.org/10.1111/j.1558-5646.1998.tb02237.x

Jaworski, B. J., & Kohli, A. K. (1993). Market orientation: antecedents and consequences. *Journal of Marketing*, *57*(3), 53–70.

Jayachandran, S., Sharma, S., Kaufman, P., & Raman, P. (2005). The Role of Relational Information Processes and Technology Use in Customer Relationship Management. *Journal of Marketing*, *69*(4), 177–192. https://doi.org/10.1509/jmkg.2005.69.4.177

Kahan, R. (1998). Using database marketing techniques to enhance your one-to-

one marketing initiatives. *Journal of Consumer Marketing*, *15*(5), 491–493. https://doi.org/10.1108/07363769810235965

Kaplan, R. S., & Norton, D. P. (1992). The balanced scorecard--measures that drive performance. *Harvard Business Review*, *70*(1), 71–79.

Kerzner, H. (2017). *Project Management Metrics, KPIs, and Dashboards: A Guide to Measuring and Monitoring Project Performance* (3rd ed.). John Wiley & Sons, Inc.

Kim, H.-S., & Kim, Y.-G. (2009). A CRM performance measurement framework: Its development process and application. *Industrial Marketing Management*, *38*(4), 477–489. https://doi.org/10.1016/j.indmarman.2008.04.008

Kim, S. H., & Mukhopadhyay, T. (2011). Determining Optimal CRM Implementation Strategies. *Information Systems Research*, *22*(3), 624–639. https://doi.org/10.1287/isre.1100.0309

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). John Wiley & Sons, Inc.

Kipping, M., & Engwall, L. (2002). *Management consulting: Emergence and dynamics of a knowledge industry* (1st ed.). Oxford University Press.

Kiron, D., & Bean, R. (2013). Organizational alignment is key to big data success. *MIT Sloan Management Review*, *54*(3), 1–6.

Kitchens, B., Dobolyi, D., Li, J., & Abbasi, A. (2018). Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data. *Journal of Management Information Systems*, *35*(2), 540–574. https://doi.org/10.1080/07421222.2018.1451957

Kotler, P., & Keller, K. L. (2012). *Marketing Management* (14h ed.). Pearson.

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). *Collaborative Customer*

Relationship Management - taking CRM to the next level (1st ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24710-4

Krizanova, A., Gajanova, L., & Nadanyiova, M. (2018). Design of a CRM Level and Performance Measurement Model. *Sustainability*, *10*(7), 2567. https://doi.org/10.3390/su10072567

Kumar, M., & Misra, M. (2020). Evaluating the effects of CRM practices on organizational learning, its antecedents and level of customer satisfaction. *Journal of Business & Industrial Marketing*. https://doi.org/10.1108/JBIM-11-2019-0502

Kumar, V., Lemon, K. N., & Parasuraman, A. (2006). Managing Customers for Value: An Overview and Research Agenda. *Journal of Service Research*, *9*(2), 87–94. https://doi.org/10.1177/1094670506293558

Kumar, Vanya, Simon, A., & Kimberley, N. (2000). Strategic capabilities which lead to management consulting success in Australia. *Management Decision*, *38*(1), 24–35. https://doi.org/10.1108/00251740010311807

Kusiak, A. (2018). Smart manufacturing. *International Journal of Production Research*, *56*(1–2), 508–517. https://doi.org/10.1080/00207543.2017.1351644

Lages, L. F., Lancastre, A., & Lages, C. (2008). The B2B-RELPERF scale and scorecard: Bringing relationship marketing theory into business-to-business practice. *Industrial Marketing Management*, *37*(6), 686–697. https://doi.org/https://doi.org/10.1016/j.indmarman.2007.05.008

Lam, S. Y., Shankar, V., Erramilli, M. K., & Murthy, B. (2004). Customer Value, Satisfaction, Loyalty, and Switching Costs: An Illustration From a Business-to-Business Service Context. *Journal of the Academy of Marketing Science*, *32*(3), 293–311. https://doi.org/10.1177/0092070304263330

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big

data, analytics and the path from insights to value. *MIT Sloan Management Review*, *52*(2), 21–32.

Lee, J. H., & Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. *Expert Systems with Applications*, *29*(1), 145–152. https://doi.org/10.1016/j.eswa.2005.01.013

Lewis, C. D. (1982). *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann.

Liberatore, M. J., & Luo, W. (2010). The Analytics Movement: Implications for Operations Research. *Interfaces*, *40*(4), 313–324. https://doi.org/10.1287/inte.1100.0502

Lilien, G. L. (2016). The B2B Knowledge Gap. *International Journal of Research in Marketing*, *33*(3), 543–556. https://doi.org/10.1016/j.ijresmar.2016.01.003

Lindgreen, A., Palmer, R., Vanhamme, J., & Wouters, J. (2006). A relationship-management assessment tool: Questioning, identifying, and prioritizing critical aspects of customer relationships. *Industrial Marketing Management*, *35*(1), 57–71. https://doi.org/10.1016/j.indmarman.2005.08.008

Ling, R., & Yen, D. C. (2001). Customer Relationship Management: An Analysis Framework and Implementation Strategies. *Journal of Computer Information Systems*, *41*(3), 82–97. https://doi.org/10.1080/08874417.2001.11647013

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development*, *2*(4), 314–319.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281–297.

Marr, B. (2012). *Key Performance Indicators: The 75 Measures Every Manager Needs to Know*. Pearson Financial Times Publishing.

Marr, B. (2017). *Why AI, Machine Learning and Big Data Really Matter to B2B Companies*. Forbes. https://www.forbes.com/sites/bernardmarr/2017/11/03/why-ai-machine-learning-and-big-data-really-matter-to-b2b-companies/

Maxham, J. G., & Netemeyer, R. G. (2003). Firms Reap what they Sow: The Effects of Shared Values and Perceived Organizational Justice on Customers' Evaluations of Complaint Handling. *Journal of Marketing*, *67*(1), 46–62. https://doi.org/10.1509/jmkg.67.1.46.18591

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 60–68.

McCarthy, E. J. (1960). *Basic Marketing: A Managerial Approach* (1st ed.). Richard D. Irwin.

McCarthy, E. J., & Perreault, W. D. (1993). *Basic Marketing: A Global-Managerial Approach* (11th ed.). Richard D. Irwin.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, *60*(6), 656–662. https://doi.org/10.1016/j.jbusres.2006.06.015

Miguéis, V. L. (2012). *Analytical Customer Relationship Management in Retailing Supported by Data Mining Techniques*. Faculdade de Engenharia da Universidade do Porto (FEUP).

Miguéis, V. L., Camanho, A. S., & Falcão e Cunha, J. (2012). Customer data mining for lifestyle segmentation. *Expert Systems with Applications*, *39*(10), 9359–9366. https://doi.org/10.1016/j.eswa.2012.02.133

Mintzberg, H. (1987). The strategy concept I: Five Ps for strategy. *California Management Review*, *30*(1), 11–24. https://doi.org/10.2307/41165263

Mirzaei, T., & Iyer, L. (2014). Application of predictive analytics in customer

relationship management: A literature review and classification. *Proceedings of the Southern Association for Information Systems Conference*, 1–7.

Mithas, S., Krishnan, M., & Fornell, C. (2005). Why Do Customer Relationship Management Applications Affect Customer Satisfaction? *Journal of Marketing*, *69*(4), 201–209. https://doi.org/10.1509/jmkg.2005.69.4.201

Mithas, S., Ramasubbu, N., & Sambamurthy, V. (2011). How Information Management Capability Influences Firm Performance. *MIS Quarterly*, *35*(1), 237–256. https://doi.org/10.2307/23043496

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: a survey. *IEEE Transactions on Neural Networks*, *13*(1), 3–14.

Moreno, J. J. M., Pol, A. P., Abad, A. S., & Blasco, B. C. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*, *25*(4), 500–506. https://doi.org/10.7334/psicothema2013.23

Nam, D., Lee, J., & Lee, H. (2019). Business analytics use in CRM: A nomological net from IT competence to CRM performance. *International Journal of Information Management*, *45*, 233–245. https://doi.org/10.1016/j.ijinfomgt.2018.01.005

Neely, A. (1999). The performance measurement revolution: why now and what next? *International Journal of Operations & Production Management*, *19*(2), 205–228. https://doi.org/10.1108/01443579910247437

Newbold, P., Carlson, W. L., & Thorne, B. (2013). *Statistics for Business and Economics* (8th ed.). Pearson Education, Inc.

Ngai, E. W. T. (2005). Customer relationship management research (1992-2002): An academic literature review and classification. *Marketing Intelligence & Planning*, *23*(6), 582–605. https://doi.org/10.1108/02634500510624147

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining

techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*, 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

O'Neill, M., & Palmer, A. (2001). Survey timing and consumer perceptions of service quality: an overview of empirical evidence. *Managing Service Quality: An International Journal*, *11*(3), 182–190. https://doi.org/10.1108/09604520110391351

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, *64*(1), 12–40.

Park, C., & Kim, Y. (2003). A framework of dynamic CRM: linking marketing with information strategy. *Business Process Management Journal*, *9*(5), 652–671. https://doi.org/10.1108/14637150310496749

Parmenter, D. (2020). *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs* (4th ed.). John Wiley & Sons, Inc.

Parvatiyar, A., & Sheth, J. N. (2001). Customer relationship management: Emerging practice, process, and discipline. *Journal of Economic & Social Research*, *3*(2), 1–34.

Payne, A. (2005). *Handbook of CRM: Achieving Excellence in Customer Management* (1st ed.). Elsevier Butterworth-Heinemann.

Payne, A., & Frow, P. (1999). Developing a Segmented Service Strategy: Improving Measurement in Relationship Marketing. *Journal of Marketing Management*, *15*(8), 797–818. https://doi.org/10.1362/026725799784772666

Payne, A., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. *Journal of Marketing*, *69*(4), 167–176. https://doi.org/10.1509/jmkg.2005.69.4.167

Peltier, J. W., Zahay, D., & Lehmann, D. R. (2013). Organizational Learning and CRM Success: A Model for Linking Organizational Practices, Customer Data Quality, and Performance. *Journal of Interactive Marketing*, *27*(1), 1–13. https://doi.org/10.1016/j.intmar.2012.05.001

Peppers, D., & Rogers, M. (2016). *Managing Customer Experience and Relationships: A Strategic Framework* (3rd ed.). John Wiley & Sons, Inc.

Quinlan, R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers.

Ranjan, J., & Bhatnagar, V. (2011). Role of knowledge management and analytical CRM in business: data mining based framework. *The Learning Organization*, *18*(2), 131–148. https://doi.org/10.1108/09696471111103731

Reimann, M., Schilke, O., & Thomas, J. S. (2010). Customer relationship management and firm performance: the mediating role of business strategy. *Journal of the Academy of Marketing Science*, *38*(3), 326–346. https://doi.org/10.1007/s11747-009-0164-y

Reinartz, W., Krafft, M., & Hoyer, W. D. (2004). The Customer Relationship Management Process: Its Measurement and Impact on Performance. *Journal of Marketing Research*, *41*(3), 293–305. https://doi.org/10.1509/jmkr.41.3.293.35991

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46.

Robinson, J. (1969). *The Economics of Imperfect Competition* (2nd ed.). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-15320-6

Rodriguez, M., & Honeycutt, E. D. (2011). Customer Relationship Management (CRM)'s Impact on B to B Sales Professionals' Collaboration and Sales Performance. *Journal of Business-to-Business Marketing*, *18*(4), 335–356.

https://doi.org/10.1080/1051712X.2011.574252

Roel, R. (1988). Direct marketing's 50 big ideas. *Direct Marketing*, *50*(May), 45–62.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms* (1st ed.). Cambridge University Press.

Sharda, R., Delen, D., & Turban, E. (2018). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective* (4th ed.). Pearson.

Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, *23*(4), 433–441. https://doi.org/10.1057/ejis.2014.17

Sheikh, A., Ghanbarpour, T., & Gholamiangonabadi, D. (2019). A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. *Journal of Business-to-Business Marketing*, *26*(2), 197–207. https://doi.org/10.1080/1051712X.2019.1603420

Sheth, J. N. (2002). The future of relationship marketing. *Journal of Services Marketing*, *16*(7), 590–592. https://doi.org/10.1108/08876040210447324

Sin, L. Y. M., Tse, A. C. B., & Yim, F. H. K. (2005). CRM: conceptualization and scale development. *European Journal of Marketing*, *39*(11/12), 1264–1290. https://doi.org/10.1108/03090560510623253

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, *21*(1), 3–8. https://doi.org/10.1177%2F002224295602100102

Starbuck, W. H. (1992). Learning by knowledge-intensive firms. *Journal of Management Studies*, *29*(6), 713–740.

Stein, A. D., Smith, M. F., & Lancioni, R. A. (2013). The development and

diffusion of customer relationship management (CRM) intelligence in business-to-business environments. *Industrial Marketing Management*, *42*(6), 855–861. https://doi.org/10.1016/j.indmarman.2013.06.004

Stringfellow, A., Nie, W., & E. Bowen, D. (2004). CRM: Profiting from understanding customer needs. *Business Horizons*, *47*(5), 45–52. https://doi.org/10.1016/j.bushor.2004.07.008

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (1st ed.). Pearson.

Tan, P.-N., Steinbach, M., & Kumar, V. (2013). *Introduction to Data Mining* (Pearson Ne). Pearson Education Limited.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tseng, S.-M., & Wu, P.-H. (2014). The impact of customer knowledge and customer relationship management on service quality. *International Journal of Quality and Service Sciences*, *6*(1), 77–96. https://doi.org/10.1108/IJQSS-08-2012-0014

Tsiptsis, K., & Chorianopoulos, A. (2009). *Data Mining Techniques in CRM: Inside Customer Segmentation* (1st ed.). Wiley.

Turban, E., Sharda, R., Delen, D., Aronson, J. E., Liang, T.-P., & King, D. (2011). *Decision Support and Business Intelligence Systems* (9th ed.). Pearson.

United Nations. (2008). *International Standard Industrial Classification of All Economic Activities - Revision 4*. United Nations Publication. https://unstats.un.org/unsd/publication/seriesm/seriesm_4rev4e.pdf

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. H. (2002). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems*, *34*(4), 471–481. https://doi.org/10.1016/S0167-9236(02)00069-6

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

Wei, J.-T., Lee, M.-C., Chen, H.-K., & Wu, H.-H. (2013). Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications*, *40*(18), 7513–7518. https://doi.org/10.1016/j.eswa.2013.07.053

Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (2010). A review of the application of RFM model. *African Journal of Business Management*, *4*(19), 4199–4206.

Werr, A., & Stjernberg, T. (2003). Exploring Management Consulting Firms as Knowledge Systems. *Organization Studies*, *24*(6), 881–908. https://doi.org/10.1177/0170840603024006004

Wilson, R., & Gilligan, C. (2005). *Strategic Marketing Management: Planning, implementation and control* (3rd ed.). Elsevier Butterworth-Heinemann.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324–342.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufman.

Wu, Jing, & Lin, Z. (2005). Research on customer segmentation model by clustering. *Proceedings of the 7th International Conference on Electronic Commerce - ICEC '05*, 316–318. https://doi.org/10.1145/1089551.1089610

Wu, Jun, Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical Problems in Engineering*, *2020*, 1–7. https://doi.org/10.1155/2020/8884227

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.

Xu, M., & Walton, J. (2005). Gaining customer knowledge through analytical CRM. *Industrial Management & Data Systems*, *105*(7), 955–971. https://doi.org/10.1108/02635570510616139

Yankelovich, D., & Meer, D. (2006). Rediscovering market segmentation. *Harvard Business Review*, *84*(2), 1–11.

Zahay, D. (2008). Successful B2B customer database management. *Journal of Business & Industrial Marketing*, *23*(4), 264–272. https://doi.org/10.1108/08858620810865843

Zineldin, M. (2006). The royalty of loyalty: CRM, quality and retention. *Journal of Consumer Marketing*, *23*(7), 430–437. https://doi.org/10.1108/07363760610712975

Appendix

Independent variable		Coefficient	OR (95%CI)	p-value
Type	SME (reference)			
	Education and research organization	0.566	1.761 (0.556, 5.579)	0.336
	Public body	0.168	1.183 (0.453, 3.090)	0.732
	Large enterprise	-0.125	0.883 (0.471, 1.654)	0.697
	Non profit	0.604	1.830 (0.656, 5.103)	0.248
Country	Other (reference)			
	Portugal	2.239	9.382 (3.061, 28.756)	**<0.001**
Economic Sector	Public Administration, Education, Health and Social Services (reference)			
	Professional, scientific and technical activities	0.983	2.672 (1.099, 6.499)	**0.030**
	Manufacturing	1.277	3.587 (1.249, 10.30)	**0.018**
	Service activities (administrative, financial and insurance, tourism, transportation, arts and other activities)	0.528	1.696 (0.591, 4.867)	0.326
	Information and communication	0.918	2.505 (0.796, 7.885)	0.116
	Other activities (Agriculture, Construction, Energy, Water, Wholesale and retail trade)	-0.243	0.785 (0.278, 2.212)	0.646
Subordinated Area	Empresas Norte (reference)			
	Empresas Sul	-0.117	0.890 (0.493, 1.605)	0.698
	Institucional	0.194	1.214 (0.482, 3.053)	0.681
	Científico	0.669	1.952 (0.651, 5.850)	0.233
	Horizon2020	0.738	2.092 (0.569, 7.694)	0.267
	Transformação Digital	-0.662	0.516 (0.135, 1.967)	0.332
	Administration	-0.473	0.623 (0.123, 3.155)	0.568
	CBC - Capacity Building and Cooperation	1.251	3.492 (0.461, 26.474)	0.226
	Support to European Institutions	-0.542	0.582 (0.106, 3.205)	0.534
	Inovação Social	1.081	2.948 (0.287, 30.269)	0.363
	RD - Regional Development	0.713	2.040 (0.092, 45.179)	0.652
Service	National candidature (reference)			
	European candidature	-0.430	0.650 (0.296, 1.431)	0.285
	Specialized services	0.510	1.666 (0.759, 3.658)	0.203
	Tax incentives candidature	0.257	1.293 (0.564, 2.963)	0.544
	National project management	-0.141	0.868 (0.341, 2.213)	0.767
	European project management	0.701	2.017 (0.538, 7.563)	0.298
	Organizational consultancy	-0.353	0.702 (0.254, 1.939)	0.495
	Funding advising	-0.029	0.971 (0.356, 2.648)	0.955
	Representation in Brussels	0.430	1.537 (0.372, 6.341)	0.552
	Business strategy consultancy	-0.067	0.935 (0.129, 6.764)	0.947
	Customized ICT development	-1.710	0.181 (0.016, 2.081)	0.170
	Implementation of standards	1.197	3.309 (0.536, 20.432)	0.198
	Support consultancy for entrepreneurship	-1.034	0.356 (0.043, 2.961)	0.339
	Outsourcing of R&D	1.064	2.897 (0.130, 64.741)	0.502

Independent variable		Coefficient	OR (95%CI)	p-value
	Other	-1.689	0.185 (0.043, 0.798)	**0.024**
Success fee	No (reference)			
	Yes	-0.311	0.733 (0.377, 1.425)	0.360
Year	(years)	-1.324	0.266 (0.056, 1.271)	0.097
Month	February (reference)			
	January	0.976	2.655 (0.647, 10.887)	0.175
	March	-0.437	0.646 (0.181, 2.311)	0.502
	April	-0.187	0.829 (0.222, 3.093)	0.780
	May	0.096	1.101 (0.287, 4.222)	0.888
	June	-0.582	0.559 (0.145, 2.153)	0.398
	July	-0.286	0.752 (0.186, 3.044)	0.689
	August	-0.022	0.979 (0.234, 4.091)	0.976
	September	-0.200	0.819 (0.212, 3.166)	0.772
	October	0.353	0.330 (0.085, 1.277)	0.588
	November	-1.109	0.330 (0.085, 1.277)	0.108
	December	1.116	3.054 (0.706, 13.212)	0.135
Covid-19	No (reference)			
	Yes	2.074	7.953 (1.560, 40.536)	**0.013**
Proposal Value	(Euros)	0.000	1.000 (1.000, 1.000)	0.199

Table 19: Coefficients, OR and p-value of logistic regression full model

Independent variable		Coefficient	OR (95%CI)	p-value
Country	Other (reference)			
	Portugal	2.499	12.171 (4.178, 35.454)	**<0.001**
Economic Sector	Public Administration, Education, Health and Social Services (reference)			
	Professional, scientific and technical activities	0.635	1.886 (0.950, 3.747)	0.070
	Manufacturing	0.583	1.791 (0.841, 3.816)	0.131
	Service activities (administrative, financial and insurance, tourism, transportation, arts and other activities)	0.220	1.247 (0.554, 2.803)	0.594
	Information and communication	0.371	1.449 (0.608, 3.449)	0.402
	Other activities (Agriculture, Construction, Energy, Water, Wholesale and retail trade)	-0.463	0.630 (0.263, 1.508)	0.299
Subordinated Area	Empresas Norte (reference)			
	Empresas Sul	-0.231	0.794 (0.460, 1.371)	0.408
	Institucional	0.011	1.011 (0.461, 2.217)	0.977
	Científico	0.739	2.094 (0.835, 5.254)	0.115
	Horizon2020	0.933	2.543 (0.833, 7.760)	0.101
	Transformação Digital	-0.268	0.765 (0.243, 2.408)	0.647
	Administration	-0.211	0.810 (0.180, 3.653)	0.784
	CBC - Capacity Building and Cooperation	1.805	6.078 (0.943, 39.181)	0.058
	Support to European Institutions	-0.951	0.386 (0.080, 1.873)	0.238
	Inovação Social	1.561	4.762 (0.506, 44.840)	0.173

Independent variable		Coefficient	OR (95%CI)	p-value
	RD - Regional Development	0.461	1.585 (0.113, 22.284)	0.733
Service	National candidature (reference)			
	European candidature	-0.332	0.718 (0.341, 1.511)	0.383
	Specialized services	0.456	1.577 (0.775, 3.209)	0.208
	Tax incentives candidature	0.106	1.112 (0.517, 2.392)	0.787
	National project management	-0.248	0.780 (0.328, 1.858)	0.575
	European project management	0.878	2.406 (0.676, 8.555)	0.175
	Organizational consultancy	-0.148	0.863 (0.335, 2.219)	0.759
	Funding advising	0.111	1.118 (0.436, 2.866)	0.817
	Representation in Brussels	0.009	1.009 (0.284, 3.583)	0.988
	Business strategy consultancy	-0.573	0.564 (0.107, 2.971)	0.499
	Customized ICT development	-2.084	0.124 (0.012, 1.292)	0.081
	Implementation of standards	0.790	2.203 (0.389, 12.486)	0.372
	Support consultancy for entrepreneurship	-1.203	0.300 (0.046, 1.977)	0.211
	Outsourcing of R&D	0.049	1.051 (0.050, 22.060)	0.975
	Other	-1.084	0.338 (0.092, 1.240)	0.102
Covid-19	No (reference)			
	Yes	0.736	2.087 (1.375, 3.167)	**0.001**

Table 20: Coefficients, OR and p-value of logistic regression feature selection model

Independent variable		Coefficient	OR (95%CI)	p-value
Country	Other (reference)			
	Portugal	1.776	5.909 (2.795, 12.490)	**<0.001**
Year	(years)	-0.707	0.493 (0.253, 0.959)	**0.037**
Covid-19	No (reference)			
	Yes	1.278	3.590 (1.811, 7.119)	**<0.001**
Proposal Value	(Euros)	0.000	1.000 (1.000, 1.000)	0.101

Table 21: Coefficients, OR and p-value of logistic regression stepwise model

Independent variable		Coefficient	OR (95%CI)	p-value
Type	SME (reference)			
	Education and research organization	1.025	2.787 (1.156, 6.715)	**0.022**
	Public body	0.151	1.163 (0.515, 2.627)	0.717
	Large enterprise	-0.072	0.931 (0.517, 1.676)	0.811
	Non profit	0.628	1.873 (0.786, 4.465)	0.157
Country	Other (reference)			
	Portugal	2.138	8.480 (3.503, 20.527)	**<0.001**
Economic Sector	Public Administration, Education, Health and Social Services (reference)			
	Professional, scientific and technical activities	0.788	2.198 (1.058, 4.568)	**0.035**
	Manufacturing	0.935	2.548 (1.038, 6.256)	**0.041**
	Service activities (administrative, financial and insurance, tourism,	0.501	1.651 (0.664, 4.106)	0.281

Independent variable		Coefficient	OR (95%CI)	p-value
	transportation, arts and other activities)			
	Information and communication	0.700	2.014 (0.744, 5.450)	0.168
	Other activities (Agriculture, Construction, Energy, Water, Wholesale and retail trade)	-0.204	0.816 (0.333, 1.995)	0.655
Service	National candidature (reference)			
	European candidature	-0.332	0.718 (0.367, 1.402)	0.331
	Specialized services	0.296	1.344 (0.675, 2.675)	0.400
	Tax incentives candidature	0.067	1.069 (0.497, 2.300)	0.864
	National project management	-0.134	0.875 (0.374, 2.044)	0.757
	European project management	1.276	3.581 (1.095, 11.712)	**0.035**
	Organizational consultancy	-0.325	0.723 (0.276, 1.895)	0.509
	Funding advising	0.091	1.096 (0.428, 2.803)	0.849
	Representation in Brussels	0.120	1.127 (0.327, 3.882)	0.849
	Business strategy consultancy	-1.084	0.338 (0.065, 1.753)	0.197
	Customized ICT development	-2.507	0.081 (0.009, 0.712)	**0.023**
	Implementation of standards	0.813	2.255 (0.399, 12.756)	0.358
	Support consultancy for entrepreneurship	-1.338	0.262 (0.043, 1.609)	0.148
	Outsourcing of R&D	-0.506	0.605 (0.036, 10.217)	0.727
	Other	-1.182	0.307 (0.086, 1.090)	0.068
Covid-19	No (reference)			
	Yes	0.763	2.144 (1.421, 3.237)	**<0.001**

Table 22: Coefficients, OR and p-value of logistic regression – model 4