FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

# Analyzing the Impact of Next Best Offer and Churn Methodologies on Customizing Promotions in the Insurance Industry

**Tomás Martins Araújo**


Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Prof. Maria Beatriz Brito Oliveira

February 9, 2023

# Abstract

The Insurance Brokerage industry is highly competitive, and maximizing profits requires effective customer relationships and retention management. In this context, developing a Next Best Offer with the Multi-Risk House Insurance and Churn prediction system using machine learning techniques can provide a significant competitive advantage.

This thesis presents two systems for an insurance brokerage firm that aims to improve profitability by identifying opportunities for cross-selling and preventing Churn. The systems are evaluated by selecting two groups of clients (treatment and control), which were contacted to demonstrate the effectiveness in identifying the Next Best Offer and predicting Churn.

The systems referred to in this study were made using data management and machine learning techniques. After receiving the data sources from the Insurance Brokerage, the data was manipulated to provide the model with the best subset of features. Additionally, to provide more information about the clients, socio-demographic data was used as external information.

Concerning the empirical analysis, various machine learning techniques were employed to determine the best suited model for each system. The evaluation metrics utilized in this study included the F1 score and its modifications, the F2 and F0.5 scores. Additionally, an examination of the most influential variables and a shap summary analysis were conducted to gain a deeper understanding of the behaviour of the machine learning model. Then, two subsets of clients were selected. The first, called treatment, is the subset given by the model with a higher probability of having Multi-Risk House Insurance or churning, while the second, called control, is, in the case of the Next Best Offer, a random selection of clients also to be contacted and, in case of churning, is a selection of clients with also a high probability of churning to monitor if they have churned or not. Regarding the results from these contacts, while in the Next Best Offer system it was possible to have two months of contacts before the delivery of this dissertation, in the churning there was still no contact made before delivering this work.

The results from the Next Best Offer contacts show that the system significantly improves the firm's ability to target customers who are likely to buy Multi-Risk House Insurance and, thus implementing this system can provide a competitive advantage for the insurance brokerage firm and support its long-term growth by optimizing critical decision-making and maximizing the benefits of customer relationships.

# Resumo

A indústria de Corretagem de Seguros é altamente competitiva, e a maximização dos lucros requer um bom conhecimento das necessidades dos clientes. Neste contexto, o desenvolvimento de um sistema de Próxima Melhor Oferta com o seguro de Habitação Multi-Risco e um sistema de Previsão de Churn utilizando técnicas de *machine learning* podem proporcionar uma vantagem competitiva significativa.

Esta tese apresenta dois sistemas para uma Corretora de Seguros portuguesa que visa melhorar a rentabilidade através da identificação de oportunidades de venda cruzada e da prevenção de Churn. Os sistemas são avaliados através da selecção de dois grupos de clientes (tratamento e controlo), que foram contactados para demonstrar a eficácia na identificação da Próxima Melhor Oferta e na previsão do Churn.

Os sistemas referidos neste estudo foram feitos utilizando gestão de dados e técnicas de *machine learning*. Após receber as fontes de dados da corretora de seguros, os dados foram tratados para fornecer ao modelo o melhor subconjunto de características. Além disso, foram também utilizados como fonte de informação externa, dados socio-demográficos que foram cruzados com a localização de cada cliente.

Relativamente aos resultados, foram testadas diversas metodologias, nomeadamente diferentes tipos de modelos de *machine learning* para obter o melhor modelo para cada caso de uso. As métricas de avaliação usadas foram o F1 bem como as suas modificações F0.5 e F2. Adicionalmente, foi realizada uma análise detalhada das variáveis mais importantes de cada um dos modelos de aprendizagem de máquina bem como da análise *shap* para uma melhor compreensão do comportamento dos modelos.

Em seguida, para obter resultados práticos da implementação dos modelos, foram selecionados dois grupos (tratamento e controlo). No caso da Próxima Melhor Oferta, o grupo de tratamento representa os clientes a contactar com maior probabilidade de terem seguro Multi-Risco Habitação enquanto o grupo de controlo representa uma amostra aleatória de clientes. No caso do Churn, o grupo de tratamento representa clientes com alta probabilidade de churn que irão ser contactados com uma proposta de renovação enquanto que o grupo de controlo representa clientes que não irão ser contactados. Relativamente aos contactos do Churn, estes apenas serão iniciados após a entrega desta dissertação.

Em suma, os resultados da Próxima Melhor Oferta mostram que o sistema melhora significativamente as hipóteses da empresa localizar clientes recetivos ao seguro Multi-Risco Habitação e consequentemente, a implementação deste sistema pode proporcionar uma vantagem competitiva para a empresa de Corretagem de Seguros e apoiar o seu crescimento a longo prazo através da optimização da tomada de decisões críticas e maximizando os benefícios das relações com os clientes.

# Acknowledgments

Esta dissertação marca o fim de uma das etapas mais importantes da minha vida, a conclusão do Mestrado em Engenharia Eletrotécnica e Computadores. Este marco não seria possível sem o apoio e ajuda de várias pessoas, às quais estou eternamente grato.

Aos meus orientadores, à professora Beatriz Brito Oliveira e ao Horácio Neri por todo o apoio e orientação no desenvolvimento deste trabalho e à LTPLabs pela oportunidade de desenvolver este projeto e pela forma como me acolheram. Um agradecimento especial ao Manager Diogo Sá por todo o acompanhamento e ajuda nestes 6 meses.

Aos meus amigos da faculdade pelos momentos inesquecíveis, noitadas e fins de semana perdidos na FEUP que nunca iremos esquecer. Um abraço especial ao Guilherme Pinheiro por todos aqueles "*extra mile*" nos trabalhos de grupo e outro ao Francisco Caetano pela paciência nos momentos de desespero e por estar sempre recetivo a ajudar.

Aos meus amigos de longa data por todo o apoio nos momentos difíceis. Marcelo, Guida, Tiqs, Borges, Trigo, João Paulo, Sara, Carol e Soto, não teria sido possível sem vocês. À malta do polo especialmente Rodrigo, Tiago e Zé pelos últimos 2 anos espetaculares. Ao grupo de corrida especialmente Jorge Basto, Matos e ao grande Jorge Teixeira pela amizade, pelas histórias de aventuras que tiveram com o meu pai e por todo o apoio.

À tia Isabel, avó Nanda e Gaspar por todo o apoio e carinho. Ao avô Mário por todas as viagens e lições. À minha mãe por todo o apoio, paciência e por todas as horas de trabalho estes anos todos para que eu pudesse e possa perseguir os meus sonhos. Ao avô Lino e ao meu Pai por se fazerem sentir tão presentes mesmo na sua ausência. Obrigado por todos os momentos que partilhamos e nunca esquecerei todos os ensinamentos e todas as aventuras.

Um último agradecimento à Joana por todo o apoio, carinho e paciência. Obrigado por acreditares sempre em mim e por tornares este caminho muito mais fácil.

Tomás Martins Araújo

*"Tantas vezes pensamos ter chegado, tantas vezes é preciso ir mais além"*

Fernando Pessoa

# Contents

# List of Figures

# List of Tables

# Acronyms and Symbols

| | |
|---|---|
| TP | True Positive |
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |
| FPR | False Positive Rate |
| AUCPR | Area Under the Precision-Recall |
| AUC | Area Under the Curve |
| GBM | Gradient Boosting Machine |
| B2C | Business to Consumer |
| B2B | Business to Business |
| B2B2C | Business to Business to Consumer |
| SQL | Structured Query Language |
| SHAP | SHapley Additive exPlanations |

# Chapter 1

# Introduction

The purpose of this chapter is to introduce the Insurance Brokerage industry, provide context for the development of the Next Best Offer and Churn Prediction project, and outline the structure of the present document. Specifically, this chapter will outline the motivations behind the creation of the project, including the need for improved customer retention and cross-selling strategies in the insurance brokerage industry and the potential benefits of using machine learning for predictive analytics. The chapter will also outline the project's objectives, including developing a machine-learning model for predicting customer Churn and identifying the clients with a higher probability of buying the Multi-Risk House Insurance. Finally, the chapter will provide an overview of the structure of the present dissertation, including a description of the topics covered in each chapter.

## 1.1 Motivation

Over the past few decades, the financial industry has been collecting and utilizing vast amounts of data, driven by the increasing competitiveness of the sector and the emergence of tech-savvy financial companies (Srivastava and Gopalkrishnan, 2015). While the insurance industry has not traditionally been at the forefront of adopting data-driven technologies, it has recently begun to leverage analytical methodologies to improve customer relations.

Maintaining a loyal customer base is crucial for the success of any insurance company. The cost of acquiring a new customer is typically much higher than retaining an existing one, making customer retention a key concern for insurance companies (Manzano-Machob, 2013). Predictive analytics can help insurance companies identify at-risk customers and intervene with targeted interventions before they Churn. Additionally, data-driven approaches can help companies identify opportunities for cross-selling and upselling to existing customers by identifying the clients with a higher probability of buying Multi-Risk House Insurance.

Given the potential benefits of using data analytics for customer retention and upselling in the insurance industry, there is a need for research on practical approaches to predicting customer Churn and identifying the Next Best Offer. This research project seeks to develop two machine

learning models for predicting customer Churn and identifying the Next Best Offer for a Portuguese Insurance Brokerage. Specifically, in this pilot project, the Next Best Offer use case focuses on finding the clients with the highest probability of purchasing Multi-Risk House Insurance. The resulting models will be trained on customer data, such as demographics and financial characteristics, providing valuable insights for the Insurance Brokerage.

## 1.2 Objectives

The main objective of this research is to develop and evaluate machine learning models for predicting customer Churn and finding the clients with the highest probability of purchasing Multi-Risk House Insurance for a Portuguese Insurance Brokerage. Developing effective models for predicting customer Churn and identifying the Next Best Offers is a valuable area of research, as it can improve customer retention and increase sales for the Insurance Brokerage. This project aims to contribute to this body of knowledge by addressing the following specific objectives:

1. Collect and prepare a dataset of customer data, including demographics, behaviour, and financial characteristics, for use in the machine learning models. This objective involves the selection and preparation of relevant data sources, as well as the creation of relevant features for model training and evaluation.

2. Develop and train machine learning models for predicting customer Churn and identifying clients with a higher probability of buying Multi-Risk House Insurance based on the collected data. This objective involves the selection and implementation of appropriate machine learning algorithms, as well as the optimization of model hyperparameters through the use of techniques such as cross-validation.

3. Evaluate the performance of the machine learning models using appropriate metrics and techniques. This objective involves the selection of relevant evaluation metrics and the application of these metrics to assess the performance of the models on the collected data.

4. Identify the factors that are most predictive of customer Churn and Next Best Offer acceptance and provide insights for the insurance brokerage seeking to improve customer retention and increase sales. This objective involves interpreting the machine learning models and identifying key drivers of Churn and Next Best Offer acceptance.

5. Development of contact interfaces to monitor and track the practical results of Churn and Next Best Offer interventions. These interfaces will enable the Insurance Brokerage to monitor the effectiveness of their customer retention efforts and identify areas for improvement. The results of the contact interventions, including the response rates and outcomes of the interventions, will be tracked through the use of these interfaces. This will provide valuable insights into the effectiveness of the interventions and help the Insurance Broker optimize their customer retention strategies.

By achieving these objectives, this research aims to significantly contribute to the field of data analytics and machine learning in the insurance industry. The results of this project will provide valuable insights into the use of predictive analytics for improving customer retention and increasing sales. They will serve as a foundation for future research in this area. Additionally, the development of practical machine learning models for predicting customer Churn and identifying the Next Best Offer has the potential to benefit the Insurance Brokerage by providing them with tools for optimizing their customer retention strategies and improving their competitive position.

## 1.3 Dissertation Structure

This dissertation is divided into six chapters, each covering a different aspect of the research project. Chapter 1, the introduction, provides a general overview of the research topic and sets the stage for the rest of the dissertation. The second chapter of this dissertation is dedicated to a comprehensive review of the relevant literature on the research topic. The literature review covers a range of topics, including marketing concepts and their application in analytics, customer relationship management, and the contextualization of the financial industry, with a focus on previous research in the banking and insurance sectors. Additionally, the literature review covers data mining and machine learning techniques, which are central to the research approach adopted in this project. The third chapter of the dissertation focuses on the research problem and context, including a discussion of the Insurance Brokerage's positioning in the Portuguese market and an overview of the project structure. The fourth chapter describes the research methodology in detail, including the data sources, models, and methods used for model development and evaluation. The fifth chapter presents the study's results, including a detailed assessment of the performance of the machine learning models and an analysis of the practical results regarding the client's contacts. The dissertation's final chapter summarises the main conclusions of the research and discusses the implications of the results for the Insurance Brokerage industry. The chapter also identifies areas for future work, including potential avenues for further analysis based on the findings of this study.

# Chapter 2

# Literature Review

This chapter addresses current knowledge about key concepts for this project by examining previously published studies. The first section 2.1 aims to explore Marketing concepts and their application with analytics and big data. Further on, Section 2.2 focuses on Customer Relationship Management and its impact on a company. Section 2.3 aims to introduce two sectors that belong to the Financial Industry. While the main focus of this work is the insurance sector, most of the past research has been performed on the banking sector, which will serve as an example of what can be achieved using analytical tools on financial firms. Moreover, Section 2.4 explores the methodology tools such as data mining and machine learning. Inside the topic of machine learning, techniques such as feature selection, dealing with imbalanced datasets, customer clustering, classification algorithms and how to evaluate them will be explored.

## 2.1 Marketing

According to the American Marketing Association (2022), Marketing is defined as the activity, set of institutions and processes for creating, communicating, delivering and exchanging offerings that have value for customers, clients, partners and societies at large. According to Lamb et al. (2012), a company can have four different philosophies influencing its marketing approach: production, sales, marketing and societal. A production orientation means that the company focus on production and distribution efficiency. Sales philosophy is based on the fact that consumers will only buy the product after intense promotion and intense sales techniques implementations. According to Armstrong (2009), this orientation is more common in industries with unrequested goods and services that people do not think they need, like insurance or organ donation. The marketing concept focuses on delivering customer wants and needs better than competitors and integrates all the organization's activities to meet customer desires. For a company to achieve this, it has to obtain information about other players, customers and markets. The societal marketing philosophy defends that marketing strategy should deliver value to customers to meet the organization's objectives and preserve the individual's and society's long-term best interests. It is an extension of the marketing concept with a stronger focus on societal gains.

The development of technology and its use in various sectors allowed the marketing field to redefine its strategies and techniques. Companies' marketing philosophy associated with the newest analytics tools gave firms more opportunities to approach their customers. Marketing Analytics emerged as one of the main areas of Business Intelligence where data collection became important to maximize firms' returns on their marketing strategies (Wedel and Kannan, 2016). Marketing analytics is defined as a field that connects marketing strategies with large-scale data. Across a broad spectrum of industries, marketing researchers use structured, and unstructured data for customer relationship management (France and Ghose, 2019). Customer relationship management is the process of optimizing profitability, revenue and customer satisfaction by targeting customer value (Armstrong, 2009). Some examples of customer relationship management systems are Next Best Offer, Customer Satisfaction Surveys, Customer Support Automation, among others. According to Lamb et al. (2012), businesses that adopt these systems are most probably market-oriented, tailoring their products and services based on data generated through interactions with customers.

## 2.2 Customer Relationship Management

This study aims to address marketing from a market-orientation point of view and explore analytical models to identify cross-selling opportunities and possible churn candidates. This section explores the current state of knowledge on these topics beginning with an overview of cross-selling and progressing to a more in-depth examination of cross-selling and retention strategies.

### 2.2.1 Cross-selling

Cross-selling is a selling technique performed by multiple companies and is defined as selling a product to a customer that has already bought a product from that company. Within customer relationship management, according to Kamakura et al. (2003), cross-selling has evolved into a valuable strategy for several reasons:

1. The concept that the cost of maintaining an existing customer is cheaper than the cost of acquiring a new one by around a factor of five.

2. On average, cross-selling initiatives have recorded response rates 2 to 5 times higher than cold sales.

3. Cross-selling broadens the customer connection scope, increasing the firm's "share of mind" and its share of the client's wallet. It also increases the customer lifetime value, the total amount of money a customer is expected to spend with the business.

4. Cross-selling increases the physical and emotional costs of switching, increasing retention by enlarging the scope of the relationship.

5. As a client purchases more goods and services from a company and extends the length of their engagement, the company has more insight into the interests and preferences of

the customer, enhancing its capacity to focus on marketing initiatives and cross-sell. The company gains a virtual local monopoly due to its information advantage and the higher switching costs. By leveraging existing customer relationships and possessing insights into their needs and preferences, a business is better positioned to compete for its customer base than its counterparts without such advantages.

For the reasons mentioned above, cross-selling is being employed by many marketing-orientated businesses as a technique for client development. According to Davenport et al. (2014), the latest cross-selling implementations include information such as past purchases, personal factors (age, gender), and previous responses to marketing campaigns, among others. With this information, marketers develop marketing strategies targeting customer preferences and identifying the Next Best Offer.

#### 2.2.1.1 Next Best Offer

Next Best Offer strategies were born based on finding the best product or service to present to the customer. They are a form of cross-selling that uses marketing analytics so that companies can recommend the best products to a specific customer (Davenport et al., 2014). They consist of business strategies combined with predictive models that use data collected from their customers.

There are examples such as the use of machine learning Algorithms for product recommendation systems. The study from Xu et al. (2014) uses the combination of customer segmentation and the Apriori Algorithm of Weka Software. The presented study builds a product recommendation system based on a customer segmentation model using clustering analysis and neural networks and then uses the Apriori algorithm to find the association rules. Association rules are valuable connections between item sets. The experimental results show that on insurance data, the product recommendation model based on customer segmentation has more accurate results. Customer segmentation not only reduces the number of association rules but also improves the focus of the recommendations. Another example is in a study performed in the banking industry, where banks are also using internal data to deepen customer relationships and build Next Best Offer models. According to Agarwal et al. (2021), Ping An Bank has created a predictive algorithm to determine the best product-per-customer ratio for every user. When a customer's needs analysis predicts a product usage ratio of eight products, but they only use two, management is prompted to get in touch with the consumer and upsell or cross-sell appropriate ecosystem products.

### 2.2.2 Churn

Churn or Churn rate is a business metric that consists of the number of clients a business has lost (Scriney et al., 2020). Customer relationship management systems predict customer Churn to retain valuable clients by recognizing comparable customer groups and offering competitive offers/services to each group. As it was previously mentioned, different studies have shown that the cost of retaining existing customers is normally lower than the cost of acquiring a new one (Kamakura et al., 2003; Manzano-Machob, 2013). The latter presented study refers to churn in the

telecom sector, where attracting new customers is estimated to be 5 to 6 times more expensive than retaining the existing ones. Another example of churn impacts is the one presented in Desirena et al. (2019). This study refers to the insurance sector and states that it is common for insurers to lose about 15% of their clients each year. Therefore, in the insurance industry, a 5% decrease in churners might lead to a considerable increase in average annual growth.

Regarding the methodology for churn problems, the latest research validated the hypothesis that machine learning is a highly efficient method to to predict the likelihood that customers will leave the company (Ahmad et al., 2019). Consequently, numerous algorithms have been tested, and the results of this study show that tree-based algorithms such as XGBoost have proven to have good accuracy results for this problem. Another study presents similar results regarding the best models for this kind of problem. According to Ullah et al. (2019), the Random Forests, a tree-based algorithm, has shown better results in the telecom sector churn dataset.

## 2.3 Financial Industry

This section presents a contextualization of the financial industry and its institutions. A financial institution has as its primary purpose to offer a wide range of investment, lending and deposit solutions both to the individual consumer and businesses. There are many types of financial institutions (banks, insurance companies, brokerage firms, and others), and although this dissertation focuses on the insurance sector, most previous research on this industry has been performed on banks. For this reason, this section examines earlier studies on both sectors.

### 2.3.1 Banking Sector

Financial firms have been collecting extensive amounts of data for decades and have recently started to use it after the big data revolution (Srivastava and Gopalkrishnan, 2015). Inside this industry, the banking sector has always had access to salaries, spending habits and investment patterns. It is one of the sectors with more access to personal information (Lau et al., 2003). Banking firms started to adopt analytical strategies to estimate their customers' financial maturity and compute the probability of adhering to new financial services. For this purpose, the relationship between variables such as household income, customer purchase patterns and household education are used to predict customers' life stages (Li et al., 2005). As a result, some services could be recommended according to the characteristics of these different stages of life. Another example is using statistical models such as logistic regression and multinomial logit models to predict the next best service for a customer (Knott et al., 2002). This study discusses the guidelines for a Next Product to Buy System in a bank and compares different statistical models. According to the results, logistic regression was the best-suited model for that particular data set. The model was then evaluated against other recommendation systems, a heuristic mail group (mail group based on wealth-related variables) and a prospect mail group (mail group for customers who do not belong to the bank). The results of this test show that the Next Product to Buy System is extremely valuable once it has achieved a 530% of Return on Investment (ratio of total profit to total

mailing costs) when the other two systems, the mailing groups, have achieved a negative Return on Investment.

Referring to churn in banking, Xie et al. (2009) proposes an improved balanced random forests algorithm and evaluates the model using lift. Lift measures how much the model outperforms a random selection. The model achieved a 7.1 lift which means that using the model gives 7.1 times a better chance at selecting a customer that will churn. The study performed by Gorgoglione et al. (2011) compares two predictive models, but rather than choosing the models based on accuracy, the bank managers selected based on their view of the problem. Their view was based on the fact that the cost of a false negative (churner who is predicted as not leaving the company) was considered substantially higher than the cost of a false positive (customer that is not going to churn but is predicted as churner). As a result, they chose the model with the highest number of relevant instances (churners) predicted. They opted to contact more customers, even loyal customers, rather than risking not identifying actual churners. After the model identified customers with a high risk of churn, the managers decided what service to offer to each customer. The research showed promising results in the automation of targetting churners.

### 2.3.2 Insurance Sector

The insurance sector is an important segment of the financial industry. It consists of companies that provide risk management options through insurance contracts. It generates a great deal of transaction data and is now facing significant growth opportunities (Xu et al., 2014). Its data consists of numerous health, life and auto insurance policies linked with the respective customers' personal information (Morik and Köpcke, 2004). Given this fact, there is a large window to explore this data for customer relationship management. As sensitive as the banking sector, insurance is a very delicate product (Lesage et al., 2020). A high level of confidence is needed from the customer side to adhere to a new service. Wrong-recommended insurance products can have an extremely negative impact on companies' reputations.

Qazi et al. (2017) have performed customer segmentation and developed a recommendation system that predicts relevant products according to the groups of clients. This study chose Bayesian Networks, a probabilistic graphical model that uses a directed acyclic graph to describe a set of variables and their conditional dependencies to model its system. The proposed model went live in the US, and after 112 days of using the model, the average conversion rate was around 20%, well above the baseline of 12% per state (the study defines it as the standard conversion rate). It is also important to mention that agents of this insurance company answered with good feedback to the model once it helped prompt the recommendation coverage for each customer. Another study proposes comparing machine learning algorithms applied to the churn problem of an insurance company (Scriney et al., 2020). The study focuses on different techniques, such as neural networks, support vector machines and decision trees. These techniques were then evaluated with accuracy, precision and recall metrics. In this particular dataset, the neural network was the best-performing model in terms of accuracy; however, decision trees remained more precise

(better at predicting relevant instances), therefore decision trees were better at identifying churners (positive class).

## 2.4 Methodology

The following section discusses past research on methodologies that fit the purpose of this study. Section 2.4.1 starts by giving a brief overview of data mining and how data usage has impacted businesses. The following section, 2.4.2, is centred around machine learning and methodologies often used on machine learning problems.

### 2.4.1 Data Mining

Nowadays, with the increase in storage capabilities, more and more data is being stored and becoming readily available for data scientists, and practitioners (Elgendy and Elragal, 2014). Data mining is the process of discovering patterns and structures in large data sets and turning them into useful information. Implementing this process is crucial for transforming traditional experience-based decision-making into data-driven decision-making (Li et al., 2022).

Different methodologies of data mining for the problem at hand have become popular, and Anand et al. (1998) proposes a solution to tackle the cross-sales problem by creating characteristic rules and filtering them with deviation detection. On the other hand, Raguseo (2018) surveyed companies about the impacts of big data and data mining techniques and discovered that companies that use big data usually achieve their efficiency goals. This study also discovered that data has an immense impact on enabling a quick response to new market conditions. Still, regarding the effect of data mining on big firms, companies like GE are placing sensors on their products to collect data and understand possible improvements. GE estimated that a reduction of 1% in fuel consumption could result in a $30 billion savings for the commercial airline industry (Davenport and Dyché, 2013).

### 2.4.2 Machine learning

Machine learning has gained much attention in the digital sphere as a critical element of digitalization solutions. It is defined as the capacity of a computer system to learn and perform tasks without being given explicit instructions. This process is performed with the applications of algorithms and statistical models to data analysis and inference. As the machine gains experience, it can build an analytical model that can relieve the human capacity to identify relationships and logic (Ray, 2019).

When training a specific machine learning model, one may select from: supervised learning, unsupervised learning, and reinforcement learning (Janiesch et al., 2021). Supervised learning requires a training data set with inputs and outputs, and the machine learning model will fine-tune its parameters according to the respective combination of inputs and outputs. In its simplest form, supervised learning refers to teaching or training a computer system using labelled data, which

indicates that the correct answer has already been assigned to specific data. Regarding unsupervised learning, this technique is supposed to detect patterns without any previous information of references and labels. An unsupervised learning algorithm will act on the data without supervision. An example of an unsupervised learning application is the clustering of data. Clustering is the grouping of data and can be used, for instance, for grouping customers using purchasing behaviour or customer personal information such as age, income, among others. Further details on clustering and segmentation will be discussed in the following Section. Another discipline of machine learning is reinforcement learning, which involves acting appropriately to maximize reward in a specific circumstance. This differs from supervised learning, where the training data includes the solution key, and the model is trained with that answer. Reinforcement learning models must learn from their mistakes in the absence of a training dataset.

Regarding supervised machine learning models, we can have two distinct types. They can be regression models where the predicted variable is a continuous quantity and classification models that predict a discrete class label. The typical evaluation metrics on the first type comes from estimating the errors like MSE (Mean Square Error) or RMSE (Root Mean Square Error), whereas the performance of the second one is evaluated by calculating the correctly and incorrectly predicted classes. These output in a few metrics such as accuracy, recall, precision and F1 score (Janiesch et al., 2021).

The following Sections will address topics related to machine learning. The Section 2.4.2.1 addresses segmentation and clustering, which is, as previously mentioned, a technique used for customer grouping in business machine-learning applications. The Section 2.4.2.2 explores how to deal with imbalanced datasets, a problem that occurs on classification machine learning tasks when a class has more samples than the others. On section 2.4.2.3, methods for performing feature selection will be explored. Feature selection is an important technique for noise reduction with datasets with large amounts of features.

Moreover, this study concentrates on classification machine learning problems and provides a succinct summary of the current state-of-the-art algorithms in Section 2.4.2.4. Furthermore, to determine the most suitable classification algorithm, it is imperative to evaluate their performance. This evaluation is achieved by comparing classification metrics, which are discussed in Section 2.4.2.5.

#### 2.4.2.1 Segmentation and Clustering

Segmentation and clustering refer to the process of separating items by applying various grouping criteria or norms (France and Ghose, 2019). These groups of separated items are defined as clusters. Consumer clusters are applied to produce a small set of groups that aggregate customers according to their behaviour and characteristics.

There are multiple algorithms for clustering, such as K-Means, Mean-Shift, DBSCAN, Hierarchical and Two-Phase Clustering. K-means divides the data into pre-assumpted k-groups (number of clusters). Its process starts at the beginning by randomly selecting centroids and data points that are assigned to the closest centroid forming the cluster. New cluster centroids are recalculated in

each iteration, and the process continues until there is no change in centroid position (Zakrzewska and Murlewski, 2005). Unlike K-means, Mean-Shift does not require any assumptions (number of centroids); hence, it is a non-parametric algorithm. This algorithm defines windows and places them randomly across the data points. Afterwards, it calculates the mean for all points in the window, moves the centre of the window to the mean location and repeats this process until there is convergence. Hierarchical clustering seeks to establish a hierarchy of clusters inside the data that resembles a tree. Starting with N clusters for N data points, the clustering process merges the nearest data points in each step so that there is one fewer cluster in that phase than there was in the step before (Kansal et al., 2018). DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. Density refers to the closeness of data points within a cluster, and this method uses this approach to discover density-connected objects that are maximal concerning density-reachability. Any object that is not a part of any cluster is defined as noise. Two-phase clustering is a technique that also possesses outlier detection and consists of two stages. The first stage is a modified k-means algorithm where if a new input pattern is far enough away from all clusters, it is assigned as a new cluster centre. With the clusters obtained in the first step serving as nodes, a minimal spanning tree is built in the second phase. The tree is then pruned by deleting the longest edge. An approach like this enables efficient outlier detection (Zakrzewska and Murlewski, 2005).

Regarding applications of these algorithms and their results, Kansal et al. (2018) compared K-Means, Hierarchical Clustering and DBSCAN on a 200 sample dataset containing two features. The results were evaluated with Silhouette Score, which outputs how well the data point has been clustered into the correct cluster. In this study, K-means was the best performer. Regarding a more driven application in topics this study is focused on: Financial industry, Zakrzewska and Murlewski (2005) proposed the comparison between DBSCAN, K-Means and the Two-Phase methods in bank customers datasets. This study presents an interesting approach once it considers high dimensionality (high number of features), outlier detection and scalability (datasets with different sizes). The results show that the K-Means algorithm is highly effective for handling huge multidimensional data sets, although it heavily depends on selecting the input parameter k. Thus, the study does not recommend K-Means for data with higher variation. On the other hand, Two-phase clustering algorithm shows interesting results on data with noise and a small number of features. Regarding DBSCAN algorithm, the wrong choice of inputs may lead to poor-quality clusters and the detection of an excessive number of outliers, thus being the most complex algorithm to implement. In what regards the selection of the input parameter k when using the K-Means method, a common way of selecting it is by using the elbow method. It works by fitting the model with a range of different values for k and then plotting the sum of squared errors (SSE) for each value of k. The SSE is the sum of the squared distances between each point and its assigned cluster centre. As the number of clusters increases, the SSE will typically decrease as each point is given a more accurate, or tighter, fit by its cluster centre. However, at some point, the decrease in SSE will level off, and the gain in accuracy will not be worth the increase in complexity of the model. This point is known as the "elbow," and the number of clusters at this point is selected as

the best value for k (Kansal et al., 2018).

#### 2.4.2.2 Imbalanced Datasets

Imbalanced Data-sets are a problem often found in real-world applications where a class contains more samples than the rest of the classes. The dominant class is named the basic class, and the other is known as the secondary class. Even though machine learning classifying algorithms are typically created to reduce error (the percentage of incorrect predictions), not all consider the balance of the classes, which could lead to unfavourable outcomes. In this scenario, some classifiers achieve poor accuracy on the minority class because they give the same importance to these classes while training (Ahmad et al., 2019).

Nevertheless, there are already solutions to deal with this problem. These solutions can be at the data, and algorithmic levels (Chawla et al., 2004). At the data level, the solutions consist of re-sampling the original data set. It can either over-sample the minority class and thus increase the number of samples of this class or under-sampling the majority class, reducing the number of its occurrences. Oversampling has the disadvantage of creating artificial data for the minority class, while under-sampling has the drawback of eliminating valuable data from the majority one (Ganganwar, 2012).

On the other hand, Ganganwar (2012) also talks about the algorithmic techniques, which may include, for instance, the adjustment of the probabilistic estimate at the tree leaf on decision trees. Essentially, this solution adjusts costs or thresholds to deal with imbalanced classes. This type of technique falls into the category of cost-sensitive learning. This presented study identified that hybrid techniques that combine data and algorithmic solutions perform best when dealing with unbalanced data.

#### 2.4.2.3 Feature Selection

Machine learning classification implementations are associated with classifying objects depending on numerous features from large data sets. With a substantial amount of features, a learning model tends to overfit, and performance tends to decrease. In order to address this problem, dimensionality reduction techniques have been studied, and feature selection is one of the practitioners' most used techniques. It consists of selecting a subset of relevant features from the original feature set according to specific relevance evaluation criteria. The main objective of feature selection is to remove irrelevant and redundant features from the dataset (Alelyani et al., 2018).

Furthermore, there are some practical implementations to solve this problem. Alelyani et al. (2018) talks about three algorithms: filter, wrapper, and embedded models. Filter models are known for separating feature selection from classifier learning to prevent the bias of a learning algorithm from interfering with the bias of a feature selection algorithm. It uses statistical tests, such as correlation, distance, and fisher score, to determine the relevance of the features to the output. On the other hand, Wrapper models evaluate the quality of chosen features based on the predicted accuracy of a preset learning method. In its simplest form, the idea of wrapper methods

is features competing against each other for a position in the final subset. An example of these algorithms is Boruta, developed by Kursa et al. (2010), which goes further than just a features' competition. Instead, in Boruta algorithm, the features face off against a randomised version of themselves (shadow features). A feature is only chosen if it outperforms the top randomised feature. The algorithm applies the following sequence:

1. Create duplicate "shadow" versions of each feature in the training set.

2. Shuffle the shadow variables to remove any dependence on the response variable.

3. Select all features with a higher maximum Z score (MZSA) than the Z scores of the shadow attributes, as determined by running a Random Forest classifier on the extended dataset.

4. Remove features with importance significantly lower than the importance of the MZSA, as determined by a two-sided equality test with the MZSA.

5. Select features with importance significantly higher than the importance of the MZSA, as determined by a two-sided test of equality with the MZSA.

6. Repeat the above steps until all features have been assigned or a predetermined number of iterations has been reached, and then remove all shadow attributes.

Despite its effectiveness in feature selection, one potential limitation of the Boruta algorithm and other wrapper models is the increased computation time required for datasets with many features. Finally, regarding embedded models, this last method was created to combine the advantages of filter and wrapper models. In other words, it concurrently achieves feature selection and model fitting. An example of an embedded method is feature importance on tree algorithms (Loh, 2014). For tree-based algorithms such as Random Forests, Gradient Boosting Machines and XGBoost, feature importance is calculated based on mean decrease impurity (Gini Index). The following formula defines it:

$$Gini = 1 - \sum_{i=1}^{n} (P(n|t))^2 \tag{2.1}$$

The second term, $P(n|t)$ is the proportion of class n observations in node t. The computation of the Gini index is the average of all of the values for each node and each tree. The lower the result, the more deterministic the feature will be. As a result, with a problem with two distinct classes: Churn and Not Churner, a node with 100% observations would have the following result:

$$1 - ((ProbabilityChurnerNodeT)^2 + (ProbabilityNotChurnerNodeT)^2) = 1 - (1)^2 - (0)^2 = 0 \tag{2.2}$$

As a result of a lower value, this feature linked with this node will be perceived as being more deterministic and having more relative importance.

#### 2.4.2.4 Classification Algorithms

There are many options when selecting machine learning algorithms for classification. In this study, we will explore three of those options: Random Forests, Gradient Boosting Machine and Extreme Gradient Boosting.

Firstly, it is appropriate to start with one algorithm already mentioned above on the churn 2.2.2 and feature selection problems 2.4.2.3, Random Forests. Random Forest is an algorithm that is made up of several decision trees, each created on different arbitrary samples of rows and columns, and it outputs the average of all decision tree outputs (Aiello et al., 2016). Besides the previous applications identified in this work, Lin et al. (2017) shows an implementation of an ensemble random forest to high-scale insurance data. This algorithm performs better than state-of-the-art algorithms, such as Support Vector Machine and Logistic Regression.

Moreover, Gradient Boosting Machine is a machine learning technique for regression and classification problems. The learning process is based on building new decision trees correlated with the negative gradient of the loss function (Natekin and Knoll, 2013). The researcher can choose this loss function, being one of the many advantages of this algorithm, its ability to be customized. In its simplest form, gradient boosting assumes that the best possible upcoming model, combined with prior models, minimizes the overall predicted error. Recent studies have shown that GBM has considerably succeeded in many practical applications. According to Vorobyev and Krivitskaya (2022), a study that focuses on detecting fraud detection on bank transactions dataset, Gradient Boosting Machine has achieved better results on different classification metrics such as accuracy, false positive rate, precision and recall than the other two models: Decision Trees and Random Forests.

According to Chen and Guestrin (2016), Extreme Gradient Boosting (XGBoost) is a decision-tree-based algorithm that uses gradient boosting to prevent overfitting by parallel processing, tree-pruning (removing part of trees that do not provide power to classify instances), missing value management and regularization (calibration of the model to minimize loss function and avoid overfitting/underfitting). This study also enhances the recent use of this algorithm to win machine learning competitions. From the 29 challenge-winning solutions published on Kaggle's blog during 2015, 17 solutions used XGBoost. Regarding its applications in business cases, specifically in the financial industry, according to Dhieb et al. (2019), XGBoost ensured better accuracy than Decision Tree Naive Bayes and Nearest Neighbor in detecting different types of fraudulent auto insurance claims. The experimental results in this study also show that XGBoost was the only algorithm that detected different types of fraud in insurance.

#### 2.4.2.5 Performance Evaluation

This chapter will address metrics often used to evaluate and compare machine learning algorithms. First and foremost, standard metrics such as accuracy, precision and recall will be explored. Then, the Lift table of a model will also be addressed, an evaluation metric that can be understood by a

non-technical audience and give a better perception of the model to the stakeholder or audience (Xie et al., 2009).

Firstly, when evaluating a classification model, it is essential to address the confusion matrix. Within the matrix, TP stands for "true positive", TN "true negative", FP "false positive", and FN "false negative". The confusion matrix can be represented by the following:

Table 2.1: Confusion Matrix

| | | Predicted | |
|---|---|---|---|
| | | Positive | Negative |
| Actual | Positive | TP | FN |
| | Negative | FP | TN |

Then, concerning the accuracy of a model, this metric identifies the number of correctly classified instances.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{2.3}$$

Other metrics often used to compare models are False Positive Rate, precision and recall.

Secondly, False Positive Rate outputs which part of data is wrongly classified. It is represented by the equation:

$$FPR = \frac{FP}{(TP+FN)} \tag{2.4}$$

Thirdly, precision returns the percentage of relevant observations regarding the total relevant instances the model retrieved.

$$Precision = \frac{TP}{(TP+FP)} \tag{2.5}$$

Finally, recall outputs the percentage of relevant observations concerning the total relevant instances. It can also be defined as the True Positive Rate.

$$Recall = \frac{TP}{(TP+FN)} \tag{2.6}$$

Another metric that is often used is the F1-score. This metric considers both precision and recall, as it is the harmonic mean of these two metrics. It uses the following formula:

$$\text{F1 score} = 2\frac{Precision \cdot Recall}{(Precision + Recall)} \tag{2.7}$$

Similar to the F1 but giving different weights to Precision and Recall, there are two more metrics: F2 and F0.5. F2 score, unlike the F1 score, which gives equal weight to precision and recall, gives more weight to recall and allows for a more nuanced model evaluation. In other words, the F2 score penalizes a model more for false negatives (predictions of negative when the actual value is positive) than false positives (predictions of positive when the true value is negative). The formula is the following:

$$\text{F2 score} = 5\frac{Precision \cdot Recall}{(4 \cdot Precision + Recall)} \tag{2.8}$$

On the other hand, the F0.5 score gives more weight to precision than to recall. In cases where false positives are considered worse than false negatives, it is crucial to give more weight to precision in evaluating a model. For example, in a use case where the goal is to predict which products will run out of stock, false positives (predictions of stock-out when there is enough inventory) can be costly, as they may result in unnecessary purchases of additional inventory. In this case, it is essential for the model to be precise and only predict stock-outs for products that are likely to run out. By giving more weight to precision, the F0.5 score can help fine-tune the model's performance to prioritize precision over recall. The formula is the following:

$$\text{F0.5 score} = 1.25\frac{Precision \cdot Recall}{(0.25 \cdot Precision + Recall)} \tag{2.9}$$

Moreover, using the metrics like precision and recall, it is possible to calculate the Receiver Operating Characteristic curve, which is a graph that displays how well a classification model performs over all classification thresholds. Its axes are True Positive Rate (y) and False Positive Rate (x). An example of what is the Receiving Operating Characteristic Curve and how it can help when choosing a model is shown in the figure 2.1:

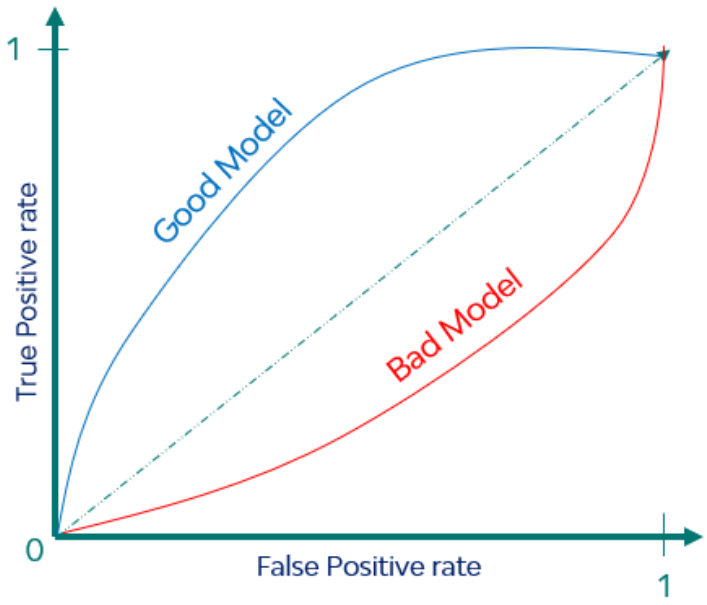



Figure 2.1: Receiving Operating Characteristic Curve

The Area Under the Curve (AUC) is the area under the Receiving Operating Characteristic Curve and measures the performance across all possible classification thresholds. It evaluates how well the model can differentiate True Positives and False Positives. However, it might not be the

most appropriate for an imbalanced dataset once a high number of True Negatives can overshadow the effects of changes in other metrics like False Positives. On the other hand, Area Under the Precision-Recall (AUCPR) Curve has higher sensitivity to True Positives, False Positive and False negatives than AUC. As the name implies, this metric calculates the area under the precision-recall curve, which is the primary distinction between this metric and Area Under the Curve. Area Under the Precision-Recall Curve assesses how well a binary classification model can distinguish between precision-recall pairs or points. These values are obtained using various thresholds on a probabilistic or continuous-output classifier. AUCPR is the average of precision-recall weighted by a threshold's probability. As a result, AUCPR is advised over AUC for severely imbalanced data sets (H2O.ai, 2016).

All these metrics are beneficial when comparing machine learning classification models. However, they might need to be understood by a non-technical audience, which is why, in a business context, it is necessary to search for metrics that can help communicate the model's performance. For instance, in marketing and cross-selling campaigns, the goal is not to predict customer behaviour for every customer but to find a suitable target group where the behaviour has a higher proportion. Therefore, in business applications, searching for models that maximize Lift is preferred over maximizing other parameters such as accuracy or Area Under the Curve (Piatetsky-Shapiro and Masand, 1999). Lift measures the increased accuracy for a particular group. This group can be, for example, the top 1% of a model-score ranked list. Lift of a subset is calculated by dividing the Response Rate of that subset by the overall Average Response Rate. Cumulative Lift takes into account all the subsets until a certain threshold. Lift is calculated by the following formulas:

$$\text{Average Response Rate} = \frac{\text{Number of Targets}}{\text{Size of Dataset}} \tag{2.10}$$

$$\text{Response Rate Subset} = \frac{\text{Number of Targets Subset}}{\text{Size of Subset}} \tag{2.11}$$

$$\text{Lift Subset} = \frac{\text{Response Rate of Subset}}{\text{Average Response Rate}} \tag{2.12}$$

The result of Lift is the number of times the model improves the chances of selecting a customer who belongs to the target group compared to a random selection. We can also assess the response rate and its cumulative since it gives us the percentage of target group samples captured.

One method for evaluating these performance metrics without explicitly splitting the data into training and test sets is K-fold Cross-Validation. K-fold Cross-Validation is a resampling procedure that involves dividing the data into a specified number (k) of folds, training the model on k-1 of the folds, and evaluating it on the remaining one. This process is repeated k times, with each fold serving as the test set in one iteration and the training set in the others. The results from all k iterations are then averaged to provide a more reliable estimate of the model's performance. The data is randomly shuffled and divided into k groups, ensuring that each data point is used for

testing and training in at least one iteration (Fushiki, 2011).

#### 2.4.2.6 Grid Search Hyperparameter Optimization

This chapter addresses the use of grid search optimization techniques in machine learning. Grid search is a method for hyperparameter optimization that seeks to find the best combination of hyperparameter values for a given model (Bergstra and Bengio, 2012). This is achieved by evaluating different combinations of hyperparameters using a performance metric based on n-fold cross-validation. There are two main types of grid search:

- Exhaustive grid search: This type of grid search involves trying every possible combination of hyperparameters to find the best-performing model. This is computationally very expensive, but it is guaranteed to find the best combination of hyperparameters.
- Randomized grid search: This type of grid search involves randomly sampling combinations of hyperparameters and training a model for each combination.

Moreover, in a study by Edwine et al. (2022), the performance of grid search hyperparameter optimization was evaluated in predicting customer churn in the telecom sector. This study showed that using grid search optimization improved the performance of the machine learning models, leading to a superior ability to predict customer churn.

## 2.5 Discussion

This literature review helps create a foundation on state-of-the-art methodologies to approach cross-selling and Churn problems in the insurance sector. The first section facilitates the comprehension of marketing concepts and the impact of analytics in marketing campaigns. The second goes in-depth about customer relationship management specifying cross-selling campaigns such as Next Best Offer and addressing Churn. The work of Desirena et al. (2019) is relevant once it addresses the churn impact in the insurance sector. Regarding the financial industry, the third section explores two sectors: banking and insurance. In the banking sector, two exciting studies deserve to be highlighted. The first one, Knott et al. (2002), where the results show that the Next Product to Buy System has a greater Return on Investment compared to more traditional methods of selecting products and customers for cross-selling. The second one, Xie et al. (2009) trained a machine learning model and obtained a Lift result of around 7. This means the model improves the chances of selecting a churned customer by seven times compared to a random selection. These studies are both interesting once they show the impacts of using machine learning in the financial industry. Finally, the fourth section delves deeper into the methodologies that will be used in this study. It explores machine learning state-of-the-art algorithms and their results on different problems as well as techniques such as feature selection, customer clustering, dealing with imbalanced datasets and comparing model's performance. This part of the review provides knowledge on different strategies and methodologies that can be applied to this problem. A relevant study

explores how XGBoost was better than other algorithms in accuracy, and differentiating types of fraudulent auto insurance claims (Chen and Guestrin, 2016). Moreover, another relevant topic is the use of the Lift metric to evaluate an algorithm, which is an essential tool when explaining the impact of machine learning models on business applications. Piatetsky-Shapiro and Masand (1999) developed an interesting work since it explains how to calculate lift and how to use its value to maximize business profitability.

# Chapter 3

# Problem Description

This project was developed in partnership with LTPLabs, an advanced analytics and business consultancy firm. LTPLabs offers a wide range of services, including operations strategy and planning, supply chain design and optimization, marketing, customer analytics, and more. This Master's thesis focuses on the application of marketing and customer analytics in the context of an Insurance Brokerage. The objective of this research is to develop analytical tools that enable the customization of offers and actions to effectively engage individual clients. In line with a customer-centric vision LTPlabs Website (2022) proposes three lines of action:

1. Gather customer insights by merging data science and business knowledge, enabling action-oriented segmentation.
2. Model customer lifetime value and churn behaviour based on the business context (e.g., contractual vs non-contractual contexts).
3. Optimize each customer's next best action/offer by incorporating previously uncovered insights and focusing on long-term customer value.

This Master's thesis focuses on all modules and proposes to implement systems for the latest two modules. In addition, the following sections will provide an overview of the Insurance Brokerage firm, the market in which it operates, the structure of the project, and the goals of the Insurance Brokerage.

## 3.1 The Company

Insurance companies provide risk management services through insurance contracts, also known as policies, which can cover a wide range of areas, including health, housing, and car insurance. Insurance brokers act as intermediaries between insurance companies and clients, earning a commission on the premium of the policies they facilitate. The premium is the price paid by the client for the insurance coverage. If a client's payments are up to date, they are entitled to make an insurance claim in the event of a covered loss or event. The insurer will review the claim and, if approved, provide the agreed-upon financial protection or compensation to the insured.

The insurance brokerage under consideration in this project is a leading player in the Portuguese market, with a market volume of approximately 50 million euros in terms of its Business to Consumer (B2C) or Business to Business to Consumer (B2B2C) clients. It partners with multiple insurance companies to offer its customers the best options. Currently, the brokerage serves over 190,000 clients and more than 300,000 policies. Its customer base is divided into three segments:

- B2C: The end consumer. These clients are reached through insurance agents or the insurance brokerage's website. Insurance agents are responsible for reaching out to their clients. Their clients cannot be contacted by any other sales representative outside of the agent. These type of customers are not included in the retail section and thus were considered in this project as customers with a close proximity.

- B2B: Companies such as museums, important cultural institutions, board-identified companies, and companies that are created as the extension of an individual.

- B2B2C: Partners that sell the insurance brokerage's products. These partners are generally big players in the retail or telco sector and can reach a vast consumer base, constituting 30% of the insurance brokerage's clients.

In this dissertation, the focus will be on B2C and B2C2B clients. Moreover, in regard its policies and the weight they have on the Insurance Brokerage Revenue, as of the year 2022, its top five classes represent the insurance categories presented in table 3.1:

Table 3.1: Insurance Brokerage Portfolio Structure

| **Insurance** | **% of Total Number of Policies** | **% of Total Commissions** |
|---|---|---|
| Auto | 69% | 67% |
| Multi-risk (Housing) | 9% | 6% |
| Health (Individual) | 7% | 14% |
| Life (Individual) | 5% | 9% |
| Work Accidents | 2% | 1% |

Most of the Insurance Brokerage's business is centred around Auto Insurance due to a partnership with a car dealership with a strong presence in Portugal (B2B2C segment) that has brought many customers. The Insurance Brokerage never acted on maximizing the customer lifetime value of these customers by cross-selling other insurance policies. Some classes of insurance, such as Health Individual and Life Individual Insurance, have noteworthy commissions percentage even though they represent some of the lowest percentages of policies.

## 3.2 The Portuguese Market

The Portuguese insurance market comprises a large number of players, with 589 companies authorized to operate in the country's total addressable market, according to the Autoridade de Supervisão de Seguros e Fundos de Pensões (2022). Of these companies, 525 are able to offer their services to customers in Portugal due to the European Union's freedom of establishment and freedom to provide services, although they do not have a physical presence in the country. The remaining 63 insurance companies have a physical presence in Portugal, with 37 of them being Portuguese companies and 26 being branches of foreign insurance companies. The Insurance Brokerage that is the focus of this dissertation is a branch of a foreign Insurance Broker.

Furthermore, the Portuguese Total Market Volume at the end of 2021 was estimated at 13.3 Billion €, and the market is considered to be divided into two parts: Life and Non-Life. Life insurance is defined as the services that provide compensation after death, and non-Life are all the other categories such as health, work accidents and others. Life insurance accounts for almost 60% of the Portuguese Market Volume. Regarding the top five classes considered earlier for the Insurance Brokerage, their representation in the Portuguese Market is the following:

Table 3.2: Portuguese Market Structure

| Insurance | % of Total Number of Policies | % of Total Premiums |
|---|---|---|
| Auto | 27% | 12% |
| Multi-risk (Housing) | 15% | 4% |
| Health (Individual) | 4% | 8% |
| Life (Individual) | 17% | 38% |
| Work Accidents | 3% | 7% |

Auto insurance is a particularly popular type of policy in Portugal due to the mandatory nature of car insurance in the country. Multi-Risk House insurance, while having a smaller number of policies, is also a significant segment of the market and may represent an opportunity for the insurance brokerage to expand its offerings. Individual life insurance policies may not be as numerous, but they still represent a significant portion of premiums and commission percentages (3.1). This suggests that these policy categories may be worth focusing on for the brokerage. In addition, the relatively low number of policies in these categories may indicate room for growth and increased market share. Overall, the various policy categories present different opportunities and challenges for the Insurance Brokerage to consider as it looks to grow and succeed in the competitive Portuguese market.

## 3.3 Project Structure

This Master's Thesis proposes the development of two systems for the Insurance Brokerage's customer relationship management: Next Best Offer and Churn Prediction. The main objective of these systems is to use machine learning techniques to identify the most suitable customers for

cross-selling and to predict which policies are at risk of churning. The project is divided into two main parts: Next Best Offer and Churn Prediction. The first part focuses on selecting a group of customers with a high likelihood of being interested in a new insurance policy. The second part aims to predict which policies are most likely to churn. Each part is then further divided into two phases: development and evaluation. In the development phase, the goal is to use data science methods to create an analytical model that meets the needs of the insurance brokerage. The evaluation phase aims to test the performance of the models in real-world conditions by contacting the customers that each model recommends. The following sections will delve deeper into the specific goals and requirements of each part of the project.

### 3.3.1 Next Best Offer

The Next Best Offer system aims to identify a subset of customers who do not currently have multi-risk Housing Insurance with the Insurance Brokerage but have a high probability of being interested in purchasing this type of policy. This subset should also have Auto Insurance and include B2C (without possessing an agent) and B2B2C clients. As mentioned earlier, selecting this subset will be divided into two phases: development and evaluation. In the development phase, there are three main steps:

1. Data Collection and Storage
2. Feature Selection and Engineering
3. Model Testing and Comparison

In the development phase of the Next Best Offer system, the first step is to collect all relevant data from the insurance brokerage and organize it in databases. The second step involves determining which information will be most beneficial for the machine learning models to consider. The third step involves testing various machine learning techniques and comparing the results using performance evaluation metrics to determine the most effective approach.

Concerning the evaluation phase, there are two groups of clients:

- A randomly chosen subset of clients will be chosen for a new offer (Control Group)
- A subset of clients with the highest probability of having multi-risk Housing Insurance chosen by the model (Treatment Group)

Once the targeted groups of customers have been identified, the team will provide the insurance brokerage with an interface to register and track the results of contacting these clients. After a period of monitoring and evaluation, the effectiveness of the process will be assessed using metrics such as:

- Percentage of clients who showed interest in both groups
- Percentage of clients who acquired insurance in both groups
- Gross in revenue generated by the contact of both groups

### 3.3.2 Churn

The Churn analysis aims to identify a subset of policies with a high likelihood of churning and to determine which clients they belong to. The subset for this analysis includes clients from the B2C (without possessing an agent) and B2B2C channels and focuses on auto insurance policies. The aim is to work with the insurance brokerage to develop a clear script for addressing these clients in an effort to prevent policy churn. As with the Next Best Offer system, identifying this subset will be divided into two phases: development and evaluation. In the development phase, there are three main steps:

1. Data Collection and Storage
2. Feature Selection and Engineering
3. Model Testing and comparison

The steps involved in the development phase of the churn analysis are similar to those described in the Next Best Offer system. However, the specific features provided to the model may differ. In the evaluation phase, a subset of around 200 policies with the highest probability of churning and a renewal date within two months of delivering the results to the insurance brokerage will be selected for analysis. The behaviour of two groups of policies will then be evaluated on a monthly basis:

- Policies that received outreach efforts (treatment group): These policies will be contacted by the insurance brokerage in an effort to prevent churn. The success of the outreach efforts will be measured by the percentage of policies that ultimately do not churn.
- Policies that did not receive outreach efforts (control group): This group will serve as a control and will allow the team to compare the churn rates of the two groups and determine the impact of the outreach efforts.

The results of the outreach efforts will be monitored and recorded in an interface provided by the team to the Insurance Brokerage. Upon completion of the monitoring and record-keeping, the effectiveness of the process will be evaluated using metrics such as:

- Churn rate: This measures the percentage of policies that cancelled their coverage or did not renew their policies.
- Retention rate: This measures the percentage of policies that the insurance brokerage retained.
- Loss in Revenue in both groups.

Unfortunately, the actual contacts with policyholders and the resulting evaluation of the outreach efforts will not be addressed in this dissertation as the due date precedes the start of the outreach process. However, the development and testing of the analytical models used to identify policies at risk of churning and target the most suitable policyholders for outreach efforts will be thoroughly covered.

## 3.4 Insurance Brokerage Objectives

From the perspective of the Insurance Brokerage, there are several key goals that it hopes to achieve through the development and implementation of the Next Best Offer and Churn Prediction systems. Firstly, the Insurance Brokerage aims to determine the potential gross revenue that can be generated through an analytical approach to cross-selling and predicting policy Churn. This information can inform the company's future approach to using customer data to improve its services. The effectiveness of these efforts will be evaluated using a range of metrics, such as those described in the Next Best Offer and Churn sub-chapters.

In addition to evaluating the financial impact of the project, the Insurance Brokerage also aims to study the overall effectiveness of its customer service teams. This will involve assessing metrics such as the percentage of sales made and the percentage of calls that were blocked or abandoned. These metrics will be analyzed using data collected through an interface provided by the LTP team and recorded by the commercial team during customer outreach efforts. At the end of the project, the LTP team will conduct a more detailed analysis of the outcomes of each call and evaluate the valuable insights that the company can gain from this project.

# Chapter 4

# Proposed Solution

In this chapter, we present the methodological approach that was used in the business cases discussed in this dissertation: Next Best Offer and Churn. These cases involve the use of machine learning techniques to predict customer behaviour and improve business outcomes for an insurance brokerage. The chapter is structured into three main sections: Data Management, Next Best Offer, and Churn. The Data Management section aims to explore the data handling techniques used to prepare the data provided by the Insurance Brokerage for analysis. Then, the Next Best Offer section intends to explore the approach and procedures that ended with a process to contact customers with the most tailored offer. Finally, the Churn section aims to analyze the methodologies used for Churn prediction.

## 4.1 Data Management

This section describes the process of acquiring and storing the Insurance Brokerage's data. The Insurance Brokerage provided the data in the form of three Excel files, which contained information about the company's Policies, Collections, and Insurance Claims. To use this data, it was necessary to transfer it to a secure and reliable storage solution. To this end, an Amazon Redshift server was used to store the raw data. Amazon Redshift is a data warehouse service part of Amazon Web Services, a broader cloud computing platform (Serdar Yegulalp, 2022). This service provides scalable, high-performance storage for large datasets. Once the data was transferred to the Amazon Redshift server, it was ready for analysis and handling. Using the Structured Query Language (SQL), the data was queried and transformed to select the features that would be used in the predictive models. The figure 4.1 summarizes the process.

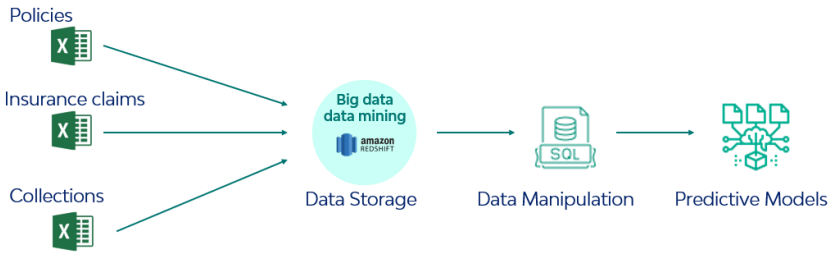



Figure 4.1: Data Management

There are three main steps when focusing on data handling:

- Identify helpful information to use in a predictive model. For example, this may include data on the types of insurance a client holds, their age, gender, and other relevant factors.
- Identify information outside the Insurance Brokerage Data that can be helpful when evaluating a client's potential of churning or adhering to a new offer.
- Identify features to segment or transform the information. The goal here is to reduce the complexity of our data.

Regarding the first goal of data handling, discussions with the stakeholder were held to understand which information could be interesting from the perspective of the experienced Insurance Brokerage's commercial team. These discussions also helped when dealing with inconsistencies in the raw data. The methodology and information differ from the Next Best Offer use case to the Churn, which will be discussed in the respective Sections. Regarding the second goal, demographic information was used on similar projects and was added to this project. This information is obtained from publicly available sources, such as Fundação Francisco Manuel dos Santos (2022) and Portal do Instituto Nacional de Estatística (2022), and was cross-referenced with the location data of each customer provided by the Insurance Brokerage. Including this information allows for a more comprehensive understanding of a client's current circumstances and is a convenient means of obtaining additional details about each client. If any patterns can be found between insurance consumption and social demographics, it can enhance the model's predictive performance. Then, regarding the third goal, this feature handling and selection considered the LTP's experience on other financial projects and some business insights from the Insurance Brokerage stakeholders. These methods differ in the two problems and will be discussed in the methodologies subsections on both these problems.

## 4.2 Next Best Offer

Regarding the Next Best Offer, it was decided to start with a pilot testing where the goal is to find customers more likely to buy Multi-Risk House Insurance. After the data handling explained in the

previous section 4.1, the following Sections will discuss the solution approach for this problem by delving into which information seems relevant to use on a predictive model. Then, the following Section will address the methodology behind building the dataset for model training and testing.

### 4.2.1 Solution Approach

This section outlines the approach to develop a predictive model for calculating the probability of a client acquiring Multi-Risk House Insurance. Firstly, it was discussed which data might be relevant to this problem. The following features were divided into three sections:

- Client Information
- Policy Data
- Socio-Demographic Factors

Regarding the Client's Information, it was essential to have some features related to the clients themselves since there was a business intuition that age and type of car might influence the client's availability towards having more insurance policies. Concerning the Policy Data, it is relevant to address how many policies and their average premium values, among other characteristics. Finally, with the Social-Demographics Factors mentioned in Section 4.1, the goal was to build a social profile of the client according to their address that could be valuable to the predictive model. These variables are indicators from each specific Portuguese region. The following sections will address the selection and building of those variables and how missing values were handled.

#### 4.2.1.1 Feature Engineering and Selection

This section presents a more detailed understanding of the variables created within Clients', Policies' and Social Demographics' Information:

**Client Information**

Age

Gender

Type of Car: Manual Segmentation by the model and brand.

**Policy Data**

Proximity: If the client has his insurance policies managed by an agent or similar within the Insurance Brokerage.

| | |
|---|---|
| Types of Insurance: | Insurance policies the client had with the Insurance Brokerage. This information is divided into variables that take the value from 0 to 1 if the client has had that class of Insurance or not. |
| Customer Lifetime: | This variable is calculated by subtracting today's date with the first policy contract with the Insurance Brokerage. |
| Active Policies | Number of active policies. |
| Inactive Policies | Number of inactive policies. |
| Total Branches | Number of branches which have active policies. |
| Policies' duration | Average duration of active and inactive policies. |
| Total Policies Premium: | The sum of active Policies' premium. |
| Average Premium: | The average a customer pays for his policy. |
| Older Policies: | The total premium paid for policies that have already been terminated. |
| Number of Claims: | Total Claims for all of a client's Policies. |
| Active Policies Ratio: | This variable was calculated by using the following formula: |

$$\text{Average Policy Active Ratio} = \frac{\frac{\text{Active Policies}}{\text{Active Distinct Classes}}}{\frac{\text{Total Policies}}{\text{Total Distinct Classes}}} \tag{4.1}$$

For all these calculations, all the policies of the multi-risk-housing class were excluded since they would create a bias in the predictive model. For instance, the average duration of policies is usually more extensive in Multi-Risk House Insurance, and the model would then use this information to unfairly predict the probability of a client having this type of insurance contract. Finally, regarding the social demographic information, the following variables were created:

**Socio-Demographics Factors**

| | |
|---|---|
| Burned Area: | Burned area in that area |
| House Thefts: | Number of house thefts in a specific area |
| Car thefts: | Number of car thefts in a specific area |
| Cost of the buildings: | Cost of the buildings in that location |
| Emergencies: | Number of emergencies in a specific area |
| Family Dimension: | Average Family Dimension in that location |
| University Education: | Number of people with university education in a specific area |
| Gross Domestic Product: | A measure of the total value of goods and services produced by a specific location |

It is essential to consider that most of these variables are calculated "per capita", meaning they are divided by 100 000 inhabitants.

#### 4.2.1.2 Management of Data Issues

When working with predictive analytics, it is essential to carefully consider the quality of the data being used to train the model. Poor data can result in a less accurate model, leading to incorrect predictions or other errors. Some common issues with datasets include missing values, inconsistent formatting, and irrelevant or duplicate data. In this specific project, among several variables, there was missing information. For instance, customers without postal code information did not have social demographics information. To address these issues, binning and clustering were compared as a means of categorization. Binning is a technique that can be used to handle missing values in a dataset and to group similar values. This involves dividing the data into groups, or bins, based on a set of criteria. For the variable "Age", the data was separated into five groups that made sense from a business perspective:

- Less than 18 years old
- Between 18 and 35 years old
- Between 35 and 65 years old
- More than 65 years old
- No available information

For the other variables such as the socio-demographic variables, the data were separated into equally sized groups based on each variable's quantiles. A group called "no information" for missing values was created. This allowed us to group the data in a way that made it easier to handle missing values and maintain the overall distribution of the data.

#### 4.2.1.3 Clustering

This section addresses the creation of a cluster algorithm to classify socio-demographic groups based on the client's location socio-demographic characteristics. The technique of clustering was utilized only for the socio-demographic variables. This experiment aimed to reduce the dataset's dimensionality while preserving the customers' socio-demographic profiles. The groups of customers were generated using the k-means algorithm from the scikit-learn application programming interface (Buitinck et al., 2013). The elbow method identified the k as being four and the clients were clustered into four different groups. Subsequently, the model's performance was evaluated under two scenarios: one in which individual binning segmentation was applied to each socio-demographic variable and another in which socio-demographic clusters were used.

#### 4.2.1.4 Predictive Model

This Section will introduce the methodology towards obtaining a predictive model capable of predicting the customer's higher probability of buying Multi-Risk House Insurance. As defined in 2.4.2, this classifies as a classification machine learning problem where a discrete class label is predicted. This problem has two distinct classes: clients with Multi-Risk House insurance and clients without Multi-Risk House Insurance. The last one accounts for 93% of the dataset, meaning it is imbalanced. The end goal is to use this model to predict the clients with a higher probability of having Multi-Risk House Insurance from the clients that do not have this type of insurance.

Regarding the application of machine learning, H2o was chosen. H2o is an open-source machine learning platform that provides tools for working with data, building and evaluating machine learning models, and deploying models in production (H2O.ai, 2022). H2o also offers a range of algorithms for tasks such as classification, regression, clustering, and deep learning. It can also be used with various programming languages, including R, Python, and Java, run on a single machine or distributed across a cluster for large-scale data processing.

Concerning the model, the Gradient Boosting Machine (GBM) model was selected after testing a variety of models and finding that it performed the best in preliminary tests. This comparison was performed on an early stage and will be further explained in the results section. It was executed with a function called Automated Machine Learning (AutoML) from the H2O.ai (2022) that provides the user with different machine-learning models for a single dataset.

To select the best suited model among the Gradient Boosting Machines, we employed a grid search optimization procedure. In this case, we utilized the randomized grid search method as it allows for efficient model selection without requiring excessive computation time.

Concerning the parameters to fine-tune the model, the chosen ones as well as their definition according to H2O.ai (2022) is described below:

**Parameters for Hyperparameter Optimization**

Learning Rate: In gradient-boosting models, the learning rate determines the step size at which the model updates its weights during training. A lower learning rate can help the model converge on a good solution, but it may also require more trees (using the number of trees parameter) to achieve the same level of fit as a model with a higher learning rate. This method of adjusting the learning rate and the number of trees can help to prevent overfitting, which occurs when a model is too complex and does not generalize well to new data. Overall, the learning rate is a crucial hyperparameter that can be fine-tuned to improve the performance of a gradient-boosting model.

| | |
|---|---|
| Maximum Depth: | The maximum depth of a tree is the maximum number of levels of nodes that can be present in the tree. Increasing the maximum depth can improve the model's performance but also increase the risk of overfitting. |
| Sampling Rate: | The sampling rate determines the proportion of the training data that is used to train each tree in an ensemble model. A lower sampling rate can reduce the training time, but it may also reduce the model's performance. |
| Number of trees: | The number of trees determines the number of individual models that are trained and combined to make predictions. Increasing the number of trees can improve the model's performance, but increases the training time. |
| Col. sampling rate: | The column sampling rate determines the proportion of the columns (i.e., features) that are used to train each tree in an ensemble model. A lower column sampling rate can help to reduce the training time and prevent overfitting, but it may also reduce the model's performance. |
| Class Sampling Fact.: | Class Sampling Factors is a technique that adjusts the training data distribution to ensure that each class is represented equally. This can help to improve the model's performance on imbalanced datasets. |

In terms of evaluating the models of the randomized grid search, the F1 metric was initially chosen as it is a more comprehensive metric. The F1 metric is advantageous for this purpose because it takes into consideration both precision and recall, providing a more complete overview of the model's performance. However, when doing the randomized grid search, the focus was to maximize how the model predicted customers who are likely to buy the Multi-Risk House Insurance, and it was used the F2 mentioned in section 2.4.2.5. As explained previously, the F2 score is a variant of the F1 score, a commonly used metric for evaluating the performance of classification models. Unlike the F1 score, which gives equal weight to precision and recall, the F2 score gives more weight to recall. It penalizes a model more for false negatives (predictions of negative when the actual value is positive) than false positives (predictions of positive when the actual value is negative). In the case of the Next Best Offer, the goal is to accurately predict the likelihood that a customer will buy Multi-Risk House Insurance. In this context, false negatives (predictions of low likelihood when the customer is interested in buying the insurance) can be particularly harmful, as they may result in missed opportunities to sell insurance to interested customers. By giving more weight to recall in the evaluation of the model, the F2 score can help to fine-tune the model's performance and improve its ability to accurately predict the likelihood of a customer buying Multi-Risk House Insurance. In addition to this evaluation of the randomized grid search models, the best model's performance was also assessed using the F1, Precision, Recall, and Accuracy metrics on a test dataset. To compare the model's ability to predict the presence of Multi-Risk House Insurance clients in relation to a random sample, the lift evaluation metric was

employed in conjunction with k-fold cross-validation using ten folds.

#### 4.2.1.5 Design of Experiments

In this section, the experiments conducted to identify the most suitable model for predicting the clients with a higher probability of buying Multi-Risk House Insurance are detailed. Firstly, the Auto Machine Learning Algorithm was employed to evaluate which type of machine learning model would be best suited for this problem. Subsequently, the cluster of demographic variables was compared to binning segmentation. After realizing this experiment, upon analyzing the plots illustrating the impact of each variable on the models' results, it was determined that some variables contained erroneous information or had been miscalculated. Consequently, the outcomes of this trial were more advantageous than the eventual result. To maintain simplicity, the F1 score of each model was the metric discussed in these early tests. In contrast, the following experiments will present a more comprehensive presentation of the results. Furthermore, another AutoML was executed and evaluated after detecting these erroneous calculations of variables. This also examined whether a tailored random search would be more effective than Auto Machine learning, which produces some models with hyperparameter tuning. Finally, a randomized grid search was conducted to find the optimum model.

#### 4.2.1.6 Contact Registration Interface

The contact registration interface aims to monitor the contacts made by the commercial team. It must have a simple design and be easy to use to save the most time for the commercial team's employees. To be eligible for being contacted, clients must meet the following conditions:

- The clients must be retail clients (B2C without agent and B2B2C) and not have their policies managed by an agent. In this project, these clients are referred to as those without proximity and can be filtered by setting the proximity feature to zero.
- They have to possess at least one policy from the Auto branch.

Moreover, it needs to suit the stakeholders' requirements once they want to measure the results from the Next Best Offer with the metrics mentioned in the Section 3.3.1 as well as evaluate other criteria. The contacts can lead to 3 results:

- Client rejection of proposals.
- Simulation. This refers to a situation where the client expresses interest in receiving an offer. After the call, the sales team will send the client multiple offers from different insurance companies to review and consider. By law, all insurance brokers in Portugal must present their clients at least three offers from three different insurance companies.
- Insurance sales.

Firstly, within the stakeholder's requirements, there are three main validations:

- Insurance Brokerage's Database Evaluation
  - Is the clients' contact information correct?
- Sales People Performance
  - From the clients who answered the call, the percentage of clients that simulated buying another insurance policy.
  - From the clients suggested the percentage of clients that bought the insurance policy. This includes the analysis of the percentage of clients who purchased from the clients who answered the call and the percentage of clients who purchased from the clients who simulated an insurance purchase.
  - What type of Insurance is preferred in the Multi-Risk House branch?
  - Was the customer interested in other insurance products besides the Multi-Risk House?
- Next Best Offer Predictive System Performance
  - How different is the conversion rate for the clients when comparing the treatment and control groups?

Regarding the tool's usability, it was asked to be built using Microsoft Excel, a standard tool known by the commercial team. As a result of the use of such a well-known tool, the interface can provide a comprehensive overview of customer service performance. Additionally, an Excel interface can generate reports and analyze data, which can help track the results from the contacts. An Excel interface also offers flexibility and scalability. It can be customized to meet the needs of the commercial team, and it can be scaled up or down as needed. To build an automated tool with an easy-to-fill form, the language Visual Basic was selected once it would be better to reduce the time it took the commercial team to input the data from the calls. In addition, eight use cases were defined according to the commercial's team director view of the customer journey. These are the following:

1. The client did not answer the call. This use case also includes clients with the wrong number in the Insurance Brokerage's database.
2. The client answered the call but immediately postponed the conversation.
3. The client answered the call but did not have an interest in simulating an insurance purchase.
4. The client answered the call and simulated an insurance purchase and is going to receive other offers from the sales team via email.
5. The client answered the call, simulated an insurance purchase and is postponing a possible purchase buy (this can occur since the client may have this Insurance with another provider and can only terminate that Insurance at a specific point in time).

6. The client answered the call and simulated an insurance purchase and is still deciding whether he will buy or not.
7. The client answered the call and simulated an insurance purchase but was not interested in buying the product.
8. The client answered the call, simulated an insurance purchase and bought the product.
9. The client did not want this Multi-Risk House Insurance but was interested in other products.

   Appendix A shows the two pages of the Interface. Figure A.1 has all the client IDs to contact. It also has a button to display the form to fill out and some graphs that give a brief overview of the results of these contacts. Figure A.2 shows a pivot table that stores the result of the contacts so that the user can check the time, day and result of historical calls. Finally, figure A.3 shows the form to fill. All these pages are in Portuguese, as this work was done for a Portuguese Insurance Brokerage.

## 4.3 Churn

The churn prediction section of this dissertation aims to identify customers at risk of cancelling their service or product. After completing the data handling and preprocessing steps, the following Sections will focus on determining the most relevant information to include in a predictive churn model. The methodology for building the dataset, for model training and testing, will also be addressed in a subsequent Section by explaining the selection and preparation of the data for analysis, ensuring that it is properly formatted and free of any errors or biases. The ultimate goal is to develop a model that can accurately identify customers at risk of churning so that proactive measures can be taken to retain their business and prevent them from cancelling their service.

### 4.3.1 Solution Approach

This section addresses the solution strategy for implementing a predictive model for identifying customers at risk of churning. In order to determine which features are most relevant to this problem, we analyzed the available data and divided it into the same categories used for the Next Best Offer prediction:

- Client Information
- Policy Data
- Socio-Demographic Factors

In order to build the predictive model for identifying customers at risk of churning, a specific methodology was followed for selecting and preparing the dataset. Specifically, it was selected a series of evaluation dates, also referred in the literature as “snapshots” (Gattermann-Itschert and Thonemann, 2021) ranging from 2016 to 2022, occurring every two months. At each of these

evaluation dates, relevant features such as the customer's age and length of time they have been a customer of the company, also known as customer lifetime, were calculated. For example, if we were evaluating whether a customer churned in the subsequent two months after the first of January of 2017, we would calculate their age and lifetime as of that specific period. Figure 4.2 explains with further detail the solution approach:

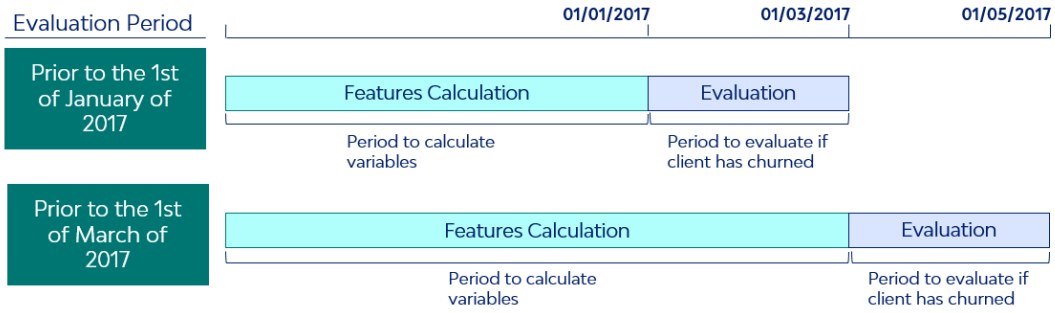



Figure 4.2: Churn Methodology

This allowed us to create a dataset with a consistent set of features for each evaluation date, which we could then use to train and test our predictive model. It is important to point out that the database only sometimes possessed the information on whether customers had churned or cancelled their contracts. The information on contract cancellation is only available in some cases, while in most cases, only the insurers have access to that information. As a result, for this project, churn was determined when payment records for a particular policy ceased, which had specific implications in some cases. It would be beneficial to know when a customer is terminating their contract. The following sections will describe in more detail the process of selecting and building these variables and how missing values were dealt with.

#### 4.3.1.1 Feature Engineering and Selection

This section provides a more in-depth understanding of the variables regarding the Client's and Policy's Information dataset used for churn prediction. The socio-demographic variables, previously discussed in the Next Best Offer section, were also utilized for churn prediction.

**Client Information**

| | |
|---|---|
| Entrance Channel: | Refers to the channel by which a customer purchases his first insurance policy with the Insurance Brokerage. |
| Age: | The age of the customer at that evaluation period. |
| Proximity: | If the client has his insurance policies managed by an agent or similar. |
| Gender | The gender of the client. |
| With agent: | If the client has an agent, it is different from the proximity variable in some cases. |

| | |
|---|---|
| OP: | Other Policies. The number of other policies the client has in that "snapshot". |
| OPSBP: | Other Policies Same Branch Group. The number of policies within the same branch group of the Policy we are evaluating in that "snapshot". |
| OPOBP: | Other Polcies Other Branch Group. The number of policies within the different branch groups of the Policy we are evaluating in that "snapshot". |
| Other Premiums: | Premiums of Policies of the different branch groups. |
| Total Premiums: | Current premium of the evaluated policy plus the Other Premiums. |
| PC 12 months: | Policies Churned in the last 12 months. Number of other Policies that the client has churned in the previous 12 months considering the evaluation date. |
| NP 12 months: | New Policies in the last 12 months. Number of new Policies that the client added to his portfolio in the previous 12 months considering the evaluation date. |
| Past Claims: | Number of claims previous to that evaluation date. |
| Claims 12 months: | Number of claims previous in the 12 months prior to the evaluation date. |
| NPS: | Next Promotor Score. It is a measure of customer loyalty and satisfaction. It is calculated by asking customers to rate the likelihood of recommending a company's product or service to a friend or colleague on a scale of 0-10. Customers who respond with a score of 9 or 10 are considered "promoters" of the company, while those who respond with a score of 0-6 are considered "detractors". |
| PCVA: | Past Compensations per Value asked. After an occurrence of a claim where the insurance will pay the client a specific value. Depending on the insurance evaluation, this value may be less or more than what the customer expects. On this variable, we are dividing the past compensations by the value asked by the customer. |
| PCAV 12 months: | On this variable, we are dividing the past compensations by the value requested by the customer in the 12 months before the evaluation date. |

**Policy Data**

| | |
|---|---|
| Policy Channel: | Refers to the channel by which a customer purchases an insurance policy. |
| Distinct Channel: | Indicates whether the Policy was purchased through a different channel than the customer's first Policy. |

| | |
|---|---|
| Renewal Month: | Indicates when the first premium payment was made and was considered by the stakeholders the month of that Policy renewal. This variable takes the value one if the renewal month is within two months after the evaluation date of that data point. |
| Branch Group: | Broad categorization of the Policy. |
| Branch: | A more specific categorization. Within each Branch Group, there are multiple Branches. |
| Product: | The most specific categorization. Within each Branch, there are multiple Products. |
| Policy Duration: | The length of the Policy. |
| Means of payment: | Method used to pay for the Policy. |
| Payments per year: | Number of payments per year. |
| Current Premium: | Current premium paid for this Policy. |
| Previous Premium: | Previous Premium paid for a policy in the same branch group. |

It is essential to base predictions about future churn on information known at the time of the evaluation. Using information that was not known at the time of the assessment would not be fair, as it would not accurately reflect the circumstances and conditions at the time. By considering only information known at the time of the evaluation, we can create a model that more accurately predicts future churn and better understand the factors that may influence churn. This approach allows us to make informed decisions and help the Insurance Brokerage take appropriate actions to mitigate churn and retain customers.

#### 4.3.1.2 Data Handling and Customer Profiling

Techniques such as segmentation binning and demographic clustering were implemented to address data management issues, as in the Next Best Offer part. Segmentation binning was applied to most of the numeric variables, including the socio-demographic, while the cluster was used to group customers according to their socio-demographic characteristics. A test to assess the effectiveness of using socio-demographic segmentation binning or cluster was subsequently conducted.

#### 4.3.1.3 Predictive Model

This Section will introduce the methodology towards obtaining a predictive model capable of predicting the customer's higher probability of churning. As defined in the Next Best Offer part, this is an example of a classification machine learning problem where a discrete class label is predicted. This problem has two distinct classes: policies churned and policies not churned. The last one accounts for 86,5% of the dataset, meaning it is imbalanced. The end goal is to use this model to predict the policies with a higher probability of churning in the following two months.

Regarding the application of machine learning, as it was in the Next Best Offer, H2O.ai (2022) was chosen.

Concerning the model, the Gradient Boosting Machine (GBM) model was selected after testing a variety of models and finding that it performed the best in preliminary tests. This comparison was performed early and will be further explained in the results section. It was executed with a function called Automated Machine Learning (AutoML) from the H2O.ai (2022) that provides the user with different machine learning models for a single dataset. The feature selection method of Boruta was also applied and the results suggested that certain demographic variables should be removed from the first dataset. Since after the preliminary tests, it was decided to follow the approach of having a socio-demographic cluster (second dataset), the results of the feature selection analysis did not significantly influence the construction of the final dataset.

Moreover, a grid search optimization procedure was employed to select the best suited model among the Gradient Boosting Machine models. In this case, we utilized the randomized grid search method as it allows for efficient model selection without requiring excessive computation time. Concerning the parameters to fine-tune the model, the chosen ones were the same ones used in the Next Best Offer. In the context of customer churn prediction, false positive predictions (i.e., predicting that a customer will churn when they do not) can be costly for a business since negotiating a contract with a customer who is not likely to Churn can lead to revenue losses. Therefore, it is essential to use a performance metric sensitive to changes in the false positive rate. F0.5 is a valuable metric in this case because it emphasizes precision more, which measures the proportion of true positive predictions made by the model out of all positive predictions. A high precision score indicates that the model makes relatively few false positive predictions, which is beneficial for minimizing the cost of customer churn. Therefore, the F0.5 score was chosen to order the models from the grid search in Churn prediction.

After all relevant variables were identified and a timeline was established for predicting policy churn, the best model was trained using the maximum amount of available data, excluding the data points for policies being evaluated for potential churn within the defined timeline. To assess the model's performance thoroughly, we employed k-fold cross-validation with k=10. To allow the model to understand the factors contributing to churn, we grouped each client into a separate group. We trained the model on these groups, allowing it to learn the characteristics of each policy and client that may have led to churn. This approach allowed us to predict policy churn and understand the underlying causes effectively.

#### 4.3.1.4 Design of Experiments

In this section, the experiments performed to identify the most suitable model for predicting customers with a higher probability of churning are detailed. Firstly, an Auto Machine Learning Algorithm was employed to evaluate which type of machine learning model would be best suited for this issue. Subsequently, the cluster of demographic variables was compared to binning segmentation. After realizing this experiment, conversations with the stakeholders resulted in the

decision only to consider policies in the renewal period in the following two months (variable renewal month equal to one). For this reason, the dataset was modified only to consider these data points. This increased the performance of our model once it eliminated variation in the renewal month variable and made the model more adequate to the use case, which was to predict churn for customers who are renewing their policy. Following this test, a meeting with the stakeholders was held to demonstrate the results and evaluate certain matters that necessitated clarification. These matters included policies for which a single payment was made or policies for which negative payments (deductions) had been incurred over the last 12 months. It was found that some policies with only one payment were cancelled prior to the renewal period. This information was then used to remove numerous policies from the model training and testing. Following the removals, a test was conducted by training the model until May 1, 2022 (including this "snapshot") and testing it on the two subsequent renewal dates: July 1, 2022, and September 1, 2022. Moreover, within these two datasets, several clients received payments from the Insurance Brokerage (signalled as a negative payment in the collections data). This indicated that these clients had terminated their contracts in the middle of the year and were therefore issued a deduction payment. These clients were excluded from the final test. The reason for not removing these clients from the training data was that they had churned; however, from the standpoint of the Insurance Brokerage, it did not make sense to contact them after they had already cancelled their agreement.

In summary, these were the experiments conducted to identify the most suitable model for predicting customers with a higher probability of churning:

- Utilizing an Auto Machine Learning Algorithm to evaluate the best suited machine learning model.

- Comparing a cluster of demographic variables to binning segmentation.

- Testing the model obtained by AutoML with the new test dataset of only considering policies in the renewal period. This test compared a model trained only on policies in the renewal period and another model trained on both policies that were on the renewal period and on policies that were not in the renewal period.

- Compare the last model to a randomized grid search.

- Test the best model on two specific evaluation dates that we know whether the client has churned and evaluate its performance.

Finally, the best model was then trained on all the available historical data and only tested on the data of the evaluation month we do not have information about whether the client has churned or not (use case that we want to predict). The results of these experiments will be detailed in Chapter 5.

#### 4.3.1.5 Contact Registration Interface

The contact registration interface is designed to monitor the contacts made by the sales team. To be eligible, clients must meet the following conditions:

- The clients must be retail clients (B2C and B2B2C) and not have their policies managed by an agent. In this project, these clients are referred to as those without proximity and can be filtered by setting the proximity feature to zero.
- Their policy with a high probability of churn must be from the Auto branch.
- The clients must be in their annual renovation period, which is the month of their first payment (the renewal month variable must equal one).

As it was the primary goal for the Next Best Offer, the contact registration interface must have a simple design and be easy to use. Since the Next Best Offer sales team remained the same for the Churn pilot test, their feedback was also considered in the design of the new contact registration interface, allowing for a solution tailored to the specific needs of the Insurance Brokerage. Moreover, the script was amended to ensure that customers are not made aware of any available superior services to what they currently have. Firstly, to ensure that the customer is satisfied with their service, it is essential to track two key factors: whether the customer is renewing their existing policy or if he has any complaints or is unsatisfied with the current service. It is also essential to identify the cause of the customer's complaints.

Additionally, the commercial team requested that the registration process should be more open-ended. The churn contact could transition to conversations about various insurance policies, and the customer may be dissatisfied with multiple services. Due to this, there was no need to define use cases, similar to what was done for the Next Best Offer part for the possible contact results. Furthermore, the calls can have four immediate outcomes:

- The customer may be unhappy with the policy suggested by the model as likely to churn.
- The client may be unhappy with other policies
- The client may be unhappy with both: the policy suggested by the model as likely to churn and other policies.
- The customer may be satisfied with their current portfolio.

Additionally, if the customer is unsatisfied with any of the services from the insurance brokerage, the call can lead to one of three possible outcomes:

- The customer may be interested in simulating an insurance purchase to substitute his current policy, which he is unsatisfied with.
- The client may reject any negotiations or simulation of a new insurance policy.
- The client may purchase an alternative service or accept a negotiation.

Finally, some procedures from the Next Best Offer version were retained, such as giving customers the option to enter a date for a follow-up conversation about the topic and tracking the number of calls made to that customer, which was well-received by the stakeholders. The appendix B shows the two pages of the Interface. Figure B.1 has all the client IDs and respective policies IDs to contact. It also has a button to display the form to fill out and some graphs that give a brief overview of the results of these contacts. Figure B.2 shows a pivot table that stores the result of the contacts so that the user can check the time, day and result of historical calls. Finally, figure B.3 shows the form to fill. All these pages are in Portuguese, as this work was done for a Portuguese Insurance Brokerage.

# Chapter 5

# Results

This chapter introduces the results of our project of the Next Best Offer and the churning system developed in this dissertation. It is divided into two sections: the Next Best Offer prediction and the Churn prediction. Both sections will present the results from our machine learning models and the results from the Next Best Offer follow-up contacts with customers.

## 5.1 Next Best Offer

This section will first present a preliminary analysis of the results from several machine learning models explored for the Next Best Offer prediction. It will then delve into the theoretical results of our chosen model, including F1 scores and other metrics. Finally, it will discuss the results from the insurance brokerage's follow-up contacts with customers and how they compare to those obtained from our machine learning model.

### 5.1.1 Model Selection

This section discusses the initial tests performed to obtain the best suited model and assess whether a socio-demographic cluster is better than binning segmentation in this use case. Section 5.1.1.1 will describe the preliminary tests, Section 5.1.1.2 will address the cluster versus binning segmentation comparison, and Section 5.1.1.3 will present the final results from our tests including the discovery and correction of biases due to miscalculated variables and the use of erroneous information.

#### 5.1.1.1 Preliminary Tests

In the field of machine learning, there is a wide variety of models available. H2o provides a tool that automates the process of model selection by training different types of models and then ordering it on the cross-validation performance results. This tool is called Automated Machine Learning (AutoML). This tool already uses some forms of hyperparameter tuning, but it does not

provide features to customize it. To obtain the best-suited model, 30 different models from these types were tested:

- Deep Learning Neural Networks
- Distributed Random Forest (DRF)
- Gradient Boosting Machine (GBM)
- Stacked Ensemble

Although this tool enables the selection of specific models to be excluded or included in the test, it does not permit the specification of the number of attempts for each model. Additionally, the metric used to order the models was the F1 score. The test ranked in the top ten, eight variations of the Gradient Boosting Machine and between them, in the seventh place, a Stacked Ensemble model:

Table 5.1: Preliminary Tests

| **Model ID** | **F1 Score** |
|---|---|
| GBM5 | 0.509 |
| GBM2 | 0.507 |
| GBM3 | 0.506 |
| GBMgrid1model3 | 0.503 |
| GBM1 | 0.498 |
| GBM4 | 0.498 |
| StackedEnsemble1 | 0.496 |
| GBM6 | 0.495 |
| DeepLearning1 | 0.493 |
| GBMgrid1model2 | 0.490 |

The numerical values represent different tests, with the inclusion of the grid in the model's name indicating that hyperparameter tuning has been employed by the tool. With these results, the Gradient Boosting Machine was chosen. The F1 score of the first model was 0.51. Finally, it is essential to note that, in this test, the socio-demographic information was used with the binning segmentation technique (as it was the initial approach decided by the team towards these variables). The following Section will compare these results to those obtained when employing a socio-demographic cluster.

#### 5.1.1.2 Socio-Demographic Analysis

Regarding comparing the socio-demographic binning segmentation and the socio-demographic cluster, it was essential to maintain the same conditions for both solutions. As a result, the datasets used for training and testing were the same for both models. Ultimately, the socio-demographic

binning segmentation yielded a F1 score of 0.509, while the clustering algorithm had a F1 score of 0.493. Consequently, it was decided to proceed with the use of the socio-demographic binning segmentation.

#### 5.1.1.3 Final Model Comparison

After determining the approach to follow with the socio-demographic segmentation, a more thorough examination of the model results was conducted. During this process, the team noticed that some variables needed to be revised or had information the model should not possess while calculating the probability of a client having Multi-Risk House Insurance. Consequently, the dataset and the values of some variables were altered. Moreover, another Automated Machine Learning algorithm was executed to obtain a model to compare to our future hyperparameter optimization using a randomized grid search. The results from this Automated Machine Learning Algorithm showed that the best seven of the thirty models were Gradient Boosting Machines with different hyper-parameters optimization selections. The best F1 score was 0.381, and the second-best type appeared in the eighth place, with an F1 of 0.367, the Stacked Ensemble. This led to the conclusion that, once again, the best model for this problem was the Gradient Boosting Machine. Regarding the best model, the table 5.2 provides a more detailed view of its performance results:

Table 5.2: Auto ML Next Best Offer Best Model

| **Perfomance Metrics** | **% Result** |
|---|---|
| F1 | 0.381 |
| F2 | 0.353 |
| Precision | 0.439 |
| Recall | 0.337 |
| Accuracy | 0.918 |

### 5.1.2 Model Fine Tunning

Model evaluation involves assessing the performance of a machine learning model and identifying areas for improvement. In this study, we selected the Gradient Boosting Machine (GBM) as our model of choice. To optimize the hyperparameters of the GBM, we conducted a random grid search based on the F2 metric. The F2 metric is a measure of the balance between precision and recall and is often used when correctly identifying the positive cases in a dataset is a priority. This is particularly relevant in a Next Best Offer predictive system, where the goal is to identify the offers that are most likely to be accepted by a customer. Accurately identifying these positive cases is crucial for the system's success, as it allows the company to target their marketing efforts and improve customer satisfaction. After performing this evaluation, we obtained the following results for the best model on the test dataset:

Table 5.3: Randomized Grid Search Next Best Offer Best Model

| **Perfomance Metrics** | **% Result** |
|---|---|
| F1 | 0.405 |
| F2 | 0.493 |
| Precision | 0.312 |
| Recall | 0.576 |
| Accuracy | 0.882 |

These results show that the hyperparameter tuning process improved several evaluation metrics compared to the previous model in Section 5.1.1.3. Specifically, the optimized GBM model had a higher F2 score and a higher recall rate, suggesting that it was better at correctly identifying the positive cases in the dataset (i.e., customers with Multi-Risk Housing Insurance). Additionally, while the accuracy of the optimized GBM model was slightly lower than that of the previous model, this decrease in accuracy may be acceptable in light of the significant improvements in the F2 score and recall rate. In general, a trade-off between precision and recall is often necessary for predictive modelling. The F2 score provides a way to balance these two metrics and identify the model that offers the best overall performance for a given problem. In this case, the optimized GBM model achieved a higher F2 score and recall rate, making it a more suitable choice for our Next Best Offer predictive system.

Moreover, a thorough analysis of the results from the optimized GBM identified the following variables as being the most impactful on the model's output.

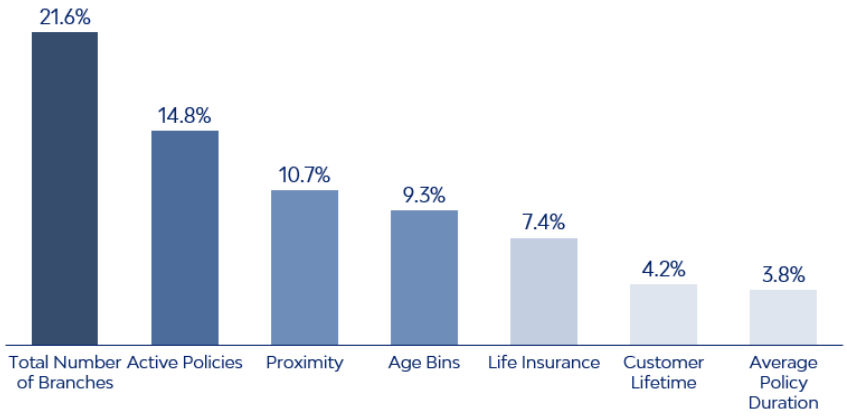



Figure 5.1: Variable Importance Next Best Offer

The results suggest that the total number of branches a client possesses with the Insurance Brokerage is the most critical factor in predicting the probability of a client having Multi-Risk Housing Insurance. The sales team corroborated this after the initial contacts were conducted.

They elucidated that customers more amenable to running simulations of Multi-Risk House Insurance tended to be those with more products and were more familiar with the company.

Moreover, the Shap Summary is also shown in figure 5.2. SHAP (Shapley Additive exPlanations) is a machine learning technique that is used to explain the output of any machine learning model. It works by determining the contribution of each feature to the final prediction. SHAP assigns a score to each feature based on the importance of that particular feature in making the prediction. This score is calculated using a game theory concept called Shapley Value, which determines each player's contribution in a game (Lubo-Robles et al., 2020). Using SHAP, machine learning models can be explained more intuitively, and humans can better understand their decisions. SHAP also provides a way to measure the importance of features and can be used to identify features that are most relevant for making predictions. As one of the objectives of this project was to provide transparent communication to stakeholders regarding the model's functioning, the variables' names were chosen to be in Portuguese.

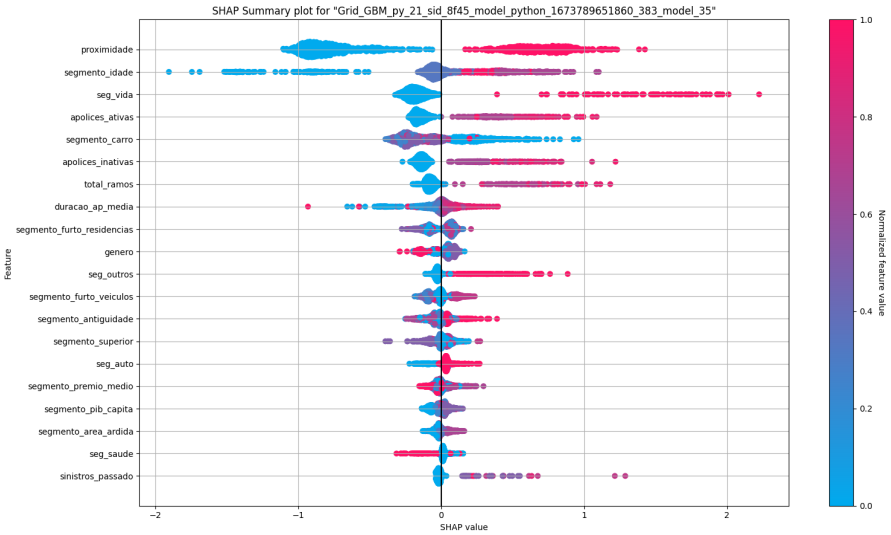
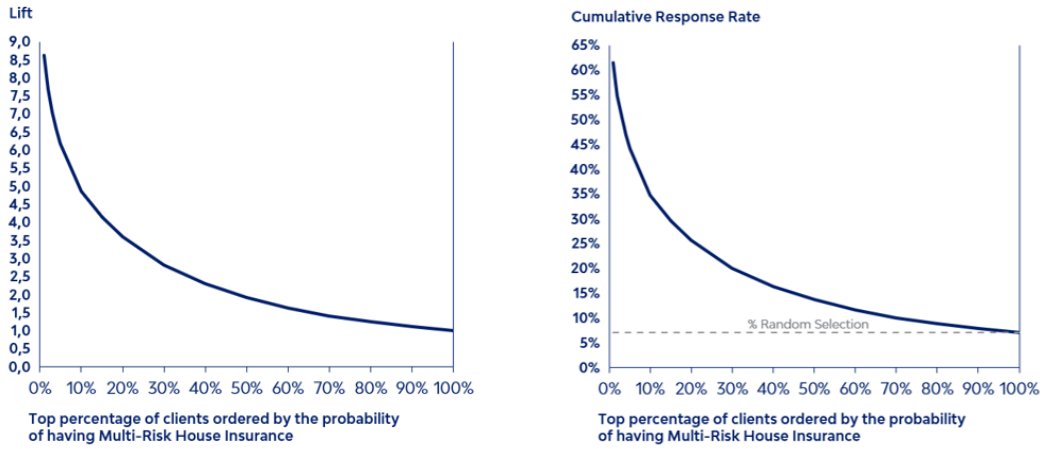



Figure 5.2: Shap Summary Next Best Offer

The SHAP summary implies that customers with more products are more likely to purchase Multi-Risk House Insurance. Furthermore, the number of active policies and the customer's proximity are significant determinants of this likelihood. Specifically, if a customer has their policies managed by an agent (proximity equals to one), they are more likely to own this type of insurance.

The results of our evaluation using cross-validation indicate that the model performs exceptionally well in terms of lift. This top 4% exhibits a lift of 6.6, indicating that the model is 6.6 times better at identifying clients who possess Multi-Risk House Insurance than a random selection process. Additionally, within this top 4%, the results show a cumulative response rate of 46%, which means that selecting the top 4% of the dataset ordered by the probability given by the model, 46% of clients have Multi-Risk House Insurance. The entire dataset has an average response rate of 7%. Figure 5.3 presents these results:




(a) Lift Results

(b) Cumulative Response Rate

Figure 5.3: Cross-Validation results ordered by the probability of having Multi-Risk House Insurance

### 5.1.3 Contacts Results

At the time of submission, the contacts had been ongoing for approximately one month. Up to this point, only the overall results had been assessed. Figure 5.4 presents the states of the contacts:


Figure 5.4: Monitorization of calls

As can be seen, around 60% of the 200 clients answered the call, while only 5% had not yet been contacted at the time of this dissertation. Of the 127 clients that answered the call, the results of the calls are illustrated in Figure 5.5:

The treatment group, which was predicted to have a higher probability of having Multi-Risk House Insurance, already has two purchases and 35 clients who have simulated an insurance purchase and will give their confirmation in subsequent calls. Notably, one of the clients who bought

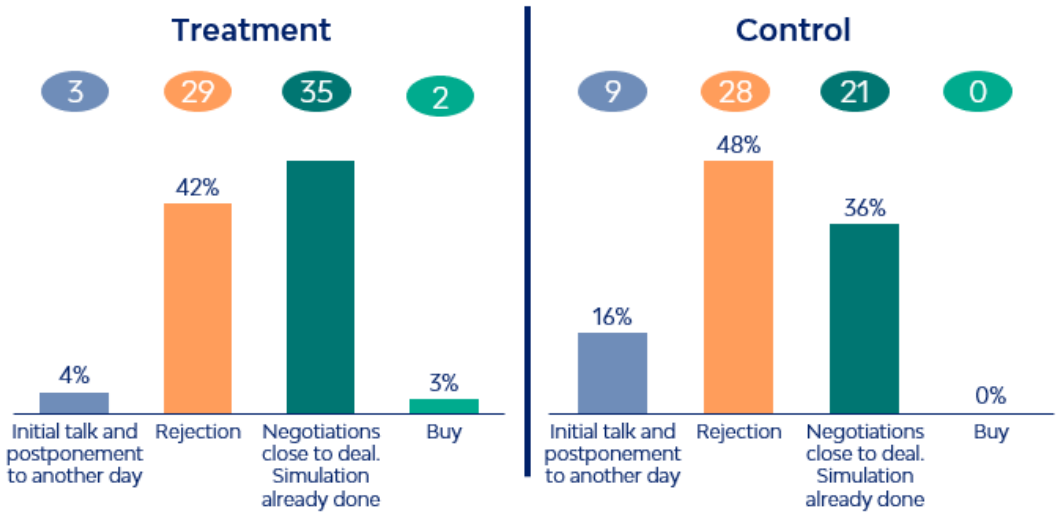



Figure 5.5: Monitorization of calls in each group

insurance purchased three products (two Multi-Risk Housing Insurance and one Health insurance). Nevertheless, since the primary goal is to ascertain the number of clients who bought Multi-Risk House Insurance, it is counted as only one purchase. Conversely, the control group, which was randomly selected, has also had some potential purchases; however, at the time of writing, there have yet to be any successful cases. It is vital to remember that not all clients who simulate an insurance purchase and consider its terms and conditions will positively respond to the negotiation.

## 5.2 Churn

In this section, we will analyze the results of various machine learning models that were explored for predicting customer Churn. We will present a preliminary analysis of the performance of these models and then delve into the results of our chosen model in more detail, including its F1 score and other evaluation metrics. Unfortunately, the follow-up contacts with customers to verify Churn outcomes will start after this dissertation's due date. As such, this dissertation will not have the actual Churn results to compare with our machine learning model's predictions in this part of the study. However, it will still be possible to provide an in-depth analysis of the model's performance using evaluation metrics such as the F1 and F0.5 scores and discuss the potential value of the model in predicting Churn and identifying at-risk customers.

### 5.2.1 Model Selection

This section discusses the initial tests conducted to obtain the best suited model and assess whether a socio-demographic cluster is better than binning segmentation in this use case. The Section 5.2.1.1 will describe the preliminary tests, Section 5.2.1.2 will address the cluster versus binning

segmentation comparison, and Section 5.2.1.3 will present the final results after only considering policies in their annual renewable period.

#### 5.2.1.1 Preliminary Tests

Regarding the initial tests, an experiment was conducted to identify the best suited model for this problem. The methodology used was the same as that employed for the Next Best Offer, which utilized an Automated Machine Learning algorithm and evaluated the results. The Automated Machine Learning algorithm experimented with four different types of models, testing thirty different variations and adjusting their hyperparameters:

- Deep Learning Neural Networks
- Distributed Random Forest (DRF)
- Gradient Boosting Machine (GBM)
- Stacked Ensemble

The results showed that the Gradient Boosting Machine was the top performer having nine variations of this model in the top ten performers:

Table 5.4: Preliminary Tests

| **Model ID** | **F1 Score** |
|---|---|
| GBM3 | 0.373 |
| GBMgrid1model2 | 0.372 |
| GBM2 | 0.371 |
| GBMgrid1model1 | 0.370 |
| GBM1 | 0.369 |
| GBMgrid1model3 | 0.369 |
| GBM5 | 0.367 |
| GBM4 | 0.366 |
| StackedEnsemble2 | 0.366 |
| GBMgrid1model4 | 0.364 |

The F1 score of the best model was 0.373. It is essential to note that, in this test, the socio-demographic information was used with the binning segmentation technique (as it was the initial approach decided by the team towards these variables). The following Section will compare these results to those obtained when employing a socio-demographic cluster.

#### 5.2.1.2 Socio-Demographic Analysis

With regards to the comparison of the socio-demographic binning segmentation versus the socio-demographic cluster, the methodology utilized was identical to that employed for the Next Best

Offer. The datasets used for training and testing were the same. Ultimately, the socio-demographic binning segmentation yielded an F1 score of 0.373, while the clustering algorithm had an F1 score of 0.381. Consequently, it was decided to proceed with the use of the socio-demographic cluster.

#### 5.2.1.3 Renewal Month Analysis

Regarding the final model comparison test, an experiment was conducted between two methods for predicting customer churn. The first model was trained on policies in their renewal period and those that were not, while the second model was trained on only policies in their renewal period. The two models were then tested with the same dataset consisting only of policies in their renewal period. The results showed an F1 score of 0.427 for the model trained with policies in the two states. This increase in performance is likely since the renewal period was already an essential variable in Churn prediction. When only testing the model in this use case, the model experienced a considerable increase in performance. Additionally, the model only trained on policies in their renewable period had an F1 score of 0.439. This led to the conclusion that using a model only trained on renewing policies at each evaluation period was the best approach. Regarding the best model, the table 5.5 provides a more detailed view of its performance results.

Table 5.5: Auto ML Churn Renewal Month Analysis Results

| **Perfomance Metrics** | **% Result** |
|---|---|
| F1 | 0.439 |
| F0.5 | 0.433 |
| Precision | 0.429 |
| Recall | 0.449 |
| Accuracy | 0.829 |

### 5.2.2 Model Fine Tunning

In machine learning, model evaluation is a crucial step in assessing the performance of a model and identifying areas for improvement. In this study, we selected the Gradient Boosting Machine (GBM) as our model for predicting Churn. To optimize the hyperparameters of the GBM, we employed a random grid search method, utilizing the F0.5 metric as the evaluation criterion. The F0.5 score, a harmonic mean of precision and recall that assigns a higher weight to precision, is a suitable metric for evaluating models in this context. This is due to the stakeholders' goal of minimizing false positives to avoid the unnecessary bother of clients who are not churning and to prevent the formation of thoughts of looking for alternatives. The F0.5 score, by giving more weight to precision, effectively penalizes the model more for false positives than for false negatives, which aligns with the desired outcome in this scenario. The results are described in table 5.6:

Table 5.6: Randomized Grid Search Churn Best Model

| **Perfomance Metrics** | **% Result** |
|---|---|
| F1 | 0.431 |
| F0.5 | 0.416 |
| Precision | 0.407 |
| Recall | 0.459 |
| Accuracy | 0.818 |

The hyperparameter tuning process revealed a decrease in all metrics apart from recall. Since this hyperparameter tuning aimed to obtain a model with improved precision and an F0.5 score, it was deemed unsuccessful. Thus, the model derived from the AutoML experiment was used to predict policies with a higher risk of churn. Moreover, a thorough analysis of the results from the AutoML GBM identified the following variables as being the most impactful on the Churn model's output:

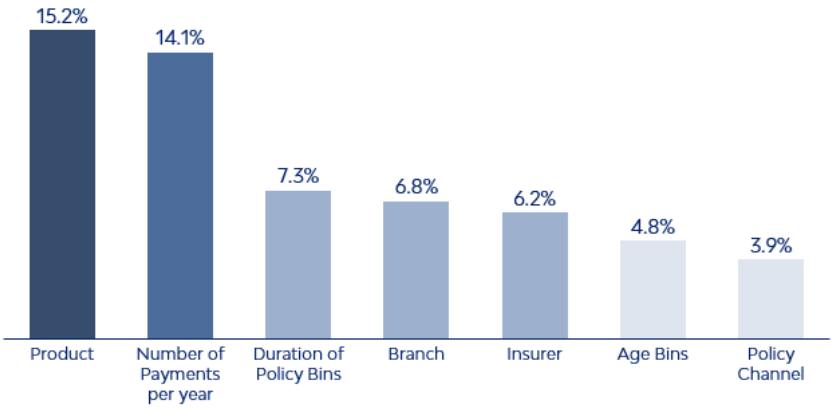



Figure 5.6: Variable Importance Churn

The results suggest that the type of product associated with the policy and the number of payments per year were the most influential variables in determining Churn. This result was expected, as some products had higher churn rates than others. Additionally, the Shap summary provided insight into the values of the number of payments that were associated with higher Churn probabilities. These results are shown in figure 5.7. As one of the objectives of this project was to provide transparent communication to stakeholders regarding the model's functioning, the variables' names were chosen to be in Portuguese.

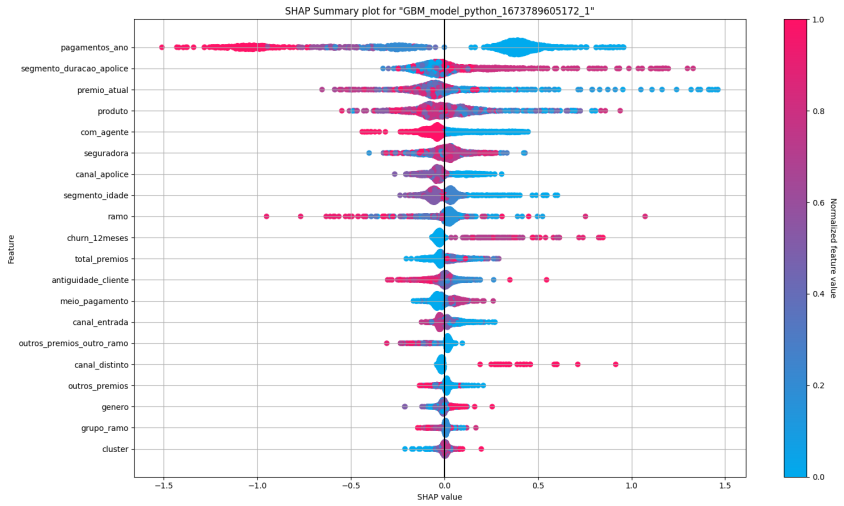
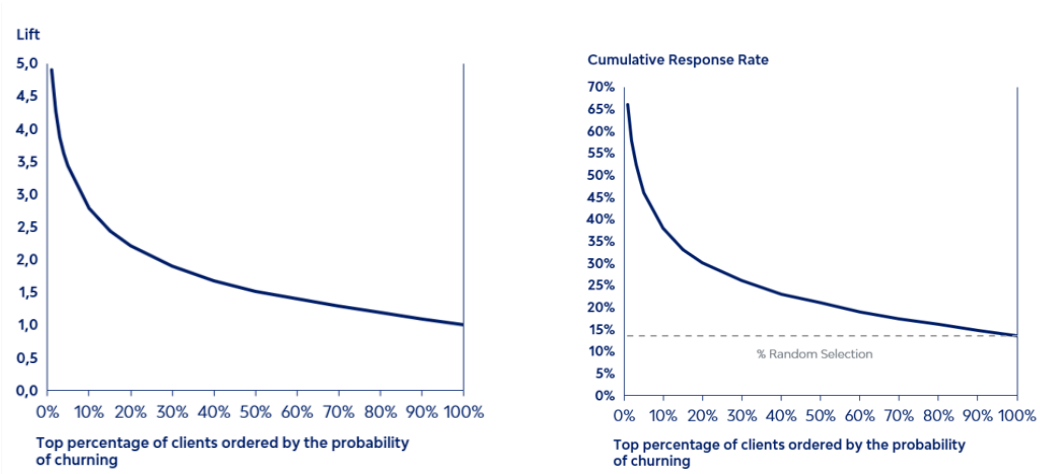



Figure 5.7: Shap Summary Churn

The SHAP analysis revealed that policies with a more frequent payment structure were less likely to Churn, and there is a premium and duration of policy segment where clients are more likely to Churn. Additionally, policies from clients who had an agent were found to be less prone to Churn. An intriguing result derived from the trend analysis of the client was the SHAP analysis of the variable: Policies Churned in the last 12 months, which showed that policies from clients with no Policies Churned in the last 12 months were less likely to Churn.

Moreover, the results of our evaluation using cross-validation indicate that the model performs admirably in the lift. This top 4% exhibits a lift of 3.62, indicating that the model is 3.62 times better at identifying churners. Additionally, within this top 4%, the results show a cumulative response rate of 49%, which means that selecting the top 4% of the dataset ordered by the probability given by the model, 49% of policies have churned. The entire dataset has an average response rate of 13,5%. Figure 5.8 presents these results.




(a) Lift Results (b) Cumulative Response Rate

Figure 5.8: Cross-Validation results ordered by the probability of churning

### 5.2.3 Evaluating Real Case Performance

In this Section, the aim is to assess the actual performance of the best suited model when tested on a real test, similar to what will occur when predicting the policies with a greater likelihood of churning and providing it to the Insurance Brokerage for contacting. As previously noted in the prior Sections, the Churn contacts will only begin after the submission of this dissertation. Nevertheless, to evaluate the effectiveness of our model, a real test was conducted wherein the model was trained until a specific evaluation date and then tested on two successive dates. Before this test, the dataset underwent some alterations, which are elucidated in Section 4.3.1.4. These changes led to a decrease in the number of churners to be taken into account on each renewal date. The periods considered for this test were July 2022 and September 2022. The average number of policies that were being evaluated was 7000, and the average number of policies that churned in both those periods was 1004 (approximately 14%). Finally, the best model generated the following results in July 2022 and September 2022 evaluation dates:

Table 5.7: Real Case Test Churn

| **Top number of policies sorted by the probability of churning** | **Actual churners July 2022** | **Actual Churners September 2022** |
|---|---|---|
| Top 50 | 16 | 22 |
| Top 100 | 36 | 36 |
| Top 150 | 50 | 49 |
| Top 200 | 62 | 62 |
| Top 250 | 79 | 76 |
| Top 500 | 152 | 151 |

The results indicate that in the top 50 policies sorted by their probability of churning as outputted by the model, 16 churned in July 2022, and 22 churned in September 2022. Taking into account the average of the results of these two evaluation dates, the following results were obtained:

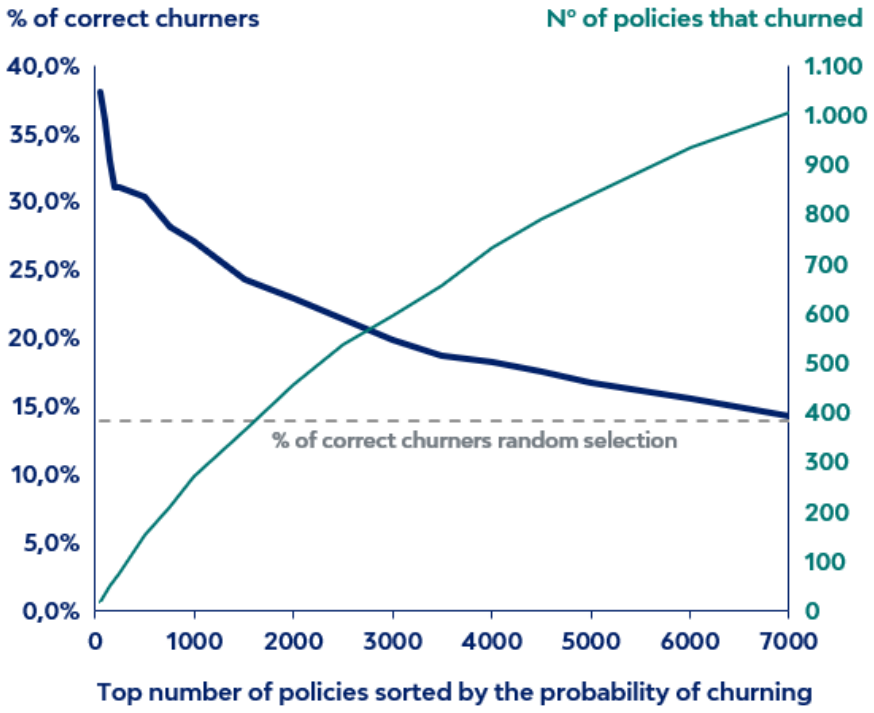



Figure 5.9: Average Results July and September 2022

These results show the percentage of correct churners in the top number of policies sorted by the probability of churning. These outcomes demonstrate how superior our model is to randomly select a subset of clients. Within the top 250, approximately 4% of our total sample size, our model can predict 2.25 times better than selecting a random sample. The percentage of accuracy expected by selecting a random sample is approximately 14%, whereas our model can accurately predict 31% of churners in the top 250. Although this may appear significantly worse than the results displayed in Section 5.2.2, it is imperative to mention that these subsets of July and September 2022 had been altered to exclude churned policies, which would not be prudent to contact. These cases were detailed in Section 4.3.1.4, and they were accurately predicted as top churners by our model. By omitting these policies, our model yields worse theoretical results; however, it is more business-oriented, only predicting cases in which it makes sense to contact the customer.

# Chapter 6

# Conclusions and Future Work

This dissertation presents the use of machine learning Models for an Insurance Brokerage operating in Portugal. It aimed to develop two machine-learning models for predicting customer Churn and identifying the clients with a higher probability of buying Multi-Risk House Insurance.

In Chapter 2, a comprehensive review of the current literature is conducted to understand the various concepts related to marketing analytics, customer relationship management, the financial industry, big data, and machine learning. The review begins with an overview of marketing concepts and their applications within analytics, followed by a discussion of customer relationship management and approaches to improving customer relationships. The financial industry is then explored, along with big data and machine learning topics. At the end of this chapter, a critical reflection is presented. Regarding machine learning models and their performance in the context of this project, Gradient Boosting Machines were found to be the most effective model, affirming the widespread utilization of Boosting techniques in machine learning problems (Vorobyev and Krivitskaya, 2022; Dhieb et al., 2019) Additionally, evaluation metrics such as Lift proved to be advantageous in elucidating to stakeholders how machine learning models can enhance customer relationship strategies.

In Chapter 3, the problem at hand is examined in detail. This includes an analysis of the Portuguese financial sector and the firm's position within the market. Furthermore, the project structure, the Next Best Offer and Churn structure, and goals are provided.

In Chapter 4, the methodological approach to obtain the solutions for both cases is detailed. Additionally, this chapter outlines the strategy utilized to address the issue of missing information on some variables. Binning segmentation was the chosen approach for all variables except the socio-demographic ones in the Churn part. For the socio-demographic variables in the Churn part, a cluster was employed. The clients who did not have associated socio-demographic information and did not belong to the socio-demographic cluster were grouped into a distinct group. This cluster was only employed in the Churn part due to its superior performance results compared to the model with the socio-demographic binning segmentation. The methodology adopted to address missing information is open to debate, and alternatives such as feature imputation based on unsupervised machine learning algorithms such as K-Nearest Neighbour can be employed, as

elucidated in Huang et al. (2017). Regarding feature selection, Boruta was employed in the Churn part, providing insights on which features would be best to omit. However, these findings were not utilised since these features were later grouped into a socio-demographic cluster. Regarding the monitoring of contacts, two Excel interfaces were developed and will serve as future work to help stakeholders integrate an interface similar to that of this project into their software systems.

In Chapter 5, the results for both use cases are presented. Gradient Boosting Machine was identified as the best suited machine learning type for both use cases. A random grid search optimized towards the F2 metric was utilized in the Next Best Offer part, resulting in improved results. However, this approach provided mediocre results for the F0.5 score in the Churn part, affirming that optimizing for a particular performance metric does not guarantee better results across all metrics. Furthermore, the academic community has yet to reach a consensus on conducting the evaluation procedure when testing multiple algorithms and incorporating hyperparameter optimization. The approach adopted in this dissertation has some merit, as the metrics selected make sense from a business perspective. Nevertheless, the methodology remains open to challenge. Furthermore, the efforts to gain insight into the model's inner workings constitute a module that can be replicated in any machine learning exercise. Aside from providing reassurance to those without technical knowledge concerning the predictions generated, it helps the model designer to verify the algorithm's decisions to ascertain whether the model's behaviour is logical, even in the case of misclassification. The assessment of the most important variables and SHAP summary was addressed for this application. However, additional analyses, such as partial dependence plots, can also be helpful when attempting to explain machine learning models. Regarding the client's contacts, by the time this dissertation was submitted, only the results from the Next Best Offer had been obtained. Despite Next Best Offer contacts being in their early stages, it is evident that there is a distinct improvement when selecting customers with higher probability as determined by the machine learning model. Currently, the treatment group is the only group with successful sales and a notable difference in clients close to purchasing the insurance.

In conclusion, several areas can be improved in the future. Firstly, the integration of data management into the Insurance Brokerage's systems so that the machine learning algorithm can accurately predict customers likely to buy the insurance and recommend the sales team to contact them. Data quality can also be enhanced; for instance, in the Churn part, the time of Churn for historical churners was calculated based on the assumption that the customer had not paid when expected. Yet, this prediction would be more precise if a churn date was recorded in the historical data. Furthermore, a machine learning model can be developed to make predictions based on the outcomes of previous contacts. Ultimately, Churn contacts will commence in the upcoming month, thus providing insight into the effectiveness of utilizing machine learning techniques and analytical approaches to predict customer Churn in the insurance sector soon.

# References


A. Agarwal, C. Singhal, and R. Thomas. Ai-powered decision making for the bank of the future. *McKinsey & Company. 2021. March. URL: https://www.mckinsey.com/industries/financial-services/our-insights/ai-powered-decision-making-for-the-bank-of-the-future*, 2021. Accessed: Oct. 28, 2022.

A. K. Ahmad, A. Jafar, and K. Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1):1–24, 2019.

S. Aiello, C. Click, H. Roark, L. Rehak, and P. Stetsenko. Machine learning with python and h2o. *H2O. ai Inc*, 2016.

S. Alelyani, J. Tang, and H. Liu. Feature selection for clustering: A review. *Data Clustering*, pages 29–60, 2018.

American Marketing Association. American marketing association website. https://www.ama.org/the-definition-of-marketing-what-is-marketing/, 2022. Accessed: Oct. 14, 2022.

S. S. Anand, A. Patrick, J. G. Hughes, and D. A. Bell. A data mining methodology for cross-sales. *Knowledge-based systems*, 10(7):449–461, 1998.

G. Armstrong. *Marketing: an introduction*. Pearson Education, 2009. ISBN 9780273713951.

Autoridade de Supervisão de Seguros e Fundos de Pensões. Asf website. https://www.asf.com.pt/NR/exeres/5378CF72-A45A-41ED-AD50-0275C4DDFB4F.htm, 2022. Accessed: Nov. 5, 2022.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

N. V. Chawla, N. Japkowicz, and A. Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

T. H. Davenport and J. Dyché. Big data in big companies. *International Institute for Analytics*, 3 (1-31), 2013.

T. H. Davenport, L. Dalle Mule, and J. Lucker. Know what your customers want before they do. *harvard Business review*, 2014.

G. Desirena, A. Diaz, J. Desirena, I. Moreno, and D. Garcia. Maximizing customer lifetime value using stacked neural networks: An insurance industry application. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 541–544. IEEE, 2019.

N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud. Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE international conference on vehicular electronics and safety (ICVES)*, pages 1–5. IEEE, 2019.

N. Edwine, W. Wang, W. Song, and D. Ssebuggwawo. Detecting the risk of customer churn in telecom sector: a comparative study. *Mathematical Problems in Engineering*, 2022, 2022.

N. Elgendy and A. Elragal. Big data analytics: a literature review paper. In *Industrial conference on data mining*, pages 214–227. Springer, 2014.

S. L. France and S. Ghose. Marketing analytics: Methods, practice, implementation, and links to other fields. *Expert Systems with Applications*, 119:456–475, 2019.

Fundação Francisco Manuel dos Santos. Pordata. https://www.pordata.pt/municipios, 2022. Accessed: Nov. 20, 2022.

T. Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2):137–146, 2011.

V. Ganganwar. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47, 2012.

T. Gattermann-Itschert and U. W. Thonemann. How training on multiple time slices improves performance in churn prediction. *European Journal of Operational Research*, 295(2):664–674, 2021.

M. Gorgoglione, U. Panniello, et al. Beyond customer churn: generating personalized actions to retain customers in a retail bank by a recommender system approach. *Journal of Intelligent Learning Systems and Applications*, 3(02):90, 2011.

H2O.ai. *Model Performance*, 2016.

H2O.ai. *H2O: Scalable Machine Learning Platform*, 2022. URL https://github.com/h2oai/h2o-3. Accessed: Dez. 2, 2022, version 3.38.0.2.

J. Huang, J. W. Keung, F. Sarro, Y.-F. Li, Y.-T. Yu, W. Chan, and H. Sun. Cross-validation based k nearest neighbor imputation for software quality datasets: an empirical study. *Journal of Systems and Software*, 132:226–252, 2017.

C. Janiesch, P. Zschech, and K. Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

W. A. Kamakura, M. Wedel, F. De Rosa, and J. A. Mazzon. Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in marketing*, 20(1):45–65, 2003.

T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury. Customer segmentation using k-means clustering. In *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*, pages 135–139. IEEE, 2018.

A. Knott, A. Hayes, and S. A. Neslin. Next-product-to-buy models for cross-selling applications. *Journal of interactive Marketing*, 16(3):59–75, 2002.

M. B. Kursa, A. Jankowski, and W. R. Rudnicki. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.

C. W. Lamb, J. F. Hair, and C. McDaniel. *Marketing*. Cengage Learning, 2012. ISBN 9781111821647.

K.-N. Lau, S. Wong, M. Ma, and C. Liu. 'next product to offer'for bank marketers. *Journal of Database Marketing & Customer Strategy Management*, 10(4):353–368, 2003.

L. Lesage, M. Deaconu, A. Lejay, J. A. Meira, G. Nichil, et al. A recommendation system for car insurance. *European Actuarial Journal*, 10(2):377–398, 2020.

L. Li, J. Lin, Y. Ouyang, and X. R. Luo. Evaluating the impact of big data analytics usage on the decision-making quality of organizations. *Technological Forecasting and Social Change*, 175: 121355, 2022.

S. Li, B. Sun, and R. T. Wilcox. Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research*, 42(2):233–239, 2005.

W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li. An ensemble random forest algorithm for insurance big data analysis. *Ieee access*, 5:16568–16575, 2017.

W.-Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82 (3):329–348, 2014.

LTPlabs Website. Ltplabs website. https://ltplabs.com, 2022. Accessed: Oct. 28, 2022.

D. Lubo-Robles, D. Devegowda, V. Jayaram, H. Bedle, K. J. Marfurt, and M. J. Pranter. Machine learning model interpretability using shap values: Application to a seismic facies classification task. In *SEG International Exposition and Annual Meeting*. OnePetro, 2020.

D. Manzano-Machob. The architecture of a churn prediction system based on stream mining. In *Proceedings of the Artificial Intelligence Research and Development 16th International Conference Catalan Association for Artificial intelligencela*, volume 256, 2013.

K. Morik and H. Köpcke. Analysing customer churn in insurance data–a case study. In *European conference on principles of data mining and knowledge discovery*, pages 325–336. Springer, 2004.

A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7: 21, 2013.

G. Piatetsky-Shapiro and B. Masand. Estimating campaign benefits and modeling lift. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 185–193, 1999.

Portal do Instituto Nacional de Estatística. Instituto nacional de estatística. http://www.ine.pt/., 2022. Accessed: Nov. 20, 2022.

M. Qazi, G. M. Fung, K. J. Meissner, and E. R. Fontes. An insurance recommendation system using bayesian networks. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 274–278, 2017.

E. Raguseo. Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, 38(1):187–195, 2018.

S. Ray. A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 35–39. IEEE, 2019.

M. Scriney, D. Nie, and M. Roantree. Predicting customer churn for insurance data. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 256–265. Springer, 2020.

Serdar Yegulalp. Info world website. https://www.infoworld.com/article/2612856/bye-bye--big-red--escaping-oracle-s-not-that-easy.html, 2022. Accessed: Dez. 5, 2022.

U. Srivastava and S. Gopalkrishnan. Impact of big data analytics on banking sector: Learning for indian banks. *Procedia Computer Science*, 50:643–652, 2015.

I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, 7:60134–60149, 2019.

I. Vorobyev and A. Krivitskaya. Reducing false positives in bank anti-fraud systems based on rule induction in distributed tree-based models. *Computers & Security*, 120:102786, 2022.

M. Wedel and P. Kannan. Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6):97–121, 2016.

Y. Xie, X. Li, E. Ngai, and W. Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3):5445–5449, 2009.

W. Xu, J. Wang, Z. Zhao, C. Sun, and J. Ma. A novel intelligence recommendation model for insurance products with consumer segmentation. *Journal of Systems Science and Information*, 2(1):16–28, 2014.

D. Zakrzewska and J. Murlewski. Clustering algorithms for bank customer segmentation. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pages 197–202. IEEE, 2005.

# Appendix A

# Next Best Offer Contact Registration Interface


Figure A.1: Next Best Offer Contact Registration Interface - First Page

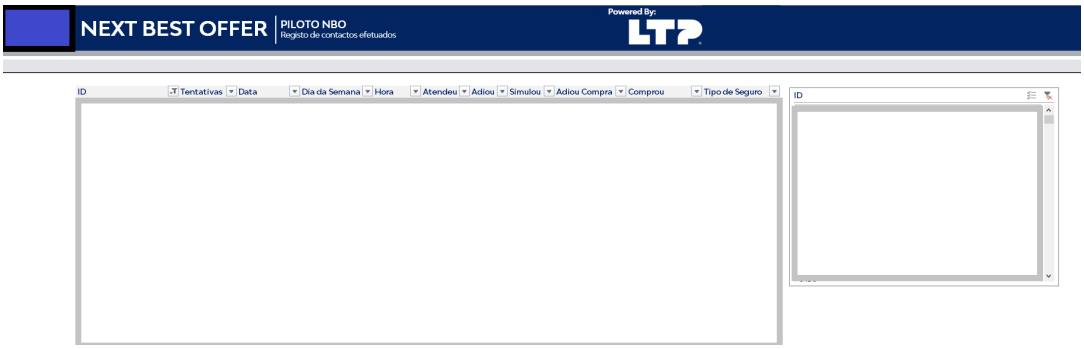



Figure A.2: Next Best Offer Contact Registration Interface - Second Page

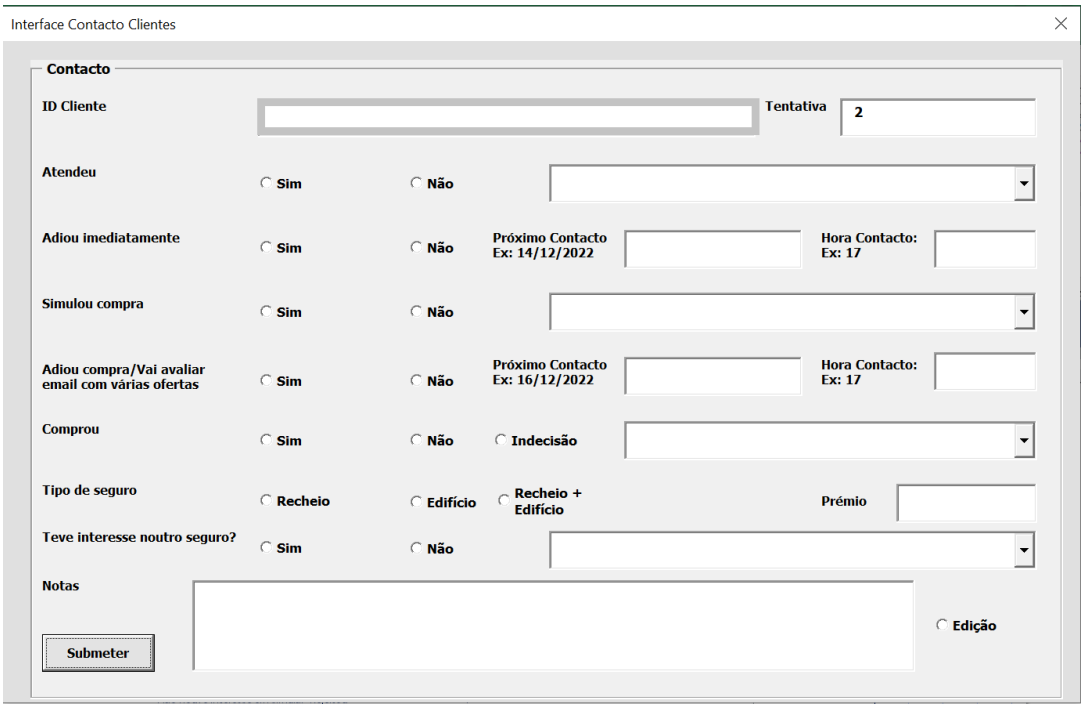



Figure A.3: Next Best Offer Contact Registration Interface - Form

# Appendix B

# Churn Offer Contact Registration Interface

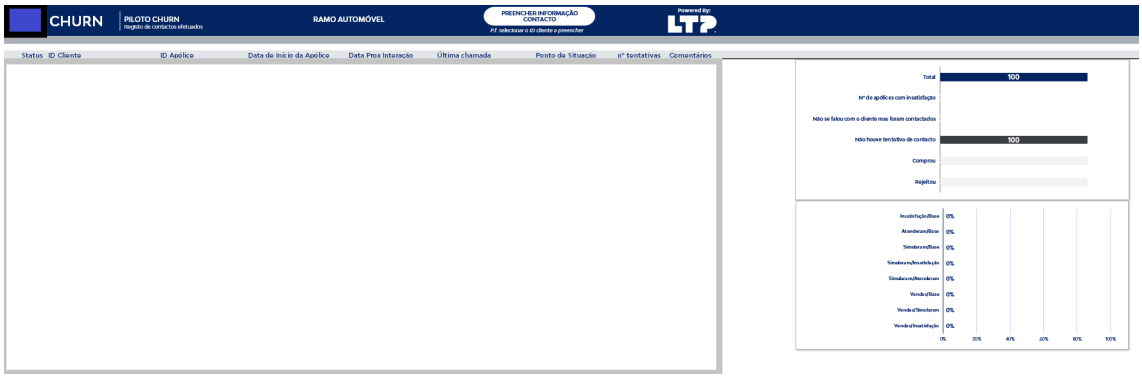



Figure B.1: Churn Contact Registration Interface - First Page


Figure B.2: Churn Contact Registration Interface - Second Page

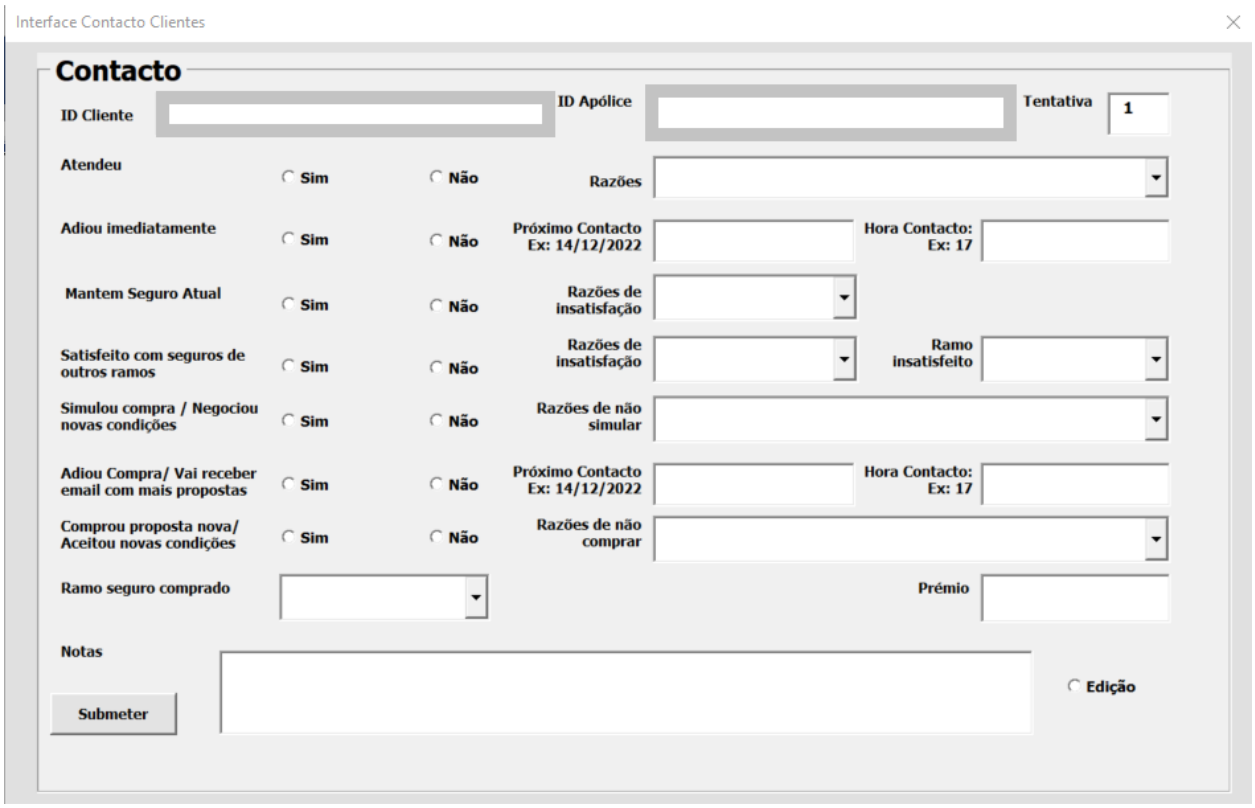



Figure B.3: Churn Contact Registration Interface - Form