NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Customer Churn Prediction in the Banking Industry

Soraia Sofia Santiago da Cunha

Internship Report

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

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

[this page should not be included in the digital version. Its purpose is only for the printed version]

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

CUSTOMER CHURN PREDICTION IN THE BANKING INDUSTRY

by

Soraia Sofia Santiago da Cunha

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

Supervisor: Prof. Frederico Jesus, PhD

November 2022

STATEMENT OF INTEGRITY

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

Soraia Cunha

Lisboa, 30 de novembro de 2022

ABSTRACT

The objective of this project is to create a predictive model that will decrease customer churn in a Portuguese bank. That is, we intend to identify customers who could be considering closing their checking accounts. For the bank to be able to take the necessary corrective measures, the model also aims to determine the characteristics of the customers that decided to leave. This model will make use of customer data that the organization already has to hand. Data pre-processing with data cleansing, transformation, and reduction was the initial stage of the analysis. The dataset is imbalanced, meaning that we have a small number of positive outcomes or churners; thus, under-sampling and other approaches were employed to address this issue. The predictive models used are logistic regression, support vector machine, decision trees and artificial neural networks, and for each, parameter tuning was also conducted. In conclusion, regarding the customer churn prediction, the recommended model is a support vector machine with a precision of 0.84 and an AUROC of 0.905. These findings will contribute to the customer lifetime value, helping the bank better understand their customers' behavior and allow them to draw strategies accordingly with the information obtained.

KEYWORDS

Customer churn prediction; Banking; Machine learning; Supervised learning

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Networks
API	Application Programming Interface
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
CLV	Customer Lifetime Value
DT	Decision Tree
KNN	K-Nearest Neighbour
SVM	Support Vector Machine

1. INTRODUCTION

The following report is the result of a six-month internship at a banking institution in Portugal. A bank is a financial institution responsible for intermediating resources between surplus and deficit agents, being essential for commercial activities. The job was to develop churn prediction models for checking accounts.

1.1. STATEMENT OF THE PROBLEM

Banks have increasingly established a variety of online and physical customer interactions to meet customers' demands and channel transactions as science and technology have improved, resulting in a great amount of user data. In the face of a large amount of data, banks must have a more thorough and accurate understanding of customer information during actual business development, analyze and predict customer attrition.

The banking industry faces the challenge of retaining customers. Customers may switch to another bank for various reasons, such as better financial services due to low rates, bank branch locations, quality of digital tools, and low-interest rates, among others (Kaur & Kaur, 2020). The internet and the explosion of social media have had a tremendous impact on the retail banking industry over the last decade, allowing customers more accessible and faster access to information. Banking activities are being increasingly digitized, and this trend appears to be unstoppable.

Churn prediction is used to identify customers who are about to leave a service provider. It costs a corporation 5 to 20 times more to keep a single customer than it does to acquire a new one (Sagala & Permai, 2021). Predictive models can enable reliable identification of likely churners shortly, which can be used to provide a retention solution. Consumers can be marketed ahead of time and on schedule to reduce the loss of bank funds.

1.2. OBJECTIVE AND SUMMARY OF THE PROCESS

This research presents a new framework for dealing with bank customer turnover. In other words, rather than reacting to customer attrition, a proactive approach is recommended to address the problem.

Customer churn prediction is the process of assigning a churn probability to each customer in a database based on an expected relationship between that customer's past data and future churning behavior (Coussement et al., 2017).

With this in mind, the overall purpose of this research is to suggest a new framework for bank churn management, so that the company's marketing efforts are directed toward retaining current customers rather than just recovering former churners. To accomplish these objectives, it is possible to summarize the process in the following way:

- In first place, data mining techniques will be used to investigate the behavior and trends of a sample of former churners over some time. Our goal in this step is to identify the most significant (independent) variables that had the biggest impact on those customers' churning. The most common churn management activity focuses on churn prediction utilizing historical churn data and predictors of customer churn (Ismail et al., 2015).

- The second stage of this procedure will focus on using the variables from the first phase as inputs to train a collection of predictive models and evaluate their performance.
- The third stage will involve assessing the models' quality by examining the classifier's performance in each model. The churn prediction models are evaluated accordingly with the metrics chosen to measure efficiency. The one that produces better results will be used to predict whether a customer will churn.
- The next phase in the process will be to test the previously trained models with new data from a sample of current consumers and then publish the results for each model based on the target variable (churner or non-churner).
- Finally, the final stage will focus on tracking the behavior of those customers over time and comparing the models' forecast quality to the actual outcome of those customers.

1.3. CONTRIBUTION TO THE COMPANY

The current study is relevant because it is a follow-up to the company's current churn-related work projects. The bank is currently studying the customer lifetime value, with churn prediction being an important element of this study since it contributes to customer characterization. Customer Lifetime Value (CLV) is a crucial indicator that companies should monitor and control. It aids in their understanding of the financial losses incurred from losing customers. Additionally, it aids them in calculating the price of attracting new customers and the potential profit from each one. The goal of this research was to aid in the development of new products and services that would add value to the company's present customers and deepen its relationship with them. The completion of such a study demonstrates that churn is a topic that causes widespread worry within the firm. Churn will continue to exist, and performing customer management is better than acquiring new customers to ensure long-term business growth and profitability (Ismail et al., 2015).

Furthermore, regardless of industry, monitoring churn should be a constant worry for any firm today, since competition is tough, and the digital era has made it easier for customers to take their business elsewhere if they want. As a result, it is to any company's best advantage to maintain a watch on its customers' behavior to foresee any unhappiness that could lead to churning. Such steps may be crucial in reaching out to those customers and, perhaps, saving the bank's relationship. The financial rewards of such initiatives are self-evident, as potential churners choose to keep their accounts with the bank. Additionally, correctly predicting the churners and convincing them to stay at the company can increase revenue, even if a churn prediction model produces a certain number of false positives (Mutanen et al., 2010).

Churn Prediction helps develop and implement strategies to manage interactions between organizations and high-risk customers to retain those customers. The work developed during the internship provided insight into the churn phenomenon and was able to draw some conclusions about the customers that closed their checking accounts at the time of the study.

2. LITERATURE REVIEW

2.1. BANKING

Increased customer happiness is already a priority for bank executives around the world. Customers' expectations and levels of demand for financial services rise in tandem with their adoption of new technology in other areas of their lives.

Retail banks must choose between aligning their services to customer expectations or risk losing those customers entirely, according to Capgemini and Efma's World Retail Banking Report 2021 (WRBR). The post-pandemic disruption has opened a new era of value-based, customer-centric banking, called Banking 4.X, while the economic ramifications of COVID-19 continue. To thrive in Banking 4.X, banks must embrace digital transformation and adopt cloud-based Banking-as-a-Service (BaaS) platform models that use APIs to embed banking in everyday life, making it more accessible and inclusive for banking customers.

Financial technology enterprises that provide internet-only financial services and do not have physical branches have gained more than 39 million customers in the last ten years. Currently, 81 percent of consumers say that simple access and flexible banking will entice them to switch from their traditional bank to a new-age financial service. Meanwhile, many traditional banks are attempting to retain and develop their customer base, and many have already begun their digitalization and cost-cutting journeys in response to the COVID-19 pandemic, which has prompted them to speed up their efforts further. Furthermore, customers want on-demand, fully digitalized experiences, hyper-personalized services, and round-the-clock assistance when confronted with pandemic-related reality (Sacchi et al., 2021).

Non-traditional financial competitors threaten traditional banks for 55 percent of bank executives. Traditional banks will find it increasingly difficult to keep their current customer base due to this varied competition scenario (Garvey et al., 2014).

2.1.1. Churn in Banking

A churned customer is someone who closes all his accounts and ceases conducting business with the bank (Chitra & Subashini, 2011). Churning consumers can result in not only the depreciation of cash but also a reduction in corporate profitability and other negative effects on operations (Karvana et al., 2019). As a result, customer-centric retention seeks to reduce churn by searching for and identifying consumers that have a high predisposition to quit the organization or predict customer attrition (De Caigny et al., 2018).

Companies that build long-term connections with their consumers may concentrate on their needs, which is more profitable than seeking new customers (Reinartz & Kumar, 2003). Companies may be able to minimize service costs as a result of this (Ganesh et al., 2000), as well as increase cross- and up-selling opportunities (De Caigny et al., 2018). Long-term customers are also less swayed by competitors' marketing initiatives (Colgate et al., 1996), and they are more inclined to buy more and promote favorable word-of-mouth (Ganesh et al., 2000). Customers who leave the company may persuade others to do so as well (Nitzan & Libai, 2011). Customers that leave due to "social contagion" may be more difficult to keep (Verbraken et al., 2014).

Detecting churn in the banking industry has unique issues. To begin with, large banks generally have tens of millions of customers on their books and human intervention-based churn reduction strategies do not scale well. Second, they are incapable of fast adjusting to changes in customer requirements. Third, although banks separate consumers into local managers, manually detecting customer trends is still challenging, especially if they handle a large number of customers. So, the development of automated systems capable of detecting non-trivial consumer behavior patterns could indicate possible churn in these enormous data sets ahead of time. The employment of machine learning techniques offers supervised learning methods that have been shown to learn non-trivial patterns in data without the need for human involvement and generalize well to previously unseen data (de Lima Lemos et al., 2022).

Neslin et al., 2006 highlight the value of measures related to churn reduction and customer retention in five basic points:

(1) Customer retention decreases the need to prospect for new customers, allowing firms to focus on strengthening connections with existing customers.
(2) Older consumers, who are more familiar with the company, tend to purchase more and, when satisfied, can refer others.
(3) Because of the additional information obtained during their consumer life cycle, serving, and sustaining long-term customers is less expensive.
(4) Long-term customers are less likely to respond to competition marketing.
(5) Customer loss is a cost of opportunity since it lowers sales and needs the acquisition of new customers to compensate for the loss.

2.2. PREDICTIVE MODELING

Predictive modeling is a statistical technique for predicting future behavior that is widely employed. Predictive modeling solutions are a type of data-mining technology that analyses historical and current data to create a model that can forecast future outcomes. Data is gathered, a statistical model is developed, predictions are generated, and the model is validated (or amended) as new information becomes available (*Definition of Predictive Modeling - Gartner Information Technology Glossary*, 2022).

Predictive modeling focuses on forecasting how a customer will behave in the future based on their previous behavior. One example of predictive modeling is predicting consumers who are likely to churn. Customer Relationship Management (CRM) data is analyzed using predictive modeling to create customer-level models that indicate the chance that a customer would perform a specific action. The acts are typically related to sales, marketing, and customer retention. A variety of models can be used to identify churners and non-churners in an organization. Traditional models or approaches (Regression Analysis) and soft computing techniques (Neural Networks) are two types of models (Shaaban et al., 2012).

2.2.1. Predictive models in churn

In the banking industry, customer churn prediction models are frequently judged simply on their predictive performance, or their ability to distinguish between churners and non-churners. More

improved strategies have been presented and assessed in recent years. de Lima Lemos et al., 2022 studied the churn prediction of customers in the banking sector using a unique customer-level dataset from a large Brazilian bank. Their work reveals that customers with a stronger relationship with the bank, that is, who have more products and services and who borrow more, are less likely to close their checking accounts. The random forests model produced consistent models with superior results in the tests. It was able to predict that 80.2% of consumers would churn in the coming months using this technique. The metric used to assess the best model was recall.

In Deng et al., 2021 paper, the purpose of the study is to analyze the quarterly user data of banks and establish a user churn prediction model using the Catboost, Lightgbm, and Random Forest techniques to improve the accuracy of prediction, to achieve the purpose of helping banks save costs. In this experience, the Random Forest performs best, followed by the Lightgbm. The random forest is difficult to overfit due to the introduction of two randomness, and it has some anti-noise capabilities. Therefore, it performs well on the testing set. It is concluded that customers with a high deposit or financial products have a low turnover rate after each quarter, therefore, they should focus on individuals with a small deposit. Another conclusion drawn from the study was that low-age customers had a greater user churn rate, according to a single-factor study of age characteristics. Banks could hold some online events to encourage young people to continue to choose their banks, while also maintaining older consumers through the following offline activities.

Hassonah et al., 2019 compared the performance of churn prediction between two machine learning algorithms which are Decision Trees and K-Nearest Neighbors algorithms. Statistical results were found to be in favor of the decision tree algorithm since accuracy, precision, recall, F-measure, and the Lift measure were found to be better in DT than in the K-NN algorithm. Both algorithms have similar accuracy results as the DT had a percentage of 93%, while the accuracy of K-NN reached 87%. The largest difference between the two algorithms was with the F1-score reaching around 33% and 73% for K-NN and DT, respectively. The cause of this difference is due to the recall (sensitivity) measure as it appears to be very low for the K-NN algorithm. Such a low value is caused by the number of K neighbors, and while trying to improve the recall by decreasing the number of neighbors, the precision was decreased to around the same ratio with the increasing recall, resulting in the same F-Measure. Results show that the AUC value for DT is slightly better than the AUC value for K-NN.

KMeans, Local Outlier Factors (LOF), and Cluster-Based Local Outlier Factors (CBLOF) are three churn prediction algorithms that have yet to be deployed for this purpose. Ullah et al., 2019 used these strategies to forecast customer attrition in the study. As far as we are concerned, the CBLOF method outperforms other techniques on average, displaying a superior performance and computing penetration density estimations are not strictly required. CBLOF is faster than its nearest competition in terms of computational complexities. Though in rehearsal, it is propose restarting the fundamental K-Mean numerous times to achieve a steady cluster result. A CBLOF might be employed when process speeds or a clustered model can be updated in data-flowing applications. Furthermore, K-Means algorithms are vulnerable to outlier detections, and K-Means is unable to find proper churner consumers because it only detects customers with the greatest distance from the rest. Furthermore, the LOF does not detect accurate customer outcomes, such as when raw data is churner/outliers, and it also does not recognize consumers that have the same densities as normal customers but are actual churner/outlier customers.

Another perspective is to investigate and analyze a variety of boosted-based methodologies to obtain the best possible customer churn model for a bank (Sagala & Permai, 2021). Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Category Boosting (CatBoost) are the three boosting-based models evaluated in this study to categorize customer turnover data. The models were trained using 80% of the data set that was chosen at random. The grid Search approach on 10-fold Cross-Validation is used to analyze and improve the models. In the end, the findings show that using LightGBM to perform produces the best possible model, with accuracy and AUC, precision, and recall scores of 91.4%, 94.8%, and 87.7%, respectively.

3. METHODOLOGY

3.1. Defining Buffer period and Time frame for Independent & dependent (Target) Variable

The buffer period is the month of the year that serves as the starting point for the definition of our Observation and Performance Window. In our situation, collecting all customers holding checking accounts as of a certain month will serve as an illustration of a buffer period (202107).

Here, we define two windows: the performance window and the observation window. The time period where we construct our target variable is referred to as the Performance window, whereas the window during which we create our independent variables is known as the Observation window.

It should be remembered that the observation window is before the buffer period month, whereas the performance window is after. As of 202107, we had two customers, A and B, each of whom held a checking account (see figure below). Customer B is still a financial institution customer whereas customer A attrited (left) after three months during the performance window (Koli, 2020).

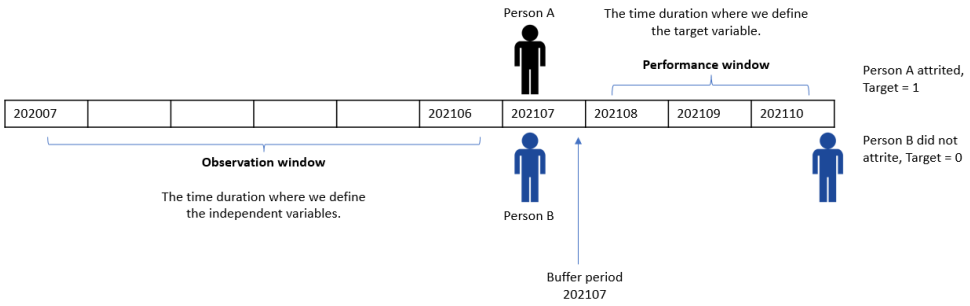

Figure 1 - Defining Observation and Performance Windows

It is vital to keep in mind that the length of observation and performance windows is determined by a variety of factors, some of which are listed below:

(1) **Availability of data**: If the data is only accessible for a little time, we can shorten both frames. Nevertheless, building a model on a big sample of data is always desirable.
(2) **The event rate during the performance window**: The number of consumers who have left the buffer period divided by the total number of checking customers is the event rate. In one month, fewer customers would attrite than in six months. A higher event rate is desirable for building a good model.
(3) **Seasonality in the data**: Certain months have a greater rate of attrition when compared to other months of the year. We would like to go through the time span with seasonal changes.
(4) **Reactiveness in dealing with Customers likely to churn**: Financial institutions may pursue to take proactive actions to decrease customer attrition. In such cases, the Performance Window should be set to a small value to predict attrition early.

3.2. BUSINESS UNDERSTANDING

As mentioned before, this project is about developing a churn model for the checking accounts. To guarantee the project's success, the first objective was to understand this business line and the project objectives. Further, this knowledge has been converted into data mining goals, and a project plan has been designed.

Customer churn, which is the tendency for consumers to stop doing business with a company in a specific time frame, has grown to be a serious issue and is one of the main difficulties that many businesses across the world have to deal with (Xie et al., 2009). Customer churn is a significant issue and one of the top issues for big businesses. Companies are working to create methods to predict probable customer churn because it has a direct impact on their revenues. To reduce customer churn, it is crucial to identify the variables that contribute to this churn (Ahmad et al., 2019). There are numerous justifications for closing a bank account. For any number of reasons, including because you find better terms at another bank, wish to keep your money in the same account as your mortgage, have too many accounts and incur high commission costs, or for any other reason.

Customers typically can close their bank accounts at any time. The financial service contract does, however, include a notice period in specific exceptional circumstances. Then it must be followed if such is the case. This means that the notification time cannot go beyond one month in compliance with the law. A registered letter must be mailed as proof that you have followed the rules to tell the bank in advance that you wish to terminate the account.

3.3. DATA UNDERSTANDING

Data was gathered once the initial business insights were gained. Here, the various variables were retrieved while considering the criteria that will be mentioned below, data from various sources were combined, and most inconsistencies were addressed immediately. One of the most time-consuming aspects of the project ended up being the below-mentioned duties, which were all carried out in SAS Enterprise Guide. The next steps were done in python using the software Anaconda Navigator. After the data was made available, a data exploration was conducted to gain first insights into the data, identify potential data quality problems, and identify tasks that needed to be completed during the data preparation phase.

When building the dataset, some criteria were established regarding the type of customers that will be present in our analysis. Thus, we only wanted customers that possessed a checking account that had at least some credit movements on the account throughout the transactional history, that is, the customer did not open an account where he did not perform any activity. In that way, we avoided having new customers, deaths, and bank employees. Moreover, the customers in our dataset are also active, meaning that they performed movements of their initiative during the last month of analysis.

On the customers defined as churners, a set of assumptions has been considered. The first is that the churn was permanent, indicating that the customer did not return to the bank after the performance window. The second is that accounts closed by the bank strategy are not considered churners. Additionally, priority was given to accounts closed by the customer initiative due to technical situations, and residency alterations, among other reasons, that is voluntary churn. We considered a

voluntary churner a customer that makes the proactive decision to close his/her checking account in the bank.

Regarding the timeline, as already mentioned, the time frame is divided into three periods: the observation window, the period where the independent variables are created, from July 2020 to June 2021. Then the buffer period is the period where the bank will implement the prediction model to get the customers that most likely will churn in the next months and share that information with the local agencies to contact those customers in order to make them continue in the bank. Most expected is a marketing strategy to be created for these specific customers. The buffer period has a duration of 1 month, that in our case, is July 2021. Finally, the performance window, the period where the target comes from, is from August 2021 to October 2021. In that way, for our target variable, the customers that have the value 1, in other words, the churners, are the ones that closed their accounts during the performance window.

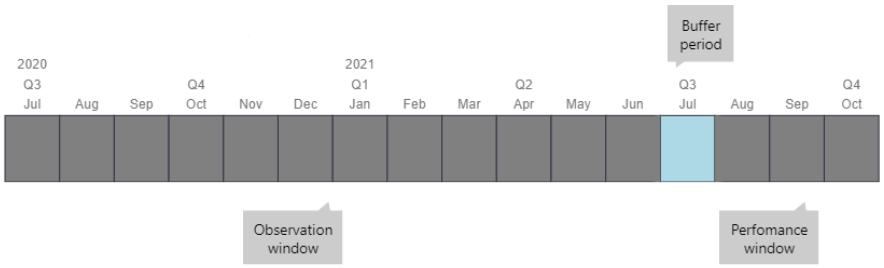

Figure 2 - Timeline of study

There were selected 110 categorical and numerical variables for this study. The variables that are part of the research belong to two groups: customer characteristics and the relation between the customer and the bank. In the first group, we have features that help us to characterize the customer, like age, gender, marital status, education level, professional situation and city, and district. In the second group, we have features related to the interaction between the customer and the bank, such as if the customer has a credit and/or debit card, when was the last time that they used the credit/debit cards if he/she has access to the home banking and the last time he/she went there, what financial products they possess, the number of movements made by the cards they have, and monetary balance.

3.4. DATA PREPARATION

Our analysis began with understanding the data we have. With this initial data, we could conclude that the bank's churn rate was 3.3% at the time of the study. During this phase, we performed a light analysis and visualizations of all the data, retrieving some possible findings better analyzed in the following steps.

Data Cleaning

Enhancing the quality of the data in an existing system is the process of data cleansing, which may be closely related to data gathering (Maletic & Marcus, 2000). User input errors, inconsistent input, missing values, misspellings, and inappropriate data production can all have an impact on the quality of the data. These issues hinder the gathering of correct analytic data, leading the researchers to make some false inferences. By addressing these issues, data cleaning raises the quality of the data for analysis (Calabrese, 2019).

Missing values

Missing values are frequently attributed to human error while handling data, machine error due to faulty machinery, respondents' refusal to answer specific questions, study dropout, and merging unrelated data (Suthar et al., 2012). The missing values problem is typically prevalent in all data-related disciplines and results in a variety of challenges, including performance deterioration, issues with data processing, and biased results (Ayilara et al., 2019). Additionally, the amount of missing data, the pattern of missing data, and the process underlying the missingness of the data all have a role in how important missing values are (Kang, 2013). Imputation is a method for handling missing values that involves replacing them with potential or estimated values in place of the missing values. Several traditional statistical and machine learning imputation techniques, such as mean, regression, and K nearest neighbor, have been suggested in the literature to handle missing values (Emmanuel et al., 2021).

Most information can be retained by replacing the missing values with new values rather than removing the variable or even the entire row. However, if a variable has many missing values, leaving it in could put the model in danger of receiving inaccurate information due to a large number of synthetic values. Therefore, depending on the properties of each variable, a different method was utilized to deal with the missing values.

It is crucial to take into account the underlying causes for the missing data while analysing the potential effects of those data shortages on registry results. Missing data are normally grouped into three categories (Mack et al., 2018):

- Missing completely at random (MCAR). The fact that data are absent when they are MCAR is unrelated to the observed and unobserved data. The most straightforward missing value type is this one. The fact that the data is missing has nothing to do with the data that has been observed or with the data that has not been observed; it is simply absent.
- Missing at random (MAR). In the case of MAR data, the absence of the data is consistently related to the observed data but not to the unobserved data. It means that other features in the dataset can predict whether data is missing.
- Missing not at random (MNAR). When data are MNAR, the fact that they are missing is systematically linked to the unobserved data, i.e., the missingness is linked to circumstances or elements that the researcher has not measured. The hardest data to find and work with both in terms of finding and using is MNAR data. The unobserved data, or the data that we don't have, is related to the missingness of the data, which is related to aspects that we failed to take into consideration.

Missing values in metric features

Depending on the type and volume of the missing data, some methods can be employed to fill in the gaps. Therefore, the following approaches can be used to deal with missing values: Case deletion is the first approach. By using this technique, all instances with at least one feature's missing value will be deleted. There are variations of this strategy, though, that take into account how much data is missing and remove instances or characteristics that have a lot of it. Identifying the qualities' relevance to the remaining data in the event of deletion is crucial. The next technique is imputation. The most popular way for handling missing data is this one, which replaces missing values with appropriate alternatives. The most prevalent replacement value options are mean imputation, median imputation, and K-nearest neighbor (KNN) imputation. The KNN imputation was the method used in this project. When a value is absent, the KNN method classifies the closest neighbors and uses those neighbors for imputation using a distance metric between instances (Maillo et al., 2017). The Euclidean distance is the most popular distance measure because it is reported to provide efficiency and productivity (Amirteimoori & Kordrostami, 2010). Other distance measures that can be used for KNN imputation include the Minkowski distance, Manhattan distance, Cosine distance, Jaccard distance, and Hamming distance.

Missing values in non-metric features

A quantitative property or a qualitative attribute of all the non-missing values is used for each value in the simple imputation strategy to replace missing values (García-Laencina et al., 2009). Simple imputation uses the mode, mean, or median of the available values to address missing data. Simple imputation techniques are most frequently utilized in research since they are straightforward and can be used as quick references (Jerez et al., 2010).

For variables for which the percentage of missing values was relatively low, the mode was the method input. According to the acquired business knowledge, other categorical variables with missing values were highly related to the customer's location. Therefore, the customer's location mode was introduced, always tracking this imputation's impact on the final variable's distribution.

Feature Engineering

Feature engineering in machine learning uses data to generate new variables that are not present in the training set. To streamline and accelerate data transformations while also improving model accuracy, it can generate new features for both supervised and unsupervised learning (Patel, 2021).

Any stage of the data can be subjected to the feature engineering process. It can convert both unprocessed raw data and processed data into usable data for a particular activity. The transformation of gathered/recorded parameters, creation of new parameter values from existing features or patterns, extraction of features from raw data, selection of features based on a specific criterion, analysis and evaluation of the usefulness of features, and automated methodologies for generating and selecting features are all parts of the feature engineering process (Dong & Liu, 2018). Another definition of feature engineering is the use of a data mining technique to extract features from unstructured data while applying domain knowledge, categorizing feature engineering as a process that is specialized to a given domain (Usmani et al., 2020).

Feature engineering is the process of creating new features from existing data to enhance the effectiveness of predictive learning. The process of creating and transforming features, or feature

engineering, is mostly manual and frequently takes the longest in a data science workflow. It is a challenging process that is carried out iteratively through trial and error and is guided by subject expertise accumulated over time. Finding appropriate characteristics is a key component of developing a strong predictive model (Khurana et al., 2016).

Outliers

Anomalies in a given dataset are known as outliers. Outliers may appear as a result of accurate or inaccurate data collection. In any case, the existence of outliers must be acknowledged and will call for specific handling. To generalize a pattern or relationship within a dataset, a representative model must be created. The presence of outliers skews the inferred model's representativeness (Kotu & Deshpande, 2019).

Today, a variety of applications need in-depth data analysis to remove outliers and guarantee system stability. Observations that sufficiently vary from typical observations are referred to be anomalies. When the number of these observations is much lower than the percentage of nominal cases—typically less than 5%—they are referred to as outliers. Most data mining methods can also gain from the process of excluding outliers from a dataset. Outlier identification techniques are very helpful for data cleaning since an outlier-free dataset enables accurate modelling tasks (Liu et al., 2004).

So, to remove the outliers, we tried different methods such as the interquartile range, the outlier detection with Local Outlier Factor (LOF) and the z-score. Comparing the three methods' results, we chose to go with the z-score, where we discarded the observations that were nine standard deviations above the mean, that is, 2.5% of the observations were removed.

Feature selection

The process of choosing the most pertinent subset of features for an effective AI/machine learning model is known as feature selection. The performance of the diagnostics model could be enhanced since redundant and/or irrelevant data is deleted from the main database during the feature selection phase. The computational load on the machine will be lowered by the feature selection process, increasing computational efficiency (Malik et al., 2021).

There are three primary feature selection paradigms. Regardless of the learning process, filter-based feature selection selects the most pertinent features. These techniques gauge the informativeness of the features by measuring their correlation, entropy, or intra/inter-class distance. The learning algorithm is used in wrapper-based feature selection to assess the quality of a subset of features. With a longer computation time than filter-based approaches, this method has the advantage of allowing the discovery of potential interaction among attributes. The advantages of both the filter and wrapper methods are intended to be combined through embedded-based feature selection. The algorithm's learning phase includes feature selection. For instance, decision trees like CART have a feature selection mechanism built in (Rosales-Pérez, 2022).

Before modeling, we must choose the best features to use in our model. Feature selection is an important action since removing irrelevant data brings benefits such as improving learning accuracy, the reduction of computation time, and facilitating an enhanced understanding of the learning model

or data. Instead of selecting only one method, we decided to try different algorithms from different methods and select the features that, in the combination of the algorithms, appear the most.

3.5. MODELING

In this phase, the used modeling techniques are selected. Also, a test design is created to build the different models and correctly assess them, understanding their value for resolving the problem at hand.

3.5.1. Imbalanced dataset

The data that we have to solve the problem is imbalanced, meaning that 3% of our data belongs to class 1 and 97% to class 0. To resolve this situation, we applied the combination of the techniques Synthetic Minority Oversampling Technique (SMOTE) and Undersampling. SMOTE is an oversampling technique that creates new minority-class synthetic samples, and Undersampling is a technique that balances uneven datasets by keeping all the data in the minority class and decreasing the size of the majority class. The idea is to combine the SMOTE technique with the Undersampling technique to increase the effectiveness of handling the imbalanced class. Removing one or both examples in these pairs (such as the examples in the majority class) makes the decision boundary in the training dataset less noisy or ambiguous. For the imbalanced dataset, first SMOTE is applied to create new synthetic minority samples to get a balanced distribution, then random Undersampling deletes examples from the majority class resulting in losing information invaluable to a model. By taking samples at random from the population of the majority class until the minority class reaches a certain percentage of the majority class, the majority class is undersampled. The minority class has a greater presence in training set at higher degrees of under-sampling, which causes the learner to encounter varied levels of under-sampling. The learner's initial bias toward the negative (majority) class is reversed in favor of the positive (minority) class by combining under-sampling and over-sampling. Classifiers are trained on a dataset that had the minority class "SMOTED" and the majority class undersampled. (Chawla et al., 2002)

3.5.2. Model selection

Initially, for this project, four predictive models were selected and constructed: a Logistic Regression, a Decision Tree, a Multilayer Perceptron model, and a Support Vector Machine model. These methods are briefly explained below:

Logistic Regression

A potent supervised machine learning approach utilized for binary classification issues is logistic regression (when the target is categorical). In essence, logistic regression models a binary output variable using the logistic function (Tolles & Meurer, 2016). The main distinction between logistic regression and linear regression is that the range of logistic regression is constrained to values between 0 and 1. Additionally, logistic regression does not require a linear relationship between the input and output variables, in contrast to linear regression (Belyadi & Haghighat, 2021).

In essence, logistic regression is a classification algorithm. Its close sibling in the regression domain, known as linear regression, is where the word "regression" originates. In supervised classification

issues, the classes are discrete, therefore, the algorithms' objective is to identify the decision borders between the classes. Examples of one class are differentiated from another by decision boundaries. Decision boundaries may have a complex and nonlinear geometric shape depending on the specific problem instance. Many machine learning algorithms make varying assumptions about the design of decision boundaries. The assumption in logistic regression is that the decision boundaries are linear. In other words, they are hyperplanes in the high-dimensional feature space, whose dimension is solely defined by the number of elements in a training example's feature vector.

The weights for the features are generally equivalent to the logistic regression model parameters. The S-shaped logistic function is used to map each weighted feature vector to a number between 0 and 1. This number is taken to mean the likelihood that an example belongs to a specific class. For the training examples to be correctly classified, the learning algorithm adjusts the weights. Here, it is unavoidable to raise the question of avoiding overfitting. Popular methods for adjusting the weights include the gradient descent approach and numerous variants. The logistic function calculates the likelihood that any unobserved example will belong to a class once the weights have been determined.

Logistic regression is frequently the first technique that is used for classification issues since it makes the oversimplified assumption that decision boundaries are linear. Also, logistic regression is thought to be less prone to overfitting due to the linear, noncomplex decision boundaries. Overfitting, as the name implies, happens when we attempt to accurately categorize each training example by randomly moving the decision border. Additionally, as gradient descent frequently operates quickly, the training stage of logistic regression is also short. All these benefits support the widespread use of logistic regression to solve various categorization issues. On the downside, the simplistic modeling assumptions may lead to underfitting for rich and complex datasets (Gudivada et al., 2016).

Decision Trees

In the machine learning community, decision trees are seen as a potential answer for categorization applications. Their appeal is a result of both their adaptability to the inference task by generating logical rules of categorization and their capacity to handle complex problems by offering a comprehensible representation that is simpler to interpret (Amor et al., 2006).

A decision tree can be generated from training data using the supervised machine learning method known as decision tree learning. A decision tree is a predictive model that maps observations about an item to judgments about its goal value. It is also known as a classification tree or reduction tree. The non-leaf nodes in the tree structures stand in for characteristics, while the branches represent the conjunctions of features that result in the classifications (also known as labels). It is simple to create a decision tree that matches the data set provided. The difficulty lies in creating effective decision trees, which usually entails using the fewest possible decision trees. Overfitting pruning can be applied to inhibit the tree from being overfitted only for the training set. This technique makes the tree general for unlabeled data and can tolerate some mistakenly labeled training data (Tan, 2015, p. 17).

Following are some of the clear benefits of employing decision trees in various classification and prediction applications, as well as some typical difficulties. First, it requires little user effort for data preparation and is simple to understand and analyze for non-technical users. Missing values in training data won't prevent data splitting for tree building, which is another characteristic that helps. Since the partitioning is based on the percentage of samples that fall inside the split ranges rather than on

absolute values, decision trees are also less vulnerable to outliers. Additionally, the performance of trees is unaffected by nonlinear correlations between parameters. Without proper pruning or restrictions on tree development, they tend to overfit the training set, which may make them rather unreliable predictors (Kotu & Deshpande, 2015).

Neural Networks

Artificial Neural Networks (ANN) architecture is based on the structure and function of the biological neural network. The neurons of ANN are arranged in several layers, just like the neurons in the brain (Sairamya et al., 2019). The input layer, output layer, and hidden layer are the layers that constitute an ANN. The nodes in the input layer must be connected to the nodes in the hidden layer, and the nodes in the hidden layer must be connected to the nodes in the output layer. The network provides the data to the input layer. The input layer sends the raw data to the hidden layer, which processes it. The output layer receives the obtained value and processes the data from the hidden layer before producing the output (Imran & Alsuhaibani, 2019). An artificial neural network technique with several layers is the multi-layer perceptron (MLP). Clearly, linear issues can be solved in a single perceptron, but nonlinear examples are not well suited to it. MLP can be used to resolve these challenging issues (Mohanty & Mohanty, 2022). MLPs may be trained to use any given nonlinear input-output mapping since they are global approximators. Moreover, it has been demonstrated mathematically that even a single hidden-layer MLP may approximate the mapping of any continuous function. As with all ANN, the input layer's number of neurons is determined by the input vector's dimension, and the output layer's number of neurons is determined by the number of classes that need to be learned. It is necessary to empirically determine the number of selected hidden layers and the number of neurons in each layer. Fewer neurons are selected for the hidden layers than for the input layer. But there is a cost associated with the number of neurons: Overtraining results from too many, whereas generalization skills are hampered by too few. Usually, the number of neurons can be gradually decreased after acceptable learning has been accomplished. Later retraining of a smaller network performs significantly better than initial training of the more sophisticated network (Meyer-Baese & Schmid, 2014).

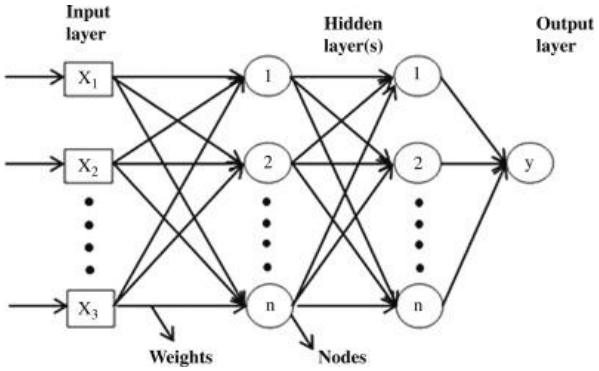

Figure 3 - General structure of an ANN (Source: Kumari & Swarnkar, 2021)

Support Vector Machine

Support vector machines (SVMs) are a group of linked supervised learning techniques that are well-liked for employing data analysis and pattern recognition for classification and regression analysis. The classifier's properties and structure influence the different methods. The most well-known SVM is a

linear classifier that determines which of two possible classes each input belongs to. A support vector machine creates a hyperplane or set of hyperplanes to categorize all inputs in a high-dimensional or even infinite space, according to a more precise specification. Support vectors are the values that are most closely related to the classification margin. The SVM aims to maximize the space between the support vectors and the hyperplane (Gove & Faytong, 2012).

Both linear and nonlinear SVM are possible. The two types of linear SVM are hard margin and soft margin. The data is entirely linearly separable by a hyperplane in hard-margin SVM, but not in soft-margin SVM. Positive slack variables are added in soft margin SVM to signify a penalty for data points that are on the wrong side of the margin border. The slack variable's value rises as its separation from the margin widens. Nonlinear SVM is utilized when the data is nonlinear, and training sets are employed to translate the original input space to a higher-dimensional feature space. The radial basis, polynomial, and sigmoidal are the ones that are more used. SVM is more robust to outliers and noise. Additionally, it provides several choices for the model parameters and kernels, which need to be tested to select the best SVM. It operates well under high-dimensional spaces. Since it is a black box algorithm, it has the drawback of being difficult to understand and interpret the final SVM model, variable weights, and individual impact (Rani et al., 2022).

3.6. EVALUATION

Precision metric

Different models were built to solve the problem studied here, but to see which was best to be applied, we chose the metric precision to compare the models. Because there were so many customers who did not churn, the high number of true negatives made the accuracy very high, which would be misleading since we wanted to focus on those customers that did leave and try to discover what we could do to prevent that from happening with other customers. Precision came in handy here. Precision is: out of all customers who were labeled as "churn," how many did we correctly label as such? This is the ratio of customers who were correctly labeled (true positives) to all customers who were labeled as "churn" (true positives and false positives). This metric made more sense for us to review both before and after handling the imbalance in the data.

In a broad sense, precision can be considered a metric that counts the number of correctly produced positive predictions. Therefore, precision determines the accuracy for the minority class. It is determined by dividing the total number of correctly predicted positive examples by the ratio of correctly predicted positive examples (Brownlee, 2020).

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + FalsePositives\ (FP)}$$

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positives (TP)	False Negative (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Table 1 - Precision and Confusion matrix

ROC Curve

The ROC curve illustrates the trade-off between specificity and sensitivity. Better performance is shown by classifiers that provide curves that are closer to the top-left corner. A random classifier is anticipated to provide points that are diagonal by default (FPR = TPR). ROC analysis is a useful method for assessing diagnostic tests and predictive models. It can be applied to quantitatively evaluate the accuracy or to compare the accuracy between various tests or forecasting models (Zou et al., 2007).

Cumulative Lift

For assessing the effectiveness of scoring models, lift analysis is frequently used (Piatetsky-Shapiro & Masand, 1999). The process is as follows (Hughes, 1995): The training dataset is used to create a prediction model, which calculates a score for each testing instance to indicate how likely it is that it falls into a particular category. Plotting the proportion of sample size versus the cumulative % of targets yields the lift cumulative curve. Intuitively, a lift cumulative curve with a larger area indicates better prediction accuracy, while one with a lift cumulative curve area of 0.5 shows performance closer to a random baseline.

3.7. DEPLOYMENT

The deployment of the model and a plan for its monitoring and upkeep are then planned. The model will finally be incorporated into the business's marketing plan after a thorough examination, which will aid it in raising its customer retention rate. It was chosen to deliver two distinct automated procedures for this section, which are detailed below. Python is the main application utilized in both.

Trimonthly predictions

The first deliverable, created as follows, intends to provide quarterly projections for the month's probable churners.

Using SAS Enterprise Guide, the data is gathered for a specific month and year. Python makes use of the final output table to reproduce all essential data preparation procedures in this dataset. The number of customers who will churn is uncertain at this point in the process. The probability that a consumer will leave is then predicted using the trained model.

Model performance monitoring

The second deliverable provides a method for continuously evaluating and further integrating the model's performance into a dashboard.

This procedure is comparable to the trimonthly forecasting procedure previously mentioned. Understanding the model's health requires comparing previous forecasts with actual events (whether consumers were lost as a result). Because of this, the last six available months' worth of data collection (in SAS Enterprise Guide), data preparation, and prediction (in Python) are all made available (considering the current date with three-month aging).

The output of the model is ultimately contrasted with the actual target. If indications of model aging appear, the model should be updated.

4. RESULTS

Below, the results from the modeling phase will be presented. The different models constructed mentioned in this section can be found in Appendix, along with the considered parameters, the used data scaling, and the percentage of the positive outcome in the dataset.

Predicting a class label is a typical aspect of classification predictive modeling. But before they can be mapped to a clear class label, many machine learning algorithms are capable of predicting a probability or scoring of class membership. This is accomplished by utilizing a threshold, such as 0.5, where all values that are equal to or higher than the threshold are mapped to one class, while all other values are mapped to a different class. The default threshold may not perform well for classification issues with significant class imbalance. As a result, changing the threshold used to translate probabilities to class labels is an easy and straightforward way to improve the performance of a classifier that predicts probabilities on an imbalanced classification task.

Considering our problem, which is the churn in banking or, more specifically, the closing of bank accounts, we have a low number of positive outcomes, since this is not a phenomenon that happens a lot, having in that way imbalanced data. With that in mind, since we have few churn cases, using the default threshold of 0.5 might not be the best scenario for the situation we are trying to solve.

To analyze the scores that we get as a result of the models, we might look at the cumulative lift, the cumulative captured response and the response rate by ventiles. The cumulative percentage of respondents who are captured in a ventile is known as the cumulative % of captured responses. The lift is the proportion of each ventiles' captured response percentage to the baseline response percentage. The response rate is the number of churners divided by the number of people who make up the ventil. So, for each of the four models, we will analyze the scores, by calculating the response rate and lift to each model to compare them and understand how the churners, or the customers that have a target equal to 1, distribute along the ventiles. Since we have a low number of churners, it makes sense to analyze the values in the top 5% and 15%. Looking at the table below the hypothesis mentioned before is confirmed because we can see that we have captured most of our churners in the top 15%. Consequently, the default threshold of 0.5 is not the best for our case. It is necessary to see what value of threshold we would get better values.

	Support Vector Machine	**Logistic Regression**	**Decision Trees**	**Neural Networks**
Cumulative Lift at 5%	13.76	7.23	8.65	11.95
Cumulative Lift at 15%	6.24	4.38	4.69	5.50
Cumulative Captured Response 5%	69%	36%	43%	60%
Cumulative Captured Response 15%	94%	66%	70%	83%
Response Rate 5%	37.2%	19.6%	23.4%	32.3%
Response Rate 15%	3.2%	5.7%	3.2%	3.8%

Table 2 - Score analysis

Therefore, to choose the right threshold according to the characteristics of our data, we analyzed the results obtained by the scores and the discrimination threshold. The probability or score at which the positive class is preferred to the negative class is known as the discrimination threshold. This is typically set at 50%, but the threshold can be changed to alter its sensitivity to false positives or other application-specific issues. In the discrimination threshold, we have the visualization of precision, recall, f1 score, and queue rate of a binary classifier. We can use the discrimination threshold, depending on the priorities of the issue we are trying to solve. Also, we can decrease the likelihood of false positives or false negatives by varying the discrimination threshold. However, doing so might degrade the performance of the classification, as shown by metrics like accuracy or the F1 score. The values of the lower and upper bounds are the 10th and 90th percentiles.

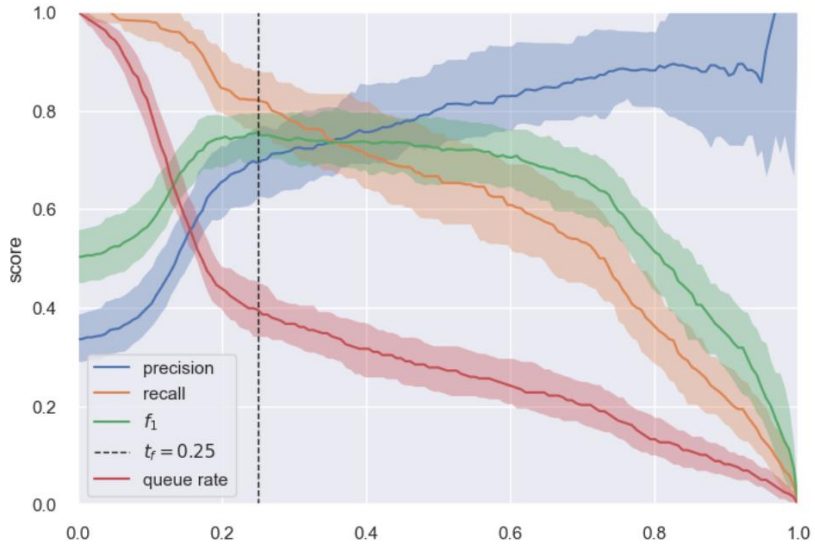

Figure 4 - Discrimination threshold

Therefore, as we can see by the results of the discrimination threshold presented in the figure above, to produce better results, we should consider defining the threshold to 0.75 since the probability at which the positive class is preferred to the negative class is equal to 0.25, meaning we have the churners in the top 25%. The threshold of 0.75 is the best value to use in our models to classify the churners.

To determine the best model to apply to the situation presented, we must compare the models of study with the metrics chosen to evaluate their performance, in this case, the precision and the area under the curve.

	Support Vector Machine	**Logistic Regression**	**Decision Trees**	**Neural Networks**
Precision	0.84	0.78	0.72	0.75
AUROC	0.905	0.836	0.805	0.884

Table 3 - Evaluation results

Based on the results obtained, the best model to apply to our situation would be the Support Vector Machine model with a precision of 0.84 and an AUROC of 0.905.

Next, we have the feature importance that describes the method that rates input features according to how well they can predict a given target variable and, in that way, the features that most contribute to our model are in the following figure.

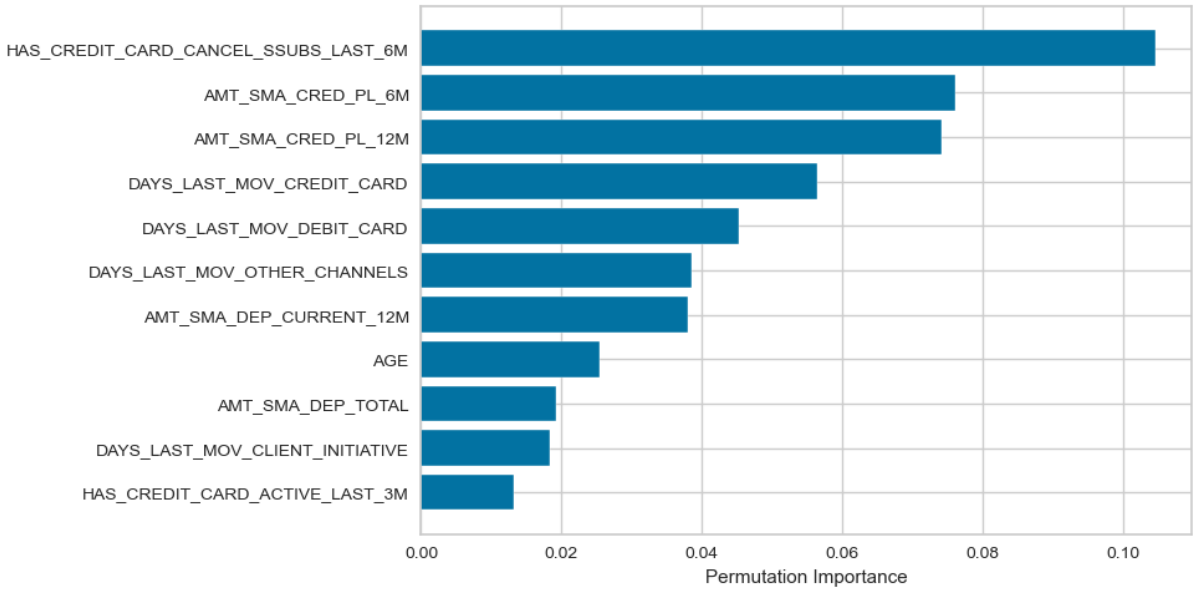

Figure 5 - Feature importance of SVM

We can observe that the utilization and the amount available of credit cards are among the top 3 most important features of the Support Vector Machine. By looking at the most important variables of the model, we can understand that these variables demonstrate the signals of the voluntary churn of the customer. For example, an indicator is the possession of an active card in the last months of analysis; discontinuing the use of the credit card can indicate the customer's intentions of leaving. Also, the study of the days of the last time the customer made a movement either with the card, used the bank channels, or proceeded to make a transfer, we can begin to understand if this is a customer who is thinking of closing the account. In these features, we can see the signs that a customer is considering closing the bank account, so the bank must pay attention to them to create proactive measures to prevent churn.

5. DISCUSSION

Customer Churn Management is one of the firm's core goals. Consistently strong retention rates can provide the organization with a competitive edge, highlight unseen issues, and offer new perspectives on the industry. These insights may affect how marketing strategies are made. It can help explain why this churn tendency exists in addition to providing a quantitative measure of the probability.

The project's main objective was to comprehend the actions of consumers who cancel their current bank accounts. In other words, we want to understand the comportment of old churners to get a pattern and predict future churners. By knowing whom the people that will close their accounts are the bank can establish measures to prevent that event. When we get as the result of the churn model prediction who are the customers most likely to churn, one strategy would be to have the agencies contact those customers and discuss with them to convince them to stay. Contacting the customers might be a good idea since they will feel valued by the bank, and that they are important, making them less likely to leave.

To get a solution to our problem, machine learning techniques were applied. The historical behavior of the customer can be used to create and train a classification model. The customer's habits can be automatically modeled by this model and then obtain the classification of is a churner or not. Information on the customer, the financial products they own, and the relationship between the bank and the customer is provided to develop the model.

To this study, it was defined a timeline that help in the characterization of the dependent and independent variables. The time period is split into three parts: the observation window, which runs from July 2020 to June 2021, is the period during which independent variables are generated. Following that, the buffer period is when the bank will put the prediction model into action to identify the customers who are most likely to leave in the coming months and share that information with the local agencies so they can get in touch with those customers and persuade them to stay with the bank. The buffer period is one month long, or July 2021 in our situation. Finally, the performance window is the time frame from which the target is derived, which runs from August to October 2021.

Understanding the company itself, its objectives, and which variables might offer useful information regarding this customer's behavior were the focus of the first section of the research. About 110 numerical and categorical variables were taken from the company's database. This stage was crucial for comprehending some of the procedures that should be done in the Data Preparation phase, where development was done with raw data. Here, tasks include removing redundant features, dealing with missing values, generating new characteristics, and preparing the data such that it would enable (and aid) the modeling phase. The next step is feature selection after the dataset has been prepared. Given the number of rows available for modeling, the objective was to use an appropriate number of features. More characteristics signify a more complex problem, which calls for more data for models to comprehend. A variety of techniques are combined with choosing the features. The Top N features are chosen based on the features that show up most frequently in the various techniques. The classification problem is challenging to complete when dealing with an imbalanced dataset, thus, undersampling and smote were applied to this situation. The default threshold of 0.5 means that all values that are equal to or higher than the threshold are mapped to one class, while all other values are mapped to a different class. With a severe class imbalance, the default threshold might not work

properly. A threshold of 0.75 would therefore be the best to employ, according to our analysis of the cumulative lift, response rate and the cumulative captured response by ventiles, and the discrimination threshold. It was necessary to construct a suitable trade-off between the percentage of detected churn customers and accurately recognized churn customers. Different models were investigated, paying close attention to the tuning of their parameters, and being able to explain their processes as much as possible. Based on measures employed to assess the models, the final model was chosen. The Support Vector Machine, with a precision of 0.84 and an AUROC of 0.905, was the final model chosen.

In addition, we could analyze that the most important aspects to investigate when concerning customer churn are the customer's age, the indication of what type of cards he/she possesses, the frequency of activities done by the customer initiative, and the amounts of funds that are in the accounts. All the information gathered here will not only contribute to the bank predicting which customers will close their accounts but also this data will be of extreme importance for the calculation of the customer's lifetime value. In other words, by calculating individual customer churn probabilities, then we can have more information to add to this measure.

5.1. IMPLICATIONS FOR BUSINESS

Churn is a metric for how many people leave an organization over a predetermined period of time. It refers to the number of customers who ceased utilizing the bank's service over a specific time period. It's never good for business when customers leave because it costs more to get new customers than it does to keep old ones.

However, not all churn is negative. Every business has unprofitable customers, and losing one is not nearly as bad as losing one of your top customers. Knowing a customer's lifetime worth will enable to assist the marketing team in determining how much to invest in each customer's retention. Simply put, spending more money to keep a customer than they are worth to the company over their lifetime is not prudent.

The bank needs to be able to predict churn rates with precision since it gives the bank a better idea of the revenue it may anticipate in the future. Additionally, churn prediction gives the ability to target a specific customer in an effort to keep them from cancelling their checking account when you can predict their likely churn rate. Predicting attrition rates can also assist your company in identifying and strengthening its weak points in customer service. You can lower turnover and increase revenue by making these enhancements.

Understanding how certain customer behaviors and characteristics affect the risk and timing of customer attrition is the goal of churn prediction. The precision of the technique utilized has a significant impact on how accurately a customer churn prediction is made. Older attrition analysis techniques concentrated on quantifying static risk-based data and measures, such as information describing a customer's current status, rather than future predictions. Accurate churn statistics would lessen the need for businesses to take a chance at losing consumers. Utilizing the customer lifetime value computation for each customer, more up-to-date and accurate attrition analysis models can forecast future consumer behavior. This would provide businesses with the flexibility to adapt to best practices that would boost customer retention. Some prediction systems can assist in identifying customers whose lifetime value has significantly decreased recently and who may churn in the future.

Companies can implement focused, proactive retention and churn prevention strategies thanks to this type of analysis.

5.2. IMPLICATIONS FOR THEORY

A company's marketing efforts are focused on keeping current customers rather than just recouping old churners because it costs five to twenty times more to attract a new customer than it does to keep an existing one. Because of that is important to predict and identify possible churners of the bank.

The advent of digital technology has made it simpler for customers to move their business elsewhere if they so want. Therefore, it is in any company's best interest to keep an eye on its customers' behavior to perhaps spot any signs of displeasure that could result in churning. By analyzing the most relevant features, we get to know the signs indicating that a customer is trying to close the account, such as starting by stopping using the credit card or a decrease in the average balance of the account. In our analysis, we discovered the indicators that the bank should be alert to, that is, we can see the signs that reflect the beginning of the customer's intention of leaving.

Reinartz & Kumar, 2003 concluded that long-term relationships with customers allow businesses to focus on meeting their requirements, which is more profitable than finding new customers. In our research, the seniority of the customers studied showed that in our situation, there is no relation between closing an account and how long the customer is in the bank.

According to de Lima Lemos et al., 2022, consumers who have more products and services and who borrow more have stronger relationships with the bank and are less likely to close their checking accounts. In our research, the number of products and services that a customer has is not a variable relevant when determining the probability of a customer being a churner.

As stated by Deng et al., 2021, older customers tend to spend more and, when happy, recommend the business to others because they are more familiar with it. An analysis of age characteristics using a single component found that low-age customers had higher user churn rates. While simultaneously retaining older customers through the following offline activities, banks might host some online events to persuade young people to keep choosing their banks. In our study, most churners were between 30 and 50 years old and it corroborates with the beforementioned research, meaning that the bank should consider an approach that uses the traditional channels for older people and the promotion of the digital channels for younger people.

6. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

This report demonstrates the work during the six-month internship in a Portuguese bank, and below is the summary with conclusions, limitations and future work.

A customer who closes all of his accounts and stops doing business with the bank is known as a churned customer. As a result, customer-centric retention looks for and identifies customers that have a high propensity to leave the company or predict customer attrition in order to lower churn. The main goal of this project was to predict the customer churn of the checking accounts from July 2020 to October 2021. Besides that, we wanted also to understand and discover the indication that the customer is leaving the bank. The analysis started by first defining the buffer period and time frame for independent and dependent (target) variables. Then, we collected all the data needed from different sources and they were transformed using the software SAS Enterprise Guide. The remainder of the research was done in python, using the Anaconda Navigator software. The following phase is the data preparation where we deal with the missing values and outliers, created new variables in feature engineering and chose the most relevant features. Since the data was imbalanced it was applied the smote and the undersampling technique. In the modeling phase, it was used four models (Decision Tree, Logistic Regression, Artificial Neural Network and Support Vector Machine). The scores were analyzed using the discrimination threshold, the response rate, the cumulative lift and the cumulative captured response and we concluded that a threshold of 0.75 was the best for the situation in the study. With that in mind, using the Precision and the AUROC metrics, we concluded that the Support Vector Machine, with a Precision of 0.84 and an AUROC of 0.905 was the best model. Also, we observed that the utilization and the amount available of credit cards are among the top 3 most important features.

Since the outcomes in terms of performance indicators may be higher, it is possible to see that the supplied solution has some room for improvement. Results are greatly reliant on the data used, and thus, some enhancements or other strategies, such as the testing of various classification algorithms, could be tried. One point would be to be able to gather more data and find hidden elements. Additional details about the consumer, such as their salary, as well as any interactions they may have had with the business to buy various financial products, could be helpful. Other predictive models and more advanced parameter tuning methods, particularly for artificial neural networks, should be investigated. Additionally, the pre-processing stage might be improved, more accurate missing values imputation methods could be taken into consideration, and different kinds of selection methods could be used.

A recommendation for future work would be to consider a rolling performance window. Although it implies that it will take several windows to create a model, the performance window's duration is fixed. Since consumer behaviour and characteristics change occasionally, understanding seasonality throughout this way might be beneficial. For instance, a specific period's churn rate is 3%. The other period could see an increase or decrease. There might be some seasonality involved. We presume that the variables are constant over time when we take a single performance window. We can simulate seasonality when we take multiple performance windows. Rolling performance windows also allow us to integrate Multiple Campaigns. Campaign data from several periods should be taken into consideration while constructing a campaign response model.

In addition, it would be of interest to add to our analysis the financial costs associated with customer churn, like the cost and its implications of losing a customer or the profit contributed by having a predictive model.

Lastly, as was possible to observe, some models present better results than others. Therefore, it could be interesting to explore other models. Three models that would be interesting to test are LightGBM (from Python's LightGBM package), lattice-based models (from Python's TensorFlow-Lattice package), and the Extreme Gradient Boosting model (from the xgboost package).

7. REFERENCES

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6

Amirteimoori, A., & Kordrostami, S. (2010). A Euclidean distance-based measure of efficiency in data envelopment analysis. *Optimization*, *59*(7), 985–996. https://doi.org/10.1080/02331930902878333

Amor, N. B., Benferhat, S., & Elouedi, Z. (2006). Qualitative Classification with Possibilistic Decision Trees. Em B. Bouchon-Meunier, G. Coletti, & R. R. Yager (Eds.), *Modern Information Processing* (pp. 159–169). Elsevier Science. https://doi.org/10.1016/B978-044452075-3/50014-5

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, *17*(1), 106. https://doi.org/10.1186/s12955-019-1181-2

Belyadi, H., & Haghighat, A. (2021). Chapter 5—Supervised learning. Em H. Belyadi & A. Haghighat (Eds.), *Machine Learning Guide for Oil and Gas Using Python* (pp. 169–295). Gulf Professional Publishing. https://doi.org/10.1016/B978-0-12-821929-4.00004-4

Brownlee, J. (2020, janeiro 2). How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification. *Machine Learning Mastery*. https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/

Calabrese, B. (2019). Data Cleaning. Em S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 472–476). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20458-5

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chitra, K., & Subashini, B. (2011). *Customer retention in banking sector*. 4.

Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer defection: A study of the student market in Ireland. *International Journal of Bank Marketing*, *14*(3), 23–29. https://doi.org/10.1108/02652329610113144

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

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009

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

Definition of Predictive Modeling—Gartner Information Technology Glossary. (sem data). Gartner. Obtido 3 de maio de 2022, de https://www.gartner.com/en/information-technology/glossary/predictive-modeling

Deng, Y., Li, D., Yang, L., Tang, J., & Zhao, J. (2021). Analysis and prediction of bank user churn based on ensemble learning algorithm. *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, 288–291. https://doi.org/10.1109/ICPECA51329.2021.9362520

Dong, G., & Liu, H. (2018). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 140. https://doi.org/10.1186/s40537-021-00516-9

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the Customer Base of Service Providers: An Examination of the Differences between Switchers and Stayers. *Journal of Marketing*, *64*(3), 65–87. https://doi.org/10.1509/jmkg.64.3.65.18028

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, *72*(7), 1483–1493. https://doi.org/10.1016/j.neucom.2008.11.026

Garvey, J., Sullivan, B., Alcocer, J., & Eldridge, A. (2014). *Retail Banking 2020: Evolution or Revolution?* PricewaterhouseCoopers LLP. https://www.pwc.com/gx/en/banking-capital-markets/banking-2020/assets/pwc-retail-banking-2020-evolution-or-revolution.pdf

Gove, R., & Faytong, J. (2012). Chapter 4 - Machine Learning and Event-Based Software Testing: Classifiers for Identifying Infeasible GUI Event Sequences. Em A. Hurson & A. Memon (Eds.), *Advances in Computers* (Vol. 86, pp. 109–135). Elsevier. https://doi.org/10.1016/B978-0-12-396535-6.00004-1

Gudivada, V. N., Irfan, M. T., Fathi, E., & Rao, D. L. (2016). Chapter 5 - Cognitive Analytics: Going Beyond Big Data Analytics and Machine Learning. Em V. N. Gudivada, V. V. Raghavan, V. Govindaraju, & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 35, pp. 169–205). Elsevier. https://doi.org/10.1016/bs.host.2016.07.010

Hassonah, M. A., Rodan, A., Al-Tamimi, A.-K., & Alsakran, J. (2019). Churn Prediction: A Comparative Study Using KNN and Decision Trees. *2019 Sixth HCT Information Technology Trends (ITT)*, 182–186. https://doi.org/10.1109/ITT48889.2019.9075077

Hughes, A. (1995). *The Complete Database Marketer: Second Generation Strategies and Techniques for Tapping the Power of Your Customer Database*. McGraw-Hill Companies,Incorporated.

Imran, M., & Alsuhaibani, S. A. (2019). Chapter 7—A Neuro-Fuzzy Inference Model for Diabetic Retinopathy Classification. Em D. J. Hemanth, D. Gupta, & V. Emilia Balas (Eds.), *Intelligent Data Analysis for Biomedical Applications* (pp. 147–172). Academic Press. https://doi.org/10.1016/B978-0-12-815553-0.00007-0

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

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. *2019 International Workshop on Big Data and Information Security (IWBIS)*, 33–38. https://doi.org/10.1109/IWBIS.2019.8935884

Kaur, I., & Kaur, J. (2020). Customer Churn Analysis and Prediction in Banking Industry using Machine Learning. *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 434–437. https://doi.org/10.1109/PDGC50313.2020.9315761

Khurana, U., Turaga, D., Samulowitz, H., & Parthasrathy, S. (2016). Cognito: Automated Feature Engineering for Supervised Learning. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 1304–1307. https://doi.org/10.1109/ICDMW.2016.0190

Koli, H. (2020, outubro 10). Checking Account Churning Prediction in BFSI Domain. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/10/the-complete-guide-to-checking-account-churn-prediction-in-bfsi-domain/

Kotu, V., & Deshpande, B. (2015). Chapter 4—Classification. Em V. Kotu & B. Deshpande (Eds.), *Predictive Analytics and Data Mining* (pp. 63–163). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-801460-8.00004-5

Kotu, V., & Deshpande, B. (2019). Chapter 2—Data Science Process. Em V. Kotu & B. Deshpande (Eds.), *Data Science (Second Edition)* (pp. 19–37). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-814761-0.00002-2

Kumari, B., & Swarnkar, T. (2021). Stock movement prediction using hybrid normalization technique and artificial neural network. *International Journal of Advanced Technology and Engineering Exploration*, *8*. https://doi.org/10.19101/IJATEE.2021.874387

Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering*, *28*(9), 1635–1647. https://doi.org/10.1016/j.compchemeng.2004.01.009

Mack, C., Su, Z., & Westreich, D. (2018). Types of Missing Data. Em *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [Internet]*. Agency for Healthcare Research and Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK493614/

Maillo, J., Ramírez, S., Triguero, I., & Herrera, F. (2017). kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data. *Knowledge-Based Systems*, *117*, 3–15. https://doi.org/10.1016/j.knosys.2016.06.012

Maletic, J. I., & Marcus, A. (2000). *Data Cleansing: Beyond Integrity Analysis*.

Malik, H., Fatema, N., & Iqbal, A. (2021). Chapter 1—Advances in Machine Learning and Data Analytics. Em H. Malik, N. Fatema, & A. Iqbal (Eds.), *Intelligent Data-Analytics for Condition Monitoring* (pp. 3–29). Academic Press. https://doi.org/10.1016/B978-0-323-85510-5.00001-6

Meyer-Baese, A., & Schmid, V. (2014). Chapter 7—Foundations of Neural Networks. Em A. Meyer-Baese & V. Schmid (Eds.), *Pattern Recognition and Signal Analysis in Medical Imaging (Second Edition)* (pp. 197–243). Academic Press. https://doi.org/10.1016/B978-0-12-409545-8.00007-8

Mohanty, M. D., & Mohanty, M. N. (2022). Chapter 5—Verbal sentiment analysis and detection using recurrent neural network. Em S. De, S. Dey, S. Bhattacharyya, & S. Bhatia (Eds.), *Advanced Data Mining Tools and Methods for Social Computing* (pp. 85–106). Academic Press. https://doi.org/10.1016/B978-0-32-385708-6.00012-6

Mutanen, T., Nousiainen, S., & Ahola, J. (2010). Customer churn prediction – a case study in retail banking. *Data Mining for Business Applications*, 77–83. https://doi.org/10.3233/978-1-60750-633-1-77

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, *43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Nitzan, I., & Libai, B. (2011). Social Effects on Customer Retention. *Journal of Marketing*, *75*(6), 24–38. https://doi.org/10.1509/jm.10.0209

Patel, H. (2021, setembro 2). *What is Feature Engineering—Importance, Tools and Techniques for Machine Learning*. Medium. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

Piatetsky-Shapiro, G., & Masand, B. (1999). Estimating campaign benefits and modeling lift. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 185–193. https://doi.org/10.1145/312129.312225

Rani, A., Kumar, N., Kumar, J., Kumar, J., & Sinha, N. K. (2022). Chapter 6—Machine learning for soil moisture assessment. Em R. C. Poonia, V. Singh, & S. R. Nayak (Eds.), *Deep Learning for Sustainable Agriculture* (pp. 143–168). Academic Press. https://doi.org/10.1016/B978-0-323-85214-2.00001-X

Reinartz, W. J., & Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*, *67*(1), 77–99. https://doi.org/10.1509/jmkg.67.1.77.18589

Rosales-Pérez, A. (2022). Chapter 20—A review on machine learning techniques for acute leukemia classification. Em A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, & O. Mendoza-Montoya (Eds.), *Biosignal Processing and Classification Using Computational Learning and Intelligence* (pp. 429–446). Academic Press. https://doi.org/10.1016/B978-0-12-820125-1.00033-6

Sacchi, M., Sezer, B., & Lednarova, J. (2021, março 25). World Retail Banking Report 2021. *Capgemini Worldwide*. https://www.capgemini.com/news/world-retail-banking-report-2021-to-create-new-value-banks-can-adopt-banking-as-a-service-to-embed-finance-in-consumer-lifestyles/

Sagala, N. T. M., & Permai, S. D. (2021). Enhanced Churn Prediction Model with Boosted Trees Algorithms in The Banking Sector. *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 240–245. https://doi.org/10.1109/ICoDSA53588.2021.9617503

Sairamya, N. J., Susmitha, L., Thomas George, S., & Subathra, M. S. P. (2019). Chapter 12—Hybrid Approach for Classification of Electroencephalographic Signals Using Time–Frequency Images With Wavelets and Texture Features. Em D. J. Hemanth, D. Gupta, & V. Emilia Balas (Eds.), *Intelligent Data Analysis for Biomedical Applications* (pp. 253–273). Academic Press. https://doi.org/10.1016/B978-0-12-815553-0.00013-6

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A Proposed Churn Prediction Model. *International Journal of Engineering*, *2*(4), 6.

Suthar, B., Patel, H., & Goswami, A. (2012). *A Survey: Classification of Imputation Methods in Data Mining*.

Tan, L. (2015). Chapter 17—Code Comment Analysis for Improving Software Quality**This chapter contains figures, tables, and text copied from the author's PhD dissertation and the papers that the author of this chapter coauthored [[3], [1], [35], [7]]. Sections 17.2.3, 17.4.3, 17.5, and 17.6 are new, and the other sections are augmented, reorganized, and improved. Em C.

Bird, T. Menzies, & T. Zimmermann (Eds.), *The Art and Science of Analyzing Software Data* (pp. 493–517). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-411519-4.00017-3

Tolles, J., & Meurer, W. (2016). Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, *316*, 533. https://doi.org/10.1001/jama.2016.7653

Ullah, I., Hussain, H., Ali, I., & Liaquat, A. (2019). Churn Prediction in Banking System using K-Means, LOF, and CBLOF. *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 1–6. https://doi.org/10.1109/ICECCE47252.2019.8940667

Usmani, R. S. A., Binti Wan Azmi, W. N. F., Abdullahi, A. M., Hashem, I. A. T., & Pillai, T. R. (2020). A novel feature engineering algorithm for air quality datasets. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(3), 1444. https://doi.org/10.11591/ijeecs.v19.i3.pp1444-1451

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505–513. https://doi.org/10.1016/j.ejor.2014.04.001

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3, Part 1), 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*, *115*(5), 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

APPENDIX

Table with the models' parameters, the used data scaling and the percentage of churn.

Model	Algorithm	Parameters	Scaling method	Churn rate
LR	Logistic Regression		Min-Max	3%
MLP	Multilayer Perceptron	activation: 'tanh', alpha: 0.001, hidden_layer_sizes: (10, 10, 10), learning_rate: 'invscaling', solver: 'lbfgs'	Min-Max	3%
DT	Decision Tree	class_weight: None, criterion: 'entropy', max_depth: 9, max_features: None, min_samples_split: 0.05, splitter: 'best'	Min-Max	3%
SVM	Support Vector Machine	kernel = 'rbf', gamma = 'scale', random_state=5, probability = True, C=50	Min-Max	3%

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

Universidade Nova de Lisboa