Erasmus School of Economics

Bachelor Thesis

Erasmus University Rotterdam

International Bachelor Economics and Business Economics

Integrating Economic Theories and Machine Learning for Effective Customer Churn Prediction in the Portuguese Banking Sector

Student: **N.T.S. Billar** **571062**

Supervisor: **B. Georgievski**

Second Assessor: **F. Prins**

Date Final Version: **15-08-2023**

Keywords: Customer churn prediction, Machine Learning, Banking

Table of Contents

1 Executive Summary

In the banking sector, customer churn research has seen a surge in machine learning utilization for predictive purposes. However, economic theories underlying churn have received relatively less attention compared to machine learning techniques. This polarization has left the integration of these dimensions unexplored, especially in the context of the Portuguese banking sector. To address this research gap, our study aims to integrate economic theories with machine learning methodologies to predict customer churn effectively in a Portuguese banking institution. The central research question guiding our investigation is:

'In what ways can the integration of economic theories and machine learning methodologies contribute to the understanding and prediction of customer churn in the context of a Portuguese banking institution?'

Our research begins by exploring supervised machine learning for predicting customer churn and its application in addressing marketing challenges. Empirical studies on churn prediction methods are then reviewed to uncover techniques used in understanding customer behavior. Our focus then narrows to banking sector studies, aiming to reveal effective churn prediction techniques in this competitive and dynamic industry. We found that prior research widely adopted supervised machine learning models to predict customer churn, which validates our use of the models. Moreover, past research revealed data cleaning and preparation methods, effective machine learning models, and evaluation metrics that will be adopted in our research.

After exploring the machine learning side of customer churn, we explore economic theories that are relevant for churn and apply them in conjunction with variables from the dataset gathered from the University of California Irvine machine learning repository, which encompasses variables from a Portuguese bank. The economic theories and variables from the dataset are used to formulate hypotheses that address the main research question:

Hypotheses	**Economic Theories**
Customers who previously had a term deposit at the bank exhibit lower churn rates in the current marketing campaign.	Switching Cost Theory
Customers who have a longer contact duration exhibit lower churn rates in the current marketing campaign.	Relationship Marketing Theory
Customers who already had a financial product at the bank exhibit reduced churn rates in the current marketing campaign.	Perceived Value Theory

Our research approach involves an in-depth exploratory data analysis, enabling us to perform statistical tests and gain valuable insights from the dataset. Subsequently, we employ various machine learning models, including Logistic Regression, Stochastic Gradient Descent, Random Forest, and Extreme Gradient Boosting, to predict customer churn and to see which features have the most predictive capability. The outcomes of the machine learning models, where our best performing model has an AUC score of 0.9510, combined with the knowledge derived from the exploratory data analysis, lead us to validate our hypotheses. Notably, we discover that customers who had a previous term deposit at the bank exhibit lower churn rates during the current marketing campaign, aligning with the switching cost theory explored in our research. Additionally, we find that customers with longer contact duration demonstrate lower churn rates in the current marketing campaign, corroborating the relationship marketing theory. Furthermore, customers who already possess a financial product with the bank exhibit reduced churn rates in the current marketing campaign, in line with the perceived value theory.

The successful validation of these hypotheses, in conjunction with the statistical significance of the variables and their relative importance in our accurate machine learning models, demonstrates the power of integrating economic theories and machine learning methodologies. The economic theories allowed us to forge hypotheses and reason further past the statistical observations and machine learning insights, granting us the framework to further explain the phenomenon of customer churn within a Portuguese banking institution.

The integration of economic theories with machine learning enhances strategic decision-making for banks, by allowing them to proactively mitigate churn threats, ensuring targeted retention and strategic resource allocation in their competitive market. The symbiotic relationship of economic theory and machine learning becomes a beacon of guidance for boards of companies, illuminating the path for informed strategic planning, adept product development, and calibrated pricing strategies. Moreover, this synergy fosters a profound understanding of customer behaviors from a marketing perspective, facilitating the design of targeted marketing campaigns, sophisticated segmentation processes, and personalized service offerings.

The findings of our research are derived from a specific dataset from a single Portuguese banking institution, which constrains the generalizability of the conclusions. To bolster the external validity and robustness of the findings, future research should encompass data from multiple banks spanning different countries. Furthermore, the integration of more diverse economic theories explaining churn and advanced machine learning models in forthcoming research holds the potential to further heighten the accuracy and explanatory power of the findings. This expanded approach would allow for a more comprehensive examination of customer churn dynamics across a range of contexts, ultimately contributing to the broad applicability of the findings.

2 Introduction

In the evolving economic landscape, the centrality of customer retention for organizational success is unequivocally recognized. Businesses, particularly in the financial service sectors like banking, thrive on the longevity of their customer relationships (Reichheld & Sasser, 1990). In their Harvard Business Review article, Reichheld and Sasser highlighted the importance of customer retention. They were ahead of their time with their analysis on the extra cost that are incurred when customers leave firms.
Consequently, customer churn - the phenomenon of customers ending their relationship with a business - poses a significant concern due to its profound financial implications such as eroding revenue streams, increasing customer acquisition costs, and potential harm to a firm's reputation and future growth prospects (Van Den Poel & Larivière, 2004).

In recent years, the advent of data science, together with its highly regarded machine learning techniques, has revolutionized predictive analytics, offering innovative ways to tackle

complex problems, including customer churn (Hastie et al., 2009). These computational methods can discern intricate patterns and forecast future customer behavior, providing businesses the opportunity to pre-emptively address potential churn (Lemmens & Croux, 2006). Lemmens and Croux, in their 2006 paper about predicting customer churn of a telecom company in the United States, further elaborate that the advancements in machine learning techniques at that time significantly improved the classification performance of traditional statistical models.

The application of machine learning techniques, like logistic regression and random forest, have shown promising capabilities in predicting churn dynamics (Coussement, Lessmann, & Verstraeten, 2017). Yet, the process from data to insights is multifaceted and requires an understanding of the dataset's characteristics, diligent data preprocessing, appropriate model selection, and discerning interpretation of results.
While the prevalence of data analytics and machine learning is acknowledged, existing research on customer churn often focuses solely on one dimension, either emphasizing the machine learning aspect or immersing itself exclusively in economic theory. This leaves the intersection of these two dimensions regarding customer churn, particularly in the Portuguese banking sector, relatively unexplored (Dam & Dam, 2021; Rahman & Kumar, 2020).

Dam and Dam (2021) delve into the economic theories of customer churn but lack significant integration of machine learning methodologies. Similarly, Rahman & Kumar (2020) discuss the role of big data analytics and machine learning regarding customer churn prediction in the banking sector, but they do not intertwine these techniques with theoretical aspects of economics.

With this context in mind, our study aims to address these gaps by integrating economic theories with supervised machine learning methodologies to effectively predict customer churn in a Portuguese banking institution. Given the significant financial implications of churn, an accurate, efficient, and economically informed prediction model could provide valuable insights for businesses to manage customer relationships, enhance retention strategies, and secure their revenue streams.

2.1 Main Research Question

To support the effective incorporation of machine learning methodologies and economic theories when predicting customer churn, this study will be guided by the following research question:

'In what ways can the integration of economic theories and machine learning methodologies contribute to the understanding and prediction of customer churn in the context of a Portuguese banking institution?'

Three hypotheses, introduced in the literature review, will be adopted to answer the main research question. The hypotheses are constructed using economic theories and variables from the dataset utilized for our research, containing data from a Portuguese banking institution.

2.2 Navigating the Thesis

The chapters of this thesis unfold in a sequential manner. The literature review provides a comprehensive understanding of supervised machine learning and its relevance for customer churn prediction. Afterwards, we will introduce the main economic theories that are used to explain the phenomenon of customer churn and introduce the hypotheses that will aid us in addressing the main research question.

Following this, we introduce the dataset, outlining the characteristics of the variables using an exploratory data analysis. A detailed account of our machine learning methodologies ensues, elaborating on the preprocessing techniques, the machine learning models employed and the evaluation metrics used. We then predict customer churn with the models mentioned in the methodology, and analyze the results from these models, discussing their performance. Upon model deployment, we utilize the findings, to answer the research question. The final chapter offers a reflective conclusion, summarizing the insights gleaned and their alignment with our initial research question.

3 Literature Review

In this literature review, we will delve into the diverse realm of machine learning techniques and their applications in the field of marketing. Our exploration begins with an elucidation of supervised machine learning. By understanding the fundamental principles and characteristics of this machine learning type, we can grasp its applicability to several machine learning problems.

Following our discussion on supervised machine learning, our attention will shift towards the specific application of supervised machine learning in marketing. We will explore how supervised machine learning techniques have been leveraged to address marketing challenges and improve decision-making processes.

Next, we will delve into empirical studies that have focused on churn prediction. By reviewing existing literature on churn prediction, we aim to uncover the methodologies, and techniques employed to forecast and understand customer churn behavior. This examination will provide insights into the state of the art in customer churn prediction and identify potential gaps or areas for further research.

Furthermore, our review will narrow its focus to empirical studies that specifically investigate churn prediction in the banking sector. The banking industry, characterized by intense

competition and customer dynamics, presents unique challenges and opportunities for churn prediction. By analyzing studies that have explored churn prediction in this sector, we aim to gain a deeper understanding of the techniques, and models that contribute to accurate churn prediction in a banking context.

Through this comprehensive literature review, we seek to consolidate knowledge and identify key insights from prior research. By synthesizing information from diverse studies, we aim to contribute to the existing body of knowledge in the field of churn prediction using machine learning techniques.

3.1 Machine Learning

3.1.1 Supervised Machine Learning

3.1.1.1 Conceptualizing Supervised Learning

Supervised learning, a corner stone of machine learning, leverages algorithms to discover patterns from labeled training data to forecast unseen or future data. In this context, 'labeled' signifies that each sample within the training dataset is equipped with a corresponding output or 'label' (Müller & Guido, 2016)

The ultimate objective of supervised learning is to construct a model that can associate input variables (features) with an output variable (label) based on the discerned input-output pairs within the labeled training data (Shalev-Shwartz & Ben-David, 2014).

Take, for example, predicting customer churn in a banking institution. The labeled training dataset would include customer features like transaction history, demographic details, and product usage for each customer. Moreover, each customer would have a label indicating whether the customer churned or not. The supervised learning algorithm's job is to ascertain the correlations between these customer features and the churn label and formulate a predictive model.

3.1.1.2 Dissecting the Supervised Learning Process

The supervised learning process is conventionally divided into two primary stages: the training phase and the testing or prediction phase. During the training phase, the model assimilates a set of inputs and their respective correct outputs. The model’s parameters are fine-tuned in an iterative manner until the algorithm can correctly link the input to the output (Müller & Guido, 2016). The adjustment of model parameters is typically achieved via the minimization of a loss function, which quantifies the disparity between the model's predictions and the actual labels.

The testing phase commences upon successful completion of the training phase, wherein the trained model's predictive performance is assessed on unseen data. This phase gauges how efficiently the model can apply the learned correlations to new data instances. The model's performance is evaluated using various metrics, such as accuracy, precision, recall for

classification tasks, and mean absolute error or root mean squared error for regression tasks (Müller & Guido, 2016).

3.1.1.3 Exploring the Major Types of Supervised Learning Tasks

Supervised learning problems primarily fall into two principal categories, namely classification and regression, each pertaining to the nature of the target variable. Classification tasks are centered around predicting a discrete output or class label. In the context of binary classification, there are two potential outcomes (Ribeiro, Singh, & Guestrin, 2016).

For instance, emails could be classified as 'spam' or 'not spam', or in a medical diagnosis scenario, patients could be classified as 'disease present' or 'disease absent'. Multiclass classification extends this notion to more than two classes, such as classifying images of handwritten digits where there are ten possible classes, one for each digit from 0 to 9.

In contrast, regression tasks involve predicting a continuous outcome variable. Examples of regression problems include predicting the price of a house given a set of features like its size, number of rooms, and location, or forecasting stock prices based on historical data (Hearty, 2016).

3.1.1.4 Weighing the Strengths and Limitations of Supervised Learning

Supervised Learning carries an array of advantages. Predominantly, it derives its value from its potent ability to predict future outcomes based on past data. In a business setting, this allows for extensive predictive modeling, enabling strategic and informed decision-making. Additionally, many supervised learning algorithms yield models that are interpretable, a highly desirable trait in scenarios where understanding the model's decision-making process is paramount, such as healthcare and financial risk assessment (Müller & Guido, 2016)

Nonetheless, supervised learning comes with its share of limitations. One of the most significant challenges pertains to the requirement for labeled data. Procuring such datasets can be expensive and time-consuming, given that the labeling process often demands the knowledge of domain experts. Besides, the quality of the training data greatly influences the effectiveness of the model – erroneous labels or data samples that are not representative of the overall population can lead to ineffective models. Overfitting is another concern where models memorize the training data to such an extent that they underperform on unseen data. Overfitting typically arises when the model is excessively complex concerning the quantity and noise level of the training data (Hearty, 2016).

Despite these challenges, the relevance of supervised learning in an array of machine learning applications remains beyond dispute. Its capacity for precise predictive modeling, coupled with the potential for model interpretability, render it an invaluable tool in a multitude of sectors, extending from banking and finance to healthcare, marketing, and beyond.

3.1.2 Applications of Supervised Machine learning for marketing purposes

In line with the focus of this thesis on predicting customer churn using supervised machine learning, it is imperative to explore various applications of supervised machine learning in a marketing context. Thus, the subsequent discussion will delve into the specific applications of supervised machine learning in marketing, such as customer segmentation, customer lifetime value (CLV) prediction, personalized recommendations, and price optimization. By harnessing the potential of these supervised machine learning models, businesses can glean invaluable insights into customer behavior, optimize their strategic endeavors, and furnish customized experiences that forge robust customer engagement, thereby fueling sustained business growth.

3.1.2.1 Customer Segmentation

In the realm of marketing, customer segmentation is a critical concept that entails classifying a company's customers into distinct groups based on shared attributes. These characteristics could include demographic data such as age or income, purchasing patterns like the frequency or quantity of products bought, or the level of engagement with the company's services (Ferrell, Hartline, & Hochstein, 2021).

The essence of customer segmentation lies in its capacity to deliver tailored marketing strategies, thereby fostering enhanced customer relations and improving business outcomes. However, traditional manual segmentation methods are labor-intensive and can fall short in capturing the multifaceted nature of customer behaviors (Palmatier & Sridhar, 2020).

This is where supervised machine learning comes into play. By applying data-driven algorithms, it is possible to predict the group a new customer would fit into based on historical customer data, enabling more nuanced and dynamic segmentation (Müller & Guido, 2016).

In a similar vein, the study conducted by Valecha et al. (2018) showcased the potential of the Random Forest algorithm in predicting consumer behavior. Their research focused on the application of the random forest algorithm to segment customers based on their attributes and subsequently predict their behavior. By leveraging historical customer data encompassing demographics, purchasing patterns, and engagement levels, they trained the random forest model to create a robust framework for customer segmentation.

The findings of the study highlighted the significance of utilizing supervised machine learning techniques like random forest in customer segmentation for predicting consumer behavior. The ability to identify distinct customer segments and forecast their behavior empowered marketers to tailor their strategies with precision. By understanding the unique needs and preferences of different customer groups, businesses could develop targeted marketing campaigns, personalized offers, and customized experiences, ultimately enhancing customer satisfaction and driving business growth.

3.1.2.2 Customer Lifetime Value (CLV) Prediction

Customer Lifetime Value (CLV) is a predictive metric of paramount importance in the economics of customer relationships. The metric attempts to quantify the total value a

business can derive from the entire future relationship with a customer. It incorporates aspects such as the expected duration of the relationship, the customer's purchase frequency, and the average revenue per transaction (Palmatier & Sridhar, 2020).

A precise understanding of CLV is invaluable for businesses, as it informs a myriad of strategies pertaining to customer retention, acquisition, cross-selling, and up-selling. For instance, by identifying high CLV customers, businesses can prioritize their marketing efforts and resources more effectively. (Ferrell, Hartline, & Hochstein, 2021).
The application of deep learning models, such as deep neural networks, has shown considerable promise in predicting Customer Lifetime Value (CLV). Chen et al. (2018) conducted a study on CLV prediction in the video game industry using a deep learning-based regression model. Their research demonstrated the efficacy of utilizing deep learning and parametric models to accurately estimate the lifetime value of customers. By leveraging extensive historical data on player behavior, the model accurately forecasted players' future spending. This predictive capability enabled the gaming platform to optimize its customer relationship management strategy and allocate resources effectively, focusing on the most valuable players.

3.1.2.3 Personalized Recommendations

Personalized recommendations play an increasingly important role in today's digital business landscape. These recommendations are suggestions of products or services made to customers, which are specifically tailored based on their historical data, including previous interactions, preferences, or behaviors (Palmatier & Sridhar, 2020).
The power of personalized recommendations lies in their ability to enhance the customer experience and foster increased engagement. They allow businesses to recommend products or services that customers are likely to be interested in, thus improving the likelihood of a purchase and strengthening the overall relationship between the customer and the business. From the customer's perspective, these recommendations simplify the decision-making process, as they reduce the need for customers to sift through countless options to find what they need or might be interested in (Palmatier & Sridhar, 2020).

In the age of data abundance, manual creation of such recommendations becomes almost impossible. Hence, advanced technologies, specifically machine learning, have been employed to automate and refine this process. Collaborative Filtering, a supervised machine learning technique, has emerged as one of the most widely used approaches in this area. By analyzing past behaviors of numerous customers and detecting patterns and similarities among them, collaborative filtering can predict what a particular customer may like based on what similar customers have liked in the past.

The findings of Nilashi et al. (2019) demonstrate the effectiveness of Collaborative Filtering. In their 2019 paper, Nilashi et al. propose a new soft computing method that combines

supervised machine learning techniques to enhance eco-friendly hotel recommendations on TripAdvisor. Evaluation of their dataset demonstrates the method's effectiveness in handling numerous user ratings and making precise recommendations. By harnessing the power of supervised machine learning, the research empowers TripAdvisor's recommendation engine to offer personalized and eco-conscious travel experiences, revolutionizing hotel selections for travelers.

3.1.2.4 Price Optimization

Price optimization is a vital component of marketing strategy, aiming to determine the optimal price point that maximizes profitability while considering various factors such as demand, competition, costs, and consumer behavior. Traditionally, pricing decisions relied on intuition or simplistic rules. However, with the advancement of technology, supervised machine learning algorithms have emerged as powerful tools for optimizing pricing strategies (Ferrell, Hartline, & Hochstein, 2021).

Price optimization involves setting prices for products or services that maximize revenue, market share, or profitability, taking into account factors that influence consumer purchasing decisions. This task requires analyzing historical sales data, market trends, customer preferences, and competitive dynamics to determine the optimal price. The challenge lies in finding the balance between setting a price high enough to maximize profit and low enough to attract and retain customers (Palmatier & Sridhar, 2020).

Supervised machine learning models have demonstrated their effectiveness in price optimization by leveraging large volumes of data to predict consumer behavior and optimize pricing strategies. These models learn from historical data to identify patterns and relationships between pricing variables and customer responses, empowering businesses to make data-driven pricing decisions. Various supervised learning algorithms have been applied in price optimization, including linear regression, decision trees, support vector machines, and neural networks (Müller & Guido, 2016).

An exemplary study that showcases the application of supervised machine learning for price optimization is the research conducted by Spedicato, Dutang, and Petrini (2018). In their research, Spedicato et al. (2018) explored various supervised machine learning methods, including decision trees, random forests, and gradient boosting, to optimize pricing decisions. They compared the performance of these methods against the traditional Generalized Linear Models (GLMs), commonly used in pricing optimization. By analyzing historical sales data, market variables, and customer attributes, they evaluated the accuracy and effectiveness of the machine learning models in identifying the optimal price points.

The findings of Spedicato et al. (2018) demonstrated the superiority of supervised machine learning methods over standard GLMs in pricing optimization. The machine learning models showcased enhanced predictive power and outperformed the traditional methods, enabling businesses to make more precise and profitable pricing decisions. By leveraging the advanced

capabilities of machine learning algorithms, companies can gain a competitive edge by optimizing their pricing strategies.

3.1.3 Empirical Studies on Machine Learning for Customer Churn Prediction

Having explored the diverse applications of supervised machine learning in various marketing domains, we now shift our focus to a specific area of interest—customer churn prediction. Customer churn, the phenomenon of customers discontinuing their engagement or terminating their relationship with a company, poses a significant challenge for businesses across industries. As this thesis centers on utilizing supervised machine learning algorithms to predict customer churn, it is essential to examine empirical studies that have investigated the effectiveness of these models in this domain. These studies provide valuable insights into the methodologies, performance, and limitations associated with applying machine learning techniques to predict and mitigate customer churn.

The first empirical study to be reviewed is the research conducted by Vafeiadis, Diamantaras, Sarigiannidis, and Chatzisavvas (2015). Their study aimed to compare and evaluate various machine learning techniques for predicting customer churn in the telecom industry, providing valuable insights into the performance of these algorithms in churn prediction scenarios.
To conduct their analysis, the researchers gathered relevant data concerning characteristics of the calls that customers made and the type of service plans they were subscribed to. The dataset was meticulously prepared and preprocessed to ensure data quality and reliability.
Vafeiadis et al. (2015) compared different supervised machine learning techniques, including logistic regression, support vector machines, decision trees, naïve bayes, and artificial neural networks. The performance of these techniques was evaluated using established metrics such as accuracy, precision, recall, and F1-score. Additionally, the researchers applied an Adaboost algorithm to some of the previously mentioned models to enhance their performance, and they succeeded in doing so.
The findings of Vafeiadis et al. (2015) highlighted the effectiveness of various machine learning techniques for customer churn prediction. Through their rigorous comparison and evaluation, they demonstrated the strengths and weaknesses of each algorithm in capturing the complex patterns of customer behavior associated with churn. The evaluation metrics provided insights into the performance of these techniques in accurately predicting customer churn. Moreover, the researchers demonstrated that using a boosting algorithm like Adaboost can significantly improve the performance of the models.

The second study to be examined is the comprehensive research conducted by Khodabandehlou and Zivari Rahman (2017). The study aimed to assess and compare the effectiveness of various supervised machine learning algorithms in predicting customer churn by analyzing customer behavior patterns. This investigation was carried out in six stages.

In the first stage, customer behavioral data was collected and prepared for analysis. The second stage involved the formation of derived variables and the selection of influential variables using a discriminant analysis method. The third stage encompassed the selection of training and testing data and reviewing their proportions. Moving on to the fourth stage, prediction models were developed using simple, bagging, and boosting versions of supervised machine learning. In the fifth stage, churn prediction models were employed and selected variables were compared. Finally, in the sixth stage, appropriate strategies were provided based on the proposed model.

Khodabandehlou and Zivari Rahman (2017) systematically compared multiple supervised machine learning techniques, including support vector machines, decision trees, and various types of artificial neural networks. The researchers also utilized bagging and boosting methods on the models. The evaluation of these models was based on well-established performance metrics, such as accuracy, precision, recall, and F1-score.

The findings of the study shed light on the performance of different supervised machine learning algorithms for customer churn prediction based on the analysis of customer behavior. Khodabandehlou and Zivari Rahman (2017) found that all types of artificial neural networks used to predict customer churn outperformed the decision tree and support vector machine models. Furthermore, they concluded that each model performed better when paired with a boosting algorithm.

The third study we investigate, by Ahmad, Jafar, and Aljoumaa (2019), focuses on customer churn prediction within the telecom industry. Their research aimed to leverage machine learning techniques on a big data platform to predict customer churn and assist telecom companies in effectively retaining their customers.

Ahmad et al. (2019) collected a vast amount of data from SyriaTel, a Syrian telecom company. The dataset included variables regarding the type of service package subscription, call detail records, and cell phone device specifications. To develop churn prediction models, Ahmad et al. (2019) utilized various machine learning algorithms, such as Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). The performance of these models was assessed using standard evaluation metrics, including Receiver Operating Characteristics (ROC) curves and the area under the curve (AUC). Additionally, the authors had to address challenges related to an imbalanced dataset, missing values, and data variety.

The findings of Ahmad et al. (2019) underscored the effectiveness of machine learning models in predicting customer churn in the telecom sector. The XGBoost model performed the best in the tests. The researchers noted that the method of data preparation, feature selection, and incorporation of mobile social network features had the most significant impact on the success of this model. The importance of the research conducted by Ahmad et al. (2019) in the telecom market is to assist companies in generating higher profits by predicting which customers are more likely to churn. Therefore, their research aimed to develop a system that accurately predicts customer churn in SyriaTel telecom company.

The last empirical study concerning customer churn to be analyzed is the study by Lalwani, Mishra, Chadha, and Sethi (2022). Their research aimed to develop an effective customer churn prediction system using a machine learning approach, addressing the challenges faced by businesses in retaining customers.

Lalwani et al. (2022) worked through phases, starting with data preprocessing and feature analysis in the first two phases. In the third phase, feature selection was taken into consideration using the gravitational search algorithm. The data was then split into train and test sets in an 80% to 20% ratio. In the prediction process, the researchers applied popular predictive models, including Logistic Regression, Naive Bayes, Support Vector Machine, Decision Trees, Random Forest Classifier, Extra Tree Classifier, and Boosting Algorithms such as AdaBoost, XGBoost, and CatBoost. Boosting and ensemble techniques were also applied to assess their effect on model accuracy. Furthermore, K-fold cross-validation was used on the train set for hyperparameter tuning and to prevent overfitting of models. Finally, the obtained results on the test set were evaluated using confusion matrices and ROC curves.

Lalwani et al. (2022) found that two ensemble learning techniques, Adaboost classifier and XGBoost classifier, outperformed the other models in the churn prediction problem. The best models demonstrated superior performance across all evaluation metrics, including accuracy, precision, F1-measure, recall, and AUC score.

By showcasing the efficacy of machine learning algorithms in accurately identifying customers likely to churn, the research conducted by Lalwani et al. (2022) makes a valuable contribution to the field of customer churn prediction.

In conclusion, the empirical studies reviewed above have demonstrated the effectiveness of supervised machine learning techniques in predicting customer churn. These studies have compared different algorithms, evaluated their performance metrics, and highlighted the strengths and weaknesses of each approach. By leveraging machine learning, businesses can proactively identify customers at risk of churn and implement targeted retention strategies, leading to improved customer satisfaction and sustainable business growth.

3.1.4 Churn prediction in the banking sector

Within the realm of customer churn prediction, an extensive body of research has investigated the application of machine learning techniques in various domains. In line with the objective of this thesis, which aims to predict customer churn in the banking sector, our focus now turns to empirical studies that have specifically examined customer churn prediction in the banking industry using machine learning approaches. Analyzing previous academic endeavors that have explored similar research questions will contribute to our understanding of customer churn dynamics in the banking sector, and ultimately enhance the accuracy and effectiveness

of our own churn prediction model. Through a comprehensive examination of these empirical studies, we can derive valuable insights and leverage the expertise and methodologies employed to advance our research in the context of customer churn prediction within the banking industry.

The first academic paper to be examined is by Bilal Zorić (2016). This research presents a comprehensive investigation of the application of neural networks, a supervised machine learning technique, in the context of customer churn prediction within the banking sector.
The hypothesis underlying this investigation posited that customers who availed themselves of a greater range of bank services would demonstrate higher levels of loyalty. To conduct this research, the author employed real-world data from a small Croatian bank, leveraging it to create a robust customer churn prediction model.
The research process encompassed several distinct phases. Firstly, the research problem was meticulously defined and subsequently translated into a data mining problem. Secondly, data gathering and preparation were conducted, involving the transformation of data into a pre-specified format and the implementation of data cleansing techniques to eliminate corrupt, inaccurate, or irrelevant records. Subsequently, the researcher constructed and evaluated the neural network model, selecting and applying various modeling techniques while optimizing model parameters for superior performance. Lastly, the knowledge derived from the model was deployed within the targeted banking environment, facilitating data-driven decision-making and strategies.
The study's conclusions shed light on the existence of a specific group, particularly young individuals such as students, who currently possess fewer than three bank products but have the potential to become highly valuable clients in the future. Considering this, the author recommended that the bank should adapt its product offerings to cater to the unique needs of this group, potentially introducing specialized products like student loans, offering favorable interest rates, and promoting the utilization of internet banking services.
The research conducted by Bilal Zorić (2016) contributes to the field of customer churn prediction in the banking industry by leveraging neural networks as a powerful tool. The study underscores the importance of customer loyalty and provides practical recommendations for banks seeking to enhance customer retention rates and develop targeted marketing strategies. By capitalizing on the insights provided by this research, banks can optimize their operations and foster sustainable growth in an increasingly competitive business landscape.

The second empirical study that will be discussed is the research conducted by Keramati, Ghaneei, and Mirmohammadi (2016). The aim of this study was to identify the features of churners in electronic banking services through the application of data mining techniques.
To investigate customer churn, the study focused on customer dissatisfaction, level of service usage, and customer-related variables within the bank's database. Various demographic

variables, including age, gender, career, and level of education, as well as transaction data from electronic banking portals such as ATM, mobile bank, telephone bank, internet bank, and USSD-based mobile banking, were extracted for analysis. Additionally, the length of customer association and customer complaints were considered in the study.

The research methodology encompassed a 6-phase approach:

1. Business understanding
2. Data understanding
3. Data preprocessing
4. Modeling
5. Evaluation
6. Deployment

In the modeling phase, the decision tree (DT) technique was chosen due to its ability to provide easily interpretable rules. The decision tree model was well-suited for identifying the features of churners and understanding the if-then rules associated with customer churn prediction. Furthermore, the decision tree algorithm was selected to handle the numerical and categorical types of data in the study.

Data preprocessing involved outlier detection and elimination to ensure data quality. Missing values were addressed using two methods: replacing them with the average value of the corresponding variable or utilizing the k-nearest neighbor (k = 5) approach. As the dataset exhibited an imbalanced distribution of churners and non-churners, a bootstrap sampling module in the RapidMiner data mining software was employed to overcome this challenge. By performing random sampling with replacement, customer record samples were obtained. The researchers employed a decision tree approach to evaluate the results and determine the best method for data cleaning.

The developed prediction model categorized the characteristics of churned customers into five distinct groups. This model successfully predicted customer churn in electronic banking services by considering the shared characteristics within each group. The model's performance was evaluated using metrics such as accuracy, recall, precision, F1-score, and ROC curve analysis, all of which yielded high scores.

The study by Keramati et al. (2016) makes a significant contribution to the field of customer churn prediction in the banking sector. By employing a decision tree model and utilizing various evaluation metrics, the researchers effectively identified the features of churners and developed a robust prediction model for customer churn in electronic banking services. This study enhances our understanding of customer churn dynamics and provides valuable insights for banks in their efforts to retain customers effectively.

The limitations of the study include the reliance on data solely from the bank's database, which may overlook other important factors influencing churn. The extraction of data was time-consuming due to the large volume and privacy concerns. Future research should explore additional data sources and incorporate qualitative methods to gain a more

comprehensive understanding of churn. By addressing these limitations, future studies can improve the accuracy and applicability of customer churn prediction models in electronic banking services.

The third empirical study to be discussed is the research conducted by Sabbeh (2018). This study aims to provide a comprehensive analysis and comparison of different machine learning techniques employed in the context of churn prediction.
In the study, various machine-learning algorithms representing different learning categories were applied to a dataset from a telecommunication company, consisting of 3333 records.
The study made use of a dataset that encompassed a wide range of customer statistical data, including 17 explanatory features related to customers' service usage, such as daily usage, international calls, and customer service calls. Notably, the dataset exhibited an imbalanced nature, with the churn class representing only 14% of the observations, while the non-churn class constituted the remaining 86%.

The machine learning techniques examined include logistic regression, CART decision tree, Naïve Bayesian, support vector machine, k-nearest neighbor, ensemble learning techniques (Ada Boost, Stochastic Gradient Boost, and Random Forest), Multi-layer Perceptron neural network, and Linear Discriminant Analysis.
The evaluation of these techniques revealed that the Random Forest and Ada Boost models demonstrated superior performance, outperforming other methods with an impressive accuracy of approximately 96%. Multi-layer Perceptron and support vector machine also yielded promising results, achieving an accuracy of 94%. Decision tree exhibited a respectable accuracy of 90%, while Naïve Bayesian and logistic regression models reached accuracies of 88% and 86.7%, respectively.

To ensure the reliability and validity of the results, the study employed a rigorous data preprocessing phase. This phase involved data transformation, data cleaning, and feature selection, all aimed at refining the dataset and enhancing the accuracy of the subsequent models. The dataset was split in a Training and testing dataset, where the samples are randomly chosen with cross validation 60% for training and 40% for testing.
Overall, this study contributes to the existing body of knowledge by offering a benchmark for the most used machine-learning techniques in the field of churn classification. The findings highlight the efficacy of ensemble-based learning methods, specifically Random Forest and Ada Boost, in accurately predicting customer churn. This knowledge can assist businesses in developing effective customer retention strategies, ultimately leading to improved customer satisfaction and long-term profitability.

The fourth empirical study that we will investigate is by Kaur and Kaur (2020). Their research focuses on investigating the application of various supervised machine learning models, including Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), and

Random Forest (RF), for the purpose of predicting customer churn within a banking dataset. The study aims to compare the performance of these models by evaluating metrics such as accuracy, recall, and others.

The study acknowledges the challenge posed by skewed data commonly encountered in banking datasets, which can adversely impact the predictive capabilities of classifiers. In response, the research employs a diverse range of classifiers to address this issue and enhance model performance.

The dataset utilized in the study is sourced from Kaggle and encompasses 28,382 records containing 21 distinct features. These features are classified into three main categories: demographic information, customer-bank relationship data, and transactional information. Notably, the minority class, representing churners, constitutes a mere 18.5% of the complete dataset, accounting for 5,260 instances.

To ensure data quality, the study undertakes comprehensive preprocessing procedures, encompassing the handling of missing values, addressing class imbalance, scaling the data, and eliminating irrelevant features. The sklearn.preprocessing package is leveraged to facilitate efficient data preprocessing.

Regarding model training, the study adopts prominent machine learning algorithms, such as LR, DT, KNN, and RF. The dataset is divided into training and testing sets, with 70% of the data allocated to model building and the remaining 30% reserved for assessing model performance.

To evaluate the effectiveness of the models, a range of performance measures are employed, including Recall, Precision, ROC, AUC, and Accuracy.

In the final analysis, ensemble techniques, including averaging and max voting, are employed to enhance the performance of the models, with the Random Forest model emerging as the most effective among the models investigated.

Overall, the research conducted by Kaur and Kaur (2020) contributes to the field of customer churn prediction in the banking industry. By comprehensively comparing the performance of different machine learning models, the study offers valuable insights into their performance in predicting customer churn within a banking dataset.

The last empirical study to be discussed is the research conducted by Vo, Liu, Li, and Xu (2021) titled 'Leveraging Unstructured Call Log Data for Customer Churn Prediction' published in the Knowledge-Based Systems journal. The study focuses on the importance of customer retention in the financial services industry and proposes a customer churn prediction model that utilizes unstructured data from phone communication.

The study collected a large-scale dataset from an Australian call center, comprising two million calls from over two hundred thousand customers. The dataset was analyzed using various text mining methods and interpretable machine learning techniques to predict customer churn risks and extract meaningful insights. The research aimed to investigate three main areas:

1. The utility of unstructured data for churn prediction.
2. Suitable approaches and techniques for leveraging unstructured data.
3. Customer profiles and characteristics associated with high churn risk and how to retain these clients.

The research makes several notable contributions. Firstly, it is one of the first attempts to incorporate unstructured text mining with structured data mining and interpretable machine learning for churn prediction in the financial services domain. Secondly, the study demonstrates that leveraging unstructured data and interpretable machine learning enables capturing a comprehensive customer preference spectrum for churn prediction. Thirdly, the study compares multiple text mining techniques to identify suitable approaches and fully leverage unstructured data for churn prediction, including exploring textual feature representations and personality traits. Lastly, the research utilizes interpretable machine learning techniques to evaluate feature importance at different levels, enabling the development of customized retention strategies for different customer segments.
In the methodology, the study employed 'term frequency-inverse document frequency' (TFIDF) as a suitable technique to derive term importance features from the call logs. Additionally, phrase embedding using the Word2Vec model was applied to understand the semantics carried by various combinations of terms. The research employed a multi-stacking ensemble model that utilized four supervised learning algorithms: Naïve Bayes, Logistic Regression, Random Forest, and Extreme Gradient Boosting. The study also considered state-of-the-art text mining approaches using BERT embedding and Bidirectional Long Short-Term Memory (BiLSTM) neural network.

The empirical results confirmed their hypothesis that unstructured data, specifically customer call logs, is valuable for customer churn prediction. The inclusion of text features significantly improved the prediction accuracy, as measured by AUC scores. The study demonstrated that text data provides insightful information for extracting customer personalities and characteristics.

In conclusion, the research highlights the utility of unstructured data, such as customer call logs, for generating meaningful insights and emphasizes the importance of utilizing interpretable machine learning techniques in all types of customer information systems.
One limitation of the research conducted by Vo, Liu, Li, and Xu (2021) is the lack of direct access to the recorded calls due to customer privacy concerns. As a result, the reliance on third-party text transcription services has led to poor quality text data, thereby hindering the effectiveness of advanced embedding models for feature extraction. To mitigate this limitation, the study focused on utilizing term-based approaches to extract text features. Despite this limitation, the researchers were able to generate meaningful insights and contribute to the field of customer churn prediction using alternative methods. Future

research could explore ways to enhance the quality and accessibility of call recordings to improve the accuracy and effectiveness of customer churn prediction models.

The reviewed empirical studies in the context of customer churn prediction in the banking industry reveal several salient themes. Firstly, the supervised machine learning methodologies mentioned demonstrated their efficacy in uncovering hidden patterns and relationships within extensive banking datasets, enabling accurate predictions of customer churn. Secondly, the studies consistently emphasize the significance of customer loyalty and the value of retaining at-risk customers. Tailoring product offerings and services to meet the specific needs of customers who utilize fewer products emerges as a recommended strategy. Moreover, the utilization of machine learning algorithms, such as ensemble methods like Random Forest and Ada Boost, shows promise in improving churn prediction accuracy. Lastly, some studies explore the integration of unstructured data, such as customer call logs, to enhance churn prediction models and extract valuable insights pertaining to customer preferences and characteristics. These findings underscore the importance of data-driven decision-making, targeted marketing strategies, and personalized customer retention efforts for banks aiming to optimize their operations and foster sustainable growth in a competitive business landscape.

3.2 Economic Theories and customer churn

Having thoroughly examined the existing scientific literature pertaining to customer churn prediction utilizing supervised machine learning techniques, our focus now shifts towards exploring the economic theories relevant to understanding customer churn. This transition allows us to bridge the gap between machine learning methodologies and the economic underpinnings that drive customer behavior, allowing us to answer our main research question.

We will commence by establishing a clear definition of customer churn within the context of our study. Subsequently, we will delve into a comprehensive exploration of economic theories, concepts, and assumptions that shed light on the phenomenon of customer churn. By drawing upon these economic frameworks, we aim to construct a solid theoretical foundation for our research, enabling us to uncover the key drivers and determinants of customer churn in the banking sector. Building upon these economic theories, we will then proceed to formulate three hypotheses which will serve as the basis for main research question, as seen in figure 1. Ultimately, findings from our machine learning models and statistical tests are used to obtain an answer to our central research question. Through this interdisciplinary approach, we endeavor to uncover novel insights that integrate economic theory and machine learning science techniques, contributing to a more comprehensive understanding of customer churn and retention dynamics within the banking industry.

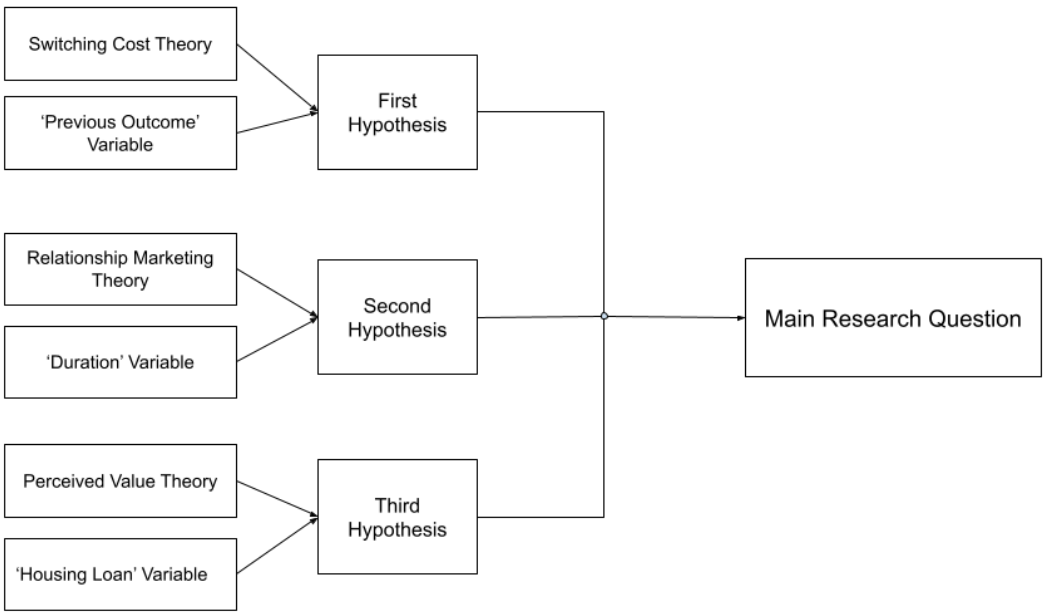

Figure 1: The Conceptual framework between the hypotheses and main research question

3.2.1 Defining Customer Churn

To commence, our exploration begins with the elucidation of Customer Churn. Customer churn, alternatively referred to as customer attrition, characterizes the phenomenon wherein customers choose to sever their business association with a company or service (Van den Poel & Larivière, 2004). In the banking industry, customer churn is typically identified when customers close their accounts or cease all banking transactions within a specific timeframe (Kaur & Kaur, 2020).

The ramifications of customer churn extend beyond the realm of customer relationships, posing significant challenges to a company's profitability and financial stability. Extensive research consistently underscores the cost-effectiveness of customer retention compared to customer acquisition. Reichheld and Schefter (2000) emphasize the strategic significance of retaining existing customers, as it is generally more cost-effective than acquiring new ones. Reichheld (2001) observes that in the financial sector even a modest 5% increase in customer retention can yield a remarkable 25% increase in profit. Reinforcing this notion, Jandaghi et al. (2011) report that the cost of acquiring a new customer is 15 times higher than retaining an existing customer. Furthermore, their findings reveal that a mere 5% increase in customer loyalty can yield profit increases ranging from 25% to 85%.

A comprehensive understanding of customer churn from an economic standpoint assumes paramount importance for strategic business planning. Organizations equipped with the ability to predict and address customer churn can proactively devise retention strategies, safeguarding their revenue streams and preserving their market share (Palmatier & Sridhar, 2020). Moreover, churn analysis offers invaluable insights into customer satisfaction, loyalty, and overall company performance, empowering informed strategic decision-making.

3.2.2 Switching Costs Theory

The first economic theory that will be explored is the Switching Costs Theory. This theory posits that customers encounter various 'costs' when considering a switch from one service provider to another, creating barriers to such changes. These costs can manifest in different forms, including transaction costs, learning costs, and artificial or contractual costs (Klemperer, 1987). For instance, when comparing two banks that offer identical checking accounts, customers may face significant transaction costs when closing their account with one bank and opening a new one with a competitor (Klemperer, 1987).
Supporting this notion, Shy (2002) conducted a study in the Finnish market for bank deposits and identified substantial switching costs associated with customer behavior in that context.

Nguyen et al. (2020) conducted a study in the context of electronic banking in commercial banks to investigate the relationships and impact of service quality, customer satisfaction, and switching costs on customer loyalty. The study gathered data from 227 electronic banking users in Hanoi City, Vietnam, primarily consisting of students and paid employees. Questionnaires with a 7-point Likert scale were utilized to collect the data, which were subsequently analyzed using the multivariate linear regression method.

The findings of the study revealed a strong and positive correlation between customer loyalty and switching costs. This implies that as the barriers to switching banks increase, customers are more likely to exhibit loyalty towards a single bank. The study contributes to the understanding of how switching costs play a role in shaping customer behavior and fostering loyalty in the (electronic) banking sector.
By utilizing the principles and tenets of the Switching Costs Theory, we propose our first hypothesis:

Customers who previously had a term deposit at the bank exhibit lower churn rates in the current marketing campaign.

3.2.3 Relationship Marketing Theory

Relationship Marketing, as defined by Lazirkha et al. (2022), is a relevant and enduring concept in modern business environments. The theory emphasizes the significance of building and nurturing strong customer relationships to enhance customer satisfaction and loyalty. Moreover, the theory recognizes that customers are not just transactional entities but valuable long-term partners. Lazirkha et al. (2022) elaborated that the theory emphasizes understanding and addressing customer needs, preferences, and expectations over time, rather than solely focusing on short-term sales or transactions. The core elements of relationship marketing include customer welfare, trust, commitment, and the importance of customer service and loyalty. Their approach aligns with the broader perspective that Relationship Marketing Theory, as proposed by Morgan and Hunt (1994), holds in fostering loyalty and mitigating churn. Embracing the principles of relationship marketing helps

businesses cultivate enduring customer relationships, leading to enhanced satisfaction, loyalty, and long-term success.
The advancement of technology across various domains, such as improved computing capabilities, enhanced accessibility to data storage, the emergence of big data, and advancements in internet infrastructure, collectively shape the potential actions available to managers in developing and strengthening relationships. Notably, the increased utilization of information technology in generating customer insights and applying them to relationship marketing initiatives has led to the prominence of customer relationship management (CRM). While relationship marketing and CRM theories are often used interchangeably, CRM represents a strategic approach to managing customer relationships through the effective utilization of technology, acknowledging that technology plays a pivotal role but not exclusively (Payne & Frow, 2017).

Payne and Frow define Relationship Marketing as the strategic management of relationships with all relevant stakeholders, encompassing not only customers but also suppliers, influencers, referral sources, and internal markets. In contrast, CRM focuses on the strategic management of relationships specifically with customers, leveraging technology as an effective tool (Payne & Frow, 2017).

In line with Relationship Quality Theory, which posits that customers with stronger relationships with the bank are less prone to churn, we propose our second hypothesis:

Customers who have a longer contact duration exhibit lower churn rates in the current marketing campaign.

3.2.4 Perceived Value Theory

Perceived Value Theory encompasses a framework that centers on the exploration of customers' subjective perceptions concerning the value or benefits obtained from a product, service, or overall customer experience (Sánchez-Fernández & Iniesta-Bonillo, 2007). The theory acknowledges the inherent subjectivity involved in customers' assessment of value, considering individual needs, expectations, and the perceived advantages offered by a particular offering relative to its associated costs (Boksberger & Melsen, 2011).

Within the confines of perceived value theory, customers evaluate the value proposition of a product or service by contemplating a diverse array of perceived benefits, encompassing dimensions such as quality, performance, functionality, convenience, emotional appeal, and social status (Aulia, Sukati, & Sulaiman, 2016). Simultaneously, they assess the sacrifices or costs entailed in acquiring or utilizing the offering, comprising monetary price, time commitment, effort expenditure, and potential risks.

The theory posits that customers make purchase decisions based on the appraisal of the net value, which denotes the perceived benefits minus the perceived costs linked to the offering. If the net value is positively perceived and surpasses that of alternative options, customers are more inclined to select the offering and exhibit sustained loyalty. Conversely, if the net value is perceived negatively or deemed inferior to alternative options, customers may opt for alternatives or discontinue their engagement with the offering.

In relation to our research, Perceived Value Theory suggests that customers who perceive higher value in the bank's offerings are less likely to churn. Hence, we propose the third hypothesis:

Customers who already had a financial product at the bank exhibit reduced churn rates in the current marketing campaign.

3.2.5 Linking hypotheses with variables

The hypotheses will be accompanied by variables from our dataset. By matching the hypotheses with their corresponding variables, we can perform statistical analyses and include these variables in our machine learning models. The first hypothesis will be paired with the variable 'previous outcome', the second hypothesis will be paired with the variable 'duration', and the third hypothesis will be paired with the variables 'previous outcome' and 'housing loan'. These variables, their use for the hypothesis and our dataset will be discussed in detail in the following chapter.

4 Research Methodology

4.1 Exploratory Data Analysis

4.1.1 The Dataset

The corpus of this research is predicated on the Bank Marketing dataset, procured from the University of California, Irvine (UCI) machine learning repository. The UCI is ubiquitously acclaimed and frequented as a trove of datasets, specifically conceived for machine learning applications. Incepted in 1987, it has proliferated over the years to boast an expansive compilation of over 500 datasets hailing from multifarious domains, rendering it one of the most antique and comprehensive archives of data extant (Lichman, 2013).

The integrity of the UCI Repository is fortified by its academic provenance and the methodical protocol it exercises for data incorporation. Each dataset is meticulously scrutinized, chronicled, and frequently correlated with academic endeavors or initiatives, thereby further bolstering the credibility of the data source. This stringent process certifies the data's dependability, accuracy, and adaptability for diverse research explorations (Dua & Graff, 2019).

The dataset for this study, taken from the UCI repository, comprises information gathered from a Portuguese banking institution, amassed over the span from May 2008 through

November 2010, concurrent with a telemarketing campaign advocating long-term bank deposits. The principal objective of the dataset is to anticipate whether a client will assent to a term deposit, an indicator that can be linked with customer attrition. The dataset encapsulates a total of 41,188 instances and 20 input attributes, along with the target variable, 'y,' which signifies whether the client ratified a term deposit.

The dataset is donated to the UCI by Moro, Cortez, and Rita, who used a more extensive version of the dataset in their 2014 paper centered on predicting the success of bank telemarketing. Their dataset consisted of 150 features regarding personal data of the bank clientele, financial product data, and socio-economic indicators.

Table 1: Description of dataset

Feature	Data Type	Description of Features
age	Numeric	Age of client
job	Multi-Categorical	Job of client ("admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
marital	Multi-Categorical	Marital status of client ("married","divorced","single"; note: "divorced" means divorced or widowed)
education	Multi-Categorical	Education level of client (basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
default	Binary Categorical	Has client credit in default ('yes', 'no')
housing	Binary Categorical	Has client housing Loan ('yes', 'no')
loan	Binary Categorical	Has client personal Loan ('yes', 'no')
contact	Binary Categorical	Communication type with client ('cellular', 'telephone')
month	Multi-Categorical	Last contact month
day_of_week	Multi-Categorical	Last contact day
duration	Numeric	Duration of last contact, in seconds
campaign	Numeric	Number of contacts performed during this campaign
pdays	Numeric	Number of days passed after client was last contacted from previous campaign (999 means client was not previously contacted)
previous	Numeric	Number of contacts performed before this campaign
poutcome	Multi-Categorical	outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
emp.var.rate	Numeric	Employment variation rate - quarterly indicator
cons.price.idx	Numeric	Consumer price index - monthly indicator
cons.conf.idx	Numeric	Consumer confidence index - monthly indicator
euribor3m	Numeric	Euribor 3 month rate - daily indicator
nr.employed	Numeric	Number of employees - quarterly indicator
y	Binary Categorical	Has the client subscribed a term deposit? ('yes','no')

Note. This table shows the variables from the UCI Bank Marketing dataset, with their corresponding data type and description.

Table 1 shows the variables of the dataset used in our research. We see that the dataset contains numeric and categorical variables. Moreover, we observe that the variables are themed regarding in demographics, marketing campaign data and socio-economic indicators.

4.1.2 Target Variable

At the heart of our study lies the dependent variable, denoted as 'y', which serves as an indicator of whether a customer has opted to subscribe to a term deposit or discontinued their patronage with the bank. This target variable assumes the values 'yes' and 'no', where 'yes' signifies a customer's subscription to a term deposit, and 'no' denotes their decision not to subscribe, leading to churn. To facilitate analysis, the values of the target variable have been transformed, wherein 'yes' is represented by a value of 1, and 'no' by a value of 0.

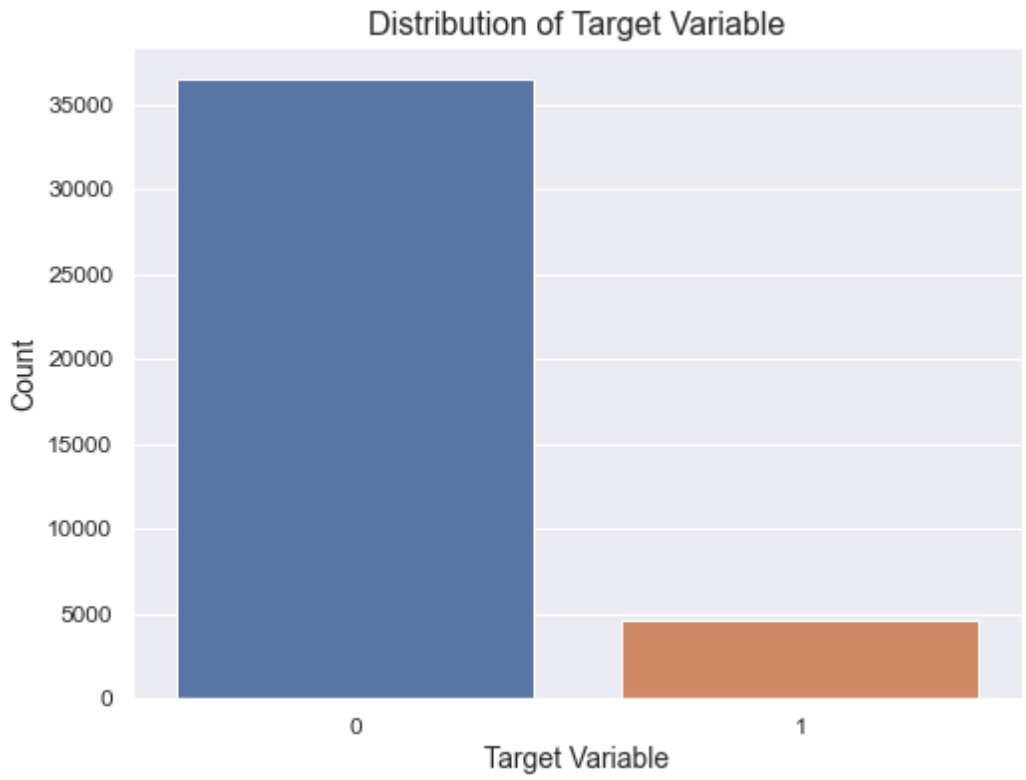

Figure 2: Count of categories for target variable

The figure 2 shows the counts of the target variable categories 0 and 1. The figure demonstrates a notable class imbalance in the target variable. The majority class, representing customers who did not subscribe (denoted by the value 0), comprises 36,548 instances, while the minority class, representing customers who did subscribe, consists of 4,640 instances. This substantial class imbalance holds significant implications from both machine learning and statistical perspectives, warranting careful consideration in subsequent analyses.

4.1.3 Numeric Variables

This section delves into the exploration of numeric variables in our dataset that will be employed in our machine learning models. These variables encompass demographics, marketing campaign, and socio-economic themes, providing quantitative insights into customer churn. Through rigorous analysis and visualization, we aim to unveil patterns, relationships, and potential predictors associated with customer behavior.
Our exploratory data analysis (EDA) commences with an examination of the distributional characteristics of these numeric variables, encompassing measures of central tendency, dispersion, and skewness. Additionally, we explore the relationships between these variables and the target variable to identify potential factors influencing churn. This initial analysis lays the groundwork for subsequent stages of our research, facilitating informed decisions in feature selection and model development.

We will utilize a histogram paired with a KDE line and box plots to analyze the distribution of the numeric variables. A KDE (Kernel Density Estimation) line is a continuous smooth curve superimposed on a histogram, depicting the estimated probability density function of a continuous variable. The presence of a KDE line aids in discerning modes, peaks, and regions of elevated density within the data, facilitating a clear comprehension of its shape and inherent characteristics.

To complement our analysis of the distribution of numeric variables, we will also incorporate box plots. Box plots provide valuable insights into the central tendency, spread, and skewness of the data. They offer a visual representation of quartile values, median, and potential outliers, allowing us to assess the variability and distributional properties of the variables. Moreover, the boxplots will include a red triangle, which represents the mean.

By combining the histogram and KDE line with box plots, we can gain a holistic understanding of the numeric variables. The histogram and KDE line reveal the overall shape and density patterns, while the box plots provide specific numerical summaries and highlight potential anomalies. This comprehensive approach enhances our ability to uncover hidden insights and make informed interpretations of the data.
Due to the expansive number of numeric variables in our dataset, we will only analyze the numeric variable used to answer the second hypothesis in this section. The appendix contains the EDA on all other numeric variables.

4.1.3.1 Duration

The variable ‘Duration' refers to the duration of the last contact made with the customer during the marketing campaign. It measures the length of time, in seconds, that the customer engaged with the bank representative or marketing personnel. The contact duration serves as

an important indicator of customer engagement and interaction intensity. A longer contact duration suggests a more extensive conversation or interaction, which may indicate a higher level of customer interest or involvement.

In our study, we include the 'Duration' variable to answer the second hypothesis, where we investigate if customers who have a longer contact duration exhibit lower churn rates in the current marketing campaign. The utilization of the 'Duration' variable in the second hypothesis facilitates us with an appropriate way to grasp the potential influence of the relationship marketing theory.

By examining the association between the contact duration and churn behavior, we aim to determine whether customers who have longer contact durations during the marketing campaign are more likely to exhibit higher engagement levels and subsequently lower churn rates.

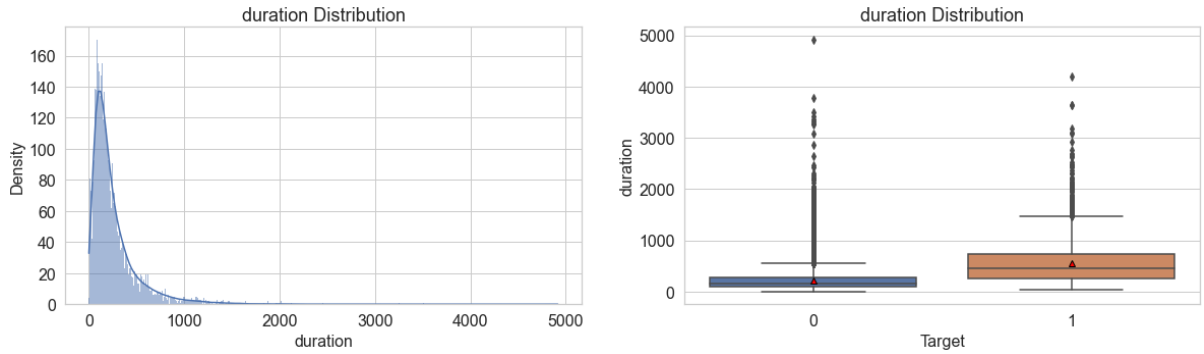

Figure 3: Histogram with KDE-line and boxplots for the variable Duration

Upon examining the histogram in the left plot of figure 3, it is evident that most of the 'Duration' values fall within the range of 0 to 1000 seconds. However, the presence of a long right tail in both the histograms and the KDE line indicates the presence of outliers that extend beyond this range.

Moving on to the boxplots in the right plot, an interesting pattern emerges. All quartiles of the group that subscribed to a term deposit exhibit higher values compared to the other group. This suggests that the contact duration for those who subscribed tends to be longer. Additionally, it is worth noting that the group of non-subscribers appears to have a higher number of outliers, indicating some exceptional instances of longer contact durations.

Table 2: Quartiles of the variable Duration

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	0	95	163.5	279	4918
1	37	253	449	741.25	4199

Note. This table shows the quartiles and whiskers of the 'Duration' variable, segmented into groups with a value 1 and 0 for the target variable. This table shows that the group of customers that have a value 1 for the target variable, indicating that they subscribed to a term deposit, has higher quartiles than the group with the value 0 for the target variable.

A closer examination of the quartiles in Table 2 reinforces this observation. Except for the maximum value, all quartiles for the group that subscribed to a term deposit are indeed higher. This finding implies that customers who showed interest in the term deposit had longer durations of contact, which could potentially be attributed to more in-depth conversations or discussions.

The disparity in contact duration between the two groups raises interesting insights into the potential influence of duration on subscription outcomes. It suggests that a longer contact duration might be a contributing factor in attracting customers to subscribe to the term deposit. Further exploration and analysis are needed to explore the underlying reasons behind this relationship and its implications for marketing strategies and customer engagement.

4.1.4 Categorical Variables

The analysis of categorical variables will be conducted using two bar plots. The first bar plot provides an overview of the count of each category, with bars color-coded based on the target variable. The color distinction helps visualize the distribution of subscribers (target variable = 1) and non-subscribers (target variable = 0) within each category.

The second bagplot focuses on proportions, showcasing the relative proportions of subscribers and non-subscribers across the categories. By examining the internal proportions of each category, we can make more accurate comparisons. For instance, we can now compare the proportions of subscribers and non-subscribers within a category that has a high customer count with that of a category that has a low customer count. This allows us to gain deeper insights into the variations in subscriber rates across different categories. This plot enables a more precise evaluation of how the target variable is distributed within the categorical variable.

Due to the extensiveness of our dataset, we will only discuss the variables needed to answer our hypotheses. The rest of the categorical variables are explored in the appendix.

4.1.4.1 Housing

The binary categorical variable 'Housing' captures whether a customer has a housing loan or not. It represents the housing loan status of customers, distinguishing between those who have a housing loan ('yes') and those who do not ('no').

The 'Housing' variable provides valuable information about customers' housing loan obligations and their financial commitments related to housing. It serves as an indicator of their housing ownership or rental status. Moreover, the ‘Housing’ variable will be used to address the third hypothesis. The variable will be used to find out if customers who already had a financial product at the bank, like a housing loan, exhibit reduced churn rates in the current marketing campaign. By using the ‘housing loan’ variable for the third hypothesis, we can discern the applicability of the perceived value theory.

Understanding the relationship between the 'Housing' variable and churn behavior enables us to gain insights into how housing loan obligations may impact customers' likelihood of attrition. Examining this variable alongside other demographic, socio-economic, and marketing factors allows us to uncover patterns and associations that contribute to a comprehensive understanding of customer churn dynamics.

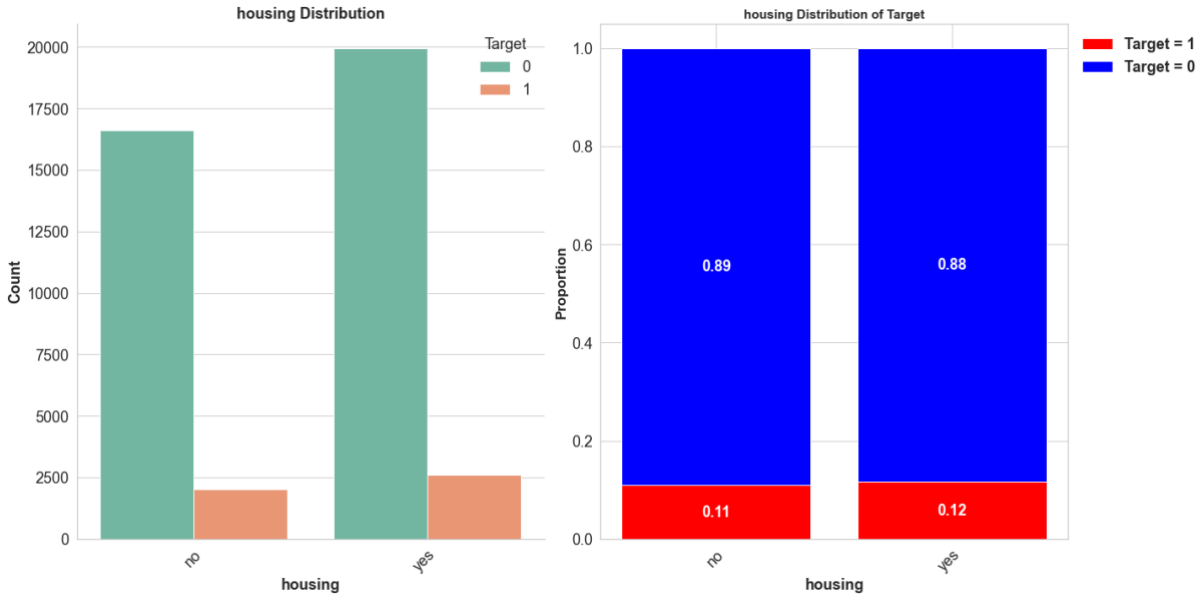

Figure 4: Count and proportions of the Housing variable

In the left plot of figure 4, we examine the distribution and proportions of the binary variable 'Housing'. It is evident that in both categories of the 'housing' variable, most customers did not subscribe to a new term deposit, indicating a churn behavior.

When we consider the proportions displayed in the right plot of figure 4, we observe that the proportions of churners and non-churners in each housing category are relatively similar. This similarity suggests that the variable 'Housing' may not possess significant predictive power in determining customers' subscription behavior.

4.1.4.2 Previous Outcome

The multi-categorical variable 'poutcome', which stands for previous outcome provides information about the outcome of the previous marketing campaign, categorizing it as 'nonexistent,' 'failure,' or 'success.' This variable plays a crucial role in understanding the effectiveness of past marketing efforts and its influence on customer churn.

The 'previous outcome' variable will be used to answer the first hypothesis, where we investigate if customers who previously had a term deposit at the bank exhibit lower churn rates in the current marketing campaign. By utilizing this variable for the first hypothesis, we can ascertain if prior adoption of the term deposit result in higher customer retention, following the switching cost theory.

Moreover, the variable will be used to answer the third hypothesis, which discerns if customers who already had a financial product at the bank exhibit reduced churn rates in the

current marketing campaign. By utilizing this variable in conjunction with the 'housing loan' variable, we can uncover the impact of prior association with the bank's financial products on customer churn, allowing us to further explore the extent of the perceived value theory.

Analyzing the 'previous outcome' variable allows us to gain insights into the success or failure of previous marketing strategies in retaining customers. By examining the impact of different campaign outcomes on customer behavior, we can identify patterns and trends that contribute to customer retention or attrition.
Understanding the relationship between the 'previous outcome' variable and customer churn behavior helps organizations refine their marketing approaches. By leveraging insights from past campaign outcomes, companies can optimize their strategies, tailor their messages, and allocate resources more effectively to increase customer retention and maximize campaign success.

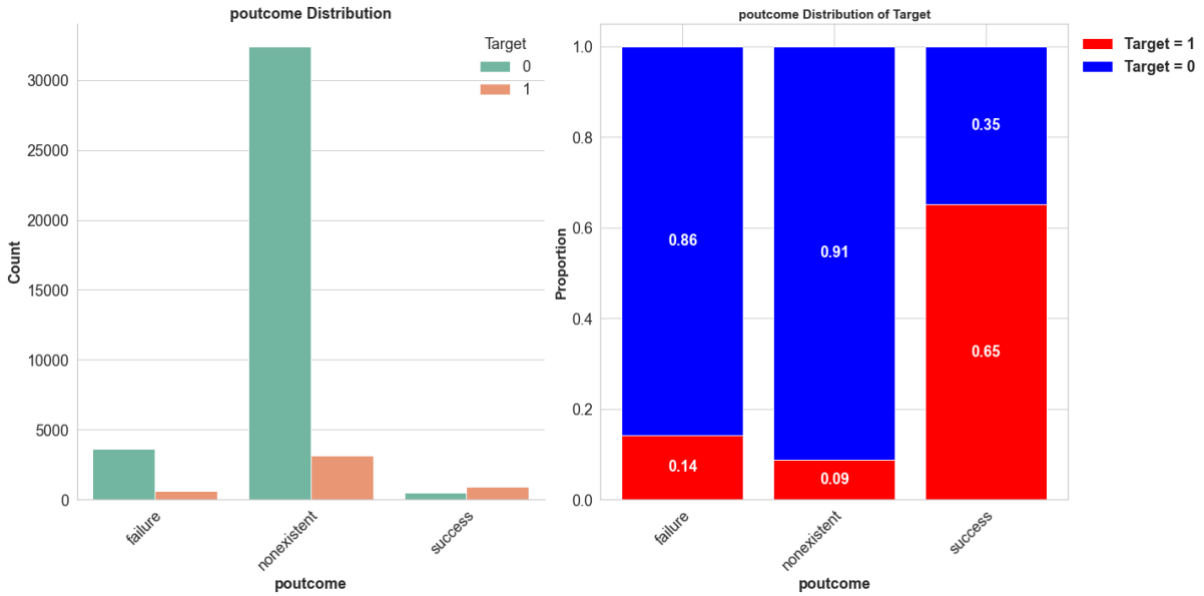

Figure 5: Count and proportions of the Previous Outcome variable

The left plot in figure 5 provides insights into the distribution of the categorical variable 'Previous Outcome,' representing the outcome of the previous marketing campaign. It is evident from the plot that most customers were not present during the previous campaign, resulting in a non-existent outcome for them. Moreover, the success category has more subscribes than non-subscribers.
Examining the right plot in figure 5, we can observe the interesting finding of the 'Success' category, which displays a higher proportion of customers subscribing to term deposits compared to the other outcome categories. This suggests a potential association between the outcome of the previous marketing campaign and the likelihood of customer subscription in the current campaign.

The difference in proportions among the outcome categories indicates that the outcome of the previous campaign could be a valuable predictor of customer behavior. Customers who had a term deposit in the previous campaign may be more inclined to subscribe to a term deposit in the current campaign, while those who had different outcomes may exhibit varying levels of interest or reluctance.

4.1.5 Missing Values

In the dataset retrieved from the UCI Machine Learning Repository, there are several variables that contain missing values. However, the dataset documentation does not provide any specific information or comments about these missing values or their nature. As a result, it is assumed that the missing values occur randomly within the dataset.

Table 3: Missing values in the banking dataset

Variable	Count Missing Values
job	330
marital	80
education	1731
default	8597
housing	990
loan	990

Note. This table shows the count of missing values for all variables in the banking dataset. The table shows that the variable default has the most missing values, followed by education.

To gain a better understanding of the extent of missing values, table 3 displays the count of missing values for each categorical variable in the dataset. The variables 'job,' 'marital,' 'education,' 'default,' 'housing,' and 'loan' have varying numbers of missing values, indicating that these variables have incomplete or unavailable information for certain observations. Missing values can arise from various sources, such as incomplete data collection during the study or non-response from clients. By addressing missing values appropriately, we can ensure the integrity of our analysis and mitigate potential biases

4.1.6 Statistical Test

4.1.6.1 Numeric

Table 4: Two-sample t-test for numeric variables

Variable	T-Statistic	P-Value
age	6.1721	0.0000***
campaign	-13.4965	0.0000***
cons.conf.idx	11.1539	0.0000***
cons.price.idx	-27.9032	0.0000***
duration	89.9672	0.0000***
emp.var.rate	-63.4337	0.0000***

euribor3m	-65.6466	0.0000***
nr.employed	-76.9845	0.0000***
pdays	-69.7221	0.0000***
previous	48.0027	0.0000***

Note. This table shows the statistical significance of the numerical variables in the dataset using the two-sample t-test, where for the p-value: * = p<0.1, ** = p<0.05, *** = p<0.01. We observe that all numeric variables are significant.

In the table 4 above, we examine the statistical significance of the numeric variables, split into a churned group if they customer did not subscribe to a term deposit and a not churned group if the customer did, using the two-sample t-test. The results of this test provide valuable insights into the differences in means between the churned and not churned groups for each variable, shedding light on factors that may contribute to customer churn.
The two-sample t-test allows us to compare the means of two independent groups and assess whether the observed differences are statistically significant.

The null and alternative hypotheses for the two-sample t-test are as follows:

- Null Hypothesis (H0): There is no significant difference in the means between the churned and not churned groups for the given numeric variable.
- Alternative Hypothesis (Ha): There is a significant difference in the means between the churned and not churned groups for the given numeric variable.

Interpreting the results, we find that for all the numeric variables analyzed, the t-statistic values are large, indicating substantial differences in means between the two groups. Furthermore, the p-values are all below the significance level, providing strong evidence to reject the null hypothesis in favor of the alternative hypothesis. These findings suggest that the differences in means for the numeric variables are statistically significant and may play a role in predicting customer churn.

For instance, the variable 'age' shows a significant difference in means, with the churned group having a higher mean age (40.91 years) compared to the not churned group (39.91 years). Similarly, the variables 'campaign', 'cons.conf.idx', 'cons.price.idx', 'duration', 'emp.var.rate', 'euribor3m', 'nr.employed', 'pdays', and 'previous' exhibit significant differences in means between the churned and not churned groups.

These results indicate that these numeric variables may serve as important factors in distinguishing between customers who churn and those who do not. By incorporating these significant numeric variables into churn prediction models, we can improve the accuracy and effectiveness of our predictions. Understanding the impact of these variables on customer behavior can guide targeted strategies for customer retention and engagement.

4.1.6.2 Binary Variables

Table 5: Two-sample Z test for proportion

Variable	Z-Score	P-Value
default	0.6172	0.5371
housing	-2.2497	0.0245**
loan	0.9064	0.3647
contact	29.3814	0.0000***

Note. This table shows the statistical significance of the binary categorical variables in the dataset using a two-sample Z test for proportion, where for the p-value: * = p<0.1, ** = p<0.05, *** = p<0.01. We observe that the 'default' and 'loan' variables are insignificant, whilst the 'housing' and 'contact' variables are significant at an alpha of 0.05.

In the table 5 above we assess the significance of various binary categorical variables in relation to churn. This is done by conducting a two-sample Z test for proportion. The results of this test provide valuable insights into the statistical significance of differences in proportions between groups, shedding light on influential factors for churn prediction.

We established the following null and alternative hypotheses for the test:

- Null Hypothesis (H0): The proportion of churners in one category of the variable is equal to the proportion of churners in another category. There is no significant difference in the proportions of churners between the two categories.
- Alternative Hypothesis (Ha): The proportion of churners in one category of the variable is not equal to the proportion of churners in another category. This suggests that there is a significant difference in the proportions of churners between the two categories.

We applied the two-sample Z test for proportion to assess the statistical significance of the observed differences. The test calculates a Z-score, which measures the deviation of the observed proportions from the expected proportions assuming equal proportions in the two groups. Additionally, the p-value indicates the probability of observing a difference as extreme as the one obtained, assuming the null hypothesis is true.

Several binary categorical variables were tested, including default, housing, loan, and contact. For each variable, we examined the Z-score and p-value to evaluate their significance in predicting customer churn.

Interpreting the results, we found that the p-value for the default variable exceeded the significance level of 0.05, indicating no significant difference in proportions between the

groups. Similarly, the loan variable showed a p-value above 0.05, suggesting no significant association with churn. However, for variables like housing and contact, the p-values were below the significance level, indicating a significant association with churn. These findings suggest that housing and contact may be important factors in predicting customer churn.

The two-sample Z test for proportion enhances our understanding of the relationship between binary categorical variables and customer churn. By assessing the statistical significance of observed differences in proportions, we can identify influential factors and develop effective machine learning models for churn prediction. The results contribute to more accurate predictions and informed decision-making in customer retention strategies. Furthermore, the findings highlight the importance of variables such as housing and contact in predicting customer churn, warranting further investigation and consideration in future studies and predictive models.

4.1.6.3 Multiple categories

Table 6: Chi-square test for independence

Variable	Chi2-Statistic	P-Value
job	960.2507	0.0000***
marital	120.7843	0.0000***
education	192.1936	0.0000***
month	3101.1494	0.0000***
day_of_week	26.1449	0.0000***
poutcome	4230.5238	0.0000***

Note. This table shows the statistical significance of the multi-categorical variables in the dataset using the chi-square test for independence, where for the p-value: * = p<0.1, ** = p<0.05, *** = p<0.01. We observe that all the multi-categorical variables are significant.

In the table 6 we explore the significance of categorical variables divided in churn and not churn groups using the chi-square test for independence. The results of this test provide valuable insights into the associations between different categories of variables and customer churn, enabling us to identify important factors for churn prediction.

The chi-square test for independence allows us to evaluate the following null and alternative hypotheses:

- Null Hypothesis (H0): There is no significant association between the categorical variable and the target variable. The proportions of churners are equal across all categories of the variable.
- Alternative Hypothesis (Ha): There is a significant association between the categorical variable and the target variable. At least one of the proportions of churners vary across different categories of the variable.

To assess the statistical significance of the associations, we conducted the chi-square test for independence. This test compares the observed frequencies in each category of the variable with the frequencies expected under the assumption of independence. The resulting chi-square statistic and p-value help us determine the strength of the association between the categorical variable and churn.

The chi-square test for independence was performed on several categorical variables, including job, marital status, education, month, day of week, and poutcome. For each variable, we examined the chi-square statistic and p-value to assess their significance in predicting customer churn.
Interpreting the results, we found that for all the variables analyzed, the p-values were below the significance level of 0.05. This indicates a significant association between these variables and the target variable. These findings suggest that the job, marital status, education, month, day of week, and 'poutcome' variables play a crucial role in predicting customer churn.

The chi-square test for independence provides valuable insights into the associations between categorical variables and customer churn. By evaluating the statistical significance of these associations, we can identify important factors for churn prediction. This information can guide the development of effective machine learning models and assist in formulating strategies for customer retention and engagement. Further exploration and consideration of these variables are warranted to enhance the accuracy of churn prediction models and improve overall customer relationship management.

4.1.6.4 Insignificant variables and machine learning models

In our pursuit of building robust machine learning models, we have chosen to drop the variables 'loan' and 'default' due to their lack of statistical significance. The decision to exclude these variables was driven by multiple factors, including their limited promise indicated by the distributions and their non-significant contribution to the predictive power of the models. By examining the distributions of 'loan' and 'default', we observed little differentiation between the categories and their respective impact on the target variable. This lack of discriminatory power, coupled with their non-significant associations in statistical tests, led us to conclude that including these variables may introduce noise and hinder the interpretability of the models.

While it is possible that these variables may have contextual relevance or interact with other predictors, our analysis did not uncover substantial evidence to support their inclusion. By excluding 'loan' and 'default', we maintain a more focused and interpretable model that is better equipped to capture the essential factors influencing the target variable.

4.1.7 Correlations

We will assess the correlations of the variables using a heatmap. A heatmap is commonly used to visualize the correlation between variables in a dataset. Correlation measures the statistical relationship between variables, indicating how changes in one variable are related to changes in another. The heatmap simplifies the interpretation of correlation matrices by color-coding cells based on the strength and direction of the correlation.

Heatmaps are widely used because they provide a clear and concise overview of complex correlation structures. The color-coded nature allows for quick identification of strong and weak correlations, as well as patterns of positive or negative relationships. This helps identify potential multicollinearity and relevant variables for further analysis (Müller & Guido, 2016).

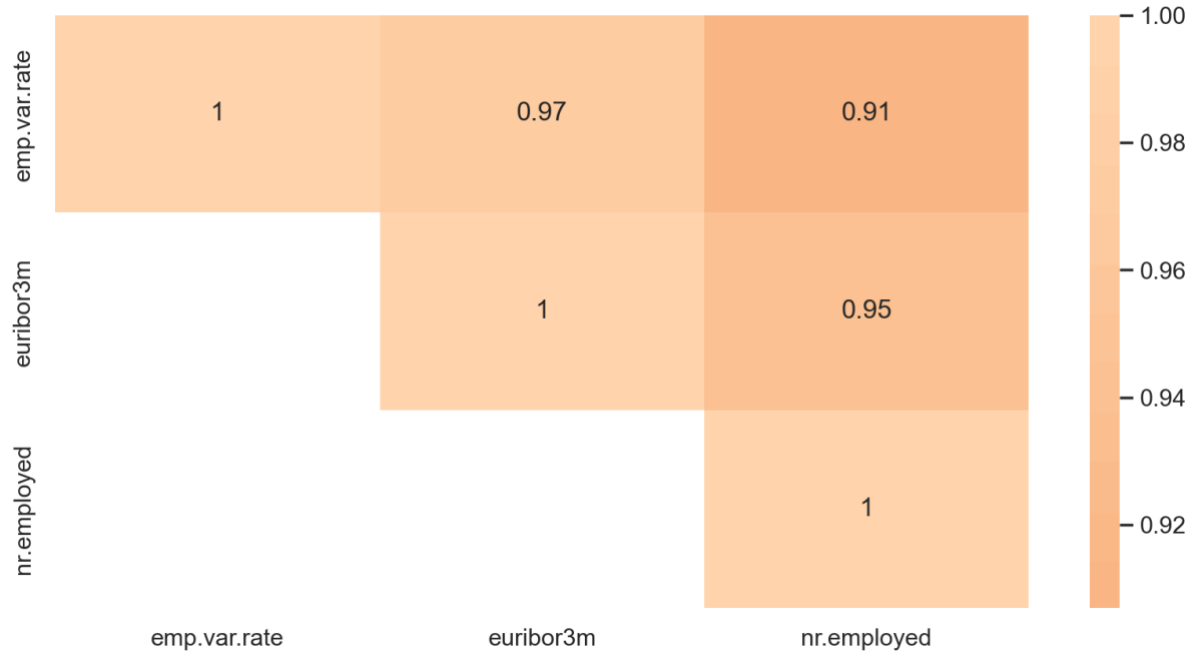

Figure 6: Heatmap of highly correlated variables

Figure 6 presents a heatmap displaying all the correlations above 0.90 among variables in the dataset. A heatmap of all the variables in the dataset can be found in Figure 32 of the appendix. Notably in figure 6, strong correlations are observed among the variables emp.var.rate, euribor3m, and nr.employed.

Emp.var.rate refers to the employment variation rate, which measures the quarterly change in employment levels. Euribor3m represents the three-month Euribor interest rate, which serves as a benchmark for euro-denominated interest rates. Nr.employed indicates the number of employees in the Portuguese bank.

While high correlations can provide valuable insights, they can also pose challenges for machine learning models. The presence of highly correlated variables can lead to

multicollinearity, where predictors become redundant, potentially resulting in unstable or biased model estimates. This multicollinearity can hinder the model's interpretability and its ability to generalize to new data.
Given the danger of high correlation, a decision has been made to remove the variables 'emp.var.rate' and 'euribor3m' from further analysis. This choice is motivated by their strong correlation with other variables, which may introduce multicollinearity. However, the variable nr.employed will be retained due to its unique nature and potential significance in predicting customer churn. Since the heatmap in Figure 32 indicates that the dataset does not have negative correlations below -0.90, we will not remove variables with extremely negative correlation.

4.2 Preprocessing

4.2.1 Packages for EDA and machine learning

In our methodology, the importation of packages plays a crucial role in enabling efficient and effective data manipulation, analysis, and modeling. These packages are widely recognized and prevalent in the data science community, providing essential tools and functionalities for various tasks. For this research, the following packages were imported:

Sci-KitLearn: Sci-KitLearn, also called sklearn, is a powerful package for machine learning tasks in the Python language. It offers a comprehensive collection of algorithms and tools for data analysis, statistical tests, feature selection, and model training. Both Śniegula et al. (2019), and Panjasuchat and Limpiyakorn (2020), in their papers about predicting customer churn in a telecom company, heavily relied on this package to construct machine learning models. The widespread usage of the package in academic studies concerning the prediction of customer churn reflect its reputation as an intuitive and versatile library that supports classification, regression, clustering, and visualization tasks.

Pandas: Pandas, as emphasized by Amuda and Adeyemo (2019) in their study about customer churn prediction in the financial sector, is a popular Python package for data manipulation and analysis. It provides flexible data structures and functions that facilitate efficient data preprocessing, transformation, and exploration. Its wide adoption by researchers, including Kaur and Kaur (2020) and Śniegula et al. (2019), underscores its significance in handling complex datasets and performing insightful analysis.

Numpy: Numpy, as acknowledged by Amuda and Adeyemo (2019), is a fundamental package for scientific computing in Python. It offers support for large arrays and matrices, along with a collection of mathematical functions. The frequent inclusion of the package in customer churn prediction studies, like Kaur and Kaur (2020) and Amuda and Adeyemo (2019), demonstrates its role in performing efficient numerical operations and supporting linear algebra computations.

Matplotlib: Matplotlib, as noted by Amuda and Adeyemo (2019), is a powerful visualization library in Python, particularly for two-dimensional plots. It provides a wide range of plot types, such as line, bar, scatter, and histogram, enabling researchers to visually explore and interpret their data. Its utilization by researchers, including Amuda and Adeyemo (2019), showcases its effectiveness in creating static, animated, and interactive visualizations.

Seaborn: Seaborn, as highlighted by Amuda and Adeyemo (2019), is a Python data visualization library built on top of Matplotlib. It offers a high-level of freedom regarding the manipulation of plot attributes, which aids in creating attractive and informative statistical graphics. Its integration with Matplotlib enhances the customizability and interpretability of visualizations, as seen by researchers like Amuda and Adeyemo (2019).

These packages complement each other in a synergistic manner. Sklearn provides a comprehensive suite of machine learning algorithms and tools, while Pandas enables efficient data manipulation and analysis. Numpy supports numerical computations, and Matplotlib and Seaborn facilitate informative data visualizations.
By leveraging the functionalities of these packages, we are equipped with a comprehensive set of tools to preprocess data, conduct in-depth data analysis, build accurate machine learning models, and effectively communicate our findings. These packages have been widely adopted in the data science community, and their utilization in academic studies supports their prevalence and effectiveness in addressing various research objectives.

4.2.2 Engineering the Target Variable

The original target variable, denoted as 'y', constituted a binary categorical variable with possible outcomes of 'yes' and 'no'. As explained in the Exploratory Data Analysis, ‘yes’ stands for having subscribed to a term deposit, and ‘no’ indicated that the customer did not subscribe and thus churned.

To facilitate a comprehensive analysis, a transformation has been applied to the target variable. Specifically, the value 'yes' has been encoded as 1, while 'no' has been encoded as 0. This conversion of the churn variable is a common practice in churn prediction studies, as observed in the works of Śniegula et al. (2019) and Kaur and Kaur (2020).
The adoption of a binary representation for customer churn aligns with established practices and empirical findings within the academic literature. Keramati, Ghaneei, and Mirmohammadi (2016) employed a binary variable to study churn in their own research, supporting the use of this approach. Similarly, Sabbeh (2018) employed customer churn as the target variable in multiple predictive models, further substantiating the suitability of a binary-dependent variable.
By building upon the scholarly foundation laid by prior researchers, our chosen approach gains credibility and reinforces the robustness of our analysis. The use of customer churn as a

binary-dependent variable is grounded in empirical evidence and accepted conventions in the field, enhancing the reliability and applicability of our findings.

4.2.3 Imputation of Missing Values with SimpleImputer

During the exploratory data analysis, it was identified that the dataset used for this study contains missing values. The presence of missing values in datasets utilized for customer churn prediction is a common occurrence, as observed in previous research conducted by Kaur and Kaur (2020), Panjasuchat and Limpiyakorn (2020), and Amuda and Adeyemo (2019).

Missing values pose a challenge for machine learning models as they can introduce biases and affect the accuracy of predictions. Incomplete data can lead to biased estimates, distorted feature importance, and hinder the model's ability to capture the true patterns and relationships within the data.

To address this issue, the SimpleImputer function from the Sklearn library was employed. The SimpleImputer replaces missing values with the most frequent value in each column, ensuring that the dataset is complete and suitable for further analysis.

Other academic papers in the field of customer churn analysis have also utilized the SimpleImputer to handle missing values. For instance, López et al. (2023), in their paper about customer churn prediction of a telecom firm in Peru, experienced missing records in the numerical variable 'INVOICING' of their dataset and used the SimpleImputer in their data preparation process. Similarly, Fujo, Subramanian, and Khder (2022), in their paper about customer churn prediction using the IBM telco dataset, encountered missing values in their numeric variables and employed the SimpleImputer to impute the missing values using each feature's mean.

The consistent adoption of the SimpleImputer technique in previous studies demonstrates its effectiveness and relevance in addressing missing values in the context of customer churn prediction.

4.2.4 OneHotEncoder for Categorical Variables

As seen in the exploratory data analysis, we have several categorical variables in our dataset that play a crucial role in understanding customer churn. Categorical variables represent qualitative attributes and often require conversion to numerical form to be effectively utilized in machine learning models.

To convert these categorical variables into a suitable format for machine learning, we employ the One-Hot Encoder, a feature encoding technique available in the Sklearn package. The One-Hot Encoder transforms categorical variables into binary vectors, enabling the representation of each category as a separate feature. This approach ensures that the raw information contained in the categorical variables is preserved while enabling compatibility with numerical-based algorithms.

Kaur and Kaur (2020) employed the One-Hot Encoder in their study to convert categorical variables into numeric ones. By utilizing this encoding technique, they were able to incorporate categorical information into their machine learning models effectively. Moreover, Fujo, Subramanian, and Khder (2022) adopted the One-Hot Encoder to convert their categorical variables into a numeric representation. They emphasized that the One-Hot Encoder is a widely employed method that utilizes binary encoding, assigning a value of 1 or 0 to indicate the presence or absence of a specific category, respectively. This conversion process transforms each category into an N-dimensional vector, where N corresponds to the number of categories within the nominal attribute. Additionally, Panjasuchat and Limpiyakorn (2020) converted all their categorical attribute values into numeric ones using the One-Hot Encoder.

Based on the findings of the aforementioned academic papers, which highlight the successful utilization of the One-Hot Encoder for converting categorical variables to numeric ones, our study will also employ this technique for handling our categorical variables. By leveraging the One-Hot Encoder, we ensure that the valuable categorical information is appropriately represented and utilized in our machine learning models for accurate customer churn prediction.

4.2.5 Handling Outliers

Outlier handling is crucial during the pre-processing stage of machine learning model building. The term outlier refers to data points that deviate heavily from the overall pattern or distribution of the dataset. Outliers can arise due to randomness, measurement errors, data entry mistakes, or rare events (Müller & Guido, 2016).

The presence of outliers in the training data can significantly impact the performance of machine learning models. Outliers in the dataset can exert a disproportionate influence on the model's fitting process, which leads to distorted parameter. These distorted parameters could dominate over other features in the dataset, resulting in a poor generalization when the model is applied to unseen data (Müller & Guido, 2016).

One way to handle the outliers is by removing them. Several academic papers that predict customer churn have included outlier removal in their preprocessing stage, with the intention of improving the performance of the churn prediction models. Notably, Lalwani et al. (2022), Keramati et al. (2016), and Khodabandehlou and Zivari Rahman (2017) have incorporated outlier removal techniques to enhance the accuracy and effectiveness of their churn prediction models. By removing their outliers, these studies intended to mitigate the influence of extreme values and improve the models' performance.

Our EDA indicated that certain variables in our dataset were identified to have outliers. To enhance the performance of our machine learning models, we have applied outlier removal techniques to the following variables:
Age: The EDA revealed the presence of outliers in the 'Age' variable. To address this, we have removed all values above 60. The cutoff point of 60 was chosen based on the left plot in Figure 15 found in the appendix, which indicated a long tail emerging from the value 60 onwards.

Campaign: The 'Campaign' variable also exhibited outliers. To improve model performance, we removed all values above 10, as indicated by the left plot in Figure 16 found in the appendix. This plot showed an extend right tail, which became prevalent around the value of 10.

Duration: Lastly, the 'Duration' variable showed outliers. The focus on increasing the performance of our models led to the removal of all values above 1000. The value of 1000 was chosen because the left plot in Figure 3 showed how the right tail became increasingly longer and thinner from this point onwards.

By removing outliers from these variables, we aim to reduce the influence of extreme values on our machine learning models. The outlier removal helps create a more reliable training dataset, allowing the models to capture the underlying patterns better, and generalize better on the test data. Considering the successful implementation of outlier removal techniques in the preprocessing phase of the academic papers mentioned before, it is valid and justified to apply similar approaches in our research.

4.2.6 Handling Imbalance with SMOTE

The EDA revealed a significant class imbalance, where the majority of customers did not subscribe to a term deposit, while the minority did. In machine learning, working with imbalanced datasets can present challenges. Traditional algorithms tend to prioritize accuracy, which can result in biased predictions favoring the majority class. Consequently, capturing patterns and making accurate predictions for the minority class becomes more challenging, particularly in domains like churn prediction (Müller & Guido, 2016).

To address the issue of class imbalance, a technique called SMOTE (Synthetic Minority Oversampling Technique) is commonly employed. SMOTE aims to balance the classes by creating synthetic examples of the minority class, thereby increasing its representation in the dataset. This technique mitigates the limitations of random oversampling, which can result in overfitting due to the replication of minority class samples.

In the paper by Mishra and Rani (2017), titled ‘Churn prediction in telecommunication using machine learning,’ the researchers recognized the challenges of class imbalance and the

drawbacks of random oversampling. They stated that SMOTE was developed to overcome the limitations of random oversampling. Although the dimensionality of their dataset was not excessively high, they used SMOTE to analyze its impact.
Similarly, in the paper by Do, Huynh, Vo, and Vu (2017) on customer churn prediction in an internet service provider, the authors addressed the class imbalance issue by applying the SMOTE oversampling technique. They highlighted the importance of reducing class imbalance before implementing predictive models and explained that SMOTE creates new instances of the minority class rather than duplicating existing ones.
The adoption of SMOTE in these academic papers, along with other studies in the field, justifies its use in our research on customer churn prediction. By applying SMOTE, we aim to improve the performance of our machine learning models by addressing the class imbalance and providing adequate representation for the minority class.

4.2.7 Mitigating Data Leakage

Data leakage refers to the inadvertent inclusion of information in the training process that would not be available during the actual prediction phase. It occurs when data from the future or information that is not representative of real-world scenarios is used, leading to overly optimistic model performance. Data leakage is a critical issue as it can result in misleading conclusions, unreliable predictions, and an inflated assessment of the model's effectiveness.

Kaur and Kaur (2020), in their paper on churn prediction in the banking industry, took precautions to avoid data leakage by appropriately splitting the dataset into training and testing data. The training data, constituting 70% of the total data, was exclusively used for model building, while the remaining 30% served as an independent evaluation set to assess the trained model's performance. This approach ensures that the model is not exposed to unseen data during the training phase, eliminating the risk of data leakage.

Moreover, Wu and Wang (2022), in their study on churn prediction for bank credit cards, implemented strategies to mitigate the risk of data leakage. The researchers utilized an 80/20 split to divide the dataset into a training set (8,101 records) and a test set (2,026 observations), ensuring independent subsets for model building and evaluation. To address data leakage concerns, Wu and Wang applied the StandardScaler from scikit-learn to standardize the numerical variables in the training set. By calculating means and standard deviations solely from the training set, they avoided incorporating information from the test set during the standardization process, safeguarding against data leakage and maintaining the integrity of the evaluation.
The original dataset exhibited class imbalance, with churned customers representing only 16.07% of the total. To handle this issue, Wu and Wang employed the Synthetic Minority Oversampling Technique (SMOTE) solely on the training set. By oversampling the minority class, they generalized its decision region, enabling the model to better capture its distinguishing characteristics.

A critical aspect highlighted by Wu and Wang was the importance of applying SMOTE exclusively to the training set after the train-test split. This approach ensured that the test set remained untainted by oversampling techniques, eliminating the introduction of duplicate or similar instances across both sets and preventing data leakage.
The careful considerations and precautions taken by Wu and Wang in their study exemplify a proactive approach to address data leakage concerns. By implementing proper data splitting, and post-split SMOTE application, they maintained the integrity of their evaluation process and produced reliable churn prediction models.

In the context of our research, precautions were taken to prevent data leakage while handling the imbalanced dataset using SMOTE. To maintain the integrity of the test set, SMOTE was exclusively applied to the training set after the train-test split. This approach avoids introducing identical or similar observations in both sets, which could lead to overfitting and overly optimistic predictions. By following this practice, in concurrence with the actions taking by Wu and Wang (2020), the test set remains representative of the original data, ensuring a reliable assessment of the model's performance without any inadvertent influence from oversampling techniques.
The adoption of such an approach aligns with the practices observed in other academic articles. Kaur and Kaur (2020) and Wu and Wang (2022) exemplify the importance of proper data splitting and preprocessing to mitigate data leakage risks.

4.3 Machine Learning

4.3.1 Logistic Regression

The first machine learning model that will be used is the Logistic Regression (LR). LR is a statistical modeling technique most often used for binary classification problems. LR is a classification algorithm, which sets it apart from a regression algorithm. In contrast to traditional regression, which focuses on predicting continuous values, LR is specifically designed for categorical or binary outcomes. It is particularly suitable for scenarios where the goal is to predict a binary outcome based on a set of independent variables. The logistic function, also known as the sigmoid function, is utilized to ensure that the predicted outcome remains within the range of 0 and 1, representing the probabilities of belonging to each class. LR is widely applicable in various scenarios. It is commonly used in binary classification problems, such as fraud detection, benign tumor detection and customer churn prediction.

An important way to combat overfitting the LR model is regularization, where the two common forms are L1 and L2 regularization. L1 regularization, or Lasso, adds the absolute value of the coefficients to the loss function. This technique promotes sparsity in the model by shrinking some coefficients to zero, effectively performing feature selection.

On the other hand, L2 regularization, known as Ridge, adds the squared magnitude of the coefficients to the loss function. It encourages a more balanced distribution of coefficient values, reducing the impact of individual variables (Müller & Guido, 2016).
The choice between L1 and L2 regularization depends on the specific problem and the desired model characteristics. The regularization strength, often denoted as 'C', influences the amount of shrinkage applied to the coefficients. A smaller 'C' value corresponds to larger shrinkage, simplifying the model and reducing the risk of overfitting. Conversely, a higher 'C' value allows the model to fit more closely to the data, potentially leading to overfitting (Müller & Guido, 2016).

In the context of customer churn prediction, Logistic Regression has been widely used by researchers. Notably, Kaur and Kaur (2020), Vo et al. (2021), and Lalwani et al. (2022) employed Logistic Regression as one of the machine learning algorithms to predict customer churn. These papers acknowledged the suitability of Logistic Regression for binary classification tasks and highlighted its ability to model the relationship between independent variables and the likelihood of customer churn.
Given the consistent usage of Logistic Regression in these studies, it is justified to utilize this technique in our own research on customer churn prediction. By building upon the findings and methodologies of these papers, our study contributes to the body of knowledge and leverages the established effectiveness of Logistic Regression in predicting customer churn.

4.3.2 Stochastic Gradient Descent

The second model that will be used to predict customer churn is the Stochastic Gradient Descent (SGD) Classifier, which is a robust linear classifier. The model tries to minimize a loss function, and this process is carried out via the stochastic gradient descent method. Both concepts will be explained below.

Firstly, gradient descent is an algorithm commonly used in machine learning to optimize and improve the performance of models. The main objective of gradient descent is to find the best set of parameters that minimize a specific measure of error or difference between the model's predictions and the actual observed data (Müller & Guido, 2016).

To further understand how gradient descent operates, let us think of it as a process of gradually adjusting the parameters of a model to reach the lowest point, or the minimum, of a hill. Here the ‘Gradient’ refers to the slope of the hill, which serves to guide the direction in which we should adjust the parameters. By iteratively moving in the opposite direction of the gradient, and thus moving to the lowest point of the hill, the algorithm fine-tunes the model's parameters to minimize the error and improve its predictive accuracy.

However, traditional gradient descent scales poorly when used on large datasets. The algorithm computes the gradient using the entire dataset at each step, which can be

computationally intensive and time-consuming. This is where Stochastic Gradient Descent comes in. SGD is a variant of gradient descent that takes a more efficient approach. Instead of using the entire dataset to compute the gradient, SGD randomly selects a small subset of data points, or even a single data point. By using a smaller portion of the data, SGD can perform updates more quickly and is better suited for large-scale and sparse machine learning problems (Müller & Guido, 2016).

Furthermore, the importance of a loss function for the SGD classifier will be discussed. The choice of a suitable loss function plays a crucial role in determining the performance of the SGD classifier. The loss function acts as a guiding principle that measures the difference between the predicted outcomes of the classifier and the actual data. Two commonly utilized loss functions are the hinge loss and the log loss functions.

The hinge loss function essentially guides the SGD classifier to correctly assign instances to their respective classes based on a designated 'margin'. The hinge loss function encourages the classifier to establish a wider margin, between the decision boundary and the data points. By doing so, it promotes a clear separation between different classes. Ultimately, the hinge loss pushes the classifier to make confident predictions that are far away from the decision boundary.
This emphasis on confident predictions away from the decision boundary enhances the classifier's effectiveness in distinguishing between different classes. It enables the classifier to achieve a more reliable and accurate classification outcome, leading to improved performance in various machine learning tasks (Shalev-Shwartz & Ben-David, 2014).

The log loss measures the difference between the predicted probabilities assigned by the classifier and the actual labels of the instances. By minimizing the log loss, the classifier is encouraged to assign higher probabilities to the correct class and lower probabilities to incorrect classes. The log loss function aids the classifier to make confident and accurate predictions. The loss function penalizes uncertain or incorrect predictions, which encourages the classifier to assign higher probabilities to the true class and lower probabilities to the other classes. Through optimizing the log loss, the classifier becomes better at distinguishing between different classes, resulting in improved performance across various machine learning applications. (Shalev-Shwartz & Ben-David, 2014).
By carefully selecting the appropriate loss function, the SGD Classifier can effectively optimize the performance of the model, leading to reliable and accurate classification outcomes in various machine learning tasks.
The SGD Classifier introduces a regularization parameter known as 'alpha', which controls the strength of regularization, similar to the 'C' parameter in logistic regression. Fine-tuning the alpha parameter is essential to achieve the ideal balance between underfitting and overfitting, ensuring the classifier's optimal performance (Müller & Guido, 2016).

The SGD Classifier has demonstrated its effectiveness in predicting customer churn, as evidenced by its frequent and successful application in academic papers. In the context of customer churn prediction, many researchers have successfully adopted the SGD Classifier in their studies, highlighting its value in this domain. Sabbeh (2018) utilized the SGD model, in her paper where she compared different machine learning models to predict customer churn in the telecom industry. Moreover, Labhsetwar (2020) utilized the SGD in a similar fashion, in his paper where he examined the performance of different machine learning models that predicted customer churn. The successful implementation of the SGD classifier by researchers in a customer churn prediction context, motivates the use of this model in our research.

4.3.3 Random Forest

The third machine learning model that will be used for churn prediction is the Random Forest Classifier (RF). RF is an ensemble learning model that is constructed from multiple decision trees. To enhance the understanding of RF, we will start by discussing the decision tree, afterwards the concept of ensemble learning will be explained.

Firstly, decision tree is a simple yet effective predictive model that resembles a flowchart-like structure. The decision tree recursively splits the data based on whether samples fulfill conditions made on the input features, for example 'is the age of the sample older or younger than 30' (Müller & Guido, 2016).
By splitting the data in such a manner, a tree-like structure is created. Here each internal node of a tree represents a condition on a feature, each branch represents outcome of the condition, and each leaf node represents a prediction or class label. The conditions that can be set in the internal node, can be of categorical or numerical nature. This attribute of the model gives it the ability to handle both feature types and capture complex relationships within the data. (Müller & Guido, 2016).

Secondly, ensemble learning is a powerful technique in machine learning where multiple models, called base learners, are combined to create a more accurate and robust predictive model. In the case of RF, the base learners are decision trees (Müller & Guido, 2016).
Random Forest, through ensemble learning, enhances the performance of the decision trees by constructing a multitude of trees and combining their predictions. Each decision tree in the RF is trained on a random subset of the training data, where the tree only has access to a random subset of the features. By using a random portion of the data and features to train, diversity amongst the trees is introduced, reducing overfitting on the dataset, and enhancing the overall predictive performance of the RF model. The final prediction of the RF is determined by aggregating the predictions of all the individual trees, where the outcome is decided through majority voting or averaging (Hearty, 2016).

One of the significant advantages of RF is its ability to handle high-dimensional data with a large number of features. It can effectively capture nonlinear relationships, handle missing

values, and perform well even with imbalanced datasets. Additionally, RF provides insights into feature importance, allowing analysts to understand the relative contribution of each feature in the prediction process (Hearty, 2016).

Several papers have successfully implemented RF for customer churn prediction, highlighting its effectiveness for this type of machine learning task. Kaur and Kaur (2020), Lalwani et al. (2022) and Vo et al. (2021) utilized RF as part of their machine learning models to predict customer churn. Similarly, López et al. (2023) and Sabbeh (2018) employed RF to analyze customer churn and retention in the telecommunications industry.
The successful implementation of RF in these studies demonstrates effectiveness of the model, when used for churn prediction tasks. The academic papers motivate the adoption of RF as a suitable model for our study. By leveraging the strengths of Random Forest and drawing insights from previous research, we aim to develop a robust and accurate predictive model for customer churn.

4.3.4 XGBoost

The last machine learning model that will be utilized to predict customer churn is the XGBoost. XGBoost, an acronym for 'Extreme Gradient Boosting,' is regarded as an advanced and highly efficient implementation of the gradient boosting algorithm.

The core concept of XGBoost is boosting. Boosting is an ensemble learning technique that combines multiple models, known as 'weak learners', because of their modest predictive power, to create a strong learner. The main premise behind boosting is to sequentially train models and use the subsequent model to correct the errors of the previous models. The models are combined through a weighted majority vote to produce the final prediction, providing a powerful approach to tackle both bias and variance in the data . By going through this iterative process multiple times, the overall predictive performance of the model increases (Hearty, 2016).

Gradient boosting is a specific type of boosting algorithm that uses gradient descent optimization to minimize the loss function. In gradient boosting, each model is trained to predict the residual errors of the previous model. By repeatedly adding new models and adjusting their weights based on the gradients of the loss function, gradient boosting iteratively builds a strong predictive model (Hearty, 2016).

These techniques bring us to XGBoost, which is an enhanced version of gradient boosting. XGBoost incorporates additional features and optimizations to deliver superior performance. It leverages advanced techniques such as parallel processing, regularization, and tree pruning to achieve exceptional predictive accuracy and speed (Hearty, 2016).

One advantage of XGBoost is its ability to handle both numerical and categorical input features effectively. It automatically handles missing values and supports various types of data transformations, reducing the need for extensive preprocessing.
Another advantage of XGBoost is its scalability. It can efficiently handle large datasets with millions of observations and thousands of input features. By utilizing parallel computing and smart memory management, XGBoost delivers fast training and prediction times.

In the context of customer churn prediction, several studies have demonstrated the effectiveness of XGBoost. To start, López et al. (2023), Vo et al. (2021), and Lalwani et al. (2022) employed XGBoost as a key machine learning algorithm to predict customer churn. These studies showcased the ability of XGBoost to handle the complexity of churn prediction tasks and achieve high predictive accuracy.

The success of XGBoost in these papers provides a strong motivation for us to adopt this model in our study. The advantages of its efficient scalability and tolerance for numeric and categorical variable makes XGBoost highly suitable for our dataset. By leveraging the proven performance of XGBoost in customer churn prediction, we can benefit from its advanced features and optimize our predictive modeling process.

4.3.5 Grid search and Hyperparameter tuning

There a several techniques that we implement when building our models. We employ grid search and hyperparameter tuning as essential techniques for optimizing the performance of our predictive models. To better understand these techniques, we will start by explaining the concept of hyperparameters.
Firstly, hyperparameters are predefined parameters that influence a machine learning model's behavior and performance. Hyperparameters are set prior to training and control aspects such as model complexity, flexibility, and regularization. Unlike internal parameters, hyperparameters are not learned from data but are specified by the user.

Grid search is a systematic approach that explores multiple combinations of hyperparameters to find the best set for our models. Before using the grid search tool, we specify in what ranges of each hyperparameters the grid search must test. The grid search exhaustively evaluates combinations of hyperparameters and gives us the set of hyperparameters that yield the highest performance, thereby 'tuning' the hyperparameters (Müller & Guido, 2016).

Academic papers on customer churn prediction, including López et al. (2023), Vo et al. (2021), Lalwani et al. (2022), and Sabbeh (2018), also utilized grid search and hyperparameter tuning in their research. By including these techniques when building their models, they identified the best configuration of hyperparameters for their predictive models, enhancing the model's performance. Their successful application of grid search and hyperparameter tuning supports the validity and relevance of using these techniques in our own research.

Regarding our research, we will apply grid search and hyperparameter tuning to our Logistic Regression, SGD, Random Forest, and XGBoost models. By exploring different combinations of hyperparameters and selecting the optimal configuration of values, we intend to enhance the performance and predictive capabilities of each model.

4.4 Evaluation Metrics

4.4.1 Confusion Matrix

The first tool that will be used to evaluate the models is the confusion matrix. The confusion matrix, commonly used in the field of classification, provides a comprehensive summary of the performance of a machine learning model (Müller & Guido, 2016).
The confusion matrix is composed of a square, where the predicted classifications and actual classification of samples are represented. The matrix consists of four key components: true positives, true negatives, false positives, and false negatives (Müller & Guido, 2016).

Firstly, true positives (TP) refer to the number of instances that were correctly classified as positive by the model. These are the cases where the model predicted a positive outcome, and the actual result was also positive. In the case of binary classification, the positive class would take on the value of 1 (Müller & Guido, 2016).
Secondly, true negatives (TN) represent the instances that were correctly classified as negative by the model. In these cases, the model predicted a negative outcome, and the actual result was indeed negative. The negative class, when viewed from a binary context, would be denoted as the value 0 (Müller & Guido, 2016).
Thirdly, false positives (FP) occur when the model incorrectly predicts a positive outcome, but the actual result is negative. These failed prediction are called Type I errors. To elaborate, one could think of a type I error as a medical test incorrectly diagnosing a healthy patient as having a disease, leading to unnecessary treatment and anxiety (Müller & Guido, 2016). Lastly, false negatives (FN) occur when the model incorrectly predicts a negative outcome, but the actual result is positive. These failed predictions are denoted as Type II errors. An example of a type II error could be a security system failing to detect an intruder, leading to a false sense of security and potential risks to the property or individuals (Müller & Guido, 2016).

The confusion matrix is frequently used to evaluate classification tasks because it provides a detailed breakdown of the model's performance across different classes. Moreover, the TP, TN, FP, and FN values can be used to calculate other performance metrics, which will be discussed in the next section (Shalev-Shwartz & Ben-David, 2014).

Furthermore, the confusion matrix has practical applications in cost-sensitive learning, where there are associated costs or consequences of different types of errors. For instance, in the

medical sector, misclassifying a critical illness as a non-illness (FN) may have severe consequences. On the other hand, misclassifying a non-illness as a critical illness (FP) could lead to unnecessary medical procedures and costs. Similarly, regarding the classification of fraudulent transaction the banking sector, misclassifying a fraudulent transaction as legitimate (FN) may result in financial losses, while misclassifying a legitimate transaction as fraudulent (FP) may inconvenience the customer. The confusion matrix helps assess the trade-offs and optimize the model's performance based on the associated costs and risks in such scenarios.

The use of the confusion matrix in empirical studies that predict customer churn is highly prevalent. Most of the customer churn prediction studies referenced in our research use the confusion matrix in their evaluation. Notable examples include the works of López et al. (2023), Kaur and Kaur (2020), Vo et al. (2021), Lalwani et al. (2022), Labhsetwar (2020), Keramati et al. (2016), Khodabandehlou and Zivari Rahman (2017), and Vafeiadis et al. (2015). The wide adoption of the confusion matrix in empirical studies that predict customer churn motivates the usage of the evaluation tool in our research.

4.4.2 Classification Report

The second evaluation tool that will be used in this research is the classification report. The classification report provides an overview of performance metrics for a classification models. It offers valuable insights into the model's accuracy, precision, recall, and F1-score, allowing for a deeper understanding of the model's classification prowess.

Firstly, the accuracy metric measures the overall correctness of the model's predictions. Accuracy is calculated by dividing the total number of correct predictions by the total number of correct and incorrect predictions summed. The metric ranges from 0 to 1, where 1 indicates that the model has perfect accuracy. In general, the correct predictions are the TP and TN values summed, while the incorrect prediction are the FP and FN summed. A higher accuracy value indicates a higher proportion of correct predictions, while a lower value suggests a higher rate of misclassifications (Müller & Guido, 2016).
The second metric is precision, which assesses the model's ability to correctly classify positive predictions. Precision is calculated by dividing TP by the sum of TP and FP, and ranges from 0 to 1, where a value of 1 signifies a perfect precision score. Precision provides insights into how precise the model is in identifying positive instances, indicating the proportion of correctly predicted positive instances among all the positive predictions made by the model (Müller & Guido, 2016).

The third metric provided by the classification report is recall. Recall, also dubbed sensitivity or true positive rate, evaluates the model's capability to correctly identify positive predictions. It is calculated by dividing the number of TP by the sum of TP and FN predictions, and ranges between 0 and 1, where the value of 1 is synonymous for a perfect recall score. Recall measures

the proportion of correctly predicted positive instances among all actual positive instances, highlighting the model's sensitivity to detecting positive instances (Müller & Guido, 2016).
The last metric that is included in the classification report is the F1-score. The F1-score combines the precision and recall metrics to provide an overall assessment of a classification model's performance. By taking the harmonic mean of the two metrics, the F1-score provides a balanced evaluation of the model's ability to make accurate positive predictions and capture all positive instances. The F1-score ranges from 0 to 1, with 1 representing perfect precision and recall, and 0 indicating poor performance in correctly predicting positive instances (Müller & Guido, 2016).

In the context of customer churn prediction, the classification report has been utilized in several academic papers. Notably, Kaur and Kaur (2020), López et al. (2023), Lalwani et al. (2022), and Vo et al. (2021) included the classification report, when they discussed the results of their customer churn prediction models. These studies used the metrics presented in the classification report analysis of the model's performance, gaining a better understanding of their models.
Given the successful implementation of the classification report in these papers, it is valid and beneficial to utilize the classification report as an evaluation tool in our research on customer churn prediction. By utilizing the metrics provided by the classification report, we can gain a better understanding of our model's performance.

4.4.3 Receiver Operating Characteristic Curve

The third evaluation tool that will be used to analyze our predictive models is the Receiver Operating Characteristic (ROC) curve. The ROC is widely used evaluation metrics in classification tasks that measure the performance of machine learning models in distinguishing between classes.

The ROC is a curve that graphically represents the performance of models at different classification thresholds. The ROC curve plots the true positive rate (TPR), which is the proportion of correctly classified positive instances, against the false positive rate (FPR), which is the proportion of incorrectly classified negative instances, as the classification threshold varies. TPR, also called sensitivity, is calculated by dividing the number of true positive predictions by the sum of true positives and false negatives. FPR, also known as specificity, is calculated by dividing the number of false positive predictions by the sum of false positives and true negatives (Müller & Guido, 2016).

The ROC curve provides insights into the trade-off between sensitivity and specificity across various decision thresholds. By adjusting the decision threshold, the balance between correctly classifying positive instances and incorrectly classifying negative instances can be controlled. The curve visualizes the discriminatory power of the model, by displaying how the TPR and FPR change with different threshold values. Models with a better classification

performance will have a higher TPR and a lower FPR across various threshold settings, resulting in a curve that is closer to the top left corner of the plot (Müller & Guido, 2016).
Analyzing the ROC curve aids in finding the optimal threshold for our classification task. The point on the ROC curve that is closest to the top left corner represents the threshold with the best balance between sensitivity and specificity (Müller & Guido, 2016).

4.4.4 Area Under Curve

The ROC curve is often paired with the area under the curve (AUC) metric. The AUC is a scalar value that summarizes the overall performance of the machine learning model by calculating the area under the ROC curve. The AUC measures the model's ability to order instances correctly, regardless of the classification threshold. The AUC ranges from 0 to 1, where a value of 1 indicates perfect classification performance, while a value of 0.5 indicates that the model is a random classifier. This means that there is no discrimination in the model, which could be seen as the model flipping coins to assign classes (Müller & Guido, 2016).

The adoption of the ROC and AUC in academic papers that predict customer churn to convey the results of their models supports the validity and applicability of these evaluation tools in our research. The empirical studies, including Kaur and Kaur (2020), López et al. (2023), Lalwani et al. (2022), Vo et al. (2021), Keramati et al. (2016), and Labhsetwar (2020), highlight the significance of ROC and AUC in evaluating and comparing customer churn prediction models. By including the ROC curve and AUC in our arsenal of evaluation tools, we can assess and validate the performance of our own models.

4.4.5 Feature Importance Bar Chart

The last evaluation tool used to gain further insights into our predictive models is the feature importance bar chart. The chart provides valuable insights into the relative importance of features that the churn prediction models consider when predicting customer churn.

For logistic regression and SGD models, the feature importance bar chart is calculated using the absolute values of the coefficients associated with each feature. The coefficients are then sorted in descending order and the top ten largest values are selected. These coefficients are then seen as the most influential features. For the logistic regression and the SGD model, it is the case that larger absolute coefficient values indicate a stronger impact on predicting customer churn. By examining the chart, we can identify the features that possess the highest coefficients and assess their importance within the model (Shalev-Shwartz & Ben-David, 2014).
Regarding the random forest model, the feature importance bar chart is generated using the '.feature_importances_' attribute provided by the sklearn package. This attribute calculates the importance of each feature based on their contribution to reducing impurity within the decision trees of the ensemble model. Features with a high importance value have a greater relevance in predicting customer churn. By using the chart, we can observe which features

have the highest importance values and analyze their significance within the model (Shalev-Shwartz & Ben-David, 2014).

The XGBoost models utilizes the '.get_score()' method from the xgboost package to obtain the feature importance values needed to create the bar chart. The method from the xgboost calculates the importance of each feature based on its frequency of appearance in the model's base learner trees. The features that are used more frequently for splitting are given a higher importance value. By observing the features with the highest importance values, we can see which features have the greatest influence on the model (Shalev-Shwartz & Ben-David, 2014).

The adoption of Feature Importance Bar Charts in academic research papers that predict customer churn further supports their relevance and applicability in our study. Papers such as Vo et al. (2021), Sabbeh (2018), Ahmad et al. (2019), and others have utilized these charts to convey the results of their customer churn prediction models. By incorporating the Feature Importance Bar Chart in our analysis, we can effectively evaluate and compare the significance of features in our models.

5 Analysis

5.1 Logistic Regression

5.1.1 Optimal Hyperparameters and Accuracy

Firstly, the tuned hyperparameters of the logistic regression model that are obtained using the grid search are {'C': 10, 'penalty': 'l1', 'solver': 'liblinear'}. Here, the parameter 'C' represents the inverse of regularization strength, with higher values indicating weaker regularization. Since the tuned hyperparameter has the value of 10, this indicates that the model performed best with moderate to weak regularization. The second hyperparameter is 'penalty', which indicates the type of regularization, has the value 'l1'. A value of 'l1' represents L1 regularization, better known as Lasso. This type of regularization encourages sparsity in feature selection, as explained in the methodology. Lastly, the 'solver' hyperparameter refers to the algorithm used for optimization, and in this case, 'liblinear' was used.
The accuracy of the logistic regression model is 0.9228471230029874, indicating that the model correctly predicted the outcome for approximately 92.28% of the samples.

5.1.2 Understanding Confusion Matrix of Logistic Regression

Table 7: Confusion Matrix of Logistic Regression

	Predicted Not Sub	**Predicted Sub**
Actual Not Sub	6808	166
Actual Sub	428	297

Note. This table shows the confusion matrix of the Logistic Regression. Here, the 'Predicted Not Sub' category refers to the number of instances where the model predicted that the customers would not subscribe to the term deposit, while the 'Predicted Sub' category refers to the number of instances where the model predicted that the customers would subscribe. The 'Actual Not Sub' category represents the actual number of customers who did not subscribe, and the 'Actual Sub' category represents the actual number of customers who did subscribe. The values in each cell indicate the count of instances that fall into each category, providing an overview of the model's performance in predicting customer subscriptions.

Table 7 reveals that the model correctly classified 6808 instances as 'Not Subscribed' (true negatives) and 297 instances as 'Subscribed' (true positives). However, it misclassified 166 instances as 'Subscribed' (false positives) and 428 instances as 'Not Subscribed' (false negatives). The findings of the table provide valuable insights into the model's performance. The high number of true negatives and true positives indicates its ability to accurately identify customers who will not subscribe and those who will subscribe, respectively. However, the presence of false positives and false negatives suggests areas where the model can be further refined to improve its predictive capabilities.

5.1.3 Dissecting Classification Report of Logistic Regression

Table 8: Classification Report of Logistic Regression

	Precision	**Recall**	**F1-Score**	**Support**
Not Subscribed	0.94	0.98	0.96	6974
Subscribed	0.64	0.41	0.5	725
Accuracy	-	-	0.92	7699
Macro avg	0.79	0.69	0.73	7699
Weighted avg	0.91	0.92	0.92	7699

Note. This table presents the classification report of the model's performance. The 'precision' score measures the accuracy of the model's predictions for each class, where a higher value indicates a higher proportion of correct predictions. The 'recall' score represents the model's ability to correctly identify instances of each class. The 'f1-score' provides a balance between precision and recall, combining both measures into a single value. The 'support' column displays the number of instances in each class. The 'Accuracy' score indicates the overall accuracy of the model and can be interpreted as the percentage of correctly classified instances. The 'Macro avg' score provides an overall evaluation of the model's performance across both classes, considering their individual contributions. The 'Weighted avg' score accounts for class imbalance and provides an overall evaluation, considering the support for each class.

Table 8 shows the classification report of the logistic regression model. Firstly, for the class 'Not Subscribed,' the precision is 0.94, indicating that 94% of the instances classified as 'Not Subscribed' were correctly predicted. The recall is 0.98, indicating that the model correctly identified 98% of the instances belonging to this class. The F1-score, which combines precision

and recall into a single metric, is 0.96, suggesting a high level of accuracy in predicting instances of the 'Not Subscribed' class. The support value of 6974 indicates the number of instances in this class.

For the class 'Subscribed,' the precision is 0.64, signifying that 64% of the instances classified as 'Subscribed' were correctly predicted. The recall is 0.41, indicating that the model identified only 41% of the instances belonging to this class. The F1-score for this class is 0.50, suggesting that the model's performance in predicting instances of the 'Subscribed' class could be improved.

The macro average F1-score, calculated by taking the average of the F1-scores of both classes, is 0.73. The weighted average F1-score, which accounts for class imbalance, is 0.92.

These results highlight the model's ability to accurately predict instances of the 'Not Subscribed' class, which are the churners. However, the model shows lower performance in predicting instances of the 'Subscribed' class, as evidenced by lower precision, recall, and F1-score. Improving the model's performance in predicting the 'Subscribed' class could be a focus for further refinement and optimization.

5.1.4 ROC AUC Evaluation of Logistic Regression

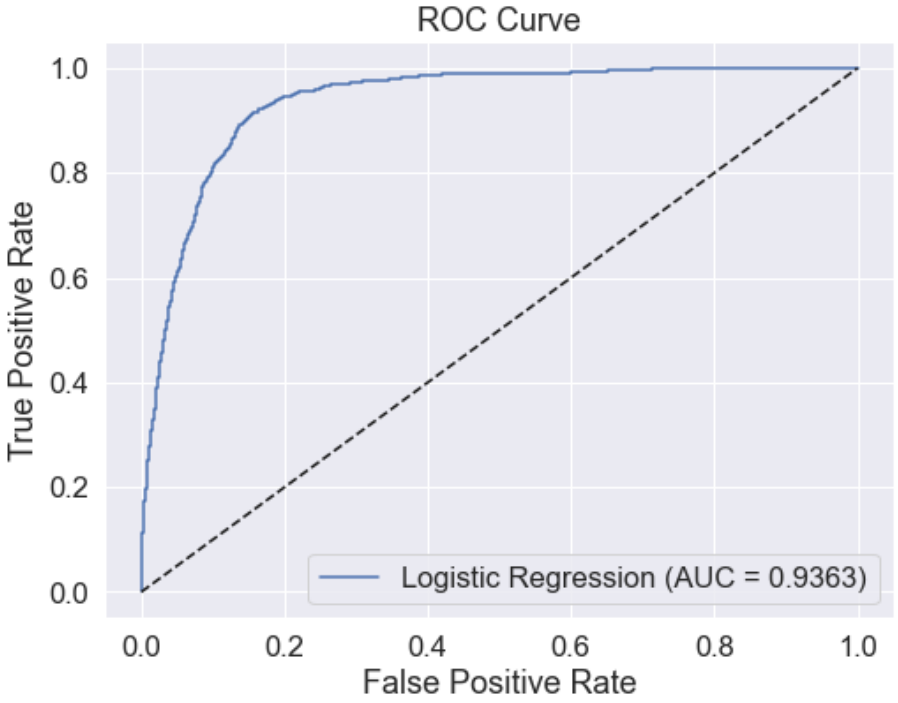

Figure 7: The Receiver Operating Characteristic curve (ROC) with corresponding Area Under the Curve (AUC)

The ROC curve, shown in figure 7, indicates that the logistic regression model has an AUC score of 0.9363. The high AUC value indicates that the model demonstrates a strong discriminatory power and is capable of accurately ranking samples in terms of their predicted probabilities.

5.1.5 Feature Importance of Logistic Regression

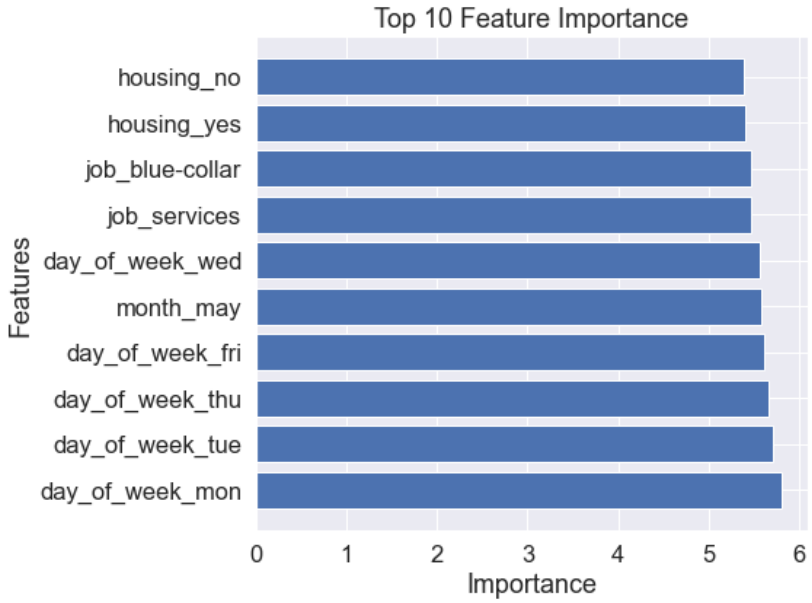

Figure 8: Top 10 important features for the Logistic Regression model

The feature importance bar chart, observed in figure 8, highlights the top 10 important features used by the model to predict customer subscription to a term deposit. Notably, most of the important features revolve around the days of the week, indicating that the day of the week in which the last contact was made helps the model predict the class in which the customers belongs to. Additionally, the type of job and whether the customer has a housing loan emerge as crucial indicators in the prediction process.

Moreover, all the importance values in the chart are relatively close to each other. This suggests that the model attributes similar levels of importance to these features when making predictions. Furthermore, the implementation of L1 regularization has effectively constrained the coefficient values, keeping them close to zero and preventing any dominant or overpowering feature influences.

5.2 Stochastic Gradient Descent

5.2.1 Optimal Hyperparameters and Accuracy

The optimal hyperparameters of the SGD model were an alpha value of 0.001, a loss function of 'log', and an L1 penalty. These hyperparameters, chosen by the grid search, enhance the model's performance and improve its predictive capabilities.

The accuracy of the SGD model was measured to be 0.8983, indicating that the model correctly predicted the outcome for approximately 89.83% of the instances. The model achieved a reasonably high accuracy, suggesting that it has the potential to make accurate predictions.

5.2.2 Understanding Confusion Matrix of SGD

Table 9: Confusion Matrix of SGD

	Predicted Not Sub	Predicted Sub
Actual Not Sub	6457	517
Actual Sub	266	459

Note. This table shows the confusion matrix of the Logistic Regression. Here, the 'Predicted Not Sub' category refers to the number of instances where the model predicted that the customers would not subscribe to the term deposit, while the 'Predicted Sub' category refers to the number of instances where the model predicted that the customers would subscribe. The 'Actual Not Sub' category represents the actual number of customers who did not subscribe, and the 'Actual Sub' category represents the actual number of customers who did subscribe. The values in each cell indicate the count of instances that fall into each category, providing an overview of the model's performance in predicting customer subscriptions.

The confusion matrix of the stochastic gradient descent (SGD) model, shown in table 9, reveals that the model achieved 6457 true negatives, correctly identifying instances as 'Not Subscribed'. This indicates that the model has a proficiency for predicting customer churn. However, the model also produced 517 false positives, misclassifying instances as 'Subscribed' when they were actually 'Not Subscribed.'

The model exhibited challenges in accurately classifying instances as 'Subscribed,' as indicated by the presence of 266 false negatives. These false negatives represent missed opportunities, where the model failed to identify customers who had actually subscribed to the service. Surprisingly, the model achieved 459 true positives, correctly identifying instances as 'Subscribed.' The model thus predicted more true positives than false negatives, indicating that the model's ability to correctly identify instances as 'Subscribed' outweighed its tendency to incorrectly classify some instances.

5.2.3 Dissecting Classification Report of SGD

Table 10: Classification report of SGD

	Precision	Recall	F1-Score	Support
Not Subscribed	0.96	0.93	0.94	6974
Subscribed	0.47	0.63	0.54	725
Accuracy	-	-	0.9	7699
Macro avg	0.72	0.78	0.74	7699
Weighted avg	0.91	0.9	0.9	7699

Note. This table presents the classification report of the model's performance. The 'precision' score measures the accuracy of the model's predictions for each class, where a higher value indicates a higher proportion of correct predictions. The 'recall' score represents the model's ability to correctly identify instances of each class. The 'f1-score' provides a balance between precision and recall, combining both measures into a single value. The 'support' column displays the number of instances

in each class. The 'Accuracy' score indicates the overall accuracy of the model and can be interpreted as the percentage of correctly classified instances. The 'Macro avg' score provides an overall evaluation of the model's performance across both classes, considering their individual contributions. It calculates the unweighted average scores for each class, providing an equal contribution to each class regardless of its size. The 'Weighted avg' score accounts for class imbalance and provides an overall evaluation, considering the support for each class.

The classification report in table 10 demonstrates a high precision score of 0.96 for instances of 'Not Subscribed', indicating that when the model predicts a customer will not subscribe, it is correct 96% of the time. Moreover, the model exhibits a high recall rate of 0.93 for 'Not Subscribed,' indicating that it correctly identified 93% of all actual 'Not Subscribed' instances. The f1-score also reflects a strong performance of 0.94 for 'Not Subscribed.'
Furthermore, the model shows relatively lower precision and recall scores for predicting instances as 'Subscribed.' The precision score of 0.47 suggests that when the model predicts a customer will subscribe, it is accurate 47% of the time. The recall score of 0.64 indicates that the model may miss a significant number of actual 'Subscribed' instances. Consequently, the f1-score for 'Subscribed' is quite low, at 0.54, reflecting the model's challenges in correctly identifying positive instances. This weakness of the model was also seen in the confusion matrix

Moreover, it is important to note that the macro-average f1-score is 0.74, suggesting room for improvement in balancing the performance between the two classes, since the model predicts the 'Not Subscribed' class better. The weighted-average f1-score is 0.9, suggesting that the model has achieved a reasonably balanced performance across precision and recall, with a strong overall harmonic mean.

5.2.4 ROC AUC Evaluation of SGD

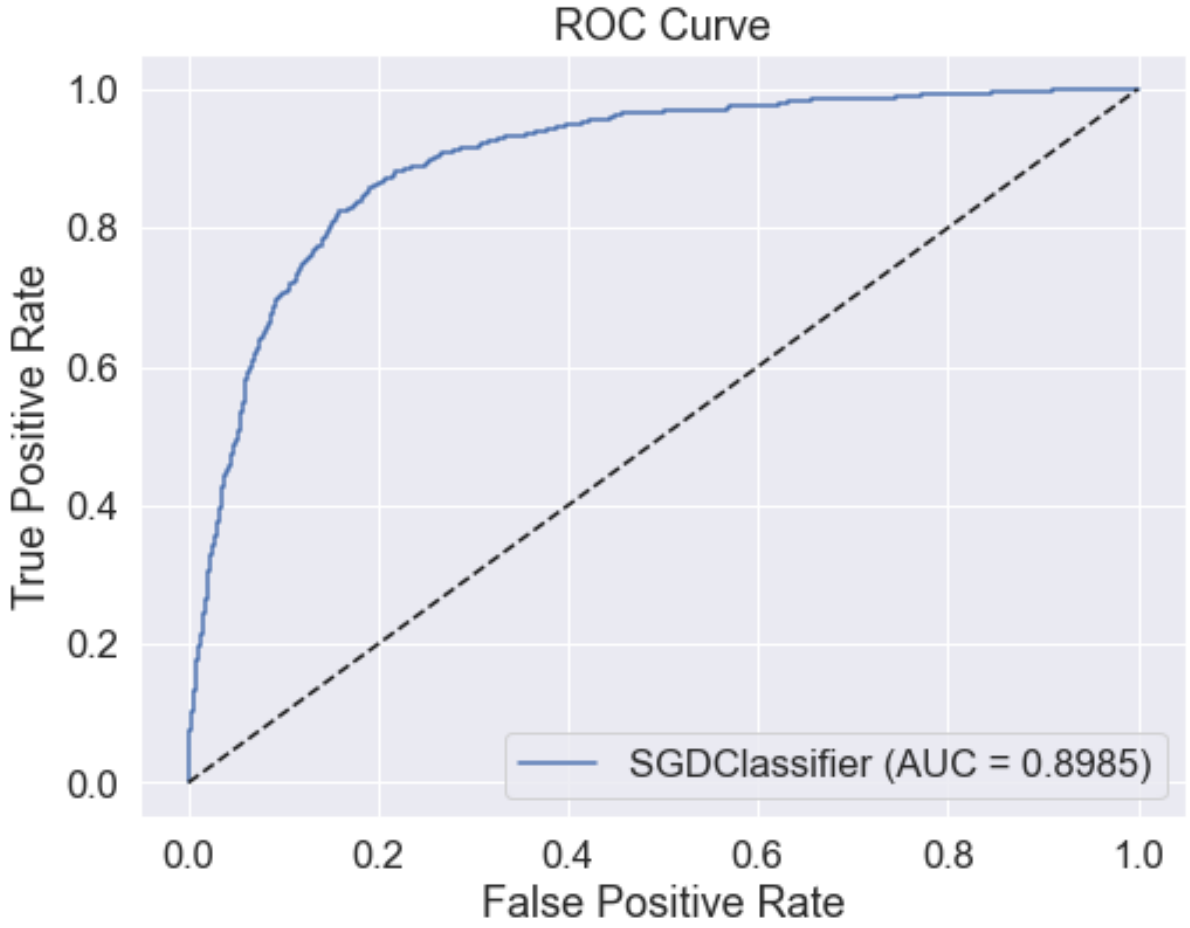

Figure 9: The Receiver Operating Characteristic curve (ROC) with corresponding Area Under the Curve (AUC)

The ROC curve of the SGD model, shown in figure 9, has an AUC score of 0.8985, indicating a reasonably good performance in predicting customer subscriptions.
the AUC of 0.8985 suggests that the SGD model has a strong ability to differentiate between customers who subscribed to a term deposit and those who did not. The high AUC score indicates that the model is able to rank instances correctly and make accurate predictions, which is crucial in customer subscription prediction tasks.

With an AUC score above 0.5, the SGD model outperforms random guessing and demonstrates a significant predictive capability.

5.2.5 Feature Importance of SGD

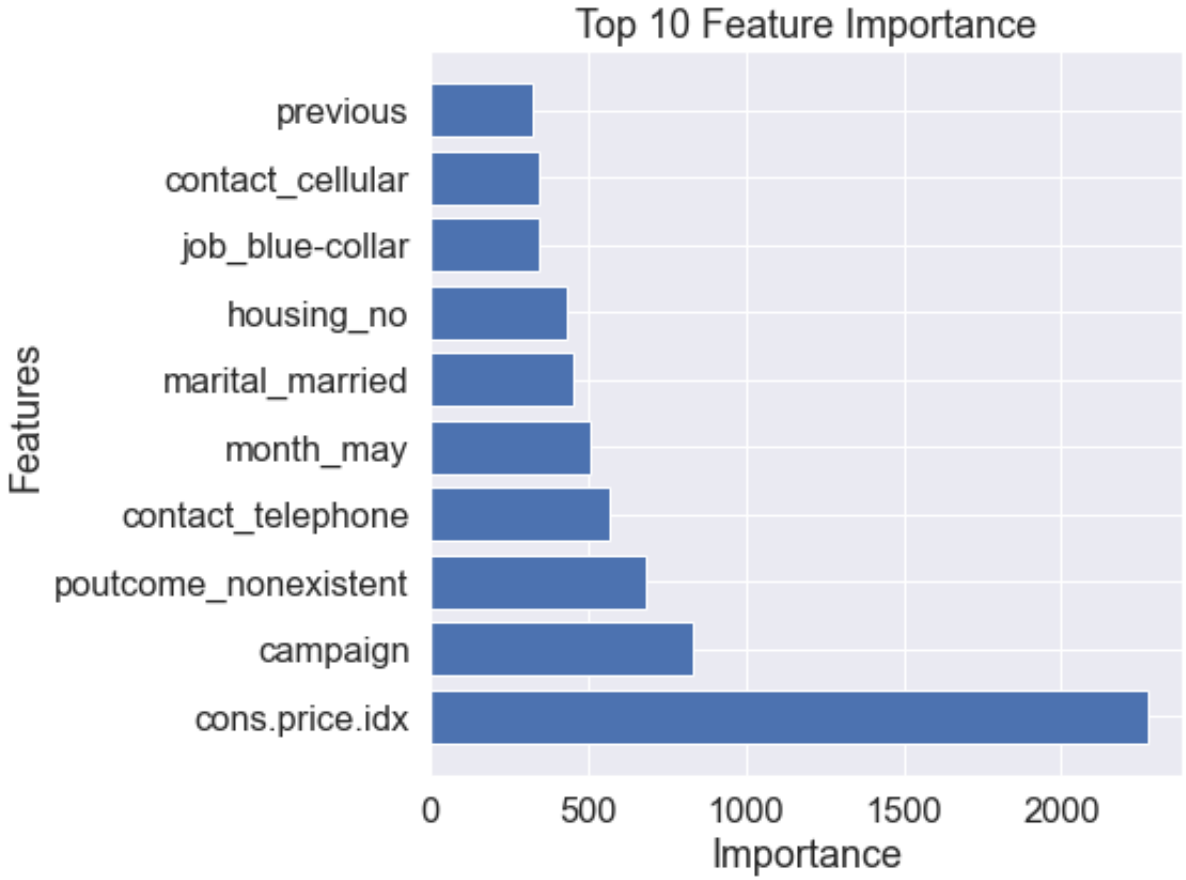

Figure 10: Top 10 important features for the SGD model

The feature importance bar chart for SGD model, displayed in figure 10, highlights the 10 most significant features used by the model for predicting customer churn. Among the important features, the consumer price index (CPI) emerges as the most influential feature, meaning that the model heavily relies on the CPI feature to differentiate between the classes. The second most important feature is 'campaign', which represents the number of contacts made during the current marketing campaign for each client. This suggests that the frequency of contacts plays a crucial role in determining customer churn.

The third most important feature used by the model is the outcome of the previous marketing campaign. The model uses the category for the customer that were not present for the previous campaign to help predict whether the person subscribed or not.
Additionally, the contact method used during the current marketing campaign emerges as another influential feature.

It is worth noting that all these important features identified by the SGD model are statistically significant. This means that there is an association between these features and the target variable, indicating their relevance in predicting customer churn. Further exploration and research into these important features could provide valuable insights into the underlying factors and patterns driving customer churn.

5.3 Random Forest Model

5.3.1 Optimal Hyperparameters and Accuracy

The Random Forest model was used with the following tuned hyperparameters: 'max_features' set to 'auto', 'min_samples_leaf' at 0.15, and 'n_estimators' set to 150.
The chosen value ‘auto’ of 'max_features' suggests that the model automatically selects the optimal number of features to consider at each split, resulting in a well-performing model. The 'min_samples_leaf' parameter indicates that each leaf node in the decision trees must contain a minimum of 15% of the total samples, which helps prevent overfitting by ensuring a sufficient number of samples in each leaf.

The high number 150 for the 'n_estimators' hyperparameter implies that the model consists of an ensemble of 150 individual decision trees, contributing to the model's overall performance. By aggregating the predictions of multiple trees, the Random Forest model can reduce the impact of individual tree errors and improve generalization.

The accuracy score of 0.8369 indicates that the model has achieved a moderate level of overall accuracy in predicting the target variable.

5.3.2 Understanding Confusion Matrix of Random Forest

Table 11: Confusion Matrix of Random Forest

	Predicted Not Sub	**Predicted Sub**
Actual Not Sub	5787	1097
Actual Sub	158	567

Note. This table shows the confusion matrix of the Logistic Regression. Here, the ‘Predicted Not Sub’ category refers to the number of instances where the model predicted that the customers would not subscribe to the term deposit, while the ‘Predicted Sub’ category refers to the number of instances where the model predicted that the customers would subscribe. The ‘Actual Not Sub’ category represents the actual number of customers who did not subscribe, and the ‘Actual Sub’ category represents the actual number of customers who did subscribe. The values in each cell indicate the count of instances that fall into each category, providing an overview of the model's performance in predicting customer subscriptions.

The Confusion Matrix presented in Table 11 provides a comprehensive overview of the model's classification outcomes. Among the predictions made by the model, there were 567 True Positives, indicating instances where the model correctly identified customers who had actually subscribed to the service. This demonstrates the model's effectiveness in capturing positive cases and identifying customers who are likely to subscribe.

Conversely, the model also produced 1097 False Positives, which signifies instances where the model incorrectly classified customers as positive (Subscribed) when they were

actually negative (Not Subscribed). This indicates a certain level of misclassification, meaning that the model's performance in identifying positive cases is not perfect.

The random forest model exhibited 158 False Negatives, representing instances where the model failed to recognize customers who had indeed subscribed to the service. Although this suggests some missed opportunities, the relatively low number of false negatives indicates that the model still has a reasonable ability to detect positive cases.

Lastly, the Confusion Matrix includes 5877 True Negatives, indicating instances where the model accurately identified customers as negative (Not Subscribed) when they were indeed negative. These correct negative predictions demonstrate the model's proficiency in distinguishing instances that are unlikely to result in subscriptions.

5.3.3 Dissecting Classification Report of Random Foreset

Table 12: Classification Report of Random Forest

	Precision	**Recall**	**F1-Score**	**Support**
Not Subscribed	0.97	0.84	0.9	6974
Subscribed	0.34	0.78	0.47	725
Accuracy	-	-	0.84	7699
Macro Avg	0.66	0.81	0.69	7699
Weighted Avg	0.91	0.84	0.86	7699

Note. This table presents the classification report of the model's performance. The 'precision' score measures the accuracy of the model's predictions for each class, where a higher value indicates a higher proportion of correct predictions. The 'recall' score represents the model's ability to correctly identify instances of each class. The 'f1-score' provides a balance between precision and recall, combining both measures into a single value. The 'support' column displays the number of instances in each class. The 'Accuracy' score indicates the overall accuracy of the model and can be interpreted as the percentage of correctly classified instances. The 'Macro avg' score provides an overall evaluation of the model's performance across both classes, considering their individual contributions. It calculates the unweighted average scores for each class, providing an equal contribution to each class regardless of its size. The 'Weighted avg' score accounts for class imbalance and provides an overall evaluation, considering the support for each class.

The Classification Report, depicted in table 12, indicates that for the 'Not Subscribed' class, the model achieved a precision of 0.97, indicating that the model correctly classified a high proportion of instances as 'Not Subscribed' out of all the instances predicted as positive. The recall for this class was 0.84, indicating that the model successfully identified a considerable proportion of actual 'Not Subscribed' instances. The corresponding f1-score, which combines precision and recall, was 0.90, suggesting a reasonable balance between precision and recall for the 'Not Subscribed' class.

In contrast, for the 'Subscribed' class , the model exhibited lower performance. The precision of 0.34 suggests that the model correctly identified only a fraction of instances as

'Subscribed' out of all the instances predicted as such. The recall for this class was 0.78, indicating that the model captured a substantial proportion of actual 'Subscribed' instances. However, the lower precision value suggests a higher rate of misclassification. The f1-score for the 'Subscribed' class was 0.47, reflecting a trade-off between precision and recall.

The macro average f1-score of the model was 0.69, indicating a moderate level of performance in capturing the balance between precision and recall for both classes. On the other hand, the weighted average f1-score was calculated as 0.86, suggesting a good overall performance in capturing the trade-off between precision and recall while considering the varying class sizes. The higher weighted average f1-score suggests that the model is performing well in terms of capturing the predictive power of both classes while accounting for their imbalanced representation in the dataset.

5.3.4 ROC AUC Evaluation of Random Forest

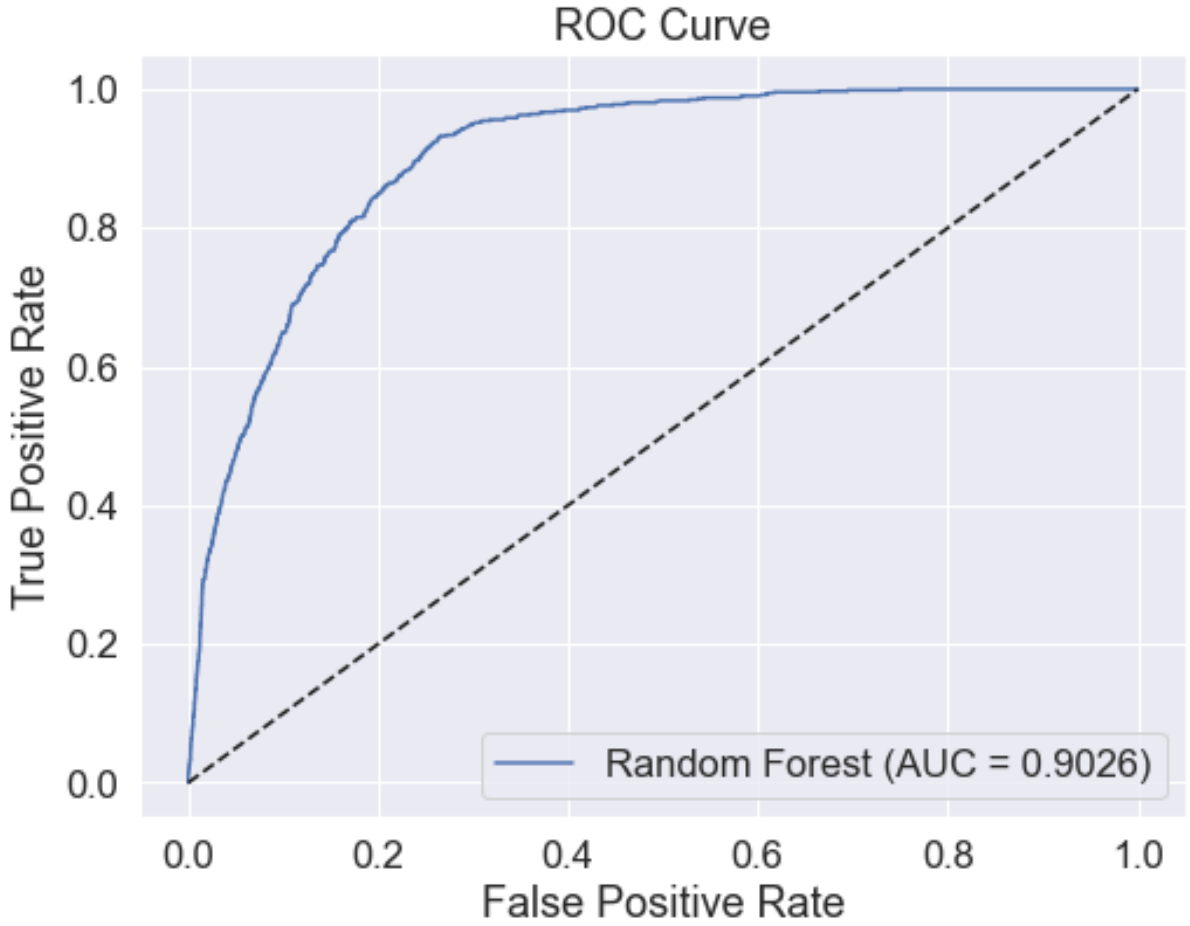

Figure 11: The Receiver Operating Characteristic curve (ROC) with corresponding Area Under the Curve (AUC)

The ROC curve, with the corresponding AUC score of 0.906 as exhibited in Figure 11, provides valuable insights into the discriminatory power of the random forest model. With an AUC score of 0.906, the random forest model demonstrates a strong ability to rank positive instances higher than negative instances on average. This suggests that the model is effective

in capturing the underlying patterns and features that differentiate between customers who subscribed and those who did not. The high AUC score signifies that the random forest model exhibits a high level of classification performance. The model accurately separates the two classes by assigning higher probabilities to instances that belong to the positive class.

5.3.5 Feature Importance of Random Forest

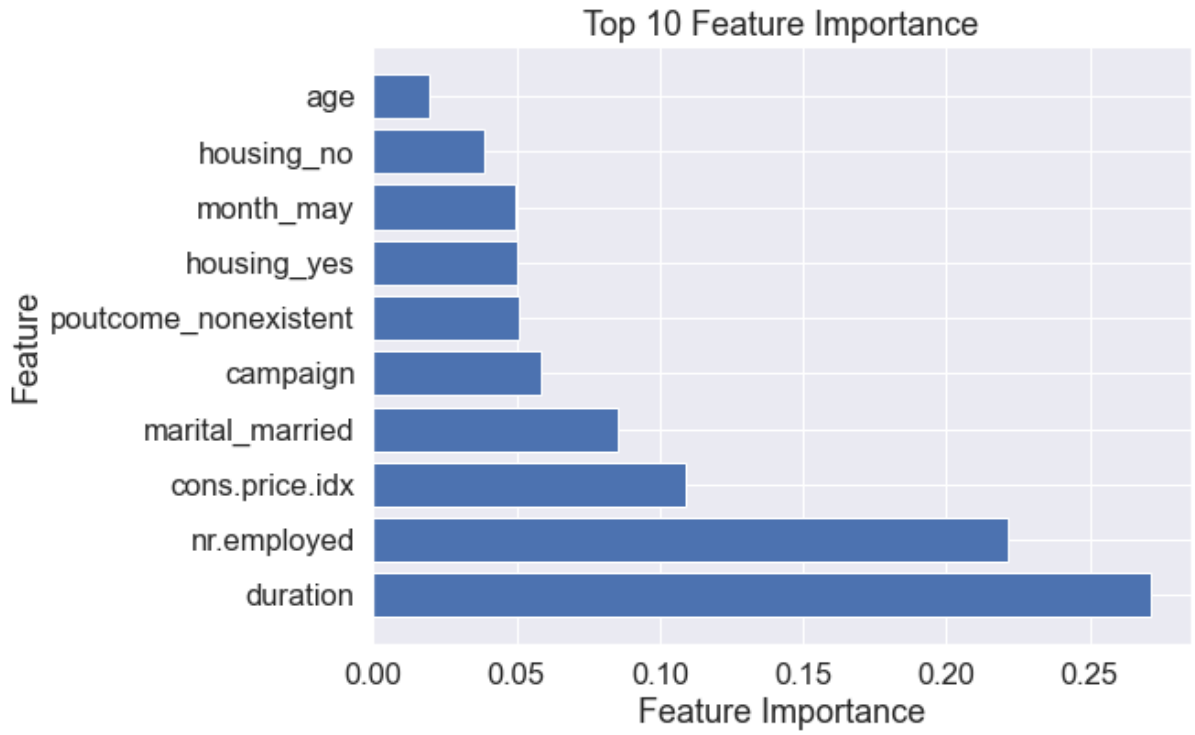

Figure 12: Top 10 important features for the Random Forest model

The feature importance bar chart, presented in Figure 12, provides insights into the significant variables utilized by the random forest model to make predictions. At the top of the list of important features is the variable 'duration,' which represents the duration of the last contact made with the customer during a marketing campaign. This variable holds the most significant impact on the model's predictions. The longer the duration of the last contact, the more influential it is in determining whether a customer will subscribe or not.

Following 'duration,' the variable 'number of employees' emerges as the second most important feature. This variable reflects a quarterly indicator that captures the number of employees within the banking institution. The model considers this information as a valuable factor in predicting customer subscription behavior.

The third most important feature identified by the model is the consumer price index, denoted as 'cons.price.idx.' The consumer price index is a monthly indicator tracks changes in the average prices of a basket of goods and services consumed by households. The model assigns a high level of importance to this variable, suggesting that fluctuations in consumer prices can significantly impact subscription decisions.

Another significant feature in the model is whether the customer is married or not. This variable plays a critical role in predicting customer behavior, with the model recognizing marital status as a factor that influences subscription decisions.

Lastly, the variable 'campaign' represents the number of contacts made during the current marketing campaign for each client. This variable ranks as the fifth most important feature in the model. The frequency of contact during the campaign holds relevance in the model's prediction, as it provides insights into the level of customer engagement and its impact on subscription outcomes.

5.4 XGBoost Model

5.4.1 Optimal Hyperparameters and Accuracy

The XGBoost model demonstrates exceptional performance with the following tuned hyperparameters: {'objective': 'binary:logistic', 'eval_metric': 'auc', 'max_depth': 6, 'learning_rate': 0.1, 'subsample': 0.9, 'colsample_bytree': 0.8, 'seed': 13}. These hyperparameters have been specifically selected to optimize the model's learning capabilities and enhance its predictive accuracy.

The ‘objective’ parameter sets the loss function to be minimized during training, with 'binary:logistic' indicating binary classification using logistic regression. The ‘eval_metric’ parameter employs the Area Under the Curve (AUC) metric, a widely used measure of the model's ability to distinguish between positive and negative instances.

The ‘max_depth’ hyperparameter controls the maximum depth of the decision trees in the ensemble, enabling the model to capture complex relationships. The ‘learning_rate’ hyperparameter determines the step size at each boosting iteration, influencing the contribution of each tree to the final prediction. The ‘subsample’ hyperparameter randomly selects a subset of samples for training each tree, mitigating overfitting and enhancing model generalization. The ‘colsample_bytree’ hyperparameter specifies the fraction of features used for training each tree, introducing diversity and robustness in feature selection. Lastly, the ‘seed’ hyperparameter ensures the reproducibility of results by initializing the random number generator.

With an high accuracy score of 0.9287, the XGBoost model showcases the model predicts classes well overall. This high accuracy indicates that the model correctly predicts the subscription behavior of approximately 92.87% of the instances.

5.4.2 Understanding Confusion Matrix of XGBoost

Table 13: Confusion Matrix of XGBoost

	Predicted Not Sub	Predicted Sub
Actual Not Sub	6777	197
Actual Sub	352	373

Note. This table shows the confusion matrix of the Logistic Regression. Here, the 'Predicted Not Sub' category refers to the number of instances where the model predicted that the customers would not subscribe to the term deposit, while the 'Predicted Sub' category refers to the number of instances where the model predicted that the customers would subscribe. The 'Actual Not Sub' category represents the actual number of customers who did not subscribe, and the 'Actual Sub' category represents the actual number of customers who did subscribe. The values in each cell indicate the count of instances that fall into each category, providing an overview of the model's performance in predicting customer subscriptions.

The confusion matrix, shown in table 13, reveals that the XGBoost model correctly predicted 6777 instances as 'Not Subscribed' (true negatives) and 373 instances as 'Subscribed' (true positives). These accurate predictions demonstrate the model's ability to correctly identify customers who are likely to subscribe or not subscribe to the service.

However, the model also had 197 false positives, meaning it incorrectly classified instances as 'Subscribed' when they were actually 'Not Subscribed.' Additionally, there were 352 false negatives, indicating instances that were mistakenly classified as 'Not Subscribed' when they were actually 'Subscribed.' These misclassifications represent instances where the model failed to capture the true subscription behavior of customers.

5.4.3 Dissecting Classification Report of XGBoost

Table 14: Classification report of XGBoost

	Precision	Recall	F1-Score	Support
Not Subscribed	0.95	0.97	0.96	6974
Subscribed	0.65	0.51	0.58	725
Accuracy	-	-	0.93	7699
Macro avg	0.8	0.74	0.77	7699
Weighted avg	0.92	0.93	0.92	7699

Note. This table presents the classification report of the model's performance. The 'precision' score measures the accuracy of the model's predictions for each class, where a higher value indicates a higher proportion of correct predictions. The 'recall' score represents the model's ability to correctly identify instances of each class. The 'f1-score' provides a balance between precision and recall, combining both measures into a single value. The 'support' column displays the number of instances in each class. The 'Accuracy' score indicates the overall accuracy of the model and can be interpreted as the percentage of correctly classified instances. The 'Macro avg' score provides an overall evaluation of the model's performance across both classes, considering their individual contributions.

It calculates the unweighted average scores for each class, providing an equal contribution to each class regardless of its size. The 'Weighted avg' score accounts for class imbalance and provides an overall evaluation, considering the support for each class.

The classification report, displayed in table 14, indicates that for the class 'Not Subscribed', the model achieved a precision of 0.95, indicating that when the model predicts a customer will not subscribe, it is correct 95% of the time. The recall score of 0.97 signifies the model's ability to correctly identify a large majority of actual 'Not Subscribed' instances. The high F1-score of 0.96, reflects a well established balance between the precision and recall metrics.

Furthermore, for the class 'Subscribed', the model achieved a precision of 0.65, indicating that among the instances predicted as 'Subscribed,' a considerable proportion were true positives. The recall score of 0.51 indicates that the model had more difficulty identifying actual 'Subscribed' instances, resulting in some false negatives. The F1-score of 0.58 reflects the trade-off between precision and recall for this class. These results suggest that the model's performance in identifying instances as 'Subscribed' is relatively lower compared to the 'Not Subscribed' class.

The macro average F1-score is 0.77, implying that the model balances precision and recall across both classes well.

The weighted average F1-score is 0.92, indicating that the model performance by accounting for the class imbalance.

5.4.4 ROC AUC Evaluation of XGBoost

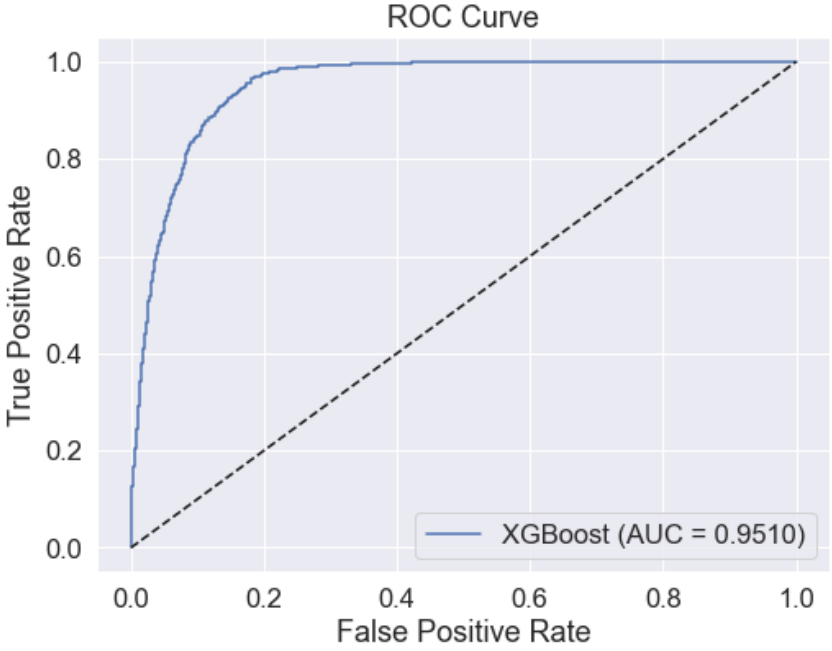

Figure 13: The Receiver Operating Characteristic curve (ROC) with corresponding Area Under the Curve (AUC)

The ROC curve, illustrated in Figure 13, reveals the exceptional performance of the XGBoost model, as it is accompanied by an AUC score of 0.9510. The high AUC score indicates that the XGBoost model exhibits a strong discriminatory power and is effective in correctly ranking the probabilities of positive and negative instances. The score highlights that the XGBoost model has a high true positive rate while keeping the false positive rate low.

5.4.5 Feature Importance of XGBoost

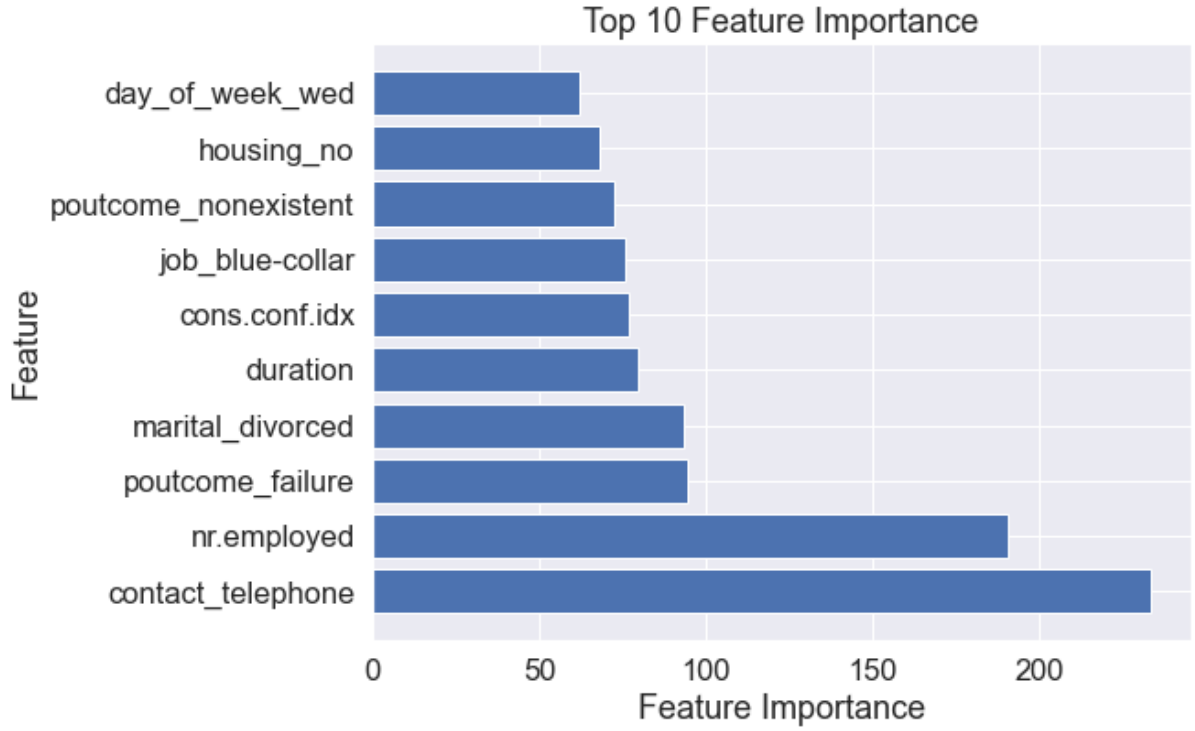

Figure 14: Top 10 important features for the XGBoost model

The feature importance bar chart, depicted in Figure 14, provides insights into the XGBoost model's top 10 most important features. These features play a crucial role in predicting the class to which a customer belongs and understanding the factors influencing customer behavior.

At the forefront of importance is the 'number of employees' variable, which holds the highest significance according to the XGBoost model. This variable represents the number of employees within the banking institution and serves as a strong indicator in predicting customer behavior.

The second most important feature is the 'month' variable, which captures the month of the year when the last contact was made with customers. The XGBoost model utilizes this feature, particularly emphasizing the month of May, to effectively differentiate customers between the classes. The significance of the month of May suggests a potential seasonality effect or a specific marketing campaign associated with this time period.

The third most important feature is the 'Employment Variation Rate,' a quarterly indicator reflecting changes in the employment market or labor market conditions. This

variable provides valuable insights into the economic context and its impact on customer behavior.

Next, the 'Contact' feature holds substantial importance, representing the communication type used for the last contact with the customer. This variable offers insights into the effectiveness of different communication channels and their influence on customer decision-making.

Lastly, the 'marital' feature, which captures the marital status of customers, emerges as a significant predictor. This variable sheds light on the role of marital status in influencing subscription behavior, potentially indicating that married individuals may exhibit different responses compared to unmarried individuals.

5.5 Summarizing performance of models

Table 15: Summary of Model Performances

	Logistic Regression	SGD	Random forest	XGBoost
Accuracy	0.92	0.90	0.84	0.93
Not Subscribed Precision	0.94	0.96	0.97	0.95
Subscribed Precision	0.64	0.47	0.34	0.65
Not Subscribed Recall	0.98	0.93	0.84	0.97
Subscribed Recall	0.41	0.63	0.78	0.51
Not Subscribed F1	0.96	0.94	0.90	0.96
Subscribed F1	0.50	0.54	0.47	0.58
Weighted avg F1	0.92	0.90	0.86	0.92
AUC	0.9363	0.8985	0.9026	0.9510

Note. This table displays the performance metrics of different machine learning models: Logistic Regression, SGD, Random Forest, and XGBoost. The rows represent various evaluation metrics, while the columns represent the respective models. The 'Accuracy' metric indicates the overall proportion of correct predictions made by each model. The 'Precision' measures the model's ability to correctly predict instances belonging to the 'Not Subscribed' and 'Subscribed' classes. The 'Recall' metric evaluates the model's ability to correctly identify instances of the 'Not Subscribed' and 'Subscribed' classes. The 'F1-score' is a combined measure of precision and recall. The 'Weighted avg F1' calculates the average F1-score, weighted by the number of instances in each class. The 'AUC' (Area Under the Curve) represents the performance of the model's Receiver Operating Characteristic (ROC) curve. It measures the model's ability to correctly rank instances of the positive and negative classes. A higher AUC score indicates better discrimination between the classes. All metrics have a scale between 0 and 1, where best performance takes on the value of 1.

The performance of the machine learning models utilized in this study is presented in table 15. The table reveals that both the random forest model and the SGD model yield comparable outcomes, with the random forest model exhibiting a slightly higher AUC score. Moreover, the logistic regression model surpasses both the random forest and SGD models, as evidenced by its superior accuracy, weighted average F1, and AUC scores. The XGBoost

model exhibits a slightly better performance than the logistic regression model, as indicated by the accuracy and AUC scores, establishing it as the top-performing model in this research.

5.5.1 Model performances and previous literature

While it is not feasible to directly compare the performance of our models with other empirical papers predicting customer churn due to differences in datasets, we can draw close comparisons by examining the results of Moro, Cortez, and Rita's 2014 paper. Utilizing an expanded version of the dataset employed in our research, they implemented a neural network model as their top-performing approach, followed by their logistic regression model.

In their study, the most crucial variable for their neural network model was the Euribor 3-month rate, followed by the duration of the call as the fifth most important feature, and the number of employees as the sixth most important variable. The top five variables in their neural network were mainly comprised of clientele and financial product data, which were not included in our dataset. Their neural network model achieved an impressive AUC score of 0.929.

We observe similarities between their findings and ours. Both studies highlight the significance of the number of employees and the duration of the call as crucial features for predicting customer churn. Furthermore, the comparable AUC scores, hovering around 90, demonstrate the efficacy of our models in capturing predictive patterns akin to those presented by Moro, Cortez, and Rita. Nonetheless, a more nuanced assessment would necessitate further alignment of methodological approaches and dataset parameters.
The primary distinction between our findings and those of Moro, Cortez, and Rita lies in the significance of the day of the week variable within our logistic regression model for predicting customer subscription to a term deposit. Surprisingly, the day of the week variable held the highest importance in our model, while it did not rank among the top 10 most important features in Moro, Cortez, and Rita's study. Despite this dissimilarity, both studies achieved comparable AUC scores, with Moro, Cortez, and Rita obtaining an AUC score of 0.900, aligning closely with our logistic regression model's AUC score.

6 Addressing the research questions

Having completed the analysis of the machine learning models, we will now use the results in combination with the findings of the EDA and economic theories to answer our research question.
We reiterate that the main purpose of our research is to find in what ways the integration of economic theories and machine learning methodologies contribute to the understanding and prediction of customer churn in the context of a Portuguese banking institution.

The investigation will proceed by addressing the hypotheses and utilizing their findings to tackle the main research question. Each hypothesis will be examined by analyzing the

respective variables. Initially, the significance tests of these variables, derived from the EDA, will be employed to determine whether there is a statistically significant difference between the group of subscribers and the group of non-subscribers for each variable.
Subsequently, the internal proportions of the variables will be evaluated to ascertain the direction of the difference between the groups. This analysis will provide insights into whether the proportion of subscribers is higher or lower for particular values of the variable.
Finally, the feature importance bar charts generated by the machine learning models will be utilized to assess the significance of the variables. This examination will indicate whether the variables hold substantial importance in the predictive models. The reliability of the feature importance chart is bolstered by the strong performance exhibited by the models, particularly the XGBoost model.

6.1 First hypothesis

The first hypothesis, constructed with the switching cost theory in mind, was:

Customers who previously had a term deposit at the bank exhibit lower churn rates in the current marketing campaign.

To answer this hypothesis, we first look at the statistical significance of the variable 'previous outcome'. Table 6 of the EDA demonstrates a high statistical significance (p-value of 0.0000**) for the variable, indicating notable differences among the various groups based on the outcome of the previous marketing campaign. This suggests that the previous outcome has a significant impact on subscription behavior in the current marketing campaign.

After the statistical significance has been confirmed, the sign regarding the influence of the variable will be discerned. Customers who had a previous outcome of 'success' in the previous marketing campaign, indicating a successful subscription to a term deposit, exhibit a higher proportion of subscribers in the current marketing campaign. This can be observed in the Figure 5 of the EDA, where the proportion for this group is notably larger compared to other groups. This substantial difference suggests a strong association between previous success and a higher likelihood of subscribing in the current campaign.

Lastly, the variable 'previous outcome' holds substantial importance across multiple machine learning models. It ranks as the sixth most important feature in the random forest model, and the third most important feature in the SGD model. Moreover, the previous outcome variable is the third most important feature in the XGBoost model, which is our best performing model. These rankings highlight the influential role of the previous outcome in predicting customer behavior and subscription outcomes across different models.

Considering these findings, it is reasonable to argue that customers who had previously subscribed to a term deposit exhibit lower churn rates in the current marketing campaign. The integration of the switching cost theory with the findings from the EDA and machine learning models, grants us the ability to statistically and theoretically uncover one of the reasons behind the customer churn.

The switching cost theory provides a valuable framework for understanding this phenomenon. It explains that customers who have already subscribed to a term deposit are more likely to exhibit lower churn rates due to the higher costs associated with switching their loan provider. The successful previous subscription indicates a positive relationship and satisfaction with the bank's offerings, which creates barriers to switching and fosters loyalty. These customers have already experienced the benefits and value associated with the term deposit, and as a result, they are more inclined to continue their subscription. They perceive the costs of switching to a different provider as inconvenient or risky, considering factors such as the need to open a new account, transfer funds, and potentially lose any benefits or preferential treatment they currently enjoy as existing customers.

6.2 Second hypothesis

The second hypothesis, motivated by the relationship marketing theory, was:

Customers who have a longer contact duration exhibit lower churn rates in the current marketing campaign.

We commence answering this hypothesis by addressing the significance value of the 'duration'. Table 4 of the EDA indicates that the variable demonstrates a high level of statistical significance (p-value of 0.000), implying a clear distinction among different lengths of contact durations concerning subscription behavior. This suggests that the duration of the last contact plays a significant role in determining customer outcomes.

Customers who subscribed to the service exhibit a higher value for the duration variable. This is evident from the box plot in Figure 3, and the table 2 which shows that the median and all quartiles for subscribers have a higher duration value compared to non-subscribers. These findings of the EDA imply that customers who have a longer duration of the last contact are more likely to subscribe to the service.

The importance of the 'duration' variable is evident when analysing the machine learning models used in this research. In the random forest model, as shown in Figure 12, 'duration' is identified as the most important feature. Additionally, in the XGBoost model, which is the best-performing model, 'duration' ranks as the fifth most important feature. This consistent recognition of the variable's significance underscores its influence in predicting customer behavior.

Based on these results and the theoretical lens of Relationship Marketing, it is plausible to assert that customers who have a longer contact duration exhibit lower churn rates in the current marketing campaign. The inclusion of the relationship marketing theory grants us the ability to reason further about the occurrence of customer churn.
The duration of the contact represents the extent of interaction between the customer and the bank, which can foster stronger relationships and deeper engagement. The tenets of the relationship marketing theory allow us to argue that customers who engage in longer duration contacts may have the opportunity to establish rapport, gain a better understanding of the bank's offerings, and receive personalized attention. This enhanced relationship quality can contribute to increased trust, loyalty, and satisfaction, thereby reducing the likelihood of churn.

6.3 Third hypothesis

The last hypothesis, influenced by the perceived value theory, was:

Customers who already had a financial product at the bank exhibit reduced churn rates in the current marketing campaign.

The first step towards answering the hypothesis is confirming the statistical significance of the 'housing loan' variable. The table 5 of the EDA shows that the housing loan variable is statistically significant, indicating a notable difference among the housing loan groups regarding subscribing or not. This suggests that the housing loan status has an influence on customer behavior.

Next, the proportion of subscribers amongst the groups of the 'housing loan' variable are compared. In Figure 4 of the EDA, we observe that the proportion of subscribers is higher for customers with a housing loan compared to those without. Although the difference in proportions may not be substantial, it still suggests a slight association between having a housing loan and a slightly higher likelihood of subscribing to the service. Even small differences in proportions can be statistically significant and have practical implications, especially with large datasets.
Thirdly, the housing loan variable consistently ranks among the top 10 most important features across all the machine learning models used in this research. This highlights its consistent impact and predictive power in determining customer behavior.

The other variable used to answer this hypothesis is 'previous outcome'. Since the summary of the results regarding this variable would be the same as the of the first hypothesis, we will abstain from repeating ourselves and focus answering the third hypothesis.

Considering the findings of both variables, it is reasonable to argue that customers who already had a financial product at the bank exhibit reduced churn rates in the current marketing campaign. Utilizing the perceived value theory, we further argue that customers churned less due to the perceived value and benefits associated with the additional banking product and service provided by the bank. Having a housing loan or term deposit establishes a pre-existing relationship between the customer and the bank, enhancing familiarity and potentially increasing the perceived value of the bank's offerings. Customers who have experienced positive interactions and favorable services related to their housing loan may develop a stronger perception of value, leading to a lower likelihood of churning.

6.4 Main Research question

The primary objective of this research was to examine the extent to which the integration of economic theories and machine learning methodologies can enhance the comprehension and predictive capabilities concerning customer churn within the context of a Portuguese banking institution.
To achieve this aim, we employed economic theories as a foundation for formulating hypotheses, which were subjected to empirical testing using a carefully curated dataset and advanced machine learning techniques.

Our analysis revealed that customers who previously had a term deposit at the bank exhibit lower churn rates in the current marketing campaign. This finding aligns with the Switching Costs Theory, suggesting that the higher switching costs associated with changing their loan provider contribute to their loyalty and reduced propensity to churn.
Furthermore, our study found that customers who have a longer contact duration with the bank also had lower churn rates. This supports the notion that increased engagement and stronger relationships contribute to a decreased likelihood of churn, as proposed by relationship marketing theories.
Additionally, we observed that customers who already had a financial product at the bank tend to exhibit reduced churn rates. This finding can be attributed to the Perceived Value Theory, as these customers have already experienced the benefits and value associated with the bank's offerings, creating switching barriers and fostering loyalty.

The successful validation of these hypotheses, in conjunction with the statistical significance of the variables and their relative importance in the machine learning models, demonstrates the power of integrating economic theories and machine learning methodologies. The economic theories allowed us to forge hypotheses and reason further past the statistical observations and machine learning insights, granting us the framework to further explain the phenomenon of customer churn within a Portuguese banking institution. The banking institutions could enhance their understanding of customer churn and develop targeted strategies to improve customer retention by adopting a similar strategy.

Overall, the integration of economic theories and machine learning methodologies provides a comprehensive framework for predicting and understanding customer churn. The insights gained from this integration enables the banking institutions to make informed decisions, implement tailored retention strategies, and foster stronger customer relationships, ultimately contributing to sustainable business growth and success.

7 Implications

The findings of our research have implications for both academic understanding and practical applications in the context of customer churn prediction within Portuguese banking institutions. The integration of economic theories and machine learning methodologies has provided valuable insights that can drive the decision-making processes and enhance customer relationship management strategies.

To commence, the practical applications of this research are noteworthy. The accurate prediction of customer churn by the models enables banking institutions to proactively identify and address potential churn risks, allowing for targeted retention efforts and personalized marketing interventions. By leveraging the insights gained from this study, banking institutions can improve their resource allocation, and improve long-term profitability.

Furthermore, the implications of our research extend to customer relationship management strategies. The integration of economic theories with machine learning methodologies enhances the understanding of customer behavior, preferences, and needs. This understanding allows banks to deploy personalized marketing strategies, improve customer segmentation, and deliver tailored services. By adopting a proactive approach to customer churn prediction, banking institutions can build stronger customer relationships, increase customer loyalty, and foster long-term customer satisfaction. Moreover, the integration of economic theories with machine learning methodologies allows banks to segment high churn-risk customers and take the appropriate measures to convert them to loyal ones.

Additionally, the insights from this research have implications for decision-making within banking institutions. The combination of economic theories and machine learning techniques provides a data-driven approach to strategic planning, product development, and pricing strategies. This allows banking institutions to make more informed decisions based on customer insights, market dynamics, and economic principles.

Overall, the implications of this research offer valuable insights for banking institutions seeking to improve customer retention, foster customer loyalty, and drive sustainable growth in an increasingly competitive market. By integrating economic theories with machine

learning methodologies, banking institutions can navigate the complexities of customer churn, optimize resource allocation, and build lasting relationships with their customers.

8 Limitations and future suggestions

Our research is subject to several limitations that open up possibilities for future research. To commence, the reliance on our dataset from a single Portuguese banking institution limits the generalizability of the results. Customer behavior and market dynamics can vary across different banks and financial institutions in Portugal, and therefore, the findings and conclusions drawn from this study may not be fully applicable to those other institutions.
Furthermore, quality of the dataset used for analysis is another limitation. The presence of missing data can introduce biases and potentially impact the robustness of the results.
Moreover, the sole focus on the Portuguese banking sector poses limitations on the transferability of the findings to other geographical regions or cultural contexts. Factors unique to the Portuguese markets, including regulatory frameworks, customer preferences, market dynamics, and demographics may influence the results and restrict their applicability to other banking sectors.
Future researchers can overcome these limitations by conducting similar studies with datasets from multiple banking institutions and different countries, which could enhance the generalizability and robustness of the findings and provide a broader understanding of customer churn dynamics.

This research acknowledges that it only examines a fraction of the economic theories and variables that contribute to the understanding of customer churn. Due to time constraints, the focus is primarily on specific variables related to customer churn, neglecting other influential factors such as socio-demographic characteristics, external market forces, and customer-specific attributes. Consequently, the predictive accuracy and comprehensive understanding of customer churn may be limited by this constrained scope of variables.
A more comprehensive exploration of economic theories and corresponding variables could have yielded additional insights into customer churn and enhanced the integration with machine learning methodologies. Future studies with fewer constraints could delve deeper into the exploration of economic theories and corresponding variables, thus further expanding the potential contributions of integrating economic theories with machine learning in the prediction of customer churn.

Moreover, methodological constraints are an important aspect to address in this research. The selection of machine learning algorithms and techniques employed in this study carry their own limitations. With the developing machine learning landscape, researchers can utilize more sophisticated models such as neural networks, ensemble methods, or deep learning

architectures. These advanced models have the potential to enhance the predictive performance and provide deeper insights into customer churn.

Furthermore, the techniques used during the preprocessing stage of the dataset, such as SMOTE and SimpleImputer, while effective, can be replaced with more sophisticated approaches. For example, researchers can explore the use of advanced imputation methods like MissForest or gain insights from newer oversampling techniques like ADASYN (Adaptive Synthetic Sampling). These advancements in preprocessing techniques offer future researchers the opportunity to improve the performance and gain more comprehensive insights into customer churn prediction.

9 Conclusion

Our study aimed to investigate in what ways the integration of economic theories and machine learning methodologies contribute to the understanding and prediction of customer churn within a Portuguese banking institution. We utilized an UCI dataset containing Portuguese banking data for statistical analysis and machine learning.

To address our main research question, we formulated three hypotheses inspired by economic theories. These hypotheses served as guiding principles in our analysis, allowing us to delve deeper into the complex dynamics of customer behavior. Through the careful selection of appropriate variables from the dataset, we tested these hypotheses and examined their implications on customer churn.

Our analysis revealed noteworthy insights from both the exploratory data analysis and machine learning models. The exploratory data analysis provided valuable statistical observations, shedding light on the relationships between variables and the influence they exerted on customer churn. The machine learning models, on the other hand, contributed predictive power and demonstrated the importance of various features in accurately predicting churn.

By integrating economic theories with machine learning methodologies, we achieved a more comprehensive understanding of customer churn. The economic theories provided a conceptual framework that went beyond mere statistical observations and machine learning insights. The theories enabled us to reason further, offering explanatory power and offering insights into the underlying mechanisms driving customer churn within the Portuguese banking context.

Furthermore, the integration of economic theories with machine learning enhances nuanced decision-making of banks, empowering them to proactively mitigate churn threats and thereby ensuring targeted retention and strategic resource allocation in an escalating competitive environment. To the boards of companies, a data-driven strategy informed by

both granular customer insights and foundational economic principles becomes a lighthouse for informed strategic planning, adept product development, and calibrated pricing strategies. From a marketing perspective, this interplay forges a deep-rooted understanding of customer behaviors, facilitating the design of tailored marketing campaigns, refined segmentation processes, and personalized service offerings. In essence, the fusion of economic frameworks with machine learning provides banks with an advanced schema to effectively navigate the intricate nuances of customer churn, cultivating lasting client relationships.

However, it is important to acknowledge the limitations of our research. Our findings are based on a specific dataset from a single Portuguese banking institution, which limits the generalizability of our conclusions. To enhance the external validity and robustness of our findings, future research should incorporate data from multiple banks across different countries. Moreover, future researchers ought to incorporate more economic theories that explain the phenomenon of customer churn, since our research was constrained to three theories. Furthermore, the integration of advanced machine learning models in forthcoming research holds the potential to yield heightened accuracy and explanatory power, increasing the comprehension of customer churn. The broader scope created by the data from multiple banks, the inclusion of additional economic theories, and more sophisticated machine learning models would allow for a more comprehensive examination of customer churn dynamics in diverse contexts and add to the generalizability of the findings.

In summary, our study demonstrated the potential of integrating economic theories and machine learning methodologies in understanding and predicting customer churn within a Portuguese banking institution. The combination of these approaches offered a more sophisticated understanding of the phenomenon, going beyond mere statistical associations and machine learning results, but uncovering the underlying mechanisms. As we move forward, it is crucial to continue exploring these avenues, expanding the scope of research to maximize generalizability and further refine our understanding of customer churn in the banking industry.

10 References

10.1 Introduction

Kiseleva, E. M., Nekrasova, M. L., Mayorova, M. A., Rudenko, M. N., & Kankhva, V. S. (2016). The theory and practice of customer loyalty management and customer focus in the enterprise activity. International Review of Management and Marketing, 6(6), 95-103.

Saran Kumar, A., & Chandrakala, D. (2016). A survey on customer churn prediction using machine learning techniques. International Journal of Computer Applications, 975, 8887.

Dam, S. M., & Dam, T. C. (2021). Relationships between service quality, brand image, customer satisfaction, and customer loyalty. The Journal of Asian Finance, Economics and Business, 8(3), 585-593.

Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In 2020 4th international conference on electronics, communication and aerospace technology (ICECA) (pp. 1196-1201). IEEE.

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276-286.

Reichheld, F. F., & Sasser Jr, W. E. (1990). Zero defections: Quality comes to services. Harvard business review, 68(5), 105-111.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. European journal of operational research, 157(1), 196-217.

10.2 Literature Review

10.2.1 Machine Learning

10.2.1.1 Supervised Machine learning

Hearty, J. (2016). Advanced machine learning with Python. Packt Publishing Ltd.

Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. ' O'Reilly Media, Inc.'.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?' Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

10.2.1.2 Applications of Supervised Machine learning for marketing purposes

Palmatier, R. W., & Sridhar, S. (2020). Marketing strategy: Based on first principles and data analytics. Bloomsbury Publishing.

Ferrell, O. C., Hartline, M., & Hochstein, B. W. (2021). Marketing strategy. Cengage Learning.

Valecha, H., Varma, A., Khare, I., Sachdeva, A., & Goyal, M. (2018, November). Prediction of consumer behaviour using random forest algorithm. In 2018 5th IEEE Uttar Pradesh section international conference on electrical, electronics and computer engineering (UPCON) (pp. 1-6). IEEE.

Chen, P. P., Guitart, A., del Río, A. F., & Periánez, A. (2018, December). Customer lifetime value in video games using deep learning and parametric models. In 2018 IEEE international conference on big data (big data) (pp. 2134-2140). IEEE.

Nilashi, M., Ahani, A., Esfahani, M. D., Yadegaridehkordi, E., Samad, S., Ibrahim, O., ... & Akbari, E. (2019). Preference learning for eco-friendly hotels recommendation: A multi-criteria collaborative filtering approach. Journal of Cleaner Production, 215, 767-783.

Spedicato, G. A., Dutang, C., & Petrini, L. (2018). Machine learning methods to perform pricing optimization. A comparison with standard GLMs. Variance, 12(1), 69-89.

10.2.1.3 Empirical Studies on Machine Learning for Customer Churn Prediction

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 1-24.

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), 1-24.

Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, 19(1/2), 65-93.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

10.2.1.4 Churn prediction in the banking sector

Bilal Zorić, A. (2016). Predicting customer churn in banking industry using neural networks. Interdisciplinary Description of Complex Systems: INDECS, 14(2), 116-124.

Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. Knowledge-Based Systems, 212, 106586.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innovation, 2, 1-13.

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. International Journal of advanced computer Science and applications, 9(2).

Kaur, I., & Kaur, J. (2020, November). Customer churn analysis and prediction in banking industry using machine learning. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 434-437). IEEE.

10.2.2 Economic Theories and customer Churn

Jandaghi, G., Amini, A., Pirani, P., Amini, Z., & Kharazi, H. (2011). Survey the role of brand in formation of customer loyalty in financial services marketing by the approach of small firms (Case study of Iran Melli bank). Far East Journal of Psychology and Business, 3(3), 50-61.

Payne, A., & Frow, P. (2017). Relationship marketing: looking backwards towards the future. Journal of services marketing, 31(1), 11-15.

Shy, O. (2002). A quick-and-easy method for estimating switching costs. International journal of industrial organization, 20(1), 71-87.

Reichheld, F. (2001). Prescription for cutting costs. Harvard Business School Publishing.

Reichheld, F., & Schefter, P. (2000). E-loyalty: Your secret weapon on the web. Harvard Business Review, 78(4), 105-113.

Sánchez-Fernández, R., & Iniesta-Bonillo, M. Á. (2007). The concept of perceived value: a systematic review of the research. Marketing theory, 7(4), 427-451.
Boksberger, P. E., & Melsen, L. (2011). Perceived value: a critical examination of definitions, concepts and measures for the service industry. Journal of services marketing, 25(3), 229-240.

Aulia, S. A., Sukati, I., & Sulaiman, Z. (2016). A review: Customer perceived value and its Dimension. Asian Journal of Social Sciences and Management Studies, 3(2), 150-162.
Nguyen, D. T., Pham, V. T., Tran, D. M., & Pham, D. B. T. (2020). Impact of service quality, customer satisfaction and switching costs on customer loyalty. The Journal of Asian Finance, Economics and Business, 7(8), 395-405.

Lazirkha, D. P., Hom, J., & Melinda, V. (2022). Quality Analysis Of Digital Business Services In Improving Customer Satisfaction. Startupreneur Business Digital (SABDA Journal), 1(2), 156-166.

10.3 Research Methodology

10.3.1 Exploratory Data Analysis

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, School of Information and Computer Science.

Lichman, M. (2013). UCI Machine Learning Repository. University of California, School of Information and Computer Science.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

10.3.2 Preprocessing

Śniegula, A., Poniszewska-Marańda, A., & Popović, M. (2019, December). Study of machine learning methods for customer churn prediction in telecommunication company. In Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (pp. 640-644).

Panjasuchat, M., & Limpiyakorn, Y. (2020, August). Applying Reinforcement Learning for Customer Churn Prediction. In Journal of Physics: Conference Series (Vol. 1619, No. 1, p. 012016). IOP Publishing.

Amuda, K. A., & Adeyemo, A. B. (2019). Customers churn prediction in financial institution using artificial neural network. arXiv preprint arXiv:1912.11346.

López, M. B. V., García, M. Y. A., Jaico, J. L. B., Ruiz-Pico, Á. A., & Hernández, R. M. (2023). Application of a Data Mining Model to Predict Customer Defection. Case of a Telecommunications Company in Peru. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 14(1), 144-158.

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters, 11(1), 24.

Mishra, K., & Rani, R. (2017, August). Churn prediction in telecommunication using machine learning. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 2252-2257). IEEE.

Do, D., Huynh, P., Vo, P., & Vu, T. (2017, December). Customer churn prediction in an internet service provider. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 3928-3933). IEEE.

Wu, C., & Wang, L. (2022). A Comparative Analysis of Churn Prediction Models: A Case Study in Bank Credit Card. Journal of Supply Chain and Operations Management, 20(2), 120.

10.3.3 Machine learning models

Labhsetwar, S. R. (2020). Predictive analysis of customer churn in telecom industry using supervised learning. ICTACT Journal on Soft Computing, 10(2), 2054-2060.

11 Appendix

11.1 Exploratory Data Analysis

11.1.1 Numeric variables

11.1.1.1 Age

Age is a commonly utilized variable in predictive models, including those for customer churn prediction. It represents the chronological age of the customers in the dataset and is measured in years. Incorporating demographic factors like age has been a common practice in modeling customer churn in various domains. For instance, Keramati, Ghaneei, and Mirmohammadi (2016) incorporated demographic variables, including age, in their predictive models for customer churn in the context of an electronic bank. Similarly, Kaur and Kaur (2020) also leveraged demographic variables, such as age, in their prediction models. By considering age as a predictor, we can explore its impact on customer churn in the banking sector and uncover potential age-related patterns, preferences, and behaviors that influence customer retention. The inclusion of age as a variable in our analysis allows us to assess its significance and contribution to predicting customer churn in our specific research context.

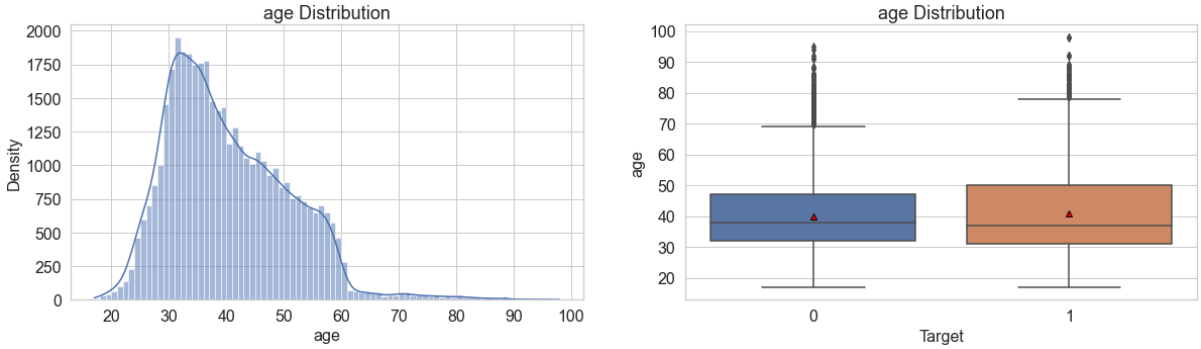

Figure 15: Histogram with KDE-line and boxplots for the variable Age

In Figure 15, the left plot presents a histogram with a superimposed KDE line, depicting the distribution of the numeric variable 'age'. The histogram indicates that the majority of customers fall within the age range of 30 to 40, with a slight rightward tail suggesting the presence of a few outliers with higher ages. Notably, the histogram illustrates a narrowing right tail beyond the age of 60, suggesting a smaller population above this age. The KDE line exhibits a similar pattern, with a peak in the densest region of the histogram and a gradually declining density towards higher ages.

On the right plot, boxplots are displayed for the age variable, separated based on whether the customer subscribed to a new term deposit or not. Both groups exhibit outliers, which are represented by individual data points beyond the whiskers of the boxplots. It is interesting to observe that the quartile group of customers who subscribed to a term deposit appears to have slightly older participants compared to the non-subscribers. Furthermore, the third quartile and the whisker indicating the maximum value are higher for the group that subscribed to the term deposit.

Table 16: Quartiles of the variable Age

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	17	32	38	47	95
1	17	31	37	50	98

Note. This table shows the quartiles and whiskers of the 'Age' variable, segmented into groups with a value 1 and 0 for the target variable.

Firstly, comparing the first quartile values (Q1) of table 16, we can observe that customers who subscribed to a term deposit (y = 1) tend to have slightly lower ages, with a Q1 of 31.0 compared to 32.0 for customers who did not subscribe (y = 0). This suggests that younger customers may be more inclined to subscribe to a term deposit.
Secondly, examining the third quartile values (Q3), we find that the age of customers who subscribed to a term deposit (y = 1) has a larger spread, with a Q3 of 50.0, compared to 47.0 for customers who did not subscribe (y = 0). This indicates a higher presence of older customers in the group that subscribed to a term deposit.
Additionally, analyzing the maximum age values, we find that both groups have outliers with advanced ages. However, the maximum age for customers who subscribed to a term deposit (y = 1) is 98.0, which is higher than the maximum age of 95.0 for customers who did not subscribe (y = 0). This suggests that a few customers with exceptionally high ages opted for a term deposit.

The presence of outliers in the right tail of the age variable can potentially impact the performance of machine learning models. Given the low density of these outliers, it may be reasonable to consider removing them to improve model performance.

11.1.1.2 Campaign

The variable 'campaign' represents the number of contacts made during the current marketing campaign for each client. It reflects the level of engagement and interaction between the bank and its customers during the campaign period. By considering the number of campaign contacts as a predictor, we can explore its relationship with customer churn. Higher values of the campaign variable may indicate multiple attempts to communicate with customers, potentially involving promotional offers, discounts, or personalized messages aimed at influencing their decision-making. Analyzing the campaign variable's relationship with customer churn allows us to examine the impact of contact frequency on customer behavior and their likelihood to churn. It provides insights into the effectiveness of the bank's marketing efforts and the role of customer engagement in influencing churn outcomes.

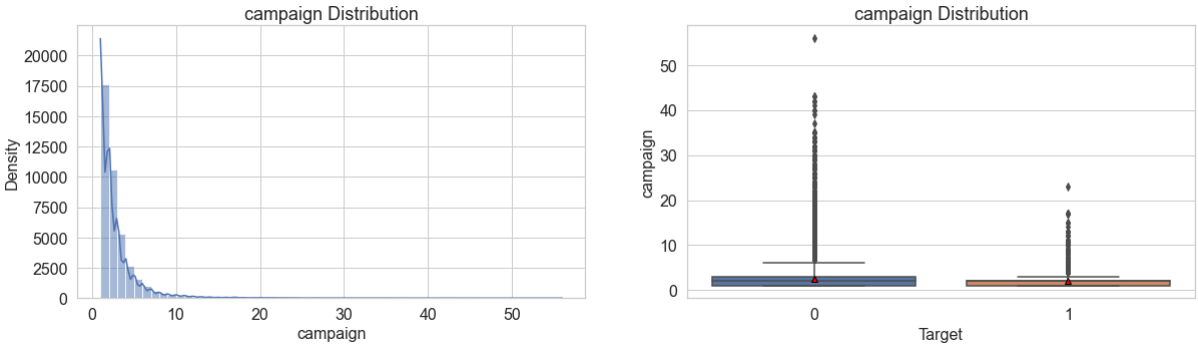

Figure 16: Histogram with KDE-line and boxplots for the variable Campaign

In Figure 16, we examine the distribution of the 'campaign' variable using a histogram in the left plot. The histogram reveals a common range of values concentrated between 1 and 10, suggesting that most customers were contacted a relatively low number of times during the campaign. However, there is a long right tail in the distribution, indicating the presence of outliers, where a small number of customers were contacted significantly more times.

On the right plot, we further explore the 'campaign' variable by dividing customers into two groups based on their subscription to the term deposit. Interestingly, we observe that the group of customers who did not subscribe to the term deposit exhibits larger outliers in the number of campaign contacts compared to the group that did subscribe. This finding is surprising because one might expect that customers who were contacted more frequently, possibly presented with discounts and promotions, would be more likely to subscribe. The presence of larger outliers in the non-subscription group suggests that there may be other factors influencing their decision or that additional targeted strategies may be needed for this group.

Table 17: Quartiles of the variable Campaign

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	1	1	2	3	56
1	1	1	2	2	23

Note. This table shows the quartiles and whiskers of the 'Campaign' variable, segmented into groups with a value 1 and 0 for the target variable.

The findings of table 17 highlight some key differences in the distribution of campaign contacts between customers who did and did not subscribe to the term deposit. It appears that the group of customers who did not subscribe to the term deposit received a wider range of campaign contacts, with some outliers receiving a significantly higher number of contacts. On the other hand, customers who subscribed to the term deposit generally had fewer campaign contacts, with a smaller variation in the number of contacts.

These insights can inform campaign management strategies, suggesting the need for targeted and personalized approaches to engage customers effectively. For customers who did not subscribe, it may be important to strike a balance between maintaining regular contact and

avoiding excessive communication. For customers who subscribed, understanding their preference for fewer campaign contacts can help optimize resources and focus on more personalized interactions to drive conversion rates.

11.1.1.3 Consumer Confidence Index

The variable 'Consumer Confidence Index', in the dataset referred to as 'cons.conf.idx', represents a monthly indicator that measures consumers' overall confidence and sentiment regarding the current and future state of the economy. It is derived from surveys conducted among a representative sample of consumers, capturing their perceptions of economic conditions, employment prospects, income expectations, and spending intentions. The Consumer Confidence Index serves as a crucial macroeconomic indicator, reflecting the level of consumer optimism and sentiment, which in turn influences their purchasing decisions and economic behavior. In the context of our study, we include the Consumer Confidence Index as a predictor to explore its relationship with customer churn. By examining the impact of consumer confidence on churn behavior, we can gain insights into how changes in economic sentiment may influence customers' propensity to switch or discontinue their banking relationship.

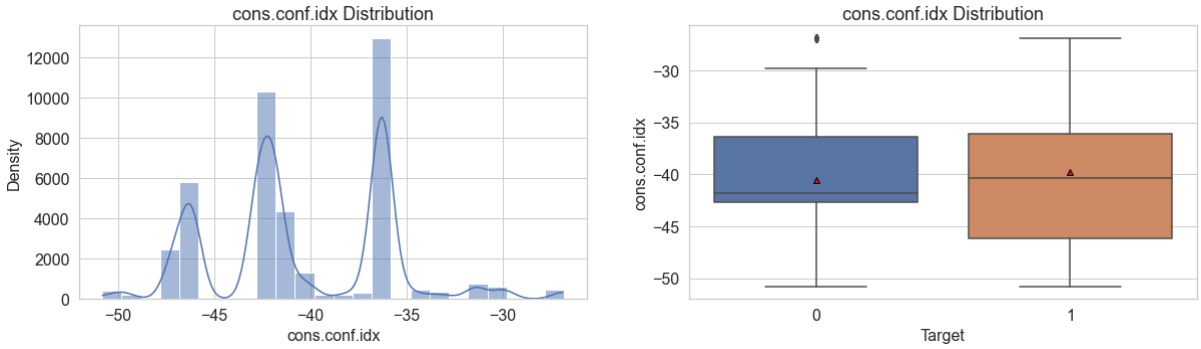

Figure 17: Histogram with KDE-line and boxplots for the variable Consumer Confidence Index

In Figure 17, the left plot reveals a distribution of the variable with the highest density observed between -40 and -35. The KDE line exhibits distinctive wave-like patterns and a rightward tail, indicating the presence of outliers.

Turning our attention to the right plot, we observe boxplots that compare the consumer confidence index between two groups: those who subscribed and those who did not. Interestingly, the first quartile of the group who subscribed has a more negative value compared to the other group, suggesting a relatively lower level of consumer confidence among those who ultimately subscribed. Additionally, the group who subscribed displays a slightly higher median value compared to the non-subscribers. It's worth noting that outliers are only present in the group of non-subscribers.

Table 18: Quartiles of the variable Consumer Confidence Index

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	-50.8	-42.7	-41.8	-36.4	-26.9
1	-50.8	-46.2	-40.4	-36.1	-26.9

Note. This table shows the quartiles and whiskers of the 'Consumer Confidence Index' variable, segmented into groups with a value 1 and 0 for the target variable.

Table 18 provides further support to the findings from the boxplot, as it confirms the higher values for the first and third quartiles of the group who subscribed to the term deposit.
The negative values of the consumer confidence index reflect a prevailing pessimism among customers regarding economic conditions during the campaign. These values suggest a cautious outlook and may influence decision-making and subscription willingness. However, the limited differentiation among groups based on the consumer confidence index indicates its potential lack of significant explanatory power in predicting subscription behavior. Further analysis and consideration of additional factors are necessary to understand the drivers behind subscription behavior.

11.1.1.4 Consumer price Index

The variable 'Consumer Price Index', in the dataset abbreviated to 'cons.price.idx', represents a monthly indicator that measures changes in the average prices of a basket of goods and services consumed by households. It serves as a key measure of inflation and reflects the purchasing power of consumers. The Consumer Price Index is calculated by comparing the current prices of a predefined set of goods and services to their prices in a base period. This index provides insights into the overall price level and inflationary trends, which impact consumers' cost of living and their purchasing decisions. In our study, we include the Consumer Price Index as a predictor to examine its potential influence on customer churn. By analyzing the relationship between the Consumer Price Index and churn behavior, we can assess whether changes in inflationary pressures and price levels have an effect on customers' decision to continue or discontinue their banking services

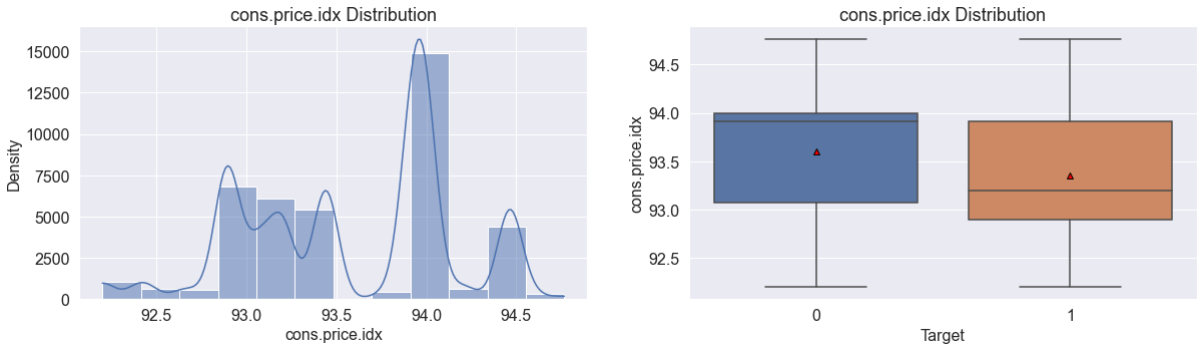

Figure 18: Histogram with KDE-line and boxplots for the variable Consumer Price Index

The left plot of Figure 18 reveals a distribution with multiple peaks, where the most prominent peak occurs at a value of 94 for the consumer price index. The presence of these peaks suggests distinct patterns or clusters within the distribution.
Upon examining the boxplots in the right plot, there is a noticeable difference in the median values between the two groups. The median of customers who subscribed to a term deposit appears to be lower compared to the other group.

Table 19: Quartiles of the variable Consumer Price Index

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	92.201	93.075	93.918	93.994	94.767
1	92.201	92.893	93.2	93.918	94.767

Note. This table shows the quartiles and whiskers of the 'Consumer Price Index' variable, segmented into groups with a value 1 and 0 for the target variable.

Referring to the quartile values presented in Table 19, we can confirm the observation of a lower median for the group of subscribers. Additionally, the first quartile of this group also demonstrates a lower value compared to the other group.
Considering the implications of these findings, the lower median and first quartile values for subscribers may suggest a different pricing perception or sensitivity among this group. This could indicate that changes in the consumer price index have a stronger impact on the decision-making process and subscription behavior of these customers. Further analysis and examination of additional factors are necessary to gain a deeper understanding of the relationship between the consumer price index and subscription outcomes.

11.1.1.5 Employment variation rate

The variable 'Employment Variation Rate', in the dataset identified as 'emp.var.rate', represents a quarterly indicator that reflects changes in the employment market or labor market conditions. It captures the fluctuations and shifts in the overall employment level, indicating the degree of volatility and instability in the job market during a specific time period. The employment variation rate serves as an economic indicator that can have implications for customer behavior and churn rates. Changes in employment conditions, such as increasing job opportunities or declining job stability, may influence customers' financial situations, confidence levels, and propensity to switch banks or discontinue their banking relationships. In our study, we include the Employment Variation Rate as a predictor to examine its potential impact on customer churn. By analyzing the relationship between the Employment Variation Rate and churn behavior, we aim to determine whether customers' likelihood of churn is influenced by fluctuations in the employment market and economic conditions.

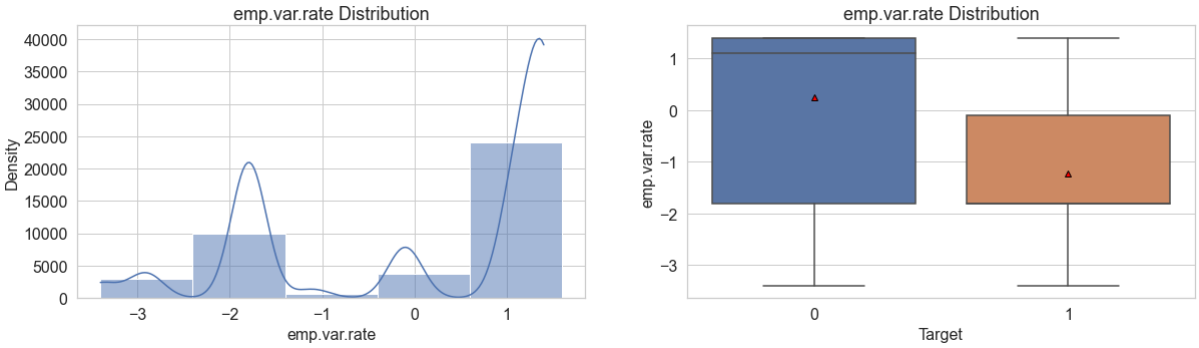

Figure 19: Histogram with KDE-line and boxplots for the variable Employment Variation Rate

Upon examining the left plot of Figure 19, we observe a histogram that displays two prominent peaks. The highest peak is located at a value of 1, while the second highest peak appears at -2. The density between these two peaks appears relatively low, indicating a potential bimodal distribution or the presence of distinct subgroups within the data.
Moving to the boxplots in the right plot, we notice some interesting observations. Firstly, the median and third quartiles for the group that did not subscribe to the term deposit are higher compared to the group that did subscribe. The median seems to have a value close to one, which is in line with the histograms having a peak at that value. Additionally, we observe that the mean value for the group that did not subscribe is also higher.

Table 20: Quartiles of the variable Employment Variation Rate

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	-3.4	-1.8	1.1	1.4	1.4
1	-3.4	-1.8	-1.8	-0.1	1.4

Note. This table shows the quartiles and whiskers of the 'Employment Variation Rate' variable, segmented into groups with a value 1 and 0 for the target variable.

Table 20 provides further insights into the quartile values for the variable. We encounter an unusual occurrence where the first quartile and the median have the same value for the group that subscribed. This suggests that a significant proportion of data points within this group are concentrated around a specific value. In other words, there is a lack of variability within the data, resulting in the overlap of the first quartile and the median.

This finding raises questions about the nature of the data distribution within the group that subscribed. It could indicate a distinct pattern or a specific characteristic of this subgroup. Further exploration and analysis are necessary to understand the underlying reasons for this occurrence and assess its potential impact on the analysis.

11.1.1.6 Euribor 3 month rate

The variable 'Euribor 3 Month Rate', in the dataset labeled 'euribor3m', refers to the daily indicator of the interest rates at which European banks lend to one another. It represents the average interest rate for a three-month period. The Euribor 3 Month Rate is a key benchmark

used in the financial industry and is closely linked to the cost of borrowing for banks and individuals. In our study, we include the Euribor 3 Month Rate as a predictor to examine its potential influence on customer churn behavior. By analyzing the relationship between the Euribor 3 Month Rate and churn rates, we aim to understand whether changes in interest rates impact customers' banking decisions and their likelihood of churn.

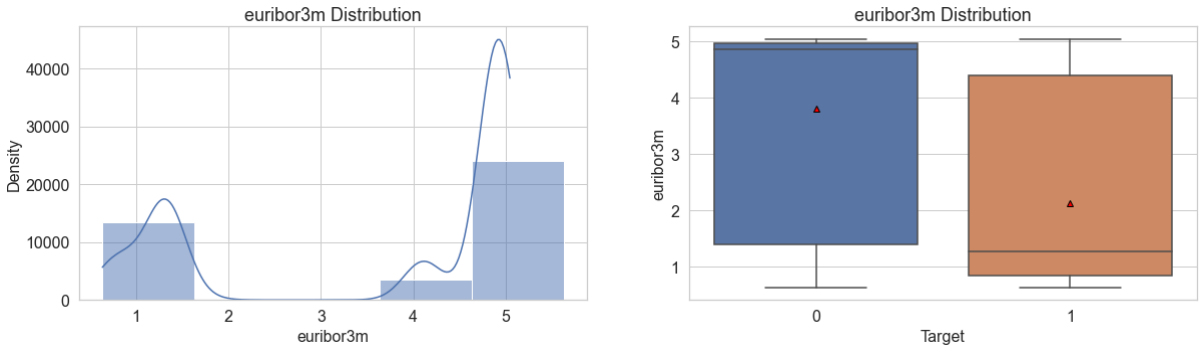

Figure 20: Histogram with KDE-line and boxplots for the variable Euribor 3 Month Rate

When investigating the left plot of Figure 20, we observe that the highest density is concentrated around the value of 5, indicating a peak in the distribution. Additionally, there is a noticeable decline in density between the values 1 and 4, suggesting a significant variation in the data within this range.
Shifting our attention to the boxplots in the right plot, we observe a distinct contrast between the medians of the two groups. The median value for the group that subscribed to the term deposit is notably lower compared to the group that did not subscribe. Furthermore, we notice that the mean value for the group that subscribed, as denoted by the red triangle, is also lower.

Table 21: Quartiles of the variable Euribor 3 Month Rate

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	0.634	1.405	4.857	4.962	5.045
1	0.634	0.849	1.266	4.406	5.045

Note. This table shows the quartiles and whiskers of the 'Euribor 3 Month Rate' variable, segmented into groups with a value 1 and 0 for the target variable.

Table 21 provides a comprehensive overview of the quartile values for the variable, shedding further light on the differences between the two groups. Notably, all quartiles for the group that subscribed to a term deposit are lower compared to the group that did not subscribe. The most substantial discrepancy is observed between the medians of the two groups, indicating a significant divergence in the central tendency of the variable's distribution.

The prominent disparity in values between the groups suggests a potentially influential role of the euribor 3 month rate in predicting subscription behavior. The lower values for the group that subscribed could indicate favorable market conditions, potentially leading to a higher

propensity for customers to subscribe to the term deposit. Conversely, the higher values for the group that did not subscribe may reflect less favorable market conditions, influencing customers to refrain from subscribing.

11.1.1.7 Number of Employees

The variable 'Number of Employees', known as 'nr.employed' in the dataset, represents a quarterly indicator that reflects the number of employees within the banking institution. It serves as a measure of the bank's workforce and can provide insights into the bank's operational capacity and resources. The number of employees can indirectly influence customer churn by impacting the bank's service quality, responsiveness, and ability to meet customer needs. In our study, we include the Number of Employees as a predictor to explore its potential association with customer churn. By analyzing the relationship between the Number of Employees and churn behavior, we aim to investigate whether a larger or smaller workforce has an impact on customer retention and loyalty.

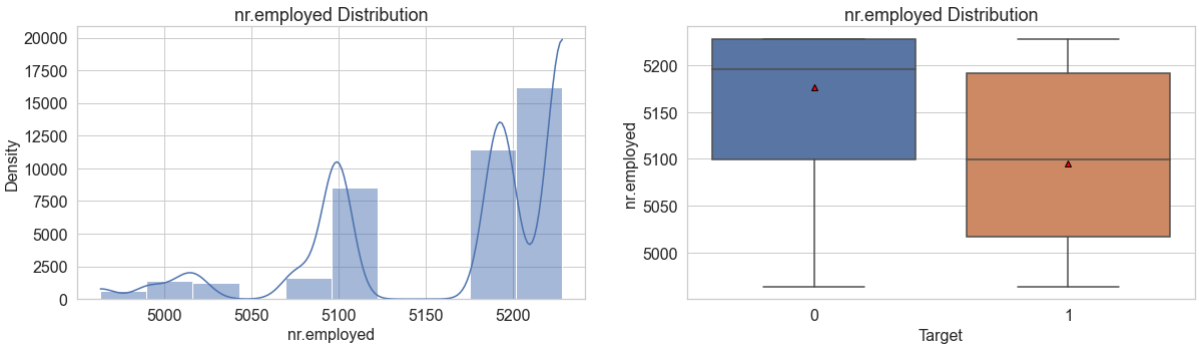

Figure 21: Histogram with KDE-line and boxplots for the variable Number Of Employees

Examining the left plot in Figure 21, we can observe that the highest density is centered around the value of 5200 for the number of employees variable. The KDE line provides further insights, revealing the presence of two distinct peaks around this value, with the majority of data points distributed above 5200. Additionally, another peak is noticeable around the value of 5100, indicating a secondary concentration of data points.

Shifting our focus to the right plot, we can observe differences in the median and mean values between the two groups. The median and mean values for the group that did not subscribe to the term deposit are higher compared to the group that subscribed. Furthermore, we notice that the first quartile of the group that subscribed appears to be lower than that of the group that did not subscribe.

Table 22: Quartiles of the variable Number Of Employees

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	4963.6	5099.1	5195.8	5228.1	5228.1
1	4963.6	5017.5	5099.1	5191	5228.1

Note. This table shows the quartiles and whiskers of the 'Number Of Employees' variable, segmented into groups with a value 1 and 0 for the target variable.

Table 22 provides a detailed overview of the quartile values, allowing us to delve deeper into the differences between the two groups. Notably, the first quartile, median, and second quartile values are all lower for the group that subscribed to the term deposit, indicating a lower range of values compared to the group that did not subscribe.

This finding suggests that the number of employees within the Portuguese bank may play a significant role in predicting customer subscription behavior. The quartile analysis reveals interesting patterns regarding the relationship between the number of employees and customer decisions.
The lower quartile values observed for the group of customers who subscribed to the term deposit may indicate more favorable employment conditions within the bank. A smaller number of employees in this context could imply efficient operations, streamlined processes, and potentially better service quality. Such conditions might influence customers positively, leading them to opt for the term deposit.
Conversely, the higher quartile values observed for the group of customers who did not subscribe to the term deposit may suggest less favorable employment conditions within the bank. A larger number of employees might indicate a relatively larger workforce, potentially associated with administrative complexities, slower decision-making processes, or limited resources. These factors could make customers more hesitant in subscribing to the term deposit.

11.1.1.8 Days passed

The variable 'Days Passed', shortened to 'pdays' in the dataset, represents the number of days that have elapsed since the customer was last contacted during a previous marketing campaign. It captures the length of time between the customer's previous interaction with the bank and the current campaign. The variable takes on a value of 999 if the client was not contacted previously.

The Days Passed variable can provide insights into the recency of customer engagement and the potential impact of previous marketing efforts on customer churn. In our study, we include the Days Passed as a predictor to examine its relationship with customer churn behavior. By analyzing the association between the Days Passed and churn rates, we aim to understand whether the duration since the last contact influences customers' likelihood of churn.

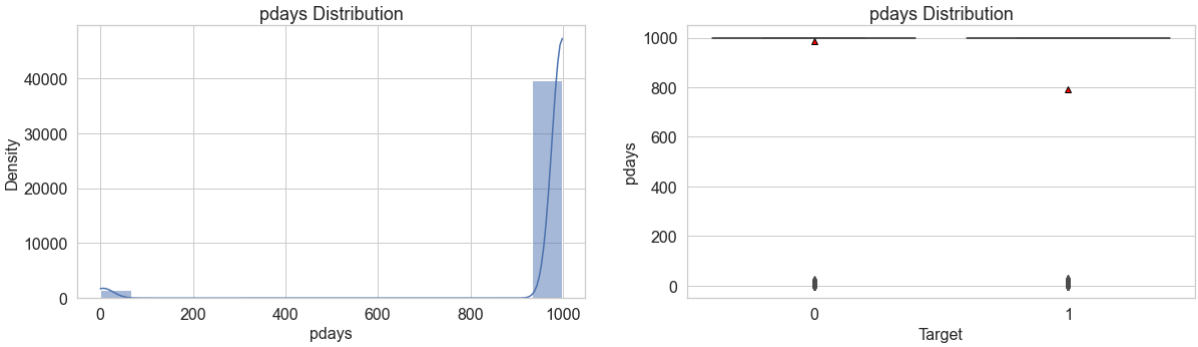

Figure 22: Histogram with KDE-line and boxplots for the variable Days Passed

The analysis of the variable 'days passed' reveals interesting insights about the previous contact history of customers. The left plot of Figure 22 illustrates that the value 999 has the highest density. Since the density is highest at this value, a significant portion of the clients in the dataset had no prior interactions with the bank.

When examining the boxplots in the right plot, we observe that both boxplots are tightly squeezed together. This indicates a lack of variation among customers in terms of the number of days passed since their previous contact, as most of them were not contacted previously. Moreover, the mean of the group that subscribed, denoted by the red triangle, appears to be slightly lower compared to the other group.

Table 23: Quartiles of the variable Days Passed

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	0	999	999	999	999
1	0	999	999	999	999

Note. This table shows the quartiles and whiskers of the 'Days Passed' variable, segmented into groups with a value 1 and 0 for the target variable.

The quartile values shown in Table 23 further emphasize this finding, as all quartiles have the value of 999. This lack of variability in the 'days passed' variable suggests that it may not possess strong explanatory power in predicting customer subscription behavior.

The prevalence of the value 999 and the limited variation in this variable's distribution may limit its ability to provide meaningful insights in predicting customer subscription behavior. The variable will undergo further testing and evaluation to assess its relevance and potential inclusion in the machine learning models. However, given the limited variation and prevalence of the value 999, there is a possibility that it may not significantly contribute to the predictive performance of the models.

11.1.1.9 Previous contact

The variable 'Previous', which stands for previous contact represents the number of contacts performed before the current marketing campaign for a particular customer. It indicates the frequency of previous interactions between the bank and the customer. The Previous Contact variable can provide insights into the customer's historical engagement with the bank and their level of familiarity with the bank's offerings and services. In our study, we include the Previous Contact as a predictor to explore its potential impact on customer churn behavior. By examining the relationship between the Previous Contact and churn rates, we aim to determine whether the frequency of prior contacts affects customers' likelihood of churn.

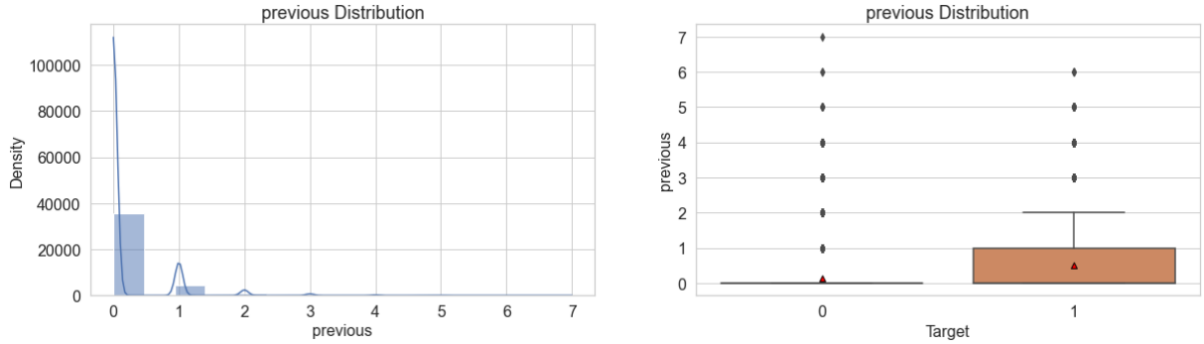

Figure 23: Histogram with KDE-line and boxplots for the variable Previous Contact

The variable 'previous contact' reveals intriguing insights when examining the left plot of figure 23. It is evident that the majority of customers in the dataset have not been contacted previously by the bank, with only a fraction having engaged in prior interactions.

The right plot provides further analysis, demonstrating noteworthy patterns. The boxplot of customers who did not subscribe indicates limited variation, with the data points tightly clustered together. In contrast, the boxplot of customers who subscribed shows that the third quartile has a value of 1, indicating a higher level of previous contact for this group. Additionally, the mean value for this group is higher.

Table 24: Quartiles of the variable Previous Contact

Target Variable (y)	Min	1st Quartile	2nd Quartile	3rd Quartile	Max
0	0	0	0	0	7
1	0	0	0	1	6

Note. This table shows the quartiles and whiskers of the 'Previous Contact' variable, segmented into groups with a value 1 and 0 for the target variable.

Table 24 presents a comprehensive view of the quartiles, emphasizing the prominence of a value of 1 in the third quartile for customers who subscribed.

This finding is particularly intriguing, as it suggests a potential relationship between previous interactions with the bank and customer subscription. The presence of previous contact may influence customer decision-making and positively impact their likelihood of subscribing to the term deposit. However, further analysis and modeling will be necessary to validate and quantify the significance of this relationship.

11.1.2 Binary categorical variables

11.1.2.1 Default

The binary variable 'Default' indicates whether a customer has a credit in default or not. It represents the default status of customers' credit obligations, distinguishing between those who have defaulted ('yes') and those who have not ('no').

The 'Default' variable is a significant factor in assessing customers' financial health and creditworthiness. It serves as an important indicator of their ability to meet their financial obligations, highlighting potential credit risks and financial stability. By examining the 'Default' variable in our analysis, we can delve into the impact of credit default on customer churn behavior and evaluate its role in predicting customer retention.

Understanding the relationship between the 'Default' variable and churn behavior provides valuable insights into the influence of credit default as a contributing factor to customer attrition. Incorporating this variable alongside other demographic, economic, and marketing factors allows us to uncover patterns and associations that contribute to a comprehensive understanding of customer churn dynamics.

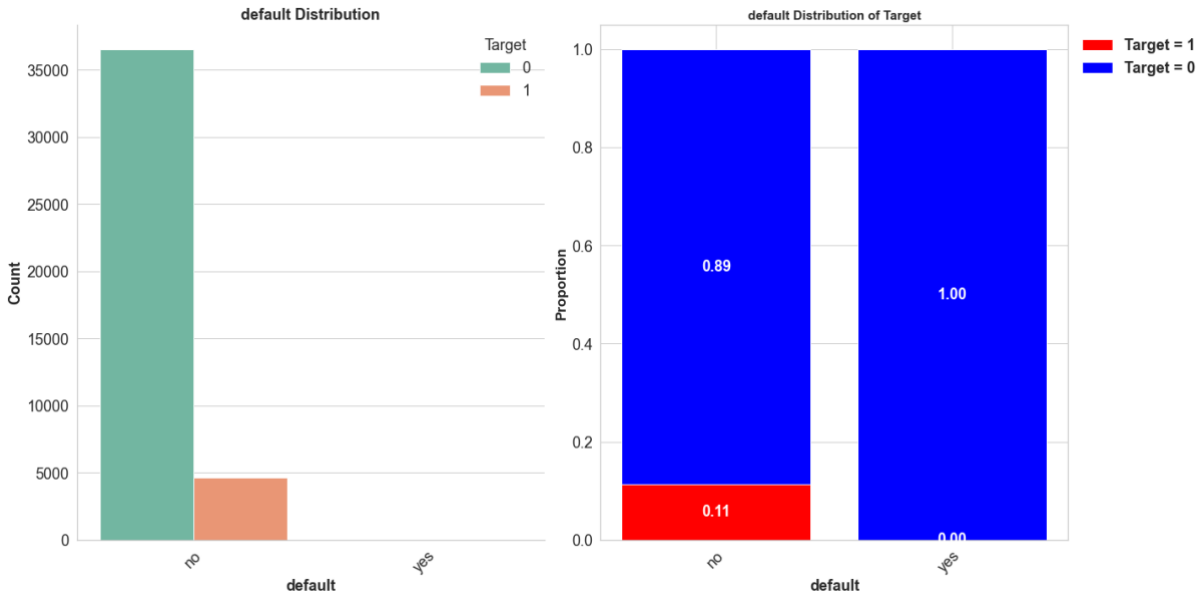

Figure 24: Count and proportions of the Default variable

In the left bar plot 24, we observe the distribution of the binary variable 'Default', color-coded based on the target variable. It is noteworthy that there are no bars representing customers

who had credit in default. This absence can be attributed to the fact that only three customers in the dataset reported having credit in default.
Furthermore, the default variable exhibits outliers, which could potentially explain the limited presence of customers with default credit. The presence of outliers will be further examined and discussed later in the analysis.
The right plot of figure 24 reinforces this observation, as the barplot representing the proportions of the target variable for the category 'yes' is entirely blue, indicating that almost all customers did not subscribe to a term deposit.
The lack of customers with credit in default suggests that this variable may not offer substantial information for predicting customers' subscription behavior to a new term deposit.

11.1.2.2 Loan

The binary variable 'Loan' indicates whether a customer has a personal loan or not. It is a categorical variable that captures the loan status of customers. The value 'yes' indicates that the customer has a personal loan, while 'no' indicates the absence of any personal loan.
The 'Loan' variable provides valuable information about customers' financial commitments and their borrowing behavior. It serves as an indicator of their credit obligations outside of housing-related loans. By considering the 'Loan' variable in our analysis, we can explore the impact of personal loans on customer churn behavior and assess its significance in predicting customer retention.
Understanding the relationship between the 'Loan' variable and churn behavior allows us to examine how customers' personal loan obligations may influence their likelihood of attrition. Analyzing this variable alongside other demographic, economic, and marketing factors helps us uncover patterns and associations that contribute to a comprehensive understanding of customer churn dynamics.

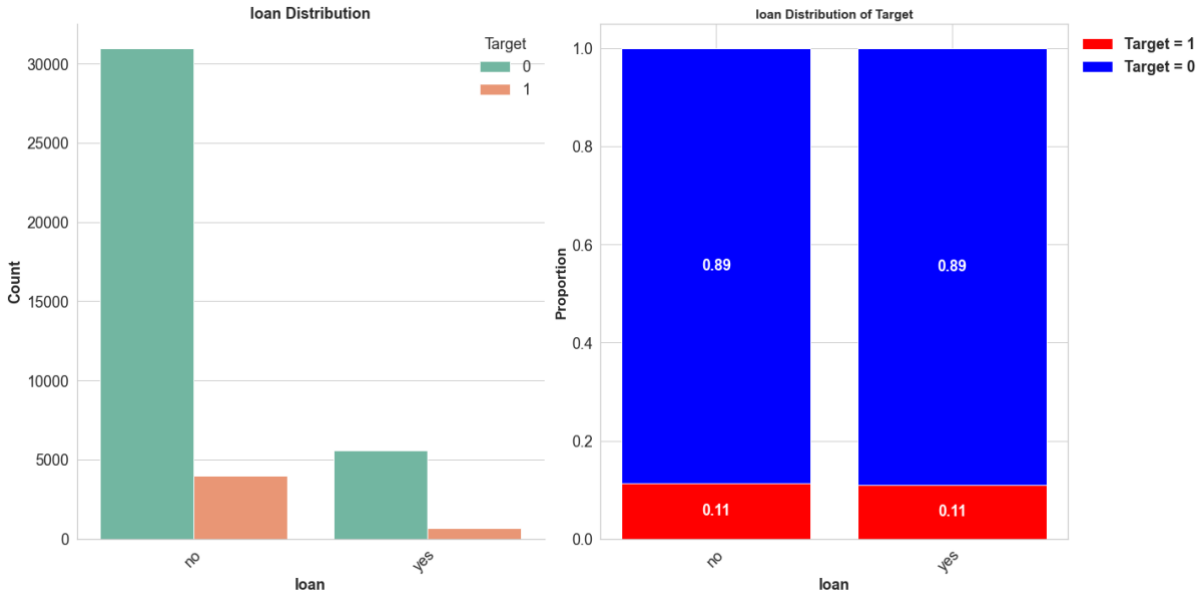

Figure 25: Count and proportions of the Loan variable

The left plot of figure 25 illustrates the distribution and proportions of the binary variable 'Loan'. It is notable that for both categories of the variable, a majority of customers did not opt for a new term deposit, indicating a churned status. The analysis reveals that most customers did not have a personal loan and did not subscribe to a term deposit.
Examining the right plot in figure 25, we observe that the proportions of subscribers and non-subscribers within each loan category are quite similar. This similarity suggests that the variable 'Loan' may not carry significant predictive power when determining customers' subscription behavior.

11.1.2.3 Contact

The variable 'Contact' represents the communication type used for the last contact with the customer. It is a categorical variable that captures the mode of communication employed during the customer interaction. The 'Contact' variable can take on values such as 'telephone' or 'cellular', representing different communication channels.
The 'Contact' variable provides insights into the methods used to reach out to customers during marketing or promotional campaigns. It helps us understand the preferred communication channels and their potential influence on customer churn behavior. By examining the 'Contact' variable in our analysis, we can explore the impact of communication modes on customer engagement and assess their significance in predicting customer retention.
Understanding the relationship between the 'Contact' variable and churn behavior allows us to investigate how different communication channels may affect customers' likelihood of attrition. Analyzing this variable alongside other demographic, economic, and marketing factors allows us to uncover patterns and associations that contribute to a comprehensive understanding of customer churn dynamics.

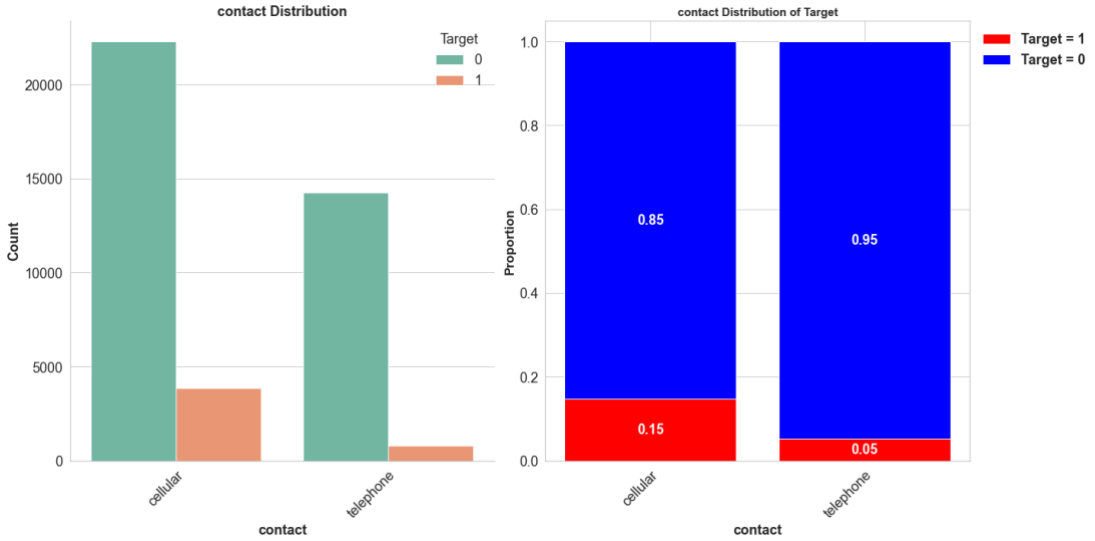

Figure 26: Count and proportions of the Contact variable

The left plot displayed in figure 26 provides an overview of the distribution of the binary variable 'Contact'. It is evident that in both categories of the variable, a significant majority of customers did not subscribe to a new term deposit, implying a churned status. Additionally, it is noteworthy that the majority of customers were last contacted using the cellular communication type and did not subscribe to a term deposit.

Shifting our focus to the right plot in figure 26, we observe that the proportion of subscribers is notably higher among customers who were contacted via the cellular communication type. Conversely, the proportion of subscribers is lower among customers contacted via the telephone. This disparity in proportions between the cellular and telephone categories suggests that the 'Contact' variable may hold some potential in predicting customer churn.

11.1.3 Multi-categorical variables

11.1.3.1 Job

The variable 'job' represents the occupation or type of job that the customer is engaged in. It is a categorical variable that captures the various employment categories or professions of the customers. The 'job' variable includes categories such as administration, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, and services.

The 'job' variable provides insights into the diverse range of occupations among the customers. It helps us understand how different job roles and employment categories may influence their behavior, preferences, and likelihood of churn. By examining the 'job' variable in our analysis, we can explore the impact of specific occupations on customer churn behavior and assess their significance in predicting customer retention.

Understanding the relationship between the 'job' variable and churn behavior allows us to investigate how certain occupations or employment categories may be associated with higher or lower churn rates. Analyzing this variable alongside other demographic, economic, and marketing factors allows us to uncover patterns and associations that contribute to a comprehensive understanding of customer churn dynamics.

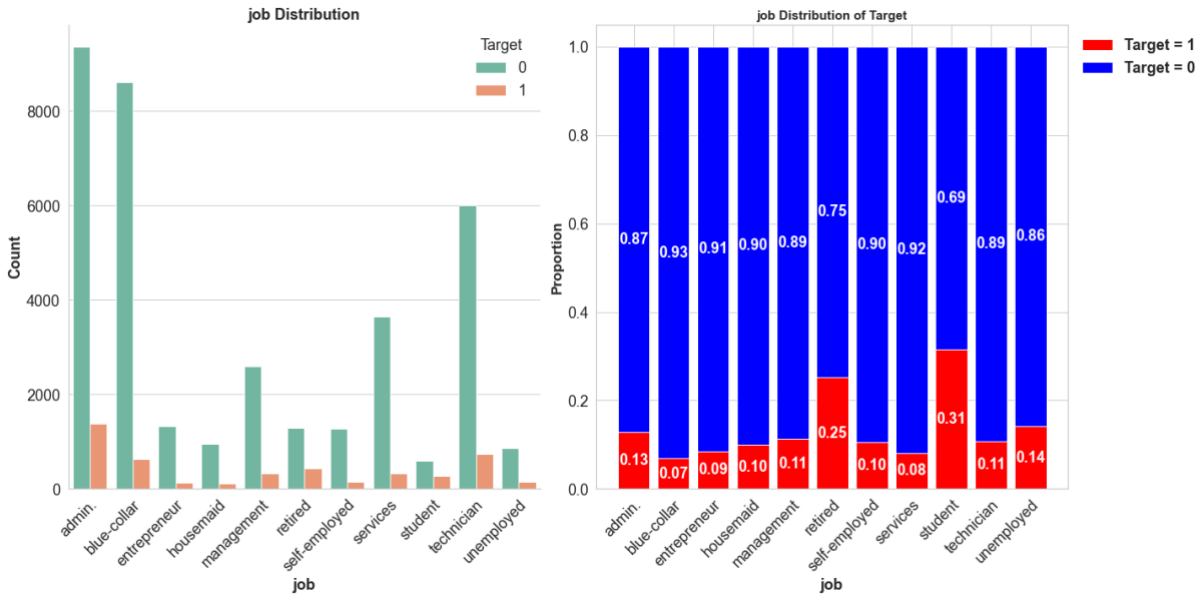

Figure 27: Count and proportions of the Job variable

The left plot in figure 27 presents a visual representation of the distribution of the categorical variable 'Job'. It is evident that across all job categories, the majority of customers did not subscribe to a new term deposit, indicating a churned status. Furthermore, it is worth noting that a significant number of customers fall into the administrative and blue-collar occupations.

When observing the right plot in figure 27, we observe interesting patterns in the proportions of churners and non-churners among the different job categories. Specifically, the retired and student categories exhibit higher proportions of churners compared to non-churners, indicating a potential relationship between these job categories and customer churn. These variations in proportions among the job categories suggest that the 'Job' variable could hold predictive power in determining customer churn.

11.1.3.2 Marital

The variable 'marital' represents the marital status of the customers. It is a categorical variable that captures the different marital status categories, including ‘married,’ ‘divorced,’ and ‘single.’

The 'marital' variable provides insights into the customers' marital status, which can influence their behavior, preferences, and likelihood of churn. By examining the 'marital' variable in our analysis, we can explore how marital status relates to customer churn behavior and assess its significance in predicting customer retention.

Understanding the relationship between the 'marital' variable and churn behavior allows us to investigate whether customers' marital status has an impact on their propensity to churn. Analyzing this variable in conjunction with other demographic, economic, and marketing factors enables us to uncover patterns and associations that contribute to a comprehensive understanding of customer churn dynamics.

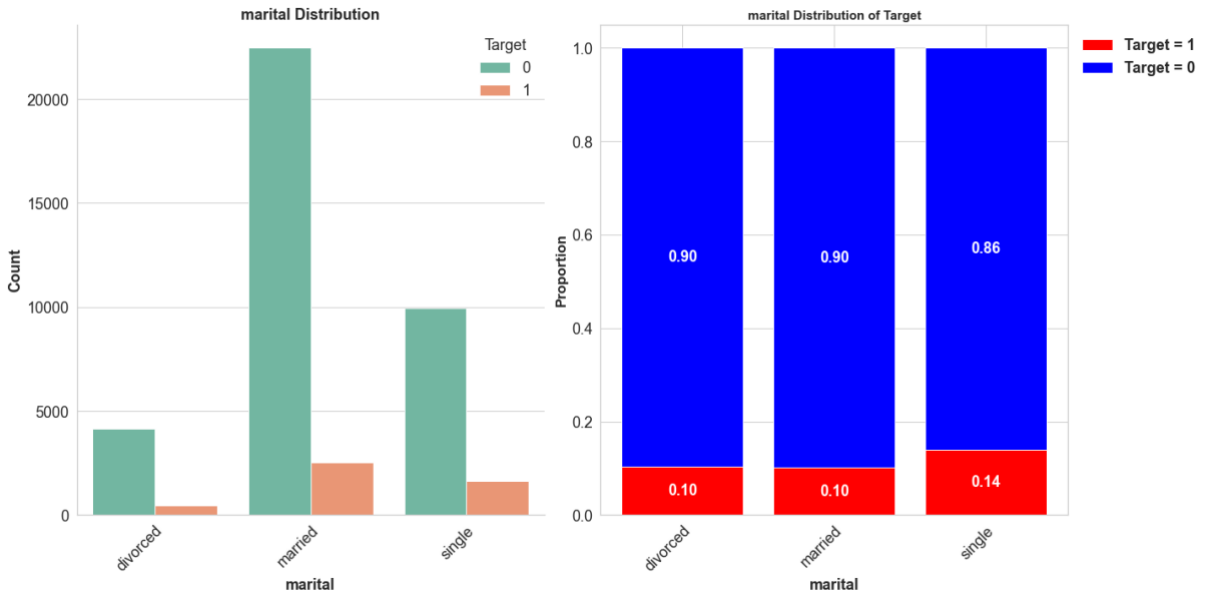

Figure 28: Count and proportions of the Marital variable

The left plot depicted in figure 28 provides insights into the distribution of the categorical variable 'Marital'. It is evident that across all marital categories, the majority of customers did not subscribe to a new term deposit, indicating churn. Furthermore, it is worth noting that a significant number of customers fall into the married category.

Shifting our attention to the right plot in figure 28, we observe interesting patterns in the proportions of churners and non-churners among the different marital categories. Specifically, the single category exhibits higher proportions of churners compared to non-churners, indicating a potential relationship between marital status and customer churn. These variations in proportions among the marital categories may suggest that the 'Marital' variable holds predictive power in determining customer churn.

11.1.3.3 Education

The variable 'education' represents the educational level of the customers. It is a categorical variable that captures different levels of education, including ‘basic.4y,’ ‘basic.6y,’ ‘basic.9y,’ ‘high.school,’ ‘illiterate,’ ‘professional.course,’ and ‘university.degree.’ These categories correspond to specific levels of education attainment.

The categories within the 'education' variable provide information about the customers' educational backgrounds. For instance, ‘basic.4y’ refers to 4 years of basic education, ‘basic.6y’ corresponds to 6 years of basic education, and ‘basic.9y’ represents 9 years of basic education. Additionally, the categories ‘high.school,’ ‘illiterate,’ ‘professional.course,’ and ‘university.degree’ indicate completion of high school, illiteracy, professional courses, and university degrees, respectively.

By examining the 'education' variable in our analysis, we can gain insights into how different levels of education relate to customer churn behavior. This allows us to assess the significance of educational attainment in predicting customer retention. By considering the 'education' variable alongside other demographic, economic, and marketing factors, we can uncover

valuable patterns and associations that contribute to a comprehensive understanding of customer churn dynamics.

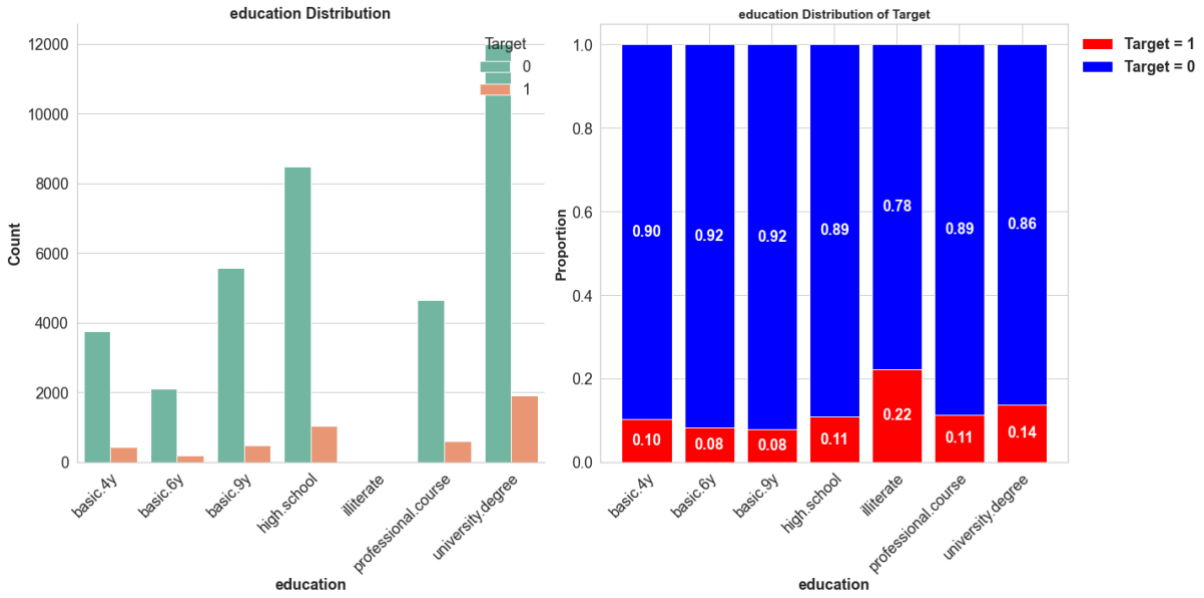

Figure 29: Count and proportions of the Education variable

The left bar plot displayed in figure 29 provides insights into the distribution of the categorical variable 'Education'. It is evident that across all education categories, the majority of customers did not subscribe to a new term deposit, indicating churn. Additionally, it is noteworthy that a significant number of customers hold a university degree. However, the 'illiterate' category stands out with no bars, indicating that only a small number of customers in the dataset reported being illiterate. Furthermore, the presence of missing values in the education variable will be addressed later in the analysis.

Shifting our focus to the right plot in figure 29, we observe interesting variations in the proportions of churners and non-churners among the different education categories. Particularly, the proportion of churners is higher in the 'illiterate' category compared to non-churners, implying that education level may play a significant role in predicting customer churn. These findings highlight the potential value of the 'Education' variable in understanding and predicting customer behavior.

11.1.3.4 Month

The variable 'month' represents the month of the year in which the last contact was made with the customers. It is a categorical variable that captures different months of the year, including ‘mar,’ ‘apr,’ ‘may,’ ‘jun,’ ‘jul,’ ‘aug,’ ‘sep,’ ‘oct,’ ‘nov,’ and ‘dec.’

The 'month' variable provides information about the timing of the last contact, allowing us to examine seasonal patterns and their potential influence on customer churn. By considering the 'month' variable in our analysis, we can explore whether certain months exhibit higher or lower churn rates and assess the impact of seasonality on customer retention.

Analyzing the 'month' variable in relation to other variables, such as campaign duration or customer demographics, can help uncover patterns and associations that contribute to a comprehensive understanding of customer behavior. This information can inform strategic decision-making and the development of targeted marketing strategies, taking into account the temporal dynamics of customer churn.

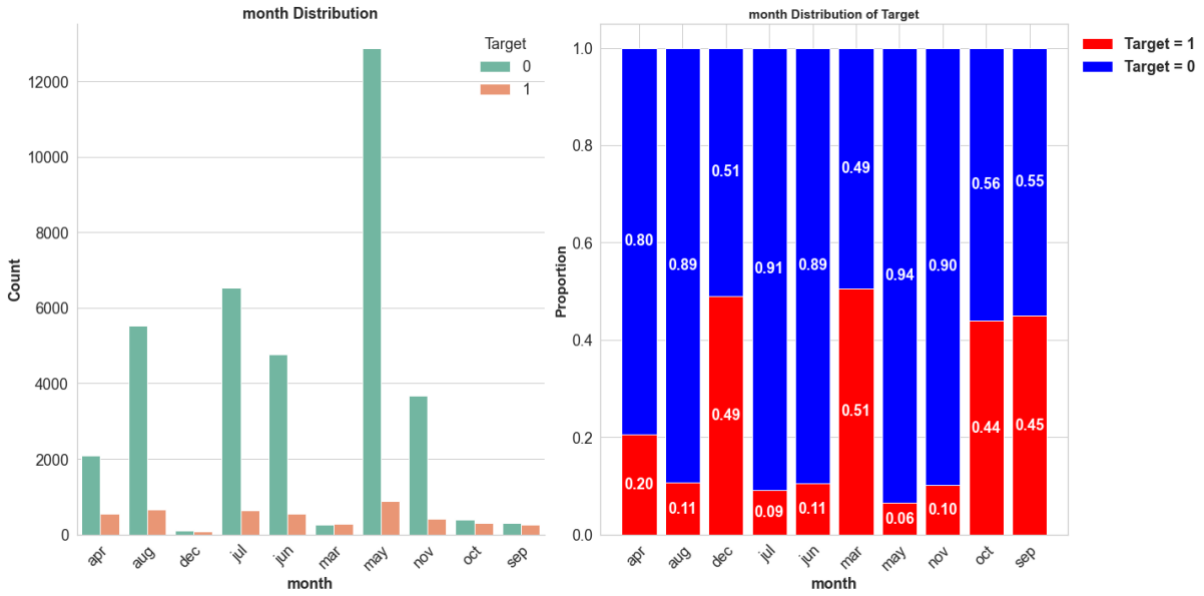

Figure 30: Count and proportions of the Month variable

The left plot in figure 30 provides insights into the distribution of customers based on the month of their last contact. It is evident that the count of customers varies across different months, indicating varying levels of customer engagement during different times of the year. Notably, the month of March stands out with the highest count of last contacts.
Shifting our attention to the right plot in figure 30, we observe interesting patterns in the proportions of churners and non-churners among the different months. Specifically, the months of December, March, October, and September exhibit a higher proportion of churners compared to other months. This suggests that the timing of customer interactions, particularly during these months, may have an impact on their decision to subscribe or churn. These findings underscore the significance of considering the 'Month' variable in predicting customer behavior and developing targeted strategies.

By analyzing the distribution and proportions of customers across different months, we gain valuable insights into the temporal aspects of customer interactions. The higher count of last contacts in March suggests a potential focus on customer engagement during that period. Additionally, the variations in churn proportions among specific months highlight the need for further investigation into the underlying factors contributing to these patterns. Incorporating the 'Month' variable in predictive models can aid in capturing the seasonality

and temporal dynamics of customer behavior, enabling more effective marketing strategies and customer retention efforts.

11.1.3.5 Day of the week

The variable 'day of the week' represents the day of the week in which the last contact was made with the customers. It is a categorical variable that captures the different days of the week, including ‘mon,’ ‘tue,’ ‘wed,’ ‘thu,’ and ‘fri.’

The 'day of the week' variable provides information about the timing and frequency of customer contacts throughout the week, allowing us to examine any potential patterns or trends in relation to customer churn. By considering the 'day of the week' variable in our analysis, we can explore whether certain days exhibit higher or lower churn rates and assess the influence of weekly dynamics on customer retention.

Analyzing the 'day of the week' variable in conjunction with other variables, such as campaign duration or customer characteristics, can help reveal associations and insights into customer behavior. This knowledge can guide marketing strategies and decision-making, enabling organizations to optimize their outreach efforts on specific days or adjust their engagement strategies based on weekly patterns observed in customer churn.

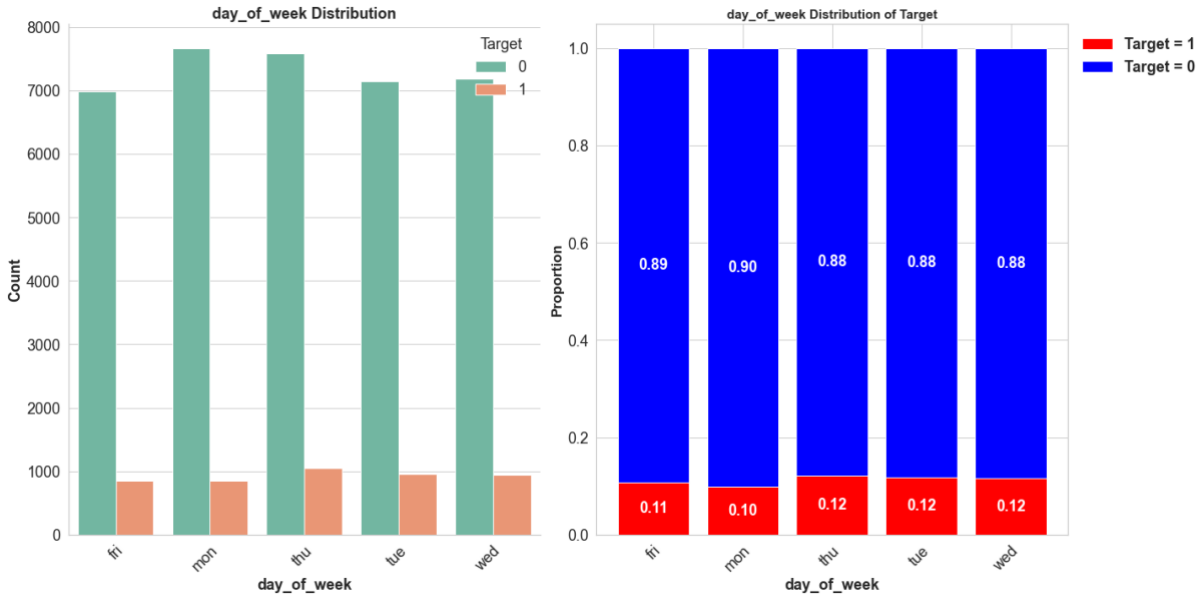

Figure 31: Count and proportions of the Day of the Week variable

The left plot in figure 31 provides valuable insights into the distribution and proportions of the categorical variable ‘Day of the Week,’ representing the last contact day of customers by the bank. The plot highlights that the majority of customers did not subscribe to a term deposit across all days of the week. Additionally, it is evident that a higher number of customers were last contacted on Mondays.

Examining the right plot in figure 31, we can observe a relatively balanced distribution of subscribers and non-subscribers across all days of the week. This suggests that the specific

day of the week when the bank last contacted the customer may not have a significant impact on their decision to subscribe or churn.

11.1.4 Correlation Heatmap

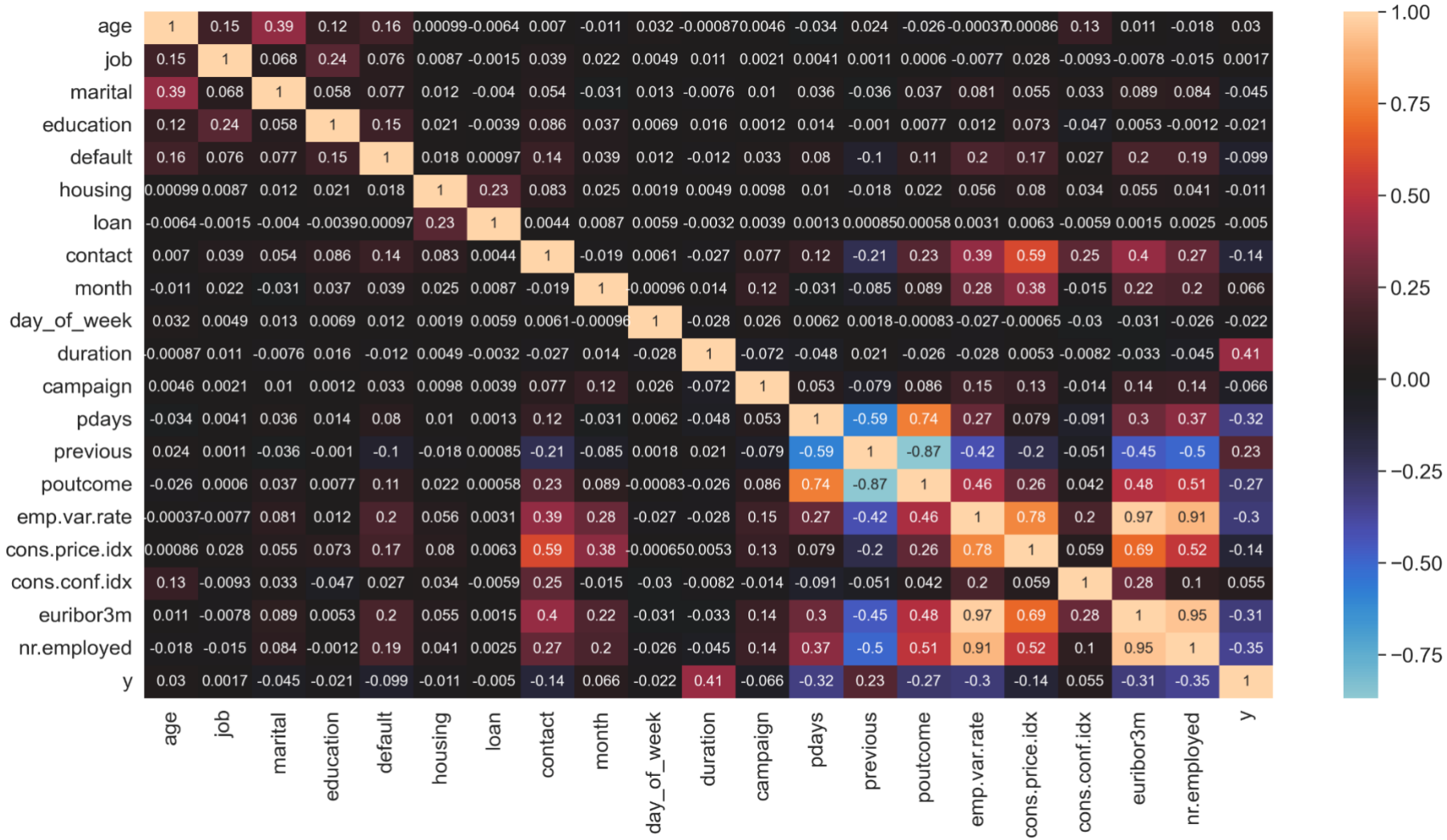

Figure 32: Heatmap of all the variables in the Dataset