T.C.
ISTANBUL AYDIN UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

MACHINE LEARNING APPROACH TO THE PREDICTION OF BANK CUSTOMER CHURN PROBLEM.

MASTER'S THESIS

Omobola Azeezat OKOCHA

Department of Software Engineering
Artificial Intelligence and Data Science Program

MARCH, 2023

T.C.
ISTANBUL AYDIN UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

MACHINE LEARNING APPROACH TO THE PREDICTION OF BANK CUSTOMER CHURN PROBLEM.

MASTER'S THESIS

Omobola Azeezat OKOCHA

(Y2013.140021)

Department of Software Engineering
Artificial Intelligence and Data Science Program

Thesis Advisor: Assoc. Prof. Dr. Ilham HUSEYINOV

MARCH, 2023

ONAY FORMU

DECLARATION

I hereby declare with the respect that the study "AUTO QUESTION ANSWERING SYSTEM USING DBPEDIA", which I submitted as a Master thesis, is written without any assistance in violation of scientific ethics and traditions in all the processes from the project phase to the conclusion of the thesis and that the works I have benefited are from those shown in the References. (20/03/2022)

Omobola Azeezat OKOCHA

FOREWORD

First, I would like to express my endless gratitude to God for being who I am right now and helping me find the patience and strength within myself to complete this thesis.

I would also like to thank my husband and children for their love and understanding during the period of this master's degree program and also for their support towards achieving my dream.

I feel very fortunate to have Dr. ILHAM HUSEYINOV as my advisor and want to express my appreciation for guiding me within the whole research process in a patient and effective manner.

Finally, I would like to acknowledge the important contribution of Istanbul Aydin University to my life, not only from an academic perspective but helping to meet great people that inspire, challenge, support and motivate me.

March, 2023 Omobola Azeezat OKOCHA

MACHINE LEARNING APPROACH TO THE PREDICTION OF BANK CUSTOMER CHURN PROBLEM

ABSTRACT

In the modern banking industry, customers have a plethora of options when it comes to deciding where to invest their money. As a result, customer retention and churn have become significant challenges for most banks. In an effort to address the issue of customer churn, this research employs various machine learning algorithms such as Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, eXtreme Gradient Boosting, and Light Gradient Boosting.

The study utilizes a feature selection technique to remove irrelevant features and identify the most relevant ones. Additionally, the resulting dataset is balanced using the SMOTE method. The performance of classifiers on balanced and imbalanced datasets is compared in terms of accuracy, recall, precision, and overall performance. The results demonstrate that no classifier outperformed others when dealing with imbalanced data (before SMOTE is applied). However, in the case of balanced data (after SMOTE is applied), the Random Forest classifier outperformed other classifiers by a significant margin.

Keywords: Customer Churn in Bank, Support Vector Machine, Random Forest, Gradient Boosting, SMOTE

BANKA MÜŞTERİ KAYIP PROBLEMİNİN TAHMİN EDİLMESİNE MAKİNE ÖĞRENME YAKLAŞIMI.

ÖZET

Modern bankacılık sektöründe, müşteriler paralarını nereye yatıracaklarına karar verme konusunda çok sayıda seçeneğe sahiptir. Sonuç olarak, müşteri tutma ve kayıp çoğu banka için önemli zorluklar haline geldi. Müşteri kaybı sorununu ele almak amacıyla bu araştırma, Lojistik Regresyon, Destek Vektör Makinesi, Rastgele Orman, Gradient Boosting, eXtreme Gradient Boosting ve Light Gradient Boosting gibi çeşitli makine öğrenimi algoritmalarını kullanır.

Çalışma, ilgisiz özellikleri kaldırmak ve en alakalı olanları belirlemek için bir özellik seçme tekniği kullanır. Ek olarak, ortaya çıkan veri seti SMOTE yöntemi kullanılarak dengelenir. Sınıflandırıcıların dengeli ve dengesiz veri kümelerindeki performansı, doğruluk, hatırlama, kesinlik ve genel performans açısından karşılaştırılır. Sonuçlar, dengesiz verilerle uğraşırken (SMOTE uygulanmadan önce) hiçbir sınıflandırıcının diğerlerinden daha iyi performans göstermediğini göstermektedir. Ancak, dengeli veriler söz konusu olduğunda (SMOTE uygulandıktan sonra), Rastgele Orman sınıflandırıcısı, diğer sınıflandırıcılardan önemli bir farkla daha iyi performans gösterdi.

Anahtar Kelimeler: Bankada Müşteri Kaybı, Destek Vektör Makinesi, Rastgele Orman, Gradient Boost, SMOTE

TABLE OF CONTENTS

ABBREVIATIONS

API **:** Application Program Interface

CCM : Customer Correlation Management

CP : Customer Preservation

FCM : Fuzzy C-Means

GB : Gradient Boosting

RF : Random Forest

ROC : Receiver Operating Characteristic

LGBM : Light Gradient Boosting Machine

LR : Logistic Regression

PFCM : Possibilistic Fuzzy C-Means

SMOTE : Synthetic Minority Oversampling Technique

SVC : Support Vector Classifier

SVM : Support Vector Machine

XGBOOST : Extreme Gradient Boosting

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. Purpose of Study

Customer churn is a term used to describe the gradual loss of clients by a corporation or organization over a period of time. However, it is more commonly known as "customer attrition". One of the most important concerns for businesses is the process of acquiring new customers and ensuring that they remain loyal to the company. While new companies focus on obtaining new customers, established businesses tend to concentrate on retaining the clients they already have so that they can potentially cross-sell to them. Freeman (1999) has pointed out that one of the most critical strategies for enhancing the value of customers is to retain them for an extended period of time. This approach allows companies to build stronger relationships with their customers and potentially increase their revenue through repeat business and cross-selling opportunities (Freeman, 1999).

In the present era, the rise of electronic commerce has resulted in a significant increase in the amount of information that is available. As noted by Peppard (Peppard 2000), the emergence of the internet as a channel for conducting business has given customers a greater sense of empowerment, as they are no longer limited to the products or services offered by a single company. This has resulted in an upsurge in competition, as rivals are only a few clicks away from each other. With customer empowerment on the rise, it is likely that the rate of customer churn will also increase (Lejeune, 2001). To mitigate this risk, businesses must ensure that they have access to the most advanced and dependable tools for analyzing customer behavior and predicting the probability of customer attrition in the future. By leveraging these tools, companies can take proactive steps to retain their customers and build stronger relationships with them, thereby bolstering their overall competitiveness in the marketplace.

As per the source (Lejeune 2001), the practice of churn management refers to a set of tactics and techniques employed by companies to establish and sustain lucrative partnerships with their current customer base. This approach typically involves a range of activities aimed at reducing customer attrition and maximizing customer retention, such as improving the quality of customer service, offering customized products or services, and incentivizing customers to remain loyal to the company. Ultimately, the goal of churn management is to help businesses maintain a steady stream of revenue from their existing customers, while simultaneously reducing the need for costly customer acquisition efforts. By implementing effective churn management strategies, companies can build stronger and more sustainable relationships with their customers, ultimately leading to greater long-term success and profitability.

The objective of this research is to develop a reliable and precise predictive model for forecasting customer churn in a banking setting, using machine learning techniques. In order to provide a better understanding of the context and significance of this study, the report begins by presenting relevant background information and statistics related to the increasing prevalence of customer churn within the banking industry. Following this, we outline our problem definition and research question, which seek to identify the most effective approaches for predicting and mitigating customer attrition in a bank. By leveraging advanced machine learning methods, we aim to provide practical insights and recommendations that can help banks improve their customer retention strategies and enhance their overall competitiveness in the market.

B. Background and Statistics

The banking sector faces significant financial consequences from customer churn as acquiring new clients can often be more costly than retaining current ones. Hence, it is imperative for banks to take preemptive measures to prevent customer attrition by implementing customer churn prediction techniques to identify individuals who may be at risk of leaving. This is especially important in a highly competitive market where customers have numerous financial service alternatives available to them. There are multiple reasons for customer turnover in the banking industry, such as unsatisfactory service, perceived insufficient value, and the availability of more attractive options from competitors. The extent of customer churn may fluctuate depending on various factors.

Customer turnover in the banking industry can also be influenced by a multitude of factors, including but not limited to, alterations in personal circumstances, such as relocation or fluctuations in income, modifications in the bank's products or services, and changes in the overall banking sector. These external and internal shifts can create a ripple effect on customer behavior, leading to churn or loyalty. Therefore, it is essential for banks to stay abreast of these changes and proactively adapt their strategies to retain their customer base.

Banks can leverage the power of machine learning algorithms to predict customer churn and retain potentially at-risk customers. These algorithms can analyze an array of data sources, including transaction history, demographic information, and customer feedback, to detect patterns and trends that indicate a higher risk of churn. By gaining a deeper understanding of each customer's unique needs and preferences, banks can improve their chances of keeping that customer. The use of machine learning algorithms in this way enables banks to personalize their approach to customer retention and provide a more targeted and effective strategy. By adopting this approach, banks can reduce customer attrition rates, increase customer satisfaction, and ultimately achieve a competitive advantage in the marketplace.

Machine learning methods have been widely used in the banking industry to predict customer churn. Numerous studies have been conducted to evaluate various machine learning techniques, such as decision trees, random forests, and Gradient Boosting, to identify the most effective approach to predicting customer attrition. These studies have also explored additional data sources and feature sets that may be crucial for accurately forecasting customer turnover. By analyzing a diverse range of factors, including transaction history, demographic information, and customer feedback, these machine learning models can uncover hidden patterns and trends that may indicate an increased risk of customer churn. The insights gained from these models can help banks make informed decisions and take proactive measures to retain customers and improve overall customer satisfaction.

Adopting machine learning techniques to predict customer churn can significantly enhance a bank's capacity to retain customers and maintain profitability in a highly competitive market. By utilizing these techniques, banks can identify at-risk customers and take necessary measures to reduce churn, which can help to improve the bottom line.

In order to fully grasp the importance of this field of research, it is crucial to examine the current statistics on the magnitude of churn and its associated costs in the banking industry. By doing so, banks can better understand the impact of customer attrition on their business and the benefits of implementing machine learning models to predict and prevent it. This can help banks to improve their customer retention rates, increase customer satisfaction, and ultimately gain a competitive advantage in the marketplace. The following is a brief discussion of current statistics on customer churn in the banking sector.

- According to a study conducted by Bain & Company, the cost of acquiring a new customer can be as much as five times higher than retaining an existing one for banks. Moreover, retaining existing customers can lead to a greater customer lifetime value since loyal clients are more likely to make repeat purchases and refer others to the bank's services. This underscores the importance of customer retention in the banking industry and highlights the potential financial benefits that can be realized by reducing churn rates. By focusing on customer retention and leveraging machine learning models to predict churn, banks can improve their profitability and gain a competitive edge in the market (Bain 2018).
- Several studies have examined customer churn rates in the banking industry, revealing that banks experience an average annual churn rate ranging from 15% to 20%. For example, a study conducted by the Federal Reserve Bank of Chicago found that the average annual customer turnover rate for American banks in 2013 was approximately 17.5%. These statistics indicate the prevalence of customer attrition in the banking industry and highlight the need for proactive measures to retain customers. By leveraging machine learning models to predict churn and taking necessary actions to retain at-risk customers, banks can reduce their churn rates, improve customer satisfaction, and ultimately drive revenue growth (Federal Reserve Bank of Chicago 2013).
- According to a study conducted by Capgemini, the average customer churn rate for banks in Europe was approximately 20% in 2015. This indicates a significant level of customer turnover in the European banking market and highlights the importance of customer retention strategies. By utilizing machine learning algorithms to predict

customer attrition and taking proactive measures to retain customers, banks can improve their customer retention rates, reduce churn-related costs, and enhance their profitability. This underscores the significance of investing in machine learning technologies for customer churn prediction in the banking sector (Capgemini, 2015).

- The International Journal of Bank Marketing reported that American banks experience an average annual customer churn rate of approximately 17.5%. The study also revealed that while demographic factors such as age and wealth were less relevant, customer satisfaction and loyalty were the most significant indicators of customer turnover. This highlights the importance of fostering strong relationships with customers, providing quality service, and maintaining high levels of customer satisfaction to reduce churn rates. By leveraging machine learning algorithms to analyze customer data and identify at-risk customers, banks can take proactive measures to retain customers and enhance their profitability (Al-Swidi et al, 2013).
- As per a study published in the Journal of Financial Services Marketing, banks in the United Kingdom experience an average customer churn rate of 15% annually. The research also found that customers who reported higher levels of satisfaction with their bank were less likely to switch to another institution. Conversely, customers who encountered problems with their bank were more likely to churn. These findings highlight the importance of prioritizing customer satisfaction and addressing customer concerns to reduce customer turnover rates. By leveraging machine learning algorithms to analyze customer feedback and identify potential churners, banks can take proactive measures to retain customers and enhance their long-term profitability (McLeod et al, 2000).
- The Journal of Financial Services Marketing published a study revealing that Australian banks experience an average customer churn rate of 20% each year. The research also found that, when compared to other factors such as age, income, and gender, customer satisfaction and loyalty were the most significant predictors of customer turnover. This underscores the importance of providing quality service and fostering strong relationships with customers to reduce churn rates. By using machine learning algorithms to analyze customer behavior and identify those at risk

of churning, banks can take proactive steps to retain customers and boost their long-term profitability (Frow et al, 2006).

C. Problem Definition

The primary challenge faced by the vast majority of businesses operating in industries where there is little or no cost involved in switching is the loss of customers. Among these industries, the banking industry is one of the top five, with a yearly turnover rate of around 20%. This problem affects the banking industry, and to solve it, we need to identify potential churners before they actually leave. To achieve this, it is of utmost importance to develop a model that predicts future churners accurately.

However, the banking industry has a unique characteristic in that there is no limit to the number of bank accounts an individual can have, making it difficult to trace and quantify customer churn. Developing a predictive model that can accurately identify customers likely to defect in the near future would be an arduous and time-consuming task.

The first step to achieve this goal is to identify what "churn" means and who a "churner" is, and then focus on anticipating "churn." Additionally, since the churn class is rare in churn datasets, dealing with this imbalance in the dataset can help to improve the performance of the predictive model. Therefore, it is crucial to tackle the issue of identifying potential churners and develop an effective model to predict churn in the banking industry.

D. Research Question

- In the realm of the banking industry, what is the meaning and definition of "customer churn"? Specifically, what constitutes a customer's decision to terminate their relationship with a bank and switch to another institution?
- What are the possible attributes that can be employed to construct a forecasting model for customer churn in the banking industry?
- What What solutions or strategies can be implemented to address the issue of data imbalance in churn datasets?

E. Research Objective

The primary aim of this study is to determine the most effective Supervised Machine Learning techniques that can accurately predict customer churn in a bank's customer data. The research aims to identify which specific methods and approaches can be employed to predict which customers are most likely to churn, allowing banks to prioritize these customers and take action to retain their loyalty. By focusing on retaining existing customers, banks can foster growth and maintain their profitability. Therefore, the research aims to contribute to the development of more efficient strategies for identifying and retaining valuable customers, leading to better outcomes for both the bank and its customers.

The goals and aims of the research can be summarized as follows:

- The objective is to gather the necessary customer data from the organization for the purpose of the research.
- The research involves comprehending the data, detecting any potential data-related concerns, and resolving them to enable the application of machine learning algorithms.
- The research entails data preparation procedures such as sampling, encoding, feature selection, and data splitting to ensure the data is in a suitable form for analysis and modeling.
- The research involves constructing various supervised machine learning models such as Logistic Regression, Support Vector Machine, Random Forest, Gradient boosting, Xtreme Gradient boosting, and Light gradient boosting, and assessing their performance on a training dataset.
- The research involves validating the constructed models on a separate validation dataset, and based on the evaluation metrics, identifying the most effective model for predicting customer churn among all the models tested.
- Next, the research will evaluate the top-performing model among all the constructed supervised models by testing it on a separate test dataset, and analyzing the results obtained.
- The research will identify any limitations encountered during the study and suggest potential areas for future research.

F. Document Outline

This particular section of the thesis document provides an overview of the entire report, which begins by clearly defining and explaining the research problem and highlighting its significance. The report will also delve into the specifics of the problem, outlining its purpose and identifying relevant research questions that will guide the research process.

Chapter 2 (Literature Review) This chapter provides an in-depth analysis of past research conducted in this field, with a focus on comparing and contrasting various supervised machine learning techniques that have been used to predict customer churn. Specifically, the chapter highlights the applications of techniques such as Logistic Regression, Support Vector Machine, Random Forest, Gradient boosting, Xtreme Gradient boosting, and Light gradient boosting, and discusses their respective specifications. Additionally, the chapter outlines the most valuable and relevant research conducted in this area.

Chapter 3 (Design and Methodology) provides a detailed description of the design and methodology that were employed to address the research problem. This chapter offers an in-depth explanation of the specific steps that were taken to conduct the study, and it delves into the details of each step.

Chapter 4 (Implementation and Results) showcases the implementation details and results obtained from the study. It provides a detailed explanation of the specific models that were selected and used for the research. Furthermore, this chapter provides a comprehensive justification for the selection of these models and evaluates their performance. The hypothesis of the research is also considered and evaluated by comparing the results obtained from the models.

Chapter 5 (Conclusion) provides an overall discussion of the research problem, the obtained results, and their evaluation. It provides a summary of the research and highlights the contribution of the research towards addressing the research question. Additionally, this chapter suggests potential areas for future research in a related field.

II. LITERATURE REVIEW

This chapter presents a comprehensive analysis and assessment of the latest studies related to customer churn prediction. The primary objective of this review is to conduct a critical appraisal of the previous research to identify areas where additional knowledge is needed and to suggest potential avenues for future investigations. Specifically, this literature review focuses on the historical application of machine learning techniques in predicting the likelihood of bank customers to churn, a topic that has been a subject of intense scrutiny over the last decade.

The terms "attrition" and "defection" are alternative expressions used to describe the phenomenon of customers leaving a bank. This issue poses a significant threat to the financial well-being of banks and their market position. As a result, predicting and preventing customer churn has become a critical priority for banks. To achieve this, banks must be able to identify customers who are at risk of leaving and take proactive measures to retain them.

Numerous methods exist for forecasting customer churn, encompassing both traditional statistical approaches such as logistic regression and survival analysis and more modern machine learning techniques. The recent surge in the popularity of machine learning techniques can be attributed to their ability to handle massive amounts of data and produce highly precise predictions.

Lovelace et al. were among the pioneers in developing a machine learning approach for predicting customer attrition. They utilized a neural network to classify customers as either churners or non-churners, based on a combination of customer attributes and transactional data. The neural network was trained using the backpropagation algorithm, which led to a high accuracy rate of 82.9% (Lovelace et al, 2001).

Brajdic et al. proposed an alternative machine learning technique for predicting customer attrition. Their approach involved utilizing a decision tree algorithm to classify customers as churners or non-churners, based on a dataset comprising customer demographics and account information. The decision tree achieved an accuracy rate of 84.4%. Other studies have explored the use of ensemble learning methods to predict bank customer churn. Yin et al. tested a boosting-based ensemble approach on a dataset of Chinese bank clients and found that it achieved an accuracy rate of 87.1%. The study also revealed that the customer's account balance, credit limit, and number of products owned were the top three most significant factors influencing customer churn, as determined by the ensemble approach (Bradjic et al 2006).

According to the research conducted by Zhao, Jing, and their team, the Support Vector Machine (SVM) approach is a highly effective classification method capable of handling real-world problems that conventional methods cannot, such as nonlinearity, high dimensionality, and local minimum issues in predicting bank customer churn. The researchers utilized a dataset containing customer demographics, account information, and transaction data to train the SVM, which resulted in an accuracy rate of 87.2% in predicting customer churn (Zhao et al, 2008).

Farquad, Ravi, and Raju observed that although the SVM is a state-of-the-art classification model, it has the disadvantage of being a "black box" model that lacks transparency and is incomprehensible to people. To address this issue, they developed a hybrid approach that involved first reducing features using SVM-recursive feature removal. After extracting the support vectors, they constructed the SVM model and generated rules using the Naive Bayes tree. By outperforming the SVM and simplifying the model without feature selection, the researchers demonstrated the effectiveness of their approach (Farquad et al, 2014).

In the study conducted by Deepika Ghanta, Guru Sree Ram Tholeti, and Sunkara Venkata Krishna, it was highlighted that one of the leading telecommunication service providers in India has achieved substantial revenues of a couple of billion dollars in just three months, while other competing service providers are struggling to improve their revenues. The study emphasizes the importance of retaining high potential customers and understanding the risks of them switching to a new service provider. The authors mention that building a standard customer churn prediction model, rather than just advertising

about their services, can significantly impact the profits of telecom service companies, as retaining existing high potential customers is easier than acquiring new customers. The study further discusses the use of various models, including logistic regression, decision trees, neural networks, and naive Bayes, to review their accuracies and compare each model for building a standard prediction system. The study reports accuracy rates of 73.54% for decision trees, 76.33% for naive Bayes, 75.01% for neural networks, and an impressive 80.88% for logistic regression (Ghanta et al, 2021).

.

In the study conducted by Sivasankar and Vijaya it was emphasized that the success of every organization or firm depends on Customer Preservation (CP) and Customer Correlation Management (CCM), which are identified as the two parameters determining the rate at which customers decide to subscribe to the same organization. The study highlights that higher service quality can reduce the chances of customer churn, and discusses the analysis and prediction of various attributes in industries such as telecommunication, banking, and financial institutions. The study also mentions the importance of customer churn forecast in helping organizations retain valuable customers and avoid failure in a competitive market. Furthermore, Sivasankar and Vijaya mention that the use of a single classifier does not result in higher churn forecast accuracy, and highlight the trend of combining unsupervised and supervised techniques for better classification accuracy. The study also emphasizes the significant role of unsupervised classification in hybrid learning techniques. The work specifically focuses on various unsupervised learning techniques, such as Fuzzy C-Means (FCM), Possibilistic Fuzzy C-Means (PFCM), and K-Means clustering (K-Means), for customer segmentation and prediction of better customer segmentation. The clusters are divided for training and testing using the Holdout method, where training is carried out by decision tree and testing is done using the generated model. The study concludes that the K-Means clustering algorithm, along with the decision tree, helps improve the results of churn prediction in the telecommunication industry, as shown in the churn prediction dataset experiment conducted by Sivasankar and Vijaya (2017).

Xiu, Li et al proposed a novel learning method called improved balanced random forests (IBRF) for churn prediction. They demonstrated the application of IBRF to address the challenge of imbalanced data distribution in churn prediction, and show that IBRF

improves prediction accuracy significantly compared to other algorithms, such as artificial neural networks, decision trees, and class-weighted core support vector machines. The authors integrated sampling techniques and cost-sensitive learning into the IBRF approach to achieve better performance than existing algorithms. Furthermore, their experiments on a real bank customer churn dataset reveal that IBRF outperforms other random forests algorithms, such as balanced random forests and weighted random forests (Li et al, 2009).

Gao et al. conducted research on the Australian banking industry, where they employed a machine learning technique to forecast customer churn (Gao et al, 2019). Similarly, Kim et al. conducted a study on the South Korean banking industry, where they utilized a machine learning approach to predict customer churn (Kim, 2020). Both studies demonstrated the effectiveness of machine learning in predicting customer attrition and identified several key variables that were strongly linked to churn. In another study by Kumar et al., machine learning was employed to forecast customer attrition in the banking sector. The authors found that their machine learning approach was able to accurately predict customer churn The focus of the research paper was to identify the key predictor variables that are crucial in effectively predicting credit card churn. Additionally, the decision tree J48 generated rules that could serve as an expert system to provide an early warning system. The data used for analysis was obtained from the Business Intelligence Cup that was hosted by the University of Chile back in 2004 (Dudyala et al, 2008).

Based on the available research, it appears that utilizing machine learning techniques for predicting bank customer churn is a highly promising method. While there is still a lot of work to be done in this area, machine learning models have shown considerable potential in accurately predicting customer attrition, which could help banks improve their client retention efforts.

III. METHODOLOGY

A. Design

In this particular chapter, we will be guided through a comprehensive understanding of the initial three machine learning workflows that have been utilized within the scope of this research.

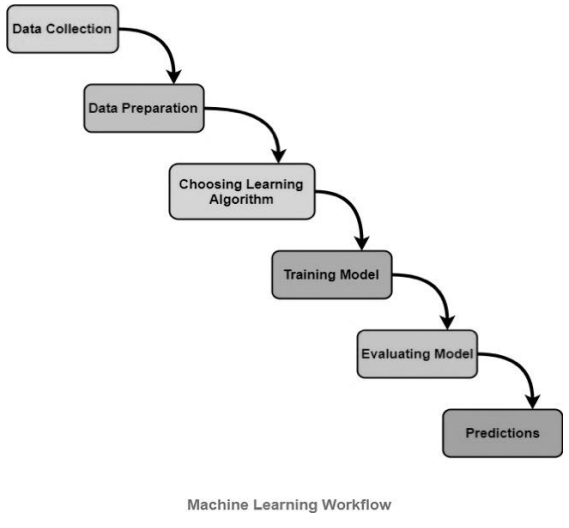

Figure 1 Machine Learning Workflow

B. Data Collection

There exist numerous techniques through which data can be gathered. The methods that can be employed to collect data encompass a wide range of possibilities and may include:

1. **Surveys and Questionnaires**: One of the methods commonly used to gather data is through surveys and questionnaires. This approach entails obtaining information from individuals by administering either structured or unstructured sets of questions. Depending on the type of questions being asked, the data gathered may take on a quantitative or qualitative form.

2. **Experiments**: When it is necessary to establish causal relationships, experiments may be designed to gather data under highly controlled conditions. This particular method is advantageous in situations where the research question demands a rigorous and systematic approach to data collection.
3. **Observational Studies**: Observational studies are a technique that involves watching individuals or objects in their natural environment without interfering with their behavior. This method can be either passive or active in nature, depending on whether the subject being observed is aware of the data collection process. In passive observational studies, data is gathered covertly, without the knowledge of the subject, whereas in active observational studies, the subject being observed is fully informed about the data gathering process.
4. **Log Files**: Log files are computer-generated records that contain information pertaining to the activities and performance of computer systems or applications. These files can prove to be quite useful in providing valuable insights into various aspects of website or application usage patterns. By analyzing the information contained within log files, it is possible to gain a better understanding of how users interact with websites or applications and identify areas for potential improvement.
5. **API Calls**: API (Application Programming Interface) calls provide a means of accessing data stored in remote systems. This technology enables users to retrieve information from a variety of online sources, including social media platforms, websites, and other web-based services. By making API calls, users can obtain data in a structured and standardized format, which can then be processed and analyzed to derive meaningful insights.
6. **Scraping**: Web scraping is a data extraction technique that involves retrieving information from websites. The data obtained from web scraping can be utilized for various purposes, such as performing sentiment analysis and conducting market research. This method involves automatically collecting data from web pages, parsing the information, and then storing it in a structured format that can be analyzed using software tools.

There are several techniques that are commonly utilized to gather data in the context of machine learning. The selection of an appropriate data collection method is dependent on various factors, including the nature of the research question, the specific type of data

that is needed, and the resources that are available to the researcher. Some of the frequently employed methods include those discussed earlier, such as surveys and questionnaires, experiments, observational studies, API calls, and web scraping.

C. Data Preparation

The process of data preparation involves converting raw data into a structure that is well-suited for use in machine learning algorithms. This step is of paramount importance for the successful implementation of a machine learning project, since the accuracy and effectiveness of the model is largely dependent upon the quality and relevance of the data utilized during the training phase. As such, careful attention must be paid to the data preparation process in order to ensure that the resulting model is as accurate and reliable as possible.

In order for a machine learning project to yield positive results, it is essential that the data utilized to train the model is of high quality and relevance. By adhering to a rigorous data preparation process, the accuracy and effectiveness of machine learning algorithms can be significantly enhanced. As such, careful attention must be given to the data preparation phase to ensure that the resulting model is as precise and reliable as possible.

The data preparation process consists of a series of distinct steps, which include Data Cleaning, Data Normalization, Feature Selection, Encoding, Sampling, and Data Splitting. Each of these steps is critical for ensuring that the data utilized during the machine learning training phase is of high quality and relevance. In the following paragraphs, we will delve into the details of each of these steps.

1. Data Cleaning

Data cleaning aims to eliminate or rectify any inaccuracies or inconsistencies in the data that could have an adverse effect on the performance of machine learning models. The data cleaning process usually involves several steps, which include:

- Detecting errors: This involves identifying mistakes such as missing values, duplicate records, outliers, and inconsistent values within the data.
- Correcting errors: This step entails correcting errors by filling in missing values, deleting duplicate records, transforming outliers, and standardizing inconsistent values.

- Verifying corrections: This step involves verifying that the errors have been accurately rectified.

By carrying out a thorough data cleaning process, it is possible to ensure that machine learning models are trained on data of high quality. This, in turn, can enhance the performance of these models and result in more precise and reliable outcomes. Thus, it is crucial to give due attention to the data cleaning step as part of the overall data preparation process.

2. Data Normalization

Data normalization is the process of transforming raw data into a standardized format to make it easier for machine learning algorithms to process. The goal of normalization is to eliminate redundant data and minimize the effects of differences in scale and variability in the data. That is: transforming the values of numeric variables so they can have similar scale, there are several methods for normalization, each with its own strengths and weaknesses, including:

- Min-Max Scaler: This method scales the values of a feature between 0 and 1. It is calculated by subtracting the minimum value of the feature from each value, and then dividing the result by the range of the feature (i.e., the difference between the minimum and maximum values).
- Standard Scaler: This method transforms the values of a feature so that they have a mean of zero and a standard deviation of one. This is achieved by subtracting the mean of the feature from each value, and then dividing the result by the standard deviation of the feature.
- Z-Score Normalization: This method is similar to the standard scaler but it standardizes the values of a feature by subtracting the mean and dividing by the standard deviation. This produces a score known as the Z-score, which represents the number of standard deviations that a value lies from the mean.
- Logarithmic Transformation: This method is used to transform features that have a skewed distribution. For example, logarithmic transformation can be applied to features that have a large range of values, such as income or population data. The logarithmic transformation helps to reduce the impact of extreme values and makes the distribution more symmetrical.

It's important to note that normalization should only be performed on the training data, as normalizing the test data could lead to information leakage, which would compromise the validity of the results.

3. Feature Selection

In most cases, when data is gathered, it is difficult to determine which features will be useful and which ones will be irrelevant or contain noise. To address this issue, feature selection is employed. This is the process of selecting a subset of the most important and relevant features from a large set of features within a dataset, for use in constructing a machine learning model. The purpose of feature selection is to eliminate features that are either redundant, irrelevant, or noisy, and to retain only those that significantly contribute to the predictive accuracy of the model.

There are two main approaches to feature selection in machine learning: filter methods and wrapper methods. Filter methods use statistical measures to evaluate the importance of each feature, such as correlation with the target variable, information gain, and chi-squared test. Wrapper methods, on the other hand, evaluate the performance of a model on a subset of features, and use this information to guide the selection process.

Benefits of feature selection include reduced computational time, improved interpretability of the model, reduced overfitting, and improved generalization of the model to new data.

It is important to note that feature selection is not a one-time process, but is often an iterative process that requires testing and evaluating different combinations of features. The final set of features selected will depend on the specific problem and dataset, as well as the machine learning algorithm used.

4. Encoding

Encoding in machine learning refers to the process of converting categorical data into numerical values that can be used as input features for a model. The purpose of encoding is to allow the machine learning algorithm to understand and work with non-numerical data.

There are several types of encoding, including:

- One-Hot Encoding: This type of encoding creates a binary representation of each category, where each feature represents a single category and is either 1 or 0 depending on the presence or absence of that category.
- Label Encoding: This encoding assigns a unique integer value to each category in the dataset. It is the simplest form of encoding, but it may result in bias if the categories are not ordered logically.
- Ordinal Encoding: This encoding assigns a numerical value to each category based on the order of the categories. For example, if the categories are low, medium, and high, low would be assigned the value 1, medium would be assigned 2, and high would be assigned 3.
- Binary Encoding: This encoding represents categories as binary digits. Each digit represents a particular level in a categorical hierarchy.

The type of encoding used depends on the type of data and the machine learning algorithm being used. Proper encoding is important to ensure that the machine learning model accurately captures the relationship between the variables and correctly predicts the target variable.

5. Sampling

Sampling is a technique used to address class imbalance in machine learning, which occurs when one class in a dataset is over-represented compared to another class. This can result in a model that is biased towards the majority class, leading to poor performance on the minority class.

There are two types of sampling techniques that can be used to address class imbalance: oversampling and undersampling.

Oversampling involves duplicating the minority class samples in the dataset to make it more balanced. This can lead to overfitting, as the model is exposed to the same samples repeatedly.

Undersampling involves reducing the number of samples from the majority class to balance the class distribution. This can result in a loss of information and decreased accuracy.

In order to effectively address class imbalance, it is important to use the appropriate sampling technique based on the specific problem and dataset. Combining both oversampling and undersampling can also be an effective strategy, creating a balanced dataset without losing too much information.

6. Data splitting

Data splitting is where the original dataset is divided into smaller subsets for different purposes, such as training, validation, and testing. The purpose of data splitting is to provide an unbiased evaluation of the model's performance, as well as to prevent overfitting, which occurs when a model learns the training data too well and fails to generalize to new, unseen data.

Here are the most common data splitting techniques in machine learning:

- Train/Test Split: The dataset is divided into two parts, the training set and the testing set. The model is trained on the training set and then evaluated on the testing set. The purpose of this split is to get a rough estimate of how well the model will perform on new, unseen data.
- k-Fold Cross Validation: In this technique, the dataset is divided into k equally sized "folds". The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, each time with a different fold being used as the test set. The average performance across all k trials is used as the final performance measure.
- Stratified k-Fold Cross Validation: This is similar to k-Fold Cross Validation, but the data is divided into folds in a way that each fold contains roughly the same proportion of samples from each class, in case the target variable is categorical. This helps to reduce the chances of biasing the model's performance evaluation by having an imbalanced distribution of classes in the folds.
- Train/Validation/Test Split: In this split, the original dataset is divided into three parts: the training set, the validation set, and the testing set. The model is trained on the training set and then evaluated on the validation set. The model's hyperparameters can be adjusted based on the results obtained on

the validation set. Finally, the model's performance is evaluated on the testing set.

The choice of data splitting technique depends on the size of the dataset, the purpose of the evaluation, and the complexity of the model.

D. Algorithms

1. Logistic Regression

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

Logistic Regression works by using an equation as a model, which is then trained on existing data so that the model can make a prediction on new data. The equation used in Logistic Regression is known as the logistic function, which is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1.

Logistic Regression is a widely used method for classification problems, especially when the dependent variable is binary. It is also used in situations where the independent variables are not normally distributed or when there are non-linear relationships between the independent variables and the binary outcome.

In summary, Logistic Regression is a powerful tool for binary classification that allows us to model the relationship between multiple independent variables and a binary outcome.

2. Support Vector Machine

Support Vector Machine (SVM) is a popular machine learning algorithm used for classification and regression analysis. It was introduced by Vladimir Vapnik in 1963. SVM is a supervised learning algorithm that is widely used for both linear and non-linear classification problems.

SVM works by finding a hyperplane that best separates the data into classes. The hyperplane is chosen such that it maximizes the margin, which is the distance between the

hyperplane and the closest data points, called support vectors. These support vectors are critical to the position of the hyperplane and determine the maximum margin. The data points closest to the hyperplane are referred to as the support vectors, and they have the greatest impact on the position of the hyperplane.

SVM is a machine learning algorithm that follows the principles of structure risk minimization (SRM) to minimize an upper bound of generalization error, as opposed to minimizing empirical error like other neural networks (Vapnik, 1999). It achieves this by using a kernel to map the input data to a higher-dimensional space, where the data points can be linearly separated in classification problems, or by finding an optimal hyperplane that optimizes the distance between two datasets in regression problems.

Figure 2 depicts the optimal hyperplane in SVM that effectively separates two datasets, and the data points close to the hyperplane are referred to as Support Vectors (SVs). The accuracy of an SVM model is greatly influenced by the choice of kernel parameters, as these parameters significantly impact the performance of the kernel method (Wang et al, 2003). The number of these parameters is determined by the margin that separates the datasets.

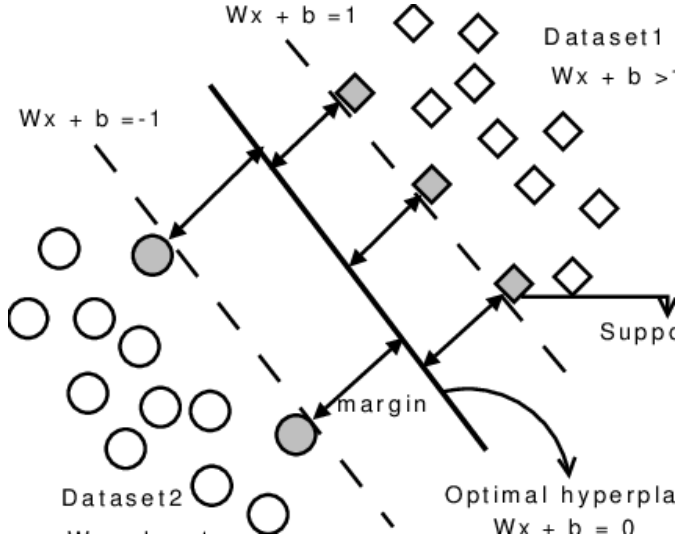

Figure. 2 Optimal hyperplane in support vector machine (Rodan et al 2014).

When training a Support Vector Machine (SVM), it involves solving a quadratic optimization problem with linear constraints. The complexity of the solution in SVM is determined by the complexity of the desired solution, rather than the dimensionality of the

input space. In other words, the computational complexity of SVM is not solely dependent on the number of input features, but rather on the complexity of the underlying problem being addressed (Rodan et al, 2014).

One of the key strengths of SVM is its ability to handle non-linearly separable data through the use of kernels. A kernel is a function that transforms the input data into a higher dimensional space, where it becomes linearly separable. The most commonly used kernels are the radial basis function kernel and the polynomial kernel.

In conclusion, SVM is a powerful machine learning algorithm that is widely used for classification and regression problems. Its ability to handle non-linearly separable data through the use of kernels, and its robustness make it a popular choice for many applications.

3. Random Forest

Random Forest is a popular machine learning algorithm that was introduced by Leo Breiman in 2001 (Breiman, 2001) It is an ensemble learning method that creates multiple decision trees and combines their predictions to form a final result. Random Forest is known for its ability to handle high dimensional data, missing values, and outliers, making it a versatile algorithm for different types of data and applications.

Random Forest works by selecting a random subset of features from the dataset to create each decision tree. This random feature selection ensures that each tree in the forest is different, reducing overfitting and improving the accuracy of the model (Cutler et al, 2004). The algorithm then trains each decision tree with a different sample of the data. Finally, the prediction of the forest is obtained by combining the predictions of all the trees, usually through a majority voting scheme (Escolano et al, 2009).

Random Forest is widely used in classification, regression, and feature selection tasks. It has been applied to many real-world problems such as financial forecasting, image classification, and fraud detection (Kelleher et al, 2015). One of the strengths of Random Forest is its interpretability, which is essential in applications where the results need to be understandable by non-technical stakeholders.

In conclusion, Random Forest is a powerful machine learning algorithm that has proven its effectiveness in various applications. Its combination of accuracy and interpretability makes it an ideal choice for many data-driven projects.

4. Gradient Boosting

Gradient Boosting is a machine learning technique that belongs to the family of ensemble methods and is used for classification and regression tasks. It combines multiple weak learners (i.e., models that are only slightly better than random guessing) to create a single strong model with improved accuracy.

The idea behind gradient boosting is to iteratively train weak models and add them together to form a stronger model. Each weak model is trained to correct the mistakes made by the previous model. The training process involves computing the gradient of the loss function with respect to the predicted outputs and updating the model parameters in the direction of the negative gradient. This process is repeated until the desired number of weak models are combined or until no further improvement can be made.

One of the most popular implementations of gradient boosting is called Gradient Boosted Regression Trees , which uses decision trees as weak learners. Gradient Boosted Regression Trees builds trees one-at-a-time, where each new tree is fit to the negative gradient of the loss function with respect to the current ensemble prediction.

The gradient boosting algorithm has several key hyperparameters, including the number of weak learners to use, the learning rate (or shrinkage), and the maximum depth of the decision trees. The learning rate controls the contribution of each weak learner to the final model, and the maximum depth controls the complexity of each individual decision tree.

Gradient Boosting has been successfully applied to a wide range of real-world problems and has been found to be a powerful and flexible machine learning technique.

5. eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting algorithm that is designed to handle large datasets and computationally expensive problems. It was created by Tianqi Chen in 2014 and has since become one of the most widely used machine learning algorithms (Chen, 2016).

Gradient Boosting is a popular machine learning technique that combines multiple weak models to create a strong prediction model. In XGBoost, the algorithm uses decision trees as base models and fits the model iteratively, in which each tree is built upon the

residuals of the previous tree. This approach allows for the prediction model to continuously improve and make more accurate predictions.

One of the key features of XGBoost is its ability to handle sparse data and missing values. This is achieved through a technique known as "column subsampling", in which XGBoost only uses a random subset of columns during each iteration of the model-building process. This reduces overfitting and leads to improved accuracy.

XGBoost also uses an advanced regularization technique known as "regularized gradient boosting", which helps to reduce overfitting and improve the generalization of the model. This technique adds a penalty term to the objective function, which helps to prevent the model from becoming too complex and reduces the risk of overfitting.

Another feature of XGBoost is its parallel processing capabilities, which allow the algorithm to take advantage of multiple cores and GPUs. This makes XGBoost much faster than traditional gradient boosting algorithms, and enables it to handle large datasets and computationally expensive problems.

In conclusion, XGBoost is an advanced implementation of gradient boosting that is designed to handle large datasets and computationally expensive problems. Its ability to handle sparse data, missing values, and overfitting, combined with its parallel processing capabilities, make XGBoost a powerful and widely used machine learning algorithm.

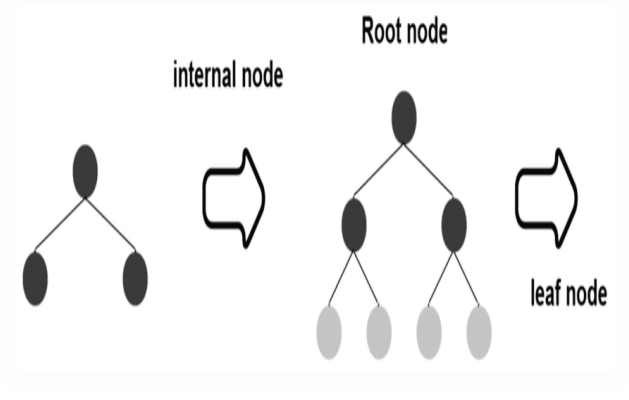

Figure 3 Xgboost Workflow (Cutler et al, 2004)

6. Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) is a machine learning algorithm that is used for gradient boosting trees. It was developed by Microsoft in 2017 to improve the speed and efficiency of gradient boosting algorithms.

LightGBM uses a novel approach to handle large datasets by implementing a histogram-based optimization algorithm. The algorithm builds a histogram of the feature values for each data sample and uses this histogram to determine the optimal split points for each feature. This approach allows LightGBM to perform much faster than traditional gradient boosting algorithms and handle larger datasets.

LightGBM also uses a technique called leaf-wise tree growth, which grows the decision tree in a leaf-wise manner rather than a level-wise manner. This results in a more accurate prediction and lower overfitting compared to level-wise tree growth algorithms.

LightGBM has been found to be particularly useful for large datasets and for binary classification problems. In comparison to other gradient boosting algorithms such as XGBoost, LightGBM has been shown to be faster and more efficient in terms of memory usage (Ke et al, 2017).

In conclusion, LightGBM is a powerful and efficient machine learning algorithm that has been shown to be effective in solving large and complex problems. It is a valuable tool for data scientists and machine learning practitioners.

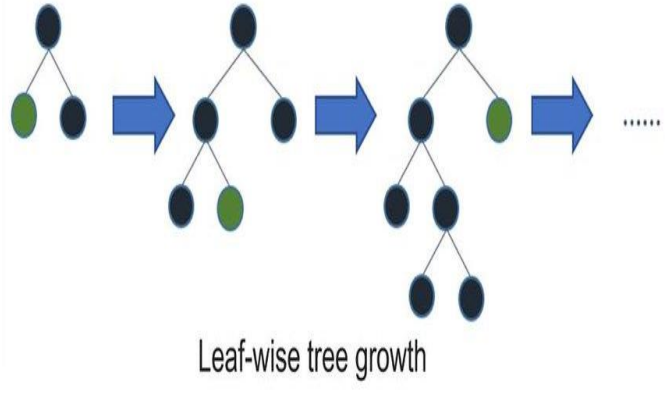

Figure 4 Leaf-wise tree growth (Yang et al, 2018).

E. Evaluation

Model evaluation is a crucial step in the machine learning process that helps assess the quality and performance of a model in solving a given problem. It involves the use of

various metrics and techniques to quantify the accuracy and effectiveness of the model in making predictions or classifications.

There are several aspects of model evaluation, including:

1. **Accuracy**: This is one of the most common metrics used to evaluate models. It measures the percentage of correct predictions made by the model compared to the total number of predictions.

2. **Confusion Matrix**: A confusion matrix is a table that summarizes the performance of a model by comparing the predicted and actual results. It helps identify false positives, false negatives, true positives, and true negatives.

3. **Precision and Recall**: Precision measures the proportion of positive predictions that are actually correct, while recall measures the proportion of positive cases that the model was able to identify.

4. **F1 Score**: The F1 score is the harmonic mean of precision and recall, and is a good metric to use when there is an uneven class distribution.

5. **ROC Curve**: The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate, and is used to evaluate the performance of binary classification models.

6. **Cross-Validation**: Cross-validation is a technique used to evaluate the performance of a model by dividing the dataset into multiple subsets and using each subset for testing and training the model.

In summary, model evaluation is an essential step in machine learning that helps assess the quality and performance of models, and helps identify areas for improvement.

IV. EXPERIMENT

A. Tools and libraries

In this research, a variety of tools and libraries were utilized for the machine learning experiments. Anaconda, a widely adopted data science platform, was used for managing Python environments and packages. Jupyter Notebook, a web-based interactive development environment, was employed for creating, documenting, and sharing the research code, and visualizations. Python, a popular programming language in the field of machine learning, was utilized for implementing the machine learning algorithms and data analysis tasks.

Specifically, Anaconda distribution version 2022.01, Jupyter Notebook version 7.10.0, and Python version 3.8.6 were used. The extensive libraries available in Python, such as NumPy, Pandas, and Scikit-Learn, and seaborn were leveraged for data manipulation, analysis, visualization, and the machine learning tasks. These libraries provided tools for preprocessing, algorithms for ML, and ways to plotting of the data.

In this chapter, we will describe the process of applying machine learning techniques in the study and provide a detailed explanation of the preprocessing steps that were carried out. The purpose of this chapter is to give a comprehensive overview of the methods used to ensure that the study's results are accurate and dependable.

1. Data Understanding

Table 1. Attributes and first five rows of data used

B	C	D	E	F	G	H	I	J	K	L	M	N
CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0

The dataset utilized for this study was created through the utilization of the Kaggle churns model. The target parameter of this dataset is a binary variable which is utilized to

indicate whether a customer has remained with the bank or has left. The dataset consists of data collected from 10,000 bank customers, out of which 7963 are categorized as positive samples, indicating customers that remained with the bank, while 2037 are categorized as negative samples, indicating customers that left the bank. The target variable is represented in binary form, with a value of 1 indicating a churned client and 0 indicating a retained client. Additionally, there are thirteen (13) attributes included in the dataset, which were derived from customer data and transactions conducted by the customer, and their meanings will be elaborated below:

Attributes:

- CustomerId — The CustomerId column contains random values and has no bearing on whether a customer will leave the bank, so it will be removed.
- Surname — the Surname column, which refers to a customer's last name, is not a relevant factor in determining churn and will be removed as well.
- CreditScore — The CreditScore column is a potentially important predictor, as customers with higher scores are less likely to leave the bank.
- Geography — The Geography column, which reflects a customer's location, is also important to retain, as it may influence their decision to stay or leave.
- Gender — The Gender column will be included as well, as it's interesting to explore whether gender plays a role in customer churn.
- Age — The Age column is definitely significant because elderly clients are more inclined to remain with their bank compared to younger ones.
- Tenure — The tenure column pertains to the length of time a customer has been a patron of the bank. Generally, senior clients exhibit greater loyalty and are less prone to switching banks.
- Balance — The balance column a strong predictor of customer turnover is also the account balance, given that individuals with larger balances in their accounts are less inclined to switch to another bank in comparison to those with smaller balances.
- NumOfProducts — The NumOfProducts column, which represents the number of products a customer has purchased through the bank, is another important factor in determining churn.

- HasCrCard — This column indicates if a customer possesses a credit card. It is likewise significant since clients who have a credit card are more loyal and are less likely to switch banks. (0=No, 1=Yes)
- IsActiveMember — We will retain this column as it indicates whether a customer is active or not, and active customers have a lower tendency to switch banks. (0=No, 1=Yes)
- EstimatedSalary — Similar to the account balance, individuals with lower salaries have a higher probability of switching banks than those with higher salaries.
- Exited — The target variable is whether or not the customer has churned or left the bank, which is what we need to forecast. (0=No, 1=Yes).

Table 2. Statistics of Data Used

	count	mean	std	min	25%	50%	75%	max
CreditScore	10000.0	650.528800	96.653299	350.00	584.00	652.000	718.0000	850.00
Geography	10000.0	0.501400	0.500023	0.00	0.00	1.000	1.0000	1.00
Gender	10000.0	0.545700	0.497932	0.00	0.00	1.000	1.0000	1.00
Age	10000.0	38.921800	10.487806	18.00	32.00	37.000	44.0000	92.00
Tenure	10000.0	5.012800	2.892174	0.00	3.00	5.000	7.0000	10.00
Balance	10000.0	76485.889288	62397.405202	0.00	0.00	97198.540	127644.2400	250898.09
NumOfProducts	10000.0	1.530200	0.581654	1.00	1.00	1.000	2.0000	4.00
HasCrCard	10000.0	0.705500	0.455840	0.00	0.00	1.000	1.0000	1.00
IsActiveMember	10000.0	0.515100	0.499797	0.00	0.00	1.000	1.0000	1.00
EstimatedSalary	10000.0	100090.239881	57510.492818	11.58	51002.11	100193.915	149388.2475	199992.48
Exited	10000.0	0.203700	0.402769	0.00	0.00	0.000	0.0000	1.00

2. Feature Selection

This study employed Pearson correlation coefficient plots to examine the relationships between variables, as illustrated in the figure 5. Notably, feature selection was incorporated into four algorithms used in the research, namely Random Forest, Gradient Boosting, Xtreme Gradient Boosting, and Light Gradient Boosting, to identify the most significant features within the dataset.

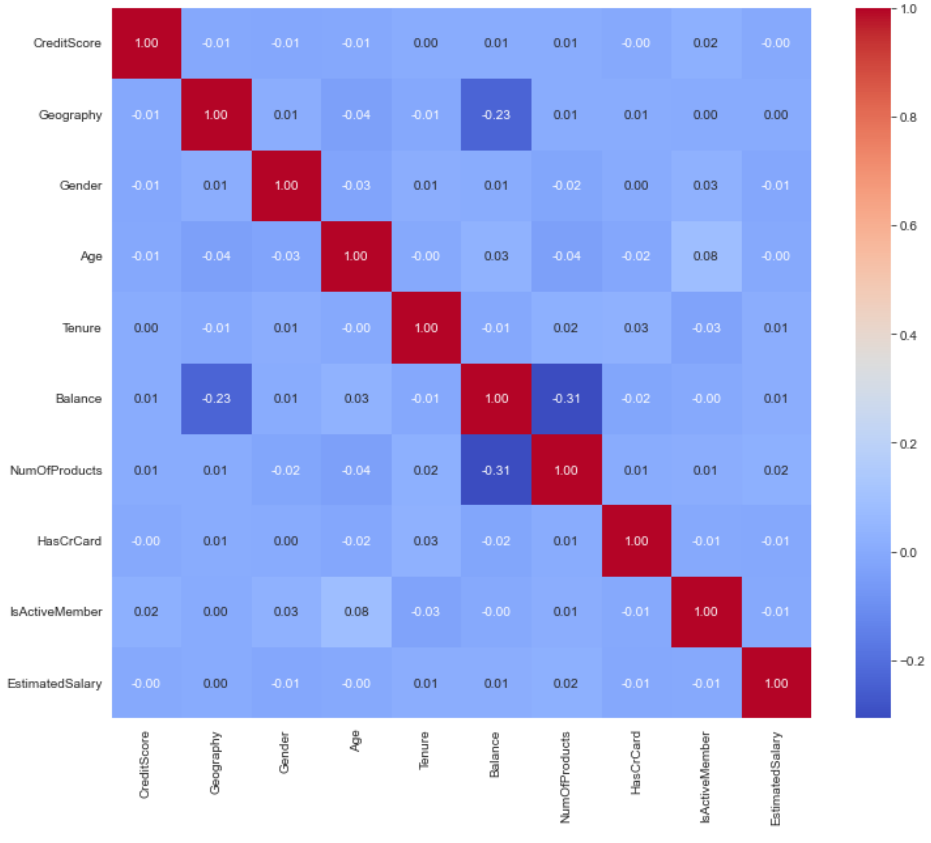

Figure 5. Heat Map of the data

The CustomerId and Surname features have been excluded from the feature selection process as they are unique to each customer. Therefore, they do not provide any meaningful information that could improve the model's performance. As a result, these features have been dropped from the dataset.

3. Encoding

In order to create machine learning models, it is typically necessary for all input and output features to be numeric. This means that category features must be transformed or encoded into integers before models can be created. In the particular dataset being used, there are two features that need to be encoded: gender and geography.

To encode the gender feature, label encoding was used. This method assigns each label a distinct integer value based on alphabetical order. In this case, the male gender was represented by the integer 1 and the female gender by the integer 0.

For the geography feature, a manual mapping approach was used. The values were mapped so that customers in France were assigned a value of 1, while all other customers in Spain and Germany were assigned a value of 0. This approach was adopted because the populations of Spain and Germany are nearly equal and much lower than that of France, as depicted in Figure 6. Therefore, it is reasonable to encode this functionality in a way that distinguishes between clients who are French and those who are not.

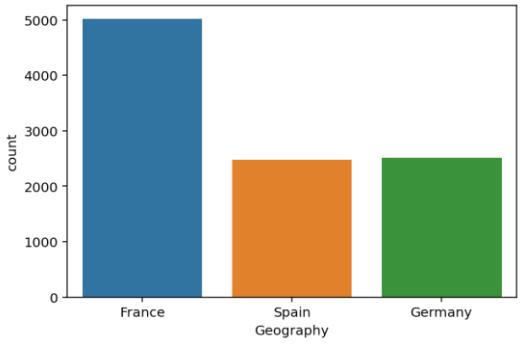

Figure 6. Count of Countries in data

4. Addressing Class Imbalance

When dealing with machine learning tasks, particularly in the case of predicting customer churn, the unequal distribution of classes can pose a problem (Liermann, 2021). This is due to the fact that there is usually more data available about customers who have remained with a company, as opposed to those who have left. As shown in Figure 5, the data is imbalanced, with one class (0 - retained) being significantly more prevalent than the other (1-churned). This uneven distribution of data can lead to biased categorization, which must be addressed. One approach to resolving class imbalance is to use data-level strategies such as oversampling and undersampling, which are discussed in chapter 3 of this research.

However, undersampling may result in incomplete data and underfitting, so oversampling is more commonly used in practice (Peter et al 2020). Therefore, the Synthetic Minority Oversampling Technique (SMOTE) is employed in this investigation to address the class imbalance issue.

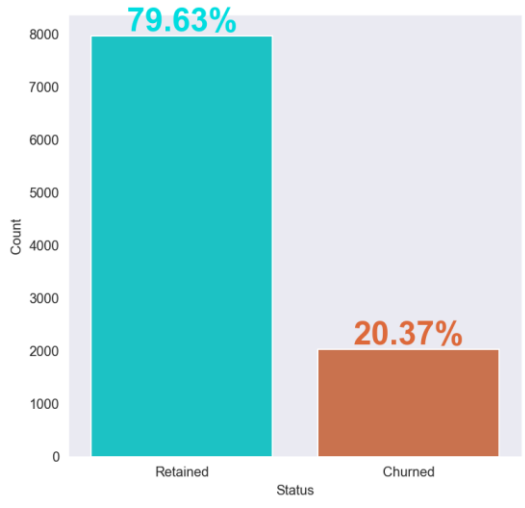

Figure 7. Target variable distribution

5. Smote

The SMOTE technique is used to increase the number of records in a dataset belonging to a minority class. It locates a comparable record to the one that needs to be up-sampled and creates a synthetic record by generating a weighted average of the original record and its nearby record. To create new synthetic examples, SMOTE selects a sample from the minority class and then identifies its k nearest neighbors (usually k=5). It then creates new synthetic examples by interpolating between the selected sample and each of its k nearest neighbors.

Compared to replicating existing minority class samples, this approach is more effective because it generates new data points that more accurately represent the underlying distribution of the minority class. SMOTE is compatible with various machine learning algorithms and has been proven to enhance performance on a wide range of classification tasks.

6. Classification

During this phase of the study, a variety of classification methods were implemented in order to make predictions about customer churn based on the pre-processed data. The primary objective of the investigation was to compare different supervised machine learning classification techniques and determine which one was the most effective. A range of classification methods were utilized in order to make predictions about customer churn, including Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting, Xtreme Gradient Boosting, and Light Gradient Boosting.

B. Results

1. Logistic Regression

According to Zhang et al, Kadam et al, and Alashwal et al, logistic regression is a highly suitable statistical technique for predicting customer churn. (Zhang et al, 2019), (Kadam et al, 2016), (Alashwal et al, 2017). This is because customer churn prediction involves analyzing the relationships between multiple independent variables and a dependent variable. Factors such as demographic information, behavioral patterns, and transactional data can all contribute to whether or not a customer is likely to churn. Logistic regression allows for the examination of these various factors and their relationship to customer churn.

The use of logistic regression in customer churn prediction also provides results that are easy to interpret. This is because the technique produces coefficients for each independent variable, which represent the strength and direction of the relationship with the dependent variable (i.e., the odds ratio). As a result, the results of the analysis can be easily understood and interpreted.

The effectiveness of logistic regression in predicting customer churn has been demonstrated in several studies. Therefore, it is considered a reliable method for predicting customer churn in various industries.

After applying this classification, the classification report is given below:

Table 3. Results from Logistic regression

	precision	recall	f1-score	support
0	0.80	0.77	0.78	2426
1	0.77	0.80	0.79	2352
accuracy			0.79	4778
macro avg	0.79	0.79	0.79	4778
weighted avg	0.79	0.79	0.79	4778

2. Support Vector Machine

The decision to use the support vector machine algorithm in this research was made based on its ability to effectively handle noise and outliers, as well as its robustness in dealing with high-dimensional data. One of the key features of the SVM is its margin-

based approach which maximizes the distance between the decision boundary and the closest data points, resulting in a more reliable prediction model.

In customer churn prediction, there are typically numerous predictor variables that can be utilized to forecast the target variable. However, SVM can effectively deal with high-dimensional data by selecting the most relevant features while ignoring the irrelevant ones, which can lead to improved accuracy in the model. Additionally, the SVM algorithm can handle non-linear relationships using kernel functions that map the input data into a higher-dimensional space where linear separation is possible.

Customer churn prediction is a complex issue that frequently involves non-linear relationships between the predictor variables and the target variable. When the SVM algorithm was applied to our data, it was found to be an effective solution for dealing with noise and outliers, as well as accurately predicting customer churn

The result is displayed below:

Table 4. Results from Support vector machine

	precision	recall	f1-score	support
0	0.84	0.83	0.83	2426
1	0.82	0.83	0.83	2352
accuracy			0.83	4778
macro avg	0.83	0.83	0.83	4778
weighted avg	0.83	0.83	0.83	4778

3. Random Forest

The Random Forest algorithm is a highly versatile and potent machine learning technique that has proven to be well-suited for predicting customer churn. It possesses a range of desirable attributes, such as the ability to process vast quantities of data, reduce overfitting, cope with missing information, offer insights into feature importance, and handle nonlinear relationships between variables, which make it an excellent choice for this particular task. Its accuracy and interpretability have made it a popular method for predicting customer churn. Additionally, due to its capability to handle intricate interactions between various factors, Random Forest is a valuable choice for predicting customer churn, where there may be complex, non-linear relationships between different factors.

The results for the application of this algorithm on our data is seen thus;

Table 5. Results from Random Forest

	precision	recall	f1-score	support
0	0.86	0.85	0.85	2426
1	0.85	0.86	0.85	2352
accuracy			0.85	4778
macro avg	0.85	0.85	0.85	4778
weighted avg	0.85	0.85	0.85	4778

4. Gradient Boosting

Customer churn prediction is a complex task that requires identifying non-linear relationships between predictors and the target variable. This can be challenging because traditional models, such as linear regression, may not be able to capture these relationships. However, gradient boosting is a powerful machine learning technique that can overcome this limitation.

One of the key strengths of gradient boosting is its ability to use decision trees to capture non-linear interactions between predictors. This means that it can identify complex relationships between variables that may not be evident from looking at individual predictors alone. By doing so, it can create a more accurate and nuanced prediction model.

In addition to handling non-linear relationships, gradient boosting is also well-suited to address other challenges associated with customer churn prediction. These include missing values, outliers, high-dimensional data, and class imbalance. By being able to handle these issues, gradient boosting can create a more robust and reliable prediction model.

Overall, the application of gradient boosting to customer churn prediction can result in a more accurate and effective prediction model that can help businesses identify and address potential customer churn before it happens.

Its result after application goes thus;

Table 6. Results from Gradient boosting

	precision	recall	f1-score	support
0	0.83	0.82	0.83	2426
1	0.82	0.83	0.82	2352
accuracy			0.83	4778
macro avg	0.83	0.83	0.83	4778
weighted avg	0.83	0.83	0.83	4778

5. Xtreme Gradient Boosting

XGBoost is an algorithm that has proven to be powerful in handling large and complex datasets. It has the ability to prevent overfitting and provide accurate predictions, making it an ideal tool for businesses to use in predicting customer churn. The algorithm employs a tree ensemble method that combines the predictions of multiple decision trees. This approach is beneficial because it improves the accuracy of the model and reduces its sensitivity to noise in the data. Moreover, XGBoost also utilizes gradient boosting, which is a technique that iteratively adjusts the weights of the training samples to focus on the harder-to-predict cases. By doing so, it further enhances the performance of the model.

One of the key advantages of XGBoost is its feature importance analysis, which provides a measure of the most influential factors that contribute to customer churn. This information is vital for businesses as it helps them prioritize which factors to address when attempting to reduce churn rates. By identifying the critical factors, businesses can take targeted action to retain their customers and minimize attrition. Overall, XGBoost's advanced methodology and feature importance analysis make it a valuable tool for businesses looking to predict and prevent customer churn.

Table 7. Results from Xgboost

	precision	recall	f1-score	support
0	0.86	0.85	0.85	2426
1	0.85	0.85	0.85	2352
accuracy			0.85	4778
macro avg	0.85	0.85	0.85	4778
weighted avg	0.85	0.85	0.85	4778

6. Light Gradient Boosting Machine

Light gradient boosting is an effective method for predicting customer churn because it utilizes an ensemble learning approach that combines multiple weak models into a stronger model. This results in improved accuracy and robustness of the model as it reduces both bias and variance. When used for churn prediction, this means that light gradient boosting is able to capture the intricate relationships between customer attributes and churn, enabling accurate predictions based on these relationships.

Furthermore, light gradient boosting is a fast and scalable algorithm that can be easily parallelized and distributed. This makes it a suitable choice for large-scale churn prediction applications, where real-time analysis of millions of customers may be required. Additionally, light gradient boosting can be trained and updated incrementally, allowing it to adapt to changes in customer behavior over time. All of these features make light gradient boosting a powerful tool for predicting customer churn in a wide range of business contexts.

Table 8. Results from Lgbm

	precision	recall	f1-score	support
0	0.86	0.84	0.85	2426
1	0.84	0.86	0.85	2352
accuracy			0.85	4778
macro avg	0.85	0.85	0.85	4778
weighted avg	0.85	0.85	0.85	4778

C. Evaluation

1. Classification Report

The classification report is a commonly used tool in machine learning to assess the performance of a classification model. This report provides an extensive summary of how well the model has classified the target variable by showing various metrics, such as precision, recall, F1-score, and support, for each class in the dataset.

Precision is a significant metric in the context of customer churn prediction, which evaluates the accuracy of the model's positive predictions by measuring the proportion of true positives among all positive predictions. In other words, precision quantifies the model's ability to correctly identify customers who are likely to churn out of all the customers predicted to churn. It can also be defined as the ratio of true positives to the sum of true and false positives.

Recall, also known as sensitivity, measures the model's ability to identify all positive cases or customers who are likely to churn by calculating the ratio of true positives to the sum of true positives and false negatives. It is focused on identifying all customers who are at risk of churning, providing a crucial evaluation metric for the model's performance.

F1 Score is an essential evaluation metric in customer churn prediction because it captures the model's performance in identifying both actual and predicted churn cases. It is a comprehensive metric that combines precision and recall to provide a balanced evaluation of the model's performance. It ranges from 0 to 1, with a score of 1 indicating perfect precision and recall, and a score of 0 indicating poor performance. F1-score is also sensitive to imbalanced datasets, which is common in customer churn prediction.

The support metric in a classification report represents the number of samples in the test set that belong to each class. It shows how many instances of each class were correctly classified by the model, indicating the number of customers who have churned and the number who have not. Support is an important metric in a classification report because it helps us understand the balance of classes in the test set and the model's ability to classify each class. A low support for a particular class can make it more challenging for the model to learn to classify that class accurately, while a high support can make it easier for the model to learn to classify that class accurately.

V. DISCUSSION AND CONCLUSION

The main aim of this thesis was to investigate the use of selected Supervised Machine Learning techniques for predicting customer churn on the bank customer data.

The first chapter provides an overview of the customer churn phenomenon in the banking industry and outlines the problem definition and research question addressed in this study.

In the second chapter, various literature sources related to the research topic were reviewed and analyzed to gain a better understanding of the objectives and goals of the research.

The third chapter delves into an extensive discussion on the design and methodology used in machine learning studies, similar to the approach employed in this research.

Finally, the fourth chapter describes the implementation and evaluation of the experiment conducted in this study.

A. Experimental Analysis

To determine the most effective machine learning classifier for predicting customer churn in a bank, an experiment was conducted using the bank customer churn dataset. Initially, the experiment utilized imbalanced data, and the results are presented in table 9 . These results showed a low recall value (<0.5), indicating that the classifiers had a significant number of False Negatives. To improve the accuracy of the experiment, a balanced dataset was employed, and the SMOTE sampling technique was utilized. The final dataset was divided into two sets: training (80%) and testing (20%), and neither dataset included the goal variable "Exited."

Table 9. Comparison Results before data is balanced

	Accuracy	Recall	Precision	f1_score
LR	0.810000	0.176904	0.615385	0.274809
SVC	0.858500	0.378378	0.836957	0.521151
RF	0.854000	0.444717	0.732794	0.553517
GB	0.856500	0.457002	0.738095	0.564492
XGB	0.841500	0.447174	0.664234	0.534508
LGBM	0.854000	0.457002	0.723735	0.560241

According to the findings presented in Table 10, it has been observed that the Random Forest machine learning technique has outperformed various other popular methods such as Logistic Regression, Support Vector Machine, Gradient Boosting, Xtreme Gradient Boosting, and Light Gradient Boosting in terms of accuracy, recall, and precision, across all the evaluation criteria employed in the experiment. This outcome further reinforces the conclusion reached in previous studies that Random Forests are currently the most precise and reliable machine learning approach in the industry.

Table 10. Comparison Results when data is balanced

	Accuracy	Recall	Precision	f1_score
LR	0.785475	0.801446	0.771592	0.786236
SVC	0.829636	0.832483	0.823381	0.827907
RF	0.853077	0.857568	0.846057	0.851774
GB	0.825869	0.827381	0.820405	0.823878
XGB	0.850147	0.851616	0.845148	0.848369
LGBM	0.850356	0.856718	0.842039	0.849315

The study aimed to identify the best classifier for imbalanced and balanced data by comparing several models, including Logistic Regression, Support Vector Machine, Gradient Boosting, Xtreme Gradient Boosting, and Light Gradient Boosting. Through analysis, it was found that the Random Forest classifier with an accuracy, recall, and f1 score of 85% was the best model for the dataset. The study also found that when the dataset was imbalanced, no classifier was best on all criteria, as shown in Table 9. However, when the dataset was balanced using SMOTE, the Random Forest classifier outperformed other classifiers, as demonstrated in Table 10. The study also concluded that SMOTE is a suitable approach for balancing data. Moreover, the most significant features in the dataset were found to be age and membership activity, which can be used by banks to improve their services and increase customer satisfaction by targeting those who are more likely to churn. These results provide valuable insights for banks to optimize their business strategies.

B. Limitations and Future Study

The quality and quantity of data are crucial for accurate predictions. However, many banks may have limited data or may not have collected the right data for churn prediction. As in this study, access to different banking data was limited.

Future research can focus on the extension of the proposed approach by incorporating unsupervised classification methods such as clustering into the solving process of customer churn problems in the field of banking.

Additionally, on a home computer, extensive grid searches and other improvements are not possible since they call for numerous iterations to discover the ideal parameters, which may have a substantial impact on the models' performance. So, it is advised that either higher-speed computers be employed in future studies or that cloud computing be used as an example.

Lastly, incorporating domain expertise, such as knowledge of banking regulations, customer behaviour, and market trends, can help to improve model accuracy and interpretability.

VI. REFERENCES

BOOKS

KELLEHER, J. D., Mac Namee, B., & D'Arcy, A. (2015). **Fundamentals of machine learning for predictive data analytics**: algorithms, worked examples, and case studies. MIT Press.

PETER Bruce, Andrew Bruce, and Peter Gedeck (2020).**Practical Statistics for Data Scientists**, 2nd Edition.

VIJAYA, J., **Computational Intelligence in Data Mining**, 2017, Volume 556 ISBN : 978-981-10-3873-0 E.

ARTICLES

ALASHWAL, A. M., Osman, I. H., & Alshamiri, M. B. (2017). Customer churn prediction using logistic regression and decision tree. **Journal of Engineering and Applied Sciences**, 12(19), 5044-5048.

AL-SWIDI, A. K., & Raza, S. (2013). Customer churn prediction in banking industry: An empirical study. **International Journal of Bank Marketing**, 31(2), 116-136.

DUDYALA Anil Kumar & V. Ravi, 2008. "Predicting credit card customer churn in banks using data mining," **International Journal of Data Analysis Techniques and Strategies**, Inderscience Enterprises Ltd, vol. 1(1), pages 4-28.

ESCOLANO, F., García, S., & Alcalá, R. (2009). An overview of random forest algorithm. Pattern recognition and image analysis, 19(2), 71-77.

FROW, P., & Payne, A. (2006). Predicting customer churn in the Australian banking industry. **Journal of Financial Services Marketing**, 11(1), 54-61.

GAO, F., Zhang, Y., & Lu, Y. (2019). A machine learning approach to predict customer churn in the Australian banking industry.

KADAM, S. B., & Sonawane, S. S. (2016). Churn prediction in telecom using logistic regression and decision tree. **International Journal of Computer Science and Information Technologies**, 7(5), 2165-2167.

KİM, J., Kim, J., & Kim, D. (2020). A machine learning approach to predicting customer churn in the South Korean banking industry.

LEJEUNE, M. A. (2001). Measuring the impact of data mining on churn management. Internet Research: **Electronic Networking Applications and Policy** , 11 (5), 375-387.

Lİ, M., Yaya, Xie., Xiu, Li., Eric, W.T., Ngai., Weiyun, Ying. (2009). Customer churn prediction using improved balanced random forests.

MCLEOD, R., & Peel, M. J. (2000). Customer satisfaction and loyalty: An investigation of their relationship in the UK financial services sector. **Journal of Financial Services Marketing**, 5(1), 40-53.

PEPPARD, J. (2000). Customer relationship management (CRM) in financial services.**European Management Journal** , 18 (3), 312-27.

RODAN, Ali & Faris, Hossam & Al-sakran, Jamal & Al-Kadi, Omar. (2014). A Support Vector Machine Approach for Churn Prediction in Telecom Industry. **International journal on information**. 17.

V.VAPNIK, "An overview of statistical learning theory," **IEEE Transactions on Neural Networks**, vol.5 ,pp.988–999,1999.

W.J.WANG, Z.B.Xu,andW.Z.Lu, "Determination of the spread parameter in the gaussian kernel for classification and regression," **Neurocomputing**,vol.55:,pp.643–663,2003.

YANG, Shenghui & Zhang, Haomin. (2018). Comparison of Several Data Mining Methods in Credit Card Default Prediction. **Intelligent Information Management Journal**. 10. 115-122. 10.4236/iim.2018.105010.

ZHANG, X., Guo, H., Li, X., & Zhang, X. (2019). Customer churn prediction for e-commerce using logistic regression and random forest. **International Journal of Data Mining and Bioinformatics**, 22(4), 358-372.

ELECTRONIC SOURCES

URL-1 "Bain & Company. (n.d.). The cost of acquiring a new customer." Retrieved from https://www.bain.com/insights/the-cost-of-acquiring-a-new-customer / , (Access Date 1 March 2022)

URL-2 "Federal Reserve Bank of Chicago. (2013). Customer churn in the U.S. banking industry". Retrieved from

https://www.chicagofed.org/publications/economic-perspectives/2013/4qtr2013/customer-churn-in-the-us-banking-industry, (Access Date: 31 January 2014).

URL-3 "Capgemini. (2015). The European customer churn landscape: How to reduce churn and increase profitability". Retrieved from https://www.capgemini.com/resource-file-access/resource/pdf/customer_churn_landscape.pdf (Access Date: 31 January 2018)

OTHER SOURCES

BRAJDIC, M., Neskovic, N., & Zivkovic, M. (2006). Decision tree-based model for predicting customer churn. Expert Systems with Applications, 30(4), 906-914.

BREIMAN, L. (2001) Random Forests. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

CHEN, T. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

CUTLER, A., & Breiman, L. (2004). Random forests. Technical report, University of California, Berkeley.

FARQUAD, M.A., Ravi, V., & Bapi, R.S. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. Appl. Soft Comput., 19, 31-40.

FREEMAN, M. (1999). The 2 customer lifecycles. Intelligent Enterprise , 2 (16), 9. Gerpott, T., Rams, W., & Schindler, A. (2001). Customer retention,loyalty, and satisfaction in the German mobile cellulartelecommunicationsmarket. Telecommunications Policy , 25, 249-269

GHANTA, Deepika & Tholeti, Guru & Krishna, Sunkara & Bano, Shahana. (2021). Machine Learning Concept-Based Prognostic Approach for Customer Churn Rate in Telecom Sector. 10.1007/978-981-33-4355-9_39.

KE, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. Y., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree.

Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3148-3156.

LIERMANN, V., Li, S. (2021). Methods of Machine Learning. In: Liermann, V., Stegmann, C. (eds) The Digital Journey of Banking and Insurance, Volume III. Palgrave Macmillan

LOVELACE, K., Japkowicz, N., & Shami, M. (2001). Using a neural network to predict customer churn. Expert Systems with Applications, 20(2), 121-127.

ZHAO, Jing; Dang, Xing-Hua (2008). [IEEE 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM) - Dalian, China (2008.10.12-2008.10.14)] 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing - Bank Customer Churn Prediction Based on Support Vector Machine: Taking a as the,Commercial Bank's VIP Customer Churn Example.

RESUME

Personal Profile

Detail-orientated Data Analyst adept at making critical decisions, managing deadlines and conducting team reviews. With expertise in analysis and quantitative problem-solving skills, dedicated to company growth and improvements.

Work Experience

Customer Service Advisor The Consumer Helpline, Swansea Jan 2023 – Present

- Offered detailed advice on Staysure travel insurance product and service benefits.
- Employed active listening and product expertise to successfully resolve inbound queries.
- Offered prompt solutions to maintain customer satisfaction.
- Adhered strictly to policies and procedures for continued company compliance.
- Recorded and processed customer data accurately.

Data Analyst Side Hustle (Nigeria) Limited April 2022 – June 2022

- Synthesized large datasets into actionable insights, elevating decision-making.
- Standardized analysis metrics and dashboards using PowerBI.
- Improved business decision-making by using Microsoft Excel to analyze customer insights.

Customer Service Rep Sokosh (Nigeria) Limited June 2016-August 2020

- Provided primary customer support to internal and external stakeholders
- Maintained customer satisfaction levels through a forward-thinking strategic focus on addressing customer queries and resolving concerns
- Offered customer advice and assistance paying attention to individual needs and wants.

Skills

- Communication – verbal, written, technical
- Problem-Solving
- Interpreting complex information
- Attention to Detail
- Computer Programming (Python & SQL)
- Analytical
- Organising – prioritising tasks
- Team working

Education

2022 Data Analysis Nano Degree: Data Analytics

Udacity

2013 Bachelor of Science: Computer Science

Al-Hikmah University Nigeria

Publications

I. Huseyinov and O. Okocha, "A Machine Learning Approach To The Prediction Of Bank Customer Churn Problem," 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/IISEC56263.2022.9998299.

Projects

- We Rate Dogs Data Analysis Using Twitter API to gather necessary Data.
- Used machine learning approach to the prediction of bank customer churn problem.
- Prosper Data Loan Exploration Using Python Programming Language.