

THE UNIVERSITY OF ECONOMICS AND LAW

FALCUTY OF INFORMATION SYSTEMS



FINAL PROJECT REPORT

DATA ANALYTICS WITH R/PYTHON

TOPIC: PREDICT CUSTOMER CHURN

USING MACHINE LEARNING IN THE BANKING

Lecturer: M.Sc Le Ba Thien

Group 7:

- | | | |
|----|------------------------|------------|
| 1. | Tran Thi Quynh Nhi | K204061441 |
| 2. | Hoang Thi Thanh Phuong | K204061445 |
| 3. | Nguyen Hoang Tinh | K204061451 |

Ho Chi Minh City, August 8, 2023

COMMITMENT

We ensure that this report “Predict Customer Churn using Machine Learning in the Banking Sector” is the original work that we researched and wrote. All the sources we have used or quoted have been indicated and acknowledged by complete references.

AUTHORS

ACKNOWLEDGMENT

We would like to express our sincere thanks to M.Sc Le Ba Thien for creating conditions for us to carry out this project as well as supporting us during the completion of the project.

Through the learning process of "Data Analytics with R/Python", we not only receive new knowledge about Data Analytics but also learn new skills and practical experiences. We really appreciate the lessons that the two teachers have taught, it will definitely be a source of motivation for us to continue to improve and develop more in the future.

During the implementation of the project, the team also encountered many difficulties and pressures in terms of time and knowledge. However, the team members tried to complete the project together, although there were still many shortcomings. We hope to receive comments and suggestions from you so that this project can be perfect.

Finally, to each member of the team, thank you for devoting time and effort on this project despite the tight demanding schedule to make this project a success.

Sincerely thank.

TABLE OF CONTENTS

COMMITMENT	ii
ACKNOWLEDGMENT	iii
TABLE OF CONTENTS	iv
LIST OF ACRONYMS.....	viii
LIST OF TABLES	ix
LIST OF FIGURES.....	x
ABSTRACT	xii
CHAPTER 1: OVERVIEW	1
1.1 Business case for the project	1
1.2 Objectives of the project.....	2
1.2.1 General Objective	2
1.2.2 Specific Objectives	2
1.3 Research Objects	2
1.4 Scope of the project	3
1.5 Value and desired outcome of the project	3
1.6 Structure of project	3
CHAPTER 2: RELATED WORD	5
2.1 Background of research	5
2.1.1 Foreign Research	5
2.1.2 Domestic Research.....	7
2.2 Gain relevant research results and model suggestions	8
CHAPTER 3: THEORETICAL BASIC	9
3.1 Data mining	9
3.2 Churn rate definition.....	9
3.3 Imbalance dataset	11

3.3.1 Definition	11
3.3.2 Examples of Imbalanced Dataset.....	11
3.4 Machine Learning.....	11
3.4.1 Logistic Regression.....	11
3.4.2 Random Forest Classifier.....	12
3.4.3 Extreme Gradient Boosting Classifier	13
3.4.4 Decision tree	13
3.4.5 KNN	14
CHAPTER 4: PROPOSED METHODOLOGY	15
CHAPTER 5: DATA EXPLORATION	17
5.1 Overview of dataset.....	17
5.1.1 Dataset.....	17
5.1.2 Handling missing data and duplicates	18
5.1.3 Numerical features about the data summary.....	19
5.2 Exploratory Data Analysis	20
5.2.1 Continuous Variables.....	21
5.2.2 Categorical Variables.....	25
5.3 Relationship of target variable.....	32
5.3.1 Age	32
5.3.2 Credit Score	34
5.3.3 Balance.....	34
5.3.4 Estimated Salary	35
5.3.5 Geography	36
5.3.6 Gender.....	37
5.3.7 Tenure	37

5.3.8 Number of Products	38
5.3.9 Card Holders	39
5.3.10 Active Members	39
5.4 Correlations Analysis	40
5.4.1 Correlations Matrix	40
5.4.2 Correlation between Credit score and EstimatedSalary	41
5.4.3 Correlation between Credit score and Age	42
5.4.4 Correlation between Credit score and Balance.....	43
5.4.5. Correlation between Age and Balance.....	44
5.4.6 Correlation between Age and EstimatedSalary	45
5.4.7 Correlation between Balance and EstimatedSalary	46
5.5 Conclusion of EDA	47
CHAPTER 6: RESULTS AND DISCUSSION	50
6.1 Data Preprocessing	50
6.1.1 Data Splitting	50
6.1.2 Drop columns.....	51
6.1.3 Feature selection	51
6.1.4 Encoding categorical Features	52
6.1.5 Scaling.....	52
6.1.6 Addressing Class Imbalance using SMOTE function	53
6.2 Machine learning modeling	54
6.2.1 Modeling	54
6.2.2 Hyperparameter Tuning	55
6.2.3 Model Finalizing	56
6.2.4 Model Evaluation.....	57

6.3 Discussion about strategy for Banking	58
6.3.1 Problem	58
6.3.2 Strategy	59
CHAPTER 7: CONCLUSION.....	63
7.1 General conclusion	63
7.2 Limited and future work.....	63
7.2.1 Limited	63
7.2.2 Future work.....	64
REFERENCES.....	65
WORK ASSIGNMENT	66
PLAGIARISM CHECKING.....	67

LIST OF ACRONYMS

No.	Acronym	Description
1	XGBoost	Extreme Gradient Boosting
2	KNN	K-Nearest Neighbors
3	EDA	Exploratory Data Analysis
4	SMOTE	Synthetic Minority Oversampling Technique
5	ROC	Receiver Operating Characteristics
6	AUC	Area Under The Curve
7	TP	True positive
8	FP	False positive
9	FN	False negative
10	TN	True negative
11	LSTM	Long Short-Term Memory
12	GRUD	Gated Recurrent Units

LIST OF TABLES

Table 5.1 The detail information of the Bank's customer dataset (authors).....	17
Table 5.2 Traning set and Test set (authors)	Error! Bookmark not defined.
Table 5.3 Feature Significance in Predicting Customer Churn	51
Table 5.4 Imbalance data.....	53
Table 5.5 Balance data.....	54
Table 5.6 The performance of different machine learning models	54
Table 5.7 Confusion matrix (actual, predicted) of these 5 models.....	55
Table 5.8 The evaluation metrics.....	56

LIST OF FIGURES

Figure 3.1 Classification of churn (Customer churn 101,2021)	10
Figure 3.2. Random Forest Classifier (Tibco)	12
Figure 4.1 Our proposed methodology (authors)	15
Figure 5.1 Missing Values by Column (authors)	19
Figure 5.2 Duplicates Values by Column (authors)	19
Figure 5.3 Numerical features by the authors.....	20
Figure 5.4 Histogram of Age	21
Figure 5.5 Histogram of Credit Score	22
Figure 5.6 Histogram of Balance.....	23
Figure 5.7 Histogram of Estimated Salary	24
Figure 5.8 Histogram of Gender	25
Figure 5.9 Histogram of Geography.....	26
Figure 5.10 Histogram of Products.....	27
Figure 5.11 Histogram of Tenure	28
Figure 5.12 Histogram of HasCrCard.....	29
Figure 5.13 Histogram of Active Member	30
Figure 5.14 Histogram of Churn rate.....	31
Figure 5.15 Correlation of Age and Churn.....	33
Figure 5.16 Correlation of Credit Score and Churn	34
Figure 5.17 Correlation of Balance and Churn.....	35
Figure 5.18 Correlation of Estimated Salary and Churn	36
Figure 5.19 Correlation of Geography and Churn.....	36
Figure 5.20 Correlation of Gender and Churn.....	37
Figure 5.21 Correlation of Tenure and Churn	37
Figure 5.22 Correlation of Products and Churn	38
Figure 5.23 Correlation of Credit Card and Churn.....	39
Figure 5.24 Correlation of Active Member and Churn	39
Figure 5.25 Correlation Matrix.....	40
Figure 5.26 Correlation between Credit score and EstimatedSalary	41
Figure 5.27 Correlation between Credit score and Age	42

Figure 5.28 Correlation between Credit score and Balance	43
Figure 5.29 Correlation between Age and Balance	44
Figure 5.30 Correlation between Age and EstimatedSalary	45
Figure 5.31 Correlation between Balance and EstimatedSalary	46
Figure 5.32 Pie chart by Age	48
Figure 5.33 Pie chart by Geography	49
Figure 5.34 Pie chart by products	49
Figure 5.35 ROC Curve and Area Under the Curve (AUC)	58

ABSTRACT

This study delves into predicting customer churn within the banking sector using advanced machine learning techniques. Customer churn prediction, an essential aspect in banking, aids in retaining a competitive edge and optimizing marketing strategies. The research employs a comprehensive banking dataset to assess the predictive power of five distinct models: Logistic Regression, Random Forest, XGBoost, Decision Trees, and K-Nearest Neighbors (KNN). Evaluation metrics such as Accuracy, Recall, Precision, and F1-Score are computed to gauge the models' performance. The study highlights Random Forest as the most accurate predictor among the evaluated models. Following Hyperparameter Tuning, the "Final Hypertuned Random Forest Classifier" achieves notable Accuracy (approximately 71.95%), F1-Score (approximately 49.41%), Recall (approximately 69.72%), and Precision (approximately 38.27%) scores. Cross-validation results further underscore the model's consistency, with an average accuracy of approximately 86.73% and a standard deviation of about 2.55%.

OVERVIEW

CHAPTER 1: OVERVIEW

In this chapter, we will introduce an overview of the topic of the final project, the goal of the project, the scope of the project, the research objects related to the topic, the results as well as the value of the project. that the group oriented after finishing the research and implementation of the project and subject.

1.1 Business case for the project

Over the past years, we have seen that many businesses have encountered the problem of customer churn because of the rapid development of data communication as well as many information sources that are reaching out to everyone, creating an increasingly fierce competition environment between businesses. According to a number of reports and studies by reputable organizations, businesses in the banking sector have recorded a higher rate of customer churn than in other industries. This situation has led to an economic slowdown and threatened the sustainability of businesses.

Currently, customer abandonment is becoming a top concern in businesses, especially in the banking sector. Retaining customers and enhancing their satisfaction is one of the key factors to ensure the growth and success of the business.

In the future, to solve the problem of customer churn, businesses need to find a variety of methods and use technology to enhance the customer experience. New technologies such as artificial intelligence, blockchain, and data-driven solutions can be applied to improve business processes and create better customer experiences. In addition, businesses need to regularly evaluate and devise strategies to retain customers and solve problems related to their services.

This is an important issue for businesses, especially in highly competitive fields such as banking, telecommunications, etc. Jobs reduce customer churn rate and retain existing customers is an important way to increase sales and profits. That's why our team decided to implement the topic "Predict Customer Churn using Machine Learning in the Banking Sector" to improve customer churn. Banking services are becoming a fiercely competitive field, so retaining existing customers as well as

OVERVIEW

attracting new ones is very important. Therefore, customer churn forecasting is a very important topic in the banking industry.

1.2 Objectives of the project

1.2.1 General Objective

The goal for the project document of the team in the customer churn forecasting project is to propose an optimal algorithm with a data set of Banks from which to help banking institutions come up with solutions. appropriate customer strategies to retain existing customers and find ways to attract new customers.

1.2.2 Specific Objectives

The specific objective of the literature topic is to build an accurate and reliable model that can maximize retention of existing customers and will be to perform in-depth analysis on the factors influencing the leaving behavior of the dataset used by the team. The team will use the Python programming language and combine with business knowledge to analyze and evaluate and offer the most practical solutions to the banking business to minimize the situation of customers abandoning their businesses, based on the results that the team analyzed.

1.3 Research Objects

Forecasting customer churn can be done using machine learning (Machine Learning) and data analytics techniques. Data analytics tools including machine learning algorithms can be used to analyze customer data such as transaction history data, customer information, product and service usage behavior, and more. thereby predicting customers will be ignored in the near future. Commonly used data analysis tools in customer churn forecasting include: Machine Learning Algorithms: Decision Trees, Random Forests, Logistic Regression, Support Vector Machines, Neural Networks, etc.

Other analytical methods: Relational analysis (correlation analysis), factor analysis (factor analysis), time series analysis (time series analysis), etc. When applying data analysis tools to forecast customer abandonment, banks need to pay attention to customer-related factors such as age, gender, income, formula for using products and

OVERVIEW

services, etc. thereby identifying appropriate strategies to retain customers and enhance customer satisfaction.

1.4 Scope of the project

For the scope of the team project, the focus will be on applying machine learning techniques to predict the likelihood that a customer will switch to a competitor's service or simply switch to another service. In addition, the team will also need to evaluate the costs and potential benefits of retaining customers versus acquiring new ones. These tools include data analysis tools, machine learning libraries, and data visualization tools. Machine learning models will be trained on large data sets to ensure that they can accurately predict the likelihood of customers switching services. Models will be evaluated and fine-tuned for best performance.

1.5 Value and desired outcome of the project

In this report, the team intends to provide companies with a comprehensive answer to the problem of reversal prediction so that they can take preventive action and provide appropriate service to customers without needing the next target item. In this study, the developer leveraged a variety of techniques to more accurately predict whether or not the customer is about to continue using the service.

These models assist companies in prioritizing the best regulations that have a significant influence on consumer choice. Helping companies make informed choices to retain customers

1.6 Structure of project

The group's report is divided into 6 monitoring content programs.

- Chapter 1 – Overview: Understanding the importance of customer churn. Before discussing predictions, it's crucial to grasp how churn affects businesses. This section explores the implications of customer churn, including financial consequences, reputation, and the costs of acquiring new customers..
- Chapter 2 - Work-related: Reviewing existing literature on predicting customer churn. A thorough analysis of previous studies provides a foundation for understanding the latest methodologies and their applications in customer churn prediction.

OVERVIEW

- Chapter 3 - Theoretical Basic: Examining the role of machine learning in churn prediction: This section explores the basics of machine learning and its integration with predicting churn. It explains commonly used machine learning algorithms like decision trees, random forests, logistic regression, and support vector machines.
- Chapter 4 – Data Exploration: Discussing data requirements and preprocessing. Effective handling and processing of customer data are crucial for accurate predictions. This part focuses on data collection, cleaning, feature engineering, and other essential steps to ensure data quality and reliability.
- Chapter 5: Proposed Model: Analyzing case studies and real-world applications. By studying successful real-world examples of organizations that have implemented machine learning-based churn prediction models, practical insights and best practices can be derived.
- Chapter 6 – Experimental Result And Analysis: Evaluating model performance metrics. The effectiveness of any predictive model depends on its ability to provide reliable results. This section discusses various performance metrics for evaluating and comparing the predictive power of different machine learning models.
- Chapter 7 – Conclusion: Future prospects and challenges: The report also touches upon potential future developments in customer churn prediction and the challenges that may arise as the field of machine learning continues to evolve.

RELATED WORD

CHAPTER 2: RELATED WORD

In chapter 2, we will make references as well as learn about research papers related to the topic that the group conducts, Customers Churn. Our team will learn about the process as well as what models and methods of conducting that research can be applied to the group's project.

2.1 Background of research

2.1.1 Foreign Research

Below are some international research papers that the group has consulted and found many models and knowledge that the group can apply to the process of implementing their project.

- Mutanen, T., Nousiainen, S., & Ahola, J. (2010). Customer churn prediction—a case study in retail banking. In *Data Mining for Business Applications*
This research will focus on customer value analysis along with customer churn prediction which will help marketing programs to target more specific customer groups. This work focuses on one of the central themes in customer relationship management (CRM): transferring valuable customers to competitors. The results of the case study suggest that using conventional statistical methods to identify churners can be successful.
- Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*
In this study, the author uses IBRF technique to predict customer leaving behavior. In this section, we present the methodological background of the technique and the evaluation criteria that we use to analyze the method's performance.
- Bilal Zorić, A. (2016). Predicting customer churn in the banking industry using neural networks. *Interdisciplinary Description of Complex Systems: INDECS*
The purpose of this paper is to present a case study on the use of one of the data mining methods, neural networks, in knowledge discovery from databases in the banking industry. In this study, the author used one of the data mining methods, neural networks, in the Alyuda NeuroIntelligence software

RELATED WORD

package to predict customer churn in the bank. Neural network is a statistical learning model inspired by neurobiology and it is used to estimate or approximate functions that may depend on a large number of inputs that are generally unknown. .

- Anil Kumar, D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*

In this paper, the authors address customer credit card churn prediction through data mining. We have developed a synthesis system that combines majority voting and involves Multilayer Perceptron (MLP), Logistic Regression (LR), decision tree (J48), Random Forest (RF), Radial network Basis Function (RBF) and Support Vector Machine (SVM) as components. Classification and Regression Trees (CART) are used for feature selection purposes. The reduced feature set is included in the classifiers mentioned above. Therefore, this article outlines the most important predictors in solving the credit card abandonment rate prediction problem.

- Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal Of Applied Microeconometrics*

This study focuses on customer churn analysis, which is an important topic in bank customer relationship management. Identifying exit customers will help management categorize potential early leavers and target customers using promotions, as well as provide insight into what factors are needed. considered when retaining customers. Although different models are used for churn analysis in the literature, this study focuses on Machine Learning models that are particularly easy to interpret and use SHAPLY Additive exPlanations (SHAP) values to support machine learning model evaluation and interpretability for customer analysis. According to the results, the XgBoost model outperforms other machine learning methods in classifying customers leaving.

RELATED WORD

- Saw Thazin Khine, Win Win Myo (2019). Customer Churn Analysis in Banking Sector

To overcome the instability and limitations of a single prediction model and predict the churn trend of high-value users, the churn prediction model for classifying bank customers is built in this study using a hybrid model of k-means and Support Vector Machine data mining methods on the bank customer churn dataset. This strategy also encourages data about comparable client groups to think about what marketing responses should be given. As a result, banks will see an increase in earnings and revenue as existing clients are kept. Additionally, the K-means clustering approach is presented with a combination model K-means-SVM that decreases support vectors and speeds up SVM training.

- Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in the telecom industry using boosting. IEEE Transactions on Industrial Informatics

This study performs a real-world study of customer churn prediction and proposes the use of reinforcement to enhance the customer churn prediction model. Unlike most of the studies that use reinforcement as a method to increase the accuracy of a given basic learner, this paper tries to separate customers into two clusters based on the weights specified by the algorithm. reinforcement math. As a result, a group of higher risk clients was identified. Logistic regression was used in this study as a baseline learner and a churn prediction model was built on each cluster, respectively..

2.1.2 Domestic Research

- Trang, N. T. T., Liên, N. T., Bích, P. T. N., & Kim, K. N. Applying machine learning methods to predict the likelihood of customers leaving credit card services.

This study predicts the likelihood of customers leaving credit card services at banks using machine learning methods. The methods used include Random Forest, SVM, Naïve Bayes, Logistic regression, and a combination of all four

RELATED WORD

methods. The analysis results show that these methods have good predictive quality with high accuracy. In particular, the forecast results by Random Forest are the best on all criteria including Accuracy, Precision, Sensitivity, Specificity and F1 score. This result can provide recommendations for bank managers in retaining customers who are using credit card services.

2.2 Gain relevant research results and model suggestions

In service businesses where there is fierce rivalry, losing consumers is a significant problem. However, firms might create large new income possibilities if they can spot consumers who are likely to quit early. By evaluating historical customer data, studies have demonstrated that machine learning approaches are useful for forecasting both churn and non-churn occurrences. This information about clients contains both past and present information. This project focuses on machine learning methods and algorithms such as Logistics Regression, Random Forest, SVM, XGBoost, and Catboost to forecast churn in the telecom industry. The telecommunications industry has several challenges, thus it's critical to adopt and adhere to best practices. Various prediction models were examined and assessed in this study utilizing quality metrics including the confusion matrix.

CHAPTER 3: THEORETICAL BASIC

This chapter covers key topics such as data mining techniques, churn rate definition, handling imbalanced datasets, and explores various machine learning algorithms, including Logistic Regression, Random Forest Classifier, Xtreme Gradient Boosting Classifier, Decision Tree and K-Nearest Neighbors, along with their applications and significance in decision-making processes.

3.1 Data mining

Data mining is a field of study that focuses on extracting useful patterns and knowledge from large volumes of data [3.4.1_1]. It involves the use of various techniques and algorithms to analyze data, discover hidden patterns, and make predictions or decisions based on the findings.

Data mining encompasses several tasks, including data preprocessing, exploratory data analysis, pattern discovery, and predictive modeling [3.4.1_2]. The process typically involves steps such as data collection, data cleaning, feature selection, and applying algorithms for pattern extraction and knowledge discovery.

Data mining aims to discover valuable insights and knowledge that enhance decision-making, process optimization, trend identification, and competitive advantage [3.4.1_1]. It finds applications in diverse fields like business, healthcare, finance, marketing, and social sciences.

Various algorithms and techniques, such as association rule mining, classification, clustering, regression analysis, and anomaly detection, are employed in data mining [3.4.1_2]. These approaches enable analysts to uncover significant patterns, relationships, and trends within data, empowering informed decision-making and actionable insights.

3.2 Churn rate definition

First of all, we need to be concerned about the concept of churning. Currently, the issue of customer retention is a matter of great concern to businesses. According to Arun Velu (2021), the situation where customers stop using the service or switch to a competitor is called churning.

THEORETICAL BASIC

Voluntary and involuntary churn are the two main categories of churn. Involuntary churn, on the other hand, occurs when a client has no influence over leaving a service, and examples of this include failure to pay, moving, and many other situations. Voluntary churn occurs when a customer leaves a service or changes to one supplied by a rival. (Krull, 2021). In practice, businesses cannot control cases of involuntary churn. Businesses need to focus on voluntary churn and come up with strategies to meet customer needs and satisfaction.

According to the article “Customer churn 101: What it is, why churn happens, and what you can do about it”, the author has classified churn in figure 3.1 Classification of churn below:

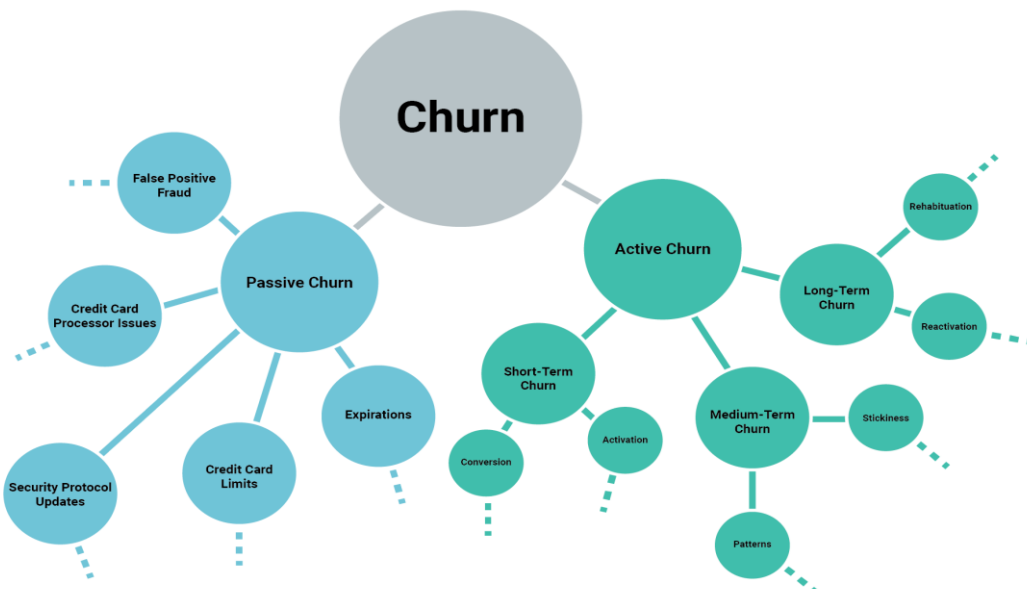


Figure 3.1 Classification of churn (Customer churn 101,2021)

When churn model implementation is concerned, churn rate is crucial. To use in the churn method, we must compute the churn rate. We must know the number of consumers who ceased using a service during a certain time frame and the number of customers who are still using the service in order to compute this rate (KDnuggets, 2021).

THEORETICAL BASIC

3.3 Imbalance dataset

3.3.1 Definition

An Imbalanced Dataset is defined as when samples from one class are considerably more or significantly less represented in the dataset than samples from the other classes, the dataset is said to be balanced. Machine learning models may become exclusive to the dominant class and disregard the minority class as a result, which might present issues with training them.

Imbalance Dataset has Precision, Recall và F1-score. These are the exact metrics that measure the performance of the model when the data is out of balance. Accuracy is the proportion of correct predictions out of all the predictions that are in the smallest class of numbers. Recall is the proportion of cases where the prediction is correct out of all the cases that are actually in the class of minimum numbers. F1-score is the harmonic average of Precision and Recall.

3.3.2 Examples of Imbalanced Dataset

Here are 2 examples related to Imbalanced Dataset

- Fraud detection in financial transactions: When compared to the number of genuine transactions, the number of fraudulent transactions is frequently quite tiny, creating an imbalance in the data set.
- Spam mail detection: The ratio of spam emails to legitimate emails is frequently quite low, which causes an imbalance in the data set

3.4 Machine Learning

3.4.1 Logistic Regression

In 1990, the phrase "logistic models" was added to the Medical Subject Headings (MeSH) thesaurus used by the National Library of Medicine to index articles for the Medline/PubMED database. Logistic regression models are defined as “statistical models which describe the relationship between a qualitative dependent variable (that is, one which can take only certain discrete values, such as the presence or absence of a disease) and an independent variable”.

Logistic regression models are used to predict a categorical variable by one or more continuous or categorical independent variables. The dependent variable can be

THEORETICAL BASIC

binary, ordinal or multicategorical. The independent variable can be interval/scale, dichotomous, discrete, or a mixture of them all.

3.4.2 Random Forest Classifier

Random Forest Classifier is a machine learning algorithm used to classify objects or predict initial values based on an input data file. This algorithm uses a set of decision trees (decision trees) to generate predictions and finally the results of these decision trees to give the final result. Each decision tree in the Random Forest Classifier makes its own prediction, and ultimately, the outcome is decided by averaging these predictions or by taking the vote results from the decision trees. This algorithm is used to minimize overfitting objects and increase model diversity. We can better visualize the Random Forest Classifier in figure 3.2 below.

There are 4 steps to build Random Forest Classification including:

- Generate a random subset of the training data set.
- Build a decision tree on this sub-dataset by selecting important features to create clusterings.
- Repeat steps 1 and 2 to generate a set of decision trees.
- Combine the results from the decision trees to get the final result.

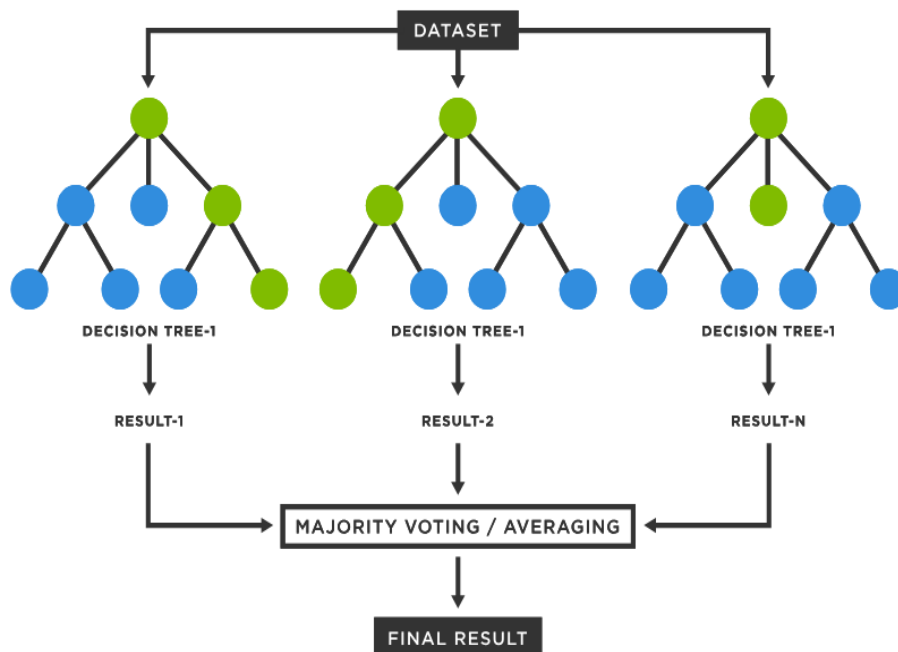


Figure 3.2. Random Forest Classifier (Tibco)

THEORETICAL BASIC

3.4.3 Extreme Gradient Boosting Classifier

XGBoost is a powerful machine learning algorithm widely used in predictive modeling tasks. It belongs to the ensemble learning methods, which combine the predictions of multiple weak learners to create a strong and accurate predictive model. According to Chen and Guestrin (2016), XGBoost leverages the boosting technique, iteratively training weak learners to correct the errors made by previous models. This iterative process focuses on challenging instances, resulting in improved predictive accuracy. [3.4.5]

XGBoost excels in handling various data types, including numerical and categorical features. It utilizes gradient boosting, where each subsequent weak learner is trained to minimize the residual errors of the previous model, enabling it to capture complex patterns effectively.

To enhance its performance, XGBoost incorporates innovative features. It employs a regularized objective function that combines a loss function with a penalty term, thus preventing overfitting and improving generalization. Furthermore, XGBoost utilizes advanced tree construction algorithms and parallel computing techniques to expedite training and enhance scalability. Another noteworthy aspect of XGBoost is its ability to handle missing values in the data. The algorithm can automatically learn how to handle missing values during the training process, eliminating the need for explicit imputation or data preprocessing steps.

Effective parameter tuning plays a critical role in optimizing XGBoost's performance. Chen and Guestrin (2016) suggest that careful adjustment of hyperparameters controlling the tree structure, learning rate, regularization terms, and subsampling ratios is necessary to prevent overfitting and achieve the best predictive results. [3.4.5]

3.4.4 Decision tree

Decision trees are foundational models in machine learning, known for their intuitive structure and ability to handle both classification and regression tasks. At its core, a decision tree is constructed through a process of recursive partitioning of the feature

THEORETICAL BASIC

space, guided by certain criteria, aiming to create a hierarchical flowchart-like structure that aids in decision-making.

The theoretical foundation of decision trees is built upon the idea of attribute selection to best discriminate between different classes or predict target values. At each internal node of the tree, a specific attribute is chosen based on measures like Gini impurity or entropy. Gini impurity measures the likelihood of misclassifying a randomly chosen element in the dataset, while entropy quantifies the level of disorder within a set of data points. By selecting attributes that reduce impurity or disorder, the tree effectively makes decisions that differentiate classes or predict values more accurately.

3.4.5 KNN

K-Nearest Neighbors (KNN) is a fundamental algorithm in machine learning that belongs to the family of instance-based learning methods. Its theoretical foundation rests on the principle of proximity-based classification or regression. In KNN, data points are represented in a multi-dimensional feature space, and the algorithm operates by comparing the distances between these points.

At its core, KNN employs a simple yet powerful concept. When presented with a new data point, KNN identifies the 'k' closest neighbors from the training dataset, where 'k' is a user-defined parameter. These neighbors contribute to the decision-making process, either by voting (in classification tasks) or by contributing their values (in regression tasks). The distance metric, such as Euclidean distance, Manhattan distance, or others, quantifies the similarity or dissimilarity between data points. This proximity-based approach assumes that similar data points tend to share similar outcomes. However, the choice of 'k' and the distance metric significantly influence the algorithm's performance and generalization.

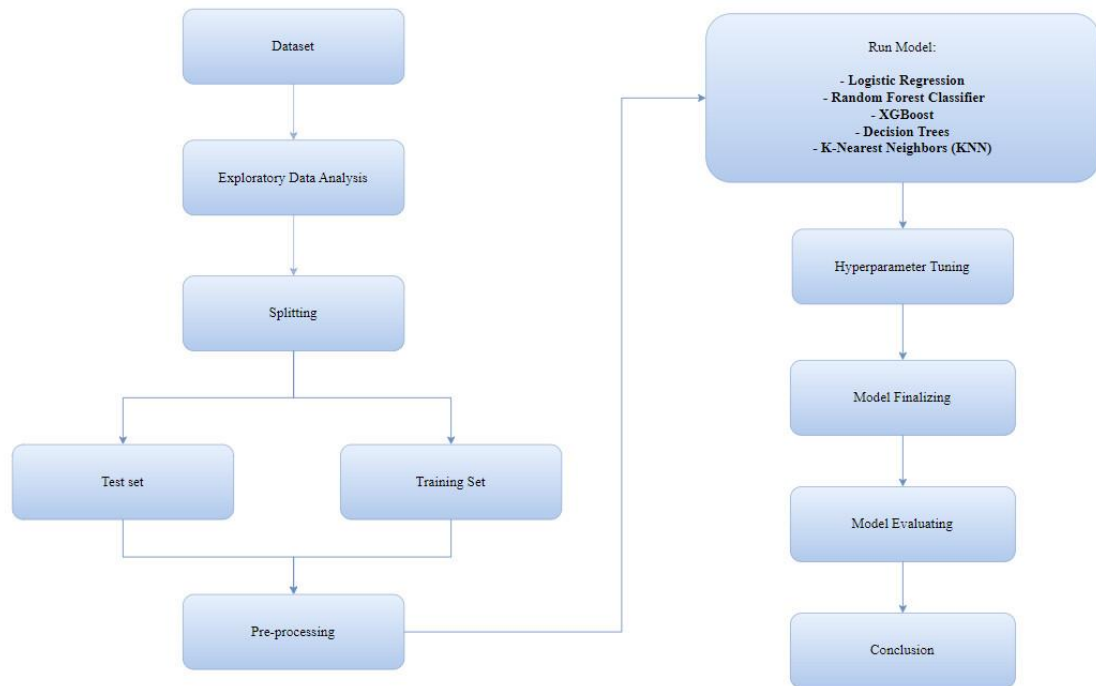
CHAPTER 4: PROPOSED METHODOLOGY

Figure 4.1 Our proposed methodology (authors)

Figure 4.1 shows an overview of the research methodology of this project. When starting with the project, we will review and select the appropriate data file as well as containing the customer data column related to the implementation topic that the team is aiming for. Next, we will focus on making the most of it and getting a brief overview of the data file that we has previously selected. In the next step, we will do the EDA process on that dataset. In it, we will visualize the data set into types of charts that show the correlation of attributes in the data file that affect customers in the process of using services at the bank. Next, we will perform data preprocessing with the content Feature Selection, Encoding Categorical Features, Scaling, Addressing Class Imbalance. After that, we will perform data separation and process customer data according to Machine Learning methods to obtain data for the following stages such as: Logistic Regression, Support Vector Classifier, Random Forest Classifier, Gradient Boosting Classifier, Xtreme Gradient Boosting Classifier, and Light Gradient Boosting Machine. In the final step, we will obtain the results of the data from which it will perform the process of comparing the results, evaluating

PROPOSED METHODOLOGY

the results obtained. At the same time, we will choose the best and most suitable model as well as give the best insights and advice for banking businesses to help reduce customer churn.

CHAPTER 5: DATA EXPLORATION**5.1 Overview of dataset****5.1.1 Dataset**

The data set that the team used to make their project thesis was taken from Kaggle which is a customer file in the bank. The data set includes 10000 rows of data equivalent to 10000 customers and includes 14 columns, each column will represent customer attributes. These customers collected data in 3 countries which are Spain, Germany and France. We have described the values of the attributes in the Table 5.1 below:

Table 5.1 The detail information of the Bank's customer dataset (authors)

No	Column Name	Description	Data type	Unique value
1	RowNumber	Numerical order	int64	10000 unique values
2	CustomerId	Identifier for each customer	int64	10000 unique values
3	Surname	Customer's surname or last name	object	2932 unique values
4	CreditScore	accumulated points	int64	
5	Geography	customer's location	object	Spain, Germany, France
6	Gender	Customer's gender	object	Male, Female
7	Age	Customer's age	int64	
8	Tenure	Number of years the customer has been with the	int64	0,1,2,3,4,5,6,7,8, 9,10

DATA EXPLORATION

		company		
9	Balance	Customer's current balance	float64	
10	NumOfProducts	Number of products customer has with the company	int64	1,2,3,4
11	HasCrCard	Whether the customer has a credit card with the company (Yes/No)	int64	1,0
12	IsActiveMember	Whether the customer is an active member (Yes/No)	int64	1,0
13	EstimatedSalary	Customer's estimated salary	float64	
14	Exited	Whether the customer churned (canceled subscription) or not (Yes/No)	int64	1,0

5.1.2 Handling missing data and duplicates

- **Missing Data**

After inspection, we found that this dataset has no missing data, figure 5.1 made this very clear.

DATA EXPLORATION

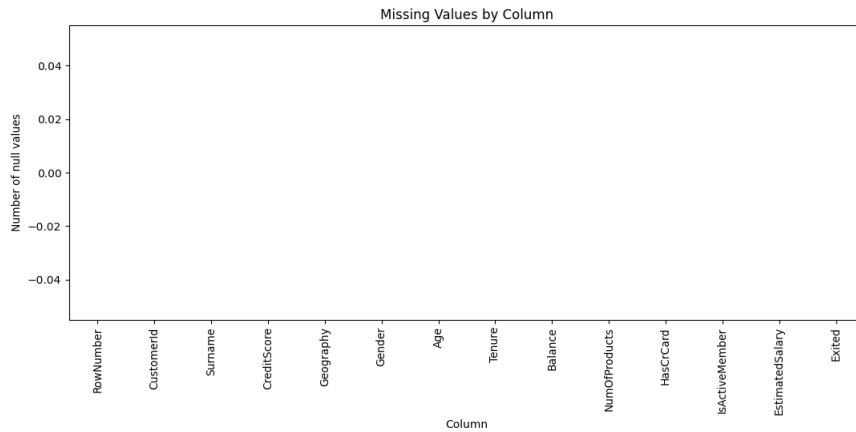


Figure 5.1 Missing Values by Column (authors)

- **Duplicates**

This dataset also has no duplicates data, which we can see in figure 5.2 below:

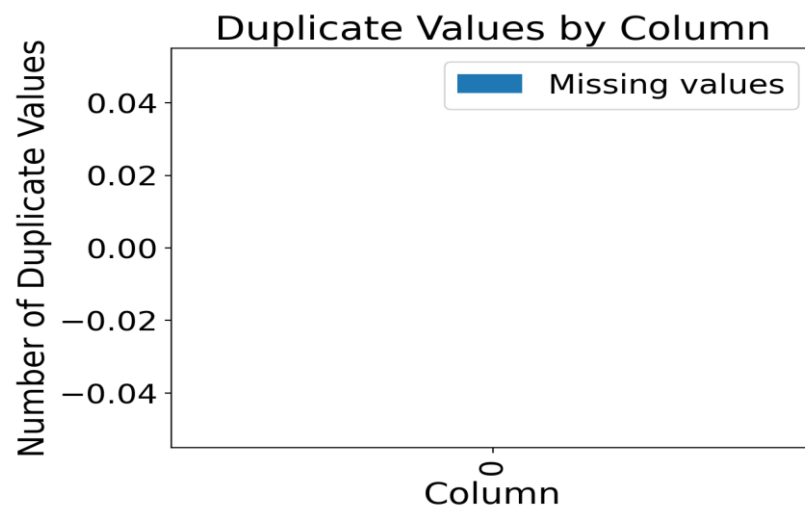


Figure 5.2 Duplicates Values by Column (authors)

5.1.3 Numerical features about the data summary

We have implemented the `describe()` method to give a statistics table of the numerical features in the figure 5.3

DATA EXPLORATION

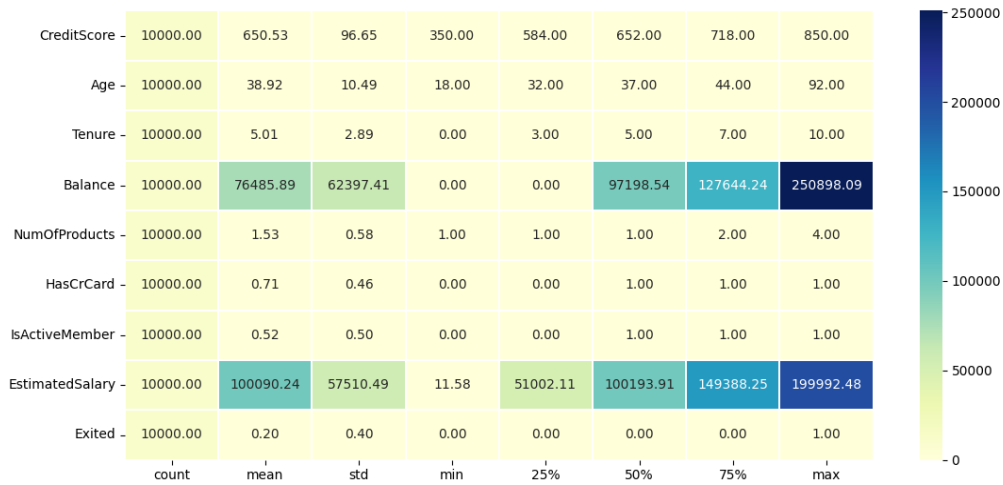


Figure 5.3 Numerical features by the authors

We calculated values like count, mean, std, min, 25%, 50%, 75% , max to get an overview of the properties in the data. After the calculation, we can draw the following observations:

The customer's credit score has a mean value of 650.53, this is an average credit score, neither too high nor too low.

The bank's customers have a diverse age range from 18 - 92 with an average value of approximately 40. Shows a diverse and highly concentrated customer base in the adult age group.

Customer loyalty is assessed quite well when the average value of tenure > 5.

At this bank, about 50% of active customers and about 20% of customers leave.

5.2 Exploratory Data Analysis

First of all, we need to distinguish between continuous and categorical variables by using numpy and pandas libraries to check the data. We then combine that with a function that checks for unique values and outputs a list of categorical and continuous variables as follows:

Categorical variables: ['Geography', 'Gender', 'Tenure', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'Exited']

Continuous variables: ['CreditScore', 'Age', 'Balance', 'EstimatedSalary']

DATA EXPLORATION

5.2.1 Continuous Variables

- Age

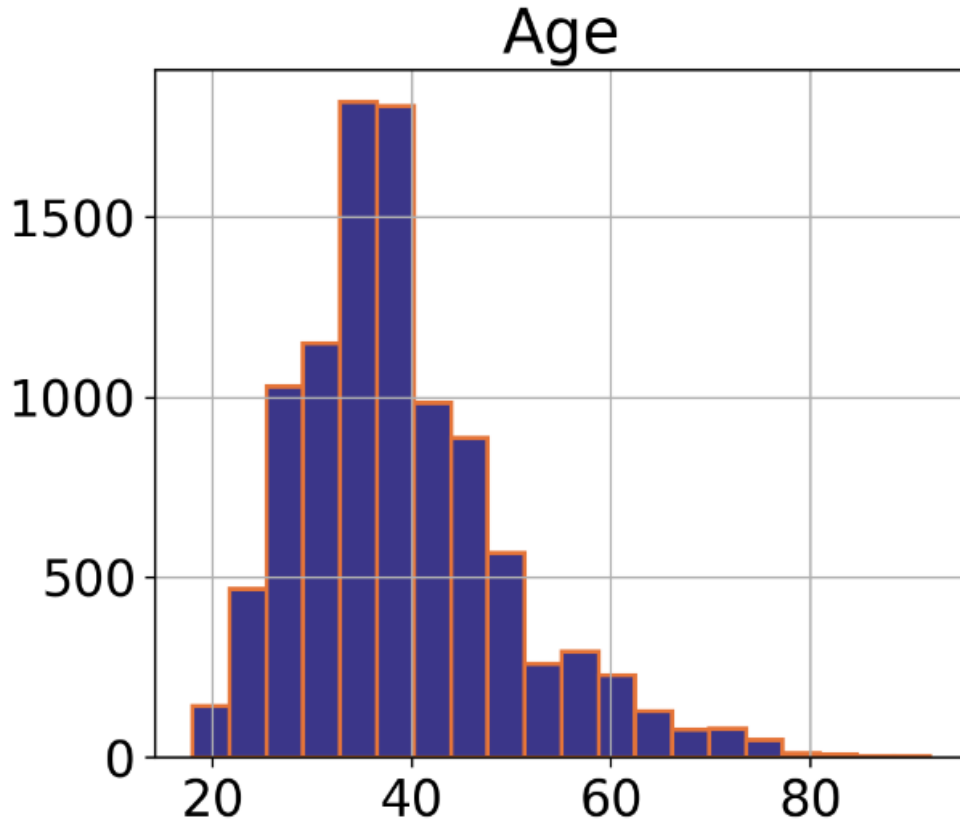


Figure 5.4 Histogram of Age

Looking at the figure 5.4 above, we can see that the young customers (18-24) are not too many but they are potential customers. According to some studies, young customers often tend to choose banks based on reputation, they appreciate safety to ensure their assets. Currently, this group of customers has not too strong financial capacity, but in the future they will continue to use banking services more.

The most concentrated age group is 32-40 years old, this is the age of middle-aged customers - people with stable incomes and high financial needs such as savings, transactions, investments, loans. .. They are willing to take risks and invest in other services. This age has accumulated enough experience and finance for them to make more risky decisions to bring better returns. In addition, this group of customers spend a lot on their families, such as land, cars, houses and children.

DATA EXPLORATION

The graph stretched to the right (heavy-tailed distribution) shows that the proportion of customers in the older age groups tends to increase significantly compared to the proportion of customers in the lower ages. This may indicate that a large number of customers have used the bank's services for a long time and tend to continue using the bank's services in the future. The average retirement age in Europe is 65 years old, this is the age when using banking services are less active because of many problems such as spending, or health, ties to children, traditional lifestyle. ... However, as banking services become more and more popular, along with the problem of population aging, we believe that this is still a potential customer file if the bank focuses on simple needs and their convenience.

- **Credit Score**

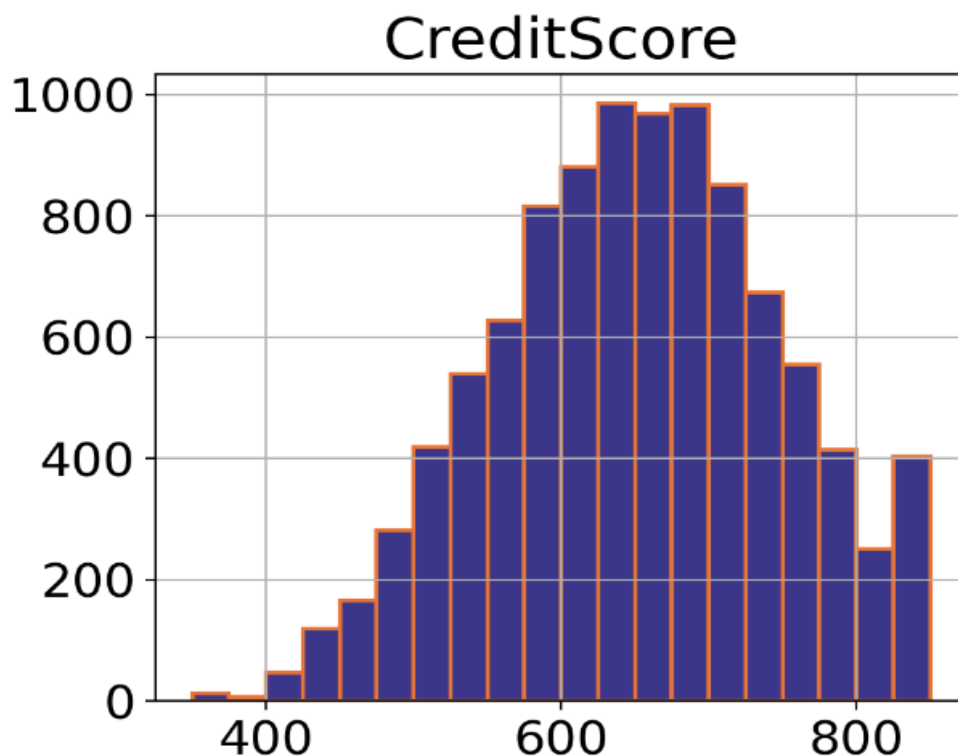


Figure 5.5 Histogram of Credit Score

Figure 5.5 shows a high percentage of credit scores from average and above (650 points). This shows that the bank's ability to manage credit risk is quite good when the customer file has a stable credit score. Partly due to the influence of special regulations in Europe on ensuring the stability of banks after the 2008 crisis, banks are quite cautious in assessing customers.

DATA EXPLORATION

Customers with below-average credit scores will often find it harder to get loans and financial services, and they'll have to pay higher interest rates. This not only ensures the safety of banks, but also causes them to lose many customers because of regulations on credit scores

However, for this bank, we clearly see the diversity of credit scores from 350 - 850, which shows that it is not only customers with high credit scores that become their customers. In other words, the bank has many services to be able to have many customers, credit score is an important factor, but the bank still welcomes customers with low scores.

- **Balance**

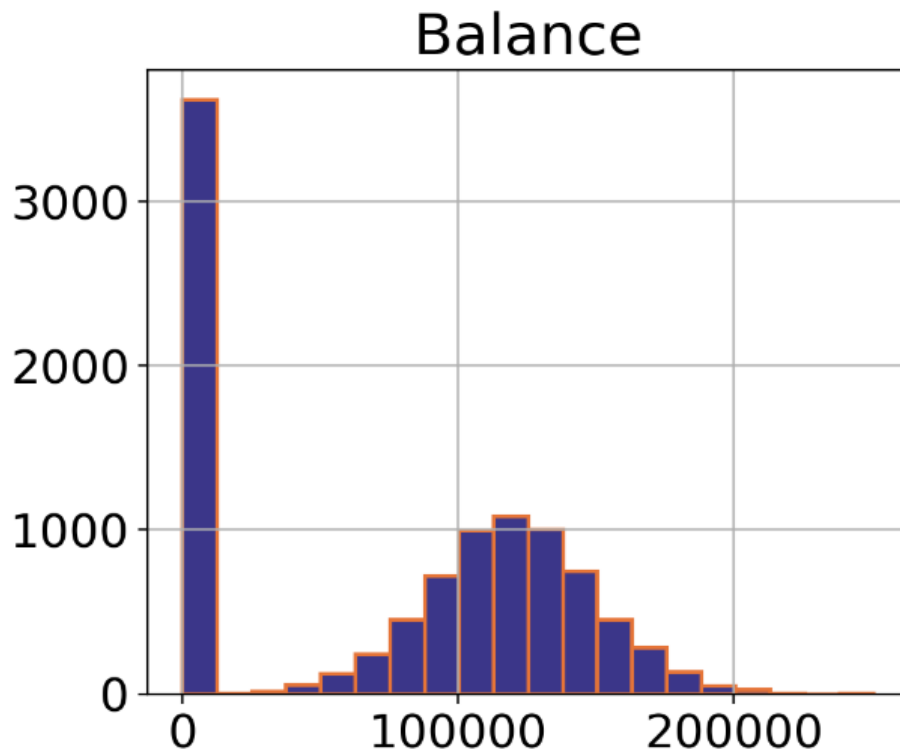


Figure 5.6 Histogram of Balance

Figure 5.6 shows that the zero remainder is very much, but if this is omitted, the balance is fairly normally distributed. The cause of the account balance being zero may be due to many customers opening the card but not using it, the number of active customers is also not high, so this number can be explained.

DATA EXPLORATION

The balance represents the amount the customer has deposited in their account, or the amount that they are borrowing from the bank. Therefore, for the bank, the balance can increase the bank's credit as well as the solvency.

The bank may review its balance to assess the customer's ability to repay the loan as well as evaluate the possibility of leaving.

Besides, the bank can also offer suitable policies and services for customers with outstanding balances.

- **Estimated Salary**

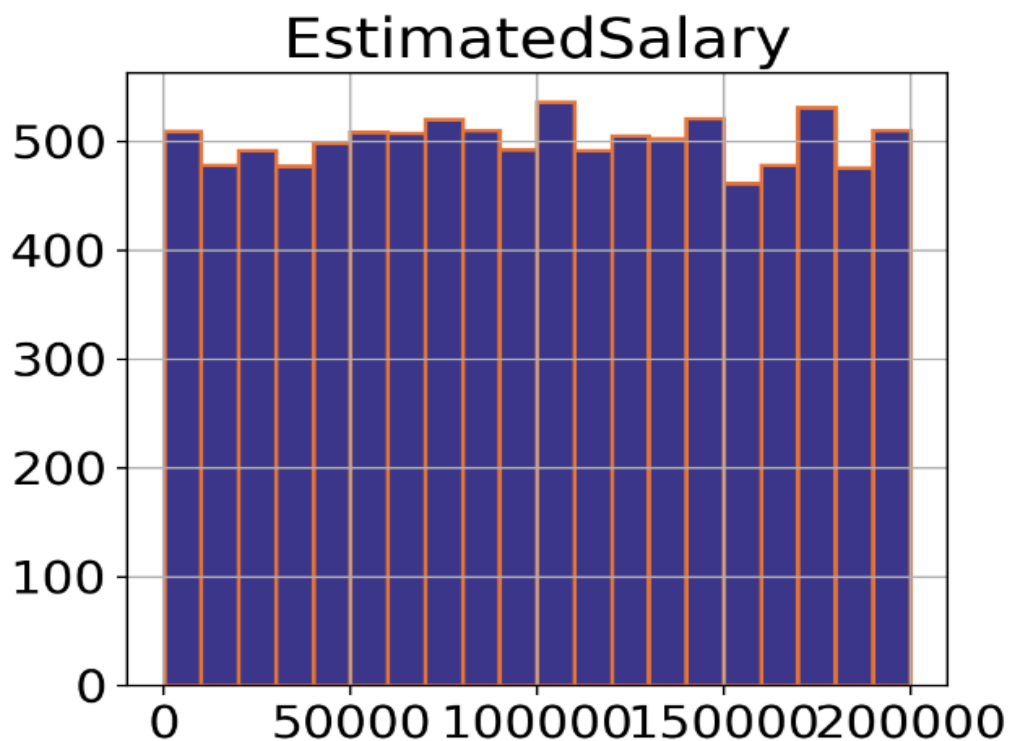


Figure 5.7 Histogram of Estimated Salary

Looking at the figure 5.7 we can see that the distribution of 'Estimated Salary' is almost uniform, however it still gives a lot of value in customer evaluation as it is a meaningful value.

People with high incomes are often able to save and invest more than those with low incomes. In addition, they often need to borrow money to buy a house, car, travel or invest in a business. They also have a need to use more advanced products and services such as credit cards, debit cards and online transfers to pay and manage their finances.

DATA EXPLORATION

Meanwhile, low-income people often have little need to use the bank's financial products and services, so they are easily excluded from the financial system and cannot access the products. and essential financial services.

Customers' income is an important factor for banks to determine their financial ability in loans, mortgages and credits. In addition, determining the customer's income helps the bank to offer more suitable services for them, for example, customers with high income will enjoy more premium services, higher interest or higher limits.

5.2.2 Categorical Variables

- **Gender**

Here, we use the `ns.countplot` function in the `seaborn` library to represent the frequency of categorical variables. We have visualized the frequency of the gender variable in the figure 5.8 below:

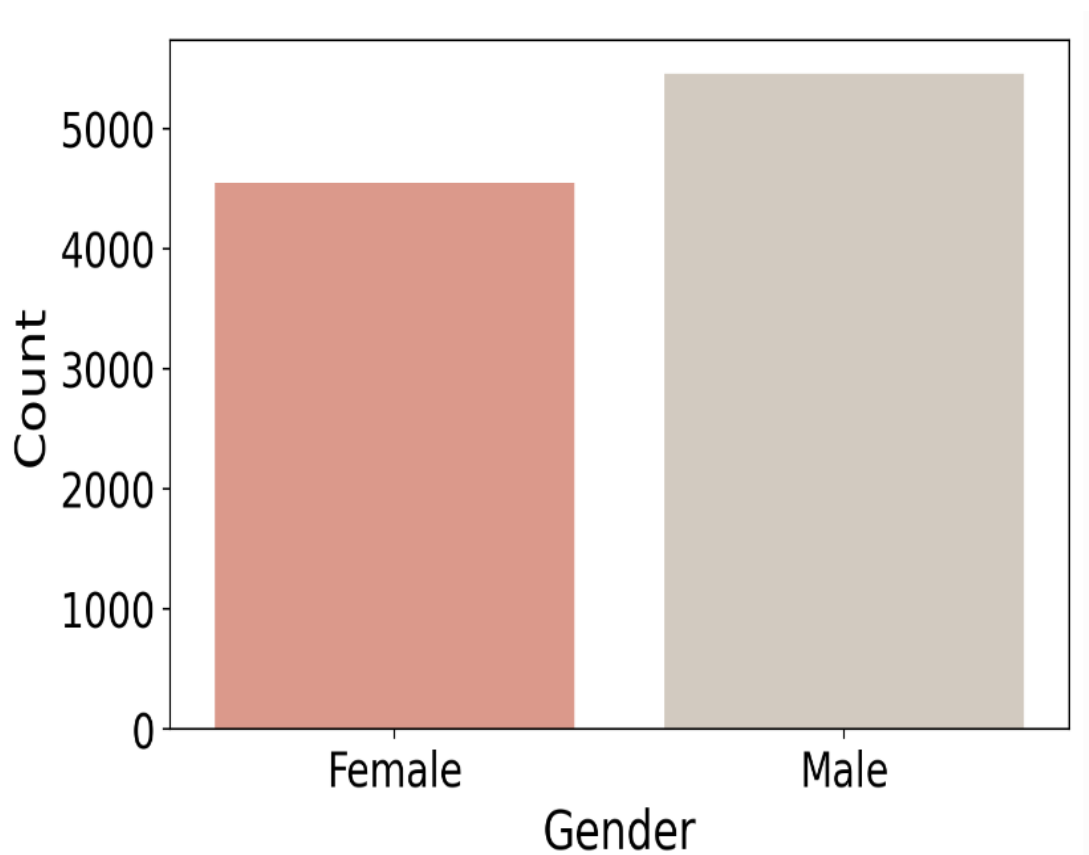


Figure 5.8 Histogram of Gender

DATA EXPLORATION

The gender ratio in this data is not too disparate, there are more male customers than female. Banking use by gender cannot accurately gauge whether men or women use banks more. However, the use of services tends to be as follows:

- Women tend to use online and mobile banking services more than men. This may be because women tend to have busier lifestyles, less time to go to the bank in person, and need more flexibility in financial management.
- Men tend to use financial services for investment and securities transactions of banks more than women. This may be because men are generally more knowledgeable about finance and investments, and tend to be interested in increasing income.

Therefore, banks can also rely on this factor to make business strategies or advertising and marketing strategies to focus on certain special needs of customers.

- **Geography**

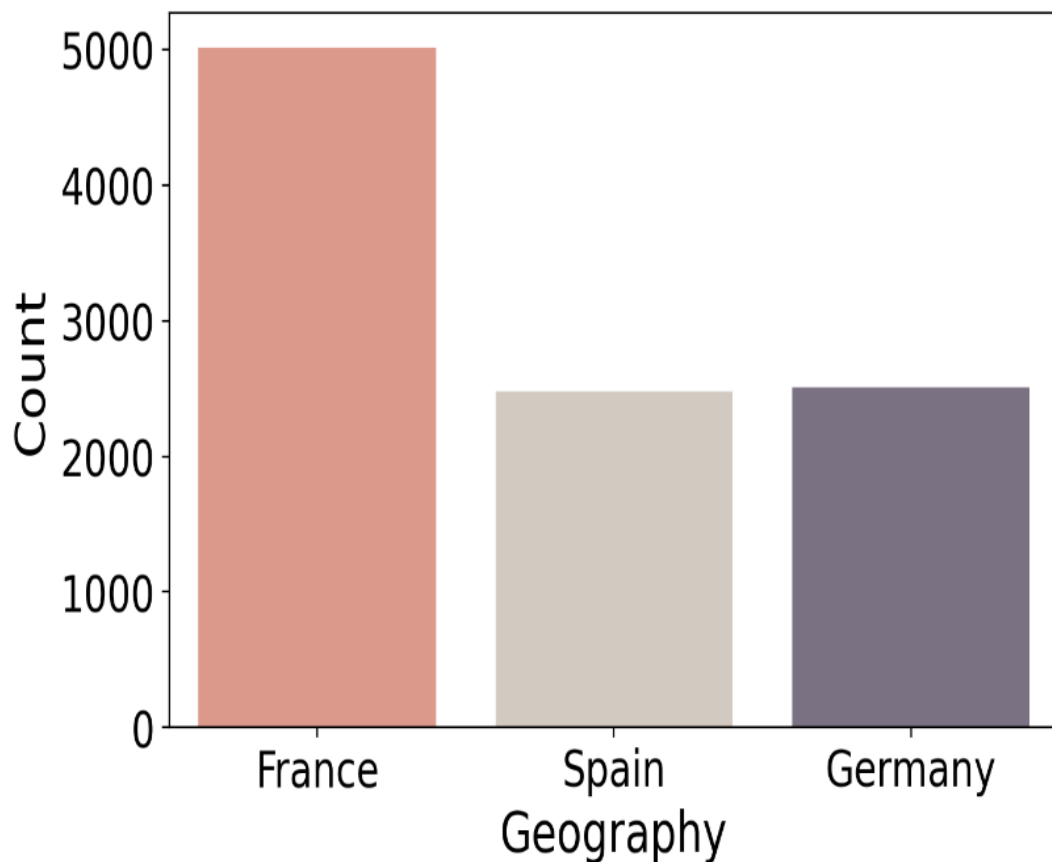


Figure 5.9 Histogram of Geography

DATA EXPLORATION

Figure 5.9 clearly shows the bank has customers in three countries (France, Spain and Germany). And most of the customers are in France. These countries belong to Western Europe with diversified developed economies. A special feature about banking here is the strict regulation after the 2008 financial crisis. This makes banks more resilient to the risk of collapse but becomes less attractive and less profitable. .

- **Number of Product**

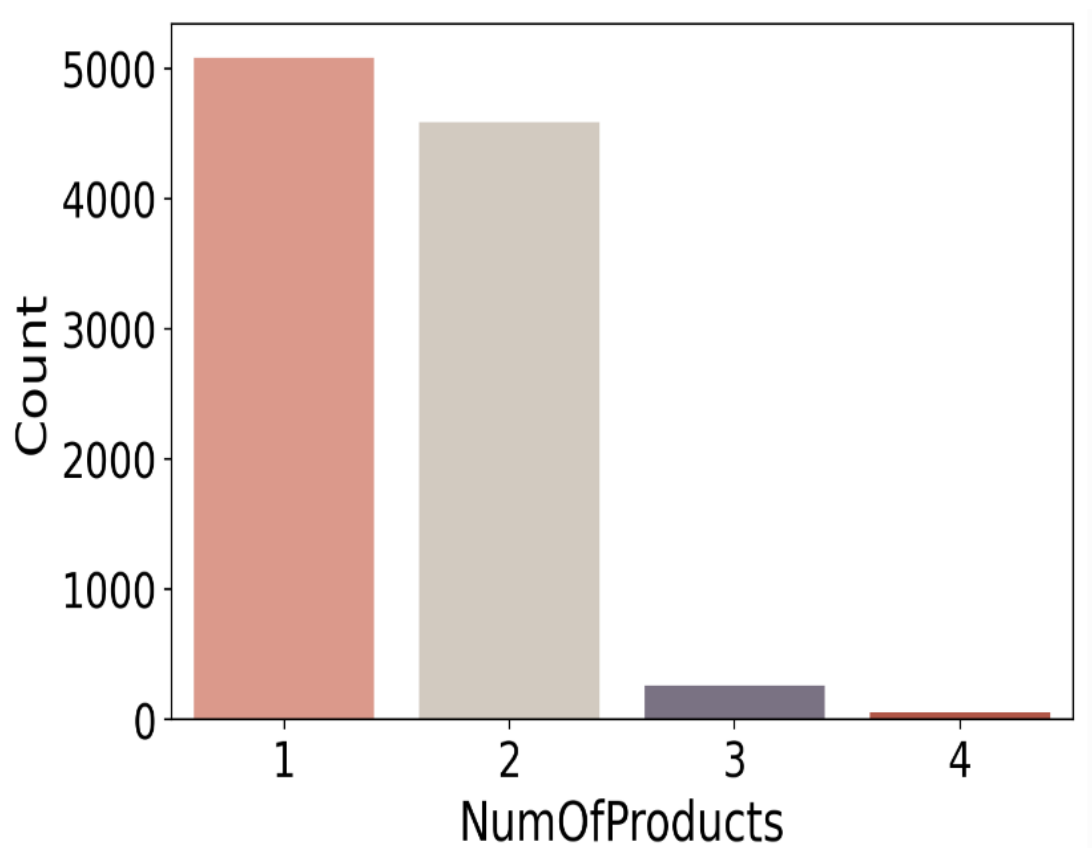


Figure 5.10 Histogram of Products

Figure 5.10 shows that Customers use 1 - 2 services much more than 3 and 4. Banks need to focus on what is the cause of the difference? And come up with strategies to promote products based on the target customer.

DATA EXPLORATION

- **Tenure**

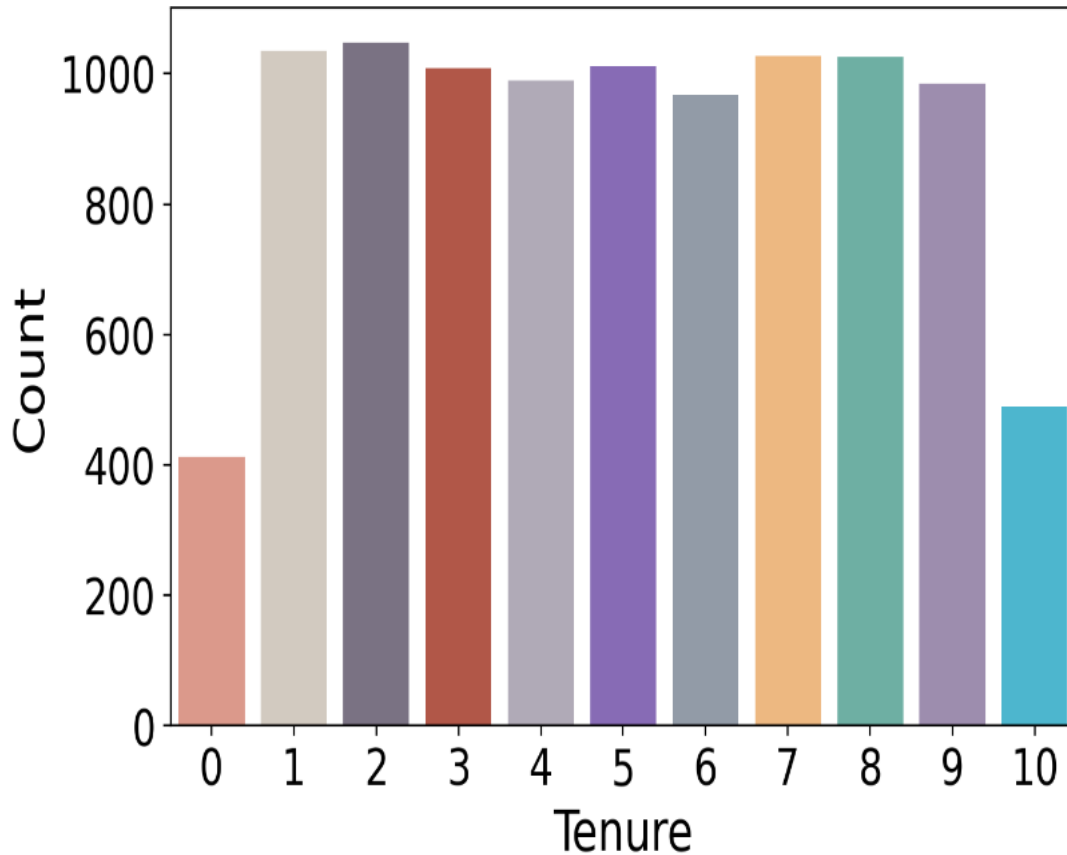


Figure 5.11 Histogram of Tenure

At figure 5.11 we can see that the lifetime of the bank is fairly evenly distributed and stable except for the value of customers leaving in the first year.

The service time in the bank shows the customer's loyalty to the bank and can show whether the customer is satisfied with the service or not. From the chart above, we can see that the bank has many long-term customers, which shows that the bank provides quite good services.

This value also helps the bank to classify the customer file to offer new services or improve service quality. For example, for long-term customers, the bank may offer special incentive programs or provide personalized financial advice to increase customer retention rate.

DATA EXPLORATION

- **Has Credit Card**

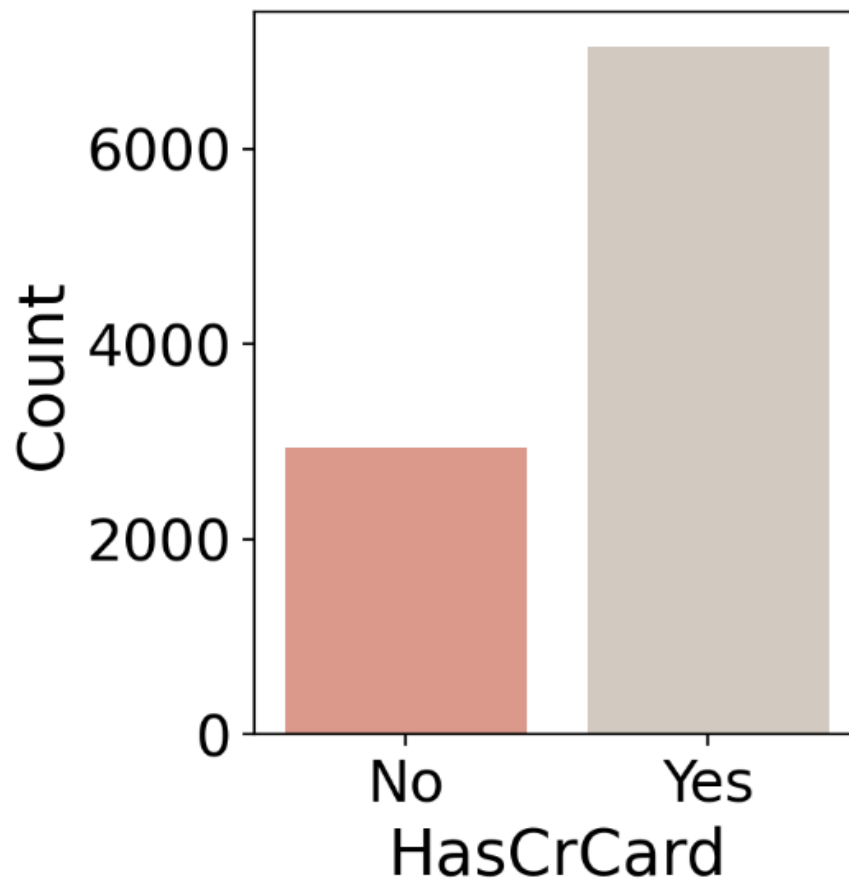


Figure 5.12 Histogram of HasCrCard

Figure 5.12 shows the majority of customers using credit cards. The use of a credit card by a customer may affect a customer's behavior using financial services. If a customer already has a credit card, it is likely that they will stay with the bank for a long time and become familiar with the customer's products and services, from which the bank can recommend new services to them or find ways to retain customers.

DATA EXPLORATION

- **Active Member**

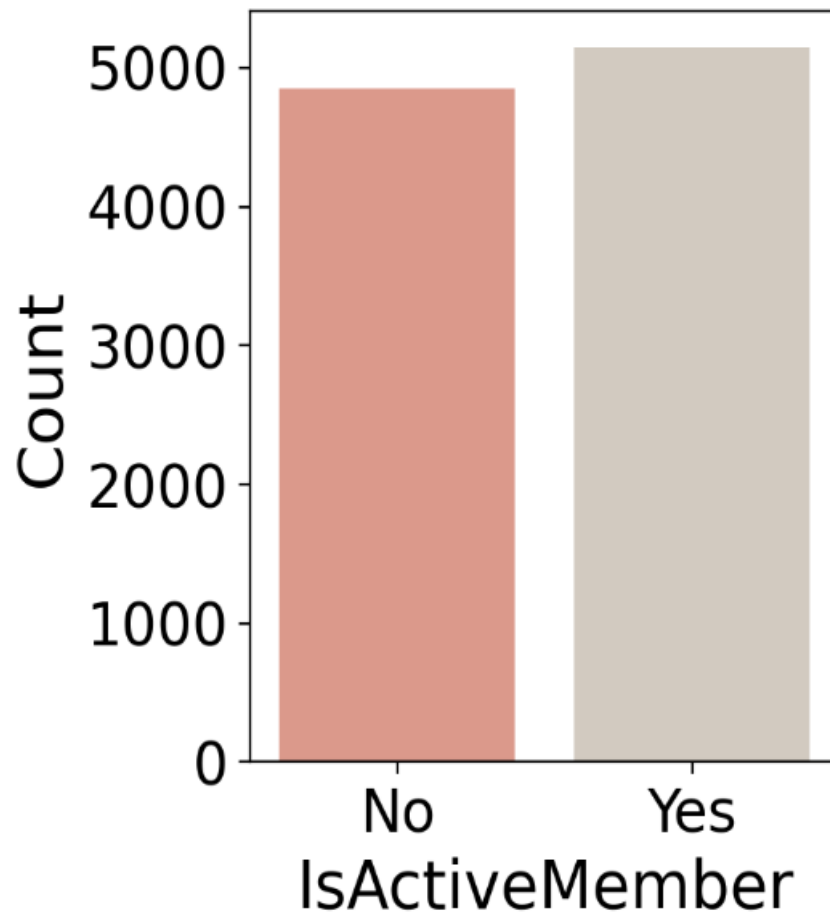


Figure 5.13 Histogram of Active Member

The number of inactive customers can be assessed against a certain time threshold of the bank, for example, within 3 or 6 months the customer has not had any transactions or activities for the services they are interested in. use.

We can see on the figure 5.13, although the number of active customers is more than the number of inactive customers, this number is not high but also approximately the same, the number of inactive customers accounts for nearly 50%. This number shows that the bank is facing challenges in maintaining relationships with customers. The bank needs to pay attention to the reason why customers are no longer active, whether they have problems using the services or no longer need to use them, so that they can take appropriate remedial measures. This is a very important variable and greatly affects the business activities of the bank.

DATA EXPLORATION

- **Exited**

This is the value that we are most interested in, the Exited variable is the target variable.

The Exited variable has two values:

Zero (0) for a customer who hasn't left and one (1) for a customer who has left. We can find out how the bank got this result, maybe the bank determines the customer is no longer active for a certain time as leaving or conducts a customer survey, it can also be they receive a request from the customer to terminate the service. However, in the dataset, there is clear data for customers to leave, so we do not need to pay too much attention to the data collection method. We have shown the number of customers leaving at figure 5.14.

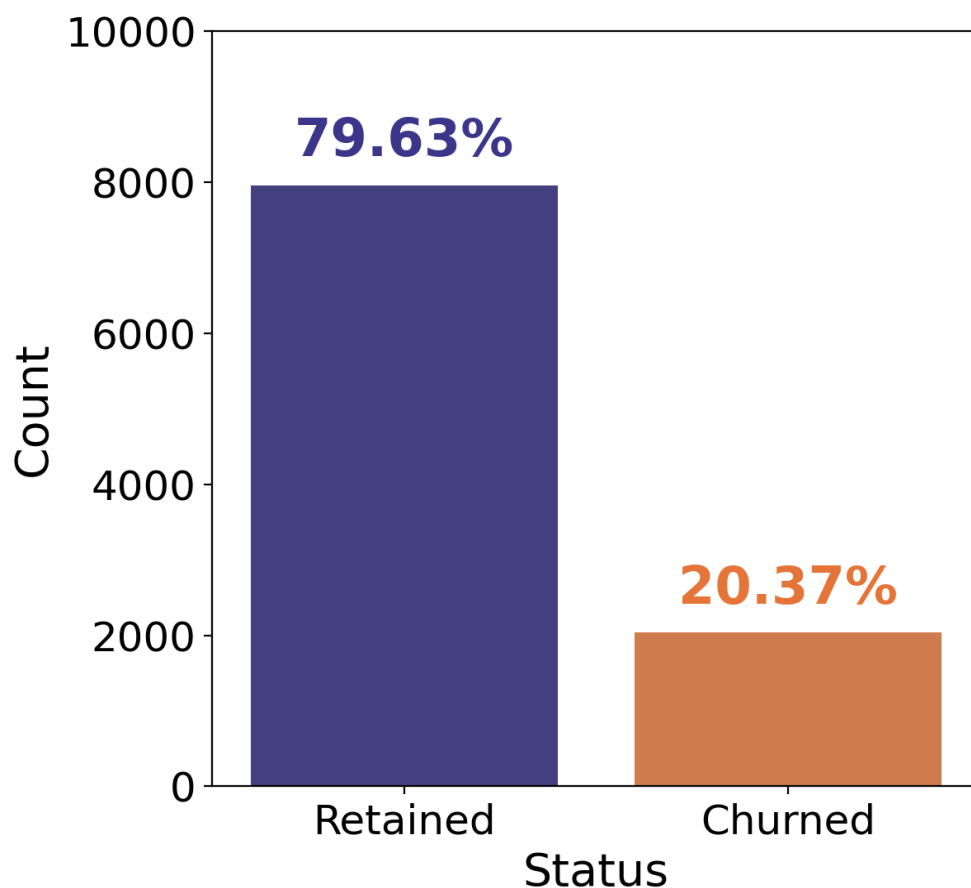


Figure 5.14 Histogram of Churn rate

DATA EXPLORATION

- It can be seen that the number of customers leaving accounts for about 20%. This churn rate is typical because according to statistics in the financial sector, the churn rate accounts for 18-20%.
- A low churn rate is not necessarily a good thing because the bank may grow slowly and find it difficult to replace new customers. If the churn rate is high, the bank may have quality and service problems.
- In addition, depending on the service that customers use, the abandonment rate should also be assessed differently. If the bank's services are aimed at long-term customers, of course the important goal is to retain customers. However, if the customer's service is as simple as opening a payment card or transferring/receiving money/, the goal of retention is not so important because customers often change banks. Therefore, to evaluate the ratios, it is necessary to consider the bank's services and strategy.

5.3 Relationship of target variable

5.3.1 Age

We use the histplot and boxplot method to show the correlation between the variables and the target variable. Based on the figure 5.15 it can be seen that the older the customer, the more likely to leave. For customers aged 40 and above, the trend will begin to leave the bank. This can evaluate the strategy as well as the incentives coming from the bank for customers that are no longer attractive so that customers can continue to use services at the bank. Besides, the bank probably did not have many surveys on middle-aged people to understand their needs and preferences, so customers aged 40 and above no longer use them banking and leave.

DATA EXPLORATION

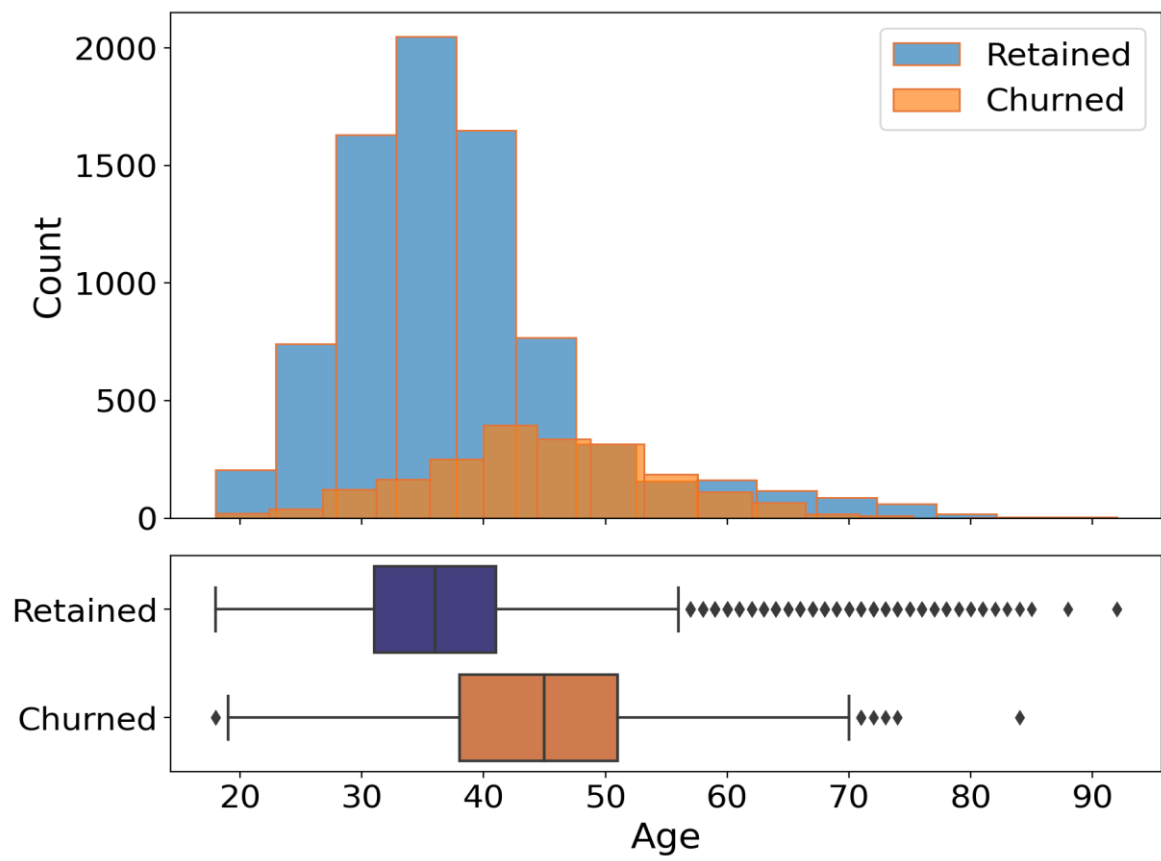


Figure 5.15 Correlation of Age and Churn

5.3.2 Credit Score

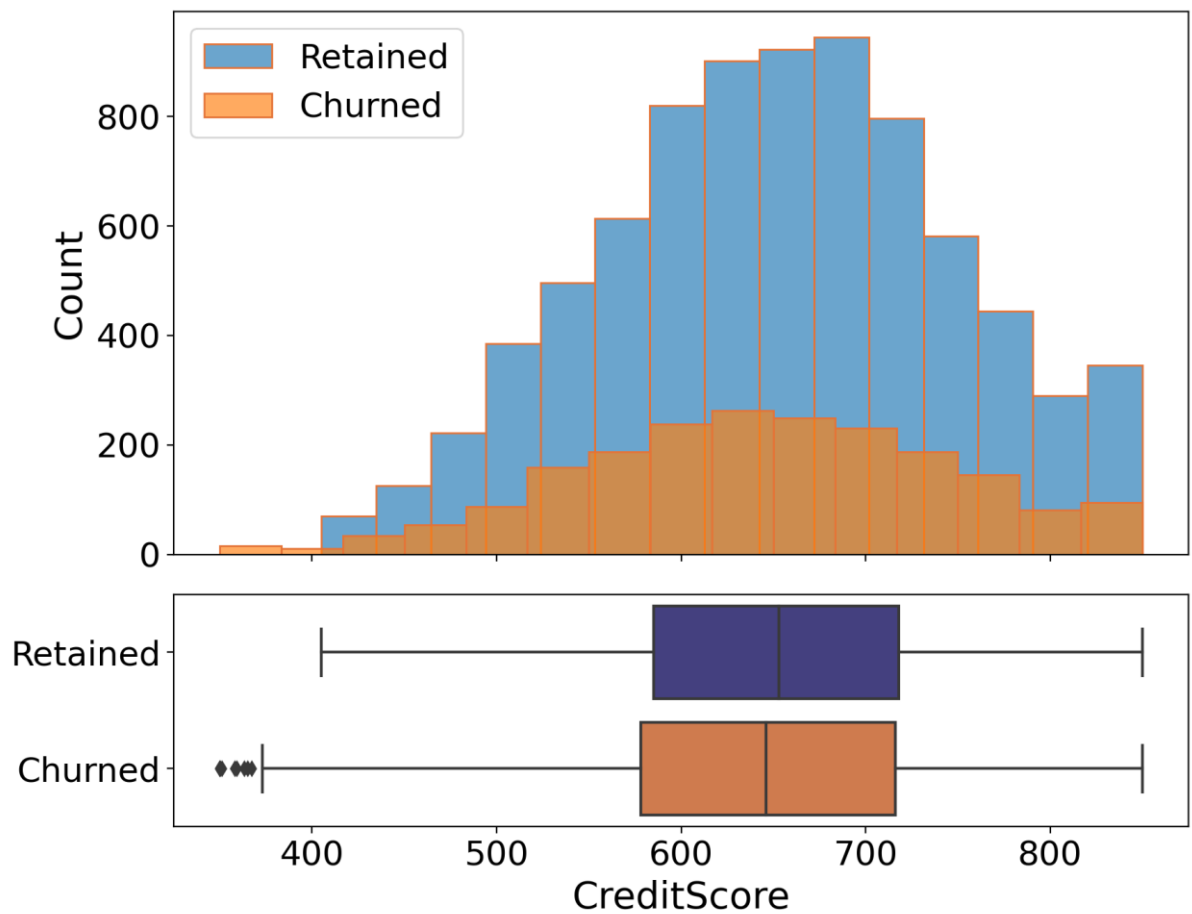


Figure 5.16 Correlation of Credit Score and Churn

The figure 5.16 shows that the credit score of customers who leave and customers who continue to use the service at the bank does not have much difference as well as a significant difference. This also assesses the bank's ability to create many opportunities to help customers accumulate their credit accumulation points in the process of using the bank.

5.3.3 Balance

With account balances of customers leaving and customers continuing to use services at the bank, the distribution is quite similar. It can be seen that the number of customers who do not leave with a low account balance is almost zero, which is understandable because the customer has just opened a bank account, so there is not much account balance. We can clearly see that in the figure 5.17 below:

DATA EXPLORATION

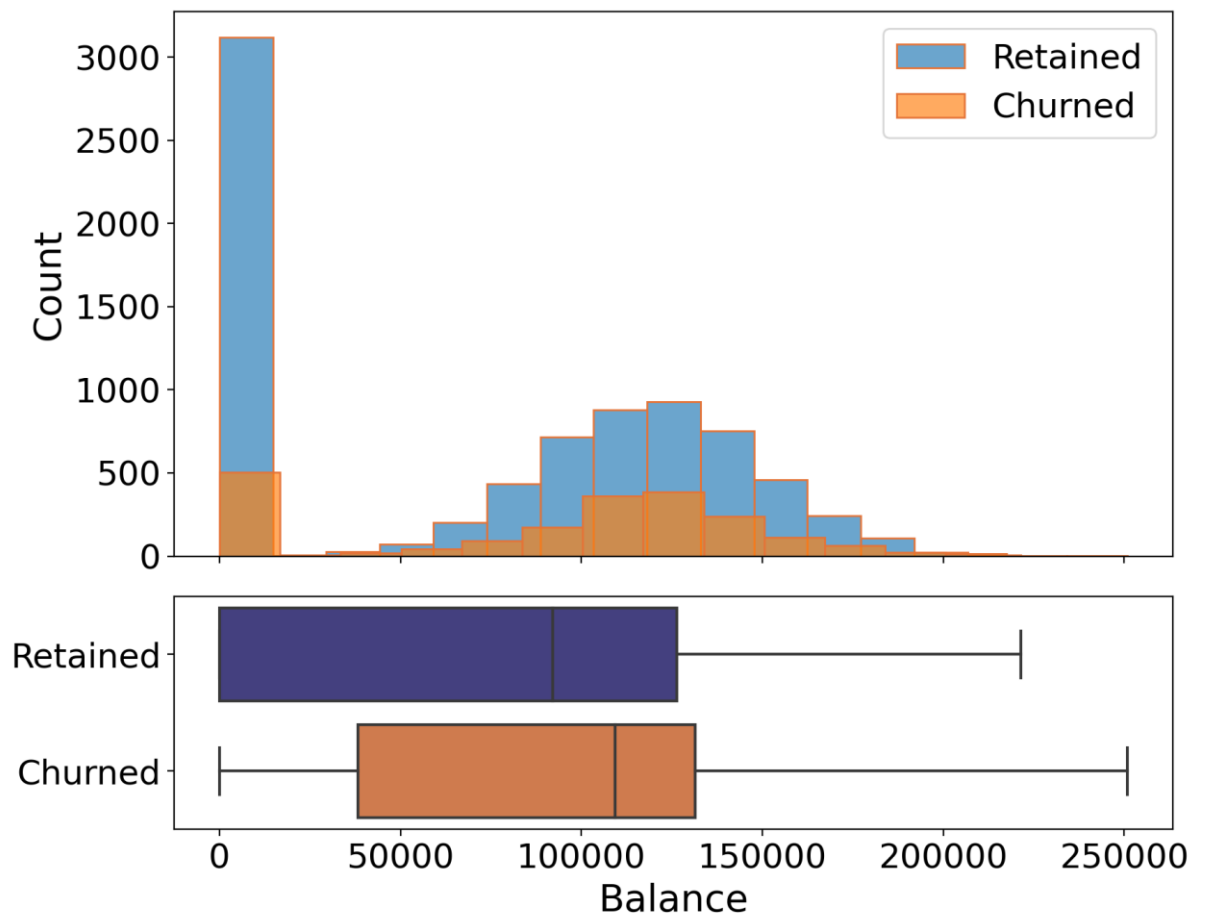


Figure 5.17 Correlation of Balance and Churn

5.3.4 Estimated Salary

The chart at figure 5.18 shows the salary distribution of customers who stay and leave the bank with almost no difference, which shows the uniformity of their salaries. Therefore, salary is not a factor to decide whether they continue to use the bank's services or not.

DATA EXPLORATION

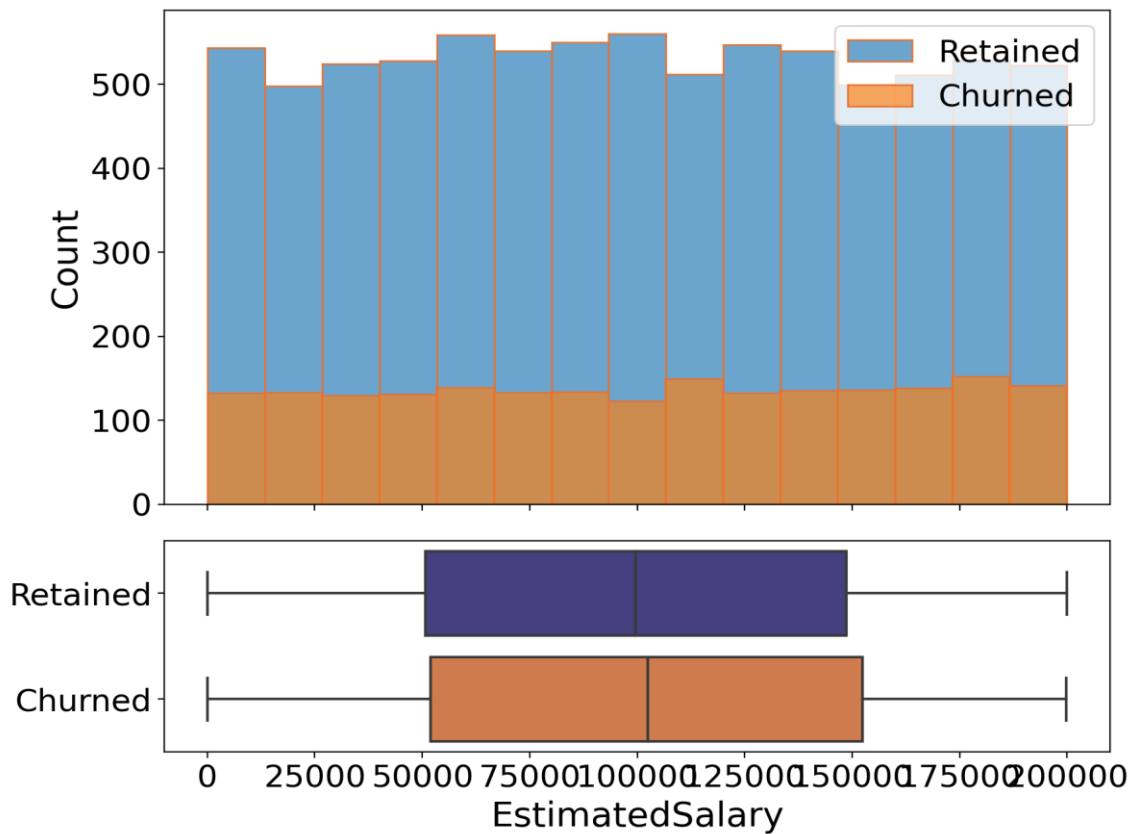


Figure 5.18 Correlation of Estimated Salary and Churn

5.3.5 Geography

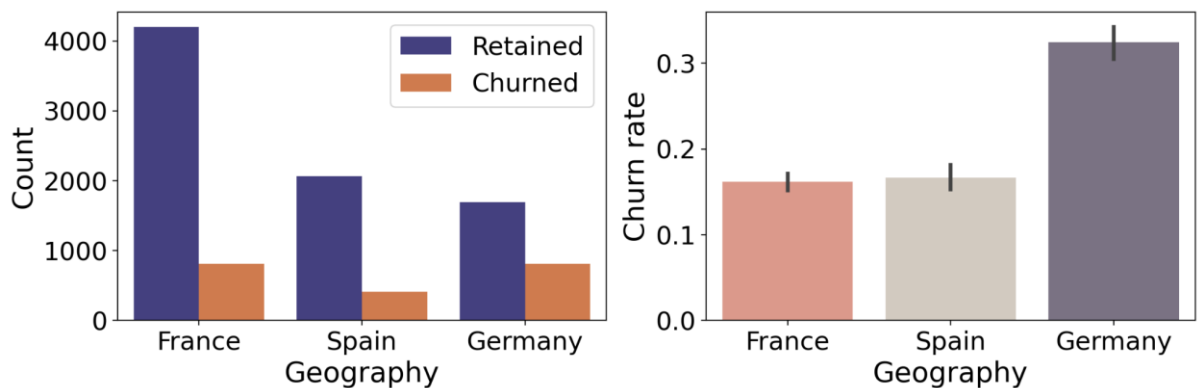


Figure 5.19 Correlation of Geography and Churn

The histogram at figure 5.19 shows that customers in Germany are more likely to leave than customers in the other two countries (the churn rate is almost twice that of Spain and France) . Many reasons could explain this finding, such as higher competition or different preferences of German customers. In Germany, banks are considered to have high stability and transparency, and the deposit insurance system

DATA EXPLORATION

is rated as the best in Europe, maybe that is the reason why customers in Germany favor the bank. customers in their own country and have a high probability of leaving the existing bank. The bank needs to focus on surveying the market and learning about the characteristics of customers in each country.

5.3.6 Gender

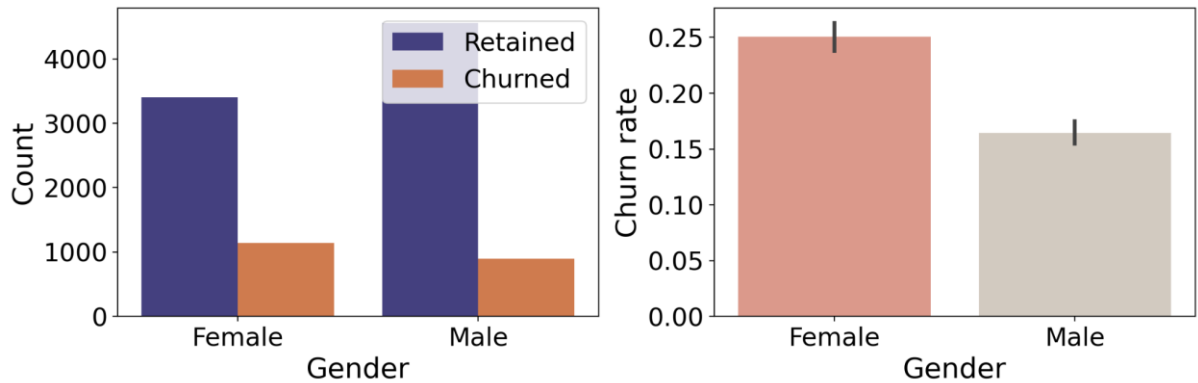


Figure 5.20 Correlation of Gender and Churn

The figure 5.20 shows that female customers are more likely to leave than men. Although there has been no official research on this issue, we can however explain that female customers have high requirements for service quality as well as purchase decisions based on emotions, so when feeling If something goes wrong, they will quickly decide to leave.

For men, they are more interested in financial factors, interest rates, risk or safety are also carefully analyzed by them, so they may be less likely to leave than women.

5.3.7 Tenure

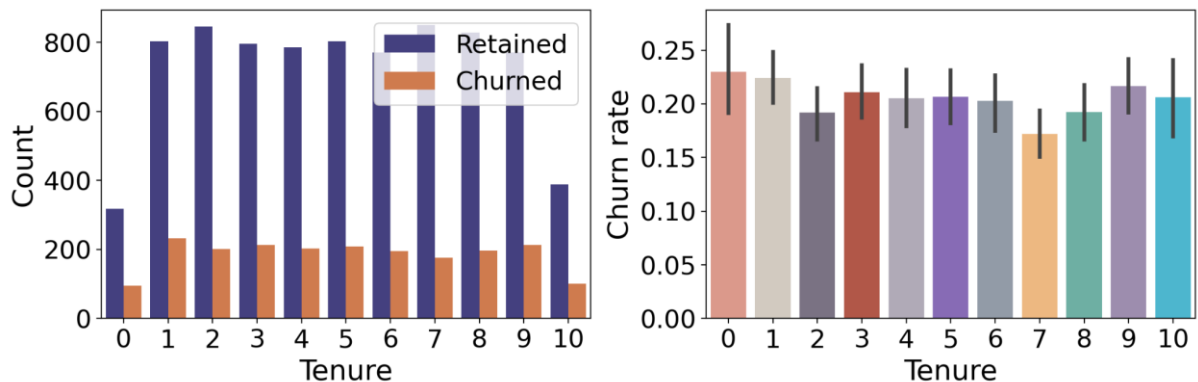


Figure 5.21 Correlation of Tenure and Churn

DATA EXPLORATION

The graph at figure 5.21 shows that the number of years of banking service has almost no effect on the churn rate. However, we can clearly see that the customer churn rate from the first year is higher than the remaining years. When they encounter problems from the start, customers will quickly leave, in contrast to long-term customers - they believe that the bank is right for them and their usage habits lead to lower churn rates.

5.3.8 Number of Products

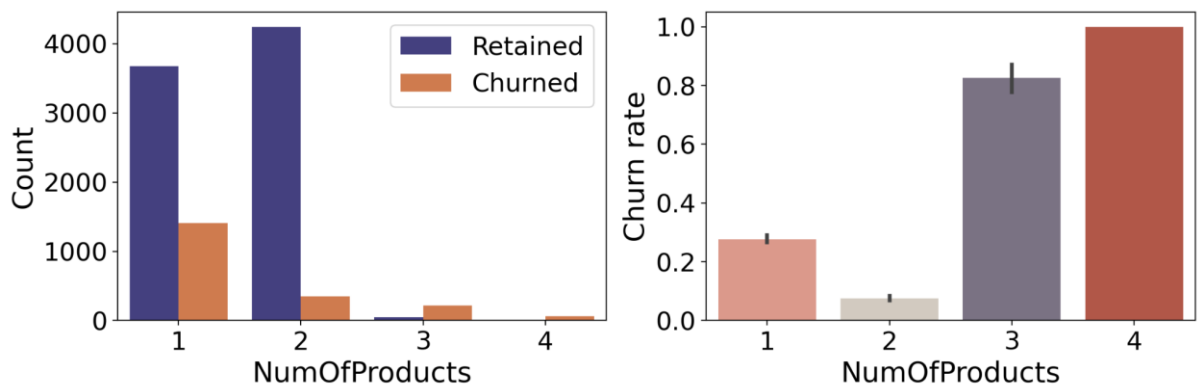


Figure 5.22 Correlation of Products and Churn

According to figure 5.22, there are many customers who only use 1-2 services. This shows that the bank does not do well in providing many of their services to customers. In particular, the rate of leaving in the file of customers using many services is very high. The reason for this may be that the bank made customers unsatisfied when using many services at the same time; This may be a process problem or the bank's strengths have a difference leading to a bad experience when customers use other services, thereby affecting customer departure. Banks also need to consider whether using multiple services at the same time causes conflicts or obstacles for customers.

DATA EXPLORATION

5.3.9 Card Holders

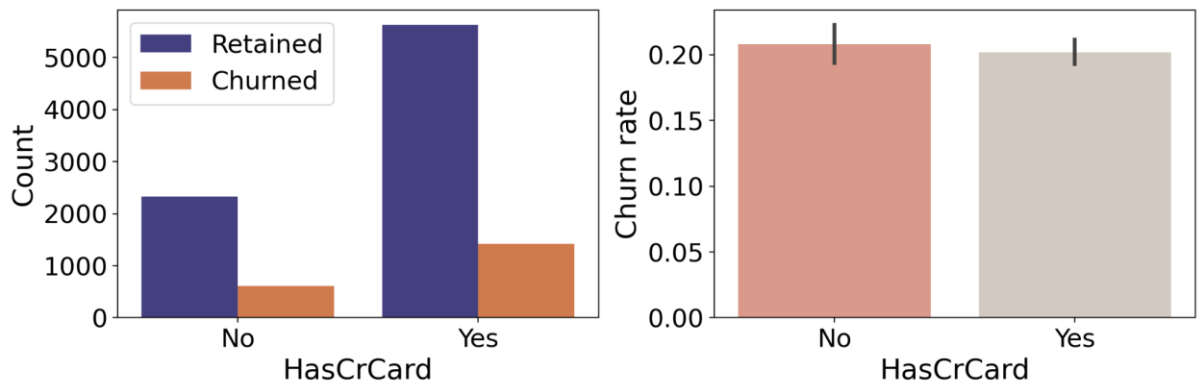


Figure 5.23 Correlation of Credit Card and Churn

The graph in figure 5.23 shows that whether or not a customer has a credit card doesn't seem to affect customer churn. This may indicate that the bank's credit service is not having serious problems.

5.3.10 Active Members

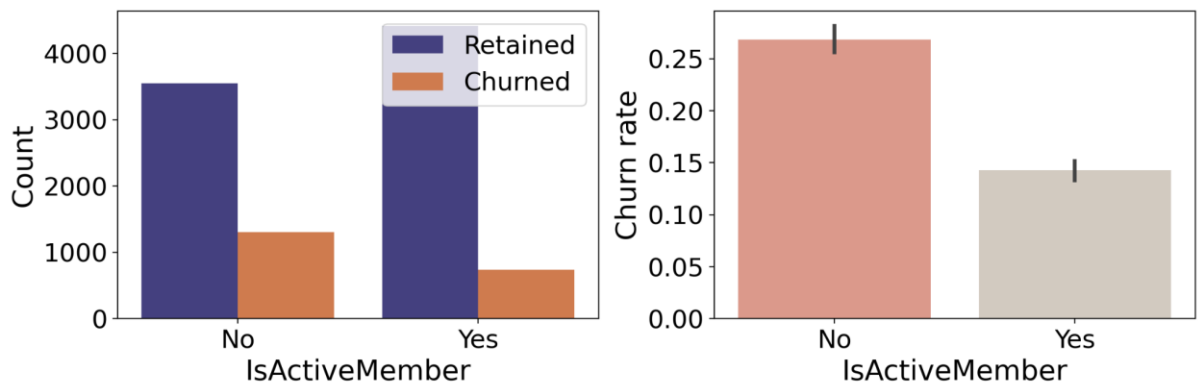


Figure 5.24 Correlation of Active Member and Churn

Looking at the figure 5.24 we see that inactive customers account for a much higher abandonment rate, which is understandable when these two values are closely related. Customers are inactive for many reasons, it may be because they no longer have the need or do not receive the attention and encouragement from the bank, or they have bad experiences that lead to them leaving the bank. Therefore, banks need to come up with policies that are more attentive to customers and promote their activities to become more active. This is a really important factor and greatly affects the churn

DATA EXPLORATION

rate of customers, once it is no longer active, most customers will not want to continue using the bank's services.

However, the active customer base is still more likely to leave the bank, so the bank needs to look at the customer experience and satisfaction to find out why they leave and fix it.

5.4 Correlations Analysis

5.4.1 Correlations Matrix

We have shown the correlation matrix between continuous variables in figure 5.25 through the `corr()` function.

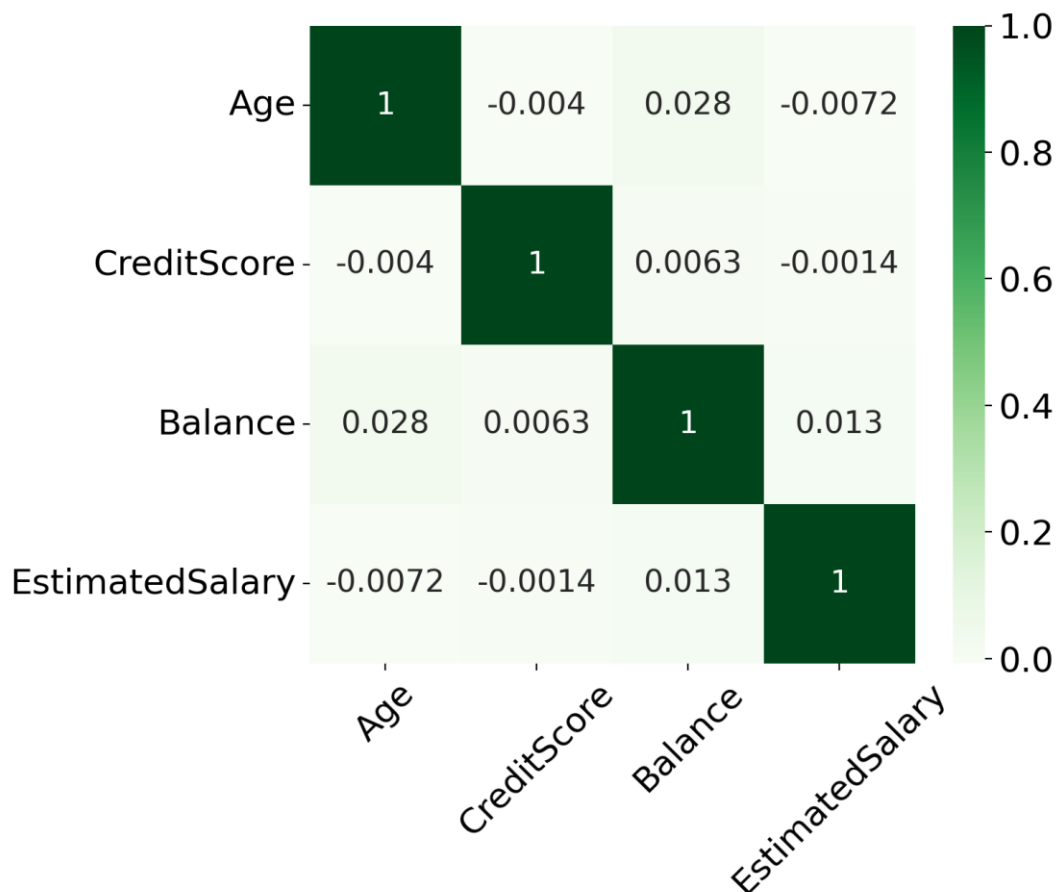


Figure 5.25 Correlation Matrix

When plotting correlation matrices, we are particularly interested in highly correlated values to find predictive models or find important variables. However, high correlation is not always good, variables with high correlation values but no theoretical value can lead to multicollinearity.

DATA EXPLORATION

Looking at the correlation matrix in figure 5.25, we can easily see that there is no significant correlation between the variables. This reassures us about the reduction of multicollinearity. Variables with low correlation should be included in the model and considered in detail.

5.4.2 Correlation between Credit score and EstimatedSalary

- Correlations coefficient: Two variables are negatively correlated, when the CreditScore variable increases, the EstimatedSalary variable decreases and vice versa.

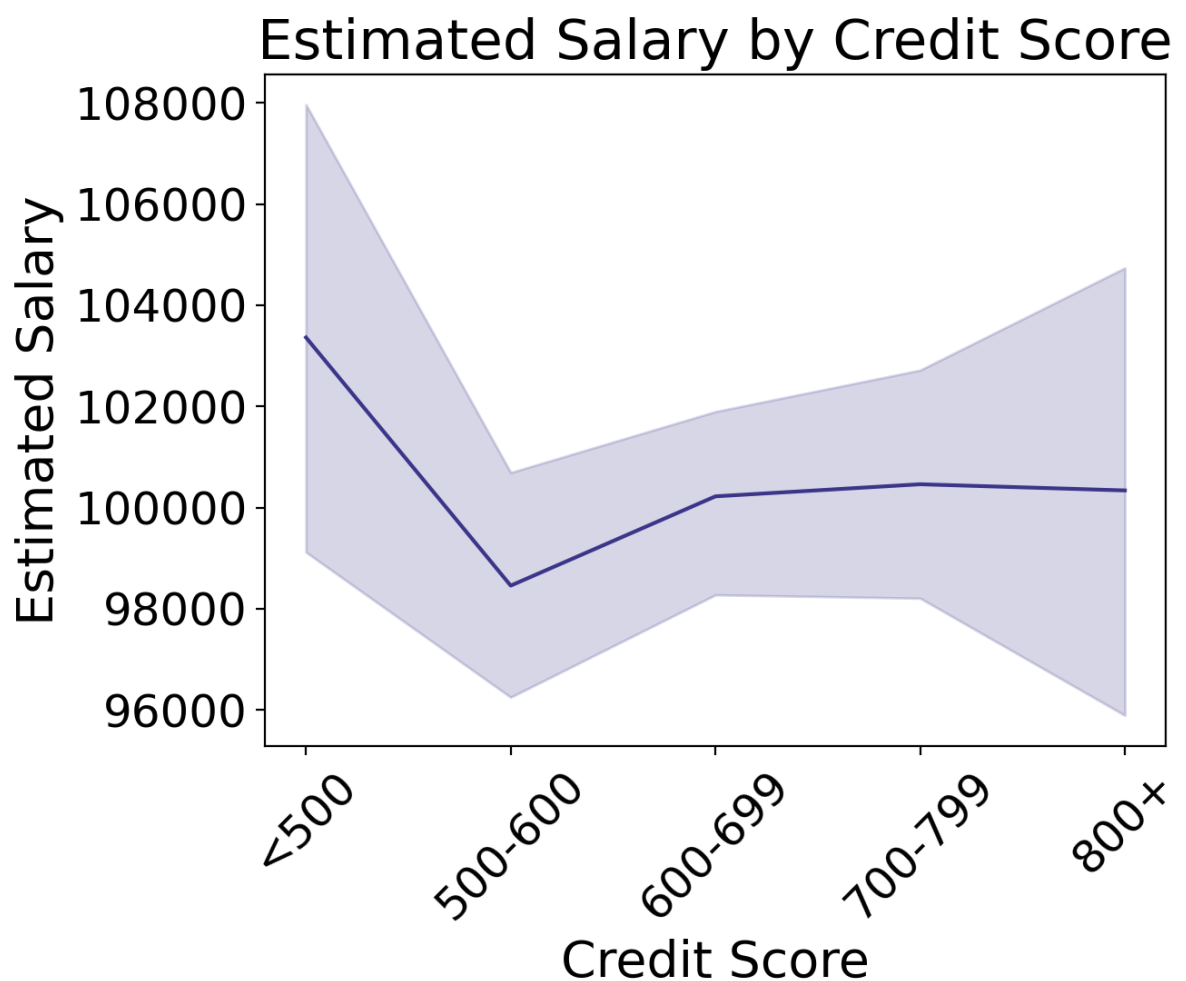


Figure 5.26 Correlation between Credit score and EstimatedSalary

Figure 5.26 is a line plot that compares Customer's CreditScore and EstimatedSalary, the team has divided Credit Score into 5 main ranges which are <500, 500-600, 600-699, 700-799 and finally 800 +. Through this chart, it can be seen that the lower the

DATA EXPLORATION

customer's credit score, the higher their salary trend, in fact, for the group of customers with the cumulative score <500 , their EstimatedSalary is 103350. Then, The customer group with accumulated points from 500-600 starts to have a lower salary than other groups, typically the salary of this group is lower than 100000, which is 98450, while the other groups have a salary above 100000.

High credit score >700 why high credit score but low salary? This can be explained by the fact that the customer can borrow less from the bank or maybe the customer pays on time and has no bad debt related to the bank or it can also be due to the customer's payment. debt to the bank before maturity as well as the term of the loan contract.

5.4.3 Correlation between Credit score and Age

- Correlations coefficient: Two variables are negatively correlated, when the CreditScore variable decreases, the Age variable increases and vice versa.

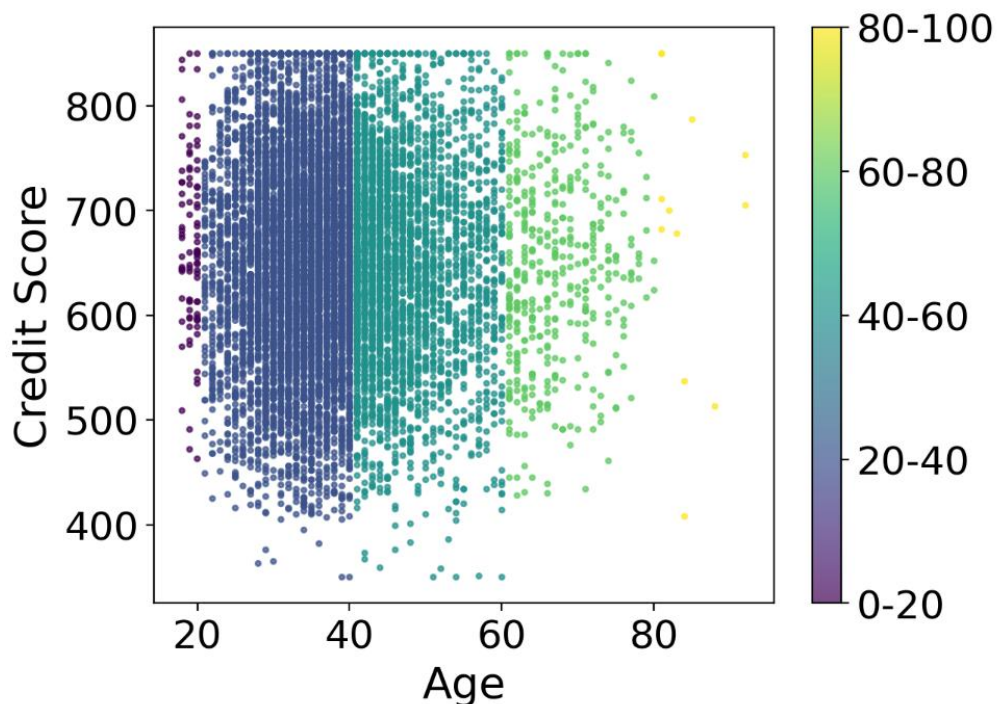


Figure 5.27 Correlation between Credit score and Age

Looking at the comparison chart between the two variables Credit Score and Age in figure 5.27 we see a clear difference between the age levels in the credit score. High credit scores are spread across 18-40 years old and gradually lower at 60 years old

DATA EXPLORATION

onwards. However, low credit scores (300-579 according to Fico's credit scale) also account for a high proportion in this age group. Due to the high concentration of customers in the age group from 3-545, the diversity in credit scores is understandable. It can be clearly seen that the credit score at the age of 35-45 is the highest among the age groups because this is the age of solid financial ability - one of the important scales to assess credit score.

5.4.4 Correlation between Credit score and Balance

- Correlations coefficient: Two variables are positively correlated, when the Balance variable increases, the Credit score variable increases and vice versa.

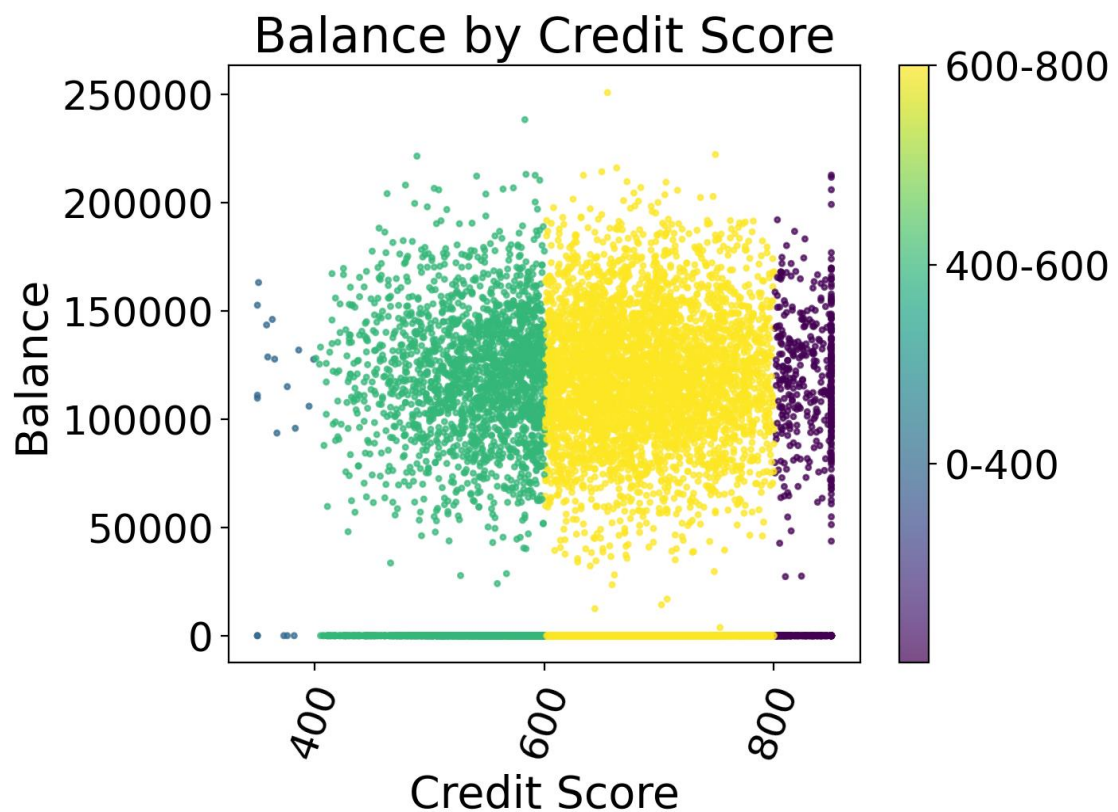


Figure 5.28 Correlation between Credit score and Balance

After the team visualized comparing CreditScore and Balance using the Scatter Plot chart, the team divided the Credit Score into 5 main ranges which are <500, 500-600, 600-699, 700-799 and finally 800 +. The team has drawn the following problems

DATA EXPLORATION

about these two columns of data. We can see the distribution of individual data points on the figure 5.28. This shows the extent of customer disparity across different balance and credit score combinations. Customers with the highest and thickest allotted account balances between 100000 and 150000 will have a credit score of about 500-700 points. From there, the bank has a strategy to accumulate points as well as a positive assessment of credit scores for customers with larger account balances.

5.4.5. Correlation between Age and Balance

- Correlations coefficient: Two variables are positively correlated, when the Age variable increases, the Balance variable increases and vice versa.

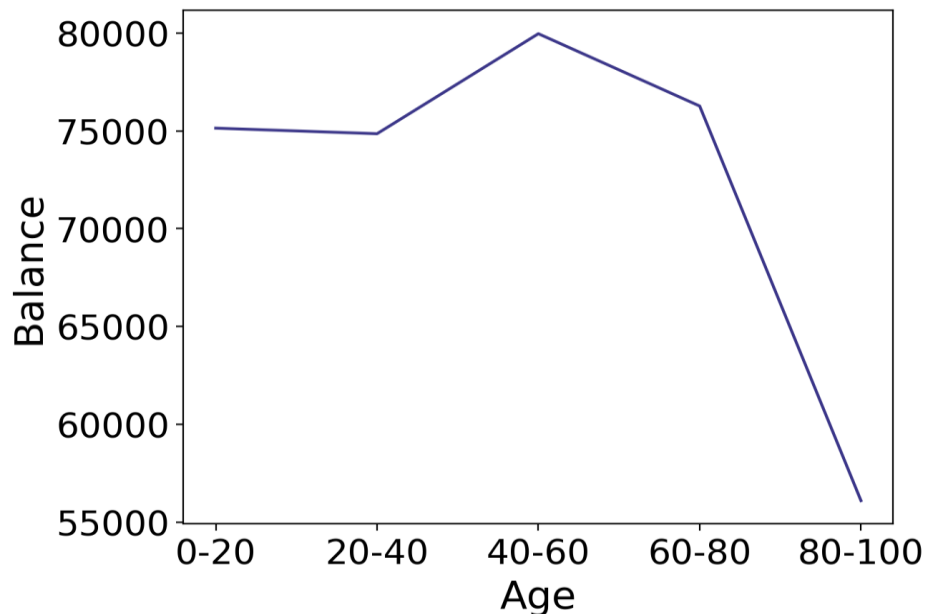


Figure 5.29 Correlation between Age and Balance

Looking at figure 5.29 we make the following comments:

Similar to the credit scores of age groups, account balances also reflect a customer's financial ability. For young customers, they already have a good balance but not the highest because this is not the time when they create the most assets. With the age from 40-60, this is the customer with the highest balance because of its financial stability and long-term needs in life. The balance is proportional to the age from 20-60 years old. Banks need to focus on this customer file to advise on savings, insurance, and investment services. And finally, the age after 80 is the age with the

DATA EXPLORATION

lowest account balance because their accumulated account is not much and decreases over time.

5.4.6 Correlation between Age and EstimatedSalary

- Correlations coefficient: Two variables Age and EstimatedSalary are negatively correlated with each other

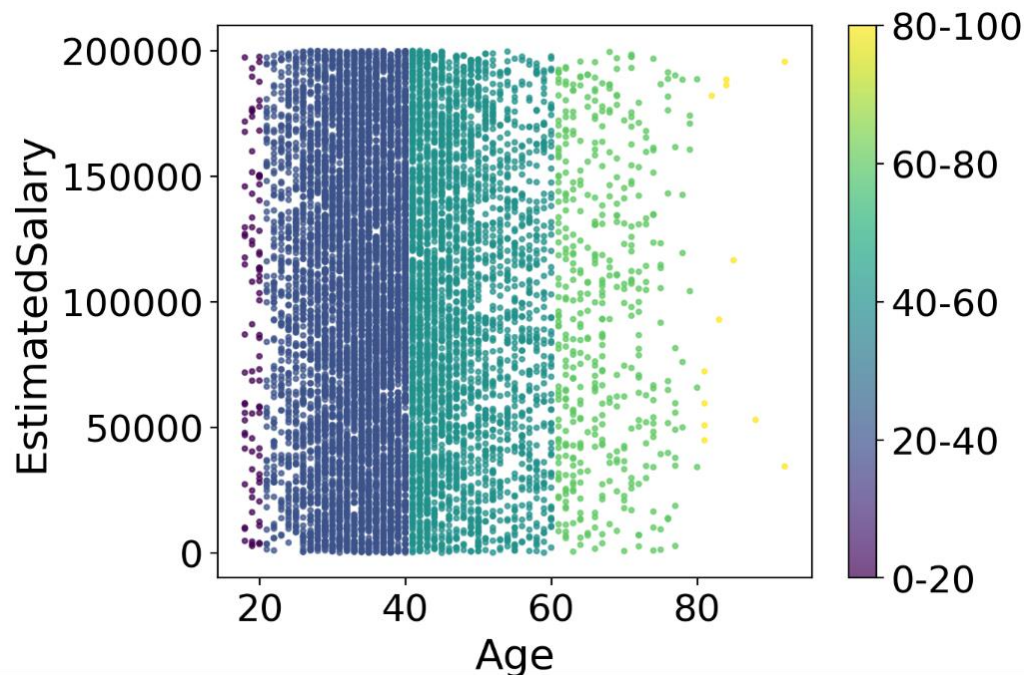


Figure 5.30 Correlation between Age and EstimatedSalary

Figure 5.30 shows the estimated salary at each age with a diverse range from high to low, but it can be seen that in the customer file 18-24 years old, the salary is not too high, this is the time. The point is that they are investing for their future direction, so they don't have a good income yet. After stabilizing the job and making certain progress, the age of 40 will be the age with the highest income. Income decreases markedly after the age of 60, when entering retirement age they will no longer have a high source of income but instead deal with old age health problems. Banks need to focus on the income characteristics of customers to assess their ability to pay as well as the potential to use other services.

DATA EXPLORATION

5.4.7 Correlation between Balance and EstimatedSalary

- Correlations coefficient: Two variables are positively correlated, when the Balance variable increases, the EstimatedSalary variable increases and vice versa.

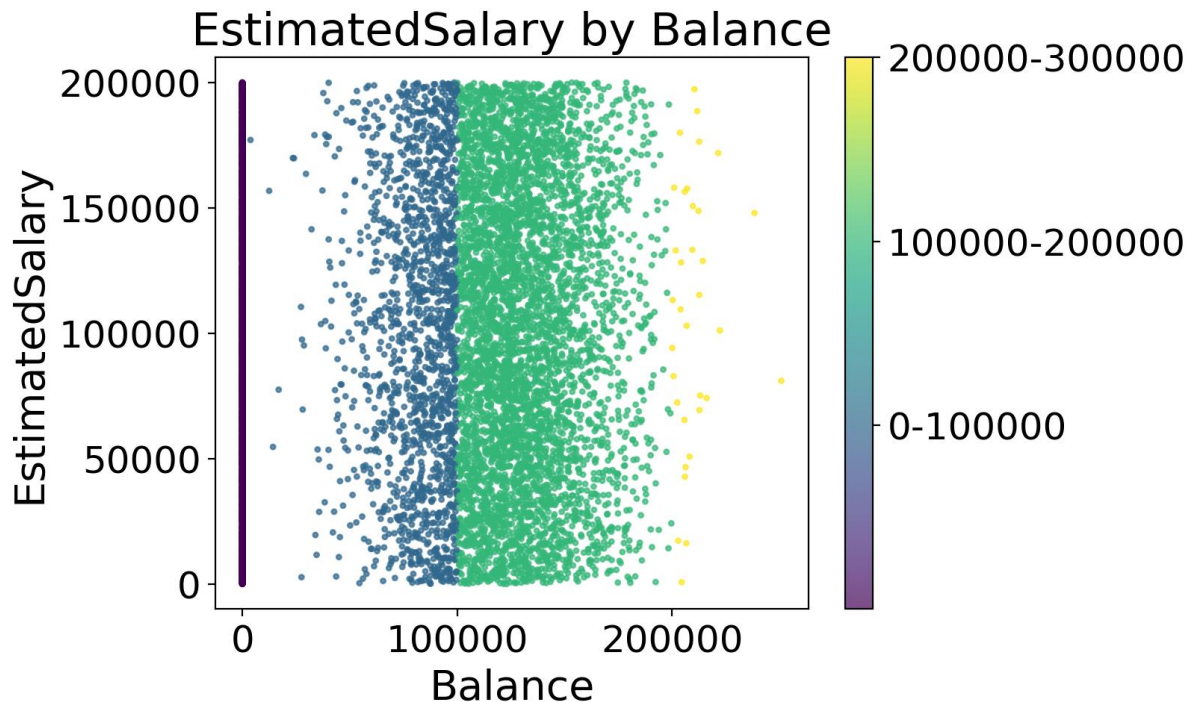


Figure 5.31 Correlation between Balance and EstimatedSalary

Looking at the figure 5.31 we clearly see that higher income corresponds to a higher average balance.

There are many studies on the negative correlation between balance and income because it is reasoned that many people tend to spend more money as income increases, leading to a decrease in account balance. However, depending on the characteristics of the population group and consumption habits, this explanation is not always correct. Especially for a bank account, only when having a good income, will users reach a high balance and use it with many financial goals, in contrast to low-income people, the need for a bank Theirs is not too high.

It is possible that users tend to manage their personal finances better and save more as their income increases. This helps them proactively manage their personal

DATA EXPLORATION

finances, and creates opportunities to invest or save money for the future. Banks can consider the correlation between these two variables to most objectively assess the financial situation of customers.

5.5 Conclusion of EDA

After performing the Exploratory Data Analysis process, we have drawn the following results:

- Technically, we have used statements to show a detailed overview of the data set, thereby making comments and choices about the variables to be analyzed. We find that salient variables have a lot of influence on target variables such as: Age; NumberOfProducts; Geography; Gender; IsActiveMember. We also use data quality techniques to ensure the best quality samples at the EDA stage. From the study of the correlation between the values with the target variable and the correlation matrix, we can choose the appropriate values to include in the model. This makes our input variables more diverse and results in a more accurate predictive model.
- On the business side, a deep understanding of the attributes of the dataset has given us interesting business insights. From there, it helps us to have an overview of the business situation of this bank:
 - The bank offers many services to diverse customer segments from many ages and countries.
 - The products and services provided by the bank are of stable quality when negative feedback such as abandonment rate or loyalty is still at an acceptable level. However, the bank also faced some difficulties in expanding many services to a single customer when the rate of customers using and giving good feedback on many services was very low.
 - The Bank has not yet taken effective measures to encourage the active activities of customers. While this is an important factor that greatly affects the business activities of customers

DATA EXPLORATION

- Finally, we can assess that the bank's business status is at a stable level, gaining the trust of the majority of customers but still facing challenges in scaling and retention. client. Currently, the competition in the financial sector is very high to attract investors, sponsors and customers, besides, banks are also affected by economic growth, inflation, politics, etc. Therefore, the bank is forced to take specific actions to maintain its position and develop better.
- Going back to the main requirement of the problem, first of all we want the bank to be able to focus on its target customers. Therefore, we have sketched out a basic customer profile of the bank with outstanding anthropometric factors so that the bank can easily approach customers and launch effective campaigns:
 - Figure 5.32 shows: Customers focus on young adults (25-40) and middle-aged adults (40-65).

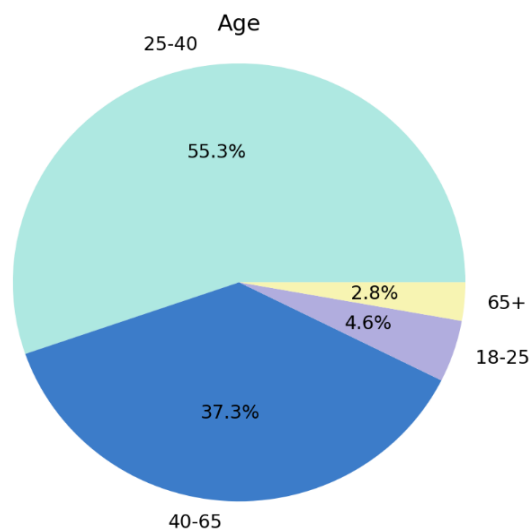


Figure 5.32 Pie chart by Age

- + Figure 5.33 shows: Customers are mainly in France with a rate of 50.1%

DATA EXPLORATION

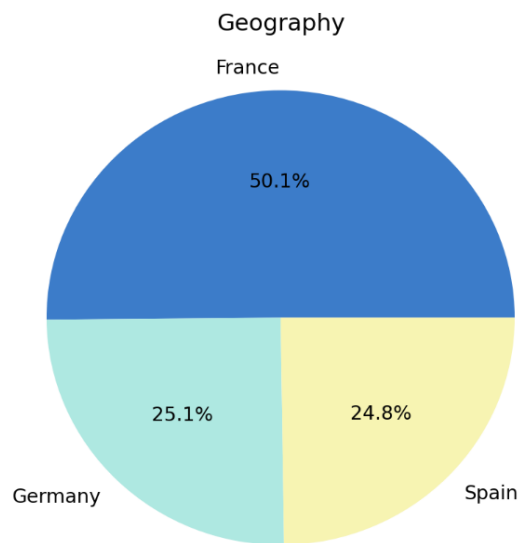


Figure 5.33 Pie chart by Geography

+ Figure 5.34 shows: Customers mainly use 1 or 2 products

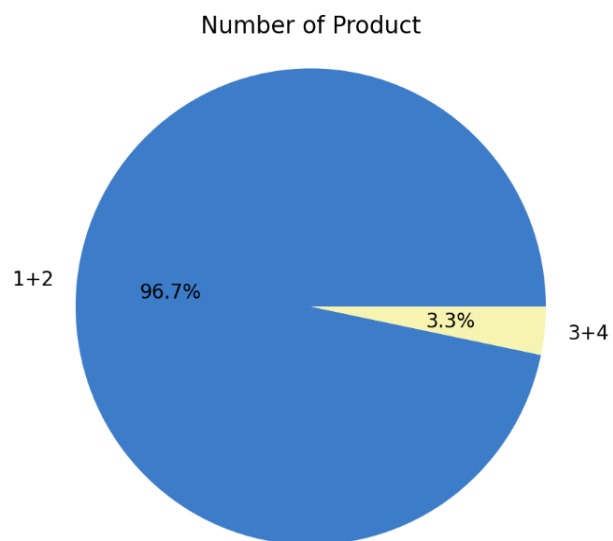


Figure 5.34 Pie chart by products

CHAPTER 6: RESULTS AND DISCUSSION**6.1 Data Preprocessing****6.1.1 Data Splitting**

During the process of building a machine learning model, splitting the dataset into two crucial parts: the training set and the test set, is referred to as the "splitting" method. This approach ensures that the model is evaluated and trained reliably and effectively.

The "Train-Test Split" method divides the original dataset into two independent subsets. The training set is used to train the model, while the test set is used to evaluate the performance of the trained model. The splitting ratio, determined by the `test_size` parameter, specifies the percentage of data that will be allocated to the test set. Depending on the scale and complexity of the model, the splitting ratio can be adjusted to suit the requirements. In this study, we chose a test size of 20%, meaning that 20% of the data will be allocated to the test set, and the remaining 80% of the data will be used to train the model.

Another essential factor in the splitting process is the randomization, controlled through the `random_state` parameter. Randomization ensures consistency and reproducibility of results, ensuring that the data splitting outcome is consistent and independent of random choices. By selecting `random_state=42` as a standard value in machine learning tasks, we ensure consistency and reproducibility of the results. Using the same `random_state` value in different runs of the program will consistently produce the same data split. This ensures that the process of evaluating and comparing models and data splits is fair and reliable.

After successfully splitting the dataset into two subsets, the training set and the test set, the results are as table 6.1:

Table 6.1 Training set and Test set (authors)

	Number of Rows	Number of Columns
--	----------------	-------------------

RESULTS AND DISCUSSION

Training set	8000	14
Test set	2000	14

6.1.2 Drop columns

Drop RowNumber, CustomerId, Surname

6.1.3 Feature selection

In the process of feature selection, the team decided to drop three columns, 'RowNumber', 'CustomerId', and 'Surname', as they were found to be irrelevant and did not contribute meaningfully to predicting customer churn.

The feature 'EstimatedSalary' was also removed from the dataset since it displayed a uniform distribution across both types of customers, indicating that it did not offer any valuable insights for distinguishing between the two groups.

Furthermore, a closer examination of the features 'Tenure' and 'HasCrCard' revealed that they exhibited similar churn rates, suggesting redundancy. To confirm this observation, a chi-square test was conducted, resulting in a small chi-square value and a p-value greater than the standard cut-off value of 0.05. The results in the table 6.2 support the original hypothesis that these two features do not provide any useful information in predicting customer churn. As a result, they are considered redundant and excluded from the final feature set.

Table 6.2 Feature Significance in Predicting Customer Churn

Variable	Chi-square	p-value
NumOfProducts	1233.595	3.767e-267
Geography	230.748	7.829e-51
IsActiveMember	195.315	2.199e-44
Gender	90.173	2.183e-21
Tenure	15.197	1.250e-01

RESULTS AND DISCUSSION

HasCrCard	0.301	5.833e-01
-----------	-------	-----------

6.1.4 Encoding categorical Features

In the context of machine learning, it is a prevalent practice to ensure that all input (and output) features are represented in numerical format. Consequently, when categorical features are present in the dataset, they require a preprocessing step known as encoding, wherein they are converted into numerical values before being utilized to train the models. By encoding categorical features, the information they carry can be incorporated into the models, contributing to accurate predictions and better overall performance.

The dataset under consideration includes two categorical features that require encoding.

For the 'Gender' feature, the scikit-learn's `LabelEncoder()` will be utilized. This method maps each unique label to an integer, with 'Male' being represented as 1 and 'Female' as 0.

As for the 'Geography' feature, a manual mapping approach will be employed. Specifically, customers in Germany will be assigned a value of one (1), while customers from France and Spain will have a value of zero (0). This decision was made based on the observation that the churn rate for customers in France and Spain is almost equal and considerably lower than that of customers in Germany. Therefore, it is reasonable to encode this feature in a manner that effectively differentiates between German and non-German customers.

6.1.5 Scaling

Feature scaling is a technique commonly employed to normalize the range of features in a dataset. It is crucial to ensure that features are on a similar scale to avoid dominance of certain features due to their magnitudes.

In this analysis, the `StandardScaler()` method was chosen to perform feature scaling. By using `StandardScaler()`, features are standardized by subtracting their mean and dividing by their standard deviation. This transformation effectively centers the features around zero and scales them to have a unit variance.

RESULTS AND DISCUSSION

The features 'CreditScore', 'Age', and 'Balance' in the train_data dataframe were selected for scaling. The StandardScaler() was applied to these specific columns using the fit_transform() method, ensuring that the scaling process is consistent and based on the training data.

6.1.6 Addressing Class Imbalance using SMOTE function

Figure 6.3 shows the dataset exhibits a significant class imbalance, with one class (0 - retained) being substantially more prevalent than the other (1 - churned):

Table 6.3 Imbalance data

Exited	
0	6356
1	1644

Class imbalance is a common issue encountered in real-world tasks. It poses a challenge in classification tasks as machine learning algorithms tend to favor the majority class, leading to models that primarily predict the dominant class. Consequently, standard evaluation metrics can become misleading in the presence of imbalanced data.

In this analysis, the SMOTE ('Synthetic Minority Oversampling Technique') algorithm was selected to address this problem. SMOTE creates synthetic instances for the minority class by finding similar records and generating new synthetic records as a weighted average of the original and neighboring instances.

To apply SMOTE, the imblearn library's SMOTE function was used with the sampling_strategy set to 'auto'. The training data (X_train) and corresponding labels (Y_train) were then resampled using SMOTE to balance the class distribution. As a result, the number of instances in the minority class (class 1 - churned) was increased to match the number of instances in the majority class (class 0 - retained). After applying SMOTE, both classes now have an equal number of instances at table 6.4:

RESULTS AND DISCUSSION

Table 6.4 Balance data

Exited	
0	6356
1	6356

6.2 Machine learning modeling

6.2.1 Modeling

After the preprocessing was completed, the data were input into the proposed machine learning models for prediction such as: Logistic Regression, Random Forest, XGBoost, Decision Tree, KNN.

These models were applied to the test set data to obtain the confusion matrix. Table 6.5 and 6.6 present the performance of different machine learning models (Logistic Regression, Random Forest, XGBoost, Decision Tree, and KNN) in predicting positive and negative outcomes based on actual results.

Table 6.5 The performance of different machine learning models

Machine learning			Predicted	
			Predicted positive	Predicted negative
Logistic Regression	Actual	Actual positive	1165	442
		Actual negative	119	274
Random Forest	Actual	Actual positive	1414	193
		Actual negative	154	239
XGBoost	Actual	Actual positive	1482	125

RESULTS AND DISCUSSION

		Actual negative	167	226
Decision Tree	Actual	Actual positive	1331	276
		Actual negative	181	212
KNN	Actual	Actual positive	1269	338
		Actual negative	119	274

Table 6.6 Confusion matrix (actual, predicted) of these 5 models

Model	Accuracy_Score	F1_Score	Recall_Score	Precision_Score
KNN	0.7715	0.545274	0.697201	0.447712
Decision Tree	0.7750	0.488636	0.547074	0.441478
XGboost	0.8540	0.607527	0.575064	0.643875
Logistic Regression	0.7195	0.494139	0.697201	0.382682
Random Forest Classifier	0.8265	0.579394	0.608142	0.553241

6.2.2 Hyperparameter Tuning

The Random Search technique was employed to find the best parameters for the Random Forest model. A set of hyperparameters, including 'n_estimators', 'criterion', 'max_features', 'bootstrap', 'oob_score', 'warm_start', 'max_depth', 'min_samples_split', and 'min_samples_leaf', was specified for the search.

During the RandomizedSearchCV process, a total of 100 candidate combinations of hyperparameters were explored using 5-fold cross-validation. The scoring metric

RESULTS AND DISCUSSION

used for evaluation was 'roc_auc'. The search was performed in parallel, utilizing all available CPU cores (n_jobs=-1) for efficiency.

The best-performing estimator discovered during the search is a RandomForestClassifier with the following parameters:

- 'bootstrap': False
- 'criterion': 'entropy'
- 'max_depth': 17
- 'max_features': 'log2'
- 'min_samples_leaf': 6
- 'min_samples_split': 10
- 'n_estimators': 200
- 'oob_score': False
- 'warm_start': False

This best estimator achieved a ROC AUC score of approximately 0.9364, indicating its superior performance in discriminating between the two classes. These optimal hyperparameters were obtained after random sampling and thorough evaluation, demonstrating the effectiveness of Random Search in finding a well-performing set of parameters for the Random Forest model.

6.2.3 Model Finalizing

After the Hyperparameter Tuning process, the final model, the "Final Hypertuned Random Forest Classifier," was selected based on its performance metrics. The evaluation metrics in table 6.7 including Accuracy Score, F1 Score, Recall Score, and Precision Score, were used to assess the model's effectiveness on the test dataset.

Table 6.7 The evaluation metrics

Model	Accuracy_Score	F1_Score	Recall_Score	Precision_Score
Final Hypertuned	0.7195	0.494139	0.697201	0.382682

RESULTS AND DISCUSSION

Random Forest Classifier				
--------------------------	--	--	--	--

The "Final Hypertuned Random Forest Classifier" achieved an Accuracy Score of approximately 71.95%, an F1 Score of approximately 49.41%, a Recall Score of approximately 69.72%, and a Precision Score of approximately 38.27%. These metrics collectively demonstrate the model's ability to make accurate predictions and effectively classify instances in the dataset.

The confusion matrix further highlights the model's performance. It correctly predicted 1165 true positive (TP) instances and made 442 false positive (FP) predictions. Additionally, there were 119 false negative (FN) predictions and 274 true negative (TN) predictions.

Cross-validation results show that the final model's accuracy is approximately 86.73%, with a standard deviation of approximately 2.55%. This indicates the model's consistency and stability in performance across multiple test folds.

Accuracy	86.7296772709054
Standard Deviation	2.5468749448512185

With these results, the "Final Hypertuned Random Forest Classifier" can be considered for deployment in real-world scenarios, as it shows promising performance in classifying instances and making accurate predictions on unseen data.

6.2.4 Model Evaluation

An AUC (Area Under the Curve) value of 0.85 suggests that the model has strong discriminative power.

This suggests that the model has a high ability to distinguish between positive and negative instances, indicating its effectiveness in making accurate predictions.

Illustrated below is the ROC Curve and Area Under the Curve (AUC), which depict the performance and discriminative capability of a classification model in distinguishing between positive and negative classes.

RESULTS AND DISCUSSION

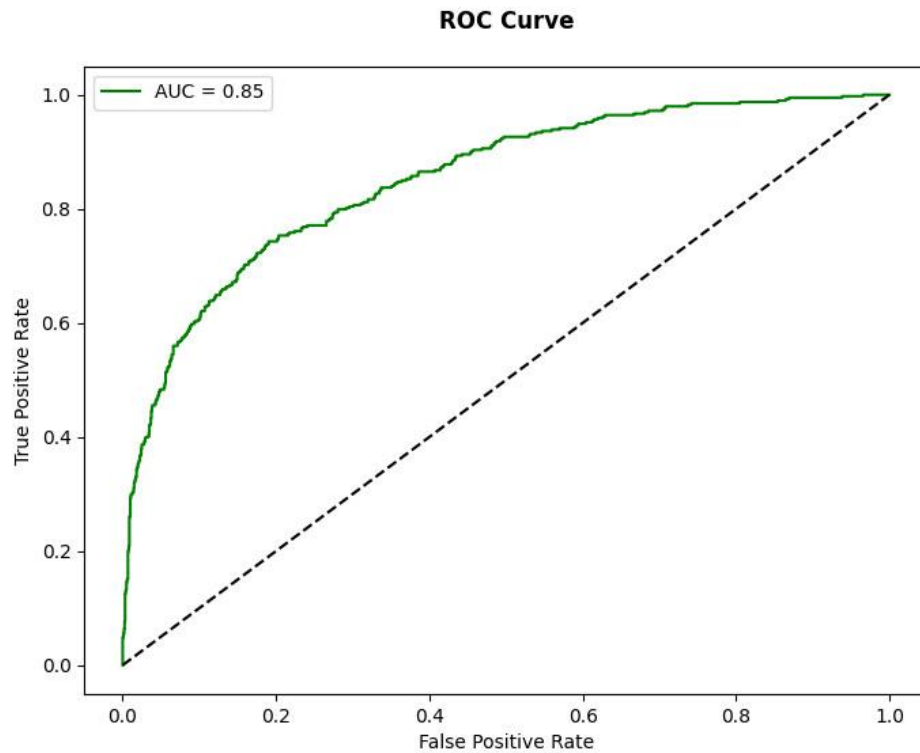


Figure 6.1 ROC Curve and Area Under the Curve (AUC)

6.3 Discussion about strategy for Banking

6.3.1 Problem

After building and defining a predictive model for a bank, we simulated a situation to better analyze the role of churn rate prediction. Let's take a closer look at the banking problem.

The director of Cactus bank realized that finding a new customer costs more and is much more difficult than retaining old customers. In particular, he also realized that the customer churn rate is at a dangerous level. What if the customer growth rate is lower than the churn rate? Surely the bank will soon encounter a crisis and bankruptcy because the number of customers is decreasing. Therefore, Cactus Bank needs to focus on customer retention! However, with current data, customer churn is what happened in the past and cannot be changed. We can't apply campaigns to departing customers, but we need to affect customers at risk of leaving in the future. By detecting early signs of customer abandonment, we can reduce churn rate.

RESULTS AND DISCUSSION

6.3.2 Strategy

We offer several strategies to improve the bank's current situation:

- First of all, even if the bank doesn't have a predictive model yet, they can control and improve the churn rate through analyzing customer churn behavior. When a customer wants to cancel a certain service, the bank certainly needs to go through 2 steps, the first step is to survey the reasons why the customer leaves and the second step is to convince the customer to stay by suggesting other options. suitable solution. If the customer is still determined to leave, the bank needs to evaluate the subjective and objective reasons. There are special reasons from the customer side such as financial factors, health, ... it is difficult for the bank to change it, but for the reasons from the bank side, they need to ask questions: Can an unsatisfied customer problem be fixed? How long does it take to fix the problem and how much does it cost? What issues should be prioritized? If the problems raised by customers can be overcome, we strongly believe that the abandonment rate will be significantly reduced for reasons from the bank.
- With current data on customer churn rate as well as the factors that are considered important affecting churn rate. Banks need to focus on those factors:
 - Older customers are more likely to leave. Therefore, banks need to come up with important strategies for taking care of elderly customers. Health and personality factors make this group of customers very disadvantaged in using the bank's services. The bank needs to keep in touch with them on a regular basis, and the employees must also be really caring, considerate and patient with them. Insurance, health, and savings services should also be suggested to older customers. In addition, the marketing campaign also needs to be subtle and appropriate, instead of digital marketing campaigns, maybe calls and sincerity will be more effective.

RESULTS AND DISCUSSION

- Customers who use multiple products have a higher risk of abandonment. The bank is having trouble expanding its services. Therefore, the bank should now pause the expansion and focus on quality issues, increasing the uniformity of quality for all services instead of just a few outstanding services. Of course, the bank has an outstanding service strategy, but the accompanying service also needs to have a stable quality for customers to consider among many factors. If there is a conflict when combining services, banks need to consider the convenience of using multiple services. Services need to be combined in a smart and convenient way. No customer is happy using 4 services with 4 completely different processes. Strategies on upselling and cross-selling also need to be focused on in the financial sector.
- Customer activity is also a very important factor. When a bank realizes that its customers are less active, the bank can implement interactive marketing, which is an important method for the bank to create and maintain relationships with customers. This includes interacting directly with customers by phone, email, text, or through social media channels. This interaction helps the bank understand customer needs and expectations, answer questions and provide personalized support. Banks can also build a strong customer community around their products/services. Create forums, chat groups, or platforms where customers can communicate with each other, share their opinions, and experiences. This will help customers to be more active.
- Balance also has an impact on customers' ability to leave, so banks need to implement strategies such as offering preferential account packages and services to customers with high balances; Create a special bonus and benefit program for customers who maintain large balances in their accounts; Consulting and providing investment

RESULTS AND DISCUSSION

solutions suitable to the financial situation of customers to increase asset value and create long-term cohesion.

- As for geography, it can be seen that the bank's market share is concentrated mainly in France and there is a negative response in Germany. Banks need to create local community events and programs to create engagement with local customers. In addition, providing additional financial services and solutions to meet the specific needs of each region. Banks in Germany, for example, are well known for their savings culture and trust in investment products. Banks can focus on developing safe and reliable savings and investment products, such as savings accounts, investment funds, and insurance solutions. At the same time, providing professional advice on investment and asset management. Investing in local advertising and media to increase brand awareness is also an important factor.
- Evaluate and predict churn rate periodically on a quarterly basis and according to the bank's special strategies.
 - A longtime customer can also leave suddenly. We can't say anything in advance, so predicting the periodic churn rate helps the bank to capture the anomalies of its customers as well as its services and products.
 - It is also important to predict and evaluate churn rate before an event or business strategy occurs. This helps the bank find the target customer file to come up with the right strategy as well as reduce costs and other resources. After the end of the campaign, the churn rate also needs to be reevaluated to check if the campaign is effective. If the churn rate drops, it means the bank has taken the right measures. Of course, we also need to pay attention and survey customer feelings to have a more comprehensive view instead of just looking at churn rate metrics.

RESULTS AND DISCUSSION

From the strategies we offer, it is hoped that Cactus Bank can improve customer retention rates as well as the quality of products and services. We believe that, if the above strategies are applied correctly, the effectiveness will not only stop at reducing the churn rate but also make the bank have more new customers. The value of the bank will also increase and become one of the reputable companies in the financial sector.

CONCLUSION

CHAPTER 7: CONCLUSION

7.1 General conclusion

In today's competitive market landscape, leveraging machine learning techniques to anticipate the likelihood of customer attrition is crucial for businesses aiming to maintain their competitive edge. By employing predictive models, organizations can tailor effective marketing strategies to mitigate potential customer loss and enhance customer retention.

The research employed a comprehensive analysis using a banking dataset, evaluating the predictive capabilities of five distinct models: Logistic Regression, Random Forest, XGBoost, Decision Trees, and K-Nearest Neighbors (KNN). Upon meticulous comparison of these models, the Random Forest prediction model emerged as the standout performer, demonstrating superior accuracy compared to its counterparts.

In essence, the study underscores the pivotal role of predictive modeling, particularly the Random Forest approach, in identifying potential customer churn in the banking industry. The insights derived from this research can aid businesses in devising strategic initiatives to retain customers and foster sustained growth in an ever-evolving market landscape.

7.2 Limited and future work

7.2.1 Limited

The study's scope has certain limitations that warrant consideration. Firstly, the analysis did not explicitly address temporal dynamics, neglecting to capture potential changes in customer behavior over time. This oversight could impact the model's effectiveness in capturing evolving churn patterns as market conditions and customer preferences shift. Additionally, the study did not comprehensively incorporate external factors, such as economic fluctuations, regulatory adjustments, or competitive market forces, which are known to exert substantial influence on customer churn. Integrating these external factors could enhance the model's predictive accuracy and provide a more holistic understanding of the dynamics driving customer attrition.

CONCLUSION

7.2.2 Future work

In the future, we could center their efforts on the implementation of sequential models, such as Long Short-Term Memory (LSTM) networks or Gated Recurrent Units (GRUs). These models have demonstrated proficiency in capturing intricate temporal dependencies, making them ideal candidates to unravel intricate patterns within customer churn behavior over time. The utilization of these sequential models has the potential to substantially enhance predictive accuracy, offering a more precise insight into the dynamics of customer churn.

REFERENCES

REFERENCES

1. Arun Velu. (2021). Customer Churn Management Using Predictive Modeling – A Machine Learning Approach. *International Journal of Emerging Technologies and Innovative Research*, Vol.8, Issue 4. DOI: ISSN:2349-5162
2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
3. Customer churn 101: What it is, why churn happens, and what you can do about it. (2021). Retrieved 07 22 , 2021, from <https://www.paddle.com/resources/customer-churn>
4. Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Morgan Kaufmann.
5. Krull, A. (2021). *Voluntary Churn Vs. Involuntary Churn / Recurly*. Retrieved 07 22 , 2023, from <https://recurly.com/blog/subscriber-retention-and-understanding-involuntary-vs.-voluntary-churn/>
6. Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*, 28(20), 597.
7. Wade, C., & Glynn, K. (2020). Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd.
8. [What is a Random Forest? | TIBCO Software](https://www.tibco.com/reference-center/what-is-a-random-forest). Retrieved 07 24 , 2023, from <https://www.tibco.com/reference-center/what-is-a-random-forest>
9. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

WORK ASSIGNMENT





WORK ASSIGNMENT

No	Full Name	Student Code	Role	Ratio
1	Trần Thị Quỳnh Nhi	K204061441	Leader	100%
2	Hoàng Thị Thanh Phương	K204061445	Member	100%
3	Nguyễn Hoàng Tính	K204061451	Member	100%

PLAGIARISM CHECKING

PLAGIARISM CHECKING

We have submitted the article for plagiarism check but have not received a response as shown below:

Tựa đề	Ngày bắt đầu	Ngày đáo hạn	Ngày gửi	Điểm có sẵn
 Check for plagiarism - Phần 1	3 Thg8 2023 - 12:51	9 Thg8 2023 - 12:00	10 Thg8 2023 - 12:51	100
Tóm tắt: Please check for plagiarism and attach proof to the final report				
 Làm mới các Bài nộp				
Tựa đề Bài nộp	ID Bài nộp Turnitin	Đã nộp	Điểm	
 Xem Biên lai Điện tử Group7_PTDL	2143354883	9/08/2023 10:12	--	 --