UNIVERSITY OF ECONOMICS AND LAW

COURSE: DATA MINING

Predict Customer Churn
Using Machine Learning In The Banking

GROUP 1

List of members of your group:

K204061445 Hoàng Thị Thanh Phương (Team leader)

K204061451 Nguyễn Hoàng Tính

K204060307 Nguyễn Trần Thúy Quỳnh

K204061450 Nguyễn Thị Huyền Thương

Supervisor:

Nguyen Thon Da Ph.D

Ho Chi Minh City, 2023

Abstract:

Customer churn, the phenomenon where customers discontinue their relationship with a business, is a critical concern in the banking sector. In an increasingly competitive landscape, understanding and mitigating customer churn is essential for sustaining profitability and fostering customer loyalty. This paper presents a comprehensive study on predicting customer churn in the banking industry through the utilization of advanced machine learning techniques. The primary objective of this research is to develop accurate and efficient predictive models that can anticipate customer churn with high precision. To achieve this, a diverse set of features, including transactional history, customer demographics, account activity, and customer interactions, are analyzed and processed. A structured dataset, extracted from real-world banking data, forms the basis for model training and evaluation. Several machine learning algorithms are employed, ranging from traditional techniques such as logistic regression and decision trees to more advanced ensemble methods like random forests and gradient boosting. Additionally, state-of-the-art techniques, including deep learning neural networks, are investigated for their potential in improving churn prediction accuracy.

The report contributes to the existing body of knowledge by systematically comparing the performance of different machine learning models in predicting customer churn within the banking sector. The experimental results demonstrate the efficacy of the proposed machine learning models in accurately predicting customer churn. Insights gained from the feature importance analysis provide valuable guidance for the banking industry to design targeted retention strategies and enhance customer satisfaction. By identifying at-risk customers early, banks can take proactive measures to retain valuable clientele and optimize their business operations.

This report underscores the significance of machine learning techniques in addressing the challenge of customer churn in the banking sector. The findings contribute to a deeper understanding of customer behavior and offer practical insights for improving customer retention strategies. As the banking industry continues to evolve, the application of advanced predictive analytics remains pivotal in maintaining a competitive edge and fostering long-term customer relationships.

Keywords: Customer churn prediction, machine learning, banking sector, predictive modeling

1. Introduction

In today's competitive business landscape, retaining customers has become a crucial factor for sustained success. Organizations across various industries invest significant resources to attract and onboard new customers, but equally important is the ability to retain their existing customer base. Customer churn, the phenomenon of customers discontinuing their relationship with a company, poses a formidable challenge for businesses seeking long-term growth and profitability.

In recent years, the advent of advanced technologies and the proliferation of data-driven strategies have paved the way for novel approaches to addressing customer churn. Among these cutting-edge methodologies, machine learning has emerged as a potent tool for businesses to predict and mitigate customer churn proactively.

This subject report aims to delve into the world of predicting customer churn using machine learning techniques. By leveraging historical customer data and powerful algorithms, organizations can gain valuable insights into customer behavior, enabling them to anticipate churn, identify potential warning signs, and implement targeted retention strategies.

The key objectives of this report are as follows:

1. *Introduction*: In the dynamic landscape of the banking industry, retaining customers is crucial for sustained success. Customer churn, the rate at which customers leave a bank for competitors, can significantly impact a bank's profitability and reputation. This report focuses on predicting customer churn using machine learning techniques in the banking sector. The introduction provides context on the importance of customer retention, outlines the goals of the report, and introduces the concept of using machine learning to address this challenge.
2. *Related Work*: This section delves into existing research and studies related to predicting customer churn in the banking sector. It explores various methodologies and models that have been applied to this problem, highlighting their strengths, weaknesses, and outcomes. By analyzing previous work, this

section sets the stage for the proposed model's novelty and potential contributions.

3. *Background*: Here, the report delves into the background of customer churn in banking, discussing key factors that influence customer decisions to leave a bank. It might touch on issues like suboptimal customer service, high fees, limited product offerings, and changing customer preferences. Understanding the underlying causes of churn is essential for designing an effective predictive model.
4. *Data Exploration*: This section involves a thorough exploration of the dataset used for modeling. It includes data sources, data collection methods, and a description of the relevant features. Exploratory data analysis (EDA) techniques are applied to gain insights into customer behavior, such as transaction history, account activity, demographics, and interactions with the bank's services.
5. *Proposed Model*: The heart of the report, this section details the machine learning model proposed for predicting customer churn. It might discuss the selection of appropriate features, the choice of algorithms (such as decision trees, random forests, logistic regression, or neural networks), and the model's architecture. The section should highlight the model's design rationale and how it addresses the specific challenges posed by customer churn in the banking industry.
6. *Experimental Result and Analysis*: In this section, the report presents the results of applying the proposed model to the dataset. It includes performance metrics such as accuracy, precision, recall, and F1-score, along with visualizations that illustrate the model's predictive capabilities. The analysis should discuss the model's strengths and limitations, including any insights gained into factors contributing to churn.
7. *Conclusion*: The report concludes by summarizing the findings and implications of the study. It emphasizes the significance of the proposed model in predicting customer churn for banks and highlights potential real-world applications. The conclusion also suggests future directions for research, such as refining the model, incorporating additional data sources, or exploring other advanced machine learning techniques.

By structuring the report around these key objectives and focusing on the specific topic of predicting customer churn using machine learning in the banking sector, you can create a comprehensive and informative document that addresses the challenges and opportunities in this critical area of business.

2. Related work

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.

1. 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.

2. 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.

3. 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 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. .

4. 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.

5. 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 SHApely 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.

6. 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.

7. 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

1. 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 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.

3. Preliminaries

3.1. Data mining

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data (Ryan S.J.d. Baker, 2010). It involves the use of statistical and mathematical techniques to discover previously unknown patterns and relationships within the data, allowing businesses to gain a competitive advantage in their industry. The data can come from a variety of sources, including databases, data warehouses, and the internet. Data mining is often used in fields such as finance, healthcare, and marketing to help businesses make more informed decisions based on their data.

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. 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 3.2 Classification of churn below:

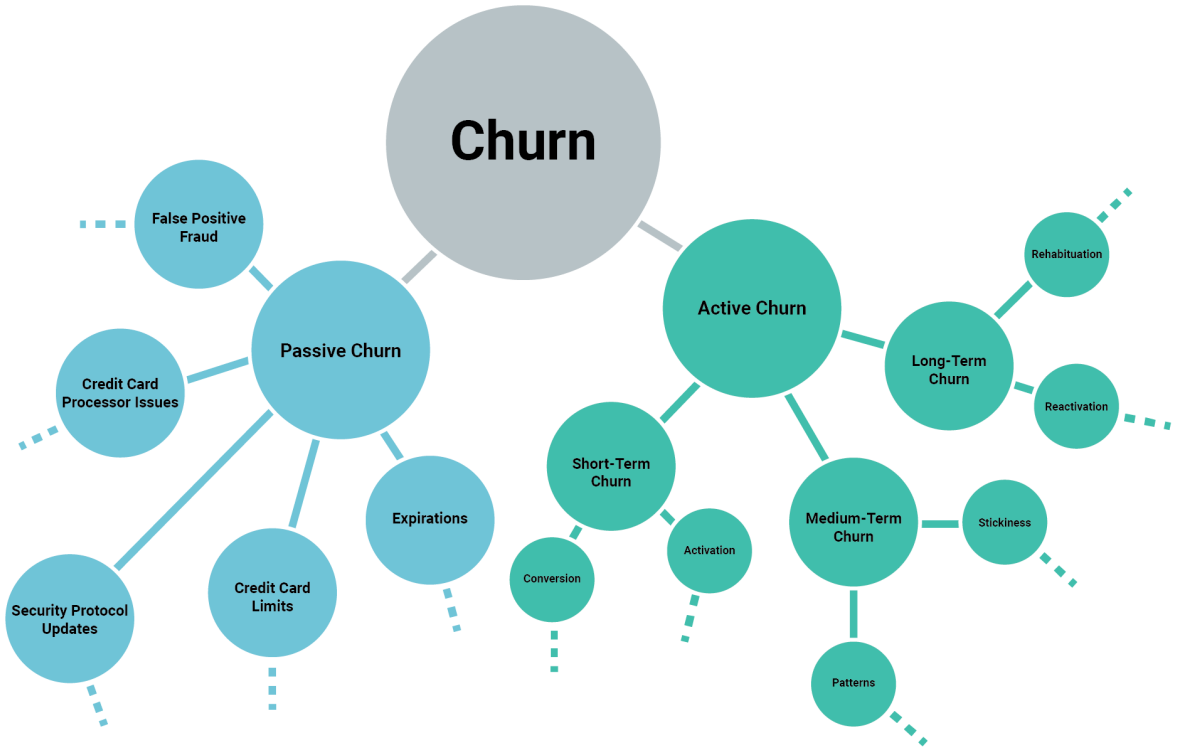

Figure 1: Classification of churn (Customer churn 101,2021)

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).

3.3. Classification algorithm

3.3.1. Association Decision Tree

Decision Tree is a popular machine learning algorithm that is used for both classification and regression tasks. It is a flowchart-like structure in which each internal node represents a decision or a test on a feature, each branch represents an outcome of the test, and each leaf node represents a class label or a predicted value.

Association rule mining is a data mining technique used to discover interesting relationships, patterns, or associations within large datasets. It focuses on identifying frequent co-occurrences or correlations between items or attributes in a dataset.

A key to our pruning is to convert the lattice of association rules into a tree of such "acting relationship" between rules. This tree is called the ADT. ADT is a tree with general rules at higher levels and specific rules at low levels, and the default rule at the root.

3.3.2. Multiclass Associative Classification

Multiclass Associative Classification (MAC) is an algorithm that reduces the number of rules discovered without drastically impacting the predictive accuracy of the classifiers.

The algorithm introduces a novel approach using the AC learning strategy for rule generation. The process involves two primary phases: rule discovery and classifier construction. During the initial phase, the MAC algorithm systematically traverses the input training dataset, focusing on rule identification and extraction. Subsequently, in the following phase, the algorithm evaluates the identified rules using the training dataset. This assessment aids in the selection of an optimal subset, which effectively serves as the representative classifier.

3.3.3. Classification based on Multiple Association Rules

Classification based on Multiple Association Rules (CMAR) is an associative classification algorithm that uses an efficient FP-tree to discover frequent itemsets and generate classification rules. CMAR is designed to be memory and space-efficient, as it consumes less memory and space compared to other associative classification algorithms. CMAR extends the concept of association rules to classification, where the discovered rules are used to classify new data instances. The algorithm first discovers a set of strong association rules that have high support and confidence, and then generates a decision tree based on the discovered rules. The decision tree is used to classify new data instances, where each internal node in the tree represents a test on an attribute, and each leaf node represents a class label.

Benefits of Classification Based on Multiple Association Rules:

- Increased Accuracy: Using multiple rules can lead to improved classification accuracy by considering a broader range of patterns in the data.
- Handling Ambiguity: In cases where different rules conflict or overlap, an aggregation mechanism can help handle ambiguity and uncertainty.

- Capturing Complex Relationships: Multiple rules can collectively capture intricate relationships between attributes and classes that might be missed by a single rule.
- Interpretability: The use of multiple rules can enhance the interpretability of the classification process, as each rule represents a distinct pattern.

CMAR brings forth substantial advancements in classification, leveraging multiple rules for accurate predictions. It introduces the novel weighted x2 technique, employs the space-efficient CR-tree for rule storage and retrieval, and incorporates the swift FP-growth approach for rule mining. These contributions collectively position CMAR as a superior option for achieving both precision and efficiency in classification tasks.

3.3.4. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a simple and popular machine learning algorithm used for both classification and regression tasks. It is a non-parametric and lazy learning algorithm, which means that it doesn't make any assumptions about the underlying data distribution and it doesn't learn a model during training. Instead, it stores all the available data and classifies new data points based on their similarity to the training data.

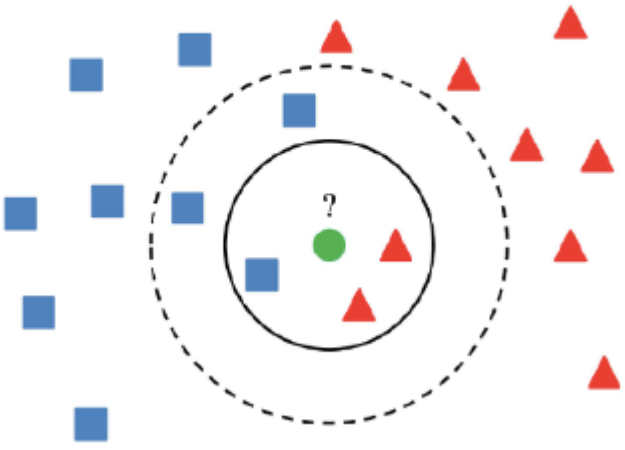

Figure 2: An example of KNN

Two crucial aspects of K-Nearest Neighbor (KNN) are worth noting:

- Initially, KNN is categorized as a non-parametric algorithm. This signifies that no presumptions are made regarding the dataset during the model's utilization. Instead, the model is exclusively fashioned from the furnished dataset.
- Furthermore, KNN sidesteps the division of the dataset into distinct training and test sets. Unlike other approaches, KNN refrains from forming generalizations

between training and testing sets. Thus, the entire training dataset is employed when the model undertakes prediction tasks.

The K-NN algorithm works as:

- Training Phase: In the training phase, K-NN stores the feature vectors and their corresponding class labels (for classification) or target values (for regression) of the available dataset.
- Prediction/Classification Phase:
  + For a given new data point (or instance) that needs to be classified or predicted, the algorithm calculates the distance (such as Euclidean distance) between this data point and all the data points in the training dataset.
  + The K-nearest data points (neighbors) with the smallest distances are selected.
  + In classification, the class labels of these K neighbors are examined, and the class label that appears most frequently among the neighbors is assigned as the predicted class for the new data point. This is often referred to as "majority voting."
  + In regression, the average or weighted average of the target values of the K neighbors is calculated, and this value is assigned as the predicted target value for the new data point.
- Choosing the Value of K: The choice of the parameter K is crucial. A smaller K can lead to noise affecting predictions, while a larger K can lead to over smoothing and ignoring local patterns. The optimal value of K often requires experimentation and cross-validation.

The initial stage of a KNN is to transform the data points into salient features, or their numerical values. Select the new data set after that, and then locate every K-neighbor and determine their distance from one another. Here are three typical calculations methods:

- Euclidean Distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Figure 3: Euclidean Distance of KNN

- Manhattan Distance

$$d(x, y) = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Figure 4: Manhattan Distance of KNN

- Minkowski Distance

$$d(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^p \right)^{\frac{1}{p}}$$

Figure 5: Minkowski Distance of KNN

K-Nearest Neighbor (KNN) offers several advantages that contribute to its popularity in the realm of machine learning. One of its primary strengths lies in its simplicity and ease of implementation. Its straightforward nature makes it an excellent choice for beginners, enabling them to quickly grasp its mechanics and apply it to various tasks. Additionally, KNN eliminates the need for an explicit training phase, as it learns from the data during the prediction process. This adaptability allows it to effortlessly incorporate new patterns without requiring retraining. Moreover, KNN's flexibility shines through as it can seamlessly handle both classification and regression tasks, proving its versatility across different problem domains.

3.3.5. L3

L3 is a lazy approach to pruning classification rules that splits the entire set of rules into three categories : harmless, useful, and harmful. The approach discards only the harmful rules from the rule base, rather than pruning the entire rule base.

By delving into L3, practitioners can fine-tune the generated rules and enhance the accuracy, interpretability, and generalization capabilities of the resulting classification model. This meticulous analysis and refinement process contributes to the overall success of the lazy pruning strategy.

The L3 algorithm has been shown to be effective in reducing the computational cost of generating and storing classification rules, while still maintaining high classification accuracy.

3.3.6. Classification Based on Associations

Classification Based on Associations consists of two parts, a rule generator (called CBA-RG), which is based on algorithm Apriori for finding association rules in (Agrawal and Srikant 1994), and a classifier builder (called CBA-CB).

Classification Based on Associations (CBA) is an algorithm used in data mining for classification tasks that employs association rule techniques to classify data. It is considered one of the most widely used algorithms in the Association Rule Classification (ARC) family.

4. Data Exploration

4.1. Overview of datasets

The team used a customer file from a bank called Kaggle to get the data set they needed to create their project thesis. A total of 14 columns, each representing a different attribute of a customer, make up the data set, which has 10,000 rows of data, or the equivalent of 10,000 customers. These clients gathered information in Spain, Germany, and France, a total of 3 nations. The Table contains a description of the values of the characteristics. The following client dataset details from the bank:

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 company	int64	0,1,2,3,4,5,6,7,8, 9,10
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

Table 4.1: The detail information of the Bank's customer dataset by the authors

4.2. Handling missing data and duplicates

4.2.1. Missing data

After inspection, we found that this dataset has no missing data, figure 6 and 7 made this very clear.

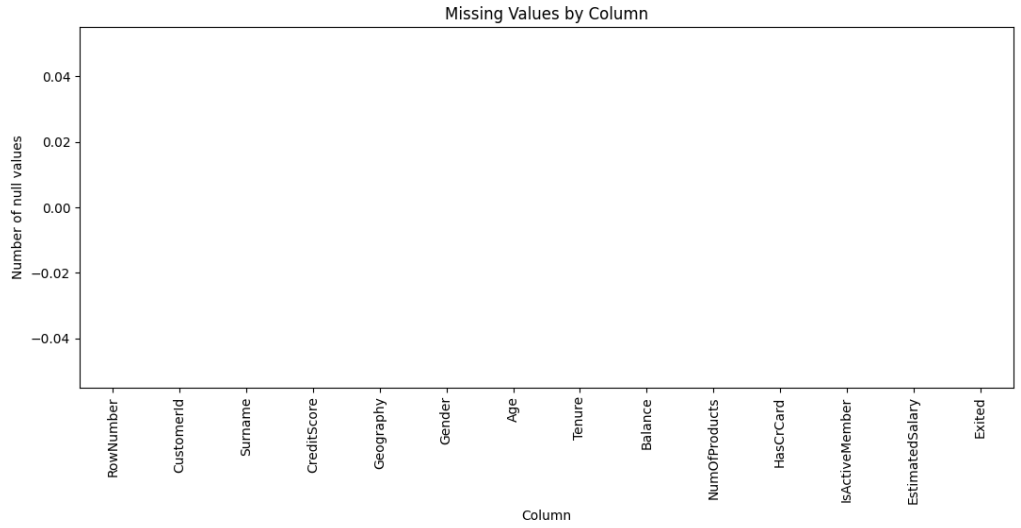

Figure 6: Missing Values by Column by the authors

4.2.2. Duplicates

This dataset also has no duplicates data, which we can see in Figure 7 below:

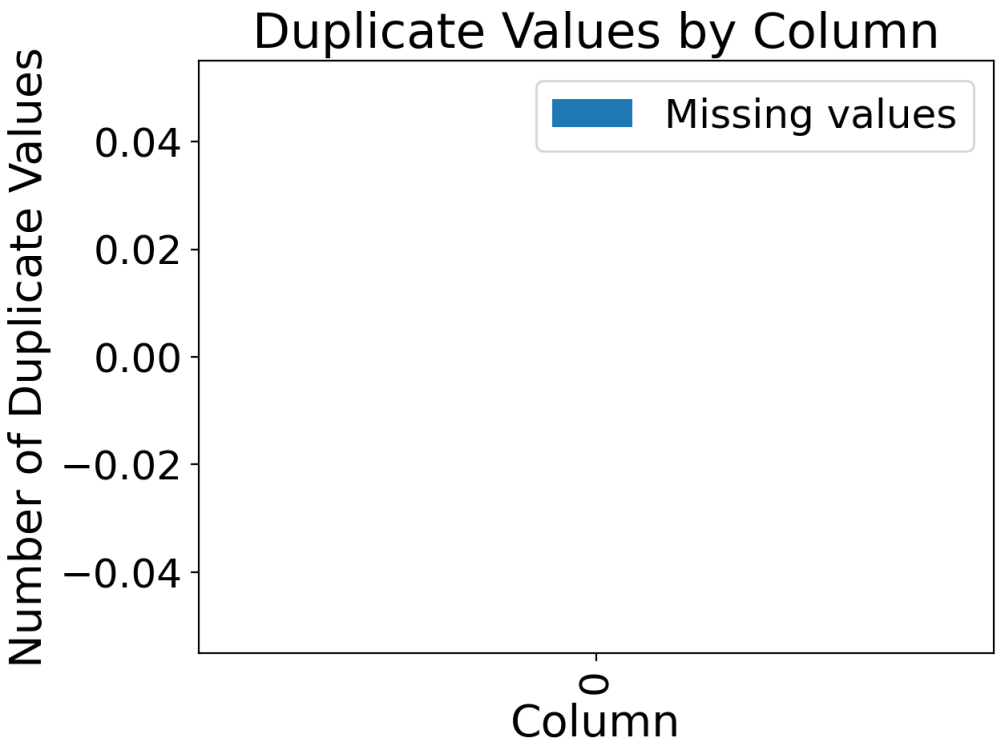

Figure 7: Duplicated Values by Column by the authors

4.3. Numerical features about the data summary

We have implemented the describe() method to give a statistics table of the numerical features in the figure 8

Figure 8: Numerical features by the authors

- To gain a sense of the characteristics of the data, we calculated numbers such as count, mean, std, min, 25%, 50%, and max. Following the computation, the following conclusions may be drawn:
- The average credit score for the consumer is 650.53, which is a typical number that is neither too high nor too low.
- Customers at the bank range in age from 18 to 92, with an average age of around 40. demonstrates an adult client base that is both diversified and heavily concentrated.
- When the average tenure value exceeds 5, the assessment of customer loyalty is pretty good.
- At this bank, 20% of clients resign and 50% of customers are still active.

4.4. Exploratory Data Analysis

To start with, it is important to differentiate between continuous and categorical variables through the utilization of numpy and pandas libraries for data verification. This process involves combining the libraries with a function that examines unique values and generates a list encompassing the categorical and continuous variables. The resulting list would be structured in the following manner:

```
categorical_vars = []
continuous_vars = []
for col in bankData.columns:
    if bankData[col].dtype == 'object' or bankData[col].nunique() <= 11:
        categorical_vars.append(col)
    else:
        continuous_vars.append(col)
print("Categorical variables:", categorical_vars)
print("Continuous variables:", continuous_vars)
```

✓ 0.0s

```
Categorical variables: ['Geography', 'Gender', 'Tenure', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'Exited']
Continuous variables: ['CreditScore', 'Age', 'Balance', 'EstimatedSalary']
```

Figure 9: Categorical and Continuous Variables by the authors

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

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

4.4.1. Continuous Variables

Age:

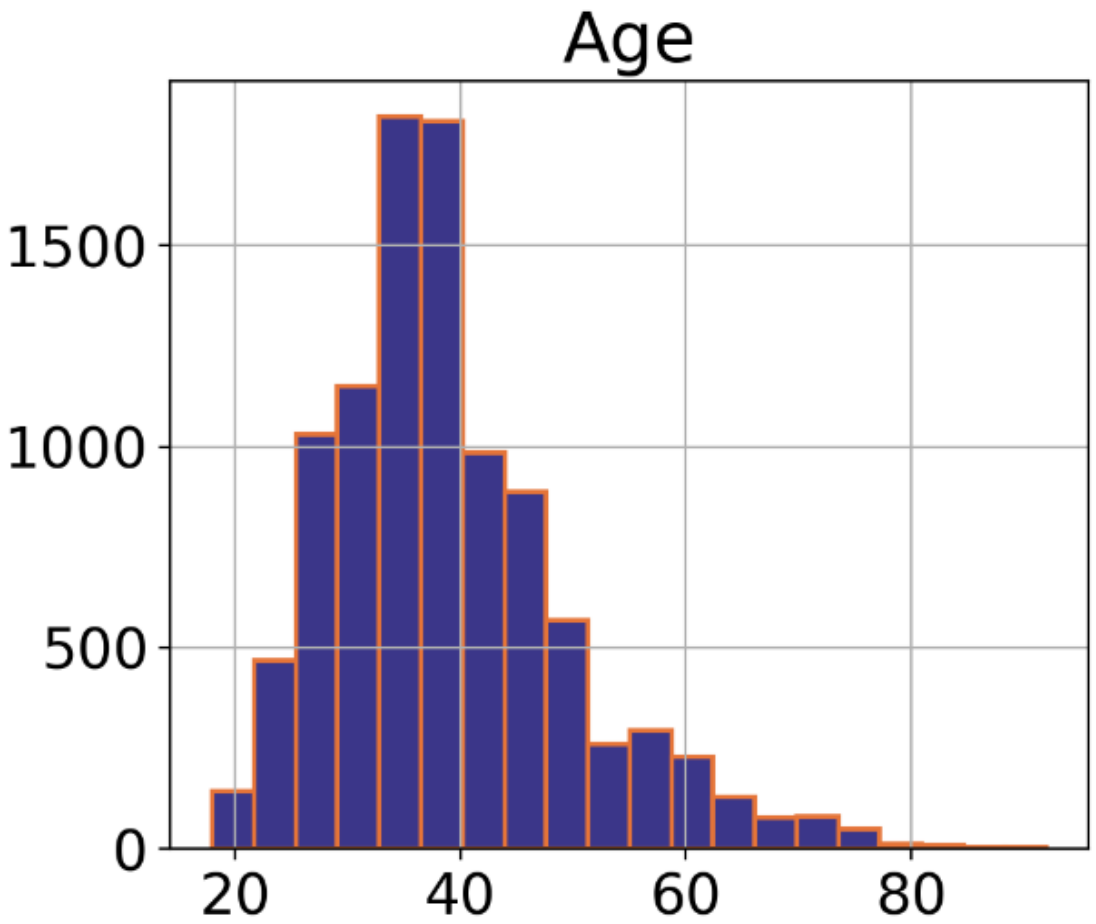

Figure 10: Histogram of Age

Looking at the figure 10 above, the population of young customers (aged 18-24) may not be substantial, but they hold significant potential as customers. Research suggests that young customers often select banks based on reputation and prioritize

safety to protect their assets. While their current financial capacity may not be robust, it is expected that they will increasingly utilize banking services in the future.

The age group with the highest concentration is between 32 and 40 years old, representing middle-aged customers who possess stable incomes and have high financial requirements such as savings, transactions, investments, and loans. They are inclined to take risks and explore additional services. This age range has accumulated sufficient experience and financial resources to make more daring decisions in order to attain greater returns. Moreover, these customers tend to spend considerably on their families, including properties, vehicles, homes, and children.

The graph exhibits a right-skewed distribution, indicating a significant increase in the proportion of customers in older age groups compared to younger ones. This might indicate that a large number of customers have been using the bank's services for an extended period and are likely to continue doing so in the future. In Europe, the average retirement age is 65, which is typically associated with reduced activity in utilizing banking services due to various factors such as expenses, health concerns, familial responsibilities, and adherence to traditional lifestyles. Nevertheless, with the increasing popularity of banking services and the challenge of an aging population, we believe that this demographic still represents potential customers for the bank, particularly if the focus is on addressing their basic needs and providing convenience.

Credit score:

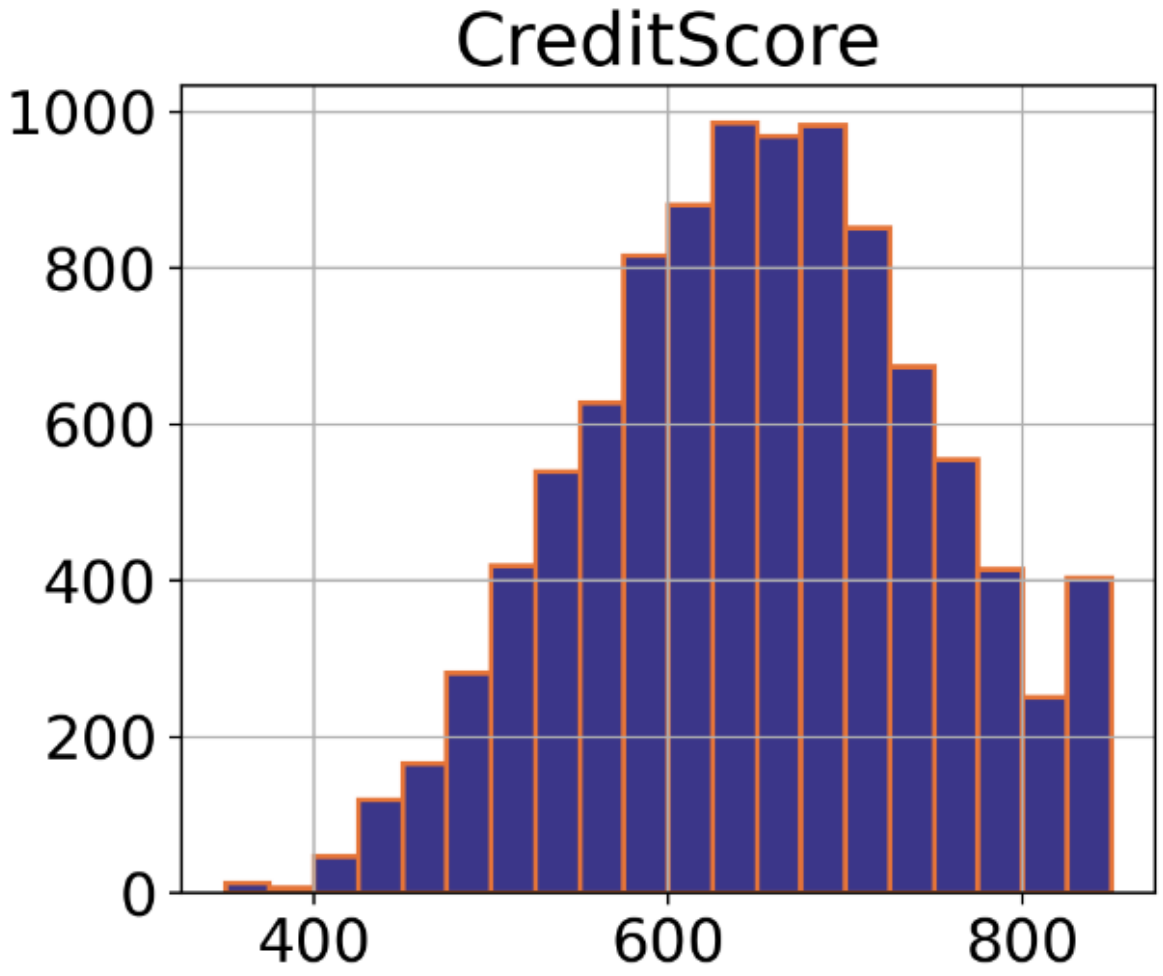

Figure 11: Histogram of Credit Score

Figure 11 shows a high percentage of credit scores from average and above (650 points). This demonstrates how well the bank can manage credit risk when a customer's file has a consistent credit score. Banks are very careful when evaluating consumers, in part because of the impact of specific rules in Europe on safeguarding the stability of banks following the 2008 crisis.

Customers with below-average credit ratings frequently experience difficulties in obtaining loans and other financial services, as well as higher interest rates. Due to laws on credit ratings, this not only secures the security of banks but also results in significant client losses. The variance of credit ratings for this bank, however, ranges from 350 to 850, demonstrating that not all of their customers have excellent credit. In other words, the bank offers a wide range of services to enable it to attract a large number of clients. While credit score is a crucial consideration, the bank nevertheless accepts clients with low scores.

Balance:

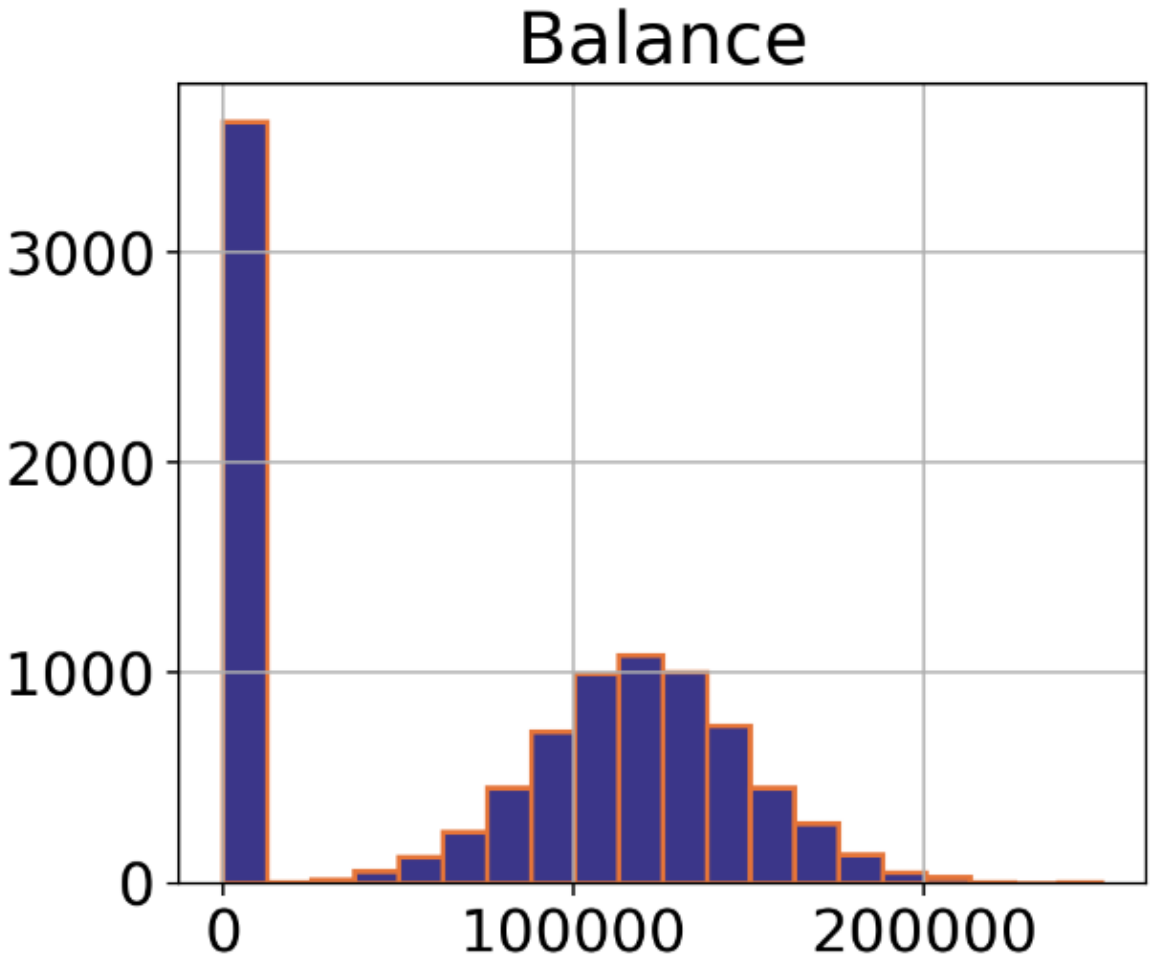

Figure 12: Histogram of Balance

Figure 12 shows that the zero remainder is very much, but if this is omitted, the balance is fairly normally distributed. The reason for some accounts having a zero balance could be attributed to customers who have opened the account but have not actively used it. Additionally, the overall number of active customers might not be significantly high, which explains this occurrence.

The account balance reflects the funds deposited by the customer or the amount borrowed from the bank. Therefore, the balance not only affects the bank's credit but also its solvency. The bank can analyze the balance to assess the customer's repayment capacity and to evaluate the likelihood of customer churn.

Moreover, the bank has the opportunity to provide tailored policies and services to customers with outstanding balances, ensuring that their needs are met appropriately.

Estimated salary:

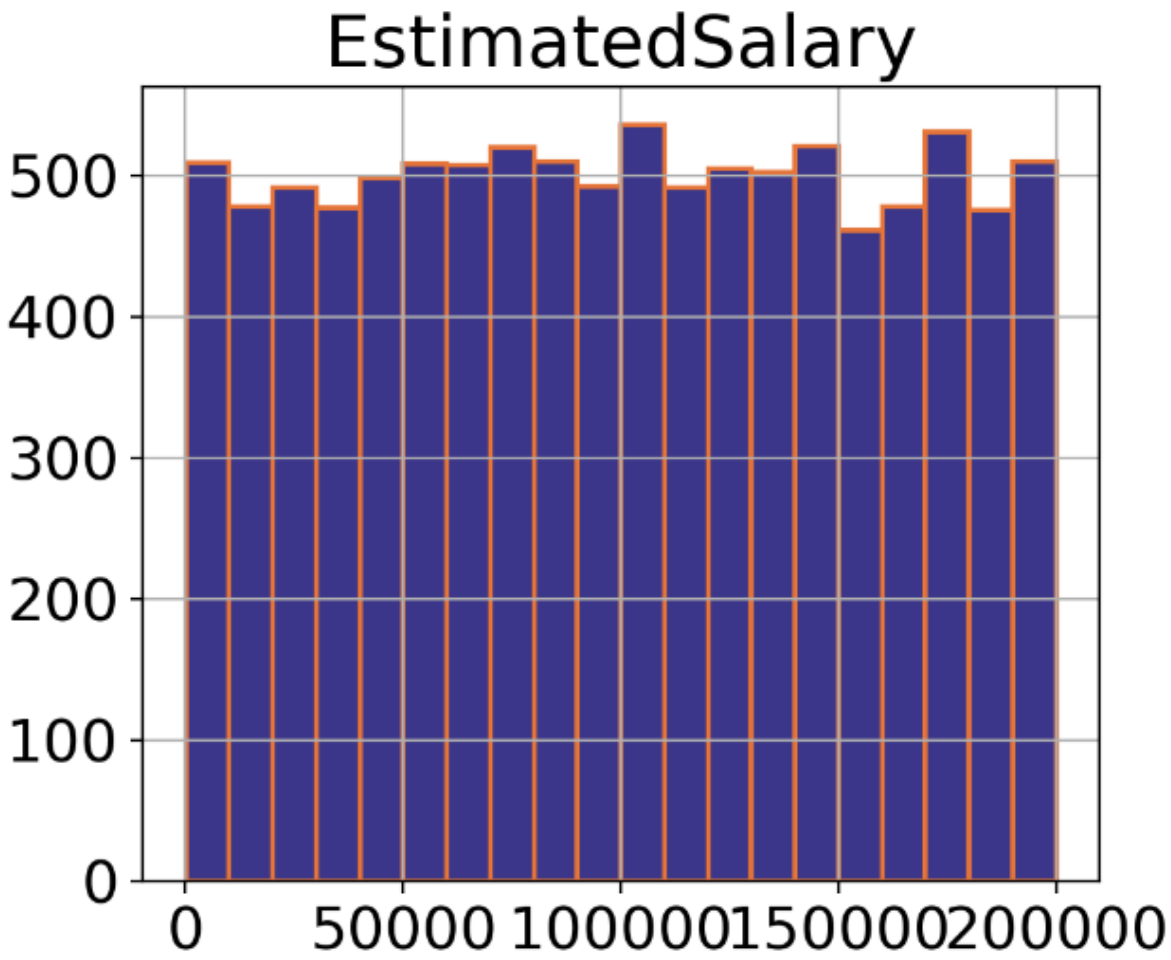

Figure 13: Histogram of Estimated Salary

Looking at the figure 13 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.

Individuals with higher incomes typically have more opportunities to save and invest compared to those with lower incomes. Additionally, they often require loans for purchasing assets like houses, cars, or for business investments. These individuals also tend to have a greater demand for advanced financial products and services, such as credit cards, debit cards, and online transfers, to efficiently manage their finances.

On the other hand, individuals with lower incomes may have limited requirements for the bank's financial offerings, resulting in their exclusion from the financial system and restricted access to essential financial services.

A customer's income plays a vital role for banks in assessing their financial capacity for loans, mortgages, and credits. Furthermore, determining a customer's income enables the bank to provide more suitable services. For instance, customers with higher incomes may enjoy premium services, higher interest rates, or increased credit limits.

4.4.2. Categorical Variables

Gender:

In this case, the frequency of categorical variables is represented using the ns.countplot function from the Seaborn library. We have visualized the frequency of the gender variable in the figure 14 below:

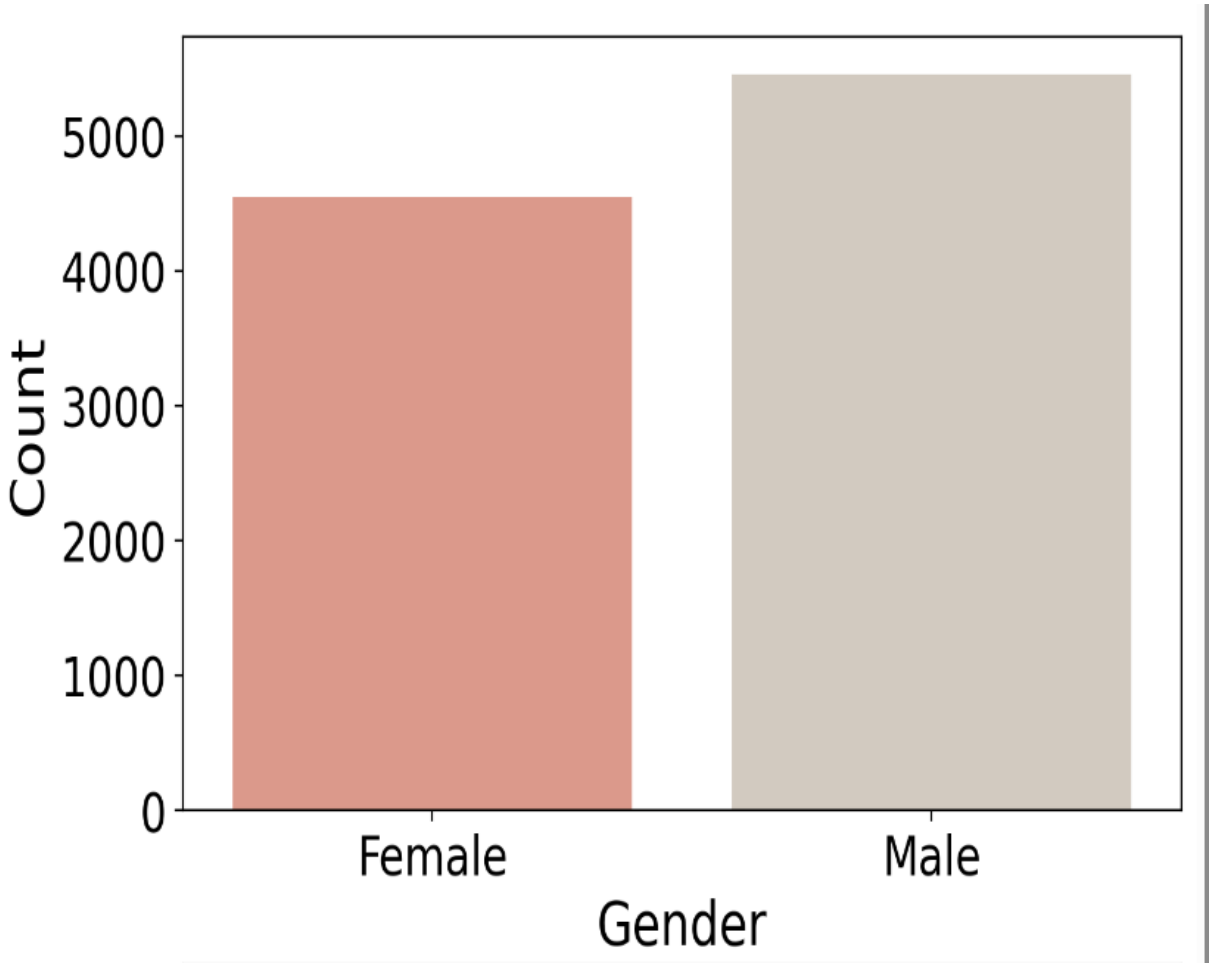

Figure 14: Histogram of Gender

The gender distribution in the provided data is not significantly skewed, although there are slightly more male customers than female customers. However, it is important to note that the utilization of banking services cannot be accurately determined solely based on gender. Nevertheless, certain usage patterns can be observed:

- Women tend to make greater use of online and mobile banking services compared to men. This trend might be attributed to women having busier lifestyles, limited time for in-person banking visits, and a greater need for flexible financial management.
- Men, on the other hand, tend to engage more with financial services related to investments and securities transactions offered by banks. This inclination might

be influenced by men typically possessing more knowledge about finance and investments, as well as a greater interest in increasing their income.

Based on these patterns, banks can consider incorporating this factor into their business strategies, advertising, and marketing efforts. This would allow them to focus on addressing the specific needs and preferences of their customers, taking into account the differing requirements of men and women.

Geography:

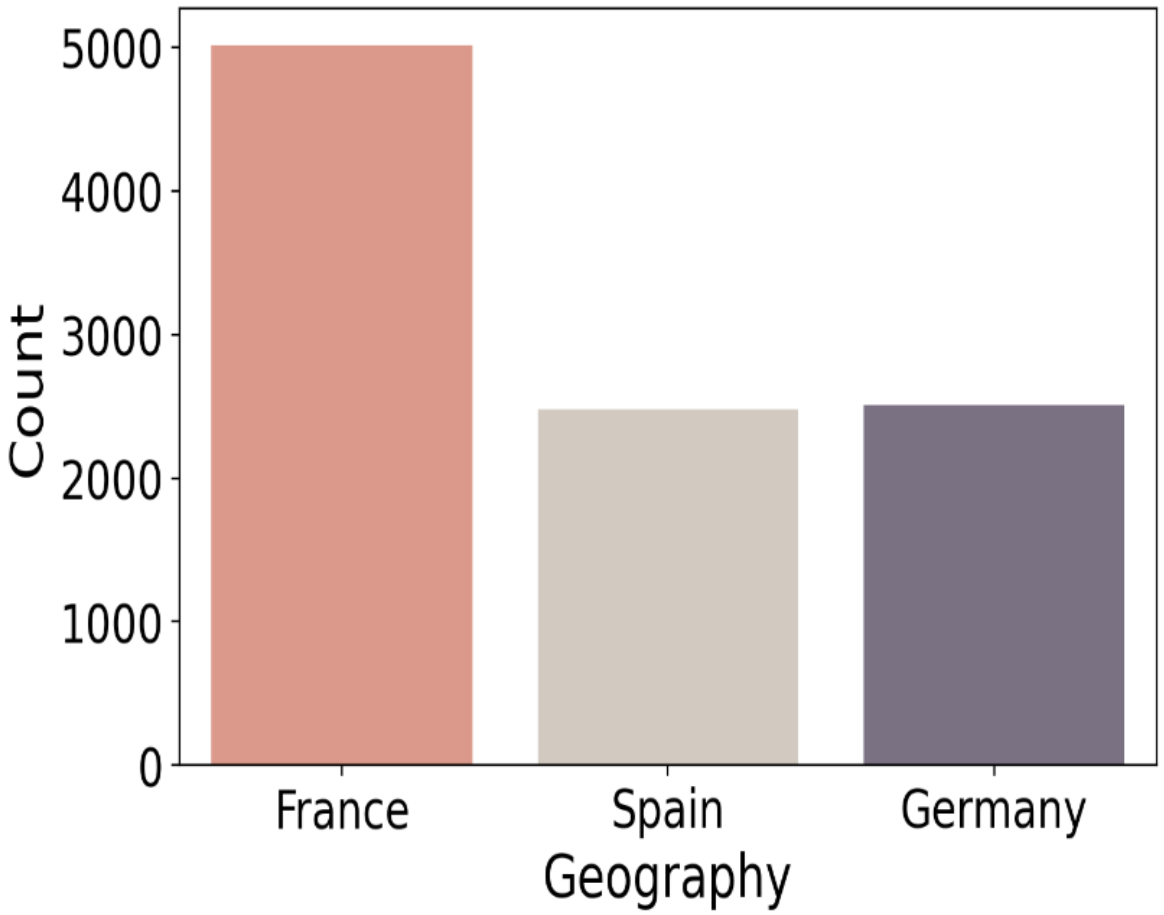

Figure 15: Histogram of Geography

Figure 15 clearly shows the bank has customers in three countries (France, Spain and Germany). And the majority of the clients are from France. These Western European nations have varied, mature economies. A unique aspect of banking in this country is the stringent regulation that followed the 2008 financial crisis. Banks are able to withstand the danger of collapse better as a result, but they lose their allure and profitability.

Number of Product:

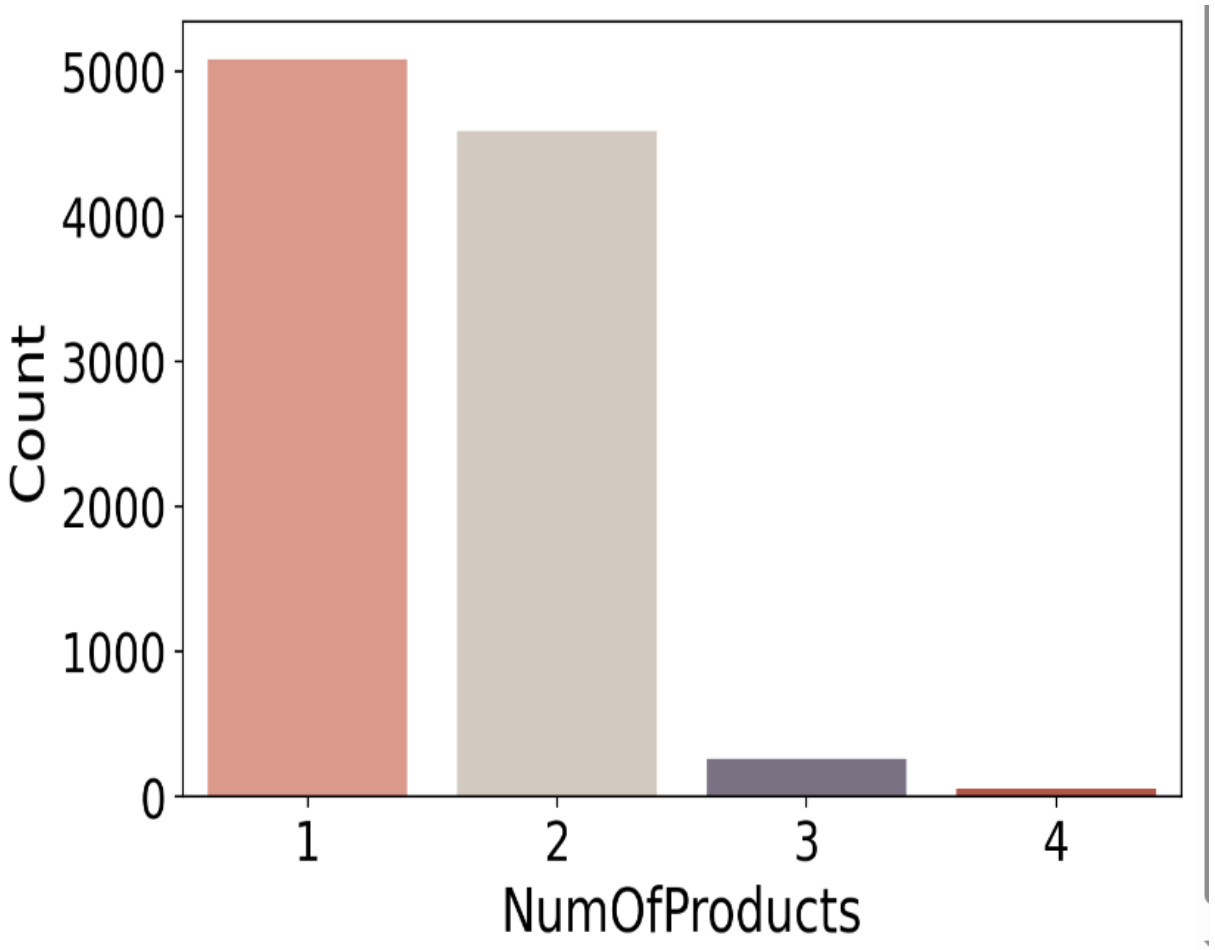

Figure 16: Histogram of Products

Figure 16 shows that Banks should investigate the underlying reasons for the significant preference of customers towards using 1 or 2 services compared to 3 and 4. This analysis is crucial in order to devise effective strategies that promote the relevant products based on the target customer segment.

Tenure:

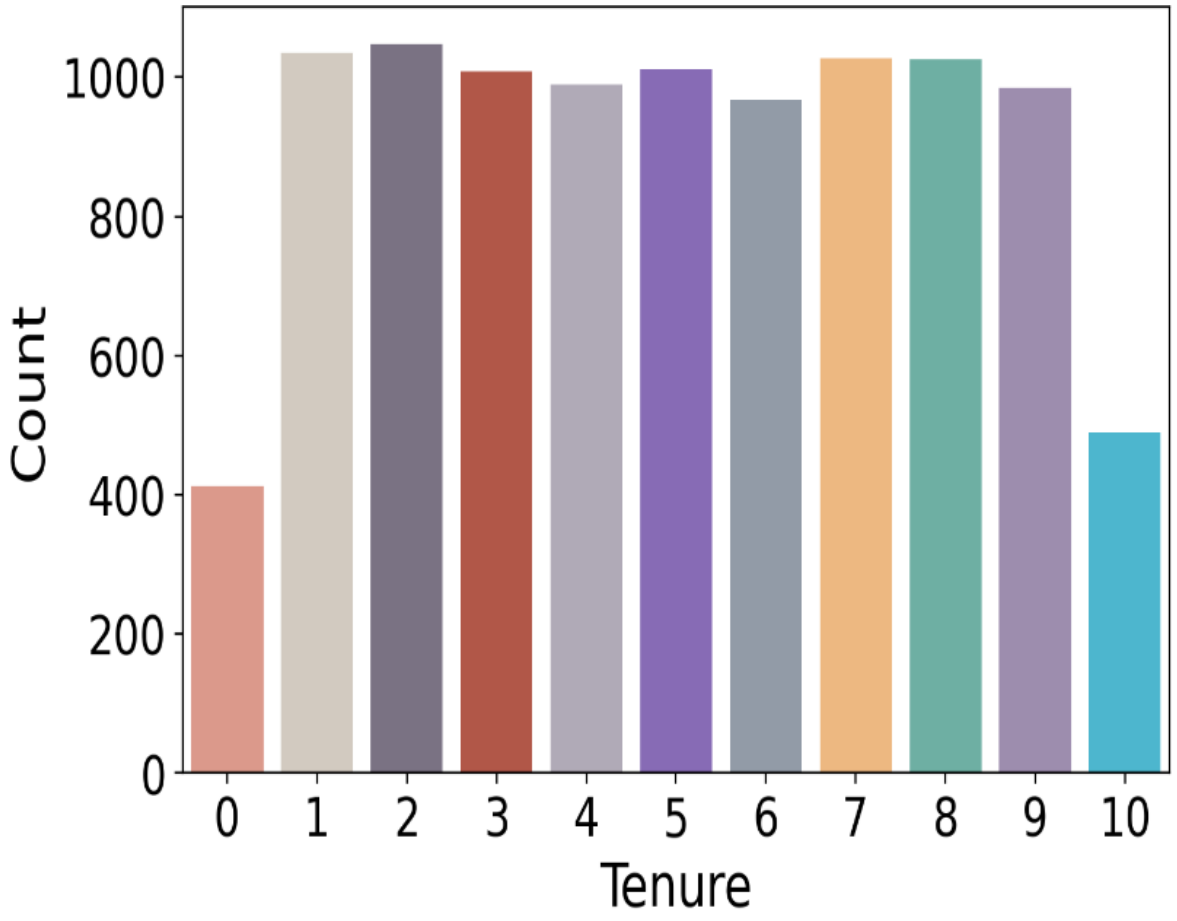

Figure 17: Histogram of Tenure

At figure 17 With the exception of a significant number of clients quitting during the first year, the bank's lifetime seems to be rather regular and stable. A customer's loyalty and contentment with the services received are indicated by the length of their connection with the bank. The data shows a substantial number of long-term clients, which may imply that the bank provides excellent services.

Additionally, this data helps the bank classify client profiles so that it may provide new services or improve the caliber of existing ones. To increase client retention rates, long-term customers could be offered specialized incentive programs or tailored financial counseling.

Has Credit Card:

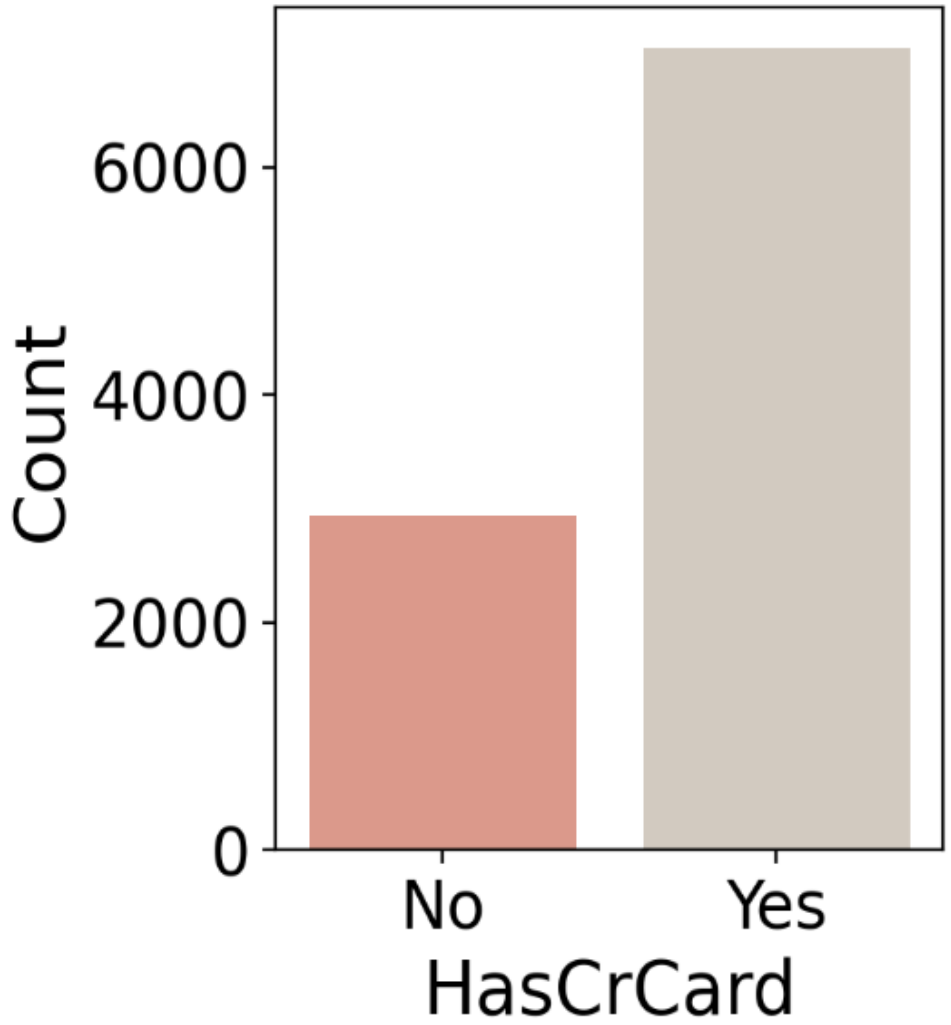

Figure 18: Histogram of HasCrCard

Figure 18 shows most clients pay using credit cards. The way a person uses their credit card might influence how they behave while utilizing financial services. Customers who already have credit cards are more likely to stay with the bank for a long time and grow familiar with its goods and services, allowing the bank to suggest additional services to them as a result. or devise strategies for keeping clients.

Active Member:

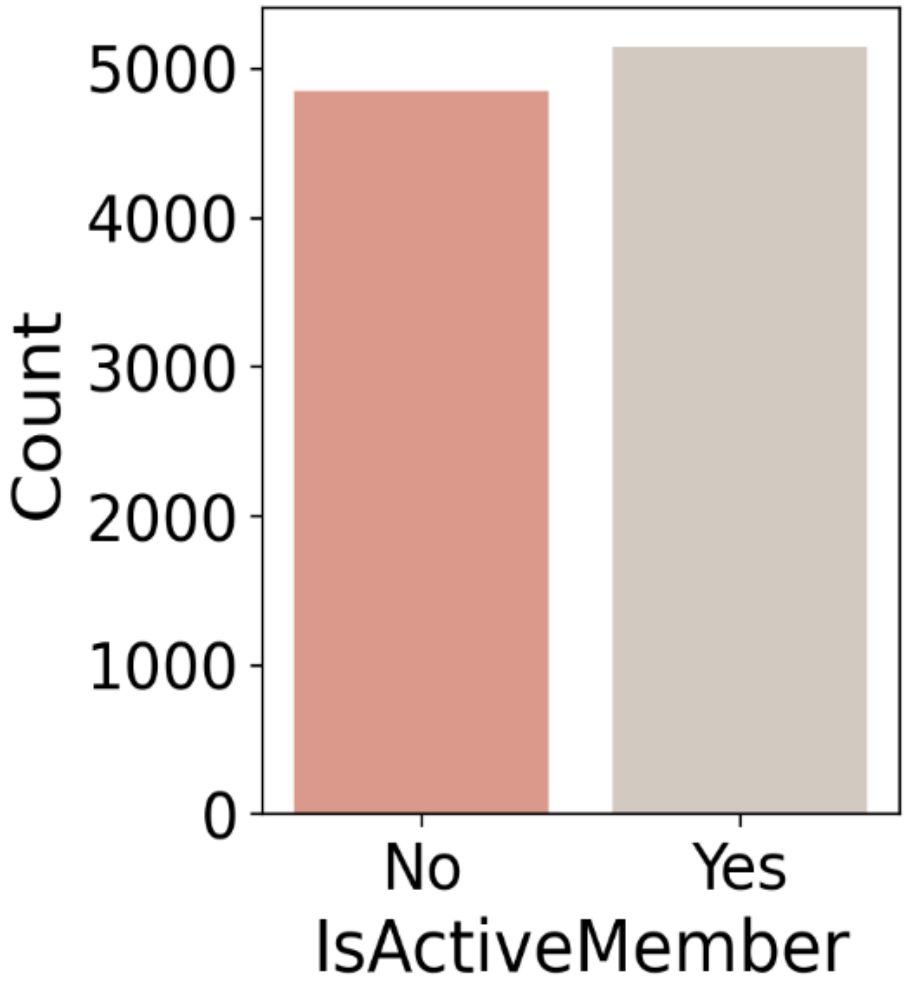

Figure 19: Histogram of Active Member

The number of inactive clients can be measured in relation to a bank-set time limit, such as 3 or 6 months, during which the customer has not engaged in any transactions or activities related to the services they are interested in using.

We can see on the figure 19, Even if there are more active customers than inactive customers, the difference is not great but also around the same, with the inactive customers making up over 50% of the total. This figure demonstrates the difficulties the bank is having in keeping its clientele. In order to take the proper corrective action, the bank must pay attention to the reasons why clients are no longer active, including if they are having difficulties utilizing the services or no longer require them. This is a crucial factor that has a big impact on the bank's business operations.

Exited:

The Exited variable is the goal variable, and its value is what we are most interested in. There are two possible values for the Exited variable: zero (for customers who haven't departed) and one (for customers who have).

We can discover how the bank arrived at this conclusion; perhaps it determined the client was no longer active for a certain period of time as departing or it carried out a customer survey. It's also possible that the bank had been given a request to stop providing the service by the consumer.

However, the dataset contains clear information about consumers who depart, so we do not need to focus too much on the process of data gathering.

We have shown the number of customers leaving at figure 20

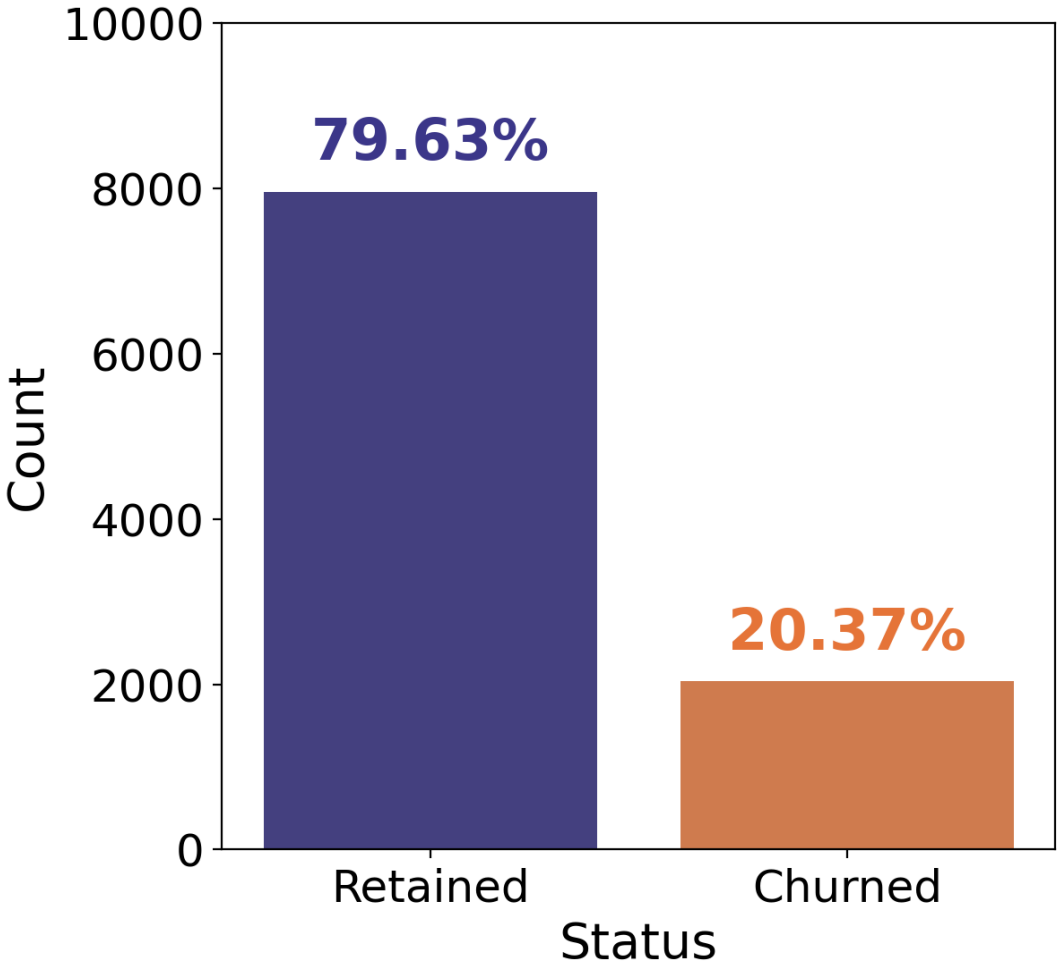

Figure 20: Histogram of Churn rate

- It is evident that around 20% of the whole client base leaves. This turnover rate is normal because it accounts for 18–20% of the financial industry's data.
- A low churn rate is not always a positive thing because the bank may expand slowly and find it challenging to bring in new clients; conversely, a high turnover rate may indicate that the bank has issues with quality and customer service.
- In addition, the desertion rate should be evaluated differently based on the service that customers utilize. Naturally, the main objective is to keep consumers

if the bank's services are geared at long-term clients. However, because consumers frequently switch banks, the aim of client retention is less critical if the service is as straightforward as creating a payment card or sending/receiving money. Therefore, the bank's services and strategy must be taken into account in order to evaluate the ratios.

4.5. Correlation

4.5.1. Relationship of target variable

Age:

To demonstrate the relationship between the variables and the target variable, we employ the histogram and boxplot methods. The figure 21 shows that the likelihood of a customer file leaving increases with age. Customers who are 40 and older will start to abandon the bank in greater numbers. This can assess the strategy and the incentives offered by the bank to clients who are no longer desirable so that they continue to utilize the bank's services. In addition, consumers over the age of 40 no longer use the bank because there probably weren't enough polls on middle-aged individuals to understand their wants and preferences. banking, then depart.

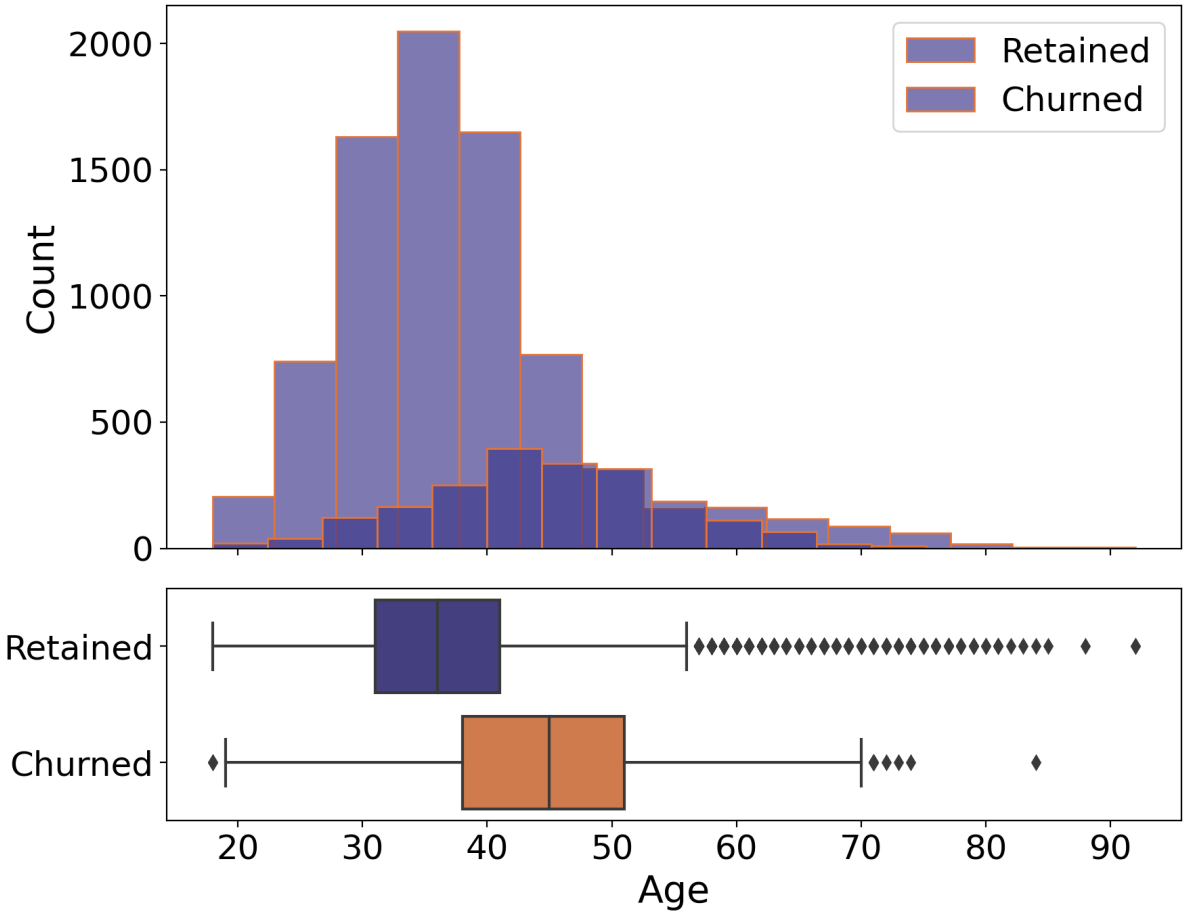

Figure 21: Correlation of Age and Churn

Credit Score:

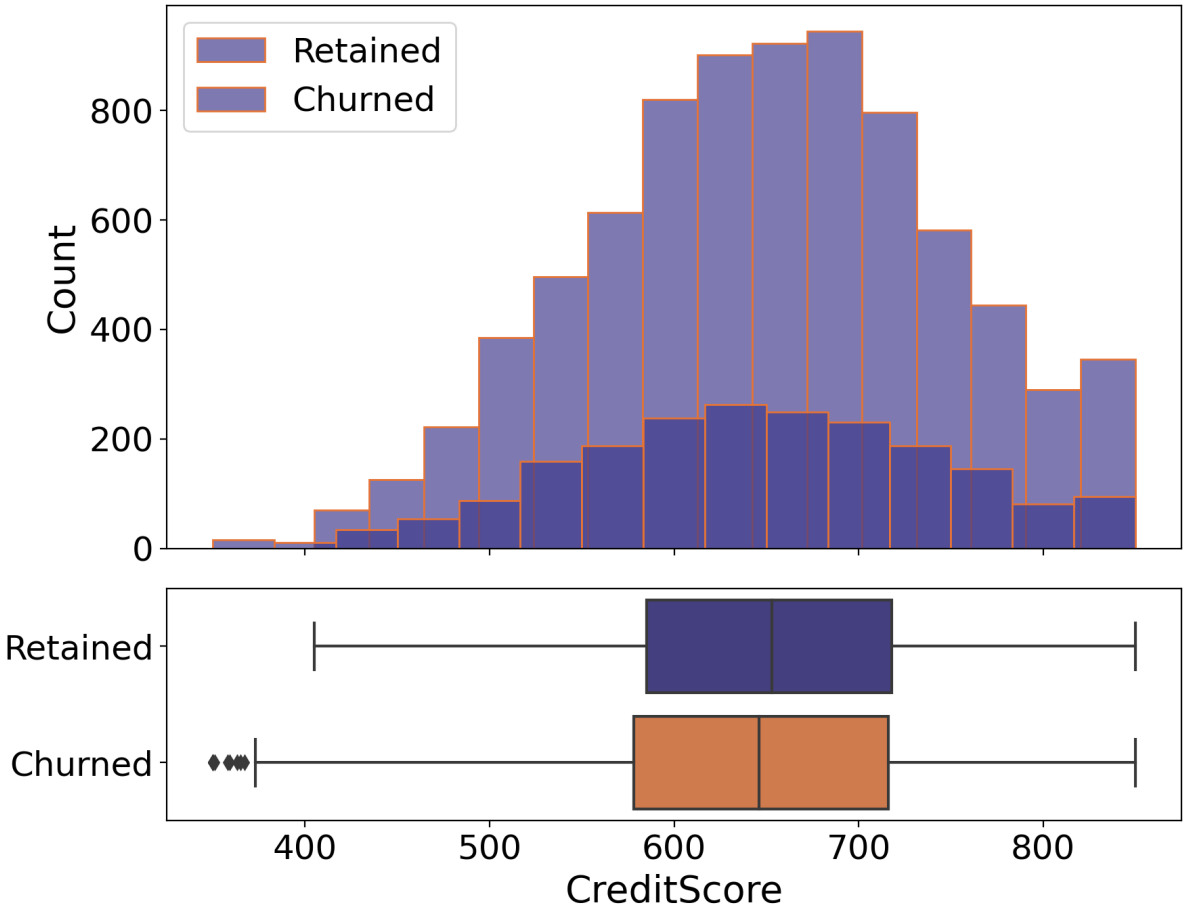

Figure 22: Correlation of Credit Score and Churn

The figure 22 shows that customers who discontinue using the bank's services and those who stay both have credit scores that differ significantly and not all that much. Additionally, it evaluates the bank's capacity to develop numerous options for clients to accrue credit accumulation points while using the bank.

Balance:

The distribution is pretty similar when comparing account balances of customers who leave and customers who continue to utilize the bank's services. It is evident that nearly no consumers depart with a low account balance, which is understandable given that the customer has just established a bank account and has a little amount. We can clearly see that in the figure 23 below:

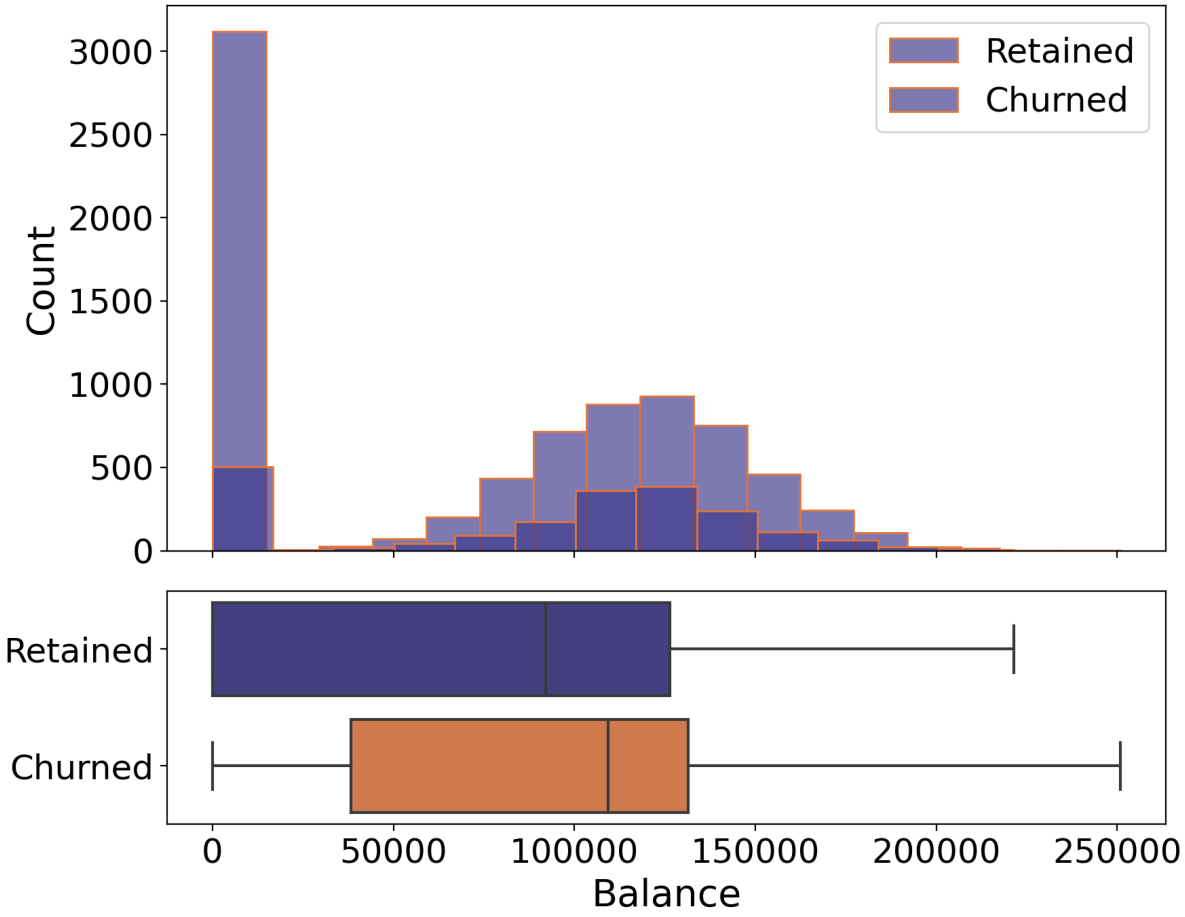

Figure 23: Correlation of Balance and Churn

Estimated Salary:

The chart at figure 24 shows the nearly same wage distribution of clients that enter and exit the bank, demonstrating the homogeneity of their pay. As a result, their decision to continue using the bank's services is not based on their compensation.

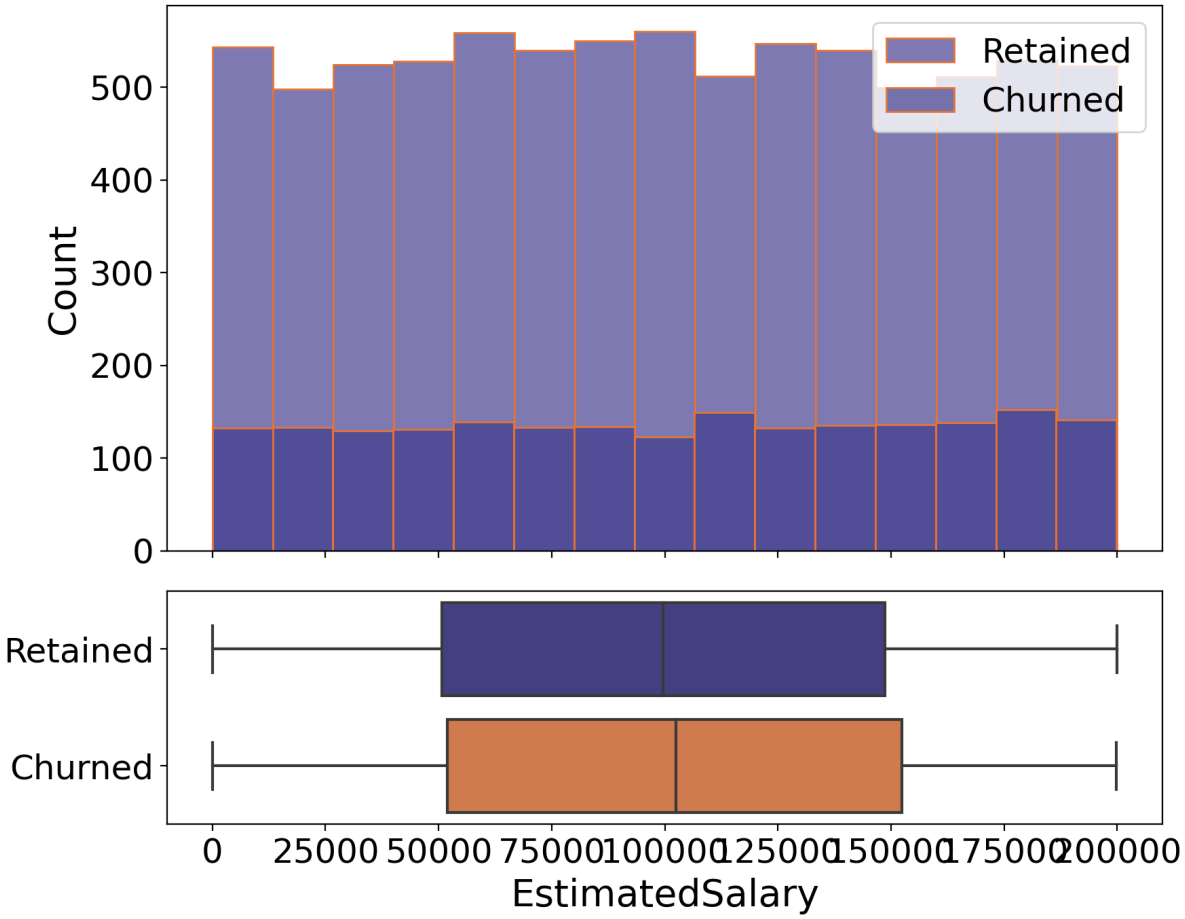

Figure 24: Correlation of Estimated Salary and Churn

Geography:

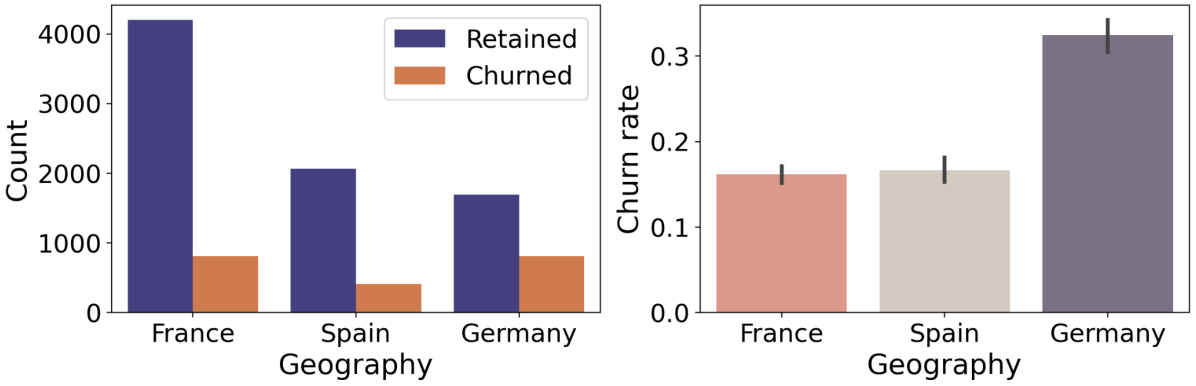

Figure 25: Correlation of Geography and Churn

The histogram at figure 25 shows that German clients are more likely to depart than customers in Spain and France are (the turnover rate is nearly twice as high). This conclusion might be explained by a variety of factors, including more competition or differing client preferences in Germany. Customers may like the bank in Germany because of the banks' high levels of stability and transparency, as well as the best deposit insurance scheme in all of Europe. consumers who are in their home nation and are very likely to switch banks.

Gender:

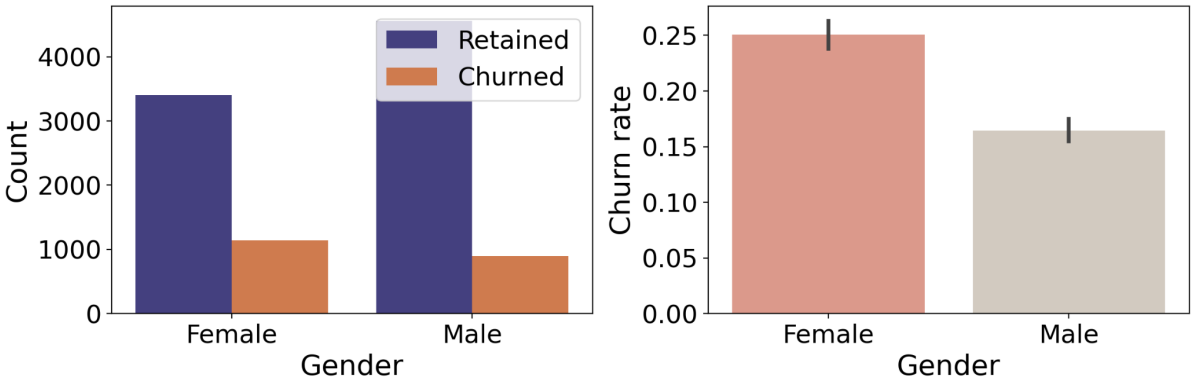

Figure 26: Correlation of Gender and Churn

The figure 26 shows that Women clients are more likely than males to depart. Despite the lack of formal study on this topic, we can nonetheless explain that female clients have high expectations for service quality and make purchases based on emotions, therefore when experiencing If something goes wrong, they'll make the decision to leave right away.

Men may be less inclined to leave than women since they are more interested in financial reasons, interest rates, danger, or safety are also carefully assessed by them.

Tenure:

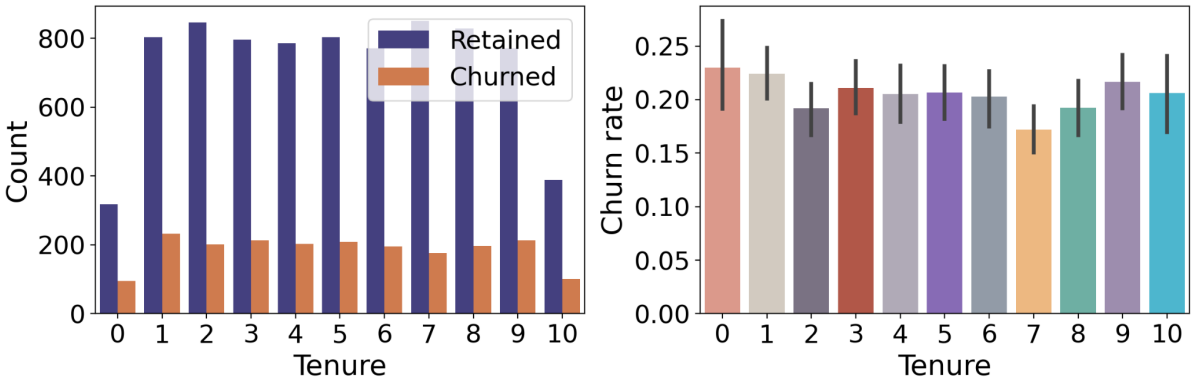

Figure 27: Correlation of Tenure and Churn

The graph at figure 27 shows that the length of time spent working in banking has essentially little impact on the turnover rate. However, it is evident that the first year has a larger rate of client attrition than the following years. Contrary to long-term clients, who are more likely to stick around because they feel the bank is the perfect fit for them and have use patterns that result in lower churn rates, consumers who have issues straight away will quit the company fast.

Number of Products:

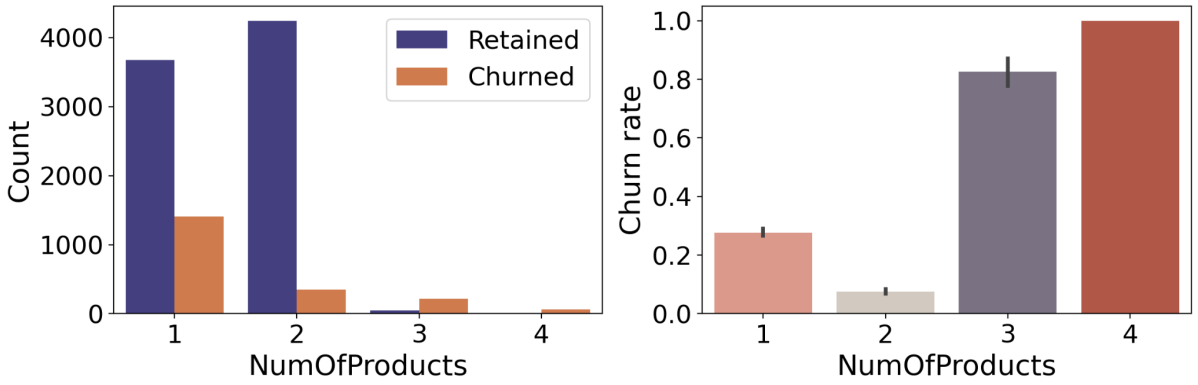

Figure 28: Correlation of Products and Churn

According to figure 28, Many clients only utilize one or two services. This demonstrates that the bank performs poorly in offering many of their services to clients. Particularly, the incidence of departure in the file of clients who utilize several services is particularly significant. The cause of this may be that the bank's processes caused consumers to be dissatisfied when using many services at once; this might be a procedural issue or a weakness in the bank's strengths that caused a negative customer experience when customers used other services, which affected customer retention.

Card Holders:

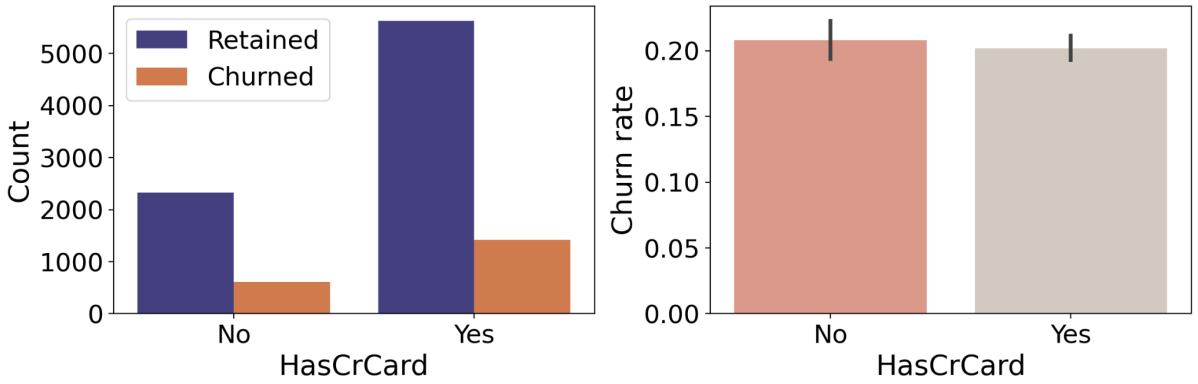

Figure 29: Correlation of Credit Card and Churn

The graph in figure 29 shows that whether or not a customer has a credit card doesn't seem to affect customer churn.

Active Members:

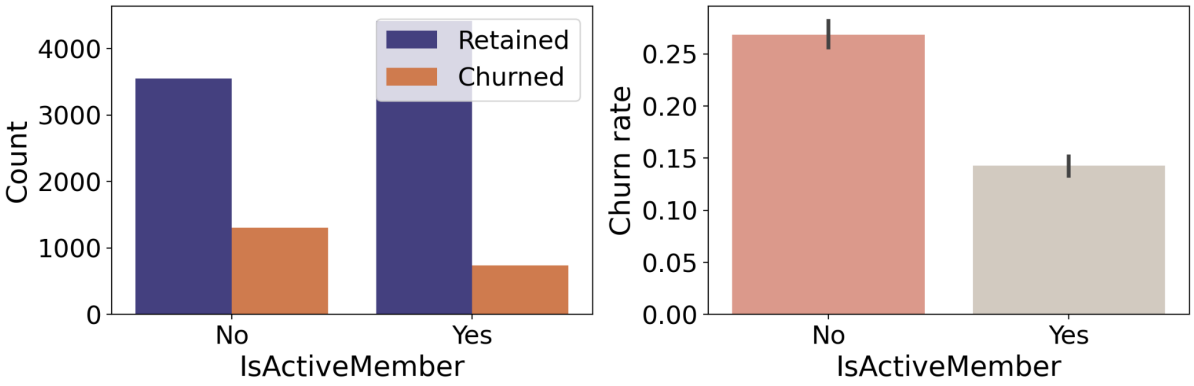

Figure 30: Correlation of Active Member and Churn

Looking at the figure 30 When these two numbers are closely associated, it makes sense that inactive customers account for a considerably greater abandonment rate. Customers become inactive for a variety of reasons, including a lack of need, a lack of attention and encouragement from the bank, or negative experiences that force them to leave the bank. As a result, banks must develop customer-focused policies and encourage more activity in their operations. The majority of clients won't want to continue using the bank's services once this element is no longer functioning, which has a significant impact on the customer churn rate. The bank must examine customer satisfaction and experience to determine why customers leave and take steps to address it as the active customer base is still more inclined to do so.

4.5.2. Correlations Analysis

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

Figure 31: Correlation Matrix

We are particularly interested in strongly correlated data when plotting correlation matrices in order to identify prediction models or significant factors. High correlation is not always a positive thing, though, as multicollinearity can result from variables with high correlation values but little theoretical significance.

We can quickly observe that there is no appreciable association between the variables by looking at the correlation matrix in figure 31 This gives us comfort that multicollinearity is being reduced. Low correlation variables must be incorporated into the model and well thought out.

4.6. Pre-processing

4.6.1. Feature Selection

During feature selection, the team decided to remove three columns, "Row Number", "Customer Id" and "Last Name" as they were deemed irrelevant and did not contribute meaningfully to the rate prediction. customer churn rate.

The 'Estimated Salary' feature has also been removed from the dataset because it shows an even distribution for both types of customers.

Furthermore, a closer examination of the 'Tenure' and 'HasCrCard' features reveals that they have similar abandonment rates, indicating redundancy. To confirm this observation, a chi-squared test was conducted, which resulted in a small chi-squared value and a p-value greater than the standard cutoff value of 0.05. The results in Table 5.3 support the initial hypothesis that these two features do not provide any useful information in predicting customer churn. Therefore, they are considered redundant and excluded from the final feature set.

Table 4.1: 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
HasCrCard	0.301	5.833e-01

4.6.2. Encoding categorical Features

In the context of machine learning, we need to ensure that all input (and output) functions are represented as numbers. Therefore, if categorical features are present in the dataset, they require a preprocessing step known as coding, where they are converted to numeric values before being used to train the model. Coding categorical features allows us to incorporate the information contained in classifier features into our model, contributing to more accurate predictions and better overall performance. The dataset under consideration contains two categorical features to be coded. LabelEncoder() from scikit-learning is used for the "Gender" function. This method maps each unique identifier to an integer. Male is represented as 1 and female as 0. A manual mapping method is used for geographic functions. Specifically, customers in Germany are assigned a value of 1, while customers in France and Spain are assigned a value of 0. This decision was made based on the observation that the customer churn

rate in France and Spain are the same and significantly lower than in Germany. So it makes sense to code this function in a way that effectively distinguishes between German and non-German clients.

4.6.3. Scaling

A typical method used to equalize the variety of features in a dataset is feature scaling. To prevent domination of specific features due to their magnitudes, it is essential to make sure that features are on a similar scale.

The StandardScaler() technique was utilized to scale the features in this analysis. Features are normalized by removing their mean and dividing by their standard deviation using StandardScaler(). The features are effectively scaled to have a unit variance and centered around zero using this transformation.

Scaling was done with the 'CreditScore', 'Age', and 'Balance' properties of the train_data dataframe. The fit_transform() method was used to apply the StandardScaler() to these particular columns, guaranteeing that the scaling procedure is consistent and based on the training data.

4.6.4. Addressing Class Imbalance using SMOTE function

The dataset demonstrates a large class imbalance, with one class (0 - kept) being far more abundant than the other (1 - churned), as seen in Figure 5.4.

Table 4.2: Imbalance data

Exited	
0	7963
1	2037

Unbalanced classes are a problem that frequently arises in real-world activities. In classification tasks, it presents a problem since machine learning algorithms frequently prefer the dominant class, resulting in models that largely predict the prevailing class. Consequently, in the presence of unbalanced data, traditional evaluation metrics can be deceptive.

The SMOTE ('Synthetic Minority Oversampling Technique') technique was chosen for this investigation to overcome this issue. By locating related records and producing new synthetic records as a weighted average of the initial and nearby examples, SMOTE develops synthetic instances for the minority class.

The sampling_strategy was set to 'auto' while using the imblearn library's SMOTE function to apply SMOTE. SMOTE was then used to resample the training data (X) and associated labels (Y) in order to balance the class. After applying SMOTE, both classes now have an equal number of instances at table 5.5:

Table 4.3: Balance data

Exited	
0	7963
1	7963

5. Proposed model

Based on the characteristics of the dataset as well as the analysis objective of this topic, the team has selected these machine learning algorithms to build predictive models: ADT; CBA; CBA2; L3; MAC; CMAR; KNN.

These are prominent algorithms in the fields of data mining, machine learning and association rule mining. In the EDA section, we explored the correlation of the values with the target variable, the results showed that there was no significant effect, however we found that the variables all had an impact on the target variable. goals and even a combination of those values and customer churn. That's why the team uses powerful classification algorithms like the one above. In addition, after processing the dataset, the team found that the dataset has 3 continuous variables and 5 important categorical variables, so it is necessary to use the above algorithms because it is suitable for continuous variables. Of course, the team evaluated based on 2 aspects of 2 types of models that are suitable for categorical and target variables, the results that the team implemented predictive models with algorithms Logistic Regression, Support Vector Classifier, Random Forest Classifier, Gradient Boosting Classifier, Xtreme Gradient

Boosting Classifier, and Light Gradient Boosting Machine all gave Accuracy results below 82%.

In this project, the team uses the model building method of Professor Philippe with the following highlights: The training and testing dataset is divided based on the k-fold cross-validation method with k=10, this value matches the number of records of the data set. The use of cross-validation reduces overfitting as well as provides an accurate and generalizable result.

5.1 CBA

The highlight in this algorithm is the minSup and minConf values. With the default code, the value minSup = 0.1 is selected, which means a rule must appear in at least 10% of the transactions to be considered valid and the value minConf = 0.5 is selected, which means a rule must be valid. have a confidence level of at least 50% to be considered valid. These two values completely match the requirements for accuracy as well as the ability to create the law that we want. The results of the CBA model are shown in the table... We can see that it's prediction ability is quite good with 80.2% accuracy.

Table 4.4: Result of CBA Classification

CLASSIFICATION RESULT ON TRAINING DATA							
Name	**Accuracy**	**Recall**	**Precision**	**F-micro**	**F-macro**	**Time**	**Memory**
CBA	0.7902	0.6397	0.6659	0.7902	0.6499	9	19.4186
CLASSIFICATION RESULT ON TESTING DATA							
CBA	0.802	0.6733	0.6907	0.802	0.6809	3	19.4186

5.2 CBA2

CBA2 is an improved version of the CBA algorithm, improving computational performance and rare sample processing capabilities. In the table... we can see that CBA2's prediction results are better than CBA's, however in terms of time and memory performance, CBA2 consumes a bit more than CBA.

Table 4.5: Result of CBA2 Classification

CLASSIFICATION RESULT ON TRAINING DATA							
Name	**Accuracy**	**Recall**	**Precision**	**F-micro**	**F-macro**	**Time**	**Memory**
CBA2	0.8062	0.5281	0.828	0.8062	0.5007	10	19.8333

CLASSIFICATION RESULT ON TESTING DATA							
CBA2	0.816	0.5508	0.8837	0.816	0.5411	2	19.8333

5.3 L3

The L3 algorithm gives lower results than CBA with the same minSup and min Conf parameters and there is no significant difference between the time and memory of these two algorithms.

Table 4.6: Result of L3 Classification

CLASSIFICATION RESULT ON TRAINING DATA							
Name	**Accuracy**	**Recall**	**Precision**	**F-micro**	**F-macro**	**Time**	**Memory**
L3	0.7963	0.5	0.3982	0.7963	0.4433	9	19.8333
CLASSIFICATION RESULT ON TESTING DATA							
L3	0.796	0.5	0.398	0.796	0.4432	2	19.8333

5.4 MAC

MAC uses data scanning to find key association rules, so it consumes more time and memory than CBA and L3. However, its prediction result is not as high as CBA.

Table 4.7: Result of MAC Classification

CLASSIFICATION RESULT ON TRAINING DATA							
Name	**Accuracy**	**Recall**	**Precision**	**F-micro**	**F-macro**	**Time**	**Memory**
MAC	0.7963	0.5	0.3982	0.7963	0.4433	20	25.3352
CLASSIFICATION RESULT ON TESTING DATA							
MAC	0.796	0.5	0.398	0.796	0.4432	4	25.8333

5.5 CMAR

CMAR uses sequence mining to mine association rules. In addition to minSup and minConf, CMAR uses the delta variable to determine the minimum magnitude of the difference between data samples within the same group. The value delta = 2 was chosen to ensure that the CMAR finds groups of data with high similarity but is still sensitive enough to distinguish data samples with larger differences.

The results of the CMAR algorithm have no difference with L3 and MAC and have average performance compared to these two algorithms.

Table 4.8: Result of CMAR Classification

CLASSIFICATION RESULT ON TRAINING DATA							
Name	**Accuracy**	**Recall**	**Precision**	**F-micro**	**F-macro**	**Time**	**Memory**
CMAR	0.7963	0.5	0.3982	0.7963	0.4433	17	20.3352
CLASSIFICATION RESULT ON TESTING DATA							
CMAR	0.796	0.5	0.398	0.796	0.4432	2	20.8333

5.6 ADT

Like MAC, ADT focuses on finding and analyzing association rules in data sets, but ADT is more general. For ADT, we chose minConf = 0.5 and minMerit = 0.5 with the desire to find more accurate and robust association rules. The prediction results are almost no different with the MAC but the time and memory performance are better improved.

Table 4.9: Result of ADT Classification

CLASSIFICATION RESULT ON TRAINING DATA							
Name	**Accuracy**	**Recall**	**Precision**	**F-micro**	**F-macro**	**Time**	**Memory**
ADT	0.7963	0.5	0.3982	0.7963	0.4433	5	17.0103
CLASSIFICATION RESULT ON TESTING DATA							
ADT	0.796	0.5	0.398	0.796	0.4432	2	17.0103

5.7 KNN

Unlike the rest of the algorithms, KNN is used to classify or predict the label of a new data point based on its nearest neighbors. Here, the default parameter is k=3, which means we will take the 3 nearest neighbors to participate in the prediction decision. The model's prediction results are very high, but it also consumes a lot of time and memory.

Table 4.10: Result of KNN Classification

CLASSIFICATION RESULT ON TRAINING DATA							
Name	**Accuracy**	**Recall**	**Precision**	**F-micro**	**F-macro**	**Time**	**Memory**
ADT	0.9898	0.9749	0.9937	0.9898	0.9839	1335	42.7987
CLASSIFICATION RESULT ON TESTING DATA							
ADT	0.996	0.9902	0.9975	0.996	0.9938	111	73.2968

6. Experimental result and Analysis

6.1. Compare results

6.1.1 Classification results on training data

#NAME:	ADT	CBA	CBA2	MAC	L3	CMAR	KNN
#ACCURACY:	0.7963	0.7902	0.8062	0.7963	0.7963	0.7963	0.9898
#RECALL:	0.5	0.6397	0.5281	0.5	0.5	0.5	0.9749
#PRECISION:	0.3982	0.6659	0.828	0.3982	0.3982	0.3982	0.9937
#KAPPA:	0	0.3017	0.0861	0	0	0	0.9679
#FMICRO:	0.7963	0.7902	0.8062	0.7963	0.7963	0.7963	0.9898
#FMACRO:	0.4433	0.6499	0.5007	0.4433	0.4433	0.4433	0.9839
#TIMEms:	4	10	3	15	3	12	1594
#MEMORYmb:	39.0993	40.6013	42.0993	49.0993	51.6013	54.6013	45.6663
#NOPREDICTION:	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 32: Classification results on training data

We have highlighted the highest-accuracy measures to easily review the algorithms that performed well in the training set. It can be clearly seen that the ratio of KNN outperforms the rest. However, the execution time is much longer The CBA2 algorithm also achieved quite good results with an accuracy of 80.62% and was one of the models with the best time performance.

6.1.2 Classification results on testing data

#NAME:	ADT	CBA	CBA2	MAC	L3	CMAR	KNN
#ACCURACY:	0.796	0.802	0.816	0.796	0.796	0.796	0.996
#RECALL:	0.5	0.6733	0.5508	0.5	0.5	0.5	0.9902
#PRECISION:	0.398	0.6907	0.8837	0.398	0.398	0.398	0.9975
#KAPPA:	0	0.3624	0.1522	0	0	0	0.9876
#FMICRO:	0.796	0.802	0.816	0.796	0.796	0.796	0.996
#FMACRO:	0.4432	0.6809	0.5411	0.4432	0.4432	0.4432	0.9938
#TIMEms:	0	0	0	2	1	1	161
#MEMORYmb:	39.0993	40.6013	42.0993	49.6013	51.6013	54.6013	5.67
#NOPREDICTION:	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 33: Classification results on testing data

The value at the test set is the one we are most interested in. Here, the KNN algorithm is still very prominent with rates from 0.99 and up. However, memory and execution time still consume too much. CBA2 continues to achieve pretty good ratings except for FMACRO. However CBA2 still has very good time and memory performance.

Whether to use CBA2 or KNN depends on the requirements of the project, for this project, our dataset is not too much and we want to give the most accurate prediction results, so we can ignore factors of time and memory. Therefore, KNN is the model that we choose.

6.2. Compare with other algorithms

We have evaluated the predictive accuracy of the classification algorithms listed in Professor Philippe's library. Here we use the Accuracy results on the training data and testing data of the algorithms and visualize it with a line graph at figure...

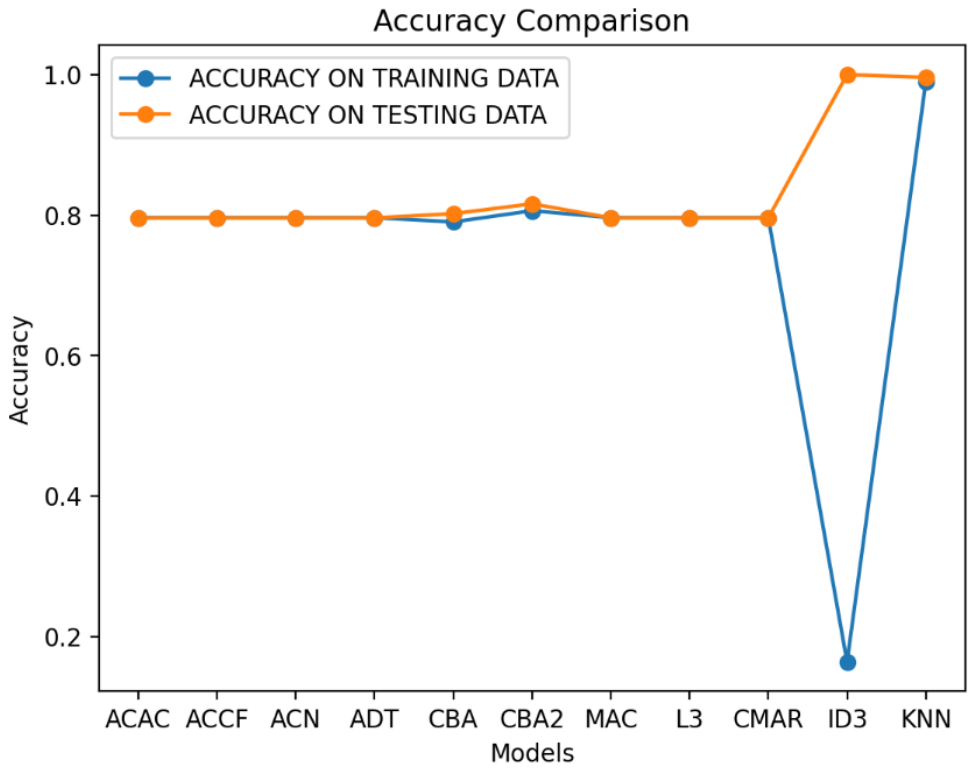

Figure 34: Compare with other algorithms

Here we can see that the prediction result of ID3 reaches result 1 in testing data, but it has very low accuracy on training data. There is a high possibility that underfitting has occurred when the model is too simple or is not capable of learning the patterns and rules in the training data. In this case, the model cannot properly learn the samples in the training set, resulting in low performance on the training set. However, since the model is not too complex, it has better generalization over the test set, resulting in higher performance.

We are quite worried about the large difference in the prediction results of ID3 and the already very high accuracy KNN, so we will use KNN to make sure the model has good results when running the new customer dataset in the future.

6.3. Rebuild kNN Classification model using k

We combined the Best Neighborhood Search and Rule of Thumb methods to find the k value. We first calculate the square root of the number of samples according to the Rule of Thumb method with k=100. The results show that the accuracy is only 57.14%. After that, we experiment with K=10 and gradually reduce k, finally k=1 achieves the best results with indices equal to 100%. Maybe because our training data is large enough, the KNN model is able to learn accurately from nearest neighbors. We decided to use k=1 to run the model.

6.4 Run model with Unknown pattern

6.4.1 Typical churn customer

Based on research that explores the influence of variables on the target variable, the team calculated a typical example for a customer who is likely to ignore specific metrics that can be in turn for each CreditScore, Location variable in turn. Reason, Gender, Age, Balance, NumofProducts, IsActiveMember are: -1.54714095593388, 1, 0, 0.48422460429935, 1.0508202865849, 4, 0. After running the prediction for this customer example, the project output received by the project team being 1 means this customer will be left from the bank. This result can prove that the predictions and judgments of the group are correct about the target variables.

6.4.2 Typical retained customer

Next, the team re-evaluates an object with features that are less likely to leave as well as target variables like the example for leaving customers as above. With the target variables Credit Score, Geography, Gender, Age, Balance, NumofProducts, IsActiveMember and specific numbers it is possible to explain for each target variable in turn as follows: -1.54714095593388, 1,1 , 0.48422460429935, 1.0508202865849, 2, 0. After the team runs the prediction for this customer example, the result is that the group gets 0 which means this customer will continue to use the service and use this bank. This result may also clarify that the assessment and predictions of the target variables for the bank's customers are correct.

Therefore, both of the group's hypotheses are made correct. The bank can focus on the factors of Geography, Gender, Number of Products, Active Members that will influence its customers to improve abandonment as well as attract files. new customer for your Bank. The focus is on studying the current economic status and financial situation of customers to come up with strategies and solutions related to issues of Geography, Gender, Number of Products, IsActiveMember related to customers.

6.5. Strategy

In order to have a more detailed and comprehensive overview of the business for this bank in order to improve its goal of reducing customer churn and attracting new customer files, the team proposed strategies for bank as follows:

- First and foremost, whenever we receive feedback from customers. Even without a predictive model yet, the bank can manage and lower the churn rate by looking at client behavior. The bank must take two actions when a customer requests to cancel a service: first, it must investigate the reasons for the departure; second, it must persuade the consumer to stay by offering alternatives. appropriate answer. The bank must assess the subjective and rational factors if the client is still adamant about leaving. It is challenging for the bank to modify it due to particular customer-side considerations like finances and health, but due to bank-side issues, they must pose questions: Can a situation with a dissatisfied consumer be resolved? How much does it cost and how long does it take to remedy the issue? What matters should be given top

priority? We are confident that if the issues brought up by consumers can be resolved, the desertion rate will be substantially lower for reasons related to 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:
    - Customers that are older are more likely to depart. Therefore, banks must develop crucial techniques for caring for their senior consumers. This category of consumers has significant disadvantages while utilizing the bank's services due to health and personality issues. Regular communication from the bank is required, and the staff must show them a lot of consideration, patience, and care. Customers over the age of 60 should also be recommended for insurance, health, and savings services. Additionally, the marketing effort must be delicate and suitable; sometimes calls and honesty will be more successful than digital marketing activities.
    - Clients that utilize numerous goods run the danger of being abandoned more frequently. The bank is having difficulty offering more services. As a result, the bank should immediately put an end to its development plans and concentrate on quality concerns, improving the uniformity of quality across all services rather than just a few exceptional ones. Although the bank has an excellent service plan, the supporting service also needs to be of a consistent quality for clients to take into account among other things. Banks must take into account the convenience of utilizing different services if there is a conflict when integrating services. A sensible and practical combination of services is required. Utilizing 4 services with 4 entirely separate procedures does not make any consumer pleased.
    - Another crucial element is customer activity. A bank can use interactive marketing, which is a crucial strategy for the bank to build and retain connections with consumers, when it notices that its customers are less engaged. This involves having direct conversations with clients over the phone, via email, SMS, or on social media. Through this engagement, the bank is better able to comprehend client requirements and expectations, respond to inquiries, and offer specialized service. Banks have the ability to create a thriving consumer base for their goods and services. Create discussion boards, chat rooms, or other online spaces where users may interact and exchange views and experiences. Customers will become more engaged as a result.
    - In order to increase asset value and foster long-term cohesion, banks need to implement strategies like offering preferential account packages and services to customers with high balances; developing a special bonus and benefit program for customers who maintain large balances in their accounts; and consulting and offering investment options appropriate to the customers' financial situations.

- Geographically, it is evident that Germany is receiving a poor reaction and that the bank's market dominance is mostly focused in France. To engage local clients, banks must organize events and initiatives in the neighborhood. Adding new financial services and solutions to address the unique requirements of each location. German banks, for instance, are widely renowned for their culture of saving and faith in investment goods. Banks may concentrate on creating dependable and secure savings and investment products, such as investment funds, savings accounts, and insurance options. simultaneously giving expert financial and asset management guidance. Another crucial element is spending money on regional media and advertising to raise brand recognition.

- Continually assess and forecast the turnover rate every quarter in accordance with the bank's unique strategies.
    - A regular consumer may abruptly stop doing business. Since nothing can be predicted in advance, the bank may better identify client abnormalities and improve its services and products by anticipating the periodic turnover rate.
    - Before an event or company plan is implemented, it is also crucial to forecast and assess the turnover rate. In order to develop the best plan and save money and other resources, the bank might use this to locate the target client file. The effectiveness of the campaign must also be determined after the campaign has ended by reevaluating the churn rate. If the bank has implemented the proper steps, the churn rate will have decreased. Of course, in addition to turnover rate measurements, we also need to pay attention and ask about customer satisfaction to get a more complete picture.

We anticipate that the Bank will be able to increase client retention rates and raise the caliber of its goods and services using the solutions we present. If the aforementioned techniques are properly implemented, we think their success will increase the number of new clients the bank receives in addition to lowering the churn rate. The bank's worth will rise as well, and it will rank among the respected businesses in the financial industry.

7.Conclusion

7.1 Conclusion

In conclusion, this subject report focuses on the crucial task of predicting customer churn in the banking industry by utilizing K-Nearest Neighbors (KNN) classification as the primary data mining technique. The study used the bank customers dataset, which provided valuable insights into customer behavior and churn patterns.

Throughout the study, we analyzed the data comprehensively, extracting meaningful features and uncovering hidden patterns that shed light on factors influencing customer churn. By using KNN classification, we built a robust predictive model that accurately

identified potential churners. This achievement demonstrates the relevance and effectiveness of data mining techniques in understanding customer behavior and enables banks to take proactive measures in retaining valuable customers.

7.2 Limited

- Data preprocessing and feature engineering play a key role in enhancing model performance and providing valuable business insights, but the quality and completeness of the original data set poses challenges.
- The approach overlooked potential changes in customer behavior over time by failing to explicitly address temporal dynamics. The model's ability to capture changing churn patterns as market conditions and customer preferences change could be impacted by this mistake.
- We have not fully considered the correlation between many variables, which leads to the failure to come up with more specific business strategies for businesses.
- The predictive model achieves very high results with k = 1, but it is possible that because the large training set leads to the model with good results, in the future there may be problems when applying on new datasets.

7.2 Future work

- Future research should incorporate additional data sources and explore more complex models like ensemble methods, deep learning, or support vector machines to potentially yield more refined churn prediction models.
- Investigating the impact of personalized marketing strategies and customer retention initiatives based on the model's predictions would also be a valuable area of exploration.
- We need to make better use of the SPMF library to take advantage of its great capabilities, thereby further developing better evaluations of the model.

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. Anil Kumar, D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*
3. Bilal Zorić, A. (2016). Predicting customer churn in the banking industry using neural networks. *Interdisciplinary Description of Complex Systems: INDECS*
4. 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
5. Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal Of Applied Microeconometrics*
6. Krull, A. (2021). *Voluntary Churn Vs. Involuntary Churn | Recurly*. Retrieved 07 22 , 2021, from https://recurly.com/blog/subscriber-retention-and-understanding-involuntary-vs.-voluntary-churn/
7. Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in the telecom industry using boosting. *IEEE Transactions on Industrial Informatic*
8. Mutanen, T., Nousiainen, S., & Ahola, J. (2010). Customer churn prediction–a case study in retail banking. *In Data Mining for Business Applications*
9. Saw Thazin Khine, Win Win Myo (2019). Customer Churn Analysis in Banking Sector
10. 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
11. Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*