UNIVERSITY OF CALIFORNIA

Los Angeles

Customer Churn Prediction
In Banking Industries: Supervised Machine
Learning Approach

A dissertation submitted in partial satisfaction of the

Requirements for the degree of Master of Applied

Statistics & Data Science

by

SUJIAN JIANG

2024

ABSTRACT OF THE DISSERTATION

Customer Churn Prediction
In Banking Industries: Supervised Machine
Learning Approach

by

SUJIAN JIANG
Master of Applied Statistics & Data Science
University of California, Los Angeles, 2024
Professor YingNian Wu, Chair

Customer churn in the banking industry occurs when clients terminate their relationship with the bank, leading to significant losses in revenue and reputation. In today's highly competitive market, retaining customers is crucial. A strong customer base not only sustains the bank's revenue but also attracts new clients through trust and referrals from satisfied customers. Therefore, identifying and preventing customer churn is a critical task for banks. Our research utilized various machine learning algorithms to predict which customers are likely to leave. By analyzing the models, we identified patterns that serve as early warning signs of churn. Based on these valuable insights, we provide banks with recommendations on effective strategies to retain their customers.

The dissertation of SUJIAN JIANG is approved.

Nicolas Christou

Oscar Madrid Padilla

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2024

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Customer churn, also known as customer attrition, refers to the phenomenon where customers stop doing business with a company or service provider. It poses a significant challenge that companies across various industries commonly encounter.

In the banking industry, customer churn occurs when customers stop using a bank's services or switch to a competing financial institution. This also includes closing savings or checking accounts, discontinuing credit cards, moving investments, or refinancing loans with another provider.

Bank client churn will cause big damage to the bank as it has multifaceted impact on bank revenue, customer acquisition costs, and overall bank reputation. One primary concern is the immediate loss of revenue flows derived from interest, transaction fees, and cross-selling opportunities. As customers discontinue their association with the bank, these vital revenue sources diminish, especially when high-value clients are involved. Therefore, understanding and managing customer churn are essential for banks to ensure financial stability and protect their reputation.

Moreover, it is valuable for banks to establish and maintain long term customer relationships. Primarily, long term stable relationships lead to an increase in Customer Lifetime Value (CLTV), as loyal clients tend to engage with multiple banking products and services over time, generating higher cumulative revenue. Long-term customers also present enhanced cross-selling opportunities. Their established trust in the bank makes them more likely to adopt additional financial products such as credit cards, investment accounts, and loans. Additionally, these satisfied clients often serve as organic brand advocates, offering valuable word-of-mouth referrals that can significantly enhance organic growth.

On the other hand, a high customer churn rate often indicates underlying problems within the bank, such as a lack of attractive products or poor customer service. Analyzing customer attrition patterns is crucial for identifying these issues and proactively addressing them. This strategic approach enables banks to enhance their product offerings, refine their customer service practices, and ultimately improve the overall customer experience, thereby reducing churn and fostering stronger, long-term customer relationships.

The primary aim of this study is to uncover patterns that lead to customer churn, operating on the premise that most customers leave a bank due to identifiable reasons, with early warning signs appearing before they sever their relationship. To detect these patterns, we employed a variety of supervised machine learning techniques. Among these were tree-based algorithms such as XGBoost, Random Forest, and AdaBoost, alongside the linear-based Support Vector Machine (SVM). These models have demonstrated high accuracy in pinpointing customers who are on the verge of churning, allowing us to analyze their behavioral trends comprehensively.

This research offers banks crucial insights into which customers are at risk of leaving, enabling them to proactively implement retention strategies. By understanding the predictive patterns, banks can refine their customer retention approaches, enhance service quality, and ultimately boost overall customer satisfaction. The insights gained from these machine learning models provide a robust framework for strengthening customer relationships and ensuring long-term growth. By leveraging these predictive analytics, banks can identify at-risk customers earlier and take timely, targeted actions to retain them, thereby minimizing churn rates and fostering a loyal customer base.

CHAPTER 2

Methodology

In this section, we are going to introduce the four supervised machine learning algorithms that we used to predict customer churn rate. Comparing the advantages and disadvantages of each model.

2.1 XGBoost

If there is a universal starting model for any machine learning project, it would likely be XGBoost. XGBoost, which stands for Extreme Gradient Boosting, is an exceptionally powerful and efficient machine learning algorithm. Developed by Tianqi Chen and Carlos Guestrin in 2016, XGBoost builds upon the Gradient Boosting Decision Tree (GBDT) framework, an ensemble method known for its iterative refinement of predictive accuracy. XGBoost enhances this framework with several optimization techniques that significantly boost both its efficiency and accuracy.

One of the key features of XGBoost is its ability to quickly access data through column block structures. This data organization method accelerates the training process by facilitating rapid data retrieval. Additionally, XGBoost employs out-of-core computation, which allows it to handle datasets larger than the system’s memory, making it particularly useful for large-scale data. Another notable feature is its parallel tree construction capability, enabling multi-threaded training that further speeds up the model-building process.

To prevent overfitting and enhance model generalization, XGBoost incorporates regularization techniques such as L1 (Lasso) and L2 (Ridge) regularization. These methods add penalties to the model's complexity, encouraging simpler and more general models. Due to these features, XGBoost has become a go-to tool for machine learning practitioners working on large-scale classification and regression tasks.

However, despite its strengths, XGBoost can be sensitive to noisy data and outliers. Proper data preprocessing, including cleaning and feature engineering, is essential to mitigate the risk of overfitting. Nonetheless, the remarkable efficiency, flexibility, and scalability of XGBoost make it an indispensable tool for achieving high predictive performance across a wide range of applications and domains. Its ability to handle complex, large-scale datasets with ease solidifies its reputation as a premier choice for machine learning projects.

2.2 Random Forest

Random Forest is a highly effective ensemble learning method introduced by Leo Breiman in 2001. This technique constructs multiple decision trees using a method known as bagging, which involves training each tree on a randomly bootstrapped subset of the data. Additionally, the algorithm selects a random subset of features for each tree, enhancing the diversity among the trees and reducing the risk of overfitting. The final predictions are made by aggregating the outputs from all the individual trees, leading to a robust and accurate outcome. One of the notable advantages of this approach is its ability to identify important features, which can be instrumental in understanding which variables significantly influence predictions.

However, despite its many strengths, Random Forest has some limitations. The training process can be computationally intensive due to the need to build and evaluate numerous trees, consuming considerable time and resources. Although the aggregation of multiple trees' predictions helps mitigate overfitting, it also results in a model that can be less interpretable, making it difficult to trace specific predictions back to their source. Furthermore, optimizing the model requires careful tuning of hyperparameters, such as the number of trees and the depth of each tree, to find the right balance for the specific dataset. Despite these challenges, Random Forest continues to be a popular choice for classification and regression tasks because of its robustness and effectiveness in various applications.

2.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm that was developed in the 1990s by Vladimir Vapnik and his colleagues. Its primary goal is to classify data by identifying an optimal hyperplane that best separates different classes within a high-dimensional space. The key concept of SVM is to maximize the margin, or the distance between the separating hyperplane and the nearest data points from each class, which are referred to as support vectors. This approach ensures a robust boundary that generalizes well to new data. SVM is versatile in handling both linear and non-linear boundaries through the use of kernel functions such as polynomial and radial basis function (RBF) kernels. These kernels effectively transform the original feature space into higher dimensions, where the classes become more easily separable.

However, SVM has its set of challenges. As the dataset size increases, the training process becomes computationally intensive, which can make it less suitable for very large datasets. Additionally, SVM requires careful parameter tuning to select the appropriate kernel function and the regularization parameter. Without proper tuning, the model's performance can degrade significantly. Another limitation is its sensitivity to imbalanced data, where the decision boundary may be biased towards the majority class, leading to suboptimal classification for the minority class. Despite these challenges, SVM remains a valuable tool, particularly for high-dimensional data problems. It has been successfully applied in various fields such as text classification, image recognition, and bioinformatics, demonstrating its efficacy in handling complex classification tasks.

2.4 AdaBoost

AdaBoost, short for Adaptive Boosting, is an influential ensemble learning technique designed to enhance classification performance by merging multiple weak learners into a single robust classifier. Developed by Freund and Schapire in 1996, AdaBoost operates through an iterative process that adjusts the weights assigned to each instance based on the classification accuracy of the previous iteration. This means that misclassified instances receive higher weights, compelling subsequent models to focus more on these challenging examples. This adaptive weighting strategy ensures that each new weak learner, typically a decision stump (a one-level decision tree), is better equipped to handle previously misclassified data points. The predictions from these weak learners are then combined using a weighted majority voting scheme, where each learner’s influence is proportional to its accuracy.

A major strength of AdaBoost is its capability to handle a variety of classification tasks with minimal parameter tuning. This adaptability allows it to reduce both bias and variance errors, leading to high predictive performance even in the presence of noisy data or when using weak base learners. Nevertheless, this same adaptability can make AdaBoost sensitive to noise and outliers, as the algorithm might overemphasize difficult samples, potentially leading to overfitting. Despite this vulnerability, AdaBoost remains a cornerstone in the field of machine learning. It not only provides a solid foundation for numerous ensemble methods but also serves as the underlying framework for more advanced boosting techniques. AdaBoost's ability to iteratively improve model accuracy by focusing on the hardest-to-classify examples exemplifies its robust approach to enhancing predictive performance across diverse datasets.

CHAPTER 3

Data and Data Preprocessing

Our dataset comprises fourteen variables and ten thousand rows, encompassing detailed customer profiles that include product interactions, demographics, and banking behaviors. Table 1 in the next page provides an overview of the dataset.

Data preprocessing for our analysis was straightforward due to the absence of missing values or duplicate rows. The initial step was to eliminate the 'RowNumber,' 'CustomerId,' and 'Surname' columns, which serve as labels for data tracking without offering any predictive value. Removing these columns reduces unnecessary computational load during model training.

Subsequently, categorical variables with fewer than four categories were converted into numerical representations using dummy encoding. This transformation ensures that the model can effectively interpret the categorical data. Lastly, all feature values were standardized to maintain consistency across variables and optimize the dataset for accurate modeling and analysis. This systematic approach lays a well-prepared foundation for subsequent machine learning tasks, ensuring that the data is clean, well-structured, and ready for in-depth analysis.

Table 1: Data Description

Variable Name	Description
RowNumber	A unique identifier for each row in the dataset.
CustomerId	Unique customer identification number.
Surname	The customer’s last name
CreditScore	The customer's credit score at the time of data collection.
Geography	The customer's country or region
Gender	The customer’s gender
Age	The customer’s age
Tenure	The number of years the customer has been with the bank.
Balance	The customer's account balance.
NumOfProducts	The number of products the customer has purchased or subscribed to.
HasCrCard	Indicates whether the customer has a credit card (1) or not (0).
IsActiveMember	Indicates whether the customer is an active member (1) or not (0).
EstimatedSalary	The customer's estimated salary.
Exited	The target variable, indicating whether the customer has churned (1) or not (0).

CHAPTER 4

Exploratory Data Analysis (EDA)

Before performing any statistical analysis, it is necessary to conduct a preliminary data analysis to see if there is any noticeable trend in the target variable. We first plotted Figure 1 to see the distribution of our target variable.

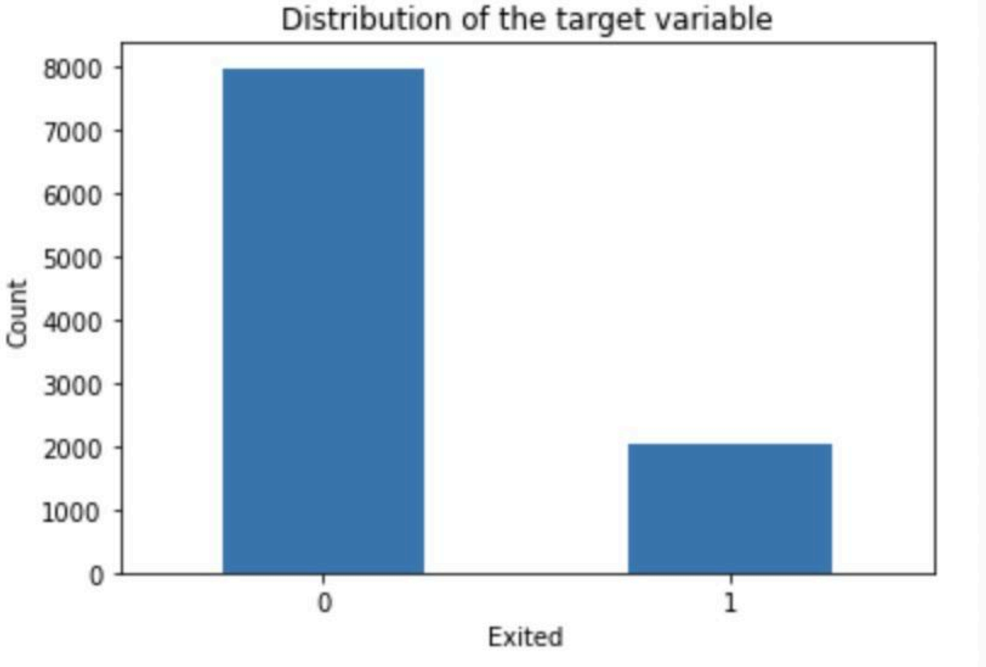


Figure 1: Distribution of Target Variable

The outcome reveals the distribution of our target variable, indicating that 80% of the customers in our dataset are non-churners, while the remaining 20% are churners. This suggests that our data suffers from imbalance.

4.1 Distribution of the Target by Categorical Variables.

In this section, we are going to get some insights about the relationships between our target variable and some categorical variables.

4.1.1 Geography and Target

Figure 2 presents a histogram depicting customer churn alongside geographical data. Initially, we did not anticipate that a customer's country or region would impact the churn rate. However, our exploratory data analysis (EDA) reveals that geographical location indeed plays a significant role. Specifically, customers in Germany have an approximate 50% churn rate, which is considerably higher than that observed in the other two European countries. This is a fairly interesting finding, we will dive deeper into this in the next section.

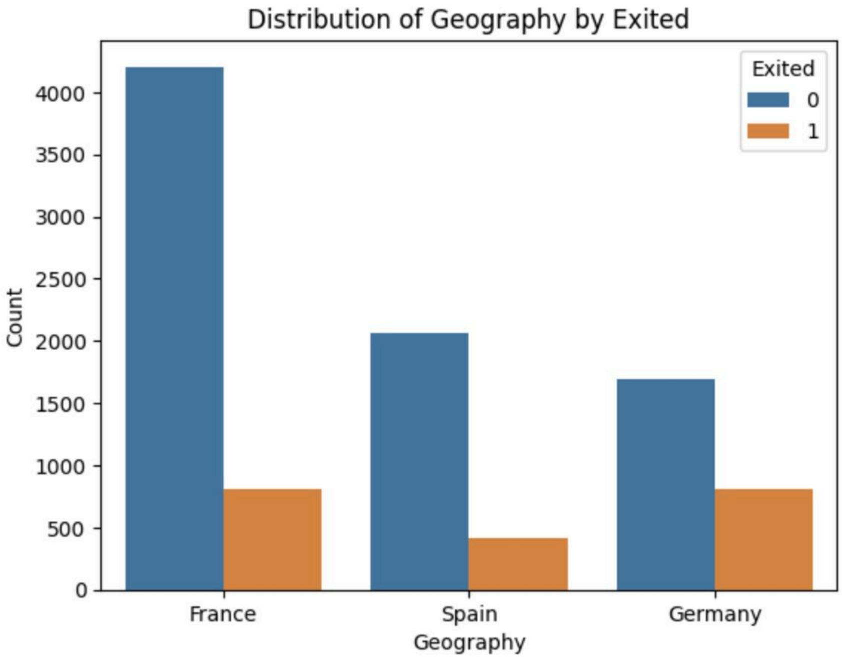


Figure 2: Histogram of Customer Churn and Customer Geographical Information

4.1.2 Gender and Target

Gender plays a crucial role in the study of human behavior due to the complex ways it shapes individuals' experiences, perceptions, and interactions with the world. Figure 3 illustrates the relationship between our target variable and gender. The data indicates that female customers generally exhibit a slightly higher churn rate than their male counterparts. This insight suggests that gender differences may play a role in customer retention and should be considered when analyzing behavioral patterns and developing tailored retention strategies.

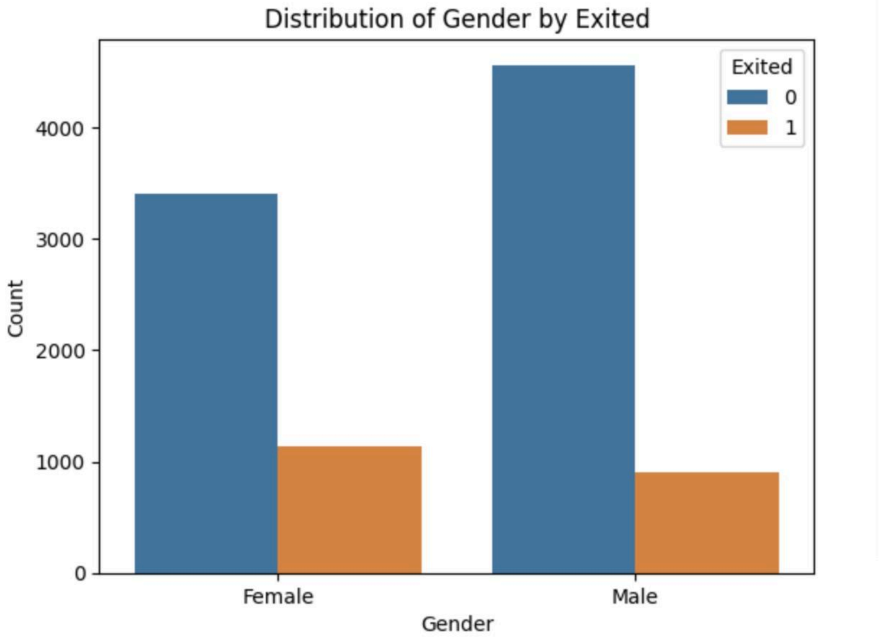


Figure 3: Distribution Plot of Gender by Target Variable

4.1.3 Has Credit Card and Target

Approximately 51% of people in Europe hold a credit card. Generally, having a credit card with a bank fosters a stronger relationship and is often seen as indicative of a lower likelihood of leaving the institution. However, as illustrated in Figure 4, the results challenge our initial assumptions: individuals with credit cards are just as likely to churn as those who do not possess one. This unexpected finding reveals that credit card ownership may not be as strong a retention indicator as initially presumed, calling for a more nuanced exploration of customer behavior.

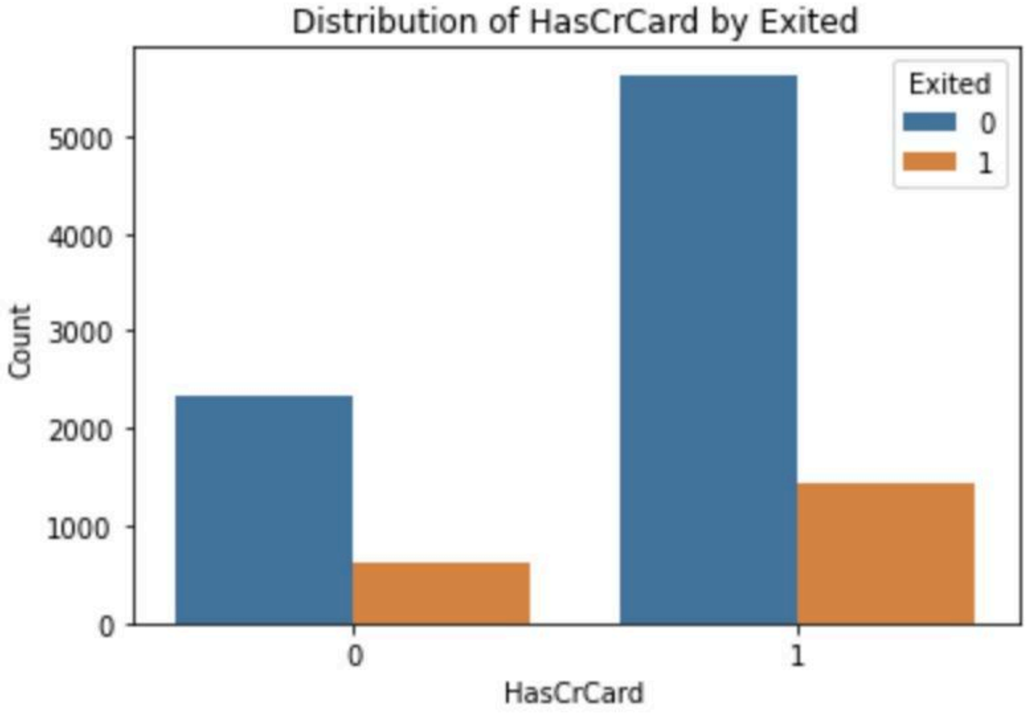


Figure 4: Distribution Plot of HasCrCard by Target Variable

4.1.4 IsActiveMember and Target

In Figure 5 below, we plotted the customer churn rate against active membership status, revealing notable differences in churn rates between active and non-active members. Nearly half of the non-active members ultimately left the bank, which is an alarmingly high proportion. This underscores the importance of keeping members actively engaged, as doing so is crucial for maintaining a healthy customer churn rate. By fostering active membership, banks can strengthen customer loyalty and significantly reduce the risk of attrition, leading to improved retention and long-term profitability.

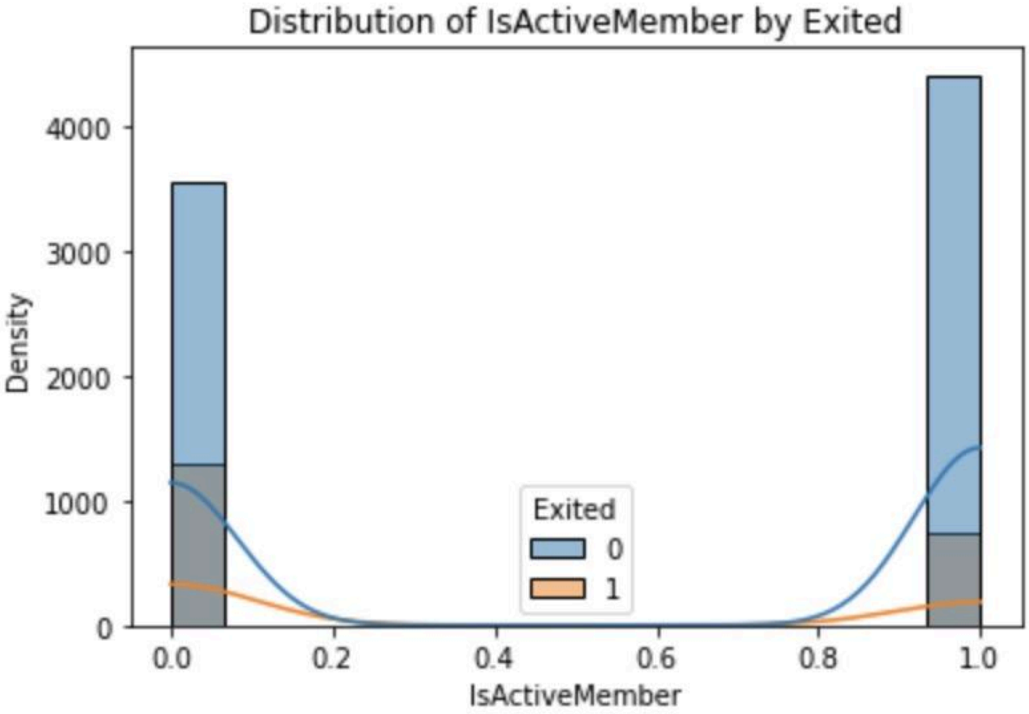


Figure 5: Distribution Plot of IsActiveMember by Target Variable

4.2 Distribution of the Target by Numerical Variables.

In the second part of our EDA, we are trying to detect some trends between continuous variables and our target.

4.2.1 Credit Score and Target

The first continuous variable we examined was the customer's credit score. We hypothesized that a person's credit score is a strong indicator of their credibility and financial status. Intuitively, individuals with low credit scores would be more prone to churn compared to those with higher credit scores. However, the findings presented in Figure 6 below contradict our initial assumptions.

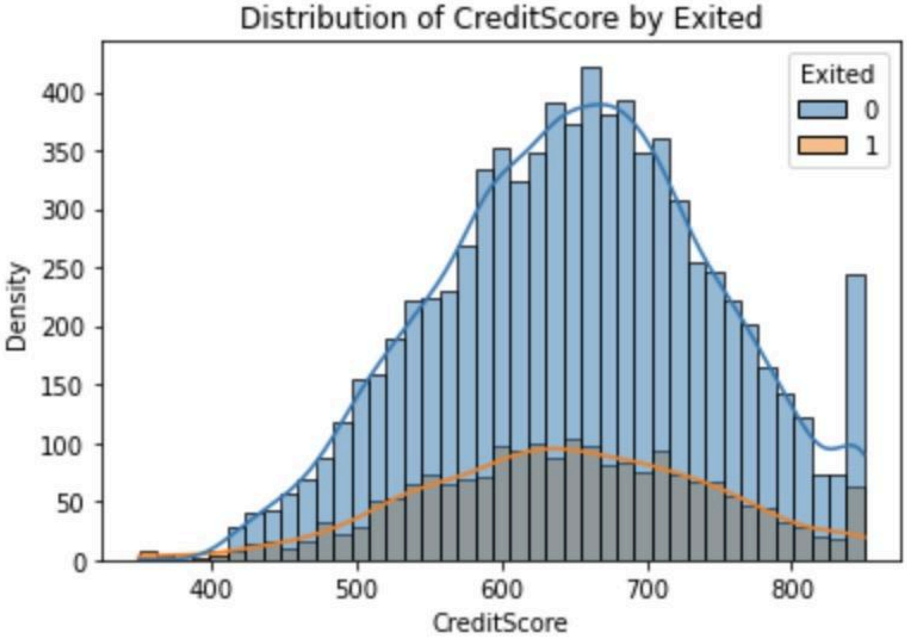
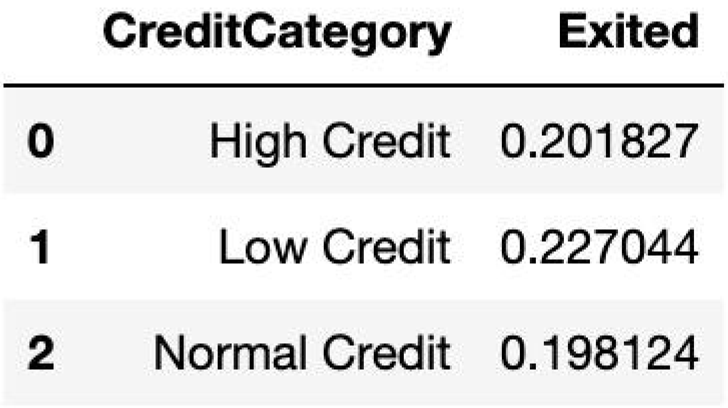


Figure 6: Distribution plot of CreditScore by Target Variable

To further investigate this, we classified customers into three groups based on their credit scores. The low credit score group includes individuals with scores below 550, the normal credit score group covers those with scores ranging from 551 to 720, and the high credit score group consists of customers with scores above 721. We then calculated the churn rate for each group. As shown in Table 2, the churn rates across these three groups do not significantly differ, challenging the assumption that credit score is a strong predictor of customer retention.



	CreditCategory	Exited
0	High Credit	0.201827
1	Low Credit	0.227044
2	Normal Credit	0.198124

Table 2: Credit Score Category Customer Churn Rate

4.2.2 Estimated Salary and Balance

Next, we considered the customer's financial situation as a potential influence on their churn rate. We visualized the relationship between the customer's estimated salary and bank balance against the target variable.

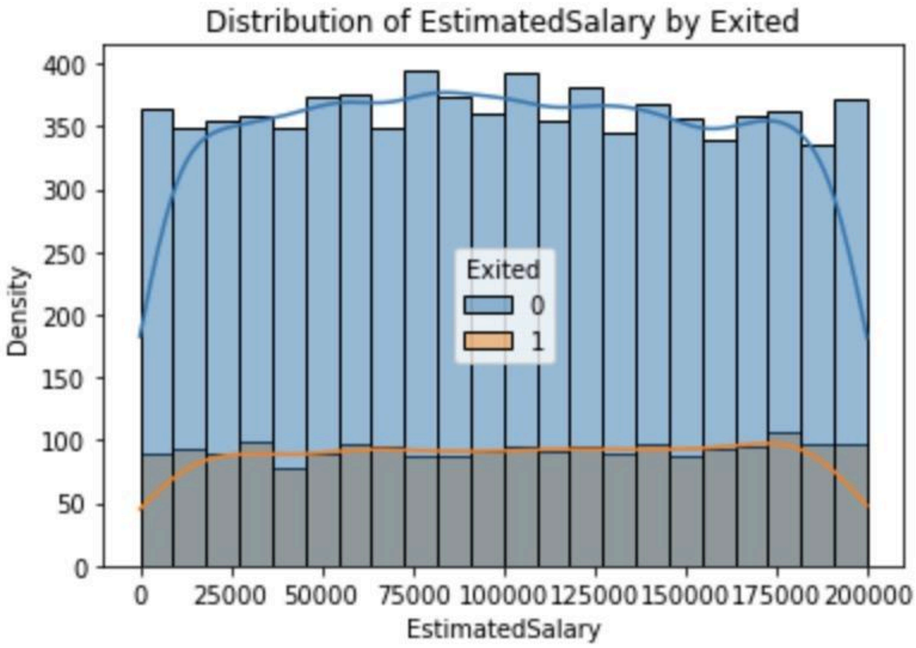


Figure 7: Distribution Plot of Estimated Salary by Target Variable

Figure 7 above illustrates that estimated salaries have almost no impact on customer churn rates. The churn rate remains consistent across all salary ranges, showing minimal variation regardless of income level.

On the other hand, in Figure 8 below, we observed that people with bank balances between 80,000 and 150,000 have higher churn rate than others. Our initial guess for this phenomenon is because of the benefits of high amounts of saving premium accounts. Different banks have different strategies to attract and maintain these premium clients.

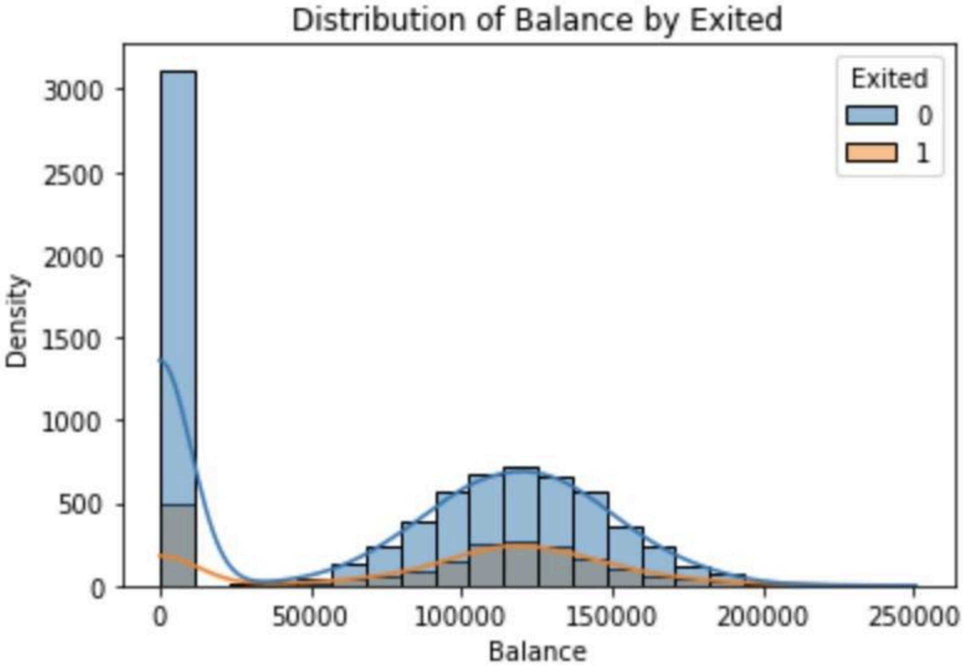


Figure 8: Distribution Plot of Balance by Target Variable

4.2.3 Age and the Target

Age is something that we cannot ignore when we are doing any sort of human behavioral study. People from different age groups think and act differently. We then plotted Figure 9 below to have an overview of the relationship between age and the churn rate.

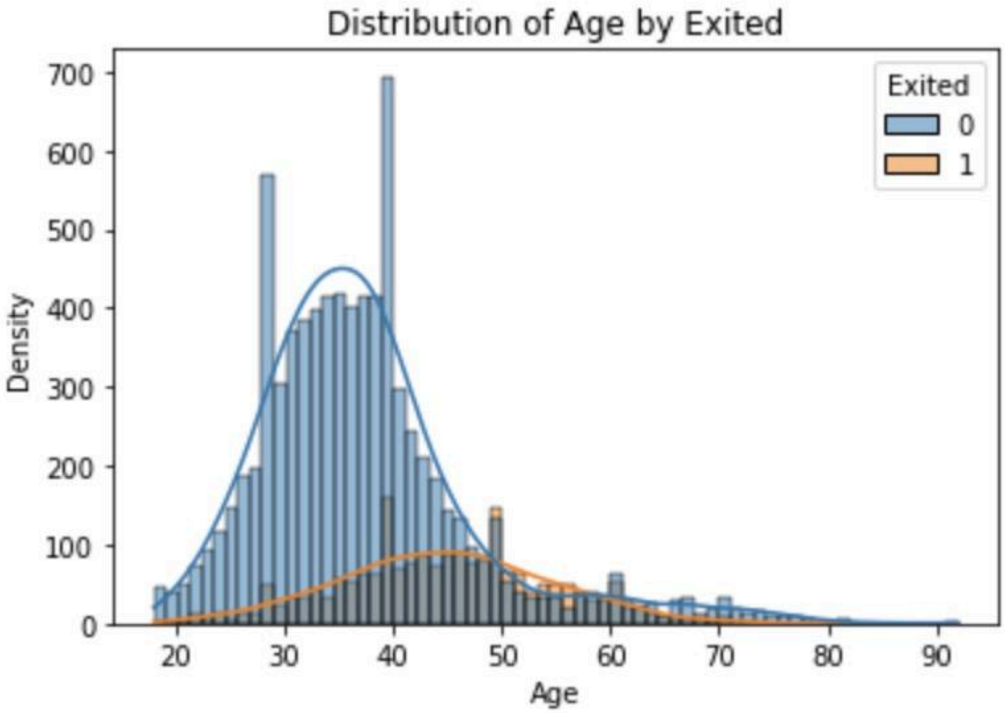
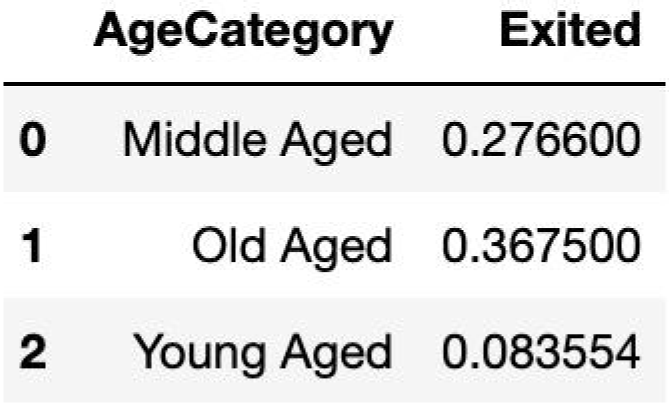


Figure 9: Distribution Plot of Age by Target Variable

The result turned out to be very surprising. People from older age groups have way higher churn rate than people in the younger age group. We then divided our customers into three age groups to further investigate this.

We have a young aged group that contains people that are 35 years old or younger, the middle aged group consists of people between the ages of 36 to 55, and the older aged group has people who are older than 56.



	AgeCategory	Exited
0	Middle Aged	0.276600
1	Old Aged	0.367500
2	Young Aged	0.083554

Table 3: Age Groups Churn Rate

In Table 3 above, we can clearly see that the churn rate differs significantly between different age groups. This could also be a result of high competitive force between banks as different banks have different benefits for senior customers.

4.2.4 Number of Products and Tenure

A strong attachment and connection to the bank significantly reduces the likelihood of customer churn. Two key indicators of customer connectivity include the number of products purchased or subscribed to and the number of years the customer has maintained a relationship with the bank. For instance, customers who have several bank products are often more integrated into the bank's ecosystem, making them less inclined to seek alternatives. Likewise, long-term customers tend to have a more established rapport, familiarity, and trust with the bank, leading to greater loyalty.

As shown in Figure 10, customers holding two bank products are the least likely to churn. This suggests a sweet spot where customers are deeply engaged yet not overwhelmed by multiple offerings. Banks can leverage this insight by encouraging customers to diversify their product holdings, strategically offering additional services that align with individual needs. Building on this knowledge, enhancing customer engagement through personalized product recommendations and targeted retention strategies can foster loyalty, deepen relationships, and ultimately reduce churn rates.

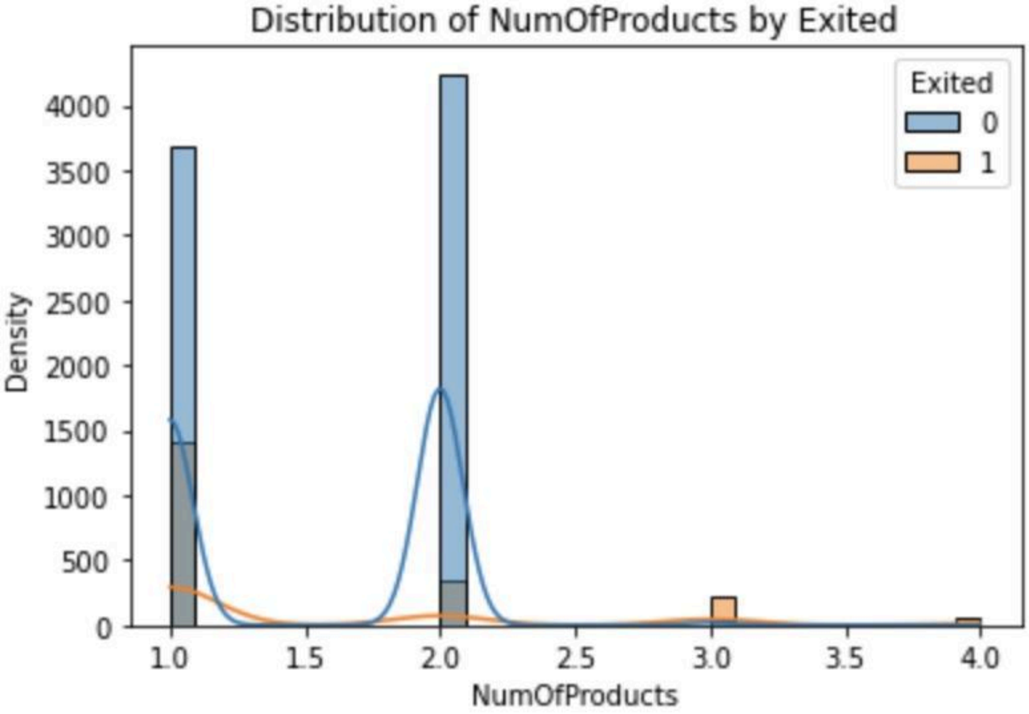


Figure 10: Distribution Plot of NumOfProducts by Target Variable

In Figure 11 below, the histogram of tenure and churn rate, we found out that churn rate is very similar across all different tenure years.

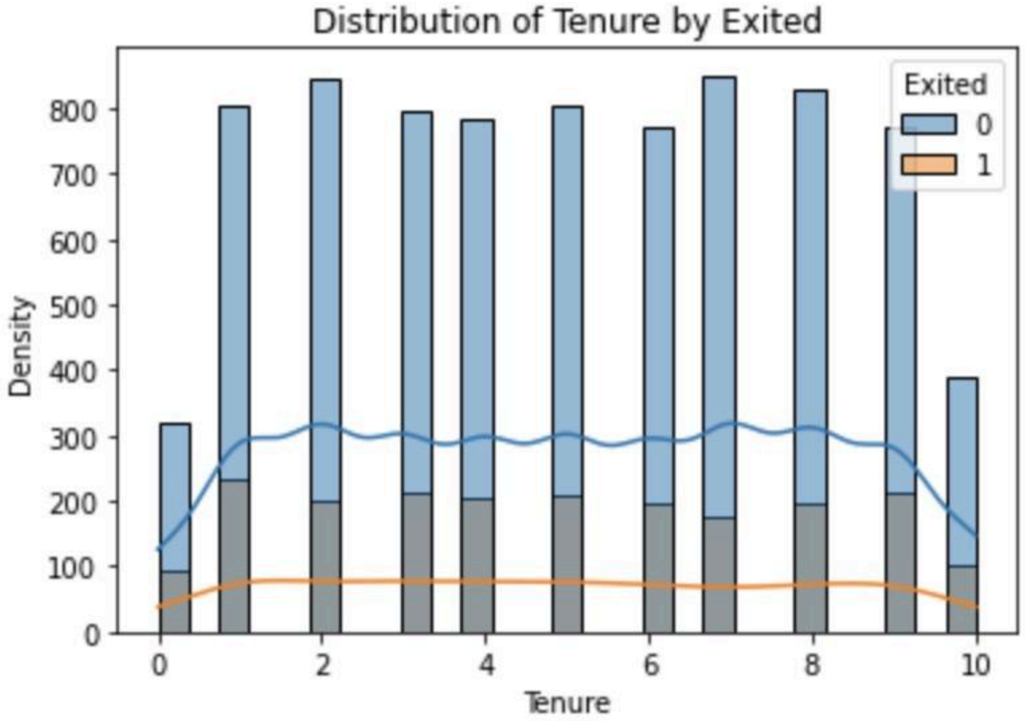
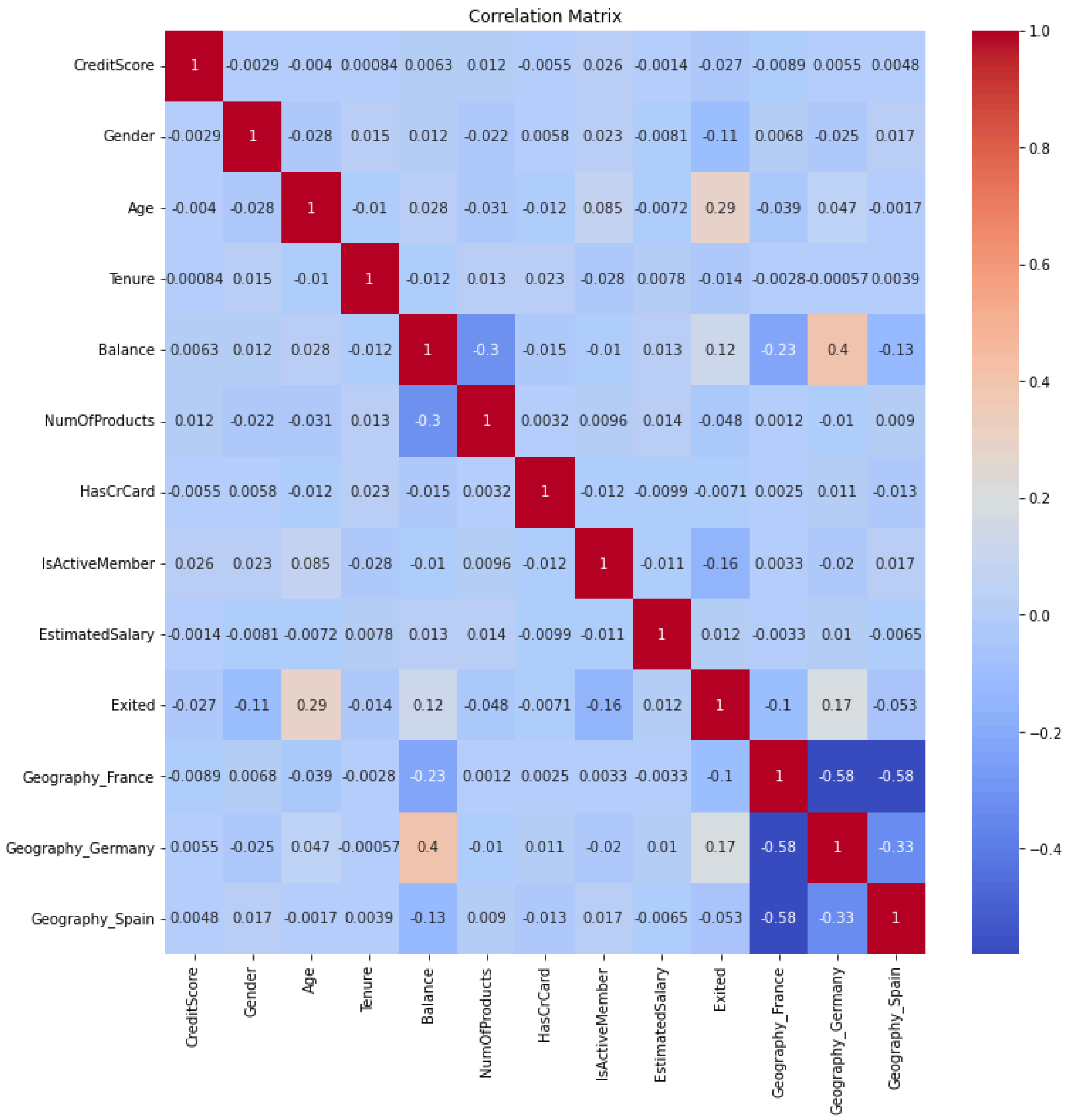


Figure 11: Distribution Plot of Tenure by Target Variable

4.3 EDA Conclusion

Finally, we plotted a correlation matrix to have an overview of all variable relationships. From Figure 12, we conclude that age, geographical location, and active member status are the most influential factors to our target variable. Gender, number of products, and the amount of balance saved in the bank are also significant factors. Other variables that we initially thought might be quite impactful to our target variable turned out to have very small impact on the customer churn rate. We have conducted more statistical analysis to further confirm this early conclusion in our next section.



	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Geography_France	Geography_Germany	Geography_Spain
CreditScore	1	-0.0029	-0.004	0.00084	0.0063	0.012	-0.0055	0.026	-0.0014	-0.027	-0.0089	0.0055	0.0048
Gender	-0.0029	1	-0.028	0.015	0.012	-0.022	0.0058	0.023	-0.0081	-0.11	0.0068	-0.025	0.017
Age	-0.004	-0.028	1	-0.01	0.028	-0.031	-0.012	0.085	-0.0072	0.29	-0.039	0.047	-0.0017
Tenure	0.00084	0.015	-0.01	1	-0.012	0.013	0.023	-0.028	0.0078	-0.014	-0.0028	-0.00057	0.0039
Balance	0.0063	0.012	0.028	-0.012	1	-0.3	-0.015	-0.01	0.013	0.12	-0.23	0.4	-0.13
NumOfProducts	0.012	-0.022	-0.031	0.013	-0.3	1	0.0032	0.0096	0.014	-0.048	0.0012	-0.01	0.009
HasCrCard	-0.0055	0.0058	-0.012	0.023	-0.015	0.0032	1	-0.012	-0.0099	-0.0071	0.0025	0.011	-0.013
IsActiveMember	0.026	0.023	0.085	-0.028	-0.01	0.0096	-0.012	1	-0.011	-0.16	0.0033	-0.02	0.017
EstimatedSalary	-0.0014	-0.0081	-0.0072	0.0078	0.013	0.014	-0.0099	-0.011	1	0.012	-0.0033	0.01	-0.0065
Exited	-0.027	-0.11	0.29	-0.014	0.12	-0.048	-0.0071	-0.16	0.012	1	-0.1	0.17	-0.053
Geography_France	-0.0089	0.0068	-0.039	-0.0028	-0.23	0.0012	0.0025	0.0033	-0.0033	-0.1	1	-0.58	-0.58
Geography_Germany	0.0055	-0.025	0.047	-0.00057	0.4	-0.01	0.011	-0.02	0.01	0.17	-0.58	1	-0.33
Geography_Spain	0.0048	0.017	-0.0017	0.0039	-0.13	0.009	-0.013	0.017	-0.0065	-0.053	-0.58	-0.33	1

Figure 12: Variable Correlation Matrix

CHAPTER 5

Statistical Modeling

This section, we introduce our model evaluation metrics, validation & hyperparameters, and model results.

5.1 Model Evaluation

To make fair comparisons and to choose the best between our models. We decided to use accuracy score and ROC AUC curve as our model evaluation metrics.

Accuracy score measures the proportion of correct predictions made by the model compared to the total number of predictions. In mathematical terms, accuracy score is given by: $Accuracy = Number\ of\ Correct\ Predictions/Total\ Number\ of\ Predictions$.

An ROC curve stands for receiver operating characteristic curve. It is a graphical representation used to evaluate the performance of a binary classification model. It plots two metrics across different decision thresholds: True Positive Rate (TPR) and False Positive Rate (FPR).

Table 4 below provides a confusion matrix, allowing key evaluation metrics to be calculated.

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

Table 4: Confusion Matrix

True Positive Rate (TPR): Also known as sensitivity or recall, it measures the proportion of actual positives correctly identified. It is calculated as: $TPR = TP/ TP + FN$.

False Positive Rate (FPR): This metric measures the proportion of actual negatives incorrectly classified as positives. It is calculated as: $FPR = FP / FP + TN$.

AUC stands for the area under the ROC curve. It provides a single output value that summarizes the overall performance of the classifier. An AUC of 1.0 indicates a perfect classifier, while 0.5 suggests a model no better than random guessing. From a probabilistic standpoint, an AUC of 0.8 indicates a 80% chance that a randomly selected churner will be ranked higher than a randomly selected non-churner.

5.2 Validation and Hyperparameters

The dataset is divided into three parts training, validation, and testing subsets—70% for training, 10% for validation, and 20% for testing—enabling a systematic approach to model evaluation and parameter tuning. Employing a ten-fold cross-validation technique on this split further enhanced the reliability of our model selection process.

Table 5 offers a comprehensive breakdown of all hyperparameters scrutinized for each algorithm during the rigorous selection procedure. By exploring a range of hyperparameter combinations, we aimed to identify the configurations that yielded optimal performance metrics.

Leveraging the optimal hyperparameters uncovered through this meticulous process, we executed our models on the testing set to gauge their real-world performance accurately. This meticulous fine-tuning ensured that each model was honed to its utmost potential, delivering robust and dependable insights into their respective capabilities.

Such a methodical approach not only enhances the accuracy and reliability of our predictive models but also instills confidence in the validity of the conclusions drawn from their performance assessments.

Table 5: Machine Learning Algorithm Hyperparameters.

Algorithm	Hyperparameter	Value
Xgboost	n_estimators	[100, 150, 200, 250]
	max_depth	[3, 5, 7, 10]
	gamma	[0, 0.01, 0.05, 0.1]
	eta	[0.01, 0.05, 0.1, 0.2]
Random Forest	n_estimators	[50, 100, 150, 200]
	max_depth	[3, 5, 7]
	criterion	[gini, entropy]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
SVM	C	[0.1, 1, 10, 20]
	kernel	[linear, rbf]
	gamma	[0.001, 0.01, 0.1]
AdaBoost	n_estimator	[50, 100, 150, 200]
	learning_rate	[0.1, 1, 5, 10]
	algorithm	[SAMME, SAMME.R]

5.3 Model Result

In this section, we present and analyze the results from an extensive benchmark study, placing emphasis on predictive performance. Additionally, we conduct a thorough feature importance analysis to ascertain whether the model's outcomes align with our early EDA conclusions. This comprehensive approach ensures that our evaluation provides not only accurate predictions but also insights into the most influential factors driving these predictions.

5.3.1 Predictive Performance

Results from these four supervised machine learning models are summarized in Table 6 below.

Metrics	XGBoost	RF	SVM	AdaBoost
Train AUC	0.9290	0.8899	0.6693	0.8537
Test AUC	0.8696	0.8676	0.6781	0.8572
Train Accuracy	0.8924	0.8699	0.7945	0.8594
Test Accuracy	0.8655	0.8615	0.8035	0.8600

Table 6: Model Predictive Performance Overview

From the result, we observed XGBoost model achieved the highest predicting accuracy of 86.55%, followed very closely by Random Forest model by 86.15% and AdaBoost model by 86%. In terms of test AUC, XGBoost is still in the leading position achieving AUC of 0.8696, followed up with Random Forest of 0.8676. Therefore, XGBoost is the most optimal model to use in our scenario.

5.3.2 Feature Importance Analysis

To gain better interpretability of the machine learning models, it is necessary to conduct feature importance analysis. Figure 13 below is a feature importance plot from the XGBoost model.

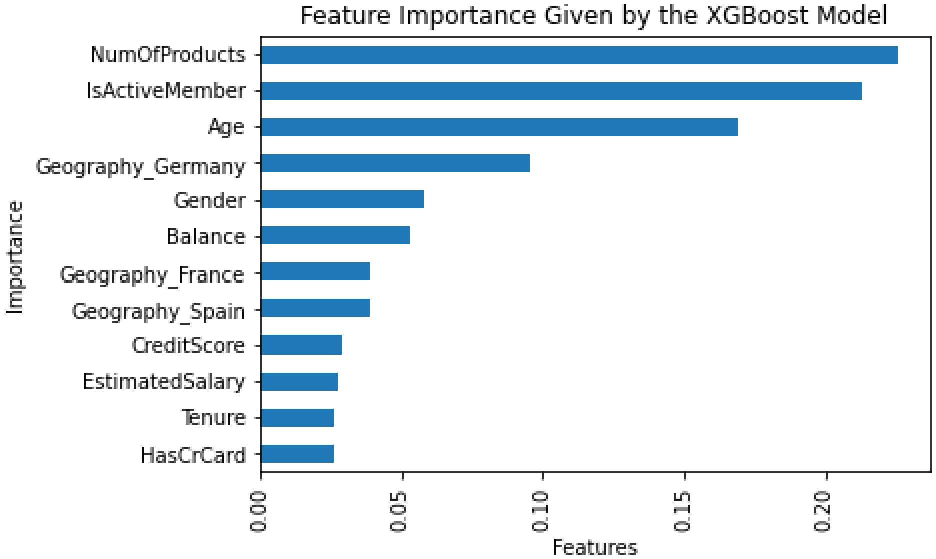


Figure 13: Feature Importance Plot of XGBoost

The results align with our initial conclusions drawn from exploratory data analysis (EDA). The number of products purchased or subscribed to from the bank, active membership status, customer age, and geographical location have proven to be the four most influential factors. Of these, the number of products holds the greatest influence in our model, as customers who own two products exhibit a much lower churn rate than those who possess only one product or more than two. Banks should consider promoting attractive products to clients who currently own only one product, while refraining from promoting additional products to those already holding two

Active membership status emerges as the second most critical factor, as non-active members exhibit nearly double the churn rate of active members, as highlighted in the EDA. Our analysis also reveals that individuals in older age groups have a significantly higher churn rate than those in younger and middle-aged groups. To retain senior clients, banks should develop tailored strategies, such as retirement plans and improved interest rates.

An intriguing discovery is the elevated churn rate among customers in Germany compared to other European nations. Additionally, customer gender and bank balance are notable factors, as female customers are more likely to churn than their male counterparts. Furthermore, customers with bank balances between 80,000 and 150,000 tend to have higher churn rates, likely due to the competitive banking market. These customers, often considered premium users, are highly valuable to banks. Offering superior services or benefits to retain their loyalty is a strategic approach to reducing churn rates among this demographic.

CHAPTER 6

Conclusion and Future Work

This study effectively predicted customer churn within the banking industry, providing valuable insights that banks can leverage to enhance customer retention through targeted business strategies. For instance, offering incentives and rewards to clients with multiple bank products can encourage loyalty and solidify customer relationships. Tailoring policies to address the specific needs of senior clients can also be instrumental in retaining this demographic, while promotional initiatives aimed at increasing member activity will strengthen customer engagement.

Furthermore, banks should prioritize understanding and meeting the unique preferences of female customers to reduce their churn rates. By identifying clients with substantial account balances, banks can proactively offer premium account benefits or personalized financial solutions to minimize the risk of churning. These proactive measures will empower banks to adopt a more comprehensive approach to customer retention, fostering loyalty, improving satisfaction, and reducing churn rates across the board.

There remain avenues to enhance our model moving forward. Our research focuses on customer behavioral patterns in banking, which are both sensitive and private, resulting in limited data accessibility. However, gaining access to additional data points would bolster the generalizability of our predictions. Furthermore, obtaining more granular data would significantly refine the model's predictive accuracy, enabling us to deliver more precise forecasts. Expanding data availability and granularity would ultimately empower us to understand customer behavior more comprehensively, paving the way for innovative retention strategies and deeper insights into churn drivers.

References

1. scikit-learn. (n.d.). sklearn.metrics.accuracy_score. Retrieved May 1, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

2. XGBoost Documentation. (n.d.). Parameters. Retrieved May 1, 2024, from https://xgboost.readthedocs.io/en/stable/parameter.html

3. scikit-learn. (n.d.). sklearn.ensemble.RandomForestClassifier. Retrieved May 1, 2024 , from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

4. Google Developers. (n.d.). ROC and AUC. In Machine Learning Crash Course. Retrieved May 1, 2024, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

5. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

6. Rigatti, Steven J. "Random forest." Journal of Insurance Medicine 47.1 (2017): 31-39.

7. Suthaharan, Shan, and Shan Suthaharan. "Support vector machine." Machine learning models and algorithms for big data classification: thinking with examples for effective learning (2016): 207-235.

8. Ying, Cao, et al. "Advance and prospects of AdaBoost algorithm." Acta Automatica Sinica 39.6 (2013): 745-758.

9. Pahul Preet Singh, Fahim Islam Anik, Rahul Senapati, Arnav Sinha, Nazmus Sakib, Eklas Hossain, Investigating customer churn in banking: a machine learning approach

and visualization app for data science and management, Data Science and Management, Volume 7, Issue 1, 2024, Pages 7-16, ISSN 2666-7649, https://doi.org/10.1016/j.dsm.2023.09.002.