FINAL PROJECT REPORT

Predicting Bank Customer Churn Using Machine Learning Models

Athaudage Sanuja Vihanga Senadeera, Justin Kyle Pedro, Mayra Geraldine Reinoso Varon,

Naveen Karan Krishna, Thi Anh Tram Le

BAN230NAA: Applied Data Mining and Modelling

Professor: Savita Seharawat

Due Date: 10th August 2025

Table of Contents

Executive Summary

This project, titled “Predicting Bank Customer Churn Using Machine Learning Models,” focuses on identifying which bank customers are most likely to leave and uncovering the key factors that drive this behavior. Using a publicly available synthetic dataset from Kaggle, we followed a structured data science workflow involving exploratory data analysis, data preprocessing, class balancing with SMOTE, and supervised model training. Three machine learning models including Logistic Regression, Decision Tree, and Random Forest were developed and evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Among the models, Random Forest performed best, achieving the highest accuracy and ROC-AUC score, indicating strong predictive power. Feature importance analysis revealed that IsActiveMember, Geography, and Age were among the most influential predictors of churn. These insights not only help prioritize customer engagement efforts but also inform targeted retention strategies using the most effective algorithm identified in this study.

Problem Statement

Customer churn continues to be a critical challenge in the banking industry, directly impacting revenue, customer lifetime value, and overall profitability. Since retaining existing customers is generally more cost-effective than acquiring new ones, banks must identify early indicators of dissatisfaction to prevent churn. This project addresses that need by developing machine learning models that can classify whether a customer is likely to leave based on historical account data and demographic attributes. Specifically, the project has three key objectives: first, to build and evaluate three supervised learning models including Logistic Regression, Decision Tree, and Random Forest to predict churn; second, to identify the most influential features that drive customer churn decisions, using model-based insights such as feature importance and coefficients; and third, to explore how demographic factors like gender, geography and age correlate with churn behavior. Together, these efforts aim to provide banks with actionable, data-driven insights that support targeted customer retention strategies.

Dataset Overview

The dataset employed in this project is the “Bank Customer Churn Prediction” dataset developed by Shubham Meshram and publicly available on Kaggle. It contains approximately 10,000 records, each representing an individual bank customer, with a mix of both numerical and categorical features that describe their demographic profile and banking behavior. Among the key numerical attributes are CreditScore, Age, Tenure with the bank, Account Balance, Number of Products held, and Estimated Salary. Categorical variables include Geography and Gender, while binary features such as HasCrCard and IsActiveMember indicate credit card ownership and account activity status, respectively. The target variable, Exited, serves as the churn indicator, where a value of 1 signifies that the customer has left the bank and 0 indicates they remain active. Although the dataset

is synthetic, it has been carefully modeled to mimic real-world customer distributions and behavioral patterns observed in retail banking environments. Initial exploratory analysis revealed a class imbalance in the target variable, with roughly 20% of customers having churned, underscoring the importance of techniques such as SMOTE to address this imbalance during model training (Meshram, 2023).

Methodology

The methodology adopted in this project was inspired by the CRISP-DM framework and comprised six distinct phases designed to ensure a systematic approach to solving the churn prediction problem. The initial phase involved a thorough understanding of the business context and the challenges posed by customer churn in the banking sector. Following this, the data exploration phase focused on examining the dataset's structure, identifying missing values, and understanding feature distributions. During data preparation, several preprocessing steps were executed to ready the data for modeling: irrelevant columns such as 'Surname,' 'RowNumber,' and 'CustomerId' were removed to avoid noise and overfitting; categorical variables including 'Geography' and 'Gender' were transformed into numerical format using one-hot encoding; and numerical features such as 'CreditScore,' 'Age,' and 'Balance' were standardized to bring them onto a comparable scale.

Addressing the inherent class imbalance in the target variable was a critical step, since only about 20% of customers had churned, the Synthetic Minority Oversampling Technique (SMOTE) was applied solely to the training dataset. This method synthetically increased minority class examples, enabling the models to learn balanced decision boundaries and reducing bias toward the majority class. Subsequently, three supervised classification algorithms were developed and trained: Logistic Regression, which included tuning of the regularization parameter to prevent overfitting;

Decision Tree Classifier, where the maximum depth was limited to control complexity; and Random Forest Classifier, which combined multiple trees with parameters such as number of estimators and max depth adjusted for optimal performance. Each model was trained on the balanced training data and evaluated on a separate test set to assess generalizability.

For model evaluation, multiple performance metrics were computed, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). Confusion matrices were also analyzed to understand the distribution of true positives, false positives, true negatives, and false negatives across models. ROC curves for all three classifiers were plotted together for visual comparison of their discrimination power. Finally, feature importance scores were extracted from each model to identify which variables had the most significant impact on predicting churn, aiding in interpretability and actionable business insights.

Findings and Interpretations

Exploratory Data Analysis (EDA) uncovered several important patterns regarding customer churn behavior in the dataset. Analysis by gender showed that females had a notably higher churn rate compared to males, with approximately 25.1% of female customers leaving the bank versus 16.5% of male customers. Geographic location was another influential factor; customers residing in Germany exhibited the highest churn rate at around 32.5%, whereas churn rates for customers in France and Spain were lower, at approximately 16.2% and 16.7%, respectively. Age was found to have a significant relationship with churn, with the highest churn rates concentrated in the 41 to 60 age groups. Specifically, customers aged 41–50 and 51–60 had churn rates exceeding 34% and 56%, respectively, indicating that middle-aged to older customers were more likely to exit the bank. Additional factors influencing churn included customer activity status and product usage; customers who were inactive (IsActiveMember = 0) showed a higher tendency to churn, and those

holding three or four products also demonstrated elevated churn rates compared to customers with fewer products. These insights were visualized through bar plots and grouped churn averages, providing clear and interpretable views of the key drivers behind customer attrition.

Model performance evaluation revealed that among the three classifiers tested including Logistic Regression, Decision Tree, and Random Forest, the Random Forest model consistently outperformed the others across multiple metrics. It achieved the highest accuracy, recall, and ROC-AUC score, indicating its superior ability to correctly identify customers likely to churn while minimizing false negatives. The Decision Tree classifier showed moderate performance, surpassing Logistic Regression in recall and F1-score but lagging slightly behind Random Forest. Logistic Regression, while exhibiting a slightly lower overall predictive capability, provided valuable interpretability through its coefficients.

Table 1: Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
Logistic Regression	0.7230	0.6549	0.7215	0.6603	0.7825
Decision Tree	0.7683	0.6793	0.7330	0.6937	0.8272
Random Forest	0.8370	0.7499	0.7700	0.7590	0.8619

Feature importance analysis supported these findings, with tree-based models consistently identifying Age, IsActiveMember status, and Balance as the most influential predictors of churn. Logistic Regression coefficients similarly highlighted these features as significant, although with varying directional impacts. The ROC curves for all three models demonstrated reasonable separability between churned and retained customers, with Random Forest achieving the highest area under the curve (AUC), confirming its robustness in this classification task.

Business Recommendations

Based on the insights uncovered through exploratory data analysis and predictive modeling, several practical and targeted business strategies are recommended to effectively reduce customer churn. First and foremost, customers located in high-churn regions, particularly Germany, should be considered a priority segment for retention efforts. Tailored engagement campaigns focused on this geographic group could include personalized communication or targeted offers to address region-specific concerns and strengthen customer relationships. Additionally, inactive customers those identified by the IsActiveMember feature as not currently engaged with the bank's services represent a critical at-risk group. Proactive outreach programs, such as personalized financial advice, reminders of product benefits, or incentives to increase account activity, can help re-engage these dormant customers and reduce their likelihood of leaving.

Furthermore, the analysis revealed that customers with a higher number of products (three or more) tend to exhibit increased churn rates, which may indicate potential service fatigue, complexity, or dissatisfaction with managing multiple products simultaneously. To address this, banks should consider simplifying product bundles or improving cross-product integration to enhance the overall experience and reduce confusion or frustration.

Finally, incorporating predictive modeling into the bank's operational systems is recommended to ensure early identification and intervention for at-risk customers. Specifically, deploying the Random Forest model within customer relationship management (CRM) platforms can enable automated flagging of high-risk customers based on their demographic and behavioral profiles. This will allow retention teams to efficiently prioritize follow-ups, design personalized retention strategies, and allocate resources effectively, ultimately improving customer loyalty and reducing churn-related losses.

Limitations

Although this project provides meaningful insights into predicting bank customer churn, several limitations must be recognized. Primarily, the dataset used is synthetic and designed to emulate realistic banking customer data, but it does not capture the full complexity and nuances of real-world customer behavior. This synthetic nature inherently restricts the extent to which the model findings can be generalized to actual banking environments. Furthermore, the available dataset includes only demographic and account-related features; important behavioral factors such as transaction frequency, customer service interactions, or digital engagement metrics were not part of the data and thus could not be leveraged to improve prediction accuracy. Another consideration is the use of SMOTE to address the class imbalance in churn outcomes. While SMOTE helps in balancing the training data and reducing model bias toward the majority class, generating synthetic samples may also introduce some distortions or bias, potentially affecting the robustness of model predictions. Finally, the project focused on three commonly used machine learning models without exploring more advanced or ensemble techniques that might offer improved performance.

Conclusion

This project demonstrated a comprehensive approach to predicting bank customer churn by applying machine learning models within a well-structured data science framework. The process involved meticulous data preprocessing, exploratory analysis to identify churn-related patterns, and the training and evaluation of three classification algorithms: Logistic Regression, Decision Tree, and Random Forest. The results highlighted key features such as age, account activity status, and balance as significant drivers of churn, reinforcing insights from exploratory data analysis. Among the tested models, the Random Forest classifier consistently delivered superior performance across multiple evaluation metrics, making it the recommended choice for practical

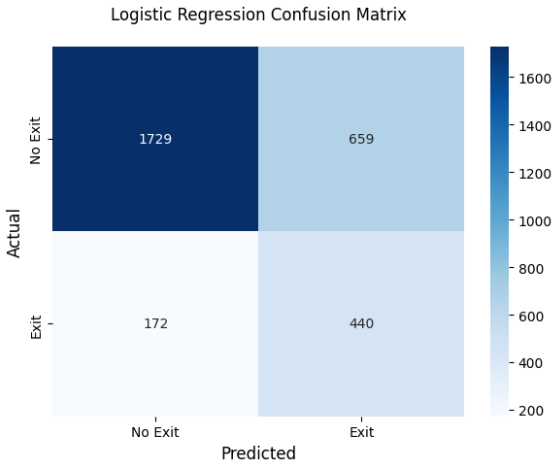
deployment. Implementing this model in banking systems would enable more precise identification of customers at risk of leaving, allowing institutions to proactively tailor retention strategies, optimize resource allocation, and ultimately improve customer loyalty and profitability over time. Despite some dataset and methodological limitations, the project underscores the value of predictive analytics in addressing critical business challenges like customer churn.

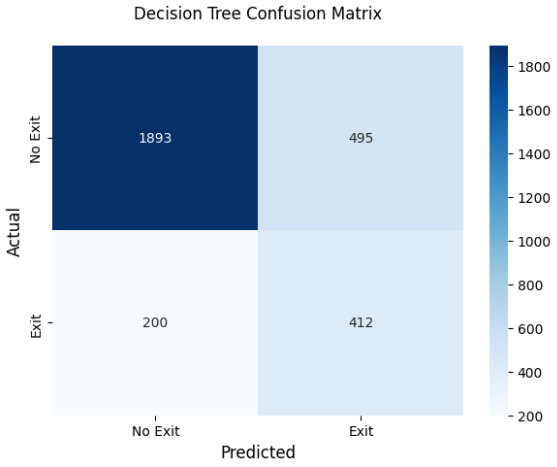
Reference

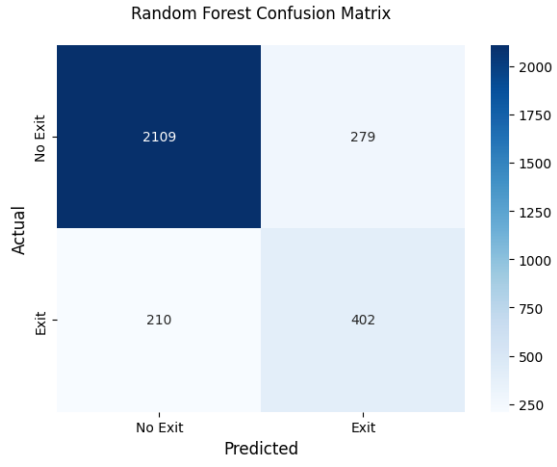
Meshram, S. (2023). *Bank Customer Churn Prediction*. Kaggle. https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction

Appendices

Appendix A

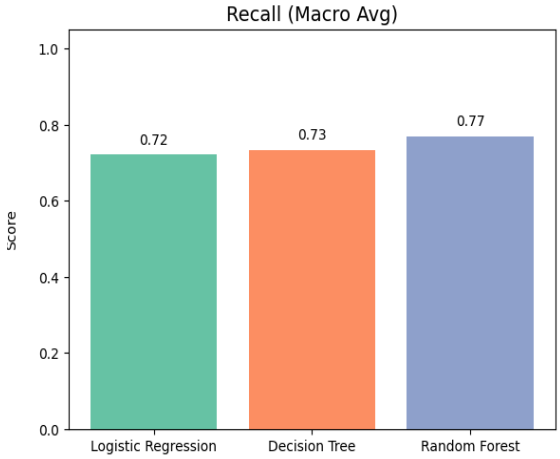


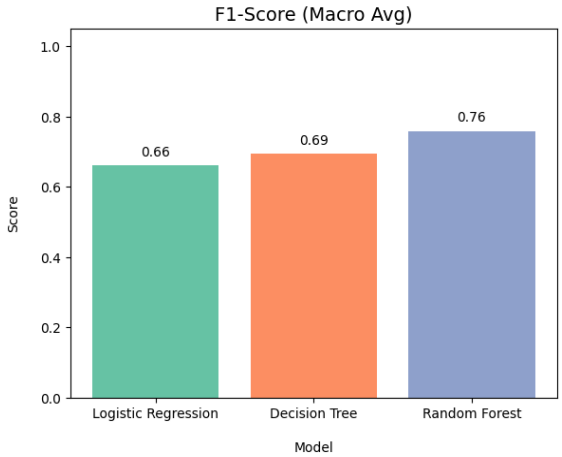




Appendix B

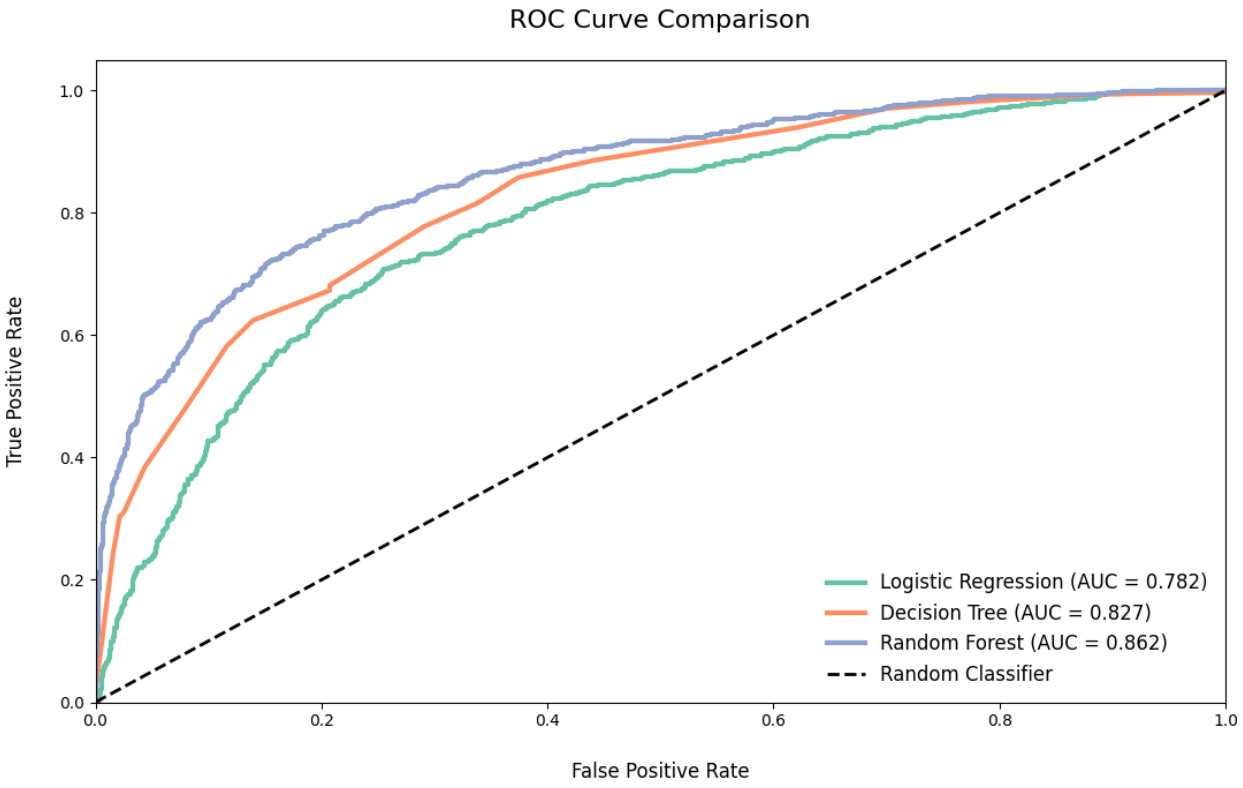

Precision (Macro Avg)
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.65
0.68
0.75
Logistic Regression
Decision Tree
Random Forest
Model


Recall (Macro Avg)
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.72
0.73
0.77
Logistic Regression
Decision Tree
Random Forest

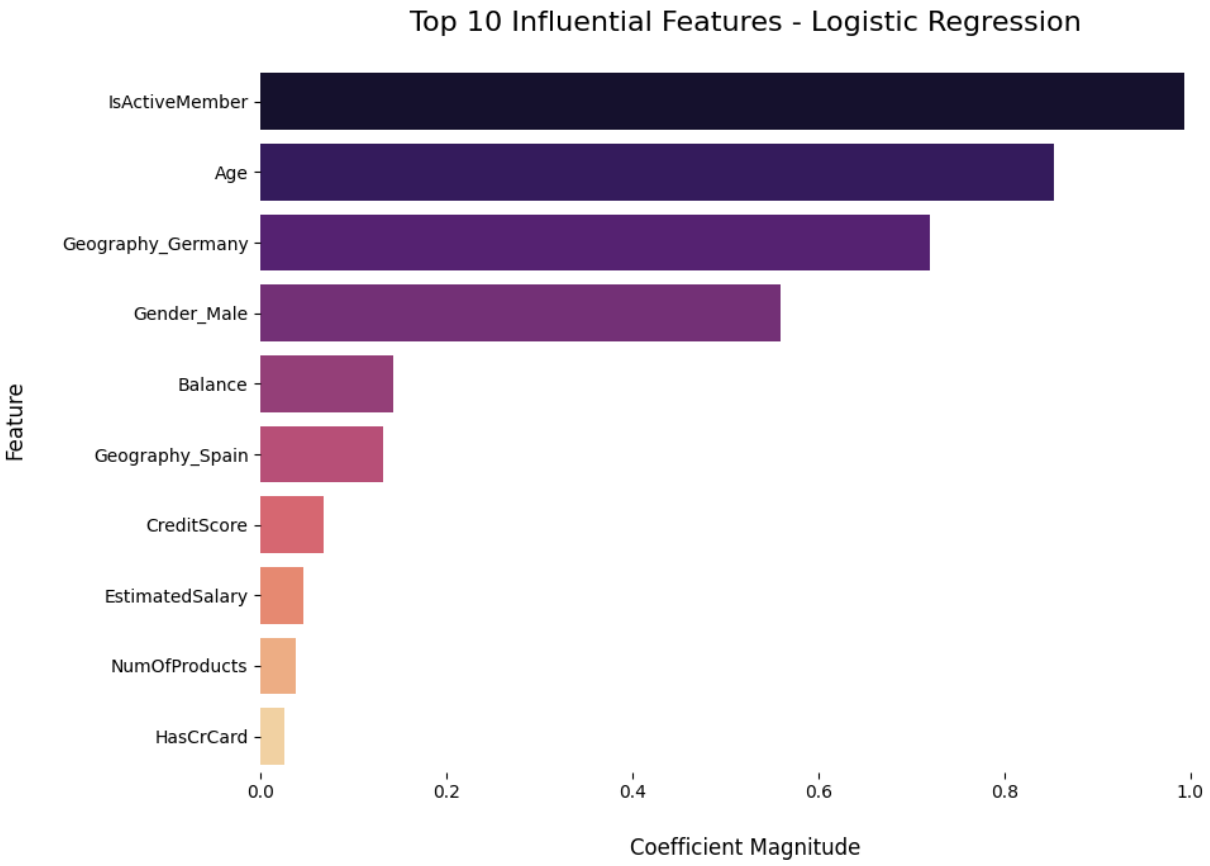

F1-Score (Macro Avg)
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.66
0.69
0.76
Logistic Regression
Decision Tree
Random Forest
Model


Logistic Regression
Decision Tree
Random Forest

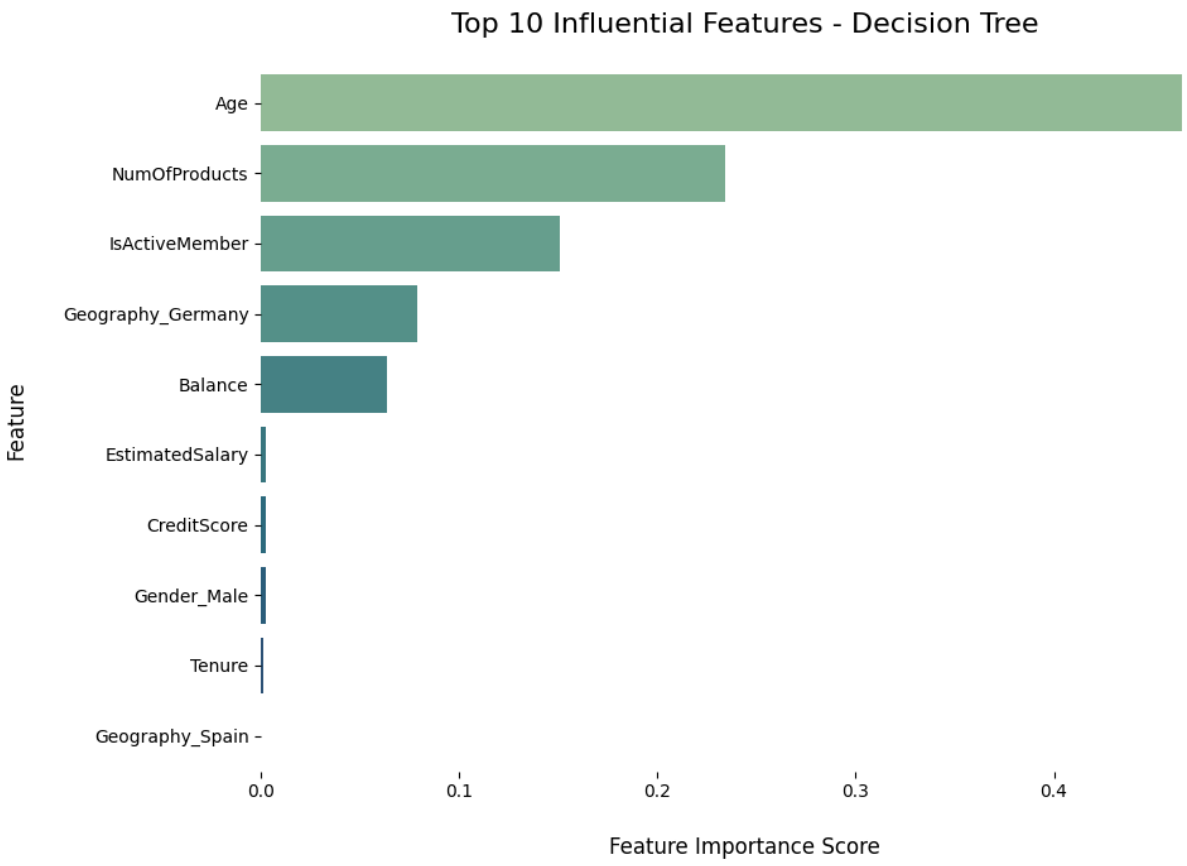
Appendix C


ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.782)
Decision Tree (AUC = 0.827)
Random Forest (AUC = 0.862)
Random Classifier

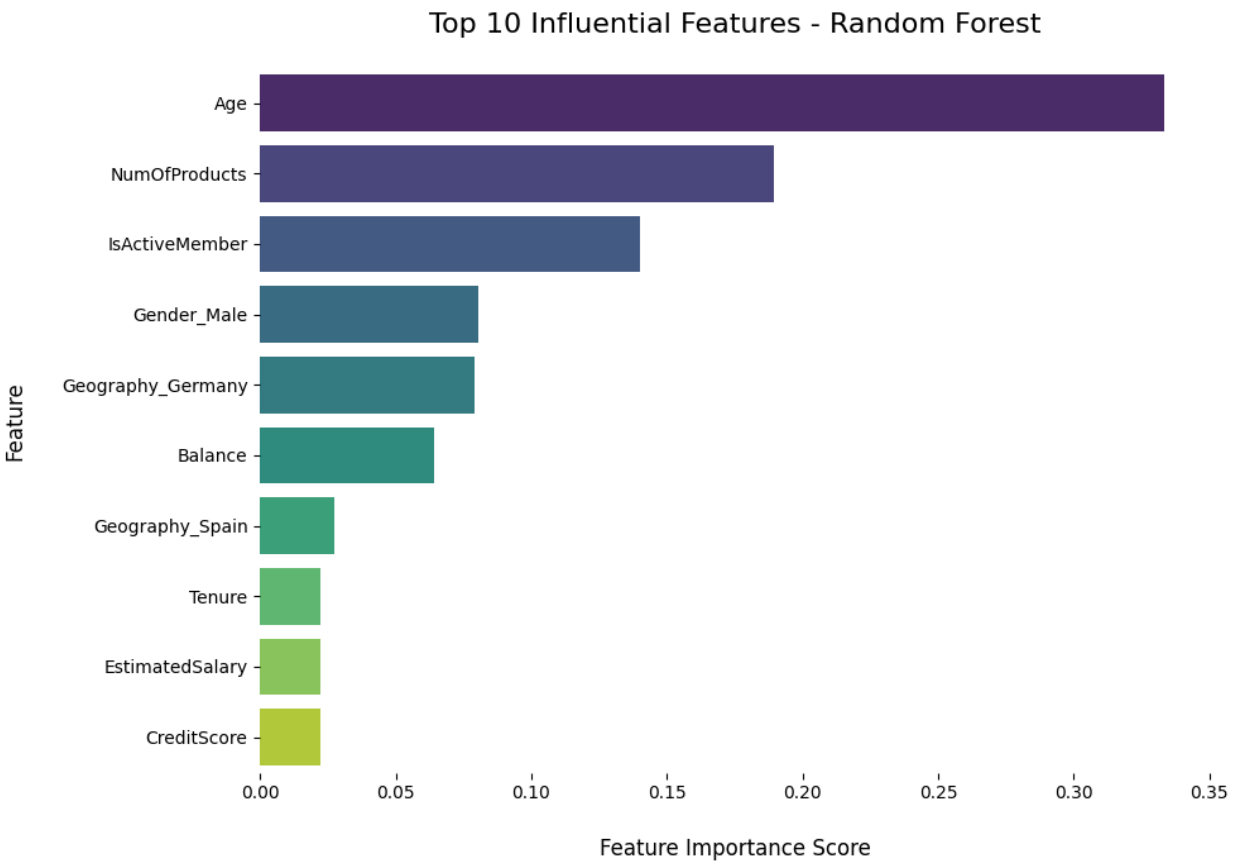
Appendix D


Top 10 Influential Features - Logistic Regression
IsActiveMember
Age
Geography_Germany
Gender_Male
Balance
Geography_Spain
CreditScore
EstimatedSalary
NumOfProducts
HasCrCard
Feature
0.0
0.2
0.4
0.6
0.8
1.0
Coefficient Magnitude

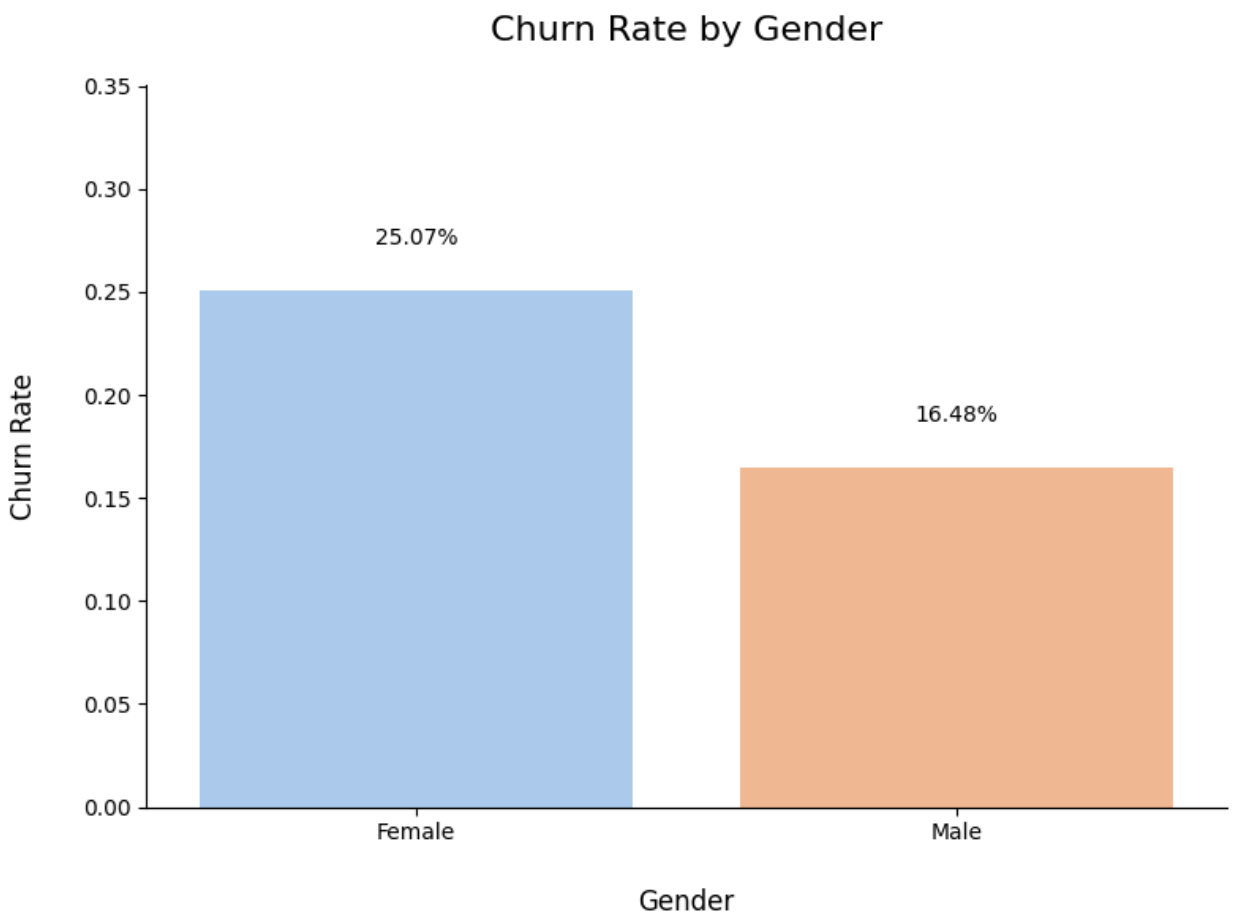
Appendix E


Top 10 Influential Features - Decision Tree
Age
NumOfProducts
IsActiveMember
Geography_Germany
Balance
EstimatedSalary
CreditScore
Gender_Male
Tenure
Geography_Spain
Feature
0.0
0.1
0.2
0.3
0.4
Feature Importance Score

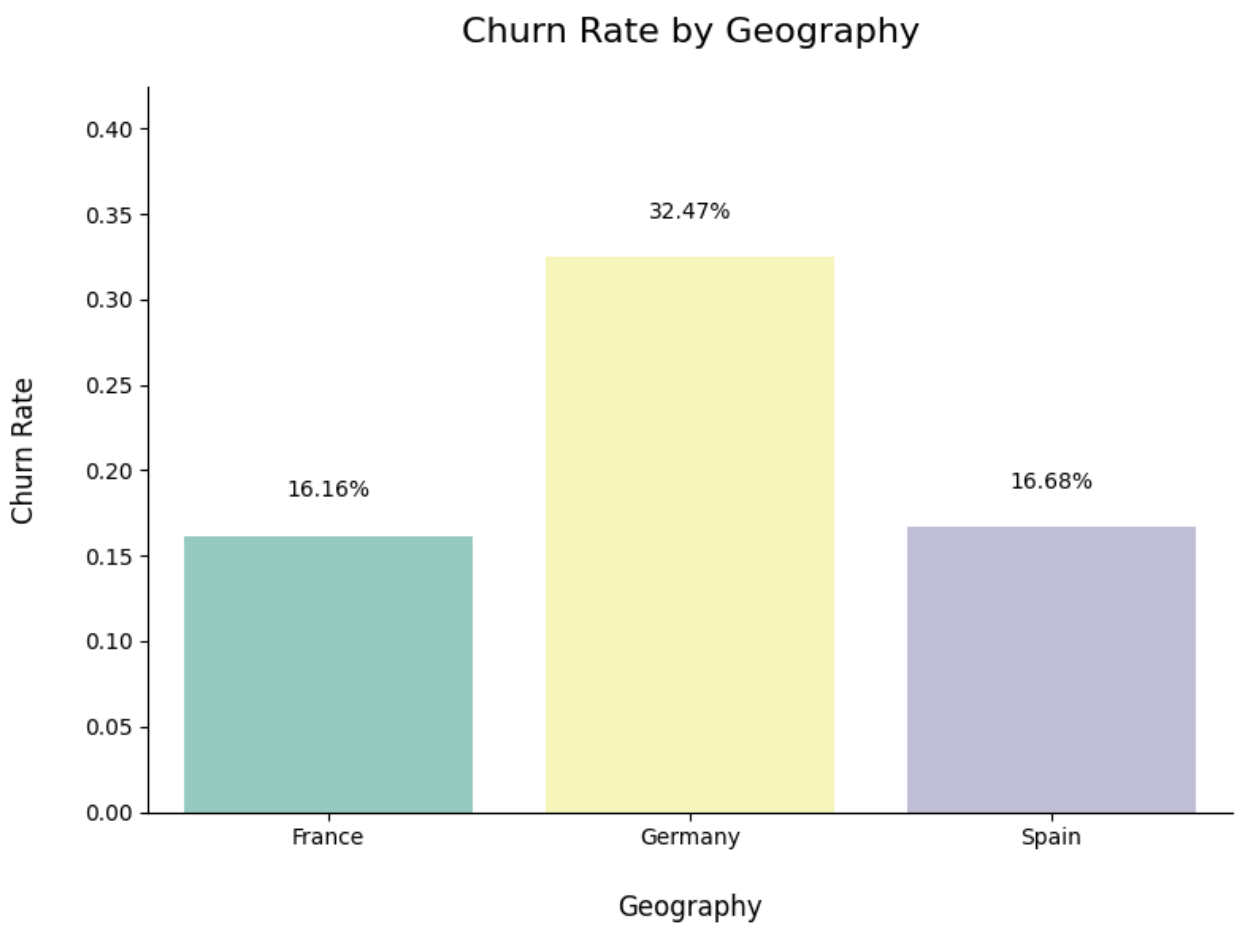
Appendix F


Top 10 Influential Features - Random Forest
Age
NumOfProducts
IsActiveMember
Gender_Male
Geography_Germany
Balance
Geography_Spain
Tenure
EstimatedSalary
CreditScore
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Feature Importance Score

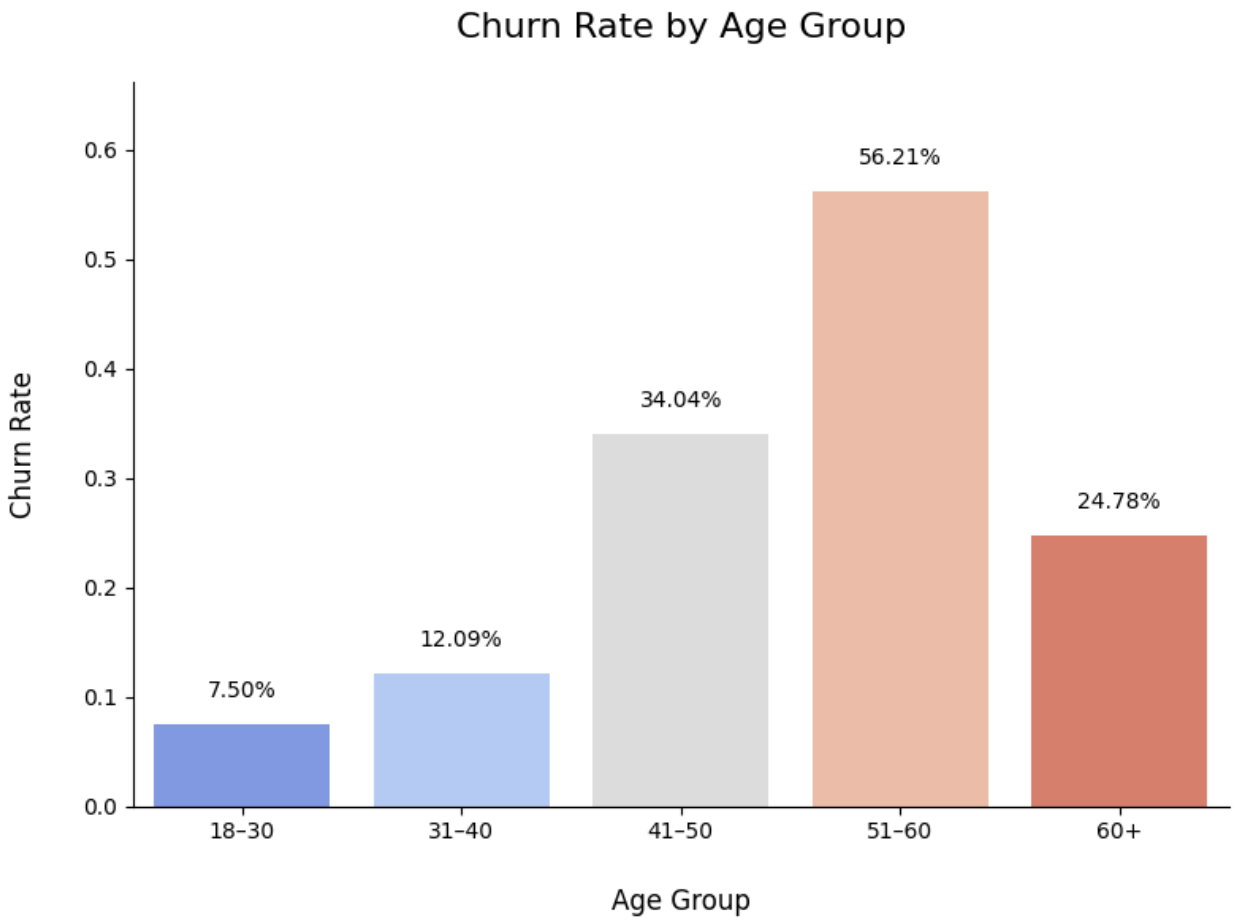
Appendix G


Churn Rate by Gender
Churn Rate
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
25.07%
16.48%
Female
Male
Gender

Appendix H


Churn Rate by Geography
Churn Rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
16.16%
32.47%
16.68%
France
Germany
Spain
Geography

Appendix I


Churn Rate by Age Group
Churn Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
7.50%
12.09%
34.04%
56.21%
24.78%
18–30
31–40
41–50
51–60
60+
Age Group