FINAL PROJECT REPORT

Predictive Analysis of Customer Churn in Banking Sector

This project aims to predict customer churn in the banking sector using machine learning models. The primary objective is to identify key factors influencing churn rates and analyze demographic patterns affecting customer retention. Using the "Churn_Modelling.csv" dataset from kaggle.com, several predictive models were implemented, including logistic regression, decision trees, random forests, and gradient boosting. Overall, the report offers insights for banking institutions to refine customer retention strategies.

Introduction

In an era where customer loyalty is pivotal to business success, reducing customer churn is a key priority. This study investigates churn predictors using banking customer data with features like geography, age, and financial behavior. The goal is to identify factors influencing churn and to develop predictive models to inform customer retention strategies.

About the Data

The dataset consists of 10,000 banking customers, including features like Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Credit Card Status, and Churn (Exited).

Preprocessing:

Categorical features like `Geography` and `Gender` were converted to factors. We also exclude the variable ‘Surname’ for confidential purposes, and the “RowNumber” variable while doing the analysis. Also, Missing values were inspected, and the dataset was split into training (80%) and testing (20%) sets.

Summary Statistics:

Descriptive statistics indicate that the average age of bank customers is 38.92, with a median of 37 years. The mean balance was around **$76,485.89**, with a high standard deviation of **$62,397.41**, indicating substantial variability in the financial status of customers. Further exploration shows that customers with higher account balances tend to remain loyal, while those with lower balances exhibit a higher propensity to churn. Age also plays a role, as younger customers are more likely to exit compared to middle-aged customers.

Visualization Insights:

Several visualizations provided insights into customer demographics and their impact on churn. **Age Distribution Histogram** and **Distribution of Churn by Age** revealed a positively skewed age distribution, with most customers being between 30 and 50 years old.

Age vs. Balance Scatter Plot illustrated the correlation between age and account balance, showing that older customers generally have higher balances.

Box Plot of Balance by Geography comparing geographic regions highlighted significant churn disparities among different locations.

Heatmap of Correlations Between Numerical Features show no variables show a strong positive correlation, indicating diverse data features suitable for predictive modeling without concerns of multicollinearity affecting performance.

Feature importance plots from the Random Forest model indicated that factors like "Age," "Balance," and "Geography" are key determinants of customer churn. Other features like EstimatedSalary and CreditScore have moderate importance, while Gender and HasCrCard are less influential.

Lastly, the **ROC curves** compare model performance, with Boosted Tree exhibiting the highest AUC, signifying superior sensitivity and specificity. Random Forest also performs well, outperforming logistic regression and CART models, which show lower AUCs.

Predictive Models employed and the evaluation metrics used:

- Logistic Regression
- CART (Decision Trees)
- Random Forests
- Gradient Boosting

Evaluation Metrics:
Accuracy
Kappa
ROC-AUC

Model Performance :

The results indicate that ensemble models like Random Forest and Boosted Tree consistently outperform the baseline logistic regression and single-tree CART models.
The **Boosted Tree** model's **AUC score** of **0.7139** makes it the strongest predictive model for identifying customers likely to churn.
Despite **logistic regression** being a simpler model, it remains valuable for identifying and understanding the influence of individual features on churn probability.
Decision trees and **random forests** provide more detailed insights into feature importance and relationships, which can inform more strategic interventions.

The overall analysis demonstrates that using ensemble learning techniques like **Random Forests** and **Boosted Trees** enhances predictive performance, particularly in **detecting customer churn**.

Factor Analysis:

Factor 1 is characterized primarily by "NumOfProducts," which has a high positive loading (0.987). Factor 2 is defined by "IsActiveMember," showing a high positive loading (0.995). Factor 3 has minor contributions from "Balance" (0.107) and "HasCrCard" (-0.129), indicating weak correlations with the overall variance.

Balance: Negative loading (-0.324) on Factor 1 and a small positive loading (0.107) on Factor 3.

NumOfProducts: High positive loading (0.987) on Factor 1, making it a primary feature.

HasCrCard: Small negative loading (-0.129) on Factor 3.

IsActiveMember: High positive loading (0.995) on Factor 2.

Overall, the analysis shows that Factors 1 and 2 are strongly influenced by "NumOfProducts" and "IsActiveMember," respectively. Factor 3 explains a minimal portion of the total variance, suggesting that additional factors may be needed to capture the underlying structure fully. The cumulative variance explained by all three factors is relatively low (26.9%).

INSIGHTS:

Key Predictors of Customer Churn in the Banking Sector

Key predictors of customer churn in the banking sector include geography, age, account balance, number of products, active membership, and credit card ownership. Geographic location significantly influences churn rates, likely due to varying banking needs and economic conditions. Age plays a crucial role, with younger customers more prone to churn, possibly driven by preferences for technology-driven services. Lower account balances indicate higher churn risk, suggesting financial vulnerability or disengagement. Conversely, customers with more products are less likely to churn, reflecting deeper integration into the bank's ecosystem. Active membership and credit card ownership are also associated with lower churn rates, highlighting the importance of customer engagement and perceived value in banking relationships.

Impact of Geography, Age, and Gender on Churn Rates

Geography: Regional differences in churn rates highlight the need for localized marketing and retention strategies. Customers in some regions may be more price-sensitive or have specific banking needs that differ from other areas.

Age: Younger customers are more prone to switching banks, while middle-aged customers tend to stay longer, suggesting that different strategies should be employed for distinct age groups.

Gender: No significant correlation between gender and churn rates was identified. This suggests that both male and female customers should receive equally targeted retention strategies.

Relationship Between Balance, Product Usage, and Churn

Account Balance:
Lower account balances correlate with higher churn rates. Customers with higher balances often have greater engagement with their bank's financial products and perceive more value in staying loyal.

Number of Products:
Higher product usage generally leads to a lower churn likelihood. Customers who use multiple banking products (e.g., savings accounts, credit cards, loans) find it more challenging to switch banks due to the disruption involved.

Recommendations:

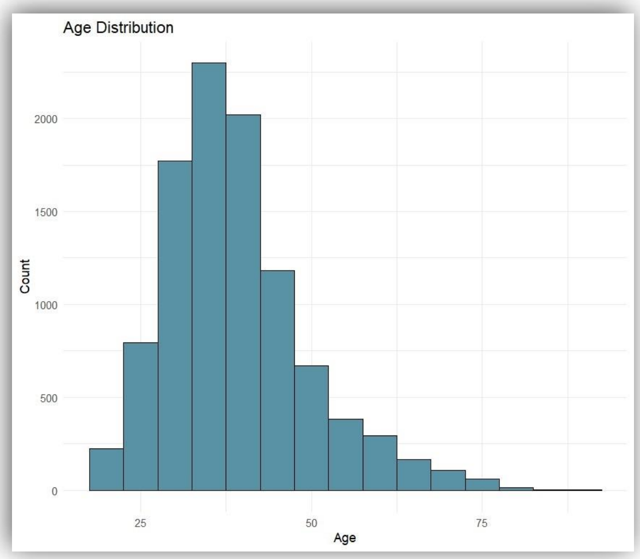
1. Focus on **Regions** with **High Churn** Rates:
    - Regional Analysis: Conduct a comprehensive analysis to identify geographic areas where customer churn is the highest.
    - Localized Services: Offer banking services that cater to regional economic conditions and consumer behavior, such as community-specific credit products or financial advisory services.
2. Target **Younger Customers** with **Lower Account Balances**:
    - Financial Literacy Programs: Implement educational programs focused on financial literacy to help younger customers understand the benefits of long-term financial planning.
    - Innovative Banking Products: Create modern, technology-driven banking products like mobile banking apps or specialized savings plans that resonate with younger audiences.
3. **Personalized Engagement** and Value-Added Services:
    - Customer Segmentation: Use data analytics to segment customers based on behavior and needs, enabling highly personalized communication and services.
    - Advisory Services: Offer tailored financial advice to help customers achieve their individual goals, reinforcing their relationship with the bank.
4. **Premium Benefits**:
    - Provide premium benefits like travel discounts, financial management tools, or higher interest rates on savings accounts to high-value customers.
5. **Active Members** and Those **Holding Multiple Products**:
    - Cross-Selling Opportunities: Leverage cross-selling strategies to introduce active members to new products that align with their existing financial habits and needs.

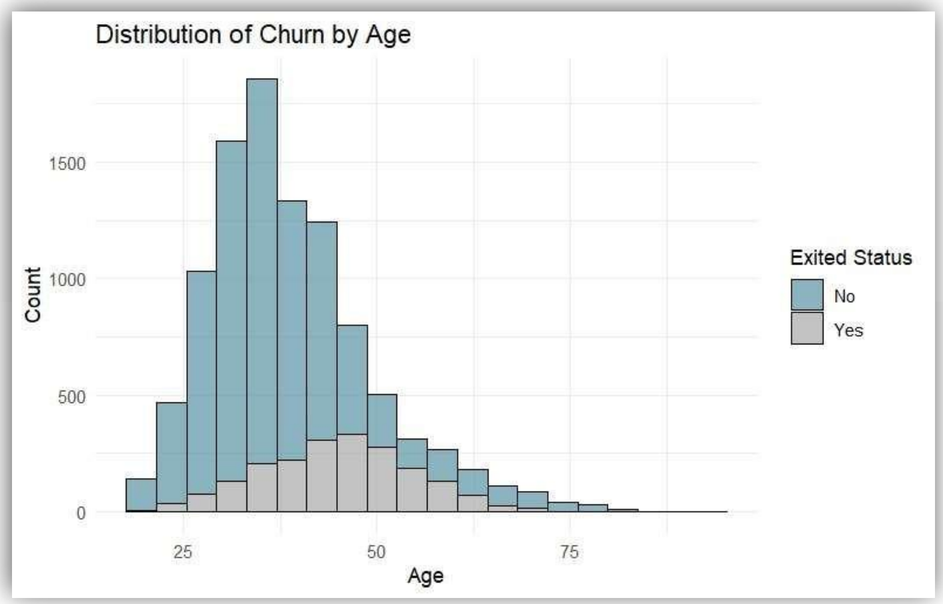
Conclusion:

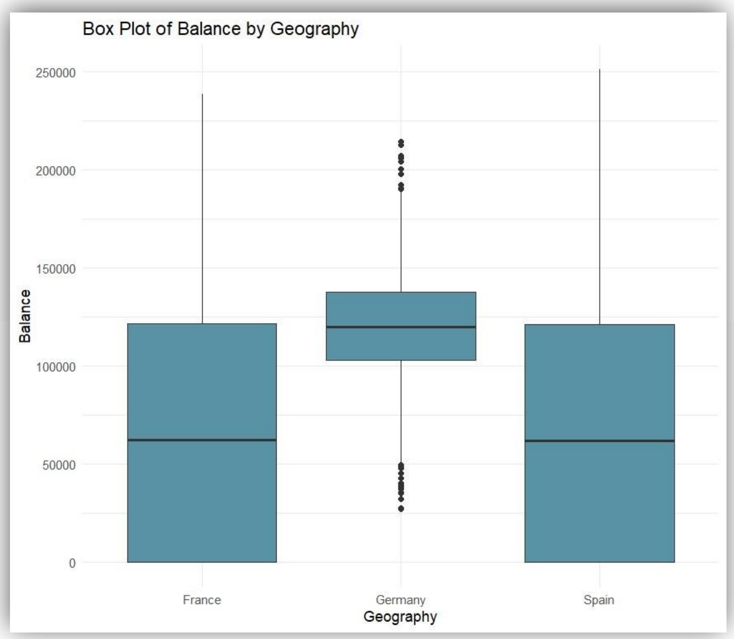
The analysis underscores the multifaceted nature of customer churn in the banking sector, revealing a combination of demographic, behavioral, and engagement factors driving churn rates. Geographic location, age, account balance, product usage, active membership, and credit card ownership emerged as key predictors of churn, highlighting the importance of personalized retention strategies tailored to diverse customer segments. While younger customers exhibit higher churn propensity, regional disparities, and varying financial behaviors further complicate churn dynamics. Notably, ensemble learning techniques like Random Forests and Boosted Trees offer superior predictive performance, enabling deeper insights into feature importance and relationships.

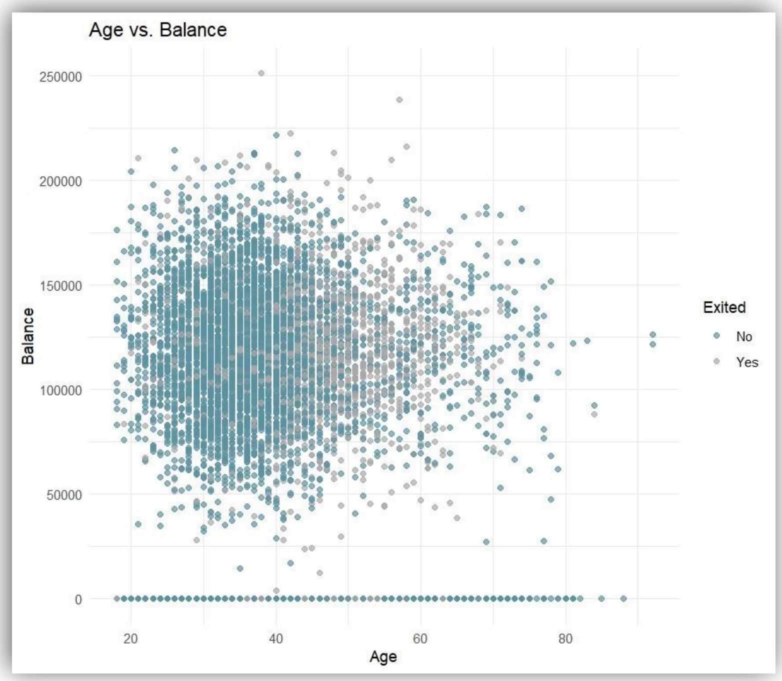
Appendices

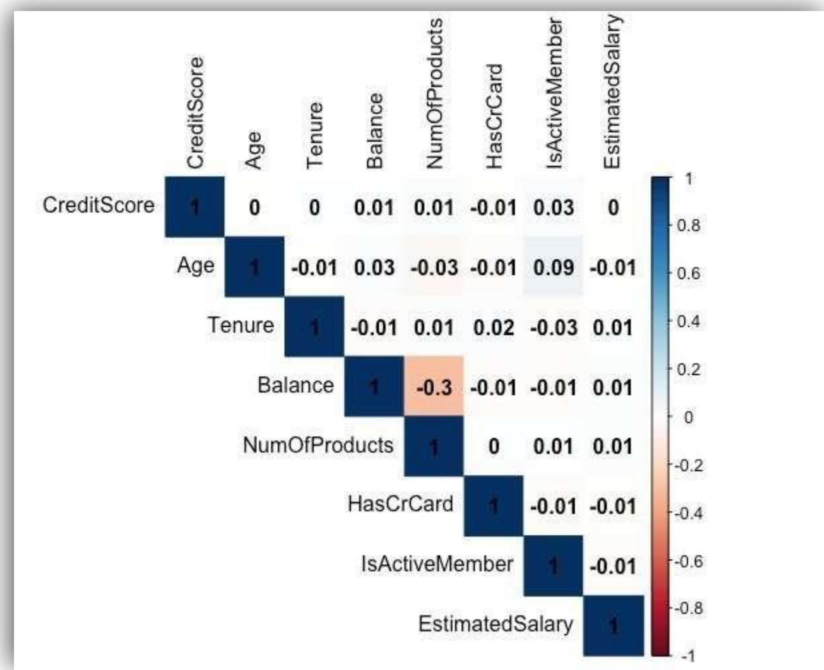
Visualisation plots:

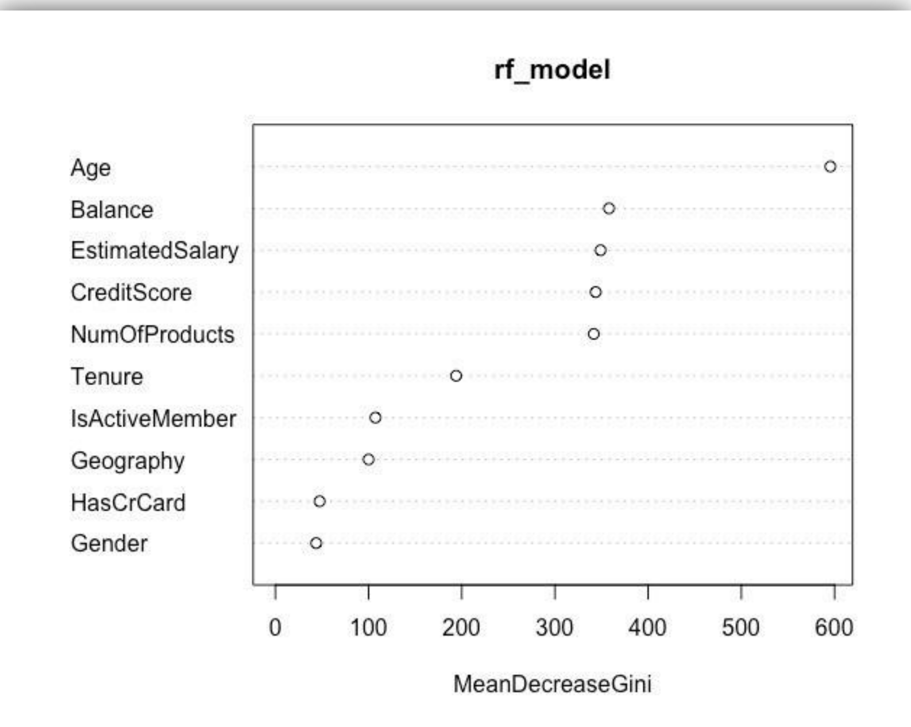


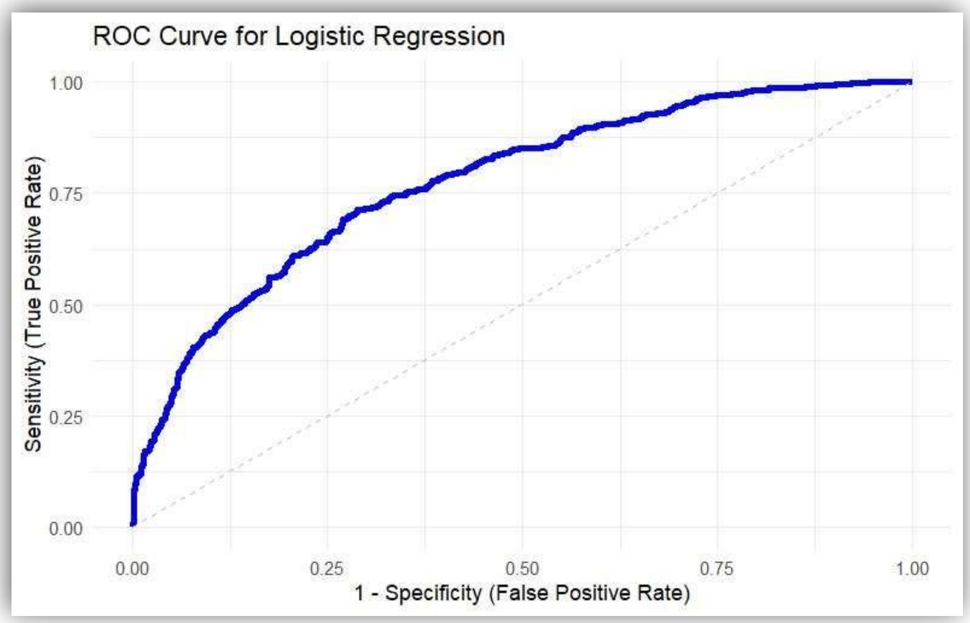


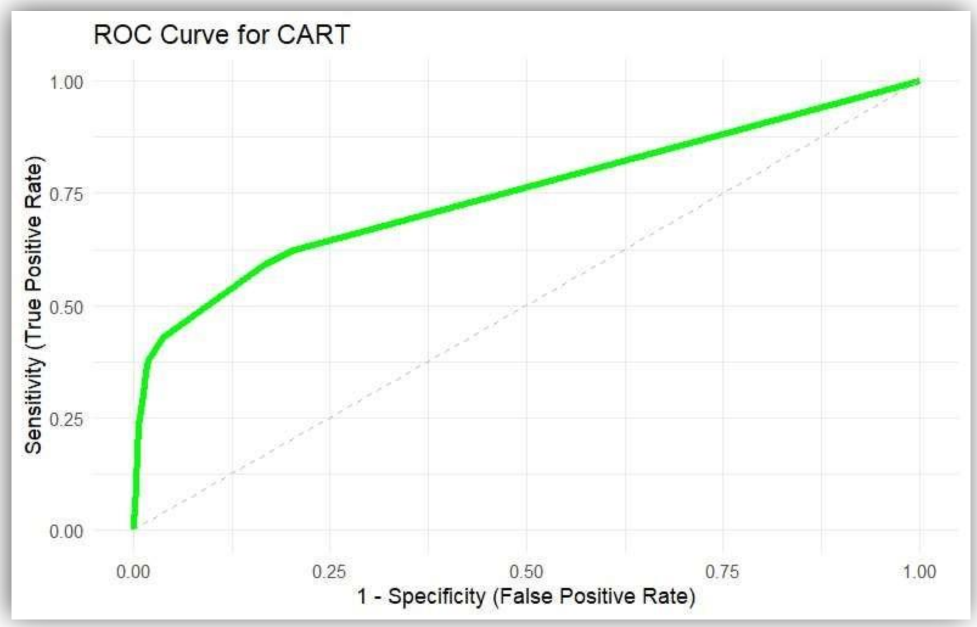

Box Plot of Balance by Geography
250000
200000
150000
100000
50000
0
Balance
France
Germany
Spain
Geography

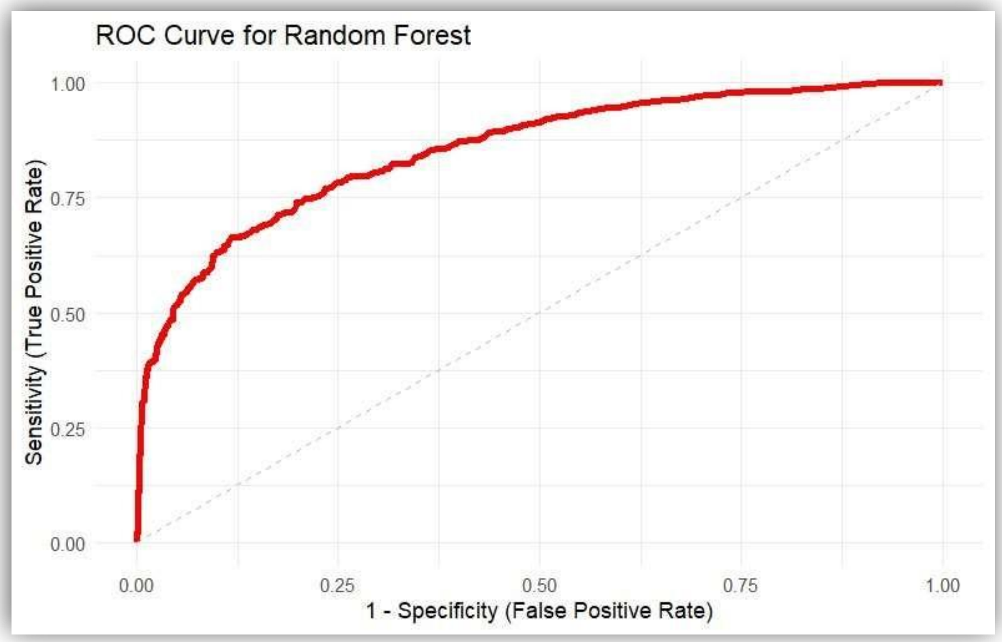

Age vs. Balance
250000
200000
150000
100000
50000
0
Balance
20
40
60
80
Age
Exited
● No
● Yes

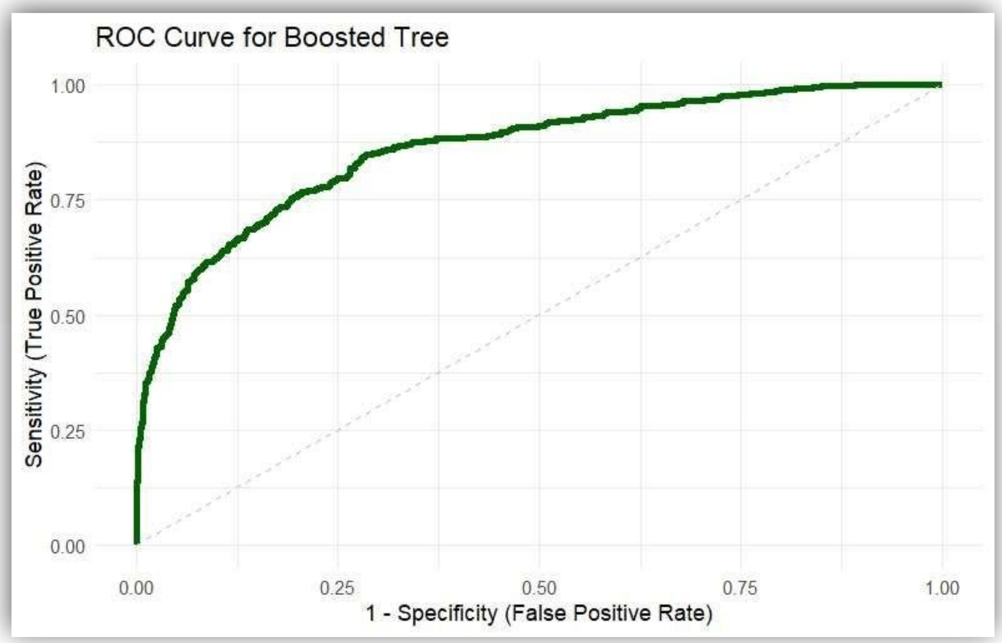

CreditScore
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
CreditScore 1 0 0 0.01 0.01 -0.01 0.03 0
Age 1 -0.01 0.03 -0.03 -0.01 0.09 -0.01
Tenure 1 -0.01 0.01 0.02 -0.03 0.01
Balance 1 -0.3 -0.01 -0.01 0.01
NumOfProducts 1 0 0.01 0.01
HasCrCard 1 -0.01 -0.01
IsActiveMember 1 -0.01
EstimatedSalary 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

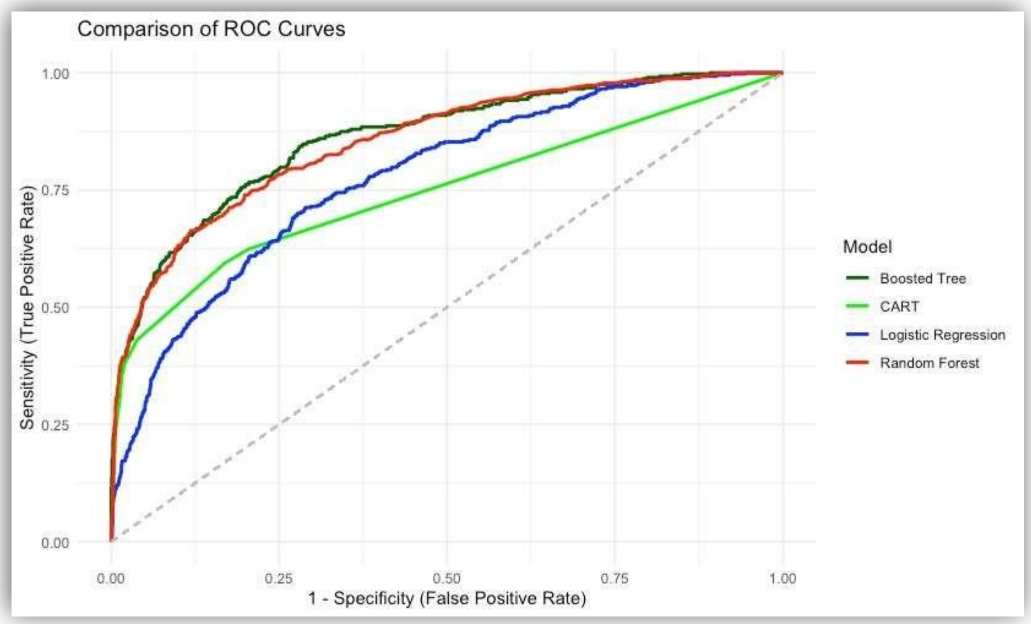

rf_model
Age
Balance
EstimatedSalary
CreditScore
NumOfProducts
Tenure
IsActiveMember
Geography
HasCrCard
Gender
0
100
200
300
400
500
600
MeanDecreaseGini


ROC Curve for Logistic Regression
1.00
0.75
0.50
0.25
0.00
Sensitivity (True Positive Rate)
0.00
0.25
0.50
0.75
1.00
1 - Specificity (False Positive Rate)


ROC Curve for CART
1.00
0.75
0.50
0.25
0.00
Sensitivity (True Positive Rate)
0.00
0.25
0.50
0.75
1.00
1 - Specificity (False Positive Rate)


ROC Curve for Random Forest
Sensitivity (True Positive Rate)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity (False Positive Rate)


ROC Curve for Boosted Tree
Sensitivity (True Positive Rate)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity (False Positive Rate)


Comparison of ROC Curves
Sensitivity (True Positive Rate)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity (False Positive Rate)
Model
Boosted Tree
CART
Logistic Regression
Random Forest