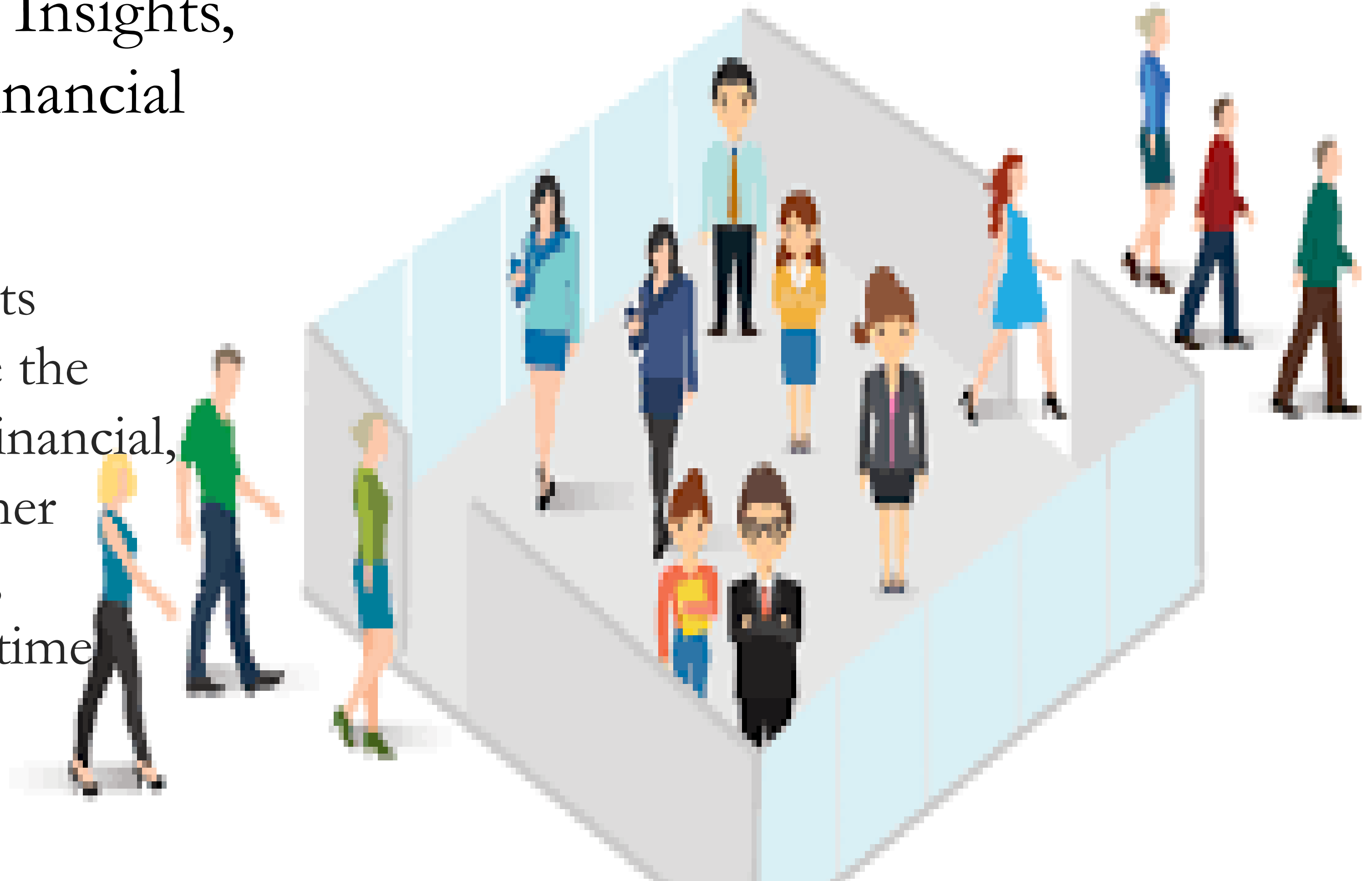
Bank Churn Prediction:

Unlocking Customer Retention Insights, Analyzing Demographic and Financial Factors

This machine learning model, predicts whether a customer will churn (leave the bank) based on their demographic, financial, and account activity data. As Customer churn is a significant issue for banks, impacting revenue and customer lifetime value.



Data Overview

Dataset Size

10000 rows and 13 columns.

Key Features

Credit_Score, Geography, Gender, Age, Tenure, Balance, Is_Active_Member, Num_Of_Products, Has_Cr_Card, Estimated_Salary

Target Variable

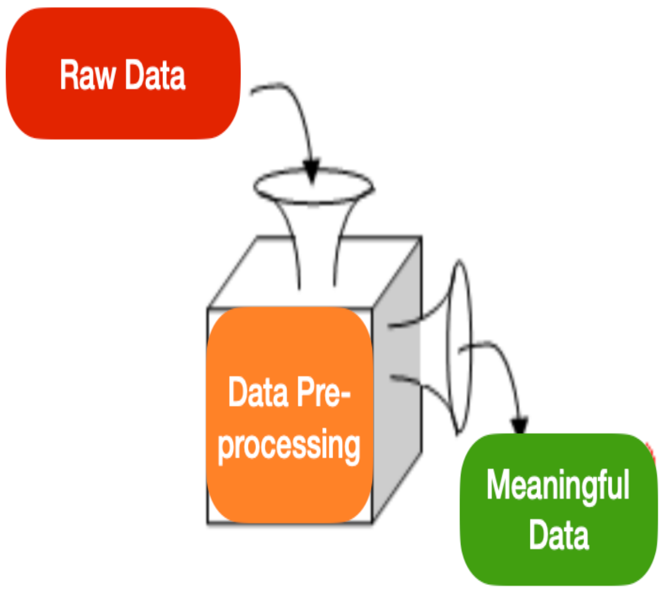
Exited(Churned = 1 / Stayed = 0), will the customer leave or Stay.

Data Preprocessing

- No missing values were found in the data
- Irrelevant columns ('Surname', 'Customer Id') were removed.

- Outlier Identification – Major outliers were removed thru the IQR method from the numerical features ('CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'EstimatedSalary’). As outlier can skew statistical analysis.

- Feature Scaling - Numerical features were standardized using ‘Standard scaler’ to have zero mean and unit variance. And Categorical features were one hot encoded using ‘ One Hot Encoder’. This converts categorical variable into numerical representations that algorithms can understand.



Exploratory Data Analysis

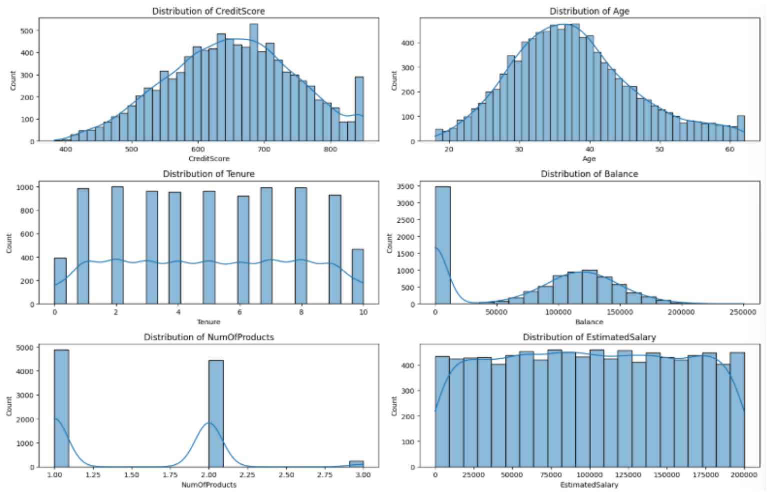
Distribution Plots

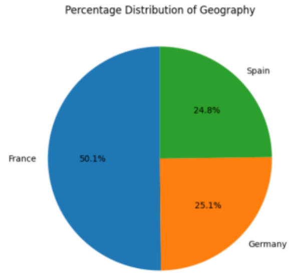
The plots help us understand the data distribution ,

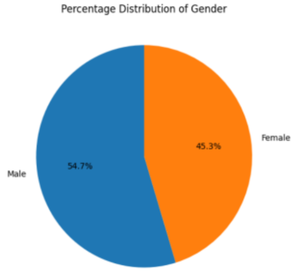
We see:

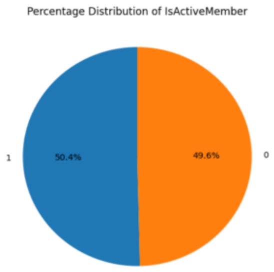
Customers population distribution
France(50.1%)> Germany(25.1%) > Spain(24.8%)
Gender distribution Male(54.7%)> Female(45.3%)
Customer have Credit cards
Yes(70.5%) > No(29.5%)
Active members (almost equally distributed)
Yes(50.4%)> No(49.6%)
The class distribution for “EXITED” has a bias towards ‘NON_Exited”

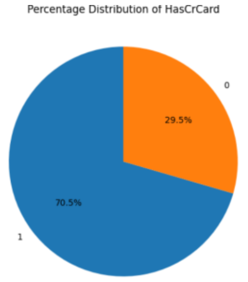
Further people in Germany have churned the most, Females have done so and who were not active and Had credit card churned more.



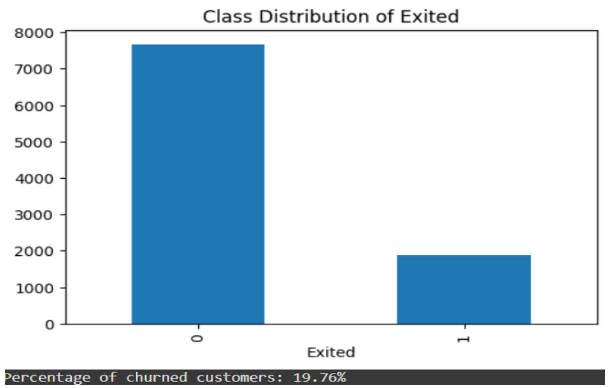


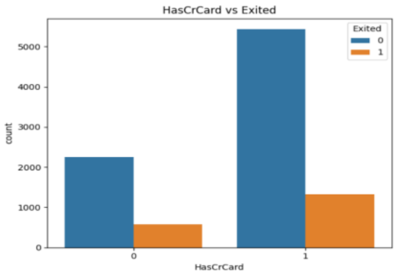


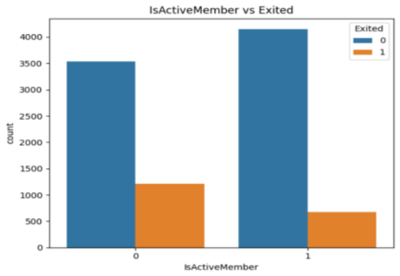


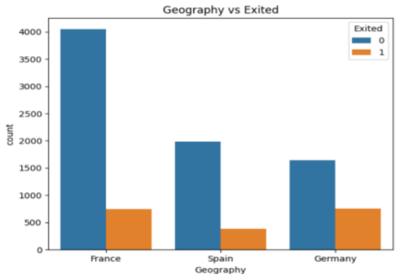


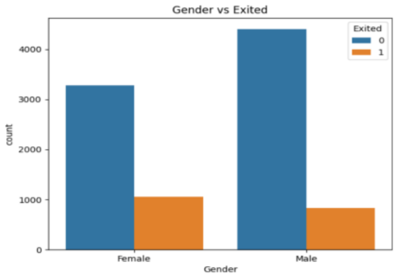
The class distribution for “EXITED” has a bias towards ‘NON_Exited”
Further people in Germany have churned the most, Females have done so and who were not active and Had credit card churned more.





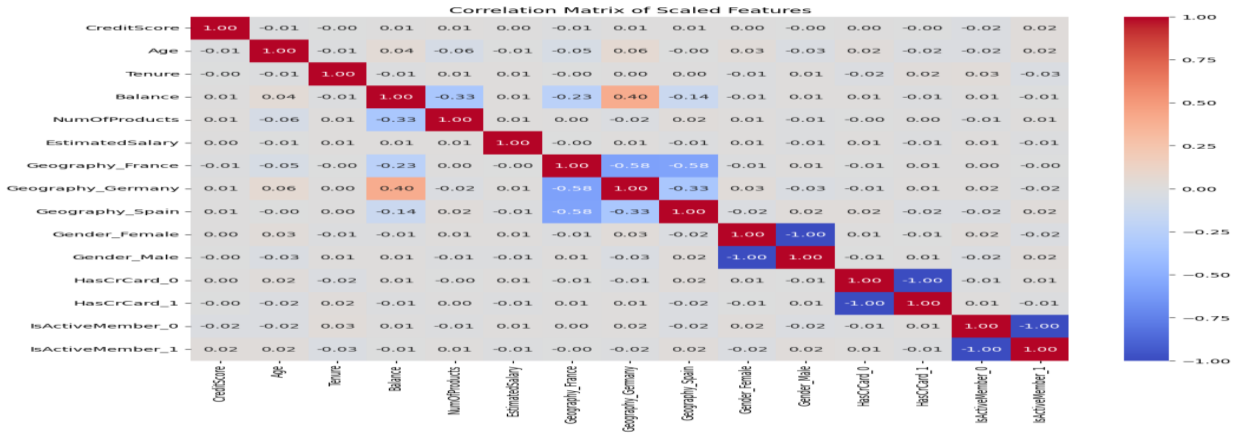






Corelation Matrix

The heatmap helps identifying relationship between the features and help In identifying the important features.



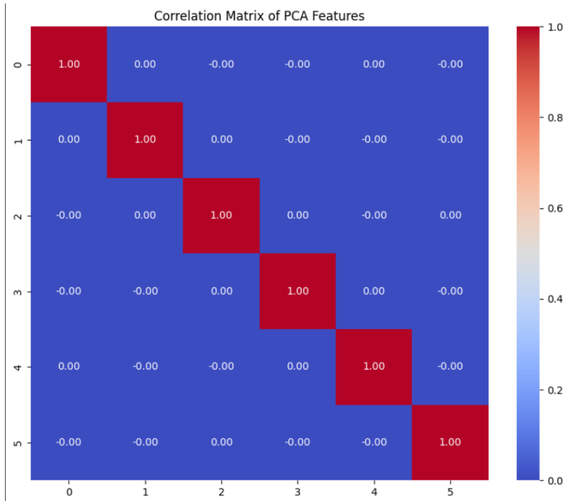
Removing the biasness

Performed SMOTE to remove the biasness from the target variable " Exited" and created 7677 data points for both the columns "Exited" and "Stayed"

```
Exited
1    7677
0    7677
Name: count, dtype: int64
```

PCA (Principal Component Analysis

PCA was performed to see what features have variance between 80%-85% and `['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'EstimatedSalary']` were the selected features



Checking for Multicollinearity

No major multicollinearity existed in the selected features.

Splitting data training models

Splitting the data into train and test sets and trained models such as Random forest, Support Vector Machines, Decision tree, ADA boost, and logistic Regression model for making predictions.

Classification reports

Made performance matrices for all the models to get precision, recall , f1_score and accuracy of the models for getting insights.

Performed cross validation

Cross validation was done on all the models to train them better by partitioning the data into multiple subsets, training the model on some subsets, and testing it on the remaining data.

Confusion matrices

Build confusion matrices for all the models to help decide the best performing model, we see that Random forest is the best model as it has the lowest “false negative” which is a concerning point in our model prediction. As it identifies the customers who will churn as non churning.

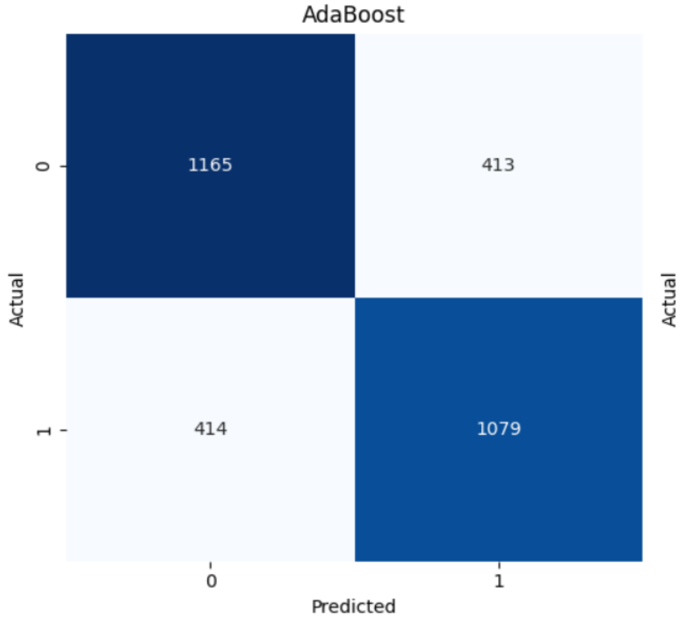
```
--- Logistic Regression ---
Accuracy: 0.6935851514164767
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.69      0.70      1578
           1       0.68      0.69      0.69      1493

    accuracy                           0.69      3071
   macro avg       0.69      0.69      0.69      3071
weighted avg       0.69      0.69      0.69      3071
```

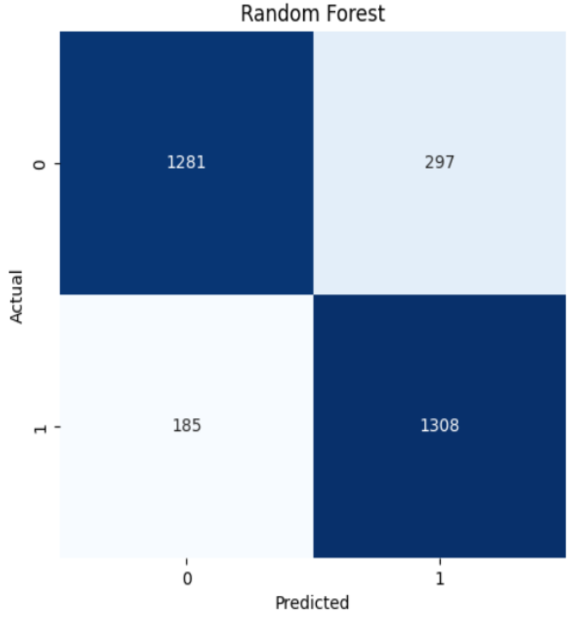
```
--- Random Forest ---
Accuracy: 0.8430478671442527
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.81      0.84      1578
           1       0.81      0.88      0.84      1493

    accuracy                           0.84      3071
   macro avg       0.84      0.84      0.84      3071
weighted avg       0.85      0.84      0.84      3071
```

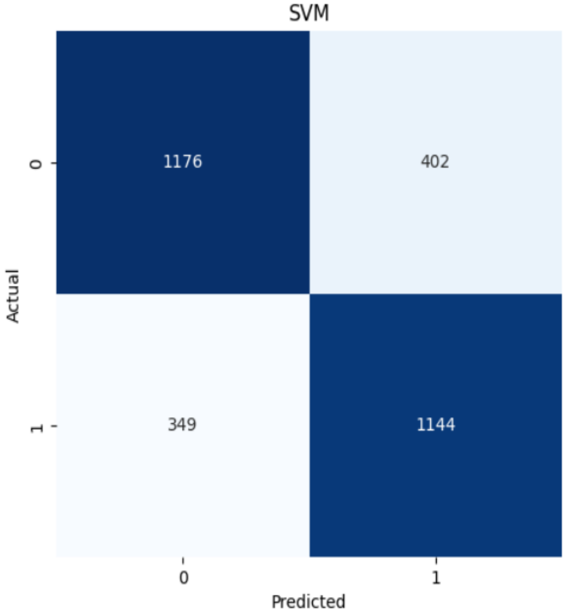
```
--- SVM ---
Accuracy: 0.755454249430153
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.75      0.76      1578
           1       0.74      0.77      0.75      1493

    accuracy                           0.76      3071
   macro avg       0.76      0.76      0.76      3071
weighted avg       0.76      0.76      0.76      3071
```

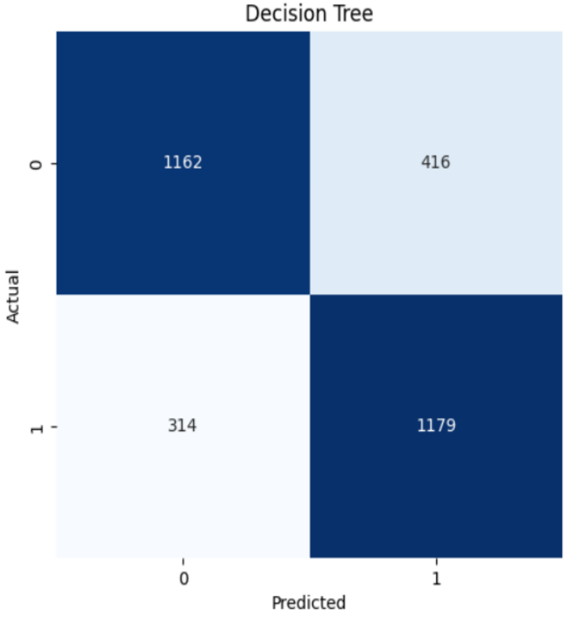
```
--- AdaBoost ---
Accuracy: 0.7307066102246825
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.74      0.74      1578
           1       0.72      0.72      0.72      1493

    accuracy                           0.73      3071
   macro avg       0.73      0.73      0.73      3071
weighted avg       0.73      0.73      0.73      3071
```

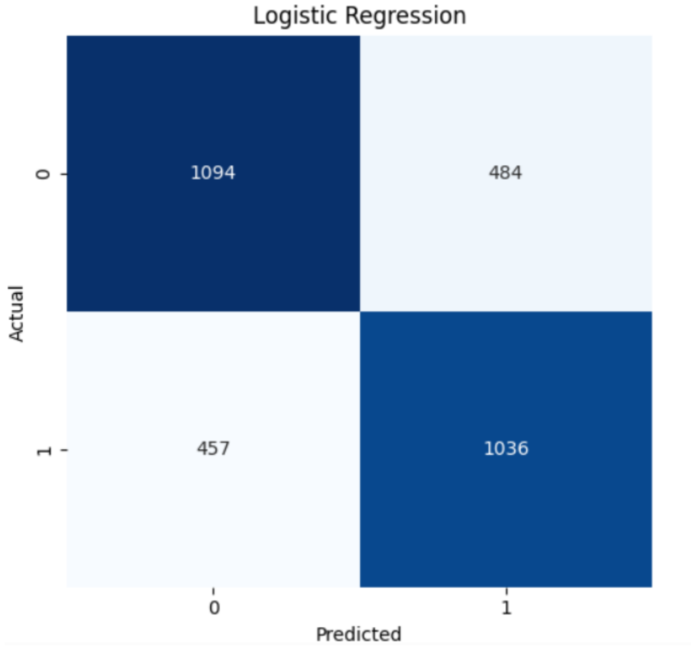
```
--- Decision Tree ---
Accuracy: 0.7622924128948225
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.74      0.76      1578
           1       0.74      0.79      0.76      1493

    accuracy                           0.76      3071
   macro avg       0.76      0.76      0.76      3071
weighted avg       0.76      0.76      0.76      3071
```





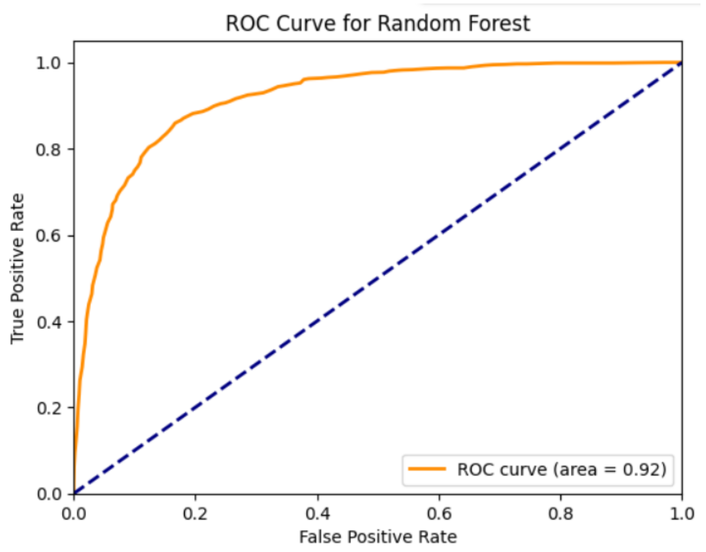






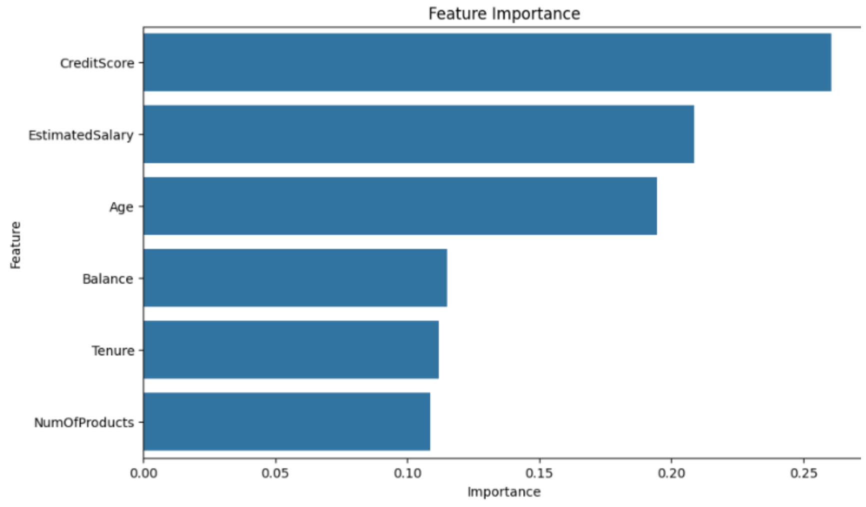
ROC curve

The roc curve has an area of 0.92 which is a quite good coverage , which indicates a high probability that the model will correctly distinguish between positive and negative classes.



Feature importance

we see, the top 6 features having influence on the predicting model are credit score > Estimated Salary > Age > Balance > Tenure > Num of Products.



5 Fold Cross-Validation and Hyperparameter Tunning

We did a 5 fold cross validation, to evaluate the model by splitting the dataset into five groups, using four for training and one for testing in each of the five iterations, and then averaging the results to get a more robust performance estimate along with parameter tunning to find the optimal set of parameters for model.

Best model

After calculating the accuracy , precision , Recall and F1-score also studying the confusion matrices of all the models we see that Random Forest is our best performing model with 84.3% percent of accuracy.

```
Random Forest:
Precision = 0.8150
Recall = 0.8761
F1-Score = 0.8444
Accuracy = 0.8430
```

```
Random Forest: Accuracy = 0.8430478671442527
SVM: Accuracy = 0.755454249430153
Decision Tree: Accuracy = 0.7622924128948225
AdaBoost: Accuracy = 0.7307066102246825
Logistic Regression: Accuracy = 0.6935851514164767

Best performing model: Random Forest with accuracy: 0.8430478671442527
```

Thank you!