Assignment Cover Sheet

Student Name and Number as per student card: Ravi Bhagiram Maurya (20065040)

Programme: Programming for Data Analysis Project

Lecturer Name: Kanza Ali Manzar

Module/Subject Title: B9DA108

Assignment Title: CA_TWO_(40%)

By submitting this assignment, I am confirming that:

- **This assignment is all my own work;**
- **Any sources used have been referenced;**
- **I have followed the Generative AI instructions/ scale set out in the Assignment Brief;**
- **I have read the College rules regarding academic integrity in the QAH Part B Section 3, and the Generative AI Guidelines, and understand that penalties will be applied accordingly if work is found not to be my/our own.**
- **I understand that all work submitted may be code-matched report to show any similarities with other work.**

Table of Content

Introduction

This project focuses on analysing customer churn data to uncover patterns and trends that contribute to customer attrition. By identifying the key factors driving churn, this analysis aims to equip the bank with actionable insights to improve customer retention strategies, ultimately enhancing business sustainability. The dataset utilized, titled **Customer Churn Records**, provides comprehensive information on customer demographics, account details, service usage, and more, enabling a robust investigation into the drivers of churn.

Objective:

1. **Understand Factors Causing Customer Churn**:
    - Identifying the key features and patterns that influence customer decisions to leave the bank.
2. **Explore and Summarize the Dataset**:
    - Performing descriptive and exploratory data analysis to uncover valuable insights and highlight the key features of the dataset.
3. **Generate Actionable Insights**:
    - Using data visualizations and summary statistics to highlight trends and provide strategic recommendations for customer retention.
4. **Examine Relationships Between Features**:
    - Analysing the relationships between independent features (input variables) and the target variable (churn), quantifying their impact on predicting customer churn.
5. **Develop a Predictive Machine Learning Model**:
    - Build and validate a machine learning algorithm capable of accurately predicting customer churn rates.
6. **Benchmark Success Metrics**:
    - Establish clear metrics to evaluate the success of the predictive model and the impact of churn-reduction initiatives.

Data Description:

Dataset Source:

- The dataset is titled Customer-Churn-Records.csv.
- The source of the dataset is Kaggle.
- https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/data

Type of Data:

- Tabular data consist of both numerical and categorical variables.
- The dataset has 18 features with a mix of data types such as numbers, categories, and binary values.

  Categorical Columns – Geography, Gender, Card Type

  Numerical columns – Age, Tenure, Balance, EstimatedSalary, Satisfaction Score, Point Earned, Credit Score

  Binary Columns (0&1) – HasCrCard, IsActiveMember, Churn, Complain

  Dropped Columns**:** RowNumber, CustomerId, Surname (non-predictive)

Features Include:

- **Customer demographics**

  Geography - Customer location which contain France, Germany, Spain.

  Gender - Gives information about customer's gender whether person is male or female & whether gender plays a role in a customer leaving the bank.

  Age - provide information about customer's age

  Surname – gives information about customer family name which is irrelevant

  EstimatedSalary - provide the information about customer current earning potential, people with lower salaries are more likely to leave the bank compared to those with higher salaries.

- **Account information**

  Tenure - number of years that the client has been a customer of the bank.

Balance- Indicate the amount of money present in customer accounts, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.

Card Type - type of card holds by the customer such as debit card, credit card.

Points Earned - the points earned by the customer for using credit card.

IsActiveMember - Customer is active or not.
HasCrCard - denotes whether or not a customer has a credit card.

CustomerId - contains random values which make each customer entry unique

- **Service-related data**

  NumOfProducts - number of products that a client has bought

  Complain - Information about customer complain to banking services

  Satisfaction Score - Score provided by the customer for their complaint resolution.

- **Churn status (target variable)**

  Exited - Whether the customer decided to leave the bank or stay.

- **Record Number Indicator**

  RowNumber - record (row) number and has no effect on the output.

Method of Data Analysis:

Libraries Used:

- pandas, NumPy for data manipulation
- matplotlib, seaborn, plotly, bokeh for visualization
- sklearn.preprocessing for encoding and standardization
- warnings to supress warning messages
- sklearn, imblearn (modelling & preprocessing)

Data Preprocessing Steps:

1. **Loading & Inspecting Data**:
   - Libraries like pandas, NumPy, matplotlib, and seaborn were used for data handling and visualization.
   - The dataset shape was checked, revealing 10,000 rows and 18 columns.

2. **Renaming the Target Column**:
    - The target column, 'Exited,' was renamed to 'Churn' for clarity and values converted from 0 and 1 to customer left and customer stayed with the Bank.

      Churn = 1 → The customer left the Bank.

      Churn = 0 → The customer stayed with the Bank.
3. **Null Values Check**:
    - Verified that the dataset contains no null values, so no null value treatment was required.
4. **Duplicate Values Check**:
    - Confirmed no duplicate entries in the dataset.
5. **Dropping Redundant Columns**:
    - Columns such as RowNumber, CustomerId, and Surname were dropped as they do not contribute meaningful information.
6. **Feature Categorization**:
    - Categorical columns identified: Geography, Gender, Card Type.
    - Numerical columns identified: CreditScore, Age, Tenure, Balance, and others.
7. **Outlier Detection and Treatment**:
    - Box plots and IQR method were used to identify and treat outliers for numerical columns like CreditScore, Age, Balance, and NumOfProducts.
8. **Model treatment**
    - Categorical features encoded using LabelEncoder and mapping.
    - Applied SMOTE to address class imbalance before training.

Data Analysis Methods

1. **Univariate Analysis**:
    - Target variable (Churn) was visualized using bar plots to understand the distribution of churned vs. retained customers.
    - Histograms were created for numerical columns to observe their distributions and bar charts created to observe the values count of categorical columns
2. **Bivariate Analysis**:
    - Relationships between categorical features and the target variable (Churn) were explored using grouped counts and bar plots (e.g., Geography, Gender, and Card Type).
    - Scatter plots were used to analyse relationships between numerical features such as CreditScore and Age.
    - Density plots were applied for features like Satisfaction Score to examine trends with churn.
3. **Churn Trends Analysis**:
    - Age-based segmentation was created to analyse churn rates across different age ranges.
    - Churn percentages were calculated for balance ranges and visualized with stacked bar plots.
4. **Summary Statistics**:
    - Descriptive statistics were generated for numerical columns to understand central tendencies and variability (mean, std, min, max, etc.).
5. **Outlier Insights**:
    - Outliers in columns like CreditScore and Age were detected and adjusted.

Feature Selection Step

The focus of the feature selection process is on identifying the most important features that contribute to predicting customer churn. Here are the steps followed:

1. **Removing Redundant Features**:
    - Features such as RowNumber, CustomerId, and Surname were removed as they do not carry meaningful information for predicting churn.
2. **Categorizing Features**:
    - Features were categorized into **categorical** (e.g., Geography, Gender, Card Type) and **numerical** (e.g., CreditScore, Age, Tenure, Balance) types for appropriate processing.
3. **Analysing Unique Values**:
    - The number of unique values in each column was examined to evaluate their variability and utility in the prediction task.
4. **Feature Relevance Assessment**:
    - Columns like HasCrCard, IsActiveMember, and Complain were assessed for their contribution to churn prediction based on exploratory data analysis (EDA).
5. **Correlation Analysis**:
    - Correlation between numerical features and the target variable (Churn) was explored to identify strongly correlated features.
6. **Distribution and Impact Analysis**:
    - Visualizations like bar plots, density plots, and scatterplots were created to understand the relationship between features (e.g., Satisfaction Score, Age) and churn trends.

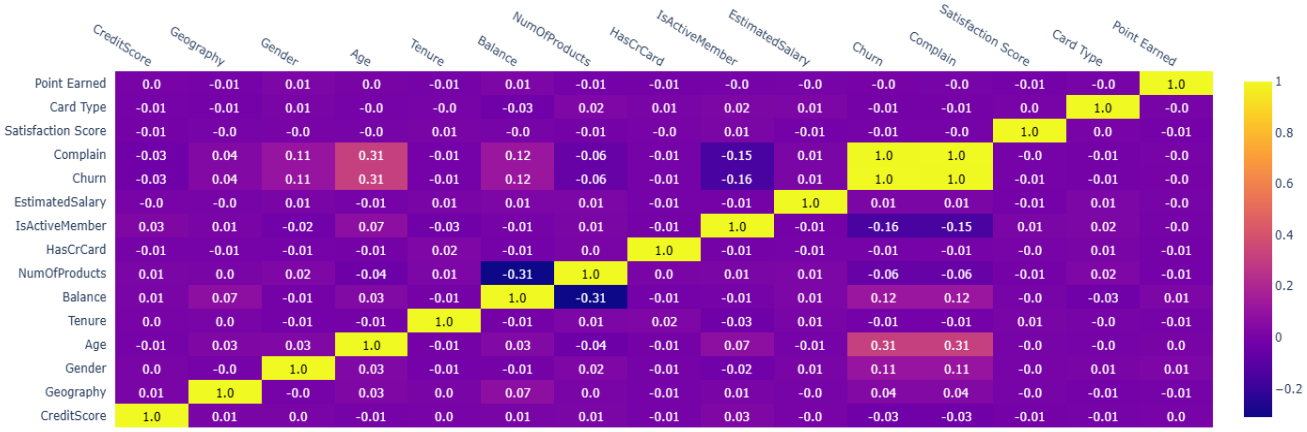
Modelling:

- Random Forest was implemented to predict Customer Churn.
- The data was divided into two sets: 70% for training and 30% for testing.

Results:

Key Visualizations:

- Heatmap of correlations illustrating the connections between numerical features and target variable



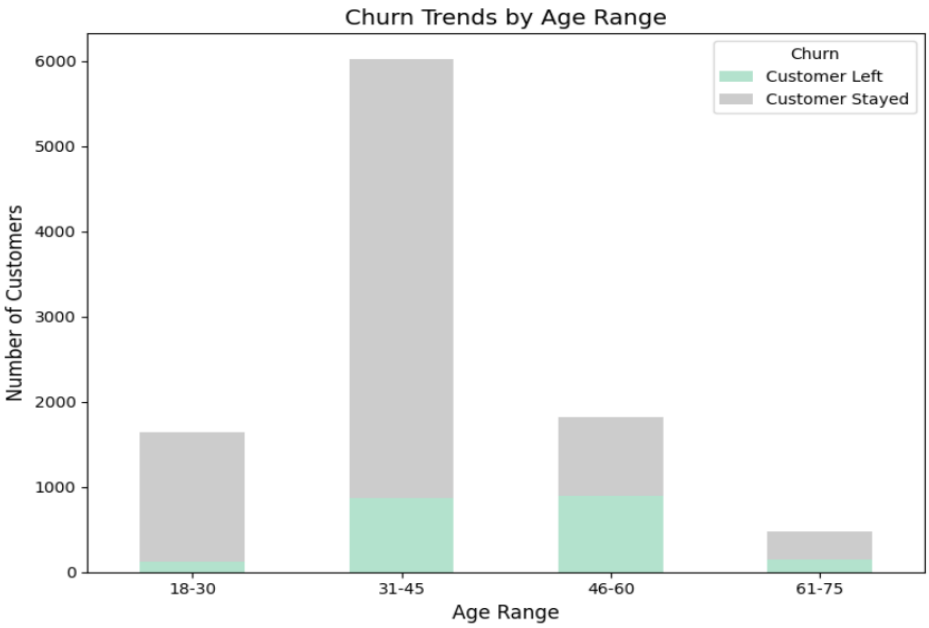
Complain and Churn (1): shows almost perfect relationship of 1 which cause data leakage or multicollinearity.

NumOfProducts and Balance (-0.31): A moderate negative correlation indicates that customers with fewer products tend to have higher balances. This might reveal different financial behaviours based on product usage.

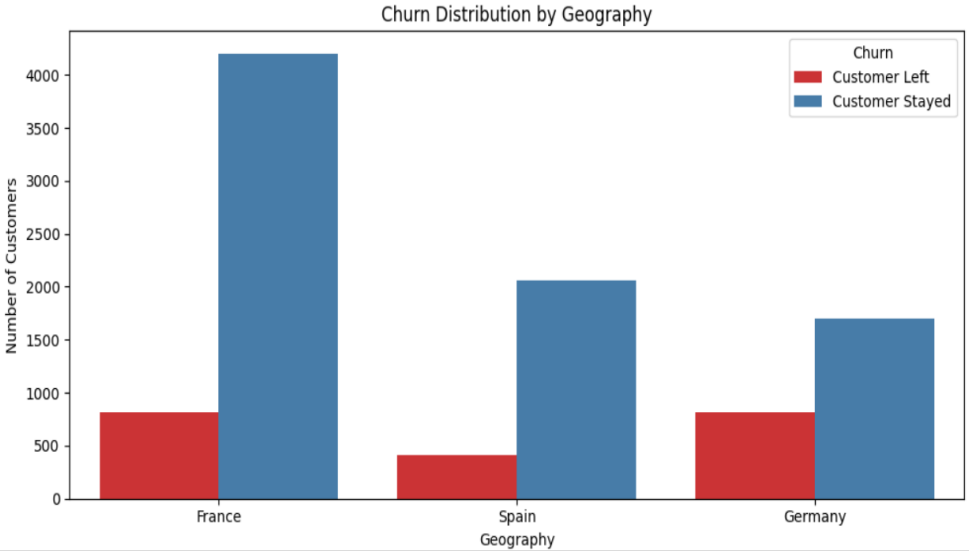
Age and Churn (0.31): A moderate positive correlation suggests that older customers are likely to churn.

IsActiveMember and Churn (-0.16): depict a weak negative relationship. This means that active members are slightly less likely to churn compared to inactive members.

- Bar charts and count plots of churn rates segmented by service and demographics



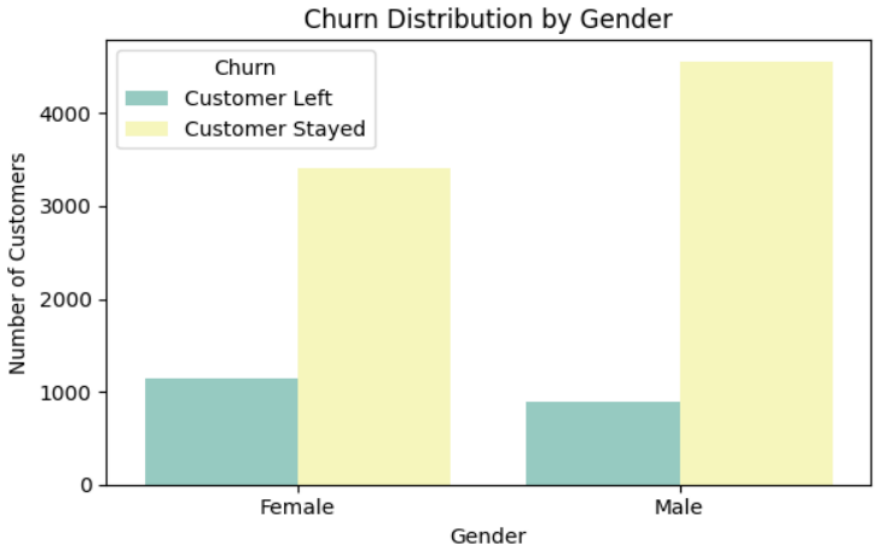
Retention is highest among 46-60 years, while 31-45 years shows notable churn. The other age groups have fewer customers with minimal churn overall.



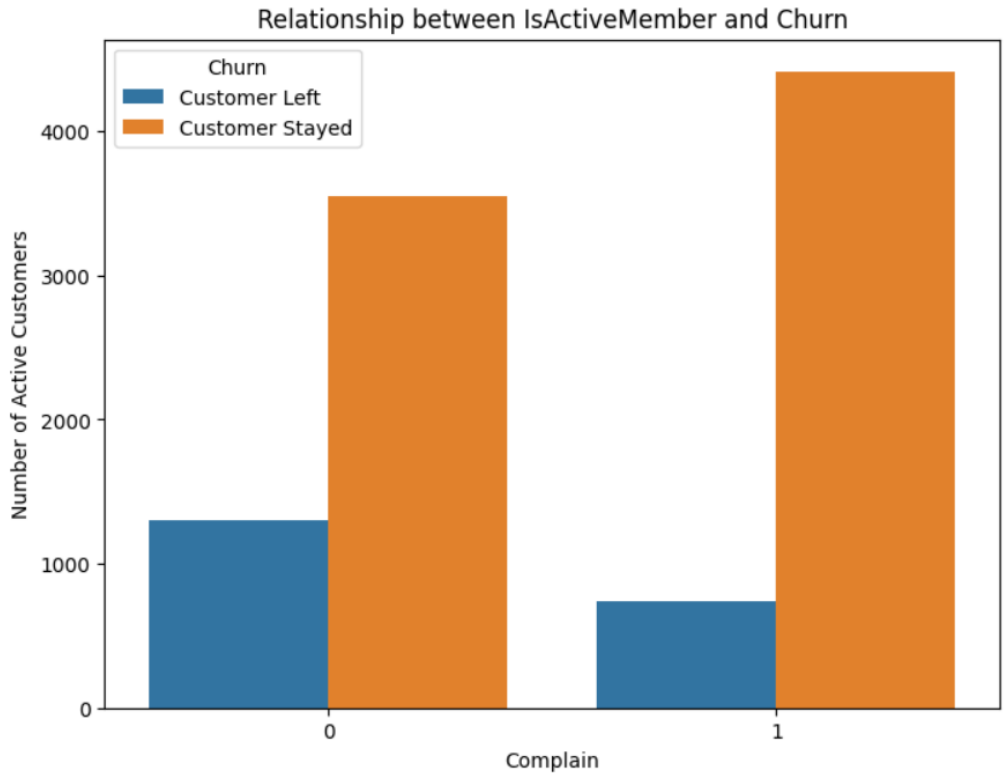
- France: Approx 20% churn (1000 out of 5000 customers).

- Spain: Around 16.7% churn (500 out of 3000 customers).
- Germany: About 25% churn (500 out of 2000 customers).

Germany has the highest churn rate, while Spain has the lowest.

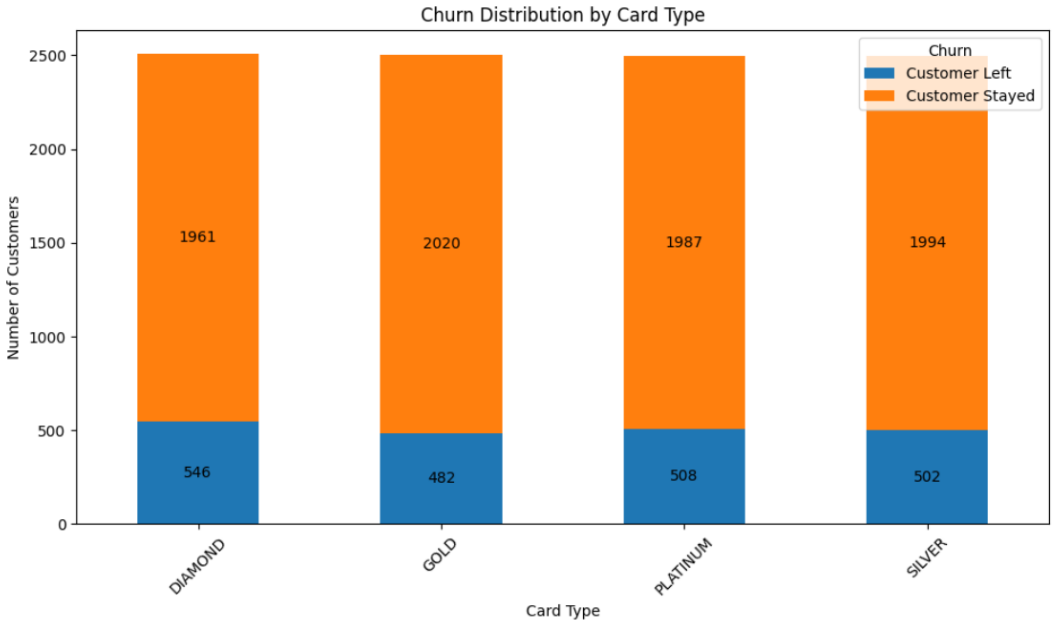


- **Females**: a little higher churn than for men.
- **Males**: Churn is lower, with more customers staying.

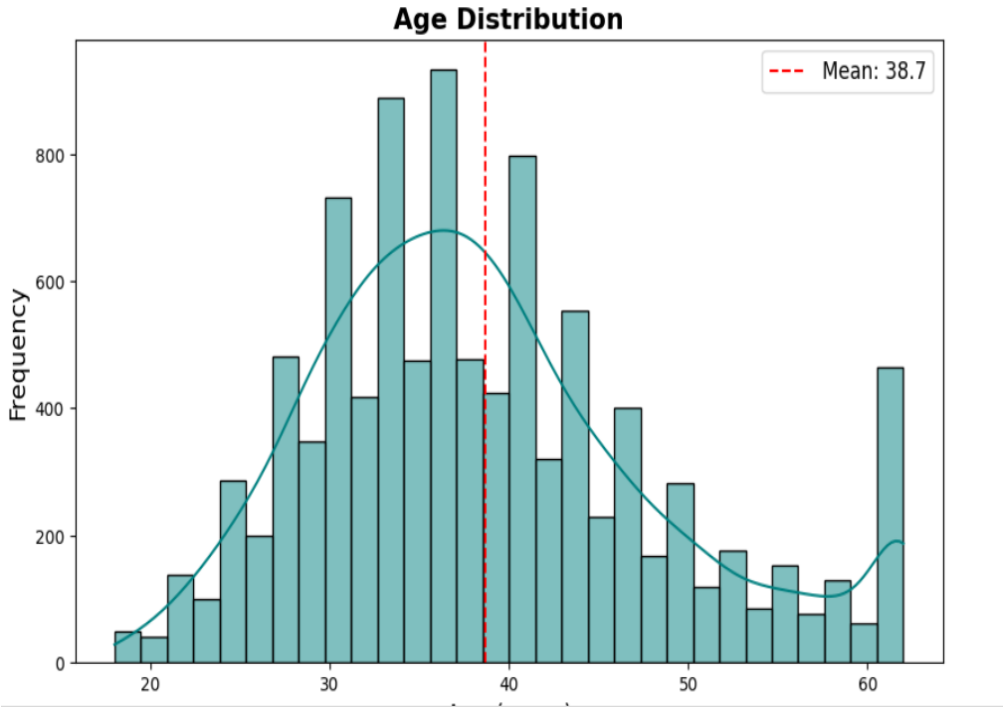


- Active members are more likely to stay, even if they complained.
- Inactive members have higher churn rates, regardless of complaints.

- Stacked bar charts displaying churn ratios across different categories



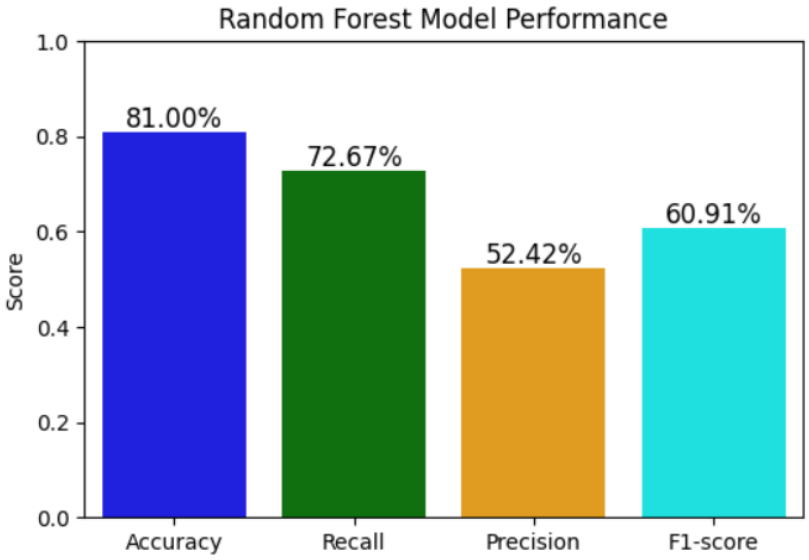
Gold cards have the highest retention (2,020 customers), and Diamond cards experience the most churn (546 customers).



The age distribution in the churn dataset shows that most customers are in their mid-30s to early 40s, with an average age of 38.7 years.

Model Performance (Random Forest):

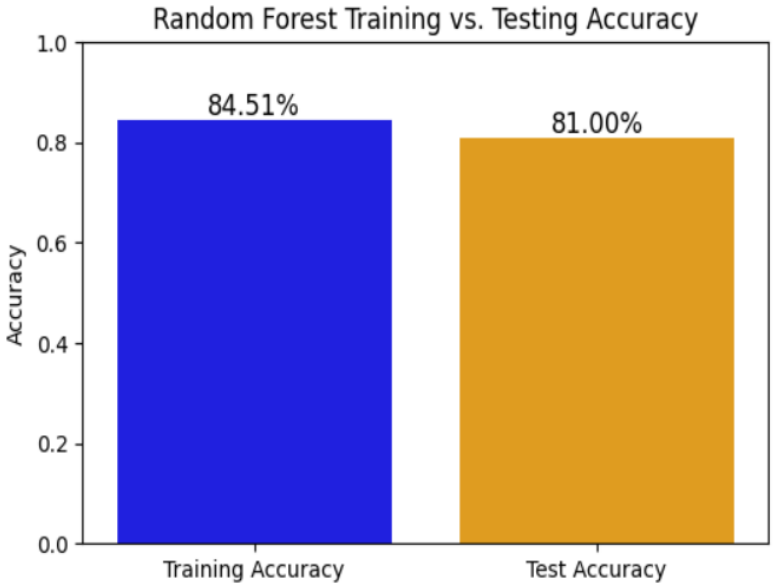
Accuracy	**81.00%**
Recall	**72.67%**
Precision	**52.42%**
F1-Score	**60.91%**

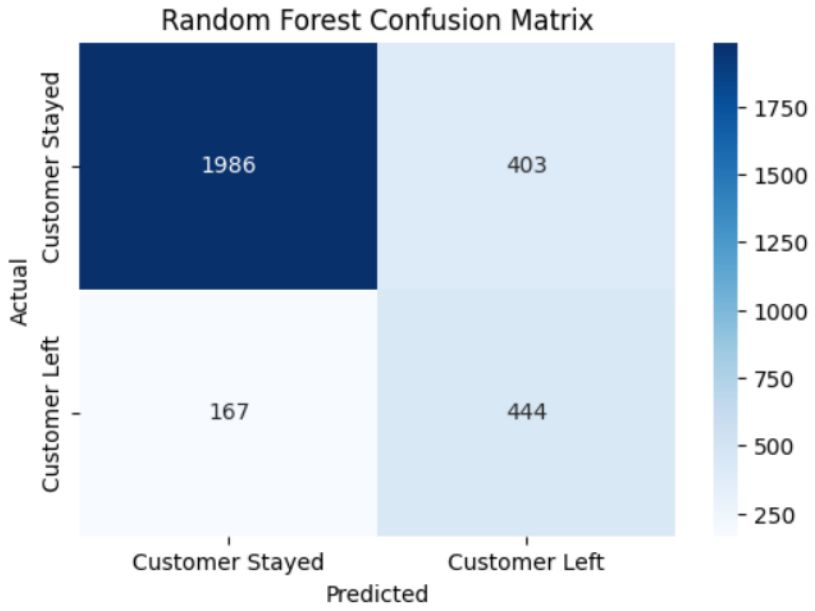


Confusion Matrix & Accuracy Charts:

- Visualized using seaborn heatmaps.

- Compared training vs testing accuracy.





Summary Statistics:

- The dataset had 10000 records and 18 features.
- Churn rate observed was approximately 20.4%.
- Features like Age, IsActiveMember, Balance, and Gender showed strong correlation with churn.

Discussion:

Patterns & Insights:

- Customers with low satisfaction or who are inactive are more likely to leave.
- Age distribution reveals churn is common in customers in their late 30s to early 40s.
- Card type impacts churn - gold card holders show highest retention.
- Tenure showed an inverse correlation with churn - the longer a customer stayed, the less likely they were to churn.

Challenges:

- Data leakage due to "Complain" column (removed).
- Class imbalance mitigated using SMOTE.
- Limited external data - future models may benefit from behavioural data (e.g., call centre logs).

Conclusion

This analysis highlights key predictors of churn such as satisfaction score, customer activity, and geographic region. A Random Forest model was built with an accuracy of 84%, helping the bank to better identify at-risk customers. Future work could involve:

- Testing advanced models like XGBoost or LightGBM
- Including time-series or transactional data
- Deploying the model in a real-time alert system

References

- Kollipara, R. (2022) *Bank Customer Churn*, Kaggle. Available at: https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/data (Accessed: 10 April 2025).
- Scikit-learn Developers (2024) *scikit-learn: Machine Learning in Python*. Available at: https://scikit-learn.org/stable/ (Accessed: 10 April 2025).
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.