Bank Customer Churn Analysis Project

Submission Date: Nov 15, 2024

Group Name:

- Buli Deressa
- Berihun Mekonnen
- Musa Shikur
- Yusuf Habib
- Yared Asefa

Summery

The Bank Customer Churn Analysis project aims to predict whether a customer will stay with the bank or leave (i.e., "churn") based on various features or attributes of the customer. The goal is to develop a predictive model that can identify customers who are at risk of churning, allowing the bank to take proactive steps to retain them (through targeted marketing, special offers, or improved services).

Dataset Overview

The dataset typically contains information about the customers, including both their personal details and transactional behavior. Common columns in a Bank Customer Churn dataset may include:

- **Customer ID**: Unique identifier for each customer.
- **Credit Score**: A measure of the customer's creditworthiness.
- **Geography**: The country or region where the customer is located.
- **Gender**: The customer's gender (Male/Female).
- **Age**: The age of the customer.
- **Tenure**: The number of years the customer has been with the bank.
- **Balance**: The account balance of the customer.
- **Number of Products**: The number of products (e.g., savings account, credit card, loan) the customer holds with the bank.
- **Has Credit Card**: Whether the customer has a credit card or not.
- **Is Active Member**: Whether the customer actively engages with the bank's services.
- **Estimated Salary**: The estimated income or salary of the customer.
- **Exited**: Target variable (1 = Churned, 0 = Retained).

Problem Objective

The primary goal of the project is to predict customer churn, i.e., whether a customer will leave the bank or stay. This is a binary classification problem, where the model needs to predict either:

- **Exited = 1**: The customer churned (left the bank).
- **Exited = 0**: The customer stayed with the bank.

Data Preprocessing

Before building the predictive model, the data typically undergoes several steps of preprocessing:

- **Handling Missing Values**: Ensure that there are no missing or null values in the dataset (e.g., filling in missing numerical or categorical data).
- **Feature Engineering**: You may create new features like `balance_per_product` or categorize customers based on credit scores.
- **Convert categorical variables**

Exploratory Data Analysis (EDA)

The analysis of the dataset helps understand trends, distributions, and relationships between features. EDA includes:

- **Distribution of Features**: Visualizing numerical columns (like `Age`, `CreditScore`, `Balance`) with histograms or boxplots.
- **Correlation Analysis**: Understanding relationships between features using correlation matrices or heatmaps.
- **Churn Analysis**: Examining the relationship between features and churn (e.g., what characteristics are common among churned customers).

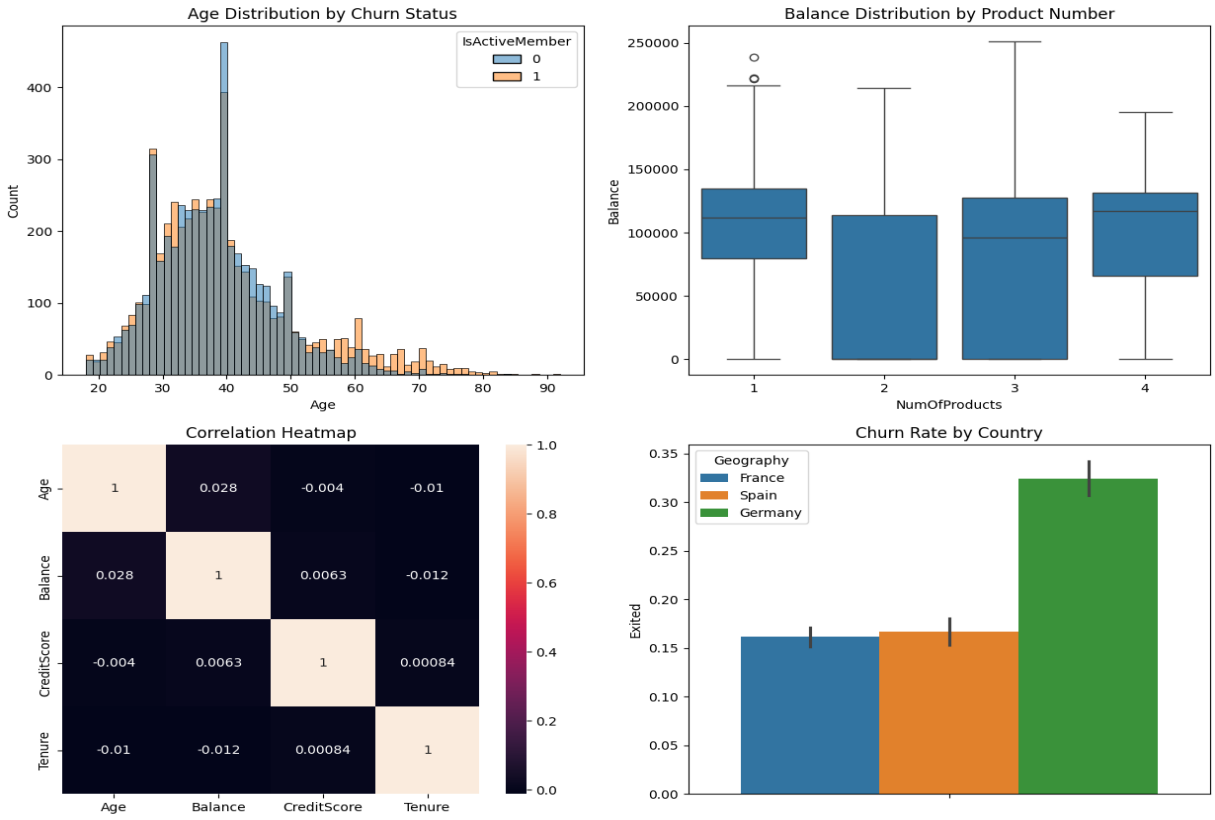


Figure 1: Exploratory Data Analysis and Visualization

Predictive Analysis

Classification Metrics

- **Recall, or true positive rate**

The true positive rate (TPR), or the proportion of all actual positives that were classified correctly as positives, is also known as recall. False negatives are actual positives that were misclassified as negatives, which is why they appear in the denominator. A hypothetical perfect model would have zero false negatives and therefore a recall (TPR) of 1.0, which is to say, a 100% detection rate.

Recall is mathematically defined as:

Recall (or TPR)=(correctly classified actual positivesa/ll actual positives)=(TP/(TP+FN))

- **Precision**

Precision is the proportion of all the model's positive classifications that are actually positive. It is mathematically defined as:

Precision=(correctly classified actual positives/everything classified as positive)e=(TP/(TP+FP))

- **F1 score**

F1 score is a measure of the harmonic mean of precision and recall. Commonly used as an evaluation metric in binary and multi-class classification

Interpretation on result of the Models

In this Modeling Activity, as can be seen from screen shot below, we check and got the performance of two Classification models that is:

- Logistic Regression Classification
- Random Forest Classification Report

Logistic Regression Classification output:

```
Logistic Regression Classification Report
              precision    recall  f1-score   support

           0       0.83      0.97      0.89      1607
           1       0.60      0.18      0.27       393

    accuracy                           0.81      2000
   macro avg       0.71      0.57      0.58      2000
weighted avg       0.78      0.81      0.77      2000

Confusion Matrix:
 [[1561   46]
 [ 324   69]]
```

Random Forest Classification output

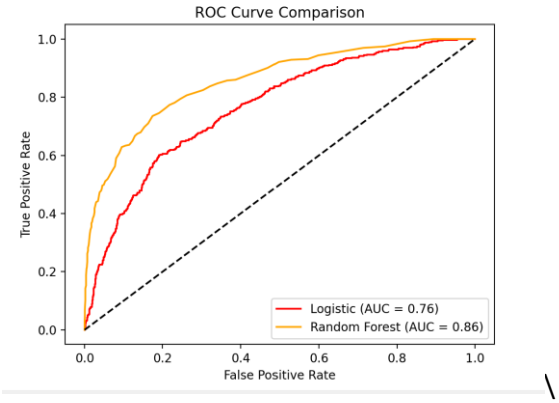
```
 Random Forest Classification Report
              precision    recall  f1-score   support

           0       0.88      0.96      0.92      1607
           1       0.76      0.45      0.56       393

    accuracy                           0.86      2000
   macro avg       0.82      0.71      0.74      2000
weighted avg       0.85      0.86      0.85      2000
```

```
Confusion Matrix:
 [[1550   57]
 [ 216  177]]
```

Based on the result Random Forest Classification outperform Logistic Regression Classification. This result also confirmed by rock curve comparison diagram, below, between the two classification models. But both model perform well.



Discussing overall effect of features on churn

Insight from summary of churn data

- **Gender vs Churn**
  Based on Analysis on the data, Females has more churn rate than men.

  Churn Rate by Gender:
  Female 0.250715
  Male 0.164559

- **AgeGroup vs Churn**
  The Analysis on the data also indicate that Seniors have more churn rate followed by Middle-aged

  Churn Rate by AgeGroup:
  Young 0.074246
  Middle-aged 0.171947
  Senior 0.459792

  Country vs Churn
  Churn Rate by Country:
  France 0.161548
  Germany 0.324432
  Spain 0.166734

Insight from Visualization

Age:

Based to the histogram visualization, there is less churn at young age and high churn rate in senior age.

Estimated Salary:

According to graph we get, Higher salary reduce churn up to some point

Recommendations

Since Analysis show at young age there will be less churn, this age is recommended. Also we have to focus on higher Estimated Salary Customers they have less churn rate. Lastly we have to work on Geography like Germany with higher churn rate so that we will reduce churn at this geographic area.