Comprehensive Data Analysis Report

SAMYAK JAIN
GitHub Repository

July 10, 2024

Bank Statements Analysis

Transaction Analysis

- Total Number of Transactions: 985
- Distribution of Transaction Amounts (Small vs. Large):

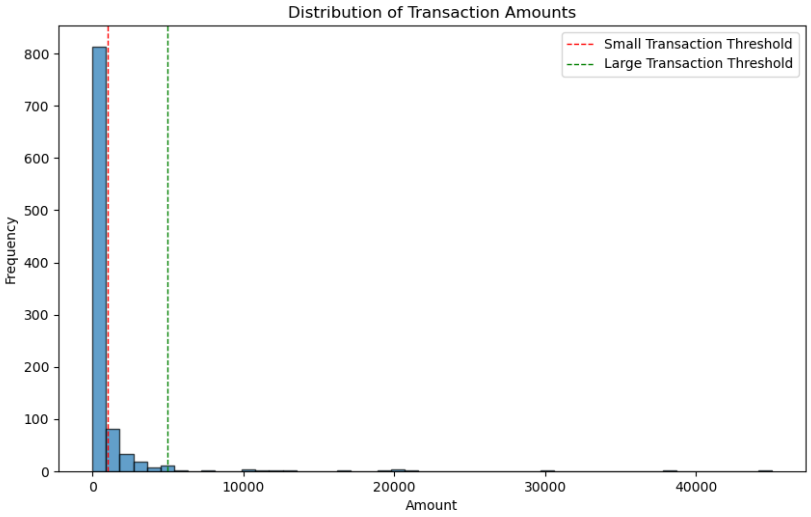

Figure 1: Distribution of Transaction Amounts

- Frequency of Different Transaction Types (Debit vs. Credit):

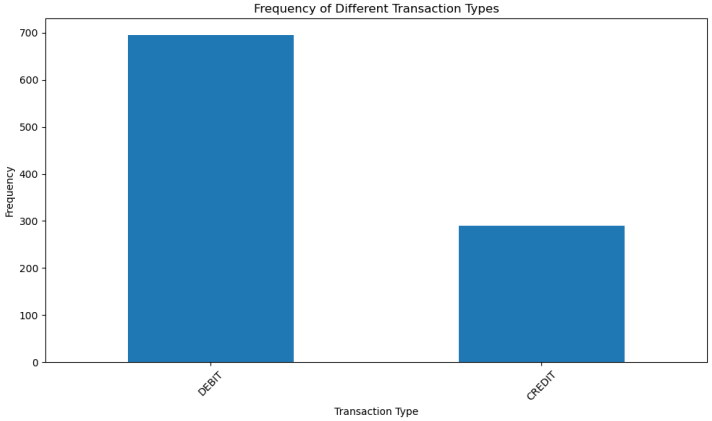

Figure 2: Frequency of Different Transaction Types

- Insights:
  - 84.87% of transactions are small (≤1000), 12.89% are medium, and 2.23% are large (>5000).
  - There are 695 debit transactions and 290 credit transactions.
  - The average transaction amount is 855.49, with a standard deviation of 3007.52.

Balance Analysis

- Trend of Account Balance Over Time:

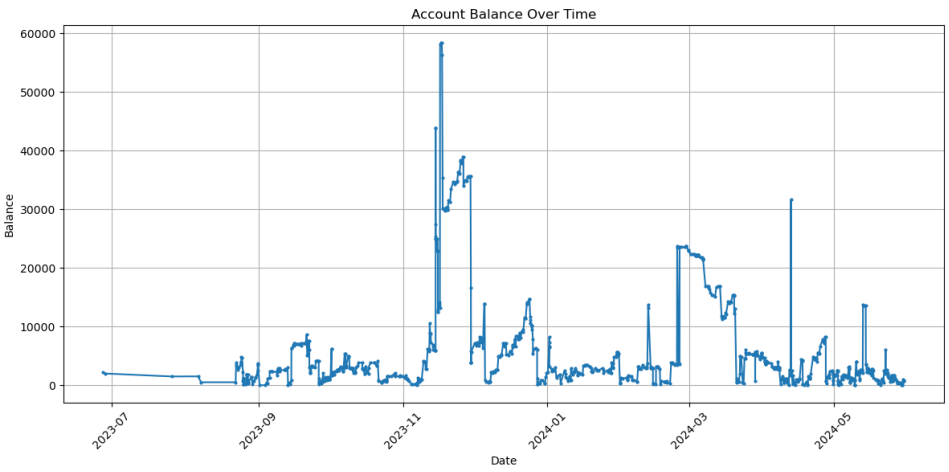

Figure 3: Trend of Account Balance Over Time

- Insights:
  - The initial balance was 2180.80 and the final balance is 761.41.
  - Over the period, the balance has decreased by 1419.39.
  - The highest balance was 58450.80 and the lowest was 0.80.

Spending Patterns

- Total Spending by Category (Debit Transactions):

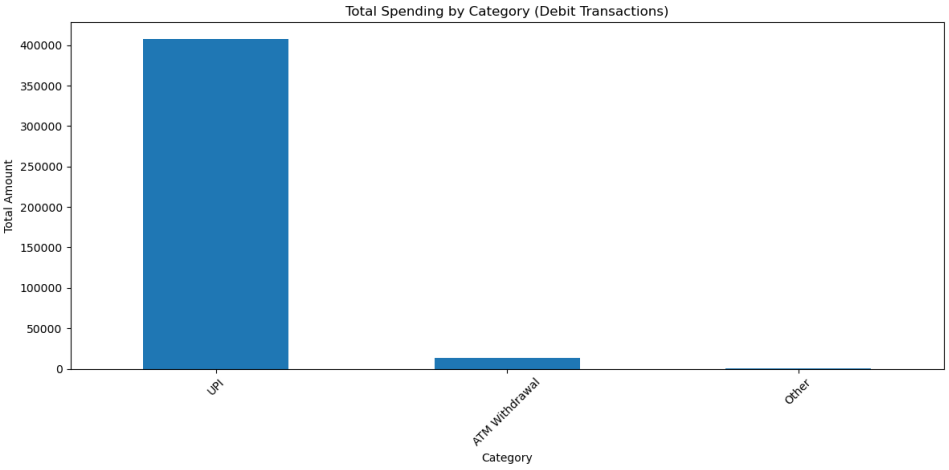

Figure 4: Total Spending by Category (Debit Transactions)

- Insights:

- The top spending category is UPI, with a total spend of 407759.90.
- The most frequent spending category is UPI, with 688 transactions.
- The category with the highest average transaction amount is ATM Withdrawal, with an average of 4500.00 per transaction.

Income Analysis

- Income Patterns Over Time:

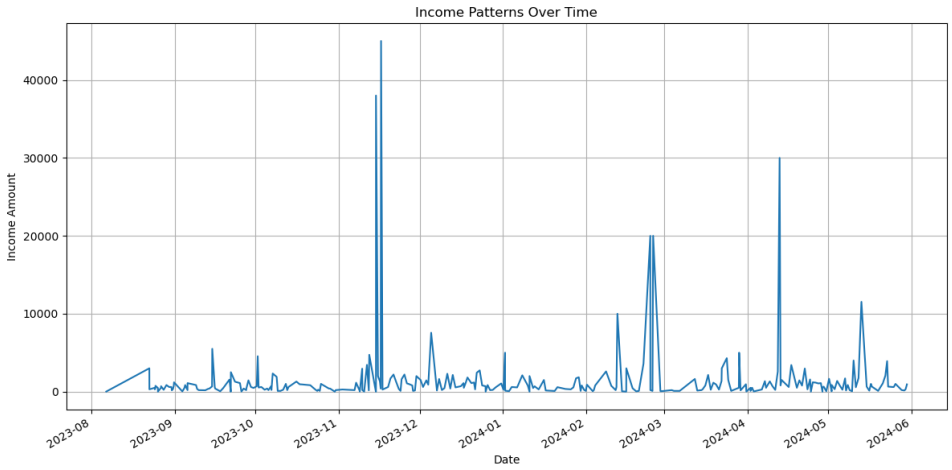

Figure 5: Income Patterns Over Time

- Insights:
  - The main source of income is from the Other category, totaling 241237.51.
  - There are 2 different categories of income sources.
  - The average income transaction is 1450.24, occurring 290 times over the period.

Alert Generation

- Alerts Generated for Low Balance or High Expenditure Periods
- Insights:
  - There are 18 unusual transactions that deviate significantly from the average.
  - The account balance fell below 5845.08 (10% of max balance) 740 times.
  - There were 31 days with unusually high spending (more than twice the daily average of 1642.37).

Office Supplies Data Analysis

Sales Analysis

- Total Sales by Product Category:

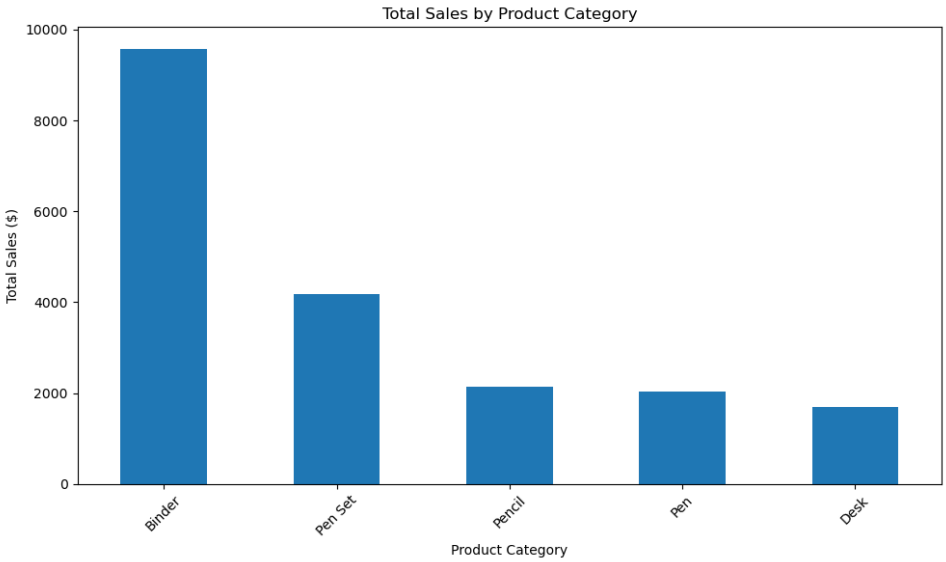

Figure 6: Total Sales by Product Category

- Insights:
  - The product category with the highest sales is Binder, followed by Pen Set and Pen.
  - Pencil is the best-selling product by units sold, but it doesn't necessarily translate to the highest revenue due to its lower unit price.
  - There's a significant difference in sales between the top-selling categories and the others, suggesting a focus on high-volume, lower-priced items.

Customer Analysis

- Top 10 Customers by Sales:

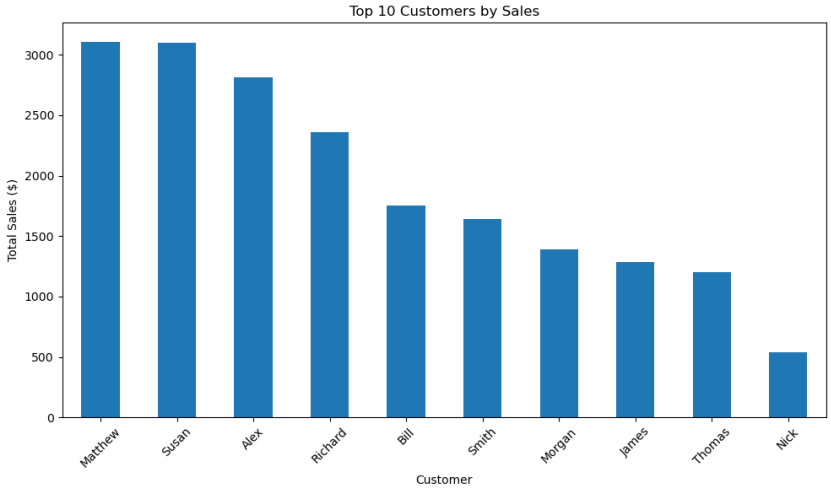

Figure 7: Top 10 Customers by Sales

- Insights:
  - Richard is the top customer by sales, followed by Susan and Alex.
  - There are 11 unique customers in the dataset.
  - Richard, Alex, and Bill have the highest purchase frequency, indicating they are frequent buyers.
  - The top customers contribute significantly more to sales than others, suggesting a focus on key account management might be beneficial.

Time Series Analysis

- Monthly Sales Trends Over the Past Year:

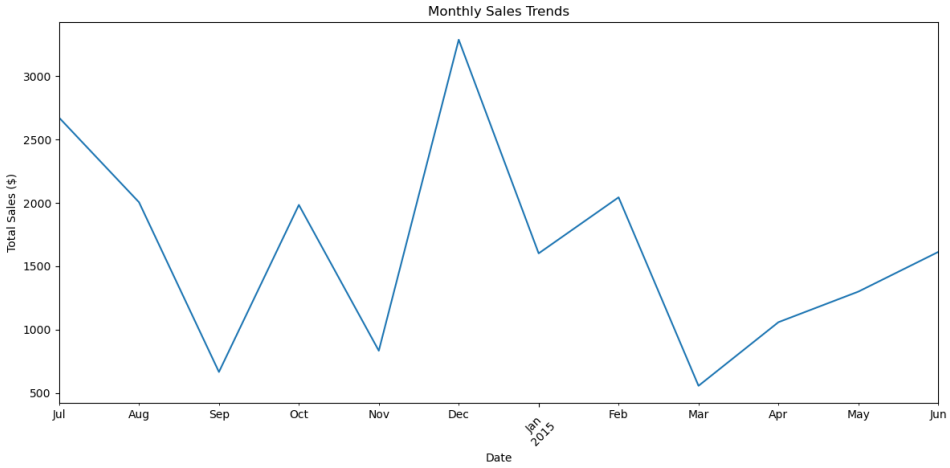

Figure 8: Monthly Sales Trends

- Insights:
  - Sales show some fluctuation over time, with peaks in December 2014 and April 2015.
  - There's no clear consistent seasonal pattern, but there seems to be a slight upward trend over time.
  - The business might want to investigate the factors contributing to the sales peaks to potentially replicate those conditions.

Geographical Analysis

- Sales Distribution by Region:

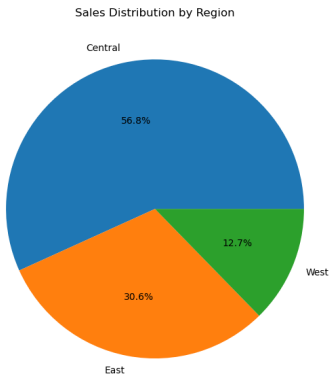

Figure 9: Sales Distribution by Region

- Insights:
  - The Central region generates the most sales, followed by the East region.
  - The West region has significantly lower sales compared to the other regions.
  - This suggests that the business might want to focus on expanding its presence or marketing efforts in the West region to boost sales.

Profit Analysis

- Total Profit by Product Category:

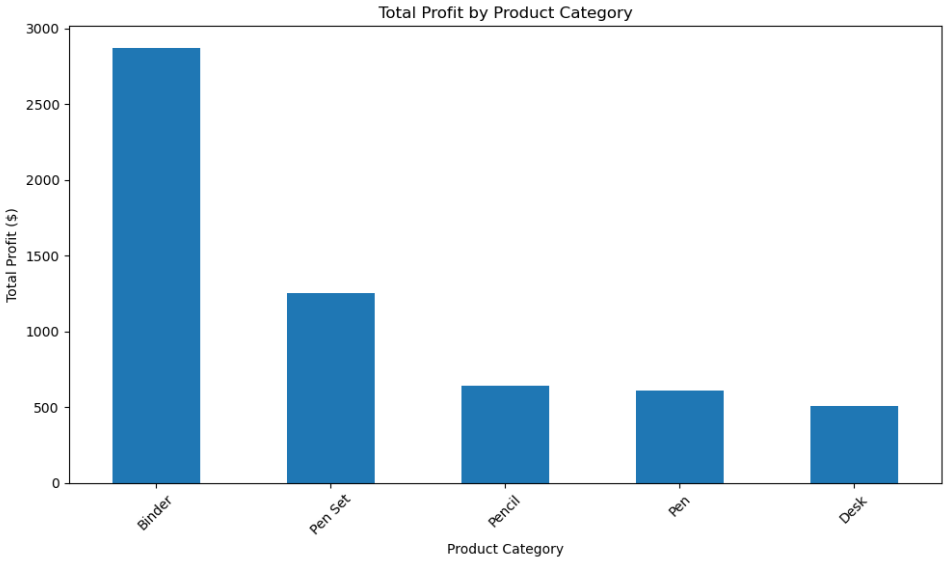

Figure 10: Total Profit by Product Category

- Insights:
  - Binder is the most profitable product category, followed by Pen Set and Desk.
  - The top 10 most profitable products closely mirror the top-selling products, which is expected given the assumed constant profit margin.
  - Despite lower unit sales, high-priced items like Desk contribute significantly to overall profits.

Churn Modelling Data Analysis

Customer Demographics

- Distribution of Customers Across Different Age Groups:

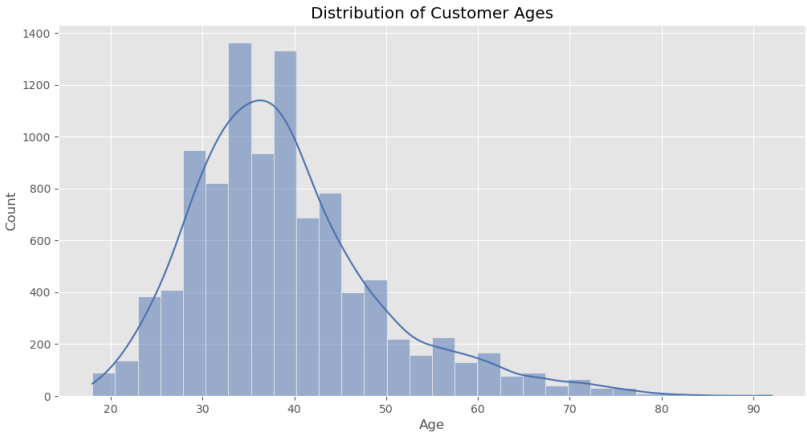

Figure 11: Distribution of Customer Ages

- Gender Distribution of Customers:

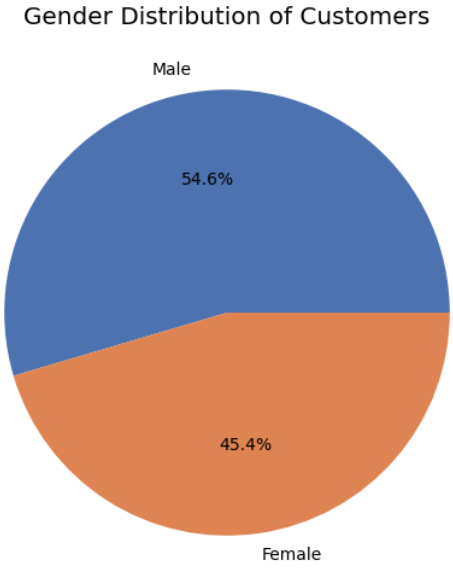

Figure 12: Gender Distribution of Customers

- Insights:
  - The average age of customers is 39 years.
  - The largest age group is 31-40 with 4451 customers.
  - The gender distribution shows 54.6% male and 45.4% female customers.
  - There's a relatively even distribution of customers across age groups from 30 to 60, with fewer very young or very old customers.

Churn Analysis

- Churned vs. Non-Churned Customers:

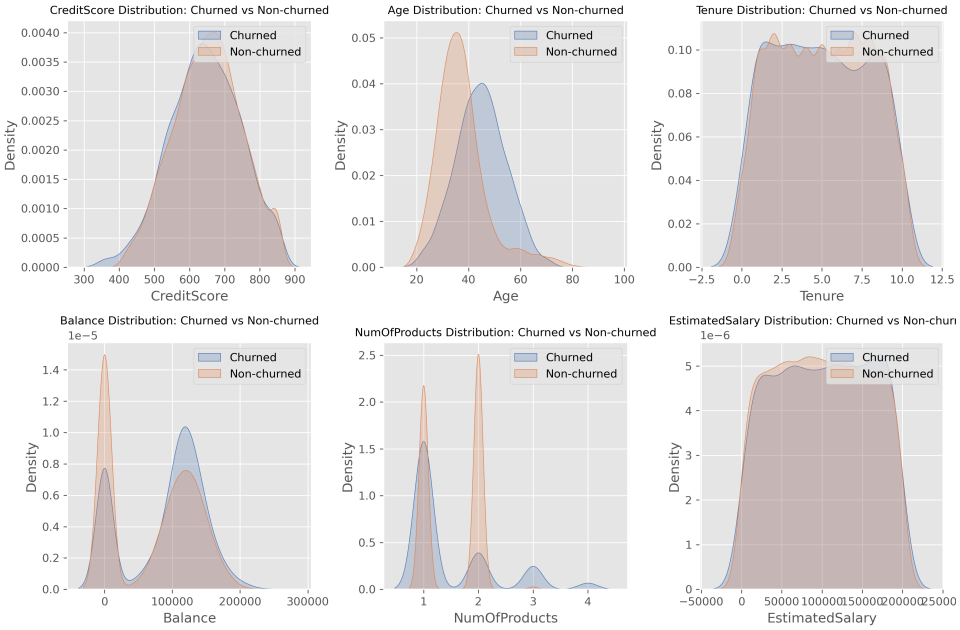

Figure 13: Churned vs. Non-Churned Customers

- Insights:
  - The overall churn rate is 20.37%.
  - Middle-aged customers (30-50) tend to have higher churn rates.
  - There's a slight difference in churn rates between genders, with females showing a slightly higher churn rate.
  - Customers from Germany have a notably higher churn rate compared to France and Spain.
  - The number of products and account balance show the strongest correlation with churn among numerical variables.

Product Usage

- Distribution of Number of Products Used by Customers:

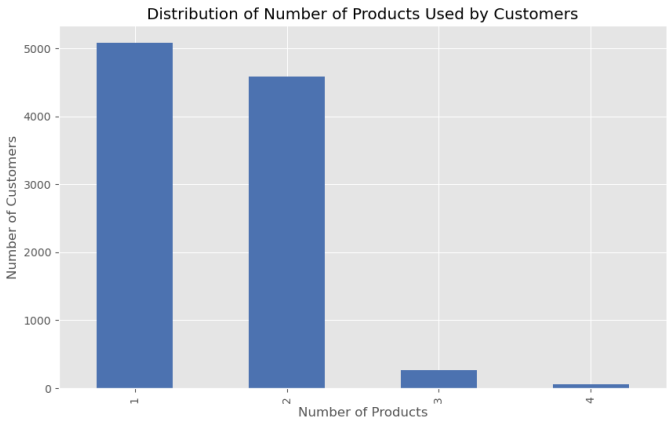

Figure 14: Distribution of Number of Products Used by Customers

- Insights:
  - The most common number of products used is 1 with 5084 customers.
  - There's a notable difference in product usage across different countries.
  - Younger customers tend to use fewer products compared to older customers.
  - Customers with a very high number of products (3 or 4) are more likely to churn, possibly due to complexity or cost.

Financial Analysis

- Account Balance Distribution:

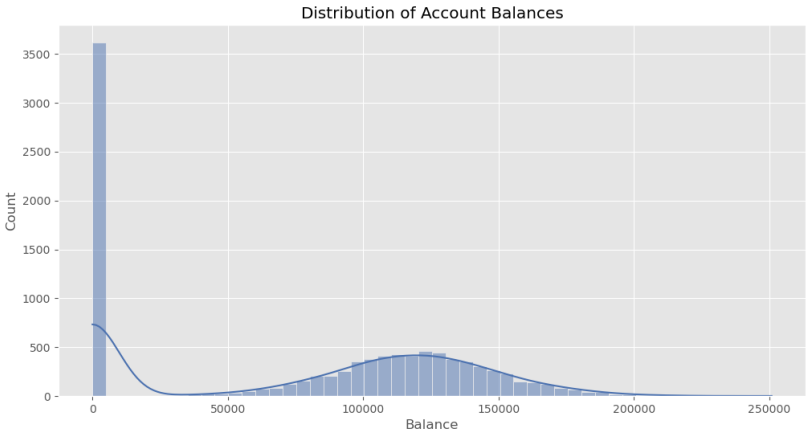

Figure 15: Account Balance Distribution

- Insights:
  - The average account balance is $76485.89.
  - There's a wide range of account balances, with a significant number of customers having very low balances.
  - Churned customers tend to have slightly higher account balances on average.
  - There's no significant difference in credit scores between churned and non-churned customers.

Predictive Modeling

- Top 10 Most Important Features (Random Forest):

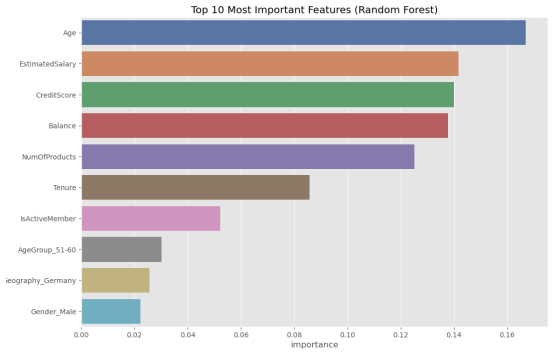

Figure 16: Top 10 Most Important Features (Random Forest)

- Insights:
  - The Random Forest model performs the best overall, with the highest accuracy and ROC AUC score.
  - The top predictors of churn are: Age, Balance, EstimatedSalary, and Geography.
  - All models struggle somewhat with recall for the churn class, indicating they miss some customers who actually churn.
  - The cross-validation results confirm that Random Forest is the most consistent performer across different subsets of the data.
  - To improve the model, we could focus on increasing the recall for churn prediction, possibly by adjusting the classification threshold or using techniques like oversampling the minority class.
- Receiver Operating Characteristic (ROC) Curve:

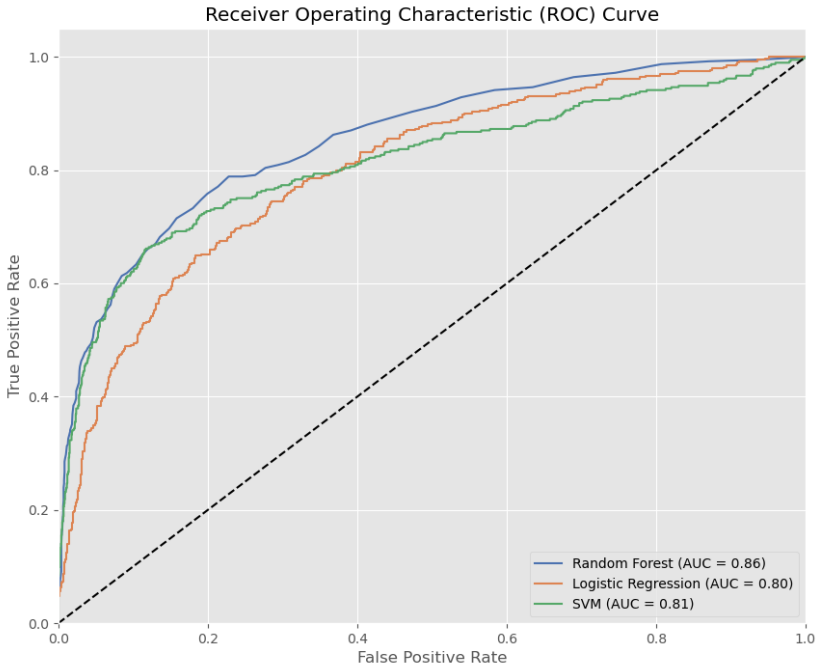

Figure 17: Receiver Operating Characteristic (ROC) Curve