Customer Churn Prediction: Bank Customers

Developed By: Raj Purohith Arjun (UIN: 535005964)

Subject Code: STAT 650

1.Objective

The main goal of this project is to perform a comprehensive analysis of the bank's customer churn data to identify the key factors that lead to customer churn. Using advanced Exploratory Data Analysis (EDA) techniques, we will visualize patterns in the dataset to answer specific business-related questions. This project aims to understand the relationships between different variables (both quantitative and qualitative) and their impact on customer retention or churn, leading to actionable business insights.

Purpose

Customer churn refers to the scenario where customers stop doing business with a company. Churn is a critical issue in industries like banking, where retaining customers is cheaper than acquiring new ones. The objective here is to analyze the customer churn data and conduct detailed statistical analysis to help the bank take preventive measures against customer attrition.

---

2. Dataset Description

The dataset used for this project is focused on bank customers and includes features related to demographics, account details, and usage. The target variable is whether the customer has churned (left the bank) or not.

Data Set Source: Kaggle - Bank Customer Churn Prediction

Dataset Characteristics:

- **Rows (Observations):** 10,000 customers
- **Columns (Variables):**
- Variables Included: The dataset contains 14 variables. Here is a summary of the key variables:
- Geography: The country where the customer is located (e.g., France, Spain) (string).
- Gender: The customer's gender (Male/Female) (string).
- CreditScore: The customer's credit score (integer).
- Age: The customer's age (integer).
- Tenure: Duration of the customer relationship with the bank (integer).

- Balance: The balance in the customer's bank account (numeric).
- NumOfProducts: Number of products purchased by the customer from the bank (integer).
- EstimatedSalary: The customer's estimated annual salary (numeric).
- IsActiveMember: Is the customer a active member (binary)
- HasCrCard : If member has credit card(binary)

Target Variable:

- Exited: Indicates whether the customer has exited the bank (1 = Exited, 0 = Not Exited).

Irrelevant Variables:

- RowNumber: Just an index, adds no value to analysis.
- CustomerId: Unique identifier, irrelevant to customer behavior or churn prediction.
- Surname: Personal identifier, not linked to churn behavior and unnecessary for analysis.

Variable Types:

- **Qualitative (Categorical) Variables:** Gender, Geography, HasCrCard, IsActiveMember.
- **Quantitative Variables:** Age, Balance, CreditScore, EstimatedSalary.

---

3. Data Pre-Processing

3.1 Importing the Dataset

In this step, we import the necessary Python libraries (pandas, numpy, seaborn, and matplotlib) and load the customer churn dataset. We display the first few rows of the dataset to familiarize ourselves with the structure.

3.2 Data Cleaning

In this step, we will:

- Handle Missing Values
- Remove Duplicates
- Correct Data Inconsistencies

3.2.1 Handling Missing Values

First, we checked for any missing values in the dataset. If missing values were present, we could either impute them using the mean, median, or mode, or remove rows/columns with missing data. For this project, we chose to drop rows with missing values for simplicity, ensuring that the dataset is clean.

Original dataset shape: (10,000, 14)
Cleaned dataset shape: (9,991, 14)

Treating Missing Values:

CreditScore: 9 missing values

After thoroughly reviewing the dataset, we identified missing values in the CreditScore column. We dropped the rows with missing values for the simplicity of the project. After addressing these missing values, the modified file was checked once again, and no null values were found.

3.2.2 Removing Duplicates

It's important to ensure that there are no duplicate records, as they can skew the analysis. We checked for duplicate rows in the dataset. Duplicates can mislead the model and analysis, so we removed any found duplicates to ensure the dataset contains unique customer records. We used the **drop_duplicates()** function to remove duplicate rows.

3.2.3 Correcting Data Inconsistencies

We checked for any inconsistencies or errors in the dataset, such as impossible values (e.g., negative age).

NOTE:
We looked for any inconsistent data, such as negative values in columns where it doesn't make sense (e.g., age). We found no negative ages in this dataset, but if we did, we would correct or impute them as appropriate.

3.2.4 Dropping Columns Not Used in Analysis

Some columns may not contribute to the analysis or prediction of churn. Thus, we dropped RowNumber, CustomerId, and Surname.

NOTE:
The columns RowNumber, CustomerId, and Surname were removed because:

- The 'RowNumber' was removed as it served merely as an index and added no value to the analysis.
- The 'CustomerId' was deemed irrelevant to customer behavior or churn prediction, as it was just a unique identifier.

- Lastly, the 'Surname' was excluded since it is a personal identifier with no linkage to churn behavior and is unnecessary for analysis.

After making these adjustments, the modified DataFrame was displayed to confirm the changes.

4. Exploratory Data Analysis (EDA)

Primary Goals of Exploratory Data Analysis (EDA)

The primary goals of EDA are to comprehensively understand the dataset, identify patterns and trends, detect outliers, evaluate missing data, explore variable relationships, and inform the data cleaning and feature selection processes. Through these steps, we aim to build a solid foundation for subsequent modeling efforts.

4.1 Univariate Analysis

A) Summary Statistics

The summary statistics offer valuable insights into the dataset's characteristics, highlighting central tendencies and variability among the features.

1. **Credit Score**: The average credit score is 650.53, indicating that most customers possess a fair credit standing. The relatively high standard deviation of 96.65 suggests considerable variability in credit scores among customers.
2. **Geography**: The majority of customers are from France, which indicates that the bank may have a significant presence in that region. This could influence targeted marketing and retention strategies.
3. **Gender**: The dataset shows a predominance of male customers, with 54.57% identified as male. This gender imbalance might inform customer relationship strategies tailored to different demographics.
4. **Age**: The average age of customers is 38.92 years, with a significant portion falling within their 30s and 40s. This demographic insight could guide marketing initiatives and product offerings aimed at this age group.
5. **Tenure**: An average customer tenure of approximately 5 years suggests that many customers have established relationships with the bank, potentially leading to brand loyalty.
6. **Balance**: The mean balance is 76,485.89, but the presence of customers with zero balances indicates that some accounts may be inactive or underutilized. This aspect requires further exploration to understand customer engagement.
7. **NumOfProducts**: With an average of 1.53 products per customer, many customers hold only one product, suggesting opportunities for cross-selling and upselling.
8. **HasCrCard**: About 70.55% of customers possess a credit card, indicating a solid base of credit card users, which the bank can leverage for targeted offers.

9. **IsActiveMember**: Around 51.51% of customers are active members, which raises questions about engagement strategies for the remaining customers.
10. **Estimated Salary**: The average estimated salary is 100,090.24, with considerable variation, indicating a diverse customer base in terms of income.
11. **Exited**: The churn rate is approximately 20.37%, suggesting that while most customers remain loyal, there is a notable portion that has exited, necessitating further investigation into the reasons behind this churn.

Insights:

- The dataset is comprehensive, with no missing values, which enhances its reliability.
- Significant variability is evident in key features, particularly balance and estimated salary, which may relate to customer retention strategies.
- The predominance of customers from France and male customers suggests specific demographic trends that could inform targeted marketing efforts.
- A concerning 20% churn rate indicates a need for strategic interventions to improve customer retention.

B) Visualizing Distributions

Visualizations play a crucial role in understanding the distribution of key variables and uncovering patterns related to customer churn.

1. **Histograms**: The histogram for age shows a skew towards younger customers, indicating that the bank may need to focus on retaining this demographic. The credit score histogram presents a normal distribution, suggesting that most customers have healthy credit scores. The balance histogram reveals a significant number of customers with zero balance, indicating potential inactivity.
2. **Density Plots**: The density plot for age confirms the peak around younger customers, while the balance plot highlights two distinct peaks—one at zero and another at higher balances. This duality in balance suggests that the bank may have a mix of inactive and highly engaged customers.
3. **Box Plots**: Box plots for credit score and age illustrate moderate variability, with the presence of lower-end outliers in credit scores. The balance box plot shows a concentration of customers with zero balance, highlighting a critical area for further exploration. The distribution of estimated salary indicates a broad range, with a few notable outliers, suggesting diverse financial circumstances among customers.

Outcome of Analysis

Through univariate analysis, we utilized various visual techniques to delve into the distribution of numerical variables. The results indicate that:

- Customer age predominantly skews younger, emphasizing the need for targeted retention strategies.
- A notable concentration of customers holds zero balances, which could correlate with higher churn rates.
- Variability in credit scores, balances, and salaries suggests diverse customer profiles that require tailored engagement strategies.

Univariate Analysis Limitations

- **Surface-Level Insights**: Univariate analysis provides insights into individual variables but lacks the ability to explore interactions between variables. Important relationships and patterns may remain undetected.
- **Potential Misleading Patterns**: Summary statistics such as mean and median may obscure underlying distributions, particularly in skewed data. This limitation can misrepresent the true characteristics of customer behaviors, especially in variables like Balance and Age.
- **Lack of Context**: Individual variable analysis does not consider contextual factors that may influence customer behavior. For example, understanding how age affects churn requires insights into the interplay between age and other demographic or behavioral factors.

4.2 Bivariate Analysis

Analyzing Relationships and Dependencies

This section explores the connections between variable pairs to detect correlations and dependencies, employing various analytical techniques.

1. Correlation Matrix

The correlation matrix is a fundamental tool for assessing the strength and direction of relationships between variable pairs, with coefficients ranging from -1 to 1:

- **Close to 1**: Strong positive correlation (as one variable increases, so does the other).
- **Close to -1**: Strong negative correlation (as one increases, the other decreases).
- **Around 0**: No correlation.

Example Insight: A correlation of 0.29 between Age and Exited suggests that older customers are more likely to leave the bank.

Key Correlations:

- **Credit Score vs. Age**: Correlation coefficient near zero, indicating minimal linear association.
- **Credit Score vs. Balance**: Correlation of around 0.01, suggesting no meaningful relationship.
- **Age vs. Balance**: Correlation of approximately 0.03, reflecting a negligible connection.
- **NumOfProducts vs. Balance**: Correlation of -0.30 indicates a weak negative relationship, suggesting that as the number of products increases, balance may slightly decrease.
- **Credit Score vs. Estimated Salary**: Essentially zero correlation, showing no influence between these variables.
- **Balance vs. Estimated Salary**: Correlation of about 0.01, indicating no significant link.

Key Insights:

1. Most variable pairs exhibit weak or no relationships, with coefficients close to zero.
2. Individual features may still hold predictive value when combined with other variables in a model.
3. Many features could operate independently when predicting outcomes like customer churn.

2. Scatter Plot Analysis

Scatter plots visualize the relationship between two quantitative variables, helping identify patterns and outliers.

1. **Number of Products vs. Balance**:
    - Most customers possess one or two products, with a higher churn rate among those with fewer products.
    - Exited customers are scattered across various balances, but those with fewer products appear to be more likely to leave.
2. **Credit Score vs. Balance**:
    - Points are evenly distributed across the credit score range, indicating little correlation.
    - Exited customers are interspersed throughout, suggesting that credit score does not effectively distinguish between those who exit and those who remain.
3. **Age vs. Estimated Salary**:

- The plot spans a wide range of ages and salary levels, revealing no clear trend.
- Exited and non-exited customers are evenly distributed across salary brackets, indicating that estimated salary may not strongly predict churn.

4. **Age vs. Balance**:
   - A significant number of customers with zero balances are present across age groups.
   - A concentration of exited customers exists among older individuals with non-zero balances, suggesting age may influence churn, especially for older customers.

Key Takeaways:

- Age appears to have a minor connection to churn, while estimated salary and credit score show no strong link.
- The presence of customers with zero balances across all groups implies that balance may not be a decisive factor in predicting exits.
- Predicting churn may require a combination of various factors rather than relying on a single variable.

3. Pair Plot Analysis

The pair plot illustrates interrelationships among several variables in the dataset, color-coded by Exited status (blue for 0, representing retained customers, and orange for 1, indicating exited customers).

Observations from the Plots:

- **Diagonal Histograms**:
  - **Credit Score**: Displays a slight skew, with many customers scoring between 600 and 700, showing no significant distinction based on exit status.
  - **Age**: Older customers exhibit a higher density of exits, as shown by a greater concentration of orange points in older brackets.
  - **Balance**: Many customers have a balance of zero, while exited customers display a more diverse balance distribution.
  - **NumOfProducts**: Most customers hold either one or two products, with a slight increase in exits among customers with more products.
  - **Estimated Salary**: The distribution is relatively even, lacking clear distinctions between exit statuses.
- **Off-Diagonal Scatter Plots**:

- **Credit Score vs. Age**: No evident pattern links these variables; points are dispersed for both groups.
- **Credit Score vs. Balance**: Similar to the previous plot, no strong correlation exists.
- **Age vs. Balance**: Older customers with varying balances are more likely to exit, evidenced by a concentration of orange points.
- **NumOfProducts vs. Balance**: Customers with three or four products tend to have higher balances; however, no clear exit pattern emerges.
- **Balance vs. Estimated Salary**: No discernible correlation between these variables, indicating independence of exit status.

Useful Insights:

- Age emerges as the most influential factor in determining customer exits, with older individuals more inclined to leave.
- The balance distribution indicates that customers with zero or excessively high balances may be at greater risk of exiting.
- Other factors, such as Credit Score, NumOfProducts, and Estimated Salary, do not exhibit strong direct correlations with exit status.

This bivariate analysis provides essential insights into the factors influencing customer behavior, particularly regarding churn. Understanding these relationships can guide strategies for retention and further model development.

Bivariate Analysis Limitations

- **Correlation vs. Causation**: While bivariate analysis identifies correlations, it does not establish causality. For instance, a positive correlation between Age and Exited does not imply that age directly causes churn.
- **Limited Scope**: Bivariate relationships may overlook more complex interactions that could be significant. For example, while Balance and Exited may show some correlation, factors like NumOfProducts could mediate this relationship.
- **Over-simplification**: Analyzing pairs of variables can lead to oversimplified conclusions, ignoring the effects of additional variables. For instance, the relationship between NumOfProducts and Exited may change significantly when other factors like Age are considered.

4.3 Multivariate Analysis

Multivariate analysis examines interactions among three or more variables, providing insights into complex relationships and patterns within the dataset.

1. Heatmap

A heatmap visualizes correlations among multiple variables using color gradients to indicate the strength and direction of relationships.

Example Insight: In the correlation heatmap for Age, Balance, NumOfProducts, and Exited, a moderate positive correlation (0.29) exists between Age and Exited, while a negative correlation (-0.30) is observed between NumOfProducts and Balance. These correlations help identify key factors related to customer churn.

Analysis of Heatmap

The **heatmap** illustrates the correlations between various numerical variables in the dataset, with values ranging from -1 to 1. Here's a summary of the key findings:

Key Observations:

- **Credit Score**: Shows no strong correlation with other variables, exhibiting very weak relationships with Exited (-0.03), Balance (0.01), NumOfProducts (0.01), and EstimatedSalary (0.00), indicating it is largely independent of these features.
- **Age**: Displays a moderate positive correlation with Exited (0.29), suggesting that older customers are more likely to leave the bank. However, it has weak correlations with other variables, which are close to zero.
- **Tenure**: Has a weak negative correlation with NumOfProducts (-0.30), indicating that longer-tenured customers tend to have fewer products. Overall, tenure shows little to no correlation with other variables.
- **Balance**: Shows a weak positive correlation with Exited (0.12), implying that customers with higher balances are somewhat more likely to exit. It is also negatively correlated with NumOfProducts (-0.30), suggesting that higher balances correspond with fewer products held.
- **NumOfProducts**: Displays a moderate negative correlation with Balance (-0.30) and a very weak negative correlation with Exited (-0.05), indicating that customers with more products are slightly less likely to leave.
- **Estimated Salary**: Shows no significant correlation with any other variable, as all correlation values are nearly zero, suggesting salary does not influence customer behavior.
- **Exited**: Exhibits notable correlations with:
    - **Age** (0.29): Older customers are more likely to exit.
    - **Balance** (0.12): Higher balance customers are slightly more likely to exit.
    - Weak correlations are seen with Credit Score (-0.03), NumOfProducts (-0.05), and Tenure (-0.01).

Insights:

- The most significant correlation is between NumOfProducts and Balance (-0.30), indicating that customers with fewer products generally have higher balances.
- Age and Exited show a moderate positive correlation (0.29), revealing that older customers tend to exit more frequently.
- Tenure has a weak connection with NumOfProducts but shows minimal impact on other variables.
- Variables such as Credit Score, Estimated Salary, and Tenure exhibit negligible correlations, indicating they do not strongly relate to other factors in this dataset.
- In summary, Age and Balance significantly influence customer churn, but a comprehensive multivariate analysis is needed to fully understand the dynamics involved.

2. Pairwise Relationships

Analysis of Pairwise Relationships

The plot illustrates pairwise relationships among variables such as Credit Score, Age, Balance, NumOfProducts, Estimated Salary, Tenure, and the target variable Exited. Points are color-coded, with blue indicating customers who did not exit (0) and orange representing those who did (1).

Key Observations:

- **Age vs. Exited**: Older customers show a higher likelihood of exiting, while younger customers tend to stay.
- **Balance vs. Exited**: Higher balance customers are more inclined to exit, whereas many with a zero balance remain.
- **NumOfProducts vs. Exited**: Most exits are from customers with 1 or 2 products, with fewer exits among those with 3 or more.
- **Credit Score vs. Exited**: No clear linear trend is observed, but certain clusters appear around specific scores.
- **Tenure vs. Exited**: Tenure among exiting customers varies widely, with no distinct trend.
- **Estimated Salary vs. Exited**: Salary does not show a strong correlation with exit status, as both groups are similarly distributed across the salary range.

Diagonal Density Plots: These plots depict the distribution of each variable for both groups (Exited = 0 and Exited = 1). For example, older customers tend to have a higher density of exits, while balance distribution indicates that those with higher balances are more likely to exit.

Insights: The pair plot effectively reveals patterns in how variables relate to the Exited status, particularly with Age, Balance, and NumOfProducts. These relationships may be valuable for predictive modeling.

3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique used for reducing the dimensionality of a dataset by converting it into a set of fewer uncorrelated variables known as principal components. These components are designed to retain the maximum amount of variance from the original data, simplifying analysis while preserving essential information.

Analysis of Output

The data has been condensed into two principal components (Principal Component 1 and Principal Component 2) from the original features: 'Credit Score', 'Age', 'Tenure', 'Balance', 'NumOfProducts', and 'Estimated Salary'.

Key Aspects of the PCA Plot:

- **Principal Components**:
    - **Principal Component 1** (x-axis) captures the highest variance, while **Principal Component 2** (y-axis) captures the second-highest variance. Together, they provide a simplified 2D representation of the dataset while retaining essential information.
- **Color Coding by Exited Status**:
    - The points are color-coded, with blue indicating customers who did not exit (0) and orange representing those who exited (1). This visualization highlights the relationship between the target variable ('Exited') and the principal components.
- **Class Separation**:
    - While some overlap exists between the blue and orange points, indicating that PCA alone does not completely separate the two classes, clusters of orange points to the right suggest that exited customers possess distinct characteristics. This implies that additional methods or more dimensions may be necessary for better classification.
- **Variance Explained**:
    - The variance captured by **Principal Component 1** and **Principal Component 2** can be assessed using pca.explained_variance_ratio_.
    - Specifically, **Principal Component 1** accounts for approximately 21.85% of the variance, and **Principal Component 2** captures about 16.89%. Together, they explain roughly 38.75% of the total variance in the dataset. While this indicates a moderate retention of the original

variability, more components may be required to capture additional variance effectively.

Insight:

The PCA plot effectively projects multi-dimensional data into 2D space, revealing clusters and relationships between features and the target variable 'Exited'. However, the observed overlap suggests that a simple PCA may not sufficiently distinguish between exited and non-exited customers. Using additional components or more sophisticated techniques could enhance the separation.

Multivariate Analysis Limitations

- **Dimensionality Challenges**: In PCA, while reducing dimensions helps visualize data, it can also lead to the loss of important information. The results may not fully capture the complexity of customer behavior across all original variables.
- **Class Overlap**: The PCA plot reveals some overlap between exited and non-exited customers, indicating that the analysis does not completely separate the two classes. This overlap suggests that additional techniques or variables may be needed for more accurate classification.
- **Interpretation Difficulty**: Multivariate techniques like PCA can make interpretation complex. While they reveal clusters and relationships, understanding the practical implications of these components can be challenging.
- **Assumptions of Independence**: PCA assumes that the principal components are uncorrelated, which might not hold true in real-world data. If the underlying relationships among variables are not adequately addressed, the results may be misleading.
- **Data Quality and Representativeness**: The accuracy of multivariate analyses relies heavily on the quality and completeness of the dataset. Any missing or inaccurate data can significantly impact the findings and their generalizability to the broader population.

5. EDA Findings and Overall Analysis

5.1 Overview of EDA Findings

The exploratory analysis of customer data yielded several important insights regarding factors associated with customer churn:

- **Demographic Correlations**: A moderate positive correlation (0.29) was found between Age and Exited status, indicating that older customers are more likely to leave the bank.
- **Balance Patterns**: Customers with higher balances were slightly more likely to exit, though the correlation was weak (0.12). Additionally, a negative correlation of -0.30 between the number of products and balance suggests that customers with fewer products tend to have higher balances.
- **Weak Relationships**: Variables like Credit Score and Estimated Salary exhibited little to no significant correlation with churn, indicating they play a limited role in predicting customer exit behavior.
- **Visual Insights**: Scatter and pair plots revealed that customers who exited shared certain characteristics, particularly regarding Age and Balance.

5.2 Enhanced Insights

The analysis uncovered several additional patterns worth noting:

- **Demographic Insights**: A significant proportion of customers who exited had lower credit scores and balances, suggesting that financial stability may play a crucial role in customer retention.
- **Geographical Trends**: Customers from specific regions exhibited higher churn rates, indicating that geographical factors may influence customer satisfaction and retention.
- **Behavioral Patterns**: The data showed that active members with multiple products were less likely to churn, highlighting the importance of customer engagement and product diversification.

5.3 Strategic Implications

The findings suggest several key implications for developing effective retention strategies:

- **Target Older Customers**: The correlation between age and churn suggests that banks should consider crafting marketing strategies aimed at retaining older customers.
- **Retention Focus on Low-Balance Customers**: Customers with lower balances appear to be at greater risk of leaving. Offering them personalized financial guidance or incentives could improve retention rates.

- **Encourage Product Diversification**: Given the negative correlation between the number of products and balance, encouraging customers to diversify their banking products may enhance their engagement and loyalty.
- **Non-impactful Factors**: Since Credit Score and Estimated Salary showed little relationship with churn, these variables may not be useful for segmentation or targeted retention strategies.

5.4 Limitations and Considerations

While this analysis provides valuable insights, several limitations must be taken into account:

- **Data Quality**: The results rely on the quality and completeness of the dataset. Missing or inaccurate data could affect the findings.
- **Correlation vs. Causation**: While correlations were identified, they do not confirm causality. More in-depth studies are necessary to establish clear cause-and-effect relationships.
- **Time Factor**: The analysis does not account for changes over time, such as fluctuations in customer behavior or external factors like economic conditions, which might impact churn rates.
- **Limited Scope**: The focus on quantitative analysis may have overlooked qualitative aspects, such as customer satisfaction and experience, which are also critical for understanding churn.

References

1. **DataCamp**. "An Introduction to Exploratory Data Analysis." Available at: DataCamp
2. **Seaborn Documentation**. "Seaborn Overview." Available at: Seaborn
3. **Matplotlib Documentation**. "Matplotlib User Guide." Available at: Matplotlib User Guide
4. **Medium**. "A Comprehensive Guide to Exploratory Data Analysis (EDA)." Available at: Medium
5. **Kaggle**. "Customer Churn Analysis with Python." Available at: Kaggle
6. **Towards Data Science**. "Exploratory Data Analysis (EDA) Visualization Using Pandas." Available at: Towards Data Science
7. **Kaggle**. "Bank Customer Churn Prediction Dataset." Available at: Kaggle Dataset