Exploratory Data Analysis (EDA) Report: Botswana Bank Customer Churn

1. Introduction

This report summarizes the Exploratory Data Analysis (EDA) performed on the Botswana Bank customer churn dataset (`botswana_bank_customer_churn.csv`). The primary objective of this EDA is to understand the data's structure, identify key patterns and relationships, uncover insights related to customer churn, and document the preprocessing steps undertaken to prepare the data for potential machine learning modeling. The analysis is based on the procedures and visualizations found in the `milestone1.ipynb` notebook.

2. Data Loading and Initial Overview

The dataset was loaded using pandas. Initial exploration involved examining its basic characteristics:

- **Dataset Source:** `botswana_bank_customer_churn.csv`
- **Dimensions:** The original dataset contains 115,640 rows and 25 columns.
- **Data Types:** Columns include a mix of integer, float, object (string), and datetime types. Specifically, `Churn Date` and `Date of Birth` were parsed as datetime objects.
- **Initial Data Glimpse:** The first few rows were inspected using `df.head()` to understand the column names and sample data.
- **Missing Values:** An initial check using `df.isnull().sum()` revealed significant missing values primarily in `Churn Reason` and `Churn Date` (101,546 missing values each). These columns are directly related to churn events; missing values here indicate customers who have not churned.
- **Summary Statistics:** `df.info()` provided a summary of non-null counts and data types, and `df.describe()` offered statistical summaries for numerical columns.

3. Data Cleaning and Preprocessing Decisions

Several data cleaning and preprocessing steps were performed based on the initial review:

1. **Dropping Irrelevant Columns:** Columns deemed unnecessary for churn prediction were removed. These included `RowNumber`, `CustomerId`, `Surname`, `First Name`, `Contact Information`, and `Address`.

2. **Feature Engineering - Age:** A new feature, `Age`, was calculated by subtracting the year of `Date of Birth` from the current year. The `Date of Birth` column was then dropped.

3. **Handling Categorical Features:**

   - `Gender`: Mapped to numerical values (Male: 1, Female: 0). Missing values in `Gender` were filled with -1, which would be treated as a distinct category if One-Hot Encoding were applied later.

4. **Handling Missing Values in Churn-Related Columns:**

   - `Churn Reason`: Missing values were filled with the string 'Not Churned', explicitly indicating customers who have not churned.
   - `Churn Date`: Missing values (NaT) were left as is, as they signify non-churned customers. This column is typically not used as a direct feature in modeling if `Churn Flag` is the target.

5. **Outlier Treatment (Capping):** For key numerical columns (`Credit Score`, `Income`, `Balance`, `Outstanding Loans`, `NumOfProducts`, `NumComplaints`, `Age`), outliers were capped. Values below Q1 - 1.5IQR or above Q3 + 1.5IQR were replaced with the respective boundary values (Q1 - 1.5IQR or Q3 + 1.5IQR). This was done to reduce the skewness and impact of extreme values on potential models.

6. **Duplicate Rows:** The notebook includes a function `clean_data` which checks for and removes duplicate rows. The output indicated no duplicate rows were found in `df_clean` after initial processing.

7. **Constant Columns:** The `clean_data` function also checks for and removes columns with only one unique value (constant columns), though none were explicitly reported as removed in the provided code execution for `df_clean`.

4. Exploratory Data Analysis & Visualizations

The EDA involved generating various visualizations to understand feature distributions and relationships:

- **Target Variable Distribution (`Churn Flag`):**
    - A pie chart showed the percentage of churned vs. non-churned customers. The dataset is imbalanced, with a significantly larger proportion of non-churned customers (approximately 87.8% non-churned, 12.2% churned based on the modeling script's subsample, which should be similar for the full dataset).
    - A countplot also visualized this imbalance.
- **Numerical Feature Distributions:**
    - **Histograms and Boxplots:** Generated for `Income` , `Balance` , `Credit Score` , `Outstanding Loans` , `NumOfProducts` , `NumComplaints` , and `Age` . These plots helped understand the central tendency, spread, and presence of outliers (before and after capping for some).
        - `Income` and `Balance` distributions were explored in detail.
    - **Skewness and Kurtosis:** Calculated for numerical columns to quantify their distribution shapes.
- **Categorical Feature Distributions:**
    - **Countplots:** Used for `Gender` , `Marital Status` , and other categorical features like `Occupation` , `Education Level` , `Customer Segment` , and `Preferred Communication Channel` to show the frequency of each category.
- **Relationship with Target Variable (`Churn Flag`):**
    - **Boxplots:** `Income` and `Customer Tenure` distributions were compared across churned and non-churned customers.
    - **Histograms:** The distribution of `Balance` was plotted separately for churned and non-churned customers, showing potential differences.
- **Correlation Analysis:**
    - **Correlation Matrix & Heatmap:** A heatmap of the correlation matrix for numerical features was generated. This helped identify linear relationships between numerical variables and with the `Churn Flag` .
        - The correlation of each numerical feature with `Churn Flag` was explicitly printed.

5. Key Findings and Insights (from `milestone1.ipynb` EDA sections)

While the notebook focuses heavily on the process of EDA (generating many plots), specific conclusive insights derived from these plots would require detailed interpretation of each. However, the general findings from the EDA steps are:

1. **Data Quality:** The dataset has some missing values, primarily in churn-related columns, which were handled appropriately. Outliers were present in several numerical features and were addressed through capping.
2. **Target Imbalance:** There's a significant class imbalance, with many more non-churners than churners. This is a critical consideration for model training and evaluation if the goal is to predict churn accurately.
3. **Feature Characteristics:** The distributions of numerical features like `Income`, `Balance`, and `Age` were visualized, providing an understanding of their spread and central tendencies. Categorical features showed varying numbers of unique values, with `Occupation` likely having many distinct categories.
4. **Potential Churn Indicators (Visual Inspection):**
    - Visual comparisons (e.g., `Balance` distribution for churned vs. non-churned) suggest some features might have different patterns for the two groups, indicating their potential predictive power. For instance, the distribution of `Balance` appeared different for churned customers.
    - The correlation analysis provided initial clues about linear relationships between numerical features and churn.

6. Preprocessing Decisions Summary for Modeling

The EDA and preprocessing steps resulted in a cleaned dataset (`df_clean` in the notebook's context, later saved as `df_clean_processed.csv` by the modeling preprocessing script). Key decisions that prepare the data for machine learning include:

- Removal of identifiers and high-cardinality, non-predictive text fields.
- Creation of an `Age` feature.
- Numerical transformation of `Gender`.
- Systematic handling of missing data, especially for `Churn Reason`.
- Outlier mitigation through capping.
- The notebook also demonstrates One-Hot Encoding for categorical variables to convert them into a machine-learning-friendly format, resulting in the `final_data` DataFrame (saved as `final_data_processed.csv`).

This EDA process provides a foundational understanding of the dataset and justifies the preprocessing choices made, setting the stage for effective model building.

7. Cleaned Dataset

The cleaned dataset, reflecting the preprocessing steps described (prior to one-hot encoding but after dropping columns, feature engineering, and outlier capping), is represented by the `df_clean` DataFrame in the notebook. This dataset was saved as `/home/ubuntu/processed_data/df_clean_processed.csv` during the previous modeling task. This file will be provided as the "Cleaned Dataset" deliverable for this EDA task.