Exploratory Data Analysis (EDA) Summary Report

1. Introduction

This report summarizes the exploratory data analysis performed on the delinquency prediction dataset. The primary goal of this analysis was to understand the dataset's structure, identify key patterns, detect anomalies and missing values, and pinpoint potential risk indicators for predicting customer delinquency.

2. Dataset Overview

This section summarizes the dataset, including the number of records, key variables, and data types. It also highlights any anomalies or inconsistencies observed during the initial review.

- **Number of records:** 500
- **Key variables:**
    - `Customer_ID`: Unique identifier for each customer.
    - `Age`: Customer's age.
    - `Income`: Customer's annual income.
    - `Credit_Score`: Customer's credit score.
    - `Credit_Utilization`: Ratio of credit used to available credit.
    - `Missed_Payments`: Number of missed payments.
    - `Delinquent_Account`: Target variable (0 = Not Delinquent, 1 = Delinquent).
    - `Loan_Balance`: Outstanding loan balance.
    - `Debt_to_Income_Ratio`: Ratio of debt to income.
    - `Employment_Status`: Customer's employment status.
    - `Account_Tenure`: Duration of the account.
    - `Credit_Card_Type`: Type of credit card held.
    - `Location`: Customer's geographic location.
    - `Month_1` - `Month_6`: Payment status for the past six months ('On-time', 'Late', 'Missed').
- **Data types:** The dataset contains a mix of numerical (float64, int64) and categorical/object data types.
- **Anomalies, duplicates, or inconsistencies observed during the initial review:**
    - **Class Imbalance:** The `Delinquent_Account` target variable is highly imbalanced, with 84% non-delinquent accounts and 16% delinquent accounts.
    - **Inconsistent Categorical Entries:** The `Employment_Status` column had inconsistent capitalization and variations (e.g., 'Employed', 'employed', 'EMP', 'retired'), which required standardization.

    - **Potential Outliers in Credit_Utilization:** A few `Credit_Utilization` values were slightly above 1.0, which might indicate extreme utilization or data entry anomalies.
    - **Skewed Distributions:** `Income`, `Loan_Balance`, and `Credit_Utilization` showed right-skewed distributions.

3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring model accuracy. This section outlines missing values in the dataset, the approach taken to handle them, and justifications for the chosen method.

- **Variables with missing values:**
    - `Income`: 7.8% missing.
    - `Loan_Balance`: 5.8% missing.
    - `Credit_Score`: 0.4% missing.
- **Missing data treatment:** Imputation was chosen as the treatment method to retain as much data as possible, given the moderate percentage of missing values.
    - **`Credit_Score`:** Imputed with the **median** due to the very small percentage of missing values, which minimizes impact and robustness to potential outliers.
    - **`Income` and `Loan_Balance`:** Imputed with the **median** values. While more advanced methods like regression or KNN imputation were considered, median imputation was chosen for its simplicity and effectiveness given the dataset size and initial exploratory phase, particularly for the skewed distributions.

4. Key Findings and Risk Indicators

This section identifies trends and patterns that may indicate risk factors for delinquency. Feature relationships and statistical correlations are explored to uncover insights relevant to predictive modeling.

- **Key findings (Patterns and Anomalies):**
    - The `Delinquent_Account` distribution highlights a significant class imbalance, which is a major challenge for predictive modeling.
    - Numerical features like `Income`, `Loan_Balance`, and `Credit_Utilization` are skewed, suggesting that transformations might be beneficial for certain models.
    - The `Employment_Status` column required standardization due to inconsistent entries, revealing common employment categories: 'Unemployed', 'Retired', 'Employed', and 'Self-employed'.
    - Payment statuses in `Month_1` through `Month_6` show a mix of 'On-time', 'Missed', and 'Late' payments, providing a historical sequence of behavior crucial for trend analysis.

- **Correlations observed between key variables:** While explicit correlation calculations were not detailed, it is expected that `Credit_Score` would have a strong inverse correlation with `Delinquent_Account`, and `Missed_Payments`, `Credit_Utilization`, and `Debt_to_Income_Ratio` would show positive correlations with delinquency.
- **Unexpected anomalies:** No highly unexpected or severe anomalies were found that would suggest significant data corruption, beyond the minor `Credit_Utilization` values slightly above 1.0.

5. AI & GenAI Usage

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns. This section documents AI-generated insights and the prompts used to obtain results.

- **Example AI prompts used:**
  - "Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency."
  - "Document your findings in bullet points for your report. Focus on: Notable missing or inconsistent data, Key anomalies, Early indicators of delinquency risk. Then, write a short paragraph (3–5 sentences) summarizing your initial data quality observations."
  - "Suggest an imputation strategy for missing values in this dataset based on industry best practices."
  - "Implement and provide a cleaned dataset."
  - "Can you provide it in excel format?"
  - "Provide all the steps as of now form the beginning to last in a ipynb file."
  - "List high-risk indicators, each with a one-sentence explanation of why it's important, as well as any insights that could impact delinquency prediction."

6. Conclusion & Next Steps

This initial EDA provided valuable insights into the dataset's quality and characteristics. The key challenges identified are the missing values, the class imbalance in the target variable, and the need for feature engineering from categorical and time-series-like variables.

Recommended Next Steps:

- **Address Class Imbalance:** Apply techniques such as oversampling (e.g., SMOTE), undersampling, or using algorithms robust to imbalance to ensure the model can effectively identify delinquent accounts.
- **Feature Engineering:**
  - Derive new features from `Month_1` to `Month_6` (e.g., total late payments, consecutive missed payments, payment trends).

    - Encode categorical variables like `Credit_Card_Type` and `Location` using appropriate methods (e.g., one-hot encoding).
- **Model Selection and Training:** Explore various machine learning models suitable for binary classification, evaluating their performance using appropriate metrics for imbalanced datasets (e.g., Precision, Recall, F1-score, AUC-ROC).
- **Model Evaluation:** Rigorously evaluate the chosen model's performance on unseen data to ensure its generalization capability.