Exploratory Data Analysis (EDA) Summary Report

1. Introduction

This report details the initial Exploratory Data Analysis (EDA) performed on the Delinquency_prediction_dataset.xlsx file. The primary purpose of this analysis is to gain a foundational understanding of the dataset's structure, identify data quality issues such as missing values, and uncover preliminary patterns and potential risk indicators related to loan delinquency. The ultimate goal is to prepare the dataset for subsequent predictive modeling to forecast 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, duplicates, or inconsistencies observed during the initial review.

Key dataset attributes:

Number of records: An exact count requires computational analysis, but based on the provided sample, the dataset contains numerous records, likely several hundred or more, each representing a customer profile.

Key variables: The dataset includes the following key variables, crucial for analyzing delinquency risk:

- Customer ID: Unique identifier for each customer.
- Age: Age of the customer.
- Income: Customer's reported income.
- Credit Score: Numerical representation of creditworthiness.
- Credit Utilization: Percentage of available credit being used.
- Missed Payments: Number of missed payments.
- Delinquent Account: Binary indicator (Yes/No) if the account is delinquent.
- Loan Balance: Outstanding loan amount.
- Debt to Income Ratio: Ratio of total debt to total income.
- Employment Status: Current employment status (e.g., Employed, Unemployed, Retired).

- Account Tenure: Duration of the customer's account.
- Credit Card Type: Type of credit card held.
- Location: Geographical location of the customer.
- Month_1 to Month_6: Payment status for the last six months (e.g., On-time, Late, Missed).

Data types:

- **Numerical:** Age, Income, Credit Score, Credit Utilization, Missed Payments, Loan Balance, Debt to Income Ratio, Account Tenure.
- **Categorical:** Delinquent Account, Employment Status, Credit Card Type, Location, Month_1 to Month_6.
- Customer ID serves as a unique identifier.

Anomalies, duplicates, or inconsistencies: Without computational tools for a full scan, precise anomalies, duplicates, or inconsistencies cannot be definitively listed. However, common issues in such datasets might include:

- Outliers in numerical features (e.g., extremely high income or credit utilization).
- Inconsistent data entry (e.g., variations in Employment Status labels).
- Potential duplicate Customer ID entries, which would require verification.
- Missing or incorrect values in critical fields, which are addressed in the next section.

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: Upon initial review of the dataset content, missing values were observed in key financial features such as:

- Loan Balance (e.g., for CUST0009, CUST0024)
- Income (e.g., for CUST0041, CUST0043)
- Debt to Income Ratio (e.g., for CUST0035, CUST0052)

Missing data treatment:

Missing Data Issue	Handling Method	Justification
Loan Balance	Impute with Median	Loan balance is a continuous numerical variable that can be skewed; the median is robust to outliers and represents a typical value.
Income	Impute with Median	Income is a continuous numerical variable often with skewed distributions; the median is a robust measure for central tendency in such cases.
Debt to Income Ratio	Impute with Predictive Model (e.g., Regression)	This ratio is critical for credit risk assessment; a predictive model can estimate missing values more accurately by leveraging other correlated features.

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:

- **Correlations observed between key variables:** While precise statistical correlations require computation, generally anticipated relationships based on industry best practices include:

  Inverse Correlation: Credit Score, Income, and Account Tenure are expected to have an inverse correlation with Delinquent Account (i.e., higher values in these features typically mean lower delinquency risk).

Positive Correlation: Credit Utilization, Missed Payments, Loan Balance, and Debt to Income Ratio are expected to have a positive correlation with Delinquent Account (i.e., higher values generally indicate higher delinquency risk).

The Month_1 to Month_6 payment statuses are directly indicative of Delinquent Account.

- **Unexpected anomalies:** Identifying unexpected anomalies precisely would require a full scan for outliers and unusual data points (e.g., a customer with a very high Credit Score but consistently Missed Payments, or an extremely low Income with a very high Loan Balance but no delinquency). Such anomalies would warrant further investigation by the analytics team.

5. AI & GenAI Usage

Generative AI tools can be instrumental in enhancing various stages of EDA. In a typical workflow, such tools would be used to:

- **Summarize key patterns in the dataset and identify anomalies:** GenAI can analyze feature distributions, relationships, and potential outliers, then generate human-readable summaries and highlight unusual observations.
- **Suggest an imputation strategy for missing income values based on industry best practices:** By understanding the data characteristics and context (e.g., financial data), GenAI can propose sophisticated imputation methods beyond simple mean/median, such as regression imputation or using K-Nearest Neighbors (KNN).
- **Generate realistic synthetic income values for missing entries using normal distribution assumptions:** For privacy-sensitive data or to augment datasets, GenAI can create synthetic data points that mimic the statistical properties and distribution patterns of real data. *It's important to note that while these are capabilities of GenAI, the specific analyses and imputations described above were conceptually outlined based on the dataset structure and common practices, rather than being actively performed by me as an AI tool during this interaction.*

6. Conclusion & Next Steps

This initial EDA provides a foundational understanding of the dataset's structure, potential data quality issues, and preliminary insights into factors influencing delinquency. The presence of missing values in critical financial metrics like Loan Balance, Income, and Debt to Income Ratio highlights the immediate need for robust data cleaning.

Key Findings: The dataset contains a comprehensive set of variables for credit risk assessment, including financial, behavioral, and demographic information. The Delinquent Account variable is the key target.

Recommended Next Steps:

- **Data Cleaning and Preprocessing:**

  Implement the proposed imputation strategies for missing Loan Balance, Income, and Debt to Income Ratio.

  Handle any other identified inconsistencies or outliers.

  Address categorical variable encoding (e.g., One-Hot Encoding for Employment Status, Credit Card Type, Location, and Month X statuses).

- **Detailed Statistical Analysis:**

  Perform comprehensive statistical tests to confirm correlations and significant relationships between features and Delinquent Account.

  Analyze the distribution of all variables and identify potential skewness or kurtosis.

- **Feature Engineering:**

  Create new features that could improve model performance (e.g., a "Payment Consistency Score" from Month_1 to Month_6, or age groups).

- **Model Selection and Training:**

  Based on EDA insights, select appropriate machine learning models for delinquency prediction (e.g., Logistic Regression, Decision Trees, Gradient Boosting).

  Split the dataset into training and testing sets.

- **Model Evaluation and Validation:**

  Rigorously evaluate model performance using relevant metrics (e.g., Accuracy, Precision, Recall, F1-Score, ROC-AUC) and cross-validation techniques.

  Interpret model results to understand feature importance and contribution to delinquency prediction.