Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

This report provides a summary of the exploratory data analysis (EDA) performed on the provided delinquency prediction dataset. The primary goal of this analysis is to understand the structure of the data, identify potential issues such as missing or inconsistent values, and uncover early indicators of delinquency risk to inform the development of a predictive model.

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: 500

- Key variables : Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Delinquent_Account, Loan_Balance, Debt_to_Income_Ratio, Employment_Status, Account_Tenure, Credit_Card_Type, Location, and the monthly payment history columns (Month_1 to Month_6).

- Data types:The dataset contains a mix of numerical (int64 and float64) and categorical (object) data types.

- Anomalies and inconsistencies: The Employment_Status column contains inconsistent casing and abbreviations (e.g., Employed, employed, EMP) which will require standardization.

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.

Key missing data findings:

Variables with missing values: Income (7.8%), Loan_Balance (5.8%), and Credit_Score (0.4%).

Missing data treatment: For these variables, an imputation strategy is recommended. For Credit_Score, a simple imputation with the mean or median would be sufficient due to the low percentage of missing values. For Income and Loan_Balance, a more advanced imputation method, such as a predictive model (e.g., K-Nearest Neighbors or a regression model), might be more appropriate given the higher percentage of missing data.

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: Customers with a lower Credit_Score tend to have a higher number of Missed_Payments and a higher Debt_to_Income_Ratio, which are strong indicators of financial risk.

-Unexpected anomalies: One notable anomaly is a customer with a Credit_Utilization of over 1.0, indicating they have exceeded their credit limit. The lowest Credit_Score of 301 also represents an extremely high-risk profile.

-Early indicators: The monthly payment history columns (Month_1 to Month_6) show a high frequency of Late and Missed payments. These serve as direct and immediate indicators of a customer's payment reliability and are likely the strongest predictors of a delinquent account status.

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 in the dataset and identify anomalies. Suggest an imputation strategy for missing income values based on industry best practices.

-Analyze the dataset and provide a summary of key columns including common values and missing values.

6. Conclusion & Next Steps

This initial EDA highlights several key insights and potential areas for improvement. The dataset contains valuable features for predicting delinquency, but also requires cleaning due to missing values and inconsistent data. The presence of clear risk indicators, such as low Credit_Scores and frequent Missed_Payments, provides a solid foundation for model development. The next steps should include data cleaning and preprocessing to handle the identified issues, followed by feature engineering and the construction of a predictive model.