Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The purpose of this report is to perform an initial exploratory data analysis on the delinquency prediction dataset. The primary goal is to understand the dataset's structure, identify data quality issues, and uncover key patterns and risk indicators related to loan delinquency.

2. Dataset Overview

This dataset contains customer financial information to predict the likelihood of a delinquent account. It includes a mix of numerical and categorical data.

Key dataset attributes:

- **Number of records:** The dataset contains 500 records.
- **Key variables:**
  - Age: The customer's age.
  - Income: The customer's annual income.
  - Credit_Score: The customer's credit score.
  - Credit_Utilization: The percentage of available credit a customer is using.
  - Missed_Payments: The number of times a customer has missed a payment.
  - Delinquent_Account: The target variable, indicating if an account is delinquent (1) or not (0).
  - Loan_Balance: The outstanding balance on a loan.
  - Debt_to_Income_Ratio: The ratio of a customer's monthly debt to their gross monthly income.
  - Employment_Status: The customer's employment status.
  - Account_Tenure: The length of time an account has been active.
  - Credit_Card_Type: The type of credit card held.
  - Location: The city where the customer resides.
  - Month_1 through Month_6: Payment status for the past six months.
- **Data types:** The dataset contains a mix of numerical (e.g., Age, Income, Credit_Score) and categorical (e.g., Employment_Status, Location, Credit_Card_Type) data.

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: Missing values for customers CUST0041, CUST0043, CUST0060, CUST0067, CUST0069, CUST0077, CUST0094.
  - Loan_Balance: Missing values for customers CUST0009, CUST0024, CUST0026, CUST0029, CUST0052, CUST0053, CUST0103, CUST0105.
- **Missing data treatment:** Given the small number of missing values and the numerical nature of the affected variables, a simple imputation strategy is recommended. Missing Income and Loan_Balance values can be imputed using the mean or median of their respective columns. This approach preserves the data and avoids introducing bias. For instance, the prompt 'Suggest a robust imputation method for numerical columns in a financial dataset, considering potential outliers.' could be used with a generative AI tool to determine the best method (mean vs. median).

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:**
  - There is a strong correlation between the number of Missed_Payments and a Delinquent_Account. Customers with a higher number of missed payments are significantly more likely to be delinquent.
  - Credit score appears to be inversely correlated with delinquency; a lower
  - Credit_Score is associated with a higher risk of having a Delinquent_Account.
  - A higher Credit_Utilization is also a potential risk indicator, as it appears to be positively correlated with the number of missed payments and delinquent accounts.
  - Customers with an Unemployed status seem to have a higher rate of delinquency compared to other employment statuses.
- **Unexpected anomalies:**
  - Some customers with a Credit_Score below 400 are not marked as delinquent, which could be an anomaly requiring further investigation.
  - The values for Credit_Utilization of 0.05 are a data entry artifact. This may indicate a specific account type or a data error, which should be investigated further before being used in a model.

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 related to delinquency and identify anomalies or unexpected data points.'
- 'Suggest an imputation strategy for missing income values based on industry best practices.'

6. Conclusion & Next Steps

The EDA revealed several strong indicators for delinquency, including the number of missed payments, credit score, and credit utilization. The data is relatively clean, but some missing values and potential anomalies were identified.

Next steps:

1. **Data Cleaning:** Implement a chosen imputation strategy to handle the missing values in the Income and Loan_Balance columns.
2. **Outlier Analysis:** Further investigate the identified anomalies, such as low credit scores without delinquency, to determine if they are valid data points or errors.
3. **Feature Engineering:** Consider creating new features from existing data, such as a payment streak score from the Month_1 to Month_6 columns, to provide more robust risk indicators for the predictive model.
4. **Model Building:** Proceed to a pre-processing and model-building phase to predict customer delinquency based on the cleaned and engineered features.