Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The purpose of this report is to conduct an exploratory data analysis (EDA) on Geldium's customer dataset. The goal is to assess data quality, address missing data, detect patterns and risk factors, and prepare the dataset for building an improved delinquency risk 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

Customer_ID, Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Delinquent_Account, Loan_Balance, Debt_to_Income_Ratio, Employment_Status, Account_Tenure, Credit_Card_Type, Location, Month_1 to Month_6 (payment history).

Numerical: Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure; Categorical: Employment_Status, Credit_Card_Type, Location, Month_1 to Month_6; Target: Delinquent_Account (binary).

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:

Income (39 missing), Credit_Score (2 missing), Loan_Balance (29 missing).

All missing values were handled using median imputation to ensure robustness against outliers while preserving central tendency.

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:

Employment status, credit card type, and location show stronger associations with delinquency rates than individual numeric features. Numerical variables like Income, Credit_Score, and Debt_to_Income_Ratio exhibit weak correlations with delinquency.

The lack of correlation between payment history and delinquency may suggest delayed delinquency, data window misalignment, or interaction effects that warrant deeper investigation.

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.
- Identify the top 3 variables most likely to predict delinquency based on this dataset. Provide brief reasoning.
- Suggest an imputation strategy for missing values in this dataset based on industry best practices.
- Detect patterns and feature relationships that could influence delinquency risk.

- 'Summarize key patterns in the dataset and identify anomalies.'

- 'Suggest an imputation strategy for missing income values based on industry best practices.'

6. Conclusion & Next Steps

The dataset is largely complete but required targeted imputation for Income, Credit_Score, and Loan_Balance. Employment_Status, Credit_Card_Type, and Location appear most predictive. Further feature engineering, especially on payment history and interaction effects, is recommended to enhance predictive model performance.