Exploratory Data Analysis (EDA) Summary Report

1. Introduction

This report presents an exploratory data analysis (EDA) of Geldium's customer dataset to support the development of a delinquency risk prediction model by Tata iQ. The goal is to assess data quality, identify missing values and anomalies, and uncover early risk indicators that influence delinquency.

2. Dataset Overview

The dataset contains financial, behavioral, and demographic attributes of customers used for assessing delinquency risk.

Key dataset attributes:

- Number of records: 500

- Key variables: 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

- Data types: Mixture of Numerical and Categorical variables

Potential anomalies include out-of-range values in Credit_Score, Age, Credit_Utilization, and Debt_to_Income_Ratio, as well as possible duplicate Customer_ID records.

3. Missing Data Analysis

Missing values can significantly affect model accuracy and were addressed using industry best practices.

Key missing data findings:

- Variables with missing values: Income, Employment_Status, Monthly Payment History (Month_1 to Month_6)

- Missing data treatment:

  - Income: Median imputation grouped by Employment_Status
  - Employment_Status: Mode imputation with an additional 'Unknown' category
  - Monthly Payment History: Treated as a distinct 'Missing' category with indicator flags

These approaches preserve data distribution, reduce bias, and retain the predictive importance of missingness.

4. Key Findings and Risk Indicators

Key correlations and risk patterns identified include:

- High Missed_Payments strongly associated with increased delinquency risk.

- Recent late or missed payments (Month_1–Month_6) indicate near-term default likelihood.

- High Credit_Utilization (>70–80%) reflects financial overextension.

- High Debt_to_Income_Ratio reduces repayment flexibility.

- Low Credit_Score is consistently linked to higher default probability.

Unexpected anomalies include customers with high Loan_Balance but very low Income, and cases of multiple missed payments where Delinquent_Account is still marked as 0.

5. AI & GenAI Usage

Generative AI tools were used to structure the exploratory analysis, summarize patterns, suggest imputation strategies, and identify risk indicators from the dataset description.

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.'

6. Conclusion & Next Steps

The dataset contains strong behavioral and financial predictors of delinquency but also exhibits potential data quality risks related to missing values, outliers, and inconsistencies. After data cleaning and validation, key behavioral and burden metrics should be prioritized in model development. Next steps include full statistical correlation analysis, feature engineering, and training of the delinquency prediction model.