Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this report is to perform exploratory data analysis (EDA) on Geldium's delinquency dataset to assess data quality, identify key risk indicators, and prepare the data for predictive modeling. This analysis supports Tata iQ's objective of improving delinquency risk prediction and intervention strategies.

2. Dataset Overview

The dataset contains customer financial and behavioral attributes used to analyze delinquency risk. It includes a mix of numerical and categorical variables related to income, credit usage, and payment behavior.

Key dataset attributes:

- Number of records: Provided in the dataset

- Key variables: Income, Credit Utilization, Payment History, Outstanding Balance, Employment Type, Delinquency Status

- Data types: Numerical (income, utilization), Categorical (employment type, loan purpose)

3. Missing Data Analysis

Several important variables contain missing values that could impact model accuracy if not handled properly.

Key missing data findings:

- Income: Missing values present; treated using median imputation due to skewness

- Credit Utilization: Missing and extreme values; handled using median or predictive imputation with capping

- Employment Type: Missing categorical values; filled using mode imputation

4. Key Findings and Risk Indicators

EDA revealed strong relationships between credit behavior variables and delinquency outcomes.

Key findings:

- High credit utilization is strongly associated with higher delinquency risk

- Customers with previous late payments are significantly more likely to default

- Low income combined with high outstanding debt increases delinquency probability

- Some anomalies (e.g., high income but delinquent accounts) require further investigation

5. AI & GenAI Usage

Generative AI tools were used to summarize dataset patterns, suggest imputation strategies, and help identify key risk indicators efficiently.

Example AI prompts used:

- Summarize key patterns and anomalies in the dataset

- Suggest imputation strategies for missing income and credit utilization values

- Identify variables most likely to predict delinquency

6. Conclusion & Next Steps

The dataset shows clear indicators of delinquency risk driven primarily by credit behavior variables. Addressing missing values, anomalies, and class imbalance will improve model reliability. Next steps include feature engineering, model development, and validation of delinquency prediction models.