Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

This report presents an Exploratory Data Analysis (EDA) of Geldium's customer dataset, conducted to support Tata iQ's analytics team in refining their delinquency risk prediction model. The primary goal of this analysis is to assess the quality, structure, and completeness of the data and to uncover early indicators that can predict the likelihood of customer delinquency. By identifying key patterns, missing data, and high-risk attributes, this EDA lays the foundation for accurate and fair predictive modeling, enabling Geldium to strengthen its financial decision-making and customer intervention strategies

2.Dataset Overview

Key dataset attributes :

Number of records: 500

Key variables: Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio, Employment_Status, Delinquent_Account, etc.

Data types:

Numerical: Age, Income, Credit_Score, Credit_Utilization, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure

Categorical: Employment_Status, Credit_Card_Type, Location, Month_1 to Month_6

Binary: Delinquent_Account

Anomalies & Issues:

Income has 39 missing values

Credit_Score has 2 missing values

Loan_Balance has 29 missing values

Payment history (Month_1 to Month_6) is in text form (Late, On-time, Missed) and will need to be converted for modelling

3.Missing Data Analysis

- Income had 39 missing values. We handled this by using median imputation, as income data is typically skewed and the median provides a more robust central tendency that reduces the impact of extreme values.
- Credit_Score had 2 missing values. We chose mean imputation because the number of missing entries was small and the data distribution appeared approximately normal, making the mean a reasonable choice.
- Loan_Balance had 29 missing values. This was treated using median imputation since loan values can vary significantly, and the median helps avoid distortion from high outliers.

4. Key Findings and Risk Indicator

Key findings:

- Missed_Payments is highly associated with Delinquent_Account = 1,it’s a strong predictor
- Credit_Utilization close to or above 0.35 appears risky in correlation with delinquencies
- Users with lower Income and higher Debt_to_Income_Ratio are more likely to be delinquent
- Unemployed users show a higher number of missed months in payment history
- The Monthly Payment History columns (Month_1 to Month_6) reveal that repeated Late or Missed payments can signal increasing risk

5. AI & GenAI Usage

Prompts used:

- “Summarize key patterns, outliers, and missing values in this dataset.”
- “Suggest an imputation strategy for missing values in this dataset based on industry best practices.”
- “Identify the top 3 variables most likely to predict delinquency based on this dataset.”

How GenAI helped:

- Generated structured summaries
- Helped choose best imputation techniques
- Highlighted key relationships for modeling

6. Conclusion & Next Steps

Conclusion:

The dataset provides a solid foundation for building a predictive delinquency model. After handling missing values through proper imputation strategies, several risk indicators—like missed payments, high credit utilization, and low income—emerged as strong signals of delinquency. The presence of structured monthly history also enables time-series or sequence modeling for payment behavior.

Next steps:

- Encode categorical columns like Month_1–Month_6 into numerical formats
- Perform feature scaling and correlation pruning
- Train and validate a classification model (e.g., Logistic Regression, Random Forest)
- Create a risk score or prediction tool for operational use