**Exploratory Data Analysis (EDA) Summary**   
**Report Template**

# 1. Introduction

The primary purpose of this report is to assess the structure, completeness and key attributes of the provided dataset while identifying any missing or inconsistent data that could affect predictions when identifying customers at-risk for delinquency.

The Goal of this report:

* Help Tata iQ’s analytics team and Geldium’s decision makers understand the current state of their data.

# 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

- Key variables:

* Missed\_Payments: Amount of missed payments up to 6, has a strong correlation to delinquency.
* Credit\_Utilization: Ratio of utilization for credit, with high utilization potentially indicating financial strain.
* Payment\_History: Consists of six individual columns denoting whether a month’s payment was On-time, Missed, or Late.. Recent Late/Missed payments are red flags.
* Delinquent\_Account: Flags a delinquent account with 1 marking delinquency and 0 no delinquency. Shows whether customer has a history of delinquency.
* Debt\_to\_Income\_Ratio: Shows how much debt a customer has compared to their income, with higher ratios suggesting repayment difficulty.
* Employment\_Status: Whether a customer is employed or not. Unemployed/self-employed individuals are at higher risk of delinquency.
* Credit\_Score: Measures the customer’s overall reliability on paying debts, with lower scores correlating with higher risk.

- Data types:

* Categorical: Employment\_Status, Credit\_Card\_Type
* Numerical: Income, Loan\_Balance

# 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:

- Missing Values: Income, Loan\_Balance and Credit\_Score have missing entries.

- Inconsistencies: Some entries in Employment\_Status and Credit\_Card\_Type are inconsistently labeled.

- Missing data treatment:

Income: Use Regression to impute by Age and Location

Employment\_Status & Credit\_Card\_Type: Impute by Mode.

Loan\_Balance: Conditional Imputation, where if Missed\_Payments and Delinquent\_Account are at 0, it’s set to $0.

# 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:

* Customers with higher Missed\_Payments are more likely to be delinquent.
* High Credit\_Utilization correlates with delinquency.
* Credit\_Score is negatively correlated with delinquency.
* Consecutive Missed/Late payments on the 6 months have a positive correlation with delinquencies.

- Unexpected anomalies:

* Extreme Credit\_Utilization may indicate errors or major outliers.
* Unusually high Credit\_Scores can be correct but merit investigation/capping.
* Inconsistencies in Employment\_Status call for data cleanup as they may cause distortion.

# 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:

- What key patterns are in the dataset? What anomalies exist? I have special interest in indicators of delinquency risk, what is most relevant? I have noticed missing values, can I get suggestions for an imputation strategy for these?

- Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency.

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

# 6. Conclusion & Next Steps

Summary of Key Findings:

* Payment behavior is strongly tied to delinquency.
* Customers with worse Financial Stability are more likely to be delinquent.
* Age group may impact delinquency, possibly owing to financial discipline or income depletion
* Geographical location may impact delinquency, likely due to local economic factors.
* Income and Credit\_Score are critical fields with some missing data.
* Values exceeding their caps, like Credit\_Utilization does, can distort analysis.

Next Steps:

* Flag customers exhibiting risky behavior for early intervention.
* Loan Approvals should gauge income stability as well as credit scores.
* Regional pricing could help mitigate risk associated with location.
* Clean up inconsistently labeled data.
* Validate the outliers In Credit\_Score and Utilization.
* Imputation to help fill in the missing values.