Forage TATA

# Exploratory Data Analysis (EDA) Summary Report

Prepared for: Tata iQ Analytics Team | Analyst: Soniya Bhatt |Dataset: `delinquency_data.csv`

## 1. Introduction

The purpose of this report is to analyze the Geldium delinquency dataset to assess data quality, uncover risk indicators, and prepare the dataset for predictive modeling. This analysis will support Tata iQ's analytics team and Geldium's decision-makers in refining their delinquency risk model and designing more effective intervention strategies.

## 2. Dataset Overview

This section provides an overview of the dataset, including its structure, types of variables, and initial observations regarding data quality.

**Key dataset attributes:**

- **Number of records:** 500
- **Key variables:**
  - Customer_ID – Unique identifier
  - Columns - Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio ,Employment_Status, Credit_Card_Type, Location,
  - Month_1 to Month_6 – monthly repayment behavior (values: On-time, Late, Missed)
  - **Delinquent_Account** – seems to be the **target column** (0 = not delinquent, 1 = delinquent)

- **Data types:**
    - **Categorical**: Customer_ID, Employment_Status, Credit_Card_Type, Location,Month_1, Month_2, Month_3, Month_4, Month_5,Month_6
    - **Numerical**:Age,Income,Credit_Score,Credit_Utilization,Missed_Payments,Delinquent_Account,Loan_Balance,Debt_to_Income_Ratio, Account_Tenure

## 3. Missing Data Analysis

Identifying and resolving missing data ensures the integrity of downstream modeling. Several variables contain missing or incomplete values that must be addressed.

**Key missing data findings:**

- **Variables with missing values:**
    - `Credit_Score:` 2 missing values
    - `Income`: 39 missing values
    - `Loan_Balance:` 29 missing values
    - `Employment_Status:` values are inconsistent
    - `Credit_Utilization >100%:`may indicate data entry errors or over-limit usage.
- **Missing data treatment:**
    - `Credit_Score`: simple imputation (median)
    - `Income`: Synthetic data generation was chosen to preserve sample size and maintain realistic value distributions aligned with industry norms.
    - `Loan_Balance`: Median imputation , if it shows normal distribution Mean imputation.
    - `Employment_Status:`Standardize values, Ensure consistent categories for classification.
    - `Credit_Utilization >100%`: Capping at 100% ,Logical upper bound for usage; anything above is risky.

# 4. Key Findings and Risk Indicators

## Key findings:

- **Correlations observed:**
  - High `Credit_Utilization` is strongly associated with `Delinquent` status.
  - Customers with missed or late `Payment_History` show increased default risk.
  - High `Debt_to_Income_Ratio` correlates with financial stress indicators.

- **Unexpected anomalies:**
  - Some entries report credit utilization >100%, which may reflect data entry errors or high-risk behaviors.
  - A few customers have `Income` values listed as zero, which should be reviewed.

- **Early risk indicators identified:**
  - Irregular payment behavior
  - Low or unstable income
  - Excessive recent credit applications (suggests financial desperation)
  - High Debt-to-Income (DTI) Ratio (Suggests financial overburden and difficulty in managing obligations)

---

# 5. AI & GenAI Usage

Generative AI tools such as ChatGPT were leveraged to automate parts of the analysis, explore missing data strategies, and summarize risk indicators.

**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."*
*"Generate synthetic income values based on normal distribution assumptions aligned with existing customer trends."*
*"List the top predictors of delinquency based on the dataset variables."*
AI insights supported faster exploration and cross-validation of assumptions.

## 6. Conclusion & Next Steps

This EDA revealed several data quality concerns, including missing values, inconsistencies, and anomalies. Key risk indicators such as high credit utilization, poor payment history, and high debt-to-income ratios were identified as critical features for delinquency prediction.

**Next Steps:**

- The dataset needs cleaning in fields like income, employment, and credit utilization.
- Imputation and synthetic data generation can fill gaps without compromising privacy.
- Key predictors like payment history, credit usage, and DTI must be prioritized in the model.
- Additional care must be taken to avoid bias while interpreting synthetic values.
- Suggest building a validation pipeline to monitor AI fairness and accuracy before deployment.