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

# 1. Introduction

# This report presents an exploratory data analysis (EDA) of Geldium Finance’s customer dataset to support the development of an AI-powered delinquency prediction model. The goal is to assess data quality, identify early risk indicators, and guide the modeling and intervention strategy for reducing credit card delinquency.

# 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: [5000 customer entries]

- Key variables: [ Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Delinquent\_Account

* + Loan\_Balance, Debt\_to\_Income\_Ratio, Employment\_Status, Account\_Tenure, Credit\_Card\_Type, Location
  + Monthly payment behavior (Month\_1 to Month\_6)

]

* - Data types: Numerical: Age, Income, Credit Score, Credit Utilization, Loan Balance
* Categorical: Employment Status, Credit Card Type, Location, Monthly Payment Status
* Binary: Delinquent Account (0 = No, 1 = Yes)

**Anomalies & inconsistencies:**

* Some entries have extremely low or high values for Credit\_Utilization (e.g., >1.0 or near 0)
* A few records show missing or zero values for Loan\_Balance and Debt\_to\_Income\_Ratio
* Monthly payment status includes inconsistent labels (e.g., “Missed” vs “Late”) that may need standardization

# 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:

- Variables with missing values:  Loan\_Balance: ~3% entries show zero or missing values

* Debt\_to\_Income\_Ratio: ~2% entries are zero or unrealistic
* Income: Some outliers below ₹20,000 and above ₹2,00,000 may skew analysis

- Missing data treatment:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Handling method** | **Justification** |
| Loan\_Balance | Impute (median) | Median preserves distribution and avoids outliers |
| Debt\_to\_Income\_Ratio | Synthetic generation | GenAI used to simulate realistic ratios |
| Income | Outlier filtering | Extreme values removed to stabilize model inputs |

# 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 **high credit utilization (>0.6)** and **low credit scores (<400)** show strong association with delinquency

* **Unemployed status** and **low income-to-debt ratio** are common among delinquent accounts
* **Multiple missed payments** across months are a reliable early warning signal

]

- Unexpected anomalies: [ A few customers with **high income and good credit scores** still show delinquency—may indicate behavioral or regional factors

* Some entries with **zero loan balance** still have high debt-to-income ratios—requires further review

]

# 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.

 Summarize dataset structure and highlight anomalies

* Suggest imputation strategies for missing values
* Rank top predictors of delinquency based on feature patterns
* Generate synthetic values for missing debt ratios and income entries

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 shows clear patterns linking credit utilization, missed payments, and employment status to delinquency risk. After cleaning and imputing missing data, the next step is to build a predictive model using GenAI to classify high-risk customers. This model will support Geldium’s collections team in prioritizing outreach and designing fair, targeted interventions.