**EDA SUMMARY REPORT – GELDIUM**

# **1. Introduction**

This report summarizes the exploratory analysis conducted on Geldium’s customer dataset, with the objective of assessing data quality, identifying missing or inconsistent values, and detecting early indicators of credit delinquency. These findings will support the development of AI-powered delinquency prediction models.

# **2. Dataset Overview**

* **Number of records:** 500
* **Key variables:**
  + Customer\_ID (Categorical)
  + Age, Income, Credit\_Score, Credit\_Utilization (Numerical)
  + Missed\_Payments, Loan\_Balance, Debt\_to\_Income\_Ratio, Account\_Tenure (Numerical)
  + Delinquent\_Account (Binary – Target)
  + Employment\_Status, Credit\_Card\_Type, Location (Categorical)
  + Month\_1 to Month\_6 (Categorical – payment behavior)
* **Data types:** 9 Numerical, 10 Categorical
* **Anomalies & Notes:**
  + Credit\_Utilization has values >1 (max = **1.025843**) – indicates overuse
  + Missing values:
    - Income: 39
    - Credit\_Score: 2
    - Loan\_Balance: 29

# **3. Missing Data Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Missing Count | Handling Method | Justification |
| Income | 39 | Median Imputation | Skewed distribution → median less sensitive to outliers |
| Credit\_Score | 2 | Mean Imputation | Very few missing values; normal-like distribution |
| Loan\_Balance | 29 | Predictive Imputation | Correlated with income, credit utilization, and missed payments |

# **4. Key Findings and Risk Indicators**

**Observed Correlations:**

* Higher **Missed\_Payments** often mean **Delinquent\_Account = 1**
* High **Credit\_Utilization** (>0.75) aligns with missed payments
* Low **Credit\_Score** (<500) + short **Account\_Tenure** (<3 yrs) = more risk

**Unexpected Anomalies:**

* Credit\_Utilization values exceed 100%
* Some high credit score customers still delinquent

**Risk Indicators:**

|  |  |
| --- | --- |
| Risk Factor | Reason for Risk |
| High Credit\_Utilization | Indicates financial stress |
| >3 Missed\_Payments | Historical pattern of delinquency |
| Debt\_to\_Income\_Ratio > 0.4 | Lower repayment capacity |
| Income < $40,000 | Increased vulnerability to missed payments |
| Unemployed status | Unstable or no income stream |

# **5. AI & GenAI Usage**

Used for:

* Summarizing patterns and missingness
* Recommending imputation strategies
* Exploring synthetic income generation options

**Example Prompts Used:**

* *"Summarize key patterns, outliers, and missing values in this dataset."*
* *"Suggest best-practice imputation strategies for missing credit data."*
* *"Generate synthetic income values using normal distribution aligned with median income bands."*

# **6. Conclusion & Next Steps**

The dataset is generally well-structured. Minor missing values and anomalies were handled using standard imputation and predictive strategies. Key features such as Missed\_Payments, Credit\_Utilization, and Debt\_to\_Income\_Ratio will significantly impact model development.

**Next Steps:**

* Finalize and apply imputation
* Engineer features such as payment behavior trends
* Begin modeling with fairness checks
* Validate outputs with Geldium’s risk experts