**Exploratory Data Analysis (EDA) Summary**

|  |
| --- |
| Name: Mujahedul Islam  LinkedIn: [www.linkedin.com/in/mi-makhdumi](http://www.linkedin.com/in/mi-makhdumi)  GitHub: <https://github.com/mimakhdumi/Credit-Card-Delinquency-Risk-Prediction/> |

# Introduction

# The purpose of this report is to conduct an exploratory data analysis (EDA) of Geldium’s dataset to assess its quality, identify gaps, and uncover early indicators of delinquency risk.

# The goal is to provide Tata iQ’s analytics team and Geldium’s decision-makers with clear insights into data patterns, anomalies, and risk factors. These findings will guide data cleaning, refinement of delinquency risk models, and the development of more effective intervention strategies.

# 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:

1. Credit\_Score- Numerical, (have 2 missing values)
2. Credit\_Utilization- Numerical,
3. Missed\_Payments- Numerical,
4. Loan\_Balance- Numerical, (Have missing values)
5. Debt\_to\_Income\_Ratio- Numerical,
6. Income- Numerical, (Have missing values)
7. Employment\_Status- Categorical,
8. Account\_Tenure- Numerical,
9. Month\_1–Month\_6- Categorical. (The given data is in ordinal categorical form, and for ML modelling we need to represent it through numerical transformation.)

There are no duplicate entries based on customer ID; each ID is unique.

# 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:

1. Credit\_Score
2. Loan\_Balance
3. Income

- For Credit\_Score: Model-based imputation involves training a model with correlated variables (such as income, repayment history, and delinquency) to estimate missing values. It is preferred because it leverages related features to maintain accuracy and variability, while avoiding the bias of simple methods like mean/median imputation, making it more reliable for financial risk modeling.

- For Loan\_Balance: the best imputation method is Model-based imputation (regression or machine learning model) using predictors such as Income, Debt\_to\_Income\_Ratio, Credit\_Utilization, and Credit\_Score. because it leverages related financial variables, preserves data relationships, and gives the most reliable estimates compared to simpler methods.

- For Income: For Income, the best imputation method is Model-based imputation using features such as:

* Employment\_Status (strongly linked to income),
* Credit\_Score (indirectly reflects repayment capacity),
* Loan\_Balance,
* Debt\_to\_Income\_Ratio,
* Credit\_Card\_Type, and possibly Location.

because it uses employment and financial indicators to estimate realistic values, avoids bias, and better preserves the data’s predictive power.

# 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:

1. **Income ↔ Debt\_to\_Income\_Ratio**
   * Higher income → usually lower debt-to-income ratio (if debts are not too high).
   * Strong negative correlation expected.
2. **Income ↔ Credit\_Score**
   * Higher income may be linked with higher credit score (better repayment ability).
   * Positive but not perfect correlation.
3. **Credit\_Utilization ↔ Credit\_Score**
   * Higher utilization (using most of the available credit) usually lowers the score.
   * Strong negative correlation.
4. **Missed\_Payments ↔ Credit\_Score**
   * More missed payments → lower score.
   * Strong negative correlation.
5. **Missed\_Payments ↔ Delinquent\_Account**
   * If missed payments are frequent, delinquency is more likely.
   * Strong positive correlation.
6. **Loan\_Balance ↔ Debt\_to\_Income\_Ratio**
   * Larger loan balance increases the debt-to-income ratio (unless income is very high).
   * Positive correlation.
7. **Employment\_Status ↔ Income / Credit\_Score**
   * Unemployed → lower income → weaker credit profile.
   * Strong relationship, but categorical.
8. **Account\_Tenure ↔ Credit\_Score**
   * Longer account history often improves credit score (shows reliability).
   * Positive correlation.
9. **Credit\_Card\_Type ↔ Income / Credit\_Score**
   * Premium cards (Gold/Platinum) usually linked with higher income and better score.
   * Positive association.
10. **Month\_1 to Month\_6 ↔ Missed\_Payments & Credit\_Score**

* Frequent late/missed entries in these months → higher missed payments count → lower score.
* Strong correlation.

# 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:

- "Summarize the dataset by listing key variables, their types (numerical/categorical), and possible value ranges."

- Suggest the best imputation methods for missing values in this dataset, considering variable types and business context."6. Conclusion & Next Steps

- “Identify correlations among numerical variables and explain which features are most strongly related."

**6. Conclusion & Next Steps**

**Conclusion**

The dataset is well-structured but has missing values in key variables. EDA shows strong links between income, utilization, missed payments, and credit score, highlighting important risk indicators.

**Next Steps**

* Impute missing values using model-based methods.
* Convert categorical/ordinal data into numerical form.
* Engineer new features and validate correlations.
* Prepare the cleaned dataset for delinquency risk modeling.