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

# 1. Introduction

Purpose: Provide an initial EDA of Geldium’s delinquency dataset to

* (1) evaluate data quality and completeness
* (2) identify early risk indicators relevant to delinquency modelling
* (3) recommend pre-processing and imputation steps required before model training

# 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: Customer\_id, Age, income, credit\_score, credit\_utilization, missed\_payment, Delinquent\_account, Loan\_balance, Debt\_to\_income\_ratio, Employment\_status, Account\_tenure, Month\_1 to Month\_6

- Data types: Categorical [Customer\_ID, Employment\_Status, Credit\_card\_Type, Location, Month\_1 to Month\_6], Numerical[Age, Income, Credit\_Score, Credit\_Utilization, Loan\_Balance, Debt\_to\_Income\_Ratio, Account\_Tenure], and Binary [Delinquent\_Account]

# 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: Income (39), Credit\_score (2), Loan\_Balance (29)

- Missing data treatment: Impute with median (robust to outlier) or use predictive imputation

# 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:

1. Credit utilization and credit score show a positive correlation with delinquency individually
2. Missed payment history is strongly associated with future delinquency, even single past missed payment increases the likelihood significantly
3. Lower credit scores (especially <600) correlate with higher delinquency risk
4. Debt-to-income ratio exhibits a moderate positive correlation with delinquency higher ratios indicate greater repayment pressure

- Importance of features (Quick Random Forest feature importance)

1. Loan Balance: importance = 0.1048
2. Income: importance = 0.0996
3. Credit score: importance = 0.0966
4. Debt to income ratio: importance = 0.08939
5. Credit utilization: importance = 0.0814
6. Age: importance = 0.0729

- Unexpected anomalies:

1. No records with **Credit Utilization > 100%** were found, indicating that utilization values fall within realistic bounds.
2. No records with **negative income values** were detected, suggesting income data is free from obvious sign-entry errors.
3. No records with **Credit\_Score < 300** were present, meaning all scores are within the expected credit-scoring range.
4. *Observation:* The absence of these anomalies suggests good data quality in these key numeric features, reducing the need for manual correction in these areas.

# 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 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

1. **Dataset completeness:** The dataset is largely clean, with minimal missing values. No anomalies such as credit utilization over 100%, negative incomes, or credit scores below 300 were detected, indicating good baseline data quality.
2. **Categorical data preparation:** Non-numeric columns (Location, Credit\_Card\_Type, Employment\_Status) were standardized and encoded for analysis, ensuring consistency and avoiding artificial category mismatches.
3. **Correlation findings:** Strong positive correlations were observed between delinquency and features like credit utilization, missed payment history, and debt-to-income ratio, while credit score showed a negative correlation (lower scores → higher delinquency risk).
4. **Payment history trends:** Month-wise payment performance (Month\_1 to Month\_6) showed that late and missed payments have a direct relationship with delinquency likelihood, even though their individual statistical correlation was moderate.
5. **Feature importance ranking:** RandomForest analysis identified Loan\_Balance, Income, Credit\_Score, and Debt\_to\_Income\_Ratio as top predictors. High importance of Customer\_ID suggests potential data leakage that should be addressed before modeling.
6. **Next steps:** Proceed with targeted feature engineering (rolling missed payment counts, recent utilization changes), further validation to prevent ID leakage, and model training using the most impactful predictors, ensuring fairness checks across demographic variables.