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

# 1. Introduction

This exploratory data analysis (EDA) has been conducted on Geldium’s customer dataset to support Tata iQ’s analytics team and Geldium’s strategic decision-makers. The primary objective is to uncover key patterns, anomalies, and early indicators related to credit delinquency. These insights will directly inform the enhancement of Geldium’s delinquency risk model and guide the development of more targeted and effective intervention strategies.

# 2. Dataset Overview

The dataset comprises 500 customer records with 19 attributes, designed to support analysis of credit delinquency patterns. It includes a mix of demographic, financial, and behavioral features such as income, credit score, loan balance, and monthly payment history. The primary outcome variable, **Delinquent\_Account**, indicates whether a customer has defaulted. Notable data quality considerations include missing values in key fields like income and loan balance, as well as the presence of categorical variables requiring transformation for modeling purposes.

**Key dataset attributes:**

**- Number of records:** 500 customer records(rows) and 19 attributes(column)

**- Key variables:** Income, Credit Score, Loan balance, and Monthly payment

**- Data types:**

| **Attribute** | **Type** |
| --- | --- |
| Age | Numerical |
| Income | Numerical |
| Credit\_Score | Numerical |
| Credit\_Utilization | Numerical |
| Missed\_Payments | Numerical |
| Loan\_Balance | Numerical |
| Debt\_to\_Income\_Ratio | Numerical |
| Employment\_Status | Categorical |
| Credit\_Card\_Type | Categorical |
| Location | Categorical |
| Month\_1 to Month\_6 | Categorical (Ordinal/Sequential)\* |

# 3. Missing Data Analysis

The missing data for the dataset has the three main attributes that are major missing values for the list

* Income - 7.8% (39 values)
* Loan\_balance - 5.8% (29 values)
* Credit\_score - 0.4% (2 values)

The attributes that are listed above have datatypes of numeric and for that the best approach to handle the missing value is for Income and Loan\_balance, the method is Median Imputation and for the Credit\_score method to be used is Mean Imputation. To use this method and justification for this for the Median used method is that the number of missing values is greater and the data distribution is more and for the Mean imputation is that data is normally distributed data.

Key missing data findings:

- Variables with missing values:

* Income - 7.8% (39 values)
* Loan\_balance - 5.8% (29 values)
* Credit\_score - 0.4% (2 values)

- Missing data treatment: Droppingr the Customer\_ID attribute is unique for every row.

# 4. Key Findings and Risk Indicators

* The dataset analysis reveals that **Missed\_Payments**, **Credit\_Utilization**, and **Debt\_to\_Income\_Ratio** are the **top 3 risk indicators** for predicting delinquency.
* These features directly reflect a customer's **financial behavior** and **repayment capacity**.
* **Missed\_Payments** is a strong signal of prior delinquency, making it highly predictive of future defaults.
* **Credit\_Utilization** above typical thresholds (e.g., >30%) often suggests overspending and increased risk.
* **Debt\_to\_Income\_Ratio** indicates the financial burden — higher ratios correlate with greater likelihood of payment delays.
* No individual feature besides the target showed a strong linear correlation (|r| > 0.2), highlighting the **need for non-linear modeling**.
* Features such as **Credit\_Score**, **Loan\_Balance**, and **Account\_Tenure** show moderate influence but need deeper interaction modeling.
* Categorical features like **Employment\_Status** and **Credit\_Card\_Type** may impact risk in specific segments.
* **Mean and median imputation** preserved data quality while handling missing values in financial attributes.
* Monthly indicators (Month\_1 to Month\_6) may carry temporal trends in customer behavior and can be valuable if aggregated.
* Overall, a **tree-based machine learning model** is recommended to capture complex feature interactions and non-linear effects effectively.

# 5. AI & GenAI Usage

* The dataset shows strong behavioral patterns around credit and debt, especially in relation to Delinquent\_Account.
* Missed\_Payments, Credit\_Utilization, and Debt\_to\_Income\_Ratio are key predictors of delinquency.
* High Credit\_Utilization and Debt\_to\_Income\_Ratio values indicate financial stress, increasing default risk.
* Income and Loan\_Balance exhibit **right-skewed distributions** with some extreme outliers.
* Several records show Credit\_Utilization over 100%, which may suggest data errors or risky borrowers.
* Class imbalance likely exists in the target variable (Delinquent\_Account), which should be addressed during modeling.
* Monthly columns (Month\_1 to Month\_6) can reveal behavioral trends over time and support delinquency prediction.
* Some anomalies in Debt\_to\_Income\_Ratio and Income could affect model performance without preprocessing.
* **Median imputation** is recommended for missing Income values due to skewness and the presence of outliers.
* Industry best practice also suggests **group-wise median imputation** using related features like Employment\_Status or Location.
* Post-imputation, log transformation or capping may be applied to reduce the influence of extreme income values.
* Overall, tree-based models (e.g., Random Forest, XGBoost) are ideal for handling feature interactions and non-linear relationships in this data.

# 6. Conclusion & Next Steps

* **Perform deeper outlier analysis** using boxplots and z-score/IQR methods for skewed financial features.
* **Engineer new features** from monthly data to capture recent trends (e.g., average missed months or payment patterns).
* **Balance the target classes** using techniques like SMOTE or class weighting if a class imbalance exists.
* **Train and evaluate models** using Random Forest or XGBoost, with cross-validation to ensure generalization.
* **Run SHAP or feature importance analysis** post-modeling to understand which variables drive predictions.
* **Collaborate with domain experts** to validate high-risk patterns and interpret financial thresholds realistically.
* **Document assumptions and preprocessing steps** clearly for reproducibility and future audits.