Exploratory Data Analysis (EDA)

1. Introduction

This report provides an Exploratory Data Analysis (EDA) of Geldium's customer dataset. The goal is to check data quality, identify missing or inconsistent values, and find early indicators of delinquency risk before building predictive models.

---

2. Dataset Overview

This dataset contains financial and behavioral information of customers used for delinquency prediction.

Key dataset attributes:

- **Number of records:** 500
- **Number of columns:** 19
- **Key variables:**
- `Customer_ID              object`
- `Age                       int64`
- `Income                  float64`
- `Credit_Score            float64`
- `Credit_Utilization      float64`
- `Missed_Payments           int64`
- `Delinquent_Account        int64`
- `Loan_Balance            float64`
- `Debt_to_Income_Ratio    float64`
- `Employment_Status        object`
- `Account_Tenure            int64`
- `Credit_Card_Type         object`
- `Location                 object`
- `Month_1                  object`
- `Month_2                  object`
- `Month_3                  object`
- `Month_4                  object`
- `Month_5                  object`
- `Month_6                  object`

- **Data types:**
  - Numerical: 9 Columns
  - Categorical: 10 Columns

Notes:

- No. of duplicates: 00
- Anomalies observed: **80 rows** in *Delinquent_Account* column are unusual (outside normal range).All other columns (Age, Income, Credit_Score, Loan_Balance, etc.) have **0 anomalies**.

---

3. Missing Data Analysis

Missing data was checked using pandas.

Missing values found in:

- [Column 3: 39]
- [Column 4: 2]
- [Column 8: 29]

Treatment applied:

- Income → imputed using **median**
- `Credit Score` → imputed using **mean**
- `Loan Balance`→ → imputed using **median**

Reason:

- Income values are usually skewed, so the median gives a more accurate central value than the mean.
- Credit scores are normally distributed, so the mean represents the typical value better than the median.
- Loan balances usually have large outliers, so the median gives a more stable and realistic replacement for missing values.

---

4. Key Findings and Risk Indicators

Risk Indicators

• **High credit utilization (>70%)** — customers using more than 70% of their credit limit have a higher chance of becoming delinquent.
• **2 or more missed payments** — missing payments repeatedly is a strong sign of repayment risk.

• **Low income (below median income)** — customers earning less may struggle to make regular payments.

Interesting Patterns

• Customers with **many missed payments** often also fall into the **low-income** group.
• Most customers have **normal credit utilization**, but a small group shows high missed-payment counts even with low utilization.
• Income levels vary widely, which may affect the performance of future prediction models.

5. AI & GenAI Usage

AI tools (ChatGPT) were used for summarizing patterns, suggesting imputation methods, and interpreting relationships.

Prompts used:

- "Summarize missing values and outliers for this dataset."
- "Suggest the best imputation method for missing income values."
- "Suggest the best imputation method for missing Credit Score values."
- "Identify key risk indicators for delinquency prediction."

6. Conclusion & Next Steps

The dataset is fully complete and shows no missing values, making it ready for analysis. Key risk indicators such as high credit utilization, low income, and repeated missed payments were identified as strong predictors of delinquency. The data also shows meaningful patterns, including the link between missed payments and lower income levels. These insights provide a solid foundation for the next steps. Moving forward, we can perform feature engineering, explore variable correlations, and prepare the dataset for building an accurate delinquency prediction model.