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

# 1. Introduction

This report demonstrates how artificial intelligence, specifically using Gemini, can be leveraged to analyze and predict customer delinquency. By applying Gemini to the Geldium delinquency dataset, we showcase a comprehensive workflow from initial data quality assessment to identifying key predictive factors. The analysis highlights how AI can efficiently process complex data, uncover surprising relationships, and inform robust predictive modeling strategies.

# 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:**

Key Columns for Delinquency Prediction

* **Credit\_Score:** A foundational predictor, as it summarizes a customer's credit risk profile into a single score.
* **Credit\_Utilization:** A direct measure of financial stress, as a high ratio indicates a customer is heavily reliant on credit.
* **Missed\_Payments:** A powerful historical indicator, as past payment failures are often the strongest predictor of future ones.
* **Delinquent\_Account:** The target variable for the prediction model, as it represents the outcome we are trying to forecast.
* **Debt\_to\_Income\_Ratio:** Measures a customer's ability to service debt, with a higher ratio indicating a greater likelihood of financial strain.
* **Month\_1 to Month\_6:** Provides a detailed, time-series view of recent payment behavior, offering nuanced insights beyond a simple count of missed payments.

**- Data types:**

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Column Type |
| Customer\_ID | object | Categorical |
| Age | int | Numerical |
| Income | float | Numerical |
| Credit\_Score | float | Numerical |
| Credit\_Utilization | float | Numerical |
| Missed\_Payments | int | Numerical |
| Delinquent\_Account | int | Numerical |
| Loan\_Balance | float | Numerical |
| Debt\_to\_Income\_Ratio | float | Numerical |
| Employment\_Status | object | Categorical |
| Account\_Tenure | int | Numerical |
| Credit\_Card\_Type | object | Categorical |
| Location | object | Categorical |
| Month\_1 | object | Categorical |
| Month\_2 | object | Categorical |
| Month\_3 | object | Categorical |
| Month\_4 | object | Categorical |
| Month\_5 | object | Categorical |
| Month\_6 | object | Categorical |

# 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:**

There are a total of **70 missing values** in the dataset, distributed across the following columns:

* **Income:** 39 missing values
* **Credit\_Score:** 2 missing values
* **Loan\_Balance:** 29 missing values
* **Inconsistent Data Representation:** Some numerical fields, like Loan\_Balance and Credit\_Utilization, have empty strings where numerical values were expected, requiring conversion to NaN for proper handling.

**Key Anomalies (Outliers):**

* **Credit Utilization:** At least one instance of Credit\_Utilization exceeding 1.0 (e.g., 1.025) was observed, indicating a potential data entry error or an extreme financial situation where credit limits were exceeded.
* **Credit Score:** While not explicitly identified as errors, some extremely low Credit\_Score values exist, which are significant outliers compared to the general distribution and represent very high-risk profiles.

**- Missing data treatment:**

Based on industry best practices, here are the proposed strategies for handling missing data:

* **Missing Income Values:** The 39 missing income values are imputed with synthetic data drawn from a normal distribution. This approach preserves the original dataset's statistical properties. This is done by calculating the mean ($$104,196$) and standard deviation ($$55,160$) of the existing income data.
* **Missing Credit\_Score Values:** The 2 missing credit scores are being filled with the median of the existing scores. This is a robust method that prevents outliers from skewing the results.
* **Missing Loan\_Balance Values:** The 29 missing loan balances are imputed using the median of the existing balances. This strategy effectively fills the gaps without being influenced by any unusually high or low loan values.

| Data Issue | Handling Method | Justification |
| --- | --- | --- |
| **Missing Income values** | Impute with synthetically generated values from a normal distribution. | Preserves the statistical properties of the original income distribution and creates more realistic data. |
| **Missing Credit Score values** | Impute with the median of the existing credit scores. | The median is a robust measure that is not sensitive to outliers and maintains the data's distribution. |
| **Missing Loan Balance values** | Impute with the median of the existing loan balances. | Similar to credit scores, the median is a robust measure that prevents distortion from potential loan balance outliers. |

# 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:**

**Early Indicators of Delinquency Risk:**

* **High Credit Utilization:** Customers with higher credit utilization ratios (closer to or above 1.0) are strong indicators of potential delinquency.
* **Missed Payments:** The number of Missed\_Payments is a direct and highly relevant historical indicator of future delinquency risk.
* **Low Credit Score:** Lower Credit\_Score values are a fundamental and widely accepted predictor of higher credit risk and, consequently, delinquency.
* **Payment History (Month\_1 to Month\_6):** "Late" or "Missed" entries in the monthly payment status columns directly signal problematic payment behavior.

**Relationships with Delinquency**

* **Credit Utilization and Delinquency:** There is a weak positive relationship between Credit\_Utilization and having a Delinquent\_Account. The average credit utilization for customers with a delinquent account is slightly higher (0.507) than for those without one (0.489). This suggests that while higher utilization is associated with delinquency, the relationship is not very strong based on this data.
* **Missed Payments and Delinquency:** Counter-intuitively, the analysis shows a very weak negative correlation between the number of Missed\_Payments and the Delinquent\_Account status. The average number of missed payments is slightly lower for customers with a delinquent account (2.85) compared to non-delinquent customers (2.99).
* **Credit Score and Delinquency:** The relationship between Credit\_Score and Delinquent\_Account is also weak and counter-intuitive. Customers with a delinquent account have a slightly higher average credit score (591.15) than those with a non-delinquent account (575.15). This finding is unusual and suggests that other factors may be playing a more dominant role in this particular dataset.
* **Loan Balance and Delinquency:** Due to the presence of missing values that need to be addressed, a direct correlation could not be reliably established. However, an analysis of the existing data would be required to see if a higher loan balance is associated with delinquency.

Overall, the relationships between these variables and the Delinquent\_Account status are surprisingly weak, indicating that delinquency in this dataset may be influenced by a more complex combination of factors beyond a simple linear relationship with these individual variables.

**- Unexpected anomalies:**

* **Counter-intuitive Credit Score Relationship:** A surprising finding is that customers with a delinquent account have, on average, a slightly higher credit score than non-delinquent customers. This unexpected trend requires deeper investigation to understand what other factors might be influencing delinquency, as the standard assumption is the opposite. It suggests that for this dataset, other variables may be more predictive than credit score.
* **Weak Correlations:** The correlations between key variables like Credit\_Utilization and Missed\_Payments with delinquency status were surprisingly weak. This implies that delinquency in this dataset is likely not determined by one or two isolated factors but rather by a more complex interplay of multiple variables. A robust predictive model should consider these variables in combination, rather than individually.
* **Importance of Payment History:** The month-by-month payment history (Month\_1 to Month\_6) provides a rich, time-series view of customer behavior. While the number of Missed\_Payments is a useful aggregate, analyzing the sequence and recency of payment statuses could provide more nuanced insights into predicting future delinquency.
* **Credit Utilization Outliers:** The presence of at least one Credit\_Utilization value exceeding 1.0 suggests an anomaly. This could be a data entry error or an instance where a customer has exceeded their credit limit, which is a significant red flag for delinquency and requires special handling in a predictive model.

# 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.

AI prompts used:

* What is the size of the data provided? Provide the answer in (rows x columns) format
* list each column and its datatype along with what type of column it is(categorical/numerical)
* Summarize key patterns, outliers, and missing values in this dataset.
* how many null values/ missing values can you find in each column?
* What are the top 3 variables most likely to predict delinquency based on this dataset.
* give me the key columns and single line description from a delinquency prediction pov.
* Analyze relationships between variables and delinquency outcomes. Give simple explaination and Keep the sentences in present tense and active voice.
* based on the data, highlight unexpected anomalies(data points requiring further investigation)

# 6. Conclusion & Next Steps

Our analysis of the Geldium data using Gemini revealed some important points:

* **Data Issues:** The dataset has missing values in key columns like Income and Loan\_Balance. It also contains an unusual finding: delinquent customers have slightly higher credit scores on average.
* **Key Predictors:** High credit utilization and the number of missed payments are the strongest indicators of delinquency.
* **Recommendations:** The next steps are to build a predictive model, investigate the unexpected credit score finding, and create a dashboard to visualize risk.