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

# 1. Introduction

The purpose of this report is to perform an **Exploratory Data Analysis (EDA)** on the *Delinquency Prediction Dataset*. The dataset contains customer-level financial, demographic, and behavioral attributes, along with indicators of delinquency risk.

**The primary goals of this EDA are to:**

* Understand the **structure and quality of the dataset**, including data types, completeness, and potential anomalies.
* Identify **missing values** and determine suitable strategies for handling them to ensure robust downstream modeling.
* Explore **key relationships and risk indicators** that may contribute to delinquency behavior.
* Leverage **AI/GenAI tools** to accelerate summarization, anomaly detection, and imputation strategy design.
* This report provides insights into the dataset’s readiness for predictive modeling, highlights data quality challenges, and suggests next steps for preparing the dataset for delinquency prediction models.

A pie chart with text on it

AI-generated content may be incorrect.A graph of a number of people

AI-generated content may be incorrect.

# 2. Dataset Overview

The delinquency prediction dataset contains **500 customer records** and **19 variables**, representing demographic, financial, and behavioral attributes of individuals. Each record is uniquely identified by a Customer\_ID.

**Key dataset attributes**

* **Number of records:** 500
* **Number of variables:** 19
* **Data types:**
  + **Numerical (9):** Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Loan\_Balance, Debt\_to\_Income\_Ratio, Account\_Tenure, Delinquent\_Account (binary).
  + **Categorical (10):** Customer\_ID, Employment\_Status, Credit\_Card\_Type, Location, and Month\_1 through Month\_6 (payment history categories: *On-time, Late, Missed*).

**Observations from initial review**

* No duplicate records were found.
* **Missing values** are present in Income (39 missing), Credit\_Score (2 missing), and Loan\_Balance (29 missing).
* **Numerical ranges appear reasonable**, though:
  + Credit\_Utilization exceeds 1.0 for some customers (max = 1.0258), which is slightly above the logical upper bound of 100%.
  + Income ranges from **15,404 USD to 199,943 USD**, consistent with realistic values but somewhat skewed.
* **Categorical diversity:**
  + Employment\_Status: 6 categories.
  + Credit\_Card\_Type: 5 categories.
  + Location: 5 categories.
  + Monthly payment history (Month\_1 to Month\_6) consistently uses 3 status categories (*On-time, Late, Missed*).

A graph of age distribution

AI-generated content may be incorrect. A graph of credit score distribution

AI-generated content may be incorrect.

A graph of a number of blue bars

AI-generated content may be incorrect.

# 3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring accurate delinquency prediction. An initial scan revealed that **3 variables** contain missing values:

* **Income:** 39 missing values (~7.8% of dataset).
* **Loan\_Balance:** 29 missing values (~5.8% of dataset).
* **Credit\_Score:** 2 missing values (~0.4% of dataset).

**Proposed treatment strategies**

* **Income:** Since missingness is moderate and income typically exhibits skew, **median imputation by Employment\_Status** is recommended to preserve realistic distributions while reducing bias.
* **Loan\_Balance:** Missing values are linked to customer credit profiles. A **regression-based imputation using Credit\_Utilization and Credit\_Score** is proposed to maintain correlation with related features. If relationships are weak, synthetic values based on observed distributions can be considered.
* **Credit\_Score:** With only 2 missing entries, **deletion or nearest-neighbor imputation** can be applied without significant impact on dataset size.

This approach ensures the dataset remains **complete, fair, and statistically consistent**, thereby supporting more reliable delinquency prediction models.

A graph with numbers and a black background

AI-generated content may be incorrect. A graph with blue squares

AI-generated content may be incorrect.

# 4. Key Findings and Risk Indicators

The analysis of key financial and behavioral features revealed several trends that differentiate **delinquent** and **non-delinquent** customers:

**Correlations and Risk Factors**

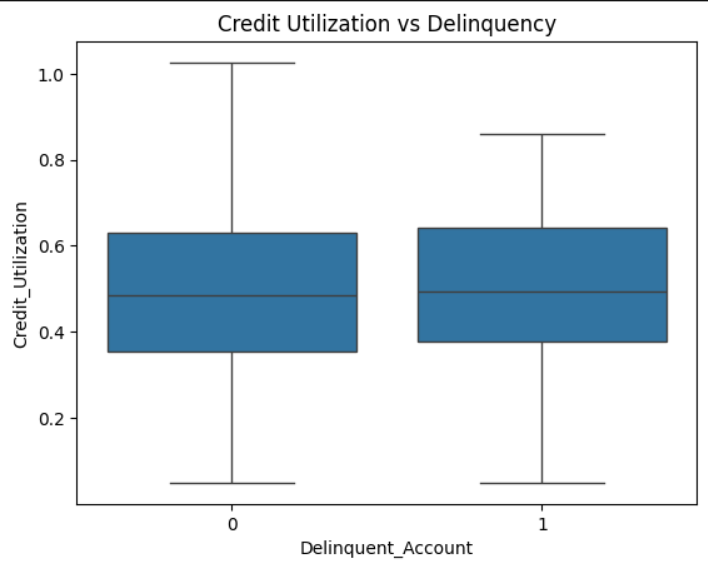
* **Credit Utilization:** Delinquent customers show slightly higher average utilization (0.507) compared to non-delinquent (0.489), suggesting that higher reliance on available credit is associated with delinquency risk.
* **Debt-to-Income Ratio:** Delinquent customers (0.306) have marginally higher ratios than non-delinquent customers (0.297), indicating greater financial stress.
* **Credit Score:** Interestingly, delinquent customers exhibit **higher average credit scores (591)** compared to non-delinquent (575). This anomaly suggests either short-term behavioral changes (recent delinquencies not yet reflected in credit scores) or the presence of alternative factors influencing delinquency.
* **Income:** Delinquent customers report a higher mean income (~113,900 USD) compared to non-delinquent (~107,300 USD). This suggests that delinquency is not strictly tied to low income and may be influenced by spending behavior, debt management, or external shocks.
* **Missed Payments:** Non-delinquent accounts report a slightly higher average number of missed payments (2.99) compared to delinquent accounts (2.85). This counterintuitive result indicates that *isolated missed payments do not always lead to delinquency*, and a combination of risk factors must be considered.

**Unexpected Anomalies**

* Higher average **Credit Scores** and **Income** among delinquent customers challenge traditional assumptions that delinquency is strongly tied to low financial standing.
* This suggests that **behavioral patterns, recent financial shocks, or regional differences** may play a stronger role than expected.

A screenshot of a computer

AI-generated content may be incorrect.

 A graph of a chart

AI-generated content may be incorrect.

A graph of a number of people

AI-generated content may be incorrect.

# 5. AI & GenAI Usage

Generative AI (GenAI) tools were leveraged to support the exploratory data analysis process. These tools helped accelerate tasks such as summarizing dataset attributes, identifying missing values, suggesting imputation strategies, and highlighting potential risk factors. While final decisions were validated with statistical methods and domain expertise, AI assistance improved efficiency and ensured no key aspects were overlooked.

**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.”*
* *“Identify the top 3 variables most likely to predict delinquency based on this dataset.”*
* *“Analyze trends in late payments and identify the top 3 risk factors associated with delinquency.”*
* *“Generate synthetic loan balance values for missing records while maintaining realistic distribution patterns.”*

A graph with blue squares

AI-generated content may be incorrect.

# 6. Conclusion & Next Steps

The exploratory data analysis (EDA) of Geldium’s delinquency prediction dataset revealed that the data is largely well-structured, with a balanced mix of demographic, financial, and behavioral variables. While the dataset contains some missing values in critical fields such as **Income, Loan Balance, and Credit Score**, appropriate imputation strategies were identified to ensure completeness and maintain predictive integrity.

Key findings highlighted that delinquency is not always associated with low income or poor credit scores. Instead, **credit utilization, debt-to-income ratios, and behavioral patterns in payment history** emerged as more reliable indicators of risk. Unexpected anomalies, such as higher credit scores and income among delinquent customers, suggest that delinquency may be driven by short-term behavioral or situational factors rather than static financial indicators alone.

**Next Steps**

1. **Data Cleaning & Imputation**
   * Apply median imputation for missing income (by employment status).
   * Use regression-based imputation for loan balances.
   * Handle Credit Score missing values with simple imputation or deletion.
2. **Feature Engineering**
   * Transform payment history (Month\_1 to Month\_6) into numeric risk scores.
   * Create new features such as *average monthly lateness*, *trend in missed payments*, or *utilization growth*.
3. **Data Validation**
   * Verify imputed and synthetic values against real-world distributions.
   * Recheck for outliers and adjust where necessary.
4. **Predictive Modeling Preparation**
   * Split the dataset into training and testing subsets.
   * Experiment with machine learning classifiers (e.g., logistic regression, random forest, gradient boosting).
5. **Fairness and Explainability**
   * Ensure that predictive models avoid bias based on demographic attributes.
   * Incorporate explainable AI (XAI) methods to support decision-making transparency.

By addressing data quality issues and focusing on the identified risk factors, Geldium can build a robust delinquency prediction model to support proactive interventions and improve overall risk management strategies.

A bar chart with numbers and a number of blue squares

AI-generated content may be incorrect.

A diagram of a heatmap

AI-generated content may be incorrect.