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

# 1. Introduction

This report presents an Exploratory Data Analysis (EDA) for Geldium’s credit risk dataset, aimed at identifying data quality issues, risk patterns, and key predictors of delinquency. The goal is to ensure a strong data foundation before deploying AI-driven delinquency prediction models.

# 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**:
  + Age – Customer’s age (Numerical)
  + Income – Annual income in USD (Numerical)
  + Credit\_Score – Score from 300–850 (Numerical)
  + Credit\_Utilization – % of credit used (Numerical, 0–1)
  + Missed\_Payments – Total missed in 12 months (Numerical)
  + Delinquent\_Account – Target: 1 = Yes, 0 = No (Binary)
  + Employment\_Status – Job type (Categorical)
  + Account\_Tenure – Years account is active (Numerical)
  + Month\_1 to Month\_6 – 6-month payment history (Categorical)
* **Data types**:
  + **Numerical**: Age, Income, Credit Score, Loan Balance, etc.
  + **Categorical**: Employment Status, Credit Card Type, Location, Month\_1–Month\_6
  + **Binary**: Delinquent\_Account
  + **Unique Identifier**: Customer\_ID
* **Initial anomalies**:
  + 39 missing values in Income
  + 29 missing values in Loan\_Balance
  + Credit\_Utilization has some values slightly over 1 (i.e., >100%)
  + No exact duplicates found (each Customer\_ID is unique)

## Summary in One Line (Optional):

The dataset is well-structured with mostly clean data, apart from a few missing entries in income and loan balance fields, and minor outliers in credit utilization.

# 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 missing values
  + Loan\_Balance → 29 missing values
  + Credit\_Score → 2 missing values

**Missing data treatment:**

| **Variable** | **Treatment Method** | **Justification** |
| --- | --- | --- |
| Income | **Median Imputation** | Median avoids skew from extreme values (e.g., high salaries) |
| Loan\_Balance | **Predictive Imputation or Median** | Can be estimated using Income, Age, Debt-to-Income; or fill with median |
| Credit\_Score | **Mean Imputation** | Very few missing; mean value keeps distribution balanced |

### Note:

Synthetic data generation was not required in this case, as missing data could be handled with traditional statistical methods.

## Summary Line:

All missing values were addressed using imputation techniques to preserve data integrity and avoid information loss during model training.

# 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.

### ****Correlations observed between key variables****:

* **High Credit Utilization** is weakly positively correlated with delinquency → Customers using >80% of their credit are at higher risk.
* **Debt-to-Income Ratio** shows a weak positive correlation with delinquency → >40% indicates possible repayment challenges.
* **Low Credit Score (<550)** is commonly found among delinquent customers → Reflects poor credit behavior.
* **Lower Income** is slightly correlated with higher delinquency risk → Financial instability may contribute to missed payments.
* **Missed Payments (Month 1 to 6)** – Customers who miss multiple payments early are more likely to default again.

### ****Unexpected anomalies****:

* Some Credit Utilization values exceed 100% → Data entry error or over-limit credit use, needs to be capped at 1.0.
* A few customers have **high Credit Scores but still show missed payments** → May require additional behavioral data (e.g., recent job loss).
* Customers with **high Income and high Debt** → Appear financially strong but over-leveraged; could be risky if interest rates rise.

## Summary Statement:

The most influential predictors of delinquency risk include credit utilization, debt-to-income ratio, and payment history. Some anomalies such as over-100% utilization and inconsistent behavior warrant deeper review before model deployment.

# 5. AI & GenAI Usage

In this project, GenAI tools such as ChatGPT were leveraged to assist in analyzing, cleaning, and interpreting the dataset. These tools played a supportive role in enhancing the accuracy, speed, and clarity of the Exploratory Data Analysis process.

**Dataset Summarization, Prompt-Based Insights,** Missing Value Strategies, Risk Indicator Detection, Documentation Assistance

### ****Why GenAI Helped:****

GenAI provided natural language explanations and automation, which made it easier to interpret complex data relationships. However, human judgment was still applied before finalizing any decision to ensure fairness and transparency.

Summary:

AI didn’t replace data science — it **augmented your work**. The combination of GenAI + analytical thinking helped deliver faster, smarter insights.

# 6. Conclusion & Next Steps

The Exploratory Data Analysis (EDA) revealed valuable insights into Geldium’s credit risk dataset. By identifying missing values, key correlations, and risk indicators, we laid a strong foundation for future predictive modeling. Income, loan balance, and credit score were addressed using imputation methods, ensuring data completeness.

Critical risk factors such as high credit utilization, elevated debt-to-income ratio, and repeated missed payments were identified as top predictors of delinquency. GenAI tools played a supporting role in summarizing findings, detecting patterns, and drafting analytical documentation efficiently.

### ****Next Steps:****

1. **Data Cleaning Finalization** – Apply imputation and fix anomalies (e.g., credit utilization >100%).
2. **Model Development** – Use the cleaned dataset to train machine learning models (e.g., logistic regression or decision trees).
3. **Validation** – Cross-check model accuracy using historical delinquency labels and performance metrics like precision, recall, and AUC.
4. **Deployment Prep** – Package the final model into a risk scoring system for Geldium’s collections team.
5. **Continuous Feedback Loop** – Monitor performance and update models as new customer behavior data becomes available.

**Summary Line:**

With clean and well-understood data, Geldium can now confidently move toward AI-driven risk modeling that is transparent, fair, and actionable