Exploratory Data Analysis (EDA) Summary Report

# Introduction

The purpose of this report is to provide an exploratory data analysis (EDA) of the Geldium Delinquency Dataset. The goal is to identify missing values, key patterns, risk indicators, and data quality issues that could influence delinquency prediction modeling. This report also documents the role of generative AI tools used in the analysis process.

# Dataset Overview

The dataset contains 500 customer records and 19 features related to credit, income, and payment history. It is a mix of categorical, integer, and float variables, making it suitable for credit risk modeling tasks.

* + Number of records: 500
  + Number of columns: 19
  + Data types: 10 categorical, 5 float, 4 integer
  + No duplicate rows detected
  + Anomalies observed: Credit Utilization values exceeding 100%

# Missing Data Analysis

Three columns were found to have missing data. Income and Loan\_Balance have moderate levels of missingness, requiring imputation, while Credit\_Score has very few missing values and can be easily handled.

* + Income: 39 missing (7.8%)  Moderate, imputation recommended
  + Loan\_Balance: 29 missing (5.8%)  Moderate, imputation recommended
  + Credit\_Score: 2 missing (0.4%)  Low, easy to impute or drop
  + Planned solutions:
  + Use the median to fill missing numeric values
  + Apply AI-generated synthetic data where appropriate for Loan Balance

# Key Findings and Risk Indicators

The analysis highlights several important patterns that may influence delinquency risk. Numerical variables such as income, credit score, and debt-to-income ratio show meaningful variation across customers. Categorical variables, especially the target variable, demonstrate imbalance that must be considered in model design.

* + Income: Average  $108,380, wide spread indicates diverse demographics.
  + Credit Score: Mean  578, with many poor-credit customers (25th percentile  418).
  + Credit Utilization: Average 49%, anomaly detected at > 100%.
  + Missed Payments: Customers average  3 missed payments, suggesting high risk.
  + Debt-to-Income Ratio: Average  30%, but max values > 50% indicate elevated risk.
  + Target imbalance: 84% non-delinquent vs. 16% delinquent accounts.

# AI & GenAI Usage

Generative AI tools were employed to summarize dataset attributes, detect anomalies, and propose imputation strategies. These tools accelerated the analysis process and ensured a structured summary.

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.'

# Conclusion & Next Steps

The Geldium Delinquency Dataset reveals strong signals for credit risk prediction, but also highlights challenges such as target imbalance, inconsistent categorical formatting, and anomalies in credit utilization. Addressing missing values and cleaning categorical data will be critical. Next steps involve feature engineering, class imbalance handling (e.g., SMOTE or stratified sampling), and testing multiple classification models (e.g., Random Forest, XGBoost, Logistic Regression).

Explainability methods such as SHAP should also be applied to ensure transparency in model outputs.