Exploratory Data Analysis (EDA) Report on Geldium's Delinquency Prediction Dataset

1. Introduction

This report presents the Exploratory Data Analysis (EDA) of Geldium's delinquency prediction dataset. The primary goal is to assess data quality, identify early risk indicators, and prepare the dataset for predictive modeling. Insights from this analysis will support Tata iQ's analytics team in refining intervention strategies.

2. Dataset Overview

Key Dataset Attributes

- **Number of records:** 500

Key Variables

- **Customer_ID:** Unique identifier for each customer
- **Income, Credit_Score, Credit_Utilization:** Financial metrics
- **Missed_Payments, Delinquent_Account:** Behavioral and target features
- **Employment_Status, Location, Account_Tenure:** Demographic and account details

Data Types

- **Numerical:** Age, Income, Credit_Score, Credit_Utilization, Loan_Balance, etc.
- **Categorical:** Employment_Status, Location, Credit_Card_Type, Month_1 to Month_6

3. Missing Data Analysis

Key Missing Data Findings

Variables with Missing Values:

- **Income:** 39 missing values
- **Credit_Score:** 2 missing values
- **Loan_Balance:** 29 missing values

Missing Data Treatment

- **Income:** Imputed using median due to skewness
- **Credit_Score:** Imputed using mean (minimal missing entries)

- **Loan_Balance:** To be imputed using regression-based imputation from related features

4. Key Findings and Risk Indicators

Key Findings

- **High Credit Utilization** correlates with higher delinquency.
- Customers with **multiple missed payments** across 6 months are more likely to be delinquent.
- **Low Credit Score** and **short Account Tenure** also indicate increased risk.

Unexpected Anomalies

Some customers with many missed payments are marked as non-delinquent, suggesting potential labeling or behavioral inconsistencies.

5. AI & GenAI Usage

Generative AI tools were employed to summarize the dataset, suggest data imputation strategies, and detect patterns.

Example AI Prompts Used

- “Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency.”
- “Suggest an imputation strategy for missing values in this dataset based on industry best practices.”
- “Identify the top 3 variables most likely to predict delinquency based on this dataset. Provide brief reasoning.”

6. Conclusion & Next Steps

This EDA revealed critical missing data areas and strong predictors of delinquency, such as credit utilization and missed payments.

Next Steps:

- Completing data imputation
- Performing feature engineering
- Building a predictive model to support Geldium in prioritizing interventions for at-risk customers.