Exploratory Data Analysis (EDA) of Geldium

Summary Report

1. Introduction

This Exploratory data analysis (EDA) report of Geldium provides comprehensive examination of a dataset designed for delinquency prediction, aiming to uncover data quality issues, understand variable distributions, and identify relationships between features and the target variable, Delinquent_Account.

2. Dataset Overview

This dataset includes 500 records, each representing a unique customer, with 19 features. The columns capture a variety of information, including customer demographics (Age, Income, Location, and Employment_Status), creditworthiness metrics (Credit_Score, Credit_Utilization, Debt_to_Income_Ratio), and current loan and account status (Loan_Balance, Account_Tenure). An essential component is the 6-month payment history, detailed in columns Month -1 through Month-6 which shows the status of each monthly payment (e.g., Late, Missed, On-time). Data types are a mix of numerical and categorical, with a few variables like Income, Credit_Score, and Loan_Balance.

Key dataset attributes**:**

- Number of records: 500
- Key variables :
    - Delinquent_Account - The target variable for prediction. A binary indicator where 1 signifies the customer has a delinquent account, and 0 signifies they do not.
    - Credit_Score - The customer's FICO-like credit score, a core measure of creditworthiness.
    - Credit_Utilization - The ratio of the customer's current outstanding credit card debt to their total credit limit. A lower value is generally better.
    - Missed_Payments - The total number of missed payments the customer has accrued over a certain historical period.
    - Income - The annual income of the customer.
    - Debt_to_Income_Ratio - The ratio of the customer's total monthly debt payments to their gross monthly income.
    - Loan_Balance - The current outstanding balance on the customer's loan.
    - Employment_Status - The current employment status of the customer (e.g., Employed, Self-employed, Unemployed).
    - Month_1 to Month_6 - A time-series of payment statuses for the last six months (e.g., Late, Missed, On-time), providing granular history.
- Data types:
    - Categorical: Employment_Status, Month_1 to Month_6, credit_card_type.

- Numerical: Delinquent_Account, Credit_Score, Loan_Balance, Missed_Payments, Credit_Utilization, Income, Debt_to_Income_Ratio.

3. Missing Data Analysis

Missing values were identified in three numerical features, all of which are important for credit risk assessment.

Key missing data findings:

- Variables with missing values:
  - Income - 39
  - credit_score - 2
  - Loan_Balance - 29
- Missing data treatment:
  - Using median imputation for filling the credit_score.
  - Using AI-generated synthetic data for the Loan balance and Income.

4. Key Findings and Risk Indicators

Key insights:

The correlations of all variables with Delinquent_Account are extremely weak (all values are very close to zero). Crucially, the coefficients for the strongest predictors of risk (e.g., Missed_Payments, Credit_Score, Credit_Utilization) are either close to zero or even have an unexpected sign (e.g., positive correlation for Credit_Score and negative for Missed_Payments).

Unexpected anomalies (Data points requiring further investigation):

- Non-delinquent but extremely high missed payments.
- Non-delinquent but very low credit score.
- Extremely low income.

5. Role of AI & GenAI

Generative AI tools were used and helped me to summarize the dataset, impute the missing data, and detect patterns in the dataset. This made the analysis more accurate and helped me analyze beyond the things.

Here are the AI prompts that I have used:

- 'Summarize me the overview of the dataset in a paragraph.'
- 'Suggest the key variables of this data set and description.'
- 'Identify and address the missing data that is critical in ensuring model accuracy.'
- 'Find the Correlations between key variables and Unexpected anomalies (Highlight data points requiring further investigation).'

6. Conclusion & Next Steps

This Exploratory Data Analysis (EDA) of Geldium revealed several key insights regarding missing data, Treatment plans (like median imputation), Correlations and Relationships between variables and Anomalies which will strongly influence the modeling strategy.

Next Steps:

- Execute Imputation: Implement the planned Median Imputation on Income, Loan_Balance, and Credit_Score.
- Handle Categorical Data: Encode all nominal features (Employment_Status, Location, Credit_Card_Type) using One-Hot Encoding.
- Encode Payment History: Convert the monthly payment status variables (Month_1 to Month_6) from descriptive categories (e.g., 'Late', 'Missed', 'On-time') into a numerical or ordinal scale to represent the severity of the payment status, which will be highly predictive
- Investigate Anomalies: The anomalous records (low Credit_Score but non-delinquent) should be investigated.

These efforts will aid Geldium in refining its analysis process and enhance data ratability for further modeling.