Exploratort Data Analysis (EDR) and Risk Profiling Summary Report

1. Introduction

This report focuses on identifying key relationships between different independent variables with the dependent variable and processing different inconsistent and abnormal values within the data with a final emphasis on what further steps should be taken.

2. Dataset Overview

Key dataset attributes:

- **Number of records:**

500 records with respect to 19 variables

- **Key variables:**

Age, Income, Credit Score, Credit Utilization, Missed Payments, Debt-to-Income Ratio, Employment Status, Account Tenure, Months 1 to 6

- **Data types:**

8 continous variables (3 int , 5 float)

10 categorical variables (object)

1 categorical variable with int data type

3. Missing Data Analysis

Key missing data findings:

A. Variables with missing values:

Income variable (29 missing records)

Credit_score variable (2 missing)

Loan_balance variable (39 missing)

B. Missing data treatment:

- Median imputation for Income and Loan_balance variable

(since there is very low correlation between missing variables and other variables so model based imputation are avoided. Also the distribution of both columns is a bit right skewed (since mode < median) so standard mode imputation is used.)

- Mean value imputation for Credit score Variable

(since there are very few values missing and the distribution of credit score is also

almost normal)

4. Key Findings and Risk Indicators

Key findings:

A. Correlations observed between key variables:

- **month_1 to month_6 :** Monthly payment behaviour of an individual over the past 6 months had the highest correlation with delinquency rate.
- **Income :** Customer income also corresponds highly with delinquency rate. Higher or lower income affects ability to repay debts.

B. Unexpected anomalies:

- **Abnormal Credit Utilization Values :** normally the values should lie within 0 to 1 (row no.s 89 , 265 , 292 , 426)
- **Insensible Correlation between Variables :** The variables “delinquent_account” and “Missed_Payments” seems kind of counter-intuitive and thus requires further investigation
- **Class imbalance in the dependent variable :** The two classes of the dependent variable are severly imbalanced (84% non-delinquent & 16% delinquent)

5. AI & GenAI Usage

Open AI’s GPT 4o was used to summarize dataset trends

Some of prompts used were:

Prompt 1 : Carefully inspect the provided dataset and briefly highlight the logical errors and unexpected anomalies that u notice in this dataset.

Prompt 2 : Suggest the best amputation methods for each of the columns in the dataset that contain null values.

6. Conclusion & Next Steps

Key Findings:

- the main factors that had the greatest impact on delinquency probability were, **annual income** and the **payment history of the past 6 months**.
- On the other hand, the inconsistencies present in the data were, **abnormal credit utilization values across several instances** and **surprising correlation between “delinquent_account” and “Missed_Payments” variables** (oppositely related).

Next Steps:

Emphasis on the abnormal values in credit utilization column as well as the counter_intuitive relation between **“delinquent_account”** and **“Missed_Payments” variables** for their justification. Furthermore, handing the data imbalance of dependent variable with appropriate measures.