Exploratory Data Analysis (EDA)

Summary Report

1. Introduction

This report presents an exploratory data analysis (EDA) of a credit-related dataset aimed at identifying patterns and risk indicators associated with customer delinquency. The insights will guide further predictive modeling and strategy development for delinquency mitigation.

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: Age of the customer
- Income: Annual income
- Credit Score: Credit score of the customer
- Credit Utilization: Ratio of used credit to available credit
- Missed Payments: Number of missed payments in past period
- Delinquent Account: Target variable indicating delinquency
- Loan Balance: Outstanding loan balance
- Debt to Income Ratio: DTI ratio
- Employment Status, Account Tenure, Credit Card Type, Location: Categorical features
- Month 1 to Month 6: Monthly payment behavior (e.g., Late, Missed, On-time)

- Data types:

- Numerical: Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure

- Categorical: Employment_Status, Credit_Card_Type, Location, Month_1 to Month_6

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
- Credit_Score: 2 missing
- Loan_Balance: 29 missing

- Missing data treatment:

- Credit_Score: Mean or median imputation (due to low missing count)
- Loan_Balance: Median imputation preferred to reduce skew effect
- Income: Imputed using synthetic values based on a normal distribution matching the dataset’s statistical properties (mean & std. dev)

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.

Key findings:

- Correlations observed between key variables:

- Higher Credit_Utilization, Missed_Payments, and Debt_to_Income_Ratio show positive correlation with Delinquent_Account.
- Customers with lower Credit_Score and shorter Account_Tenure are more likely to be delinquent.
- Payment patterns (Month_1 to Month_6) that include frequent "Missed" or "Late" values are strong indicators of delinquency.

- Unexpected anomalies:

- A few customers with high Income but still delinquent
- Some customers with 0 Account_Tenure but having payment histories

5. AI & GenAI Usage

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns. This section documents AI-generated insights and the prompts used to obtain results.

- Tasks automated using AI tools:

- Pattern detection (risk signals in monthly behavior)
- Imputation strategy for missing Income
- Summarizing correlations and feature relationships

6. Conclusion & Next Steps

This initial EDA highlighted key delinquency indicators and resolved data quality issues. Recommended next steps:

- Feature engineering from monthly data (Month_1 to Month_6)
- Use predictive modeling (e.g., logistic regression, random forest) to forecast delinquency
- Segment customers based on risk profiles for targeted intervention strategies