Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this report is to explore and analyse the customer dataset to identify patterns, missing values, and key risk indicators that can help predict loan delinquency. This analysis will inform data preparation and feature selection for machine learning models aimed at predicting delinquency risk.

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: Numerical
    - Income: Numerical (contains missing values)
    - Credit_Score: Numerical
    - Credit_Utilization: Numerical (0–1 range, 1 anomaly >1)
    - Missed_Payments: Numerical
    - Delinquent_Account: Binary target variable (0 = No, 1 = Yes)
    - Loan_Balance: Numerical (contains missing values)
    - Employment_Status: Categorical
    - Month_1 to Month_6: Categorical (On-time, Late, Missed)
- Data types:
    - Categorical: Employment_Status, Credit_Card_Type, Location, Month_1 to Month_6
    - Numerical: Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure
    - Binary: Delinquent_Account

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 (7.8%)
    - Loan_Balance: 29 missing (5.8%)
    - Credit_Score: 2 missing (0.4%)

Missing data treatment:

- Income: Due to the higher missing rate, imputation using median or model-based techniques is recommended to preserve dataset size and distribution.
- Loan_Balance & Credit_Score: Since missingness is low, mean/median imputation is acceptable.
- No columns were removed based on missing data.

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:

- Missed_Payments and Credit_Utilization are positively correlated with Delinquent_Account.
- Credit_Score is negatively correlated with delinquency — lower scores are linked to higher risk.
- Debt_to_Income_Ratio and Account_Tenure show moderate correlations with financial behavior.
- Categorical analysis reveals that certain employment types and card types may associate with increased delinquency risk.

Unexpected anomalies:

- One record has Credit_Utilization > 1.0, which may indicate data entry error.
- Rare or underrepresented categories in Employment_Status and Credit_Card_Type may need to be grouped during modeling.

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.

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.”
- “Identify top 3 factors correlated with Delinquent_Account based on statistical patterns.”

6. Conclusion & Next Steps

The dataset presents meaningful features such as Missed_Payments, Credit_Utilization, and Credit_Score that are highly relevant for predicting loan delinquency. Proper treatment of missing values and potential outliers will significantly enhance data quality.

Next Steps:

- Handle missing values and encode categorical features
- Normalize/scale numerical variables
- Perform feature engineering for modeling
- Apply classification algorithms to predict Delinquent_Account