Exploratory Data Analysis (EDA) Summary

1. Introduction

The purpose of this report is to perform an exploratory data analysis (EDA) on the delinquency dataset to assess data completeness, understand key variables, and identify early risk indicators that may impact delinquency prediction. The analysis aims to support a more structured, data-driven approach for identifying at-risk customers and improving collection strategies.

2. Dataset Overview

This dataset contains customer-level information related to loan characteristics, customer profiles, and repayment behavior. An initial review was conducted to understand the structure of the data and identify any inconsistencies.

Key dataset attributes:

- **Number of records:** 500 customers
- **Key variables:**
  - Customer profile: Age, gender, location, employment status, income
  - Loan details: Loan amount, tenure, interest rate, EMI amount
  - Repayment behavior: Days Past Due (DPD), number of missed EMIs, delinquent flag
- **Data types:**
  - Numerical: Loan amount, EMI amount, DPD, income
  - Categorical: Gender, location, employment status, delinquent flag

No duplicate records were observed during the initial review. Minor inconsistencies were noted in customer profile fields.

3. Missing Data Analysis

Identifying missing values is critical to ensure reliable delinquency assessment and future modeling accuracy.

Key missing data findings:

- **Variables with missing values:** Income and employment-related fields
- **Missing data treatment:**
  Missing values were not removed at this stage. For future modeling, imputation

using median income values or segment-based imputation (based on employment type or location) is recommended to avoid data loss.

4. Key Findings and Risk Indicators

This section identifies trends and patterns that may indicate risk factors for delinquency. Feature relationships were examined to uncover insights relevant to predictive modeling.

Key findings:

- Customers with higher **Days Past Due (DPD)** and a greater number of **missed EMIs** consistently fall into higher delinquency risk categories.
- Repayment behavior variables such as missed payments and payment delays are stronger indicators of delinquency risk compared to demographic attributes like age or location.
- Based on six-month payment behavior, customers were segmented into three risk groups:
    - **Low Risk:** 146 customers
    - **Medium Risk:** 195 customers
    - **High Risk:** 159 customers
- A significant portion of customers falls under Medium and High Risk categories, indicating elevated overall delinquency exposure.

Unexpected anomalies:

- A small number of customers exhibit **high loan amounts but low DPD**, suggesting stable income sources or external repayment support. These cases require further investigation to better understand underlying repayment capacity.

5. AI & GenAI Usage

Generative AI tools were used to assist in summarizing dataset characteristics, identifying potential risk indicators, and recommending appropriate missing data treatment strategies.

Example AI prompts used:

- *"Summarize key patterns in the delinquency dataset and identify risk indicators."*
- *"Suggest an imputation strategy for missing income values based on lending industry best practices."*

AI-generated insights were reviewed and validated against business understanding before being incorporated into the analysis.

6. Conclusion & Next Steps

The exploratory analysis indicates that repayment behavior variables, particularly DPD and missed EMIs, are the most critical indicators of delinquency risk. Risk segmentation shows that a majority of customers fall into Medium and High Risk categories, highlighting the need for proactive intervention strategies.

Recommended next steps:

- Impute missing income and employment data to improve model readiness
- Use risk segmentation to prioritize collection efforts
- Develop early-warning indicators based on payment behavior trends
- Incorporate additional behavioral history to enhance delinquency prediction accuracy