**Exploratory Data Analysis (EDA) Summary**

Introduction

The purpose of this report is to analyze the customer dataset with the objective of identifying risk factors for delinquency. The analysis highlights data quality issues, trends, correlations, and anomalies that may influence predictive modeling and risk assessment strategies.

Dataset Overview

This dataset contains financial, demographic, and behavioral attributes of 500 customers. It is structured to support delinquency risk prediction.

Key dataset attributes:

• Number of records: 500

• Number of variables: 19

Data types:

• Text: 1 (Customer ID)

• Numeric: 8 (Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Loan\_Balance, Debt\_to\_Income\_Ratio, Account\_Tenure)

• Categorical: 10 (Delinquent\_Account, Employment\_Status, Credit\_Card\_Type, Location, Month\_1–Month\_6)

Data quality observations:

• Missing cells: 70 (0.7% of dataset)

• Duplicate rows: None

Missing Data Analysis

Identifying missing data is critical for ensuring model robustness.

Key missing data findings:

• Income: 39 missing (7.8%)

• Credit\_Score: 2 missing (0.4%)

• Loan\_Balance: 29 missing (5.8%)

• Suggested missing data treatment:

• Income: Imputation using median/mean grouped by Employment\_Status or Location.

• Credit\_Score: Impute with median or predictive imputation using correlated features.

• Loan\_Balance: Median imputation or model-based imputation using Income, Credit\_Utilization, and Debt\_to\_Income\_Ratio.

Key Findings and Risk Indicators

Correlations & patterns:

• Credit\_Score inversely related to Missed\_Payments and Delinquent\_Account.

• Debt\_to\_Income\_Ratio and Credit\_Utilization are positively correlated with delinquency likelihood.

• Customers with higher Loan\_Balances and lower Credit\_Scores show greater delinquency risk.

Employment Status anomalies:

• Inconsistent labels observed: "Employed," "employed," and "EMP" should be standardized.

• Retirement and unemployment groups show higher delinquency risk compared to employed/self-employed customers.

Behavioral risk factors:

• A significant number of customers (15.4%) reported zero missed payments, while others had up to 6 missed payments.

• Monthly payment patterns (Month\_1–Month\_6) show repeated “Late” and “Missed” statuses in high-risk customers.

AI & GenAI Usage

Ydata\_profiling was leveraged to summarize dataset insights

Conclusion & Next Steps

The dataset is generally clean, with minimal missing data and no duplicates. However, categorical inconsistencies (Employment\_Status) and moderate missing values in financial attributes (Income, Loan\_Balance) require treatment.

Next steps:

• Standardize categorical variables (Employment\_Status).

• Apply appropriate imputation strategies for missing values.

• Conduct further correlation and feature importance analysis.

• Build and evaluate predictive for delinquency prediction.