Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this report is to summarize key findings from the delinquency prediction dataset, identify data quality issues, highlight patterns and anomalies, and uncover early indicators of delinquency risk. The analysis provides insights that will guide the development of predictive models and risk assessment strategies.

2. Dataset Overview

The dataset contains **500 customer records** with a **16% delinquency rate**. No duplicate records were found, indicating good data integrity.

Key dataset attributes:

- **Number of records:** 500
- **Key variables:** Credit_Score, Income, Loan_Balance, Credit_Utilization, Missed_Payments, Debt_to_Income_Ratio, Account_Tenure, Payment History (Month_1–Month_6), Customer Demographics (Employment_Status, Card_Type, City)
- **Data types:** Mix of **categorical** (e.g., Employment_Status, City, Card_Type) and **numerical** (e.g., Credit_Score, Income, Loan_Balance, Debt_to_Income_Ratio)

Key observations: Delinquent accounts show slightly lower average credit scores (591 vs. 575), higher debt-to-income ratios (0.306 vs. 0.297), and distinct payment history patterns. Interestingly, missed payments alone do not strongly differentiate delinquent from non-delinquent customers.

3. Missing Data Analysis

Critical Feature Gaps:

- **Income:** 39 missing values (7.8%)
- **Credit_Score:** 2 missing values (0.4%)
- **Loan_Balance:** 29 missing values (5.8%)
- **Credit_Utilization & Missed_Payments:** No missing values
- **Payment History (Month_1–Month_6):** No missing values

Recommended Treatment Approaches:

- **Income:** Impute with **KNN** – preserves relationships with other financial variables.
- **Credit_Score:** Impute with **Median** – very few missing values, robust and simple.
- **Loan_Balance:** Impute with **KNN** – captures relationships for more accurate imputation.
- **Credit_Utilization & Missed_Payments:** No treatment needed.
- **Payment History:** No treatment needed.

No variables require removal as none exceed 50% missing values. No synthetic data generation is required.

4. Key Findings and Risk Indicators

Obvious Patterns:

- Higher missed payments (3+) strongly correlate with delinquency.
- High credit utilization (>60%) indicates elevated delinquency risk.
- Low credit scores (<500) align with higher delinquency rates.
- High debt-to-income ratio (>40%) strongly linked to delinquency.

Unexpected Findings:

- Credit utilization has very weak correlation with missed payments (0.02).
- High credit score customers surprisingly show delinquency rates similar to lower score customers.
- Income levels do not significantly impact delinquency risk.

High-Risk Segments:

- Business credit card holders (21% delinquency rate).
- Unemployed individuals (19%).

- Los Angeles residents (20%).
- New accounts (<5 years) show higher delinquency.

Key Variables for Prediction:

- Missed_Payments (most predictive)
- Debt_to_Income_Ratio
- Credit_Score
- Account_Tenure (longer tenure reduces risk)

5. AI & GenAI Usage

AI tools were applied to summarize dataset characteristics, impute missing values, and extract key patterns. Example prompts and results include:

- **Prompt:** "Summarize key patterns in the dataset and identify anomalies."
  → **Insight:** Identified weak correlation between credit utilization and missed payments.
- **Prompt:** "Suggest an imputation strategy for missing income values based on industry best practices."
  → **Insight:** KNN imputation chosen for Income and Loan_Balance due to moderate missing rates.

6. Conclusion & Next Steps

The dataset demonstrates clear delinquency risk indicators, especially missed payments, high debt-to-income ratios, and low credit scores. However, unexpected findings around income and high-credit-score customers suggest the need for deeper exploration. Next steps include:

1. Implementing imputation strategies for missing values.
2. Building predictive models emphasizing missed payments, credit score, and DTI ratio.
3. Conducting further investigation into surprising patterns (e.g., why high credit score customers still show delinquency).

4. Exploring demographic and geographic risk adjustments for high-risk groups.