Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this report is to perform a comprehensive Exploratory Data Analysis (EDA) on the **Delinquency Prediction Dataset** to uncover early indicators of financial risk and identify variables most predictive of account delinquency. The goal is to understand the data quality, variable relationships, and behavioral or financial patterns that influence delinquency outcomes. This analysis aims to guide the development of a predictive model that can help financial institutions proactively detect at-risk customers, improve credit risk management, and enhance decision-making through data-driven insights.

2. Dataset Overview

This dataset contains **500 records** representing individual customer financial data and credit behavior.

Key dataset attributes:

- **Number of records:** 500
- **Key variables:**
    - **Customer_ID:** Unique identifier for each customer
    - **Age:** Age of the customer
    - **Income:** Annual income in USD
    - **Credit_Score:** Creditworthiness measure (300–850 scale)
    - **Credit_Utilization:** Ratio of used to available credit
    - **Missed_Payments:** Number of missed payments over the past 6 months
    - **Delinquent_Account:** Target variable (1 = delinquent, 0 = non-delinquent)
    - **Loan_Balance:** Current outstanding loan amount

- **Debt_to_Income_Ratio:** Ratio of debt payments to income
- **Employment_Status:** Employment category (e.g., Employed, Unemployed, Self-employed, etc.)
- **Account_Tenure:** Number of years since account opening
- **Credit_Card_Type:** Type of card (e.g., Gold, Silver, Platinum, etc.)
- **Location:** Customer's location/city
- **Month_1–Month_6:** Payment history for past six months

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–Month_6

---

3. Missing Data Analysis

Key missing data findings:

- **Variables with missing values:**
  - `Income` → 39 missing
  - `Credit_Score` → 2 missing
  - `Loan_Balance` → 29 missing

Missing data treatment:

- **Imputation method:**
  - For numerical columns (`Income`, `Credit_Score`, `Loan_Balance`): Median imputation was recommended due to the presence of outliers.
  - Categorical variables had no missing values.

Justification:
Median imputation helps preserve the data's central tendency without being skewed by extreme values, maintaining dataset stability for downstream model training.

---

4. Key Findings and Risk Indicators

Correlations observed between key variables:

- **Higher Credit_Utilization** is positively correlated with **Missed_Payments** and **Delinquent_Account**, indicating that overused credit lines increase delinquency risk.

- **Low Credit_Score** strongly associates with **higher Missed_Payments** and **greater Loan_Balance**, confirming its reliability as a risk indicator.

- **Debt_to_Income_Ratio** above 0.4 often aligns with delinquency cases, suggesting over-leveraging behavior.

- **Employment_Status**: Unemployed individuals show a higher proportion of delinquent accounts.

Unexpected anomalies:

- A few customers had **Credit_Utilization > 1.0**, suggesting potential data entry or reporting errors.

- Some accounts showed **zero Missed_Payments but low Credit_Score**, possibly indicating legacy credit issues not captured in recent months.

---

5. AI & GenAI Usage

Generative AI tools were utilized to accelerate the EDA process by identifying relationships, summarizing dataset statistics, and recommending data treatment strategies.

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.”*
- *“Analyze which features most strongly influence delinquency rates.”*

AI assistance helped validate findings, identify correlations, and document results more efficiently.

---

6. Conclusion & Next Steps

Summary of key findings:

- Approximately **16%** of customers are delinquent, often linked with **low credit scores**, **high utilization**, and **high debt-to-income ratios**.
- **Income** and **Loan_Balance** imputation are necessary preprocessing steps.
- **Employment_Status** and **Credit_Utilization** are strong categorical and numerical predictors respectively.

Next Steps:

1. Perform **feature scaling and encoding** for model readiness.
2. Apply **correlation-based feature selection** or **SHAP analysis** to identify top predictors.
3. Develop baseline models (e.g., Logistic Regression, Random Forest).
4. Monitor fairness and bias (e.g., by location or employment type).
5. Validate model using **stratified cross-validation** to ensure balanced delinquency representation.