Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of Exploratory Data Analysis (EDA) on this Delinquency Prediction Dataset is to understand the structure and quality of the data. It helps us explore the types of information available, such as customer demographics, credit usage, and payment behaviour. Through EDA, we can identify missing values, unusual data, and patterns that may affect model performance.

The main goals of EDA are to discover key patterns, spot potential issues, and extract useful insights from the data. This includes identifying missing or unusual values, examining relationships between different variables, detecting trends linked to delinquency, and deciding which features may be valuable for building predictive models. This step prepares the dataset for building accurate and reliable delinquency prediction models.

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 patterns include a wide range of ages, incomes, credit scores, and account tenures, with substantial variation in financial variables.
- Some notable issues are the presence of missing values in 'Income', 'Loan Balance', and occasionally 'Credit Score', which require attention through imputation or exclusion.
- Outliers appear in 'Credit Utilization', where some values exceed 1.0—an unusual scenario in typical credit data—along with possible entry inconsistencies in categorical fields like 'Employment Status'.
- Additionally, the dataset’s target variable ('Delinquent Account') is imbalanced, with far fewer delinquent cases, which could bias any predictive modeling. Addressing these data quality issues is essential before building reliable delinquency prediction models.
- The variables that most likely to predict delinquency: income, account tenure, credit score

Key dataset attributes:

- Number of records: [500]

- Key variables:

Variable	Type	Description
Customer_ID	Categorical	Unique customer identifier
Age	Numerical	Age in years
Income	Numerical	Annual income (currency not specified)
Credit_Score	Numerical	Standard credit score (typically ranges from 300-850)
Credit_Utilization	Numerical	Fraction of credit limit currently being used
Missed_Payments	Numerical	Number of missed payments
Delinquent_Account	Binary (0/1)	Target: 1 = Account became delinquent, 0 = Not delinquent
Loan_Balance	Numerical	Outstanding loan amount
Debt_to_Income_Ratio	Numerical	Ratio of total debt payments to income
Employment_Status	Categorical	Employment (EMP/self-employed/unemployed/etc.)
Account_Tenure	Numerical	Years the account has been open
Credit_Card_Type	Categorical	Type (Student/Standard/Gold/Platinum/Business)
Location	Categorical	City (e.g., Los Angeles, Chicago, New York, Houston, etc.)
Month_1 ... Month_6	Categorical	Payment status for each of previous 6 months (On-time/Late/Missed)

- Data types: [Categorical, Numerical, Binary, String/ID.]

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.

Mostly missing values are found in the columns like income, loan balance, credit score.

Key missing data findings:

- Variables with missing values:

Field	imputation	Reason
Income	Median	Better for numeric data and less sensitive to outliers.
Loan balance	Median	Data will be stable and less sensitive to outliers
Credit score	Mode / median	Small percent of data is missing. mode would be better

- Missing data treatment: [Deletion, Imputation, Synthetic Data, etc.]

- Median imputation is recommended for missing Income and Loan Balance due to skewness and outliers.
- Credit Score missing values can be filled with median or mode since missingness is minimal.
- Adding binary missingness flags for each imputed feature helps models learn patterns in missing data.
- Avoid dropping rows with missing values to prevent bias and data loss.
- For categorical variables (none missing here), use mode or “Unknown” if needed.

Generating synthetic income values for missing entries using normal distribution. After imputation, validate that values remain realistic and model performance improves

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:

- **Correlations observed between key variables**:

- **Credit card utilization and missed payments**: if the utilization of credit card is high then the risk of missing payments is high and therefore the risk of delinquency.

- **Missed payments and delinquency**: if missed payments are frequent, then it could be the indication of delinquency.
- **Credit score and delinquency**: though it may be weak, it is also factor that could be the risk of delinquency
- **Debt-income ratio and delinquency:** if income is low but debt is high, then it could be a factor of delinquency risk

High credit card utilization strongly leads to risk of delinquency and remaining three variables show weak correlation with the delinquency risk factor.

- **Unexpected anomalies:**

- Some delinquent customers show zero missed payments, suggesting missing or hidden risk factors.
- Delinquency rates peak at five missed payments then drop, indicating a nonlinear or unexpected pattern.
- Credit score and debt-to-income ratio have almost no direct association with delinquency in this data.

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.'

6. Conclusion & Next Steps

Summarizing the key points below:

- Missing data points: income, loan, credit score.
- Key patterns: high use of credit card, frequent missed payment leads to risk of delinquency and income to debt ratio can be factor too.
- Data anomalies: A notable anomaly is the presence of delinquent customers who report zero missed payments, indicating the possibility of missing behavioral risk factors or incomplete data.

Recommended steps for further investigation:

- Evaluate data quality for missed payments and delinquency classification to ensure all relevant factors are captured and correctly coded.
- Explore why delinquency rates drop among customers with the highest missed payment counts, checking for policy interventions or sample bias.
- Regularly update and monitor the imputed values to avoid introducing bias and validate model predictive performance on fresh data.
- Consider advanced modeling to capture subtle interactions and patterns missed by linear correlations.