**Exploratory Data Analysis (EDA) Summary**   
**Report**

# 1. Introduction

This report presents a comprehensive analysis of the provided delinquency prediction dataset, aiming to uncover key patterns and risk factors associated with customer delinquency. The primary goals are to assess the quality of the raw data, identify and address any missing or inconsistent information, and pinpoint the most influential variables that contribute to delinquency outcomes. Ultimately, this analysis seeks to establish a foundational understanding that can inform the development of accurate predictive models and enhance strategies for early intervention and risk management.

# 2. Dataset Overview

This dataset comprises 500 records, detailing various customer attributes relevant to delinquency prediction. It features a mix of numerical variable, alongside categorical variables and monthly payment statuses. Notable anomalies included an outlier in ***Credit utilization*** (a value slightly exceeding 1.0) and a significant class imbalance within the ***Delinquent Account*** target variable, where only 16% of accounts were marked as delinquent. Inconsistencies were also identified in the ***Employment Status*** field, which contained varied spellings for the same category, although no duplicate records were found in the dataset.

Key dataset attributes:

Number of records: 500 records

|  |  |  |
| --- | --- | --- |
| Key attributes | Description | Data Type(s) |
| Customer\_ID | Unique identifier for customer | Categorical  (Object) |
| Age | Customer’s age in years | Numerical  (Integer) |
| Income | Annual income of the customer in USD | Numerical  (Float) |
| Credit\_Score | Customer’s credit score, typically ranging from 300-850 | Numerical  (Float) |
| Credit\_Utilization | Percentage of available credit currently in use. | Numerical  (Float) |
| Missed\_Payments | Total number of missed payments in the past 12 months | Numerical  (Integer) |
| Delinquent\_Acount | Indicator of whether the customer has a delinquent account | Numerical  (Integer) |
| Loan\_Balance | Total outstanding loan balance in USD | Numerical  (Float |
| Debt\_to\_Income\_Ratio | Ratio of total debt to income, expressed as a percentage | Numerical  (Float) |
| Employment\_Status | Current employment status | Categorical  (Object) |
| Account\_Tenure | Number of years of the customer has had an active account | Numerical  (Integer) |
| Credit\_Card\_Type | Type of credit card held | Categorical  (Object) |
| Location | Customer region or city of residency | Categorical  (Object) |
| Month\_1 to Month\_6 | Payment history over the past 6 months | Categorical  (Object) |

# 3. Missing Data Analysis

Missing values were identified in ***Income***, ***Loan\_Balance*** and ***Credit\_Score*** fields. The approach involved a targeted imputation strategy: median imputation for ***Credit\_Score*** due to its minimal missingness, and a more comprehensive regression imputation combined with missing indicator variables for ***Income*** and ***Loan\_Balance*** to preserve data relationships and capture any inherent predictive signals form the missingness itself.

Key missing data findings:

|  |  |  |
| --- | --- | --- |
| Variables with missing values | Missing data treatment | Counted missing values |
| Income | Regression-Based with Missing Indicator | 39 missing values |
| Loan\_Balance | Regression-Based with Missing Indicator | 29 missing values |
| Credit\_Score | Median | 2 missing values |

# 4. Key Findings and Risk Indicators

The analysis of the dataset revealed several significant correlations between key variables and delinquency outcomes. The strongest individual predictor identified was ***month\_4*** payment status, indicating that recent payment behavior is a primary driver of delinquency risk. Other notable correlations include ***Income*** and ***Account Tenure***, suggesting their roles in financial stability and customer loyalty. Furthermore, expected patterns were observed where ***Unemployed*** individuals and those with ***High Debt-to-income Ratios*** showed elevated delinquency rates, while higher ***Credit Utilization*** generally correlated with increased risk.

Unexpected anomalies:

* ***Missed\_Payments*** Count Disrepancy: Accounts with zero missed payments sometimes exhibited a higher delinquency rate than those with a few missed payments, suggesting that the precise sequence or severity payment issues, rather than just the count, might be more critical.
* Anomalous High-End Bins for ***Credit\_Utilization*** and ***Debt\_to\_Income\_Ratio***: Account with credit utilization exceeding 100% and Debt to income ratio above 50% paradoxically showed 0% delinquency.
* Counter-Intuitive ***Income*** and ***Credit\_Score*** Patterns: The highest delinquency rates were observed in the third quartile of income rather than the lowest, and customers with “Good” credit score showed a slightly higher delinquency rate than those with “Poor” or “Fair” scores.

# 5. AI & GenAI Usage

This section documents the iterative process of generating insights, highlighting the AI's analytical contributions and the specific prompts utilized to guide the investigation. The insights presented throughout this report, from initial data quality observations to the identification of risk factors and recommended strategies, were developed through a collaborative approach, leveraging AI capabilities for efficient data processing and pattern recognition.

Below are three key prompts that guided the generation of the analytical results in this report:

**Prompt 1:** **Initial Data Overview & Problem Identification**

"Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency."

(This prompt initiated the data exploration phase, helping to understand the fundamental characteristics and challenges of the dataset.)

**Prompt 2:** **Top Predictor Identification**

"Identify the top 3 variables most likely to predict delinquency based on this dataset. Provide brief reasoning."

(This prompt focused the analysis on identifying the most impactful features for the predictive modeling objective.)

**Prompt 3:** **Pattern & Risk Factor Detection**

"Detect pattern and risk factors, your next goal is to uncover patterns and key risk factors that influence delinquency. Action: List high-risk indicators, each with a one-sentence explanation of why it’s important, as well as any insights that could impact delinquency prediction."

(This prompt drove the in-depth analysis of variable relationships with delinquency, leading to the identification of critical risk indicators and unexpected trends.)

# 6. Conclusion & Next Steps

This initial analysis of the delinquency prediction dataset, comprising 500 records with a mix of numerical and categorical attributes, had provided crucial insights into data quality, patterns and risk factors. Key findings include notable missing values in ***Income***, ***Loan\_Balance***, and ***Credit\_Score***, inconsistencies in ***Employment\_Status***, and a critical class imbalance in the Delinquent\_Account target variable. High-risk indicators such as recent missed/late payments, unemployed status and high debt to income ratios were identified as significant predictors of delincquency rates at the extremes of ***Credit\_Utilization*** and ***Debt-to-Income\_Ratio***, and surprising patterns in Income and Credit\_Score correlations, highlight areas requiring deeper investigation.

Recommended Next Steps:

1. Data Preprocessing and Cleaning:

* Implement the proposed imputation strategy for missing values in Income, Loan Balance and Credit Score.
* Standarize and consolidate categorical features like Employment\_Status to ensure consistency.
* Investigate and decide on a strategy for handling the identified outliers, particularly the Credit Utilization value exceeding 1.0 and the unexpected 0% delinquency rates in extrem bins of Credit Utilization and Debt-to-income ratio.

1. Feature Engineering:

* Develop more sophisticated features from existing data, such as aggregating monthly payment statuses into a comprehensive payment behavior score or creating interaction terms between highly predictive variables.

1. Advanced Analysis of Anomalies:

* Conduct a deeper dive into the unexpected findings regarding Income, Credit Score, and the anomalous high-end bins for Credit Utilization and Debt-to-income ratio to understand their root causes and inform appropriate data handling or model interpretation.

1. Model Development and Evaluation:

* Select and train appropriate machine learning models for binary classification, focusing on algorithms robust to class imbalance (e.g., Gradient Boosting Machines, Random Forests).
* Employ strategies to address the class imbalance, such as oversampling, undersampling, or utilizing class weights within the chosen models.
* Evaluate model performance using metrics suitable for imbalanced datasets (e.g., AUC-ROC, Precision-Recall curves, F1-Score, Precision, Recall) to ensure accurate assessment of delinquenct prediction capabilities.