Rounak Saha

Exploratory Data Analysis (EDA) Summary Report

1. Introduction

This report summarizes the exploratory data analysis (EDA) conducted on the Delinquency_prediction_dataset.csv. The primary purpose is to understand the dataset's structure, identify key variables, assess data quality (including missing values), and uncover significant patterns and risk indicators related to customer delinquency. The insights gained will serve as a foundation for subsequent predictive modeling efforts.

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 dataset attributes:

- Number of records: The dataset has 500 rows and 19 columns.
- Customer_ID: Unique identifier for each customer. (Categorical)
- Age: Customer's age in years. (Numerical)
- Income: Annual income of the customer in USD. (Numerical, may contain missing values)
- Credit_Score: Customer's credit score, typically ranging from 300 to 850. (Numerical)
- Credit_Utilization: Percentage of available credit currently in use. (Numerical, 0-100%)
- Missed_Payments: Total number of missed payments in the past 12 months. (Numerical)
- Delinquent_Account: Indicator of whether the customer has a delinquent account. (Binary: 0=No, 1=Yes)
- Loan_Balance: Total outstanding loan balance in USD. (Numerical)
- Debt_to_Income_Ratio: Ratio of total debt to income, expressed as a percentage. (Numerical, 0-100%)
- Employment_Status: Current employment status (e.g., 'Employed', 'Unemployed', 'Self-Employed'). (Categorical)

- Account_Tenure: Number of years the customer has had an active account. (Numerical)
- Credit_Card_Type: Type of credit card held (e.g., 'Standard', 'Gold', 'Platinum'). (Categorical)
- Location: Customer's region or city of residence. (Categorical)
- Month_1 to Month_6: Payment history over the past 6 months: 0 = On-time, 1 = Late, 2 = Missed. (Categorical)
- Anomalies/Inconsistencies: Missing values were identified in several key columns. No other significant anomalies or duplicates were explicitly detected during this initial review. These gaps could skew model predictions if not properly addressed.

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.

Key missing data findings:

- Variables with missing values:
    - **Income:** 39 missing values out of 500 entries.
    - **Credit_Score:** 2 missing values, minimal but worth noting for a complete financial profile.
    - **Loan_Balance:** 29 missing values, potentially impacting loan-related analyses.
- Missing data treatment:

Table of Missing Data Issues and Handling Methods

Feature	Missing Data Count	Handling Method	Justification
Income	39	MICE	MICE accounts for the uncertainty in imputed values and preserves relationships between variables, which is crucial for complex financial data that may have skewed distributions or interdependencies
Credit_Score	2	Impute with median	Median is robust to outliers and ensures fairness in income distribution.

Table of Missing Data Issues and Handling Methods			
Feature	**Missing Data Count**	**Handling Method**	**Justification**
Loan_Balance	29	MICE	MICE accounts for the uncertainty in imputed values and preserves relationships between variables, which is crucial for complex financial data that may have skewed distributions or interdependencies

Advantages of MICE:

- **Accounts for Uncertainty:** MICE generate multiple plausible values for each missing entry, reflecting the inherent uncertainty in the imputation process.
- **Preserves Relationships:** Since MICE uses other variables in the dataset to predict missing values iteratively, it maintains the underlying correlations between variables.
- **Versatility:** It can handle different types of variables (continuous, categorical) in a dataset simultaneously.

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 Score & Income with Delinquency Risk:** The correlation between customer income and delinquency risk is a very weak positive value of ~ (0.04), indicating that income alone has minimal impact on predicting delinquency. This counter-intuitive finding suggests that higher income does not necessarily translate to lower risk, potentially due to factors like larger financial obligations or spending habits among higher-income groups. The weak relationship highlights that income is not a strong standalone predictor of delinquency in this dataset and suggests the need to explore interactions with other variables, such as credit utilization or employment status, to better understand the factors influencing delinquency risk.
    - **Missed Payments with Delinquency Risk:** The correlation between missed payments and delinquency risk is very weak and negative (-0.03), suggesting that

more missed payments are slightly associated with a lower likelihood of delinquency, which is counter intuitive. This unexpected finding may stem from complexities in the data, such as the Delinquent Account flag representing severe cases rather than general missed payments or the influence of other overshadowing factors. The relationship may also be non-linear, requiring deeper investigation. Missed payments alone do not strongly predict delinquency risk, emphasizing the need to explore interactions with other variables or validate the measurement definitions in the dataset.

- Top 3 Risk Factors Associated with Delinquency

    - **Credit Card Type – Business Cards (Delinquency Rate: 21.29%)**
        - **Insight:** Accounts associated with business credit cards show the highest delinquency rate. This suggests that individuals using business cards might have higher financial exposure or variable income patterns that increase the likelihood of delinquency. It may also reflect financial risks inherent to entrepreneurial ventures or small businesses.
    - **Employment Status – Unemployed (Delinquency Rate: 19.35%)**
        - **Insight:** As expected, unemployed individuals exhibit a higher delinquency rate. This emphasizes the importance of stable income in managing financial obligations and highlights unemployment as a significant risk factor for delinquency.
    - **Location – Los Angeles (Delinquency Rate: 19.62%)**
        - **Insight:** Customers residing in Los Angeles have a notably higher delinquency rate compared to other regions. This could be attributed to regional economic factors, higher cost of living, or localized challenges in financial stability. It may also point to demographic or socioeconomic patterns specific to the area.

- Customers with more than 4 missed payments and credit utilization above 50% exhibit a significantly higher delinquency rate of 20.63%. This indicates that these two factors are strong predictors of delinquency risk. Identifying and monitoring such high-risk customers can enable proactive interventions, such as tailored repayment plans or financial counseling, to mitigate the likelihood of default and enhance overall credit portfolio health.
- **Trends in Late Payments**: The number of 'Late' payments remained relatively consistent across the six observed months:

    - Month 1: 159
    - Month 2: 173
    - Month 3: 169
    - Month 4: 181
    - Month 5: 151

- Month 6: 172

This indicates that a significant portion of customers consistently exhibit late payment behavior, with Month 4 showing the highest count of late payments.

- Unexpected anomalies:

  - The very weak correlation between Missed_Payments and Delinquent_Account, and the slightly higher average Income and Credit_Score for delinquent accounts compared to non-delinquent ones, were unexpected. These suggest that the definition or criteria for 'Delinquent_Account' might be nuanced, or that other interacting factors are at play, warranting further investigation into these relationships.
  - Customers with 0 Account_Tenure and multiple missed payments.

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:

- **Data Summary and Missing Values:**
  - "Analyze this dataset and provide a summary of key columns, including common patterns and missing values."
  - "Identify missing values in this dataset and recommend the best imputation strategy based on industry best practices."
- **Correlation and Predictive Analysis:**
  - "Analyze the correlation between customer income, Credit Score and delinquency risk, summarizing key findings in simple terms."
- **Late Payment Trends:**
  - "Analyze trends in late payments and identify recurring patterns that could predict future delinquencies."
- **Risk Factor Identification:**
  - "Identify and quantify the top 3 risk factors associated with delinquency based on the dataset."
- **Advanced Imputation Strategies:**
  - "Recommend advanced techniques, such as MICE, for imputing missing numerical data in the dataset based on industry standards."

6. Conclusion & Next Steps

The exploratory data analysis (EDA) highlights critical insights into the delinquency dataset, emphasizing the importance of addressing missing data, exploring variable interactions, and focusing on specific high-risk groups. Key findings include:

- **Top Risk Factors:** Business credit card holders, unemployed individuals, and residents of Los Angeles exhibit the highest delinquency rates, requiring focused interventions.
- **Missed Payments and Credit Utilization:** Customers with over 4 missed payments and credit utilization above 50% are at significantly higher risk, reinforcing the need for targeted monitoring.
- **Unexpected Patterns:** Weak or counter-intuitive correlations between variables like missed payments, income, and delinquency risk suggest non-linear relationships or nuances in dataset definitions.
- **Consistent Late Payment Trends:** A significant portion of customers consistently exhibit late payments, indicating behavioral patterns that may require customer education or financial support initiatives.

Recommendations

- **Data Quality Improvements:**
  - Continue using MICE and robust methods to handle missing data while ensuring consistency in variable definitions to avoid misinterpretation.
  - Reassess the definition of "Delinquent Account" to clarify unexpected correlations.
- **Risk Mitigation Strategies:**
  - Design tailored repayment plans or financial counseling for high-risk groups, such as business card holders, unemployed individuals, and customers in Los Angeles.
  - Implement automated alerts for customers with high credit utilization or frequent missed payments to encourage timely repayments.
- **Behavioral Interventions:**
  - Analyze late payment behaviors in-depth and develop campaigns to incentivize timely payments, such as discounts, fee waivers, or rewards.
- **Predictive Model Enhancements:**
  - Incorporate interactions between variables (e.g., income and credit utilization) into predictive models to better identify at-risk customers.
  - Use advanced machine learning techniques to capture non-linear patterns and improve prediction accuracy.

Next Steps

- **Refine Data Analysis:**
  - Conduct further exploratory analysis on unexpected findings, such as the weak correlation between missed payments and delinquency risk.
  - Perform additional validation of data definitions to ensure clarity and accuracy.
- **Model Development:**
  - Build and test predictive models, incorporating key findings and interactions, to proactively identify high-risk customers.
  - Evaluate model performance and fine-tune parameters for improved precision.
- **Policy and Program Implementation:**
  - Roll out intervention programs for high-risk customers identified through analysis.
  - Introduce a pilot program to reward consistent on-time payments and assess its impact on delinquency rates.
- **Ongoing Monitoring and Reporting:**
  - Establish a framework for continuous monitoring of delinquency rates, trends in late payments, and the impact of intervention strategies.
  - Regularly update and refine predictive models as new data becomes available.