Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this report is to perform an Exploratory Data Analysis (EDA) on a customer-level dataset provided for delinquency prediction. The primary goal is to understand the structure, completeness, and behavioral patterns within the data that may be linked to loan delinquency. This includes identifying key risk indicators, assessing data quality, and providing insights that will inform downstream predictive modeling efforts for Tata iQ and Geldium.

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: 250
- **Key variables:**
    - Age: Customer's age (Numerical)
    - Income: Annual income (Numerical)
    - Credit_Score: Customer’s credit score (Numerical)
    - Credit_Utilization: Proportion of credit used (Numerical)
    - Missed_Payments: Number of historical missed payments (Numerical)
    - Loan_Balance: Outstanding loan balance (Numerical)
    - Debt_to_Income_Ratio: Financial leverage ratio (Numerical)
    - Employment_Status: Employment type (Categorical)
    - Account_Tenure: Number of years with the lender (Numerical)

    - Delinquent_Account: Target variable indicating if the customer is delinquent (0 = No, 1 = Yes)
- **Data types:**
    - Numerical: Age, Income, Credit Score, Loan Balance, Account Tenure, Missed Payments, Credit Utilization, Debt-to-Income Ratio
    - Categorical: Employment_Status, Credit_Card_Type, Location
    - Binary: Delinquent_Account

Observations:

- The dataset appears clean with no duplicates.
- Only minor missing values are found in Income, Loan_Balance, and Credit_Score.
- No outliers or inconsistencies were found during the initial scan, but further validation is recommended during modeling.

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
    - Loan_Balance: 29 missing values
    - Credit_Score: 2 missing values
- **Missing data treatment:**
    - For **Income** and **Loan_Balance**, which have a moderate number of missing values, **median imputation** is recommended to preserve data consistency without introducing bias from outliers.

- For **Credit_Score**, due to the very low number of missing entries, **row deletion** or **median imputation** would be equally acceptable, with minimal impact on model training.
- All other variables are complete, requiring no intervention.

These strategies ensure minimal information loss while maintaining the integrity and predictive potential of the dataset.

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:

- **Delinquency Imbalance**: 84% of customers are non-delinquent, and 16% are delinquent. This class imbalance should be accounted for during model training.
- **Correlations observed:**
  - Weak positive correlations exist between delinquency and features such as Income, Credit_Score, Credit_Utilization, and Debt_to_Income_Ratio.
  - Account_Tenure shows a mild negative correlation, suggesting that customers with shorter account histories may have a higher likelihood of becoming delinquent.
- **Behavioral Indicators**:
  - Delinquent customers tend to have slightly higher income and credit scores, which is somewhat counterintuitive.
  - Higher Credit_Utilization and Debt_to_Income_Ratio among delinquent accounts may indicate over-leveraging.

Unexpected anomalies:

- Contrary to expectations, delinquent customers have slightly **lower average missed payments** than non-delinquent customers, which may indicate other underlying behavioral or policy-related factors not captured in the dataset.
- The low correlation values suggest that delinquency may be driven by a **combination of weak signals across multiple variables**, rather than by any single dominant factor.

These findings highlight the importance of combining multiple features in modeling and the potential need for feature engineering or temporal analysis to better detect early signs of delinquency.

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.

Use of AI Tools:

Generative AI (ChatGPT) was leveraged to:

- Interpret variable relationships and statistical patterns in the data
- Recommend strategies for handling missing values
- Assist in identifying early indicators of risk based on grouped means and correlation analysis

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.”
- “Interpret differences in average values between delinquent and non-delinquent customers.”

- “Which features show correlation with a target variable like Delinquent_Account?”

This assisted approach ensured that insights were both data-driven and professionally worded, allowing for efficient, high-quality reporting aligned with industry standards.

6. Conclusion & Next Steps

This exploratory data analysis provided valuable insights into the structure, quality, and predictive relevance of the delinquency prediction dataset. The data is generally clean and well-prepared, with only minor missing values and no critical anomalies.

Key findings:

- The target variable Delinquent_Account is imbalanced, with only 16% delinquent cases.
- Credit utilization, debt-to-income ratio, and account tenure show modest relationships with delinquency risk.
- Some patterns—such as higher income and credit score among delinquent customers—are counterintuitive and may warrant deeper investigation.
- No strong linear correlations exist, suggesting that delinquency prediction will require multi-feature modeling.

Next steps:

- Apply data preprocessing: Impute missing values and scale numerical features.
- Address target imbalance using resampling techniques or weighted loss functions.

- Perform feature engineering, especially around time-based patterns in Month_1 to Month_6.
- Begin model development using classification algorithms like logistic regression, random forest, or gradient boosting.
- Evaluate performance using appropriate metrics (e.g., precision, recall, AUC) given the class imbalance.

This analysis forms a strong foundation for building a reliable delinquency prediction model to support Tata iQ and Geldium's decision-making process.