Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

Purpose for this report is to detect pattern, inconsistent data , missing value in delinquency_prediction_dataset sheet and inform the correct pattern value should be placed at the relatable position for correct accuracy in dataset.

2. Dataset Overview

I have summarized 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: There is 500 customer record datasets.
- Column Descriptions
  Customer_ID
  (Unique identifier for each customer.)
  Age
  (Customer's age in years)
  INCOME
  (Annual income of the customer in USD.)
  Credit_Score
  (Customer's credit score, typically ranging from 300 to 850.)
  Credit_Utilization
  (Percentage of available credit currently in use.)
  Missed_Payments
  (Total number of missed payments in the past 12 months.)
  Delinquent_Account
  (Indicator of whether the customer has a delinquent account.)
  Loan_Balance
  (Total outstanding loan balance in USD.)
  Debt_to_Income_Ratio
  (Ratio of total debt to income, expressed as a percentage.)
  Employment_Status
  (Current employment status (e.g., 'Employed', 'Unemployed', 'Self-Employed').)
  Account_Tenure
  (Number of years the customer has had an active account.)
  Credit_Card_Type
  (Type of credit card held (e.g., 'Standard', 'Gold', 'Platinum').)
  Location

(Customer's region or city of residence.)
Month_1 to Month_6
(Payment history over the past 6 months: 0 = On-time, 1 = Late, 2 = Missed.)

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: only three column having missing value:

- INCOME:39
- LOAN_Balance:29
- Credit_Score:2

- Missing data treatement:

- Use **median imputation** for Income and Loan_Balance.
- Use **median** (or mean) for Credit_Score

4. Key Findings and Risk Indicators

I have 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:

Income vs. Debt-to-Income Ratio (Negative correlation)

- Higher **Income** usually lowers the **Debt-to-Income Ratio (DTI)**.
- Borrowers with high income relative to debt are at lower risk.

Loan_Balance vs. Debt-to-Income Ratio (Positive correlation)

- Higher **Loan_Balance** increases **DTI**.
- Larger outstanding loans make repayment harder.

Credit_Utilization vs. Credit_Score (Negative correlation)

- High **Credit Utilization** (using more of available credit) usually lowers **Credit Score**.
- Over-utilization is a strong risk signal.

Missed_Payments vs. Credit_Score (Negative correlation)

- More **Missed Payments** → lower **Credit Score**.
- One of the strongest delinquency indicators.

Credit_Utilization vs. Missed_Payments (Positive correlation)

- High **Credit Utilization** tends to align with more **Missed Payments**.

- Unexpected anomalies:

income anomalies

- Very **high income with very high DTI ratio** → suggests income is incorrectly reported or expenses are extreme.
- **Zero or negative income** entries → need validation

Credit Score anomalies

- Scores outside the valid range (300–850).
- **High Credit Score + many missed payments** → inconsistent pattern worth checking.

Loan_Balance anomalies

- **Extremely high balances** compared to average income → may distort models.
- **Zero balance but still high utilization or missed payments** → inconsistent.

Credit Utilization anomalies

- Should be between **0 and 1 (0%–100%)**.
- Values >1 (over 100%) → borrower exceeding credit limit, or data entry error.

Missed Payments anomalies

- Very high counts (e.g., >12 in 6 months) → unrealistic, possible data issue.
- **Zero missed payments + delinquent status = "Yes"** → mismatch.

Debt-to-Income Ratio anomalies

- Negative or >1.0 (i.e., >100%) → indicates error or unusual case.

Employment Status mismatches

- Marked as *"Unemployed"* but **high income/loan balance**.
- *"Student"* with high loan balances → may need closer inspection.

5. AI & GenAI Usage

I have used Chatgpt , copilot and gemini ai tool for work ,I have used these tools were used to summarize the dataset, impute missing data, and detect patterns. This have used documents AI-generated insights and the prompts used to obtain results.

Example AI prompts used:

- "Analyze this dataset and provide a summary of key columns ,including common patterns and missing values."
- 'Identify missing values in dataset and recommend the best imputation strategy based on industry best practices'
- 'Analyze the correlation between customer income and delinquency risk , summarizing key finding in simple terms'

6. Conclusion & Next Steps

Key Findings

- **Missing Values:** Found in three columns — *Income (7.8%)*, *Loan_Balance (5.8%)*, and *Credit_Score (0.4%)* — requiring imputation.
- **Correlations:**
    - **Negative:** Income vs. Debt-to-Income Ratio, Credit Score vs. Credit Utilization.
    - **Positive:** Loan Balance vs. DTI, Credit Utilization vs. Missed Payments.
    - Strongest delinquency risk factors → **High Missed Payments, High Utilization, High DTI with Low Income, and Low Credit Score.**
- **Unexpected Anomalies:**
    - Out-of-range values (e.g., Credit Utilization > 1, invalid Credit Scores).
    - Inconsistent patterns (e.g., High Credit Score but many missed payments).
    - Employment/income mismatches (e.g., Unemployed with high income).

Recommended Next Steps

1. **Data Cleaning & Imputation**
    - Apply median/KNN imputation for missing *Income, Loan_Balance, Credit_Score*.
    - Flag or correct anomalies (negative/invalid values, utilization > 100%).
2. **Feature Engineering**
    - Create risk indicators: *High Utilization Flag, High DTI Flag, Delinquency Risk Score*.
    - Normalize/scale numeric features for model input.
3. **Exploratory Modeling**
    - Use **logistic regression, decision trees, or random forest** to predict delinquency probability.
    - Compare accuracy across models.
4. **Further Investigation**
    - Validate suspicious data entries with source system.
    - Segment customers by risk levels (Low, Medium, High).