Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The purpose of this report is to perform an Exploratory Data Analysis (EDA) on the Delinquency Prediction Dataset as part of the GEA-AI Internship project under TCS and Geldium. The primary goal is to understand key patterns, detect anomalies, and identify potential risk indicators associated with customer delinquency, which will later support predictive modeling.

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: 500

- Key variables:

- ✓ Customer_ID: Unique customer identifier
- ✓ Age: Customer's age
- ✓ Income: Monthly income
- ✓ Credit_Score: Credit rating score
- ✓ Credit_Utilization: % of credit used
- ✓ Missed_Payments: Count of missed payments
- ✓ Loan_Balance: Current loan amount
- ✓ Debt_to_Income_Ratio: Financial pressure indicator
- ✓ Employment_Status, Credit_Card_Type
- ✓ Location: Categorical variables
- ✓ Account_Tenure: Age of account in months
- ✓ Delinquent_Account: Binary outcome
- ✓ Month_1 to Month_6: Monthly payment behavior

- Data types:

- ✓ **Categorical (object):** Employment_Status, Credit_Card_Type, Location, Month_1 to Month_6, Credit_Utilization, Loan_Balance, Age

- ✓ **Numerical (float64):** Income, Credit_Score, Missed_Payments, Debt_to_Income_Ratio, Delinquent_Account, Account_Tenure

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, Credit Score, Loan Balance

- Missing data treatment: Median imputation for numeric columns

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: Showed weak to moderate relationships among variables.

- Unexpected anomalies:

- ✓ Some rows with 0 credit utilization and high loan balance
- ✓ Object types for Age, Credit_Utilization, Loan_Balance

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.'

AI prompts used:

- ✓ "Summarize key patterns in delinquency dataset"
- ✓ "Suggest imputation strategies for financial variables"

6. Conclusion & Next Steps

EDA uncovered missing values, incorrect data types, correlations, and key delinquency indicators.

- ✓ Convert object columns to appropriate numeric types
- ✓ Feature engineering
- ✓ Model training and validation