Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

The purpose of this report is to analyze the data quality and structure of Geldium’s customer dataset to assess its readiness for predictive modeling. The primary goal is to **identify missing values, detect anomalies, and uncover early risk indicators** that will inform the development of an AI-powered delinquency risk model. This analysis ensures that subsequent predictions are built on a foundation of reliable, clean, and ethically processed data.

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 rows.
- Key variables: Annual_Income, Credit_Score, Debt_to_Income_Ratio, Credit_Utilization_Ratio, Late_Payment_Count, Delinquent (Target Variable).
- Data types:

Numerical: Age, Annual_Income, Credit_Score, Loan_Balance, Utilization Ratios.

Categorical: Employment_Status, Location, Credit_Card_Type.

- **Anomalies Observed:**

- **Credit Score Inversion:** A counter-intuitive pattern was detected where customers marked as "Delinquent" have a slightly *higher* average Credit Score (591) than non-delinquent customers (575). This contradicts standard financial logic and requires investigation into data extraction methods.

- **Payment History Logic:** Customers listed with "On-time" payments in Month_1 showed a higher delinquency rate (19.7%) compared to those marked as "Late" (11.3%), indicating potential labeling errors.

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:

Annual_Income: 39 missing records (approx. 8%).

Loan_Balance: 29 missing records.

Credit_Score: 2 missing records.

- Missing data treatment:

Annual_Income: Imputation (Median).

- *Justification:* Income data is often skewed by high earners. Using the median prevents outliers from distorting the "typical" income value, providing a more accurate baseline for Debt-to-Income calculations.

Loan_Balance: Imputation (Median).

- *Justification:* Similar to income, loan balances can vary wildy. Median imputation preserves the central tendency of the debt profile without reducing the sample size.

Credit_Score: Removal (Row Deletion).

- *Justification:* Since only 2 records are missing (0.4%), removing them has a negligible impact on sample size and avoids introducing noise into a highly sensitive variable.

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 Utilization Ratio: This is the strongest early indicator. Customers who eventually default have a higher average utilization (50.7%) compared to those who do not (48.8%).

Debt-to-Income (DTI) Ratio: Higher DTI ratios show a positive correlation with delinquency, indicating that customers with higher debt burdens relative to their income are at greater risk.

- Unexpected anomalies:

  - **Age Distribution:** The dataset contains no significant age outliers (Max age 74), but Age showed a weak correlation with delinquency, suggesting that younger vs. older demographics behave similarly in this specific sample.

  - **High Credit Score Defaults:** The presence of high credit scores among delinquent accounts suggests that Credit_Score alone may not be a reliable predictor for this specific batch, forcing the model to rely more heavily on behavioral data like Utilization and DTI.

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.

- *"Analyze this dataset to identify the top 3 variables most likely to predict delinquency and highlight any logical inconsistencies in the Payment History columns."*
- *"Suggest an imputation strategy for missing 'Annual_Income' values in a financial risk dataset, considering the potential for skewed distribution."*
- *"Generate a summary of risk indicators, specifically comparing the mean values of Credit Utilization between delinquent and non-delinquent customers."*

6. Conclusion & Next Steps

The exploratory analysis confirms that while the Geldium dataset contains valuable behavioral signals—specifically **Credit Utilization** and **Debt-to-Income Ratio**—it suffers from significant data integrity issues. The logical contradictions in Credit Scores and Payment History suggest potential system errors that could confuse a predictive model.

Recommended Next Steps:

1. **Data Validation:** Consult with Geldium's IT/Data engineering team to clarify the "Credit Score Inversion" and Payment History labeling errors.

2. **Imputation:** Apply Median Imputation to Annual_Income and Loan_Balance to resolve missing values.

3. **Feature Engineering:** Prioritize Credit_Utilization and DTI as primary features for the predictive model, given their reliability over Credit Score in this specific dataset.

4. **Model Selection:** Proceed with Random Forest or Gradient Boosting models, which can better handle the non-linear relationships observed in the data.