Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this report is to summarize the initial exploratory data analysis performed on the customer financial dataset. The primary goal is to identify data quality issues, anomalies, and key features that indicate a customer's risk of account delinquency. These findings will inform the necessary data preprocessing and feature engineering steps required to build a robust predictive model.

2. Dataset Overview

This section summarizes the dataset, including the number of records, key variables, and initial observations regarding data types and consistency.

Key dataset attributes (Before Preprocessing):

- **Number of records:** 500
- **Key variables:**
  - `Delinquent_Account`: Target binary variable (0=No, 1=Yes).
  - `Income`, `Credit_Score`, `Loan_Balance`: Core financial metrics.
  - `Month_1` to `Month_6`: Time-series payment history (string categories).
  - `Credit_Utilization`, `Debt_to_Income_Ratio`: Leverage indicators.
- **Data types:** Mix of Numerical (e.g., Age, Income), Binary (Delinquent_Account), and Categorical (e.g., Employment_Status, Month_1).

Anomalies and Inconsistencies:

- **Target Imbalance:** The most significant anomaly is the severe class imbalance in the target variable, with \$84\%\$ **Non-Delinquent** and only \$16\%\$ **Delinquent** accounts.
- **Categorical Inconsistencies:** The `Employment_Status` column contains spelling and case variations (e.g., 'EMP', 'employed', 'Employed') which require standardization.
- **Payment History Encoding:** The `Month_1` to `Month_6` columns are currently strings and must be converted to numerical ordinal values to reflect the increasing risk (On-time < Late < Missed).

3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring model accuracy. The following columns were found to have missing values.

Key missing data findings:

Variable	Count Missing	Percentage Missing
Income	39	7.8%
Loan_Balance	29	5.8%
Credit_Score	2	0.4%

Missing data treatment:

Variable	Chosen HandlingMethod	Justification
Income	**Median Imputation**	Median is robust against outliers and preserves the central tendency of the income distribution.
Loan_Balance	**Median Imputation**	It minimizes distortion from extreme loan balances, maintaining data realism for a financial feature.
Credit_Score	**Median Imputation**	Simple and safe for a small number of missing values; avoids skewing the score distribution.

4. Key Findings and Risk Indicators

This section identifies trends and patterns that may indicate risk factors for delinquency, providing insights relevant to predictive modeling.

Key findings:

- **Correlations observed between key variables:**
  - An expected inverse correlation exists between **Credit Score** and the likelihood of delinquency. Lower scores will strongly correlate with higher delinquency rates.
  - A high correlation is expected between the frequency of **Missed Payments** and **Total Payment Risk** (once engineered) with the `Delinquent_Account` status.
  - Higher **Credit Utilization** and **Debt-to-Income Ratio** are correlated with financial strain and, consequently, higher delinquency risk.
- **Unexpected anomalies:**
  - No statistically impossible numerical anomalies were found (e.g., negative income).
  - The maximum **Credit Utilization** reaching ≈ 1.03 is an extreme but valid data point, indicating severe over-leveraging in a small segment of the customer base.

High-Risk Indicators:

- **Payment History:** Any 'Missed' or repeated 'Late' status in the six-month history is the most direct and powerful signal of current financial instability.

- **Credit Score:** Scores in the "Poor" range ($300-579$) are highly predictive of future delinquency probability.
- **High Leverage:** Customers with high financial leverage indicators (high Credit Utilization or DTI) are strong candidates for high-risk flags.

5. AI & GenAI Usage

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns, accelerating the initial EDA phase.

Example AI prompts used:

- 'Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency.'
- 'Suggest an imputation strategy for missing income values based on industry best practices.'
- 'List high-risk indicators, each with a one-sentence explanation of why it's important, as well as any insights that could impact delinquency prediction.'

6. Conclusion & Next Steps

Summary: The dataset requires critical cleaning and engineering before modeling due to missing values, categorical inconsistencies, and the necessity of encoding payment history. The primary challenge is the severe 84% vs. 16% class imbalance in the target variable.

Recommended Next Steps:

1. **Execute Preprocessing:** Implement **Median Imputation** for missing values and **Standardization/One-Hot Encoding** for all categorical features (`Employment_Status`, `Credit_Card_Type`, etc.).
2. **Feature Engineering:** Create a single **Total Payment Risk Score** from the monthly payment data.
3. **Modeling Strategy:** Split the data, and use **SMOTE** on the training set to resolve class imbalance, followed by training a classification model (e.g., Random Forest).