**Exploratory Data Analysis (EDA)   
Summary Report**

# 1. Introduction

This exploratory data analysis (EDA) examines a credit customer dataset to assess its viability for predicting credit delinquency risk. The primary objective is to identify key patterns, data quality issues, and predictive indicators that will inform the development of an effective delinquency prediction model. Through comprehensive analysis of customer demographics, payment behaviors, and credit characteristics, this report aims to establish a foundation for accurate risk assessment and early intervention strategies.

# 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.

1. **Number of records: 500**
2. **Key columns: Variables that most likely affect high delinquency**

* Credit\_Utilization
* Missed\_Payments
* Months(Month\_1 to Month\_6)
* Debt\_to\_Income\_Ratio
* Employment\_Status
* Credit\_Score
* Loan\_Balance

1. **Column Datatypes:**

* **Numerical columns** - Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Delinquent\_Account, Loan\_Balance, Debt\_to\_Income\_Ratio, Account\_Tenure
* **Categorical columns** - Customer\_ID, Employment\_Status, Credit\_Card\_Type, Location, Month\_1, Month\_2, Month\_3, Month\_4, Month\_5, Month\_6

# 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:**

**I. Variables with missing values:**

|  |  |
| --- | --- |
| **Column/Variable name** | **No. of missing values** |
| Income | 39 |
| Loan\_Balance | 29 |
| Credit\_Score | 2 |

**II. Missing Data Handling Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Issue Description** | **Recommended Handling Method** | **Justification** |
| Income | missing values (~2-3% of records) | Impute (Predictive) | Income is a crucial predictive variable; removing records would lose valuable data. Predictive imputation preserves relationships with other variables like Age, Employment\_Status, and Location. |
| Loan\_Balance | missing values (~2-3% of records) | Impute (Median by Group) | Using the median value grouped by Credit\_Card\_Type or Employment\_Status is a robust and simple method that avoids the influence of outliers. |
| Credit\_Score | A single missing value (CUST0116) | Impute (Median) | With only one missing instance, removing the entire record is wasteful. The median is a stable, quick estimate that won't distort the overall distribution. |

# 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:**

1. **Payment Behavior Deterioration** - A clear pattern emerges where customers with multiple Missed\_Payments (e.g., 4+) and a Delinquent\_Account (flag = 1) almost universally show a history of "Late" or "Missed" payments in the Month\_1 to Month\_6 columns. This confirms that recent payment history is a powerful leading indicator.
2. **Credit Utilization as a Critical Signal** - High Credit\_Utilization (e.g., >0.7) is strongly correlated with poor payment behavior. Customers with utilization above 0.8 frequently exhibit "Late" or "Missed" payments, suggesting they are over-leveraged and struggling to meet obligations.
3. **Employment Status and Income Stability** - Customers marked as Unemployed or with missing Income data show a higher propensity for missed payments. Conversely, those labeled retired or Employed (standardized) show more stable payment patterns, highlighting the importance of income verification and stability.
4. **Debt-to-Income (DTI) Paradox** - While a high Debt\_to\_Income\_Ratio (>0.4) is often linked to delinquency, several records with a suspiciously perfect 0.1 ratio still show missed payments. This suggests that either: The 0.1 value is an imputation or placeholder for missing data, making the field unreliable for those records.Other factors like cash flow problems or unexpected expenses are the true drivers, not captured by DTI alone.
5. **Credit Score is Predictive but Not Perfect** - Low Credit\_Score values (e.g., <400) are a good general predictor of risk. However, there are notable exceptions where individuals with high credit scores (>750) still become delinquent, indicating that a good historical score does not guarantee future performance if current circumstances change (e.g., job loss).

**Anomalies:**

**The following data points represent significant anomalies that could skew model performance and should be investigated for data entry errors or unique edge cases:**

* **Credit\_Utilization** - Some values exceed 1.0 (e.g., CUST0090: 1.026, CUST0266: 1.025), which is unusual since credit utilization is typically a ratio between 0 and 1.
* **Credit\_Score** - A few very low scores (e.g., 301, 302) and very high scores (e.g., 847) may indicate data entry errors or extreme cases.
* **Debt\_to\_Income\_Ratio** - Values like 0.1 appear suspiciously low and may represent imputed values.
* **Employment\_Status** - Inconsistent capitalization (e.g., "employed", "Employed", "EMP", "Self-employed", etc.) — this should be standardized.
* **Age vs. Income** - Instances of very young customers (e.g., Age 19-22) with exceptionally high incomes (e.g., >$180,000). While possible, these are statistical outliers.

# 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, outliers, and missing values in this dataset. Highlight any fields that might present problems or anomalies for modeling delinquency. Identify the top variables most likely to predict early indication of delinquency risk based on this dataset. Provide brief reasoning.
* Create a simple table listing 2–3 missing data issues. For each one, include your chosen handling method and a one-line justification for why you selected it. Determine the best treatment approach for each case:

**Remove**: Drop columns with excessive missing data.

**Impute:** Fill in missing values using mean, median, or predictive modeling.

**Generate synthetic data:** Create realistic values while maintaining fairness and distribution patterns.

* Create a python file that applies our plan for handling missing data. Add comments for understandability. After transformation make sure to check for missing values if any. Export the transformed dataset as an Excel file.

# 6. Conclusion & Next Steps

## Key Findings Summary

The exploratory analysis reveals that this dataset is well-suited for credit delinquency prediction, with strong predictive signals identified across multiple dimensions. Payment behavior patterns (Missed\_Payments, monthly payment history), credit utilization levels, and employment stability emerge as the most reliable early indicators of delinquency risk. While the dataset contains manageable data quality issues—primarily missing values in Income (7.8%), Loan\_Balance (5.8%), and Credit\_Score (0.4%)—these can be effectively addressed without compromising model integrity.

Notable anomalies include credit utilization ratios exceeding 1.0, inconsistent employment status formatting, and some extreme outliers in the age-income relationship that warrant further investigation but do not invalidate the overall dataset quality.

## Recommended Next Steps

* **Data Transformation Implementation** - Apply the missing data handling plan using predictive imputation for Income, median-by-group imputation for Loan\_Balance, and simple median imputation for Credit\_Score. Standardize categorical variables and address identified anomalies.
* **Feature Engineering** - Create derived variables such as payment consistency scores, utilization trend indicators, and risk flags to enhance predictive power beyond the existing features.