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

# 1. Introduction

This report presents a comprehensive analysis of the **Delinquency Prediction Dataset**. The primary purpose of this analysis is to identify and summarize the key characteristics of the dataset's columns, with a specific focus on understanding data patterns, identifying missing values, and assessing data quality. The insights gained from this analysis will serve as a foundational step toward building a predictive model for account delinquency.

# 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.

**Dataset attributes:**

* **Number of records**: 500
* **Key variables**: The dataset contains 19 variables. The key variables for predicting delinquency include **Age**, **Income**, **Credit\_Score**, **Credit\_Utilization**, **Missed\_Payments**, and **Loan\_Balance**. The **Delinquent\_Account** column serves as the target variable for the analysis.
* **Data types**: The dataset contains a mix of data types:
  + **Numerical**: Age, Income, Credit\_Score, Credit\_Utilization, Missed\_Payments, Delinquent\_Account, Loan\_Balance, Debt\_to\_Income\_Ratio, and Account\_Tenure.
  + **Categorical**: Customer\_ID, Employment\_Status, Credit\_Card\_Type, Location, and the monthly payment status columns (Month\_1 to Month\_6).
* **Data anomalies**: The Employment\_Status column contains inconsistent values for "employed" (e.g., 'Employed', 'employed', 'EMP'), which requires standardization. There are no duplicate records.

# 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:

* -**Missing values**: Missing values were identified in three columns:
  + Income: 39 missing values
  + Loan\_Balance: 29 missing values
  + Credit\_Score: 2 missing values

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.

**Correlations observed between key variables:**

The analysis of the dataset revealed some weak positive correlations with the **Delinquent\_Account** variable, which may serve as risk indicators. However, it's important to note that all correlations are very low, suggesting that no single variable strongly predicts delinquency on its own.

* **Credit Utilization**: This variable has the highest positive correlation with delinquency (approximately **0.065**). A higher credit utilization ratio indicates a customer is using a larger portion of their available credit, which is a known risk factor.
* **Income**: There is a weak positive correlation (approximately **0.049**) between higher income and delinquency. While this seems counterintuitive, it might be an indication that higher-income individuals are taking on more credit and, therefore, more risk.
* **Credit Score**: Contrary to what might be expected, the correlation between a higher credit score and delinquency is very low and positive (approximately **0.030**). This suggests that a simple linear relationship doesn't exist in this dataset, and the model would likely need to incorporate non-linear relationships.
* **Other Variables**: All other numerical variables, including Age, Debt\_to\_Income\_Ratio, Loan\_Balance, and Missed\_Payments, show very weak to no correlation with Delinquent\_Account.

**Unexpected anomalies:**

* **Low Credit Scores**: The dataset contains an unusually low credit score of **302**. While such scores exist, they are at the very bottom of the FICO scale (300-850) and are worth investigating to ensure data integrity.
* **High Credit Utilization**: There are instances where the Credit\_Utilization is over 1.0 (e.g., **1.025**). This indicates that the customer has used more than their total credit limit, which is a significant red flag for financial distress and is a strong indicator of high risk.
* **Employment Status Inconsistency**: The same employment status is represented by different spellings and abbreviations (e.g., 'Employed', 'employed', 'EMP'), requiring data cleaning for accurate categorical analysis.

# 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:

- 'Analyze the data set and find the correlation between key variables.'

- 'Summarize the data set and mention the number of records, missing values, anomalies present.'

# 6. Conclusion & Next Steps

Based on the analysis, the dataset shows several key characteristics and potential risk indicators for a delinquency prediction model. While some expected correlations were weak, the findings provide a solid foundation for further development.

**Summary of Key Findings:**

* **Data Quality**: The dataset is of generally good quality, but it does contain missing values in the **Income**, **Loan\_Balance**, and **Credit\_Score** columns, as well as inconsistencies in the **Employment\_Status** column. These issues must be addressed before modeling.
* **Risk Indicators**: The **Credit\_Utilization** variable shows the strongest positive correlation with delinquency, indicating that this feature is a critical risk factor. The presence of very low credit scores and utilization ratios over 1.0 are also significant red flags.
* **Weak Correlations**: Many of the other numerical variables, such as **Age**, **Income**, and **Credit\_Score**, show very weak correlations with the target variable. This suggests that a simple linear model may not be sufficient and that a more complex model capable of capturing non-linear relationships will be needed.

**Recommended Next Steps:**

1. **Data Cleaning and Preprocessing**:
   1. **Handle Missing Values**: Impute missing values in the Income, Loan\_Balance, and Credit\_Score columns. A mean, median, or more sophisticated imputation method (e.g., K-nearest neighbors) should be considered.
   2. **Standardize Categorical Data**: Clean and standardize the Employment\_Status column to ensure consistency.
   3. **Feature Engineering**: Create new variables that could improve the model's predictive power. For example, a categorical variable for Credit\_Score ranges (e.g., 'Poor', 'Fair', 'Good', 'Excellent') might be more informative than the raw numerical value.
2. **Exploratory Data Analysis (EDA)**:
   1. **Visualize Relationships**: Create visualizations such as scatter plots and heatmaps to better understand the relationships between variables, especially between features and the target variable.
   2. **Analyze Anomalies**: Further investigate the data points with extremely low credit scores and high credit utilization to determine if they are valid outliers or data entry errors.
3. **Model Building**:
   1. **Choose a Model**: Select a machine learning model appropriate for classification, such as Logistic Regression, a Decision Tree, or a more advanced model like Gradient Boosting or a Neural Network, given the weak linear correlations.
   2. **Train and Evaluate**: Train the chosen model on the cleaned data and evaluate its performance using metrics like accuracy, precision, recall, and the F1-score to assess its ability to predict delinquency.