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

# 1. Introduction

The purpose of this report is to explore a customer financial dataset with the goal of identifying the most important factors contributing to **delinquency**. This EDA will support building a predictive model to help financial institutions detect high-risk customers early.

# 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:

* Age: Customer’s age (Numerical)
* Income: Annual income (Numerical)
* Credit\_Score: Numeric score indicating creditworthiness (Numerical)
* Credit\_Utilization: Ratio of credit used to credit limit (Numerical)
* Missed\_Payments: Number of missed payments in past 6 months (Numerical)
* Loan\_Balance: Total loan balance (Numerical)
* Debt\_to\_Income\_Ratio: Ratio of debt to income (Numerical)
* Employment\_Status: Employment condition (Categorical)
* Account\_Tenure: Number of years the customer has held the account (Numerical)
* Delinquent\_Account: Binary indicator of delinquency (Target variable

- Data types:

* **Numerical**: Age, Income, Credit\_Score, Credit\_Utilization, Loan\_Balance, etc.
* **Categorical**: Employment\_Status, Credit\_Card\_Type, Location, Month\_1 to Month\_6
* **Duplicates**: 0 duplicate records
* **Inconsistencies**: Credit\_Utilization exceeds 1.0 in some rows (max = 1.0258), which is unrealistic and should be capped or investigated.

# 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: 39 missing values
* Credit\_Score: 2 missing values
* Loan\_Balance: 29 missing values

- Missing data treatment:

* Income: Impute using median or use predictive imputation based on Employment\_Status and Location.
* Credit\_Score: Impute with mean or median (only 2 missing).
* Loan\_Balance: Impute with median or model-based predictions using correlated financial variables.

# 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:

* Higher Credit Utilization correlates with higher delinquency likelihood.
* Missed Payments is a strong direct indicator of delinquency.
* Lower Credit Score and higher Debt-to-Income Ratios are associated with higher delinquency.

- Unexpected anomalies:

* Credit\_Utilization > 1 is logically inconsistent.
* Distribution of delinquent accounts appears imbalanced (~16% delinquent, 84% not).
* Further review may be needed for income values below ₹20,000 or extremely high loan balances.

# 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.'

# 6. Conclusion & Next Steps

**Key findings**:

* The top 3 features most predictive of delinquency are:
  1. **Missed\_Payments**
  2. **Credit\_Utilization**
  3. **Credit\_Score**
* Missing data is manageable through targeted imputation.
* A few outliers and inconsistencies require capping or validation.

**Recommended next steps**:

* Handle missing values via imputation.
* Cap or scale Credit\_Utilization to a valid range [0, 1].
* Encode categorical variables (e.g., One-Hot or Label Encoding).
* Train baseline classification models (e.g., Logistic Regression, Random Forest, XGBoost).
* Evaluate performance using metrics like ROC-AUC, Recall, and F1-score.