Exploratory Data Analysis (EDA)
Summary Report

1. Introduction

The purpose of this Exploratory Data Analysis (EDA) report is to thoroughly investigate a dataset containing financial and behavioural attributes of customers. The primary goal is to identify key patterns, trends, and anomalies that could help in building a predictive model for **delinquency risk**—that is, the likelihood that a customer may default or miss future payments.

This report aims to:

- Understand the structure and quality of the data.
- Identify variables most predictive of delinquency.
- Detect early signs of risk based on historical financial behaviour.
- Highlight any data quality issues such as missing or inconsistent entries that might impact modelling performance

2. Dataset Overview

This dataset consists of detailed financial and behavioural information about individual customers. It is designed to support **predictive modelling of delinquency**, which refers to the **likelihood that a customer fails to meet financial obligations**, such as making loan or credit payments on time.

The dataset includes **demographic, credit, income, employment, and behavioural features** along with a delinquency flag (***Delinquent_Account***) that indicates whether a customer has been delinquent (1) or not (0). The goal is to analyse this data to find patterns that can help forecast future delinquencies.

➡ **Key dataset attributes:**

- Number of records: 501
- Key variables:

  - ○ **Customer_ID:** Unique identifier for each customer
  - ○ **Age:** Age of the customer in years
  - ○ **Income:** Annual income (may have missing entries)
  - ○ **Credit_Score:** Ranges typically from 300 to 850
  - ○ **Credit_Utilization:** Percent of used credit (0–100%)
  - ○ **Missed_Payments:** Count of missed payments in past year
  - ○ **Loan_Balance:** Outstanding loan value
  - ○ **Debt_to_Income_Ratio:** Debt as percentage of income
  - ○ **Employment_Status:** Job status (Employed, Self-employed, etc.)
  - ○ **Delinquent_Account:** Target variable (0 = No, 1 = Yes)
  - ○ **Month_1 to Month_6:** Past 6 months' payment behaviour (0 = On-time, 1 = Late, 2 = Missed)

- **Data types:**
  - ○ **Numerical:** Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure
  - ○ **Categorical:** Employment_Status, Credit_Card_Type, Location, Monthly payment history (Month_1–Month_6)
  - ○ **Binary:** Delinquent_Account

➡ **Initial Data Findings**

◇ **Notable Missing or Inconsistent Data:**

- **Income:** Contains multiple missing values which could skew the debt-to-income ratio.
- **Employment_Status:** Some unexpected blank or unrecognized categories (e.g., typos or nulls).
- **Credit_Utilization:** A few extreme values exceeding 100%—potential data entry errors.

◇ **Key Anomalies:**

- **Credit_Score:** Some values fall outside the typical 300–850 range.

- **Debt_to_Income_Ratio:** Outliers observed beyond 100%, which is not logically valid.
- **Month_1 to Month_6:** Mixed categories; some customers show all on-time while others have consistent missed payments—helps indicate clusters or groups at risk.

◇ Early Indicators of Delinquency Risk:

- High Missed_Payments strongly correlates with Delinquent_Account = 1.
- Poor Credit_Score (<600) often paired with high Credit_Utilization.
- Patterns in Month_1 to Month_6 showing back-to-back missed or late payments align with delinquency flags.

Summary:

On initial examination, the dataset is generally well-organized, but a few data quality concerns stand out. The Income column has noticeable gaps, while variables like Debt_to_Income_Ratio and Credit_Utilization contain some unrealistic or extreme values that may need to be cleaned or capped. From a predictive standpoint, several fields—particularly those tied to past payment behavior and credit profile—appear highly relevant for identifying delinquency risk. Notably, patterns of missed payments and low credit scores consistently align with delinquent accounts. These early observations will play a key role in shaping the data preparation and feature engineering strategy for modeling.

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
  - Loan_Balance
  - Credit_Score
- ◇ **Missing data treatment:**
  - Income: **Synthetic Imputation**
    - *Justification:* Income is a crucial predictive variable with moderate missingness. We assumed a normal distribution and generated synthetic values using AI-assisted imputation to maintain realism, fairness, and distributional integrity.

- Loan_Balance: **Median Imputation**
    - *Justification:* Loan balance is prone to skewness due to a few customers carrying very high debt. Using median imputation helps retain robust central tendency without being affected by outliers. This method ensures more stable downstream predictions, particularly in financial risk models. It is suitable when the missingness is moderate but non-random.
- Credit_Score: **Mode Imputation**
    - *Justification:* Credit score is a numerical metric with very low missingness in this dataset. Mean imputation was chosen for simplicity and because the impact on distribution is negligible. This method maintains consistency across the credit score scale and ensures that imputed values reflect the overall population trend without distorting variability.

- **Summary Table of Missing Data Handling:**

Variable	**Treatment Method**	**Justification**
Income	Synthetic Imputation (Normal Distribution)	Preserves natural income variation; useful for modelling financial behavior fairly.
Loan_Balance	Median Imputation	Reduces bias from outliers; stable for skewed financial data.
Credit_Score	Mean Imputation	Minimal missingness; preserves numerical scale and interpretability.

4. Key Findings and Risk Indicators

Identifying 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, debt-to-income ratio, and income exhibit a weak but positive correlation with delinquency. These features collectively point to financial stress as a key risk driver. Account tenure and missed payments show slight negative correlations, indicating that longer customer relationships and historical repayment

behavior are generally protective. While each individual correlation is modest, together they highlight multifactorial risk patterns valuable for prediction.

- **Unexpected anomalies:**
  A subset of customers recorded **multiple missed payments (4 or more)** but were not labeled as delinquent. For example:

  - Customer_ID: CUST0006 had 6 missed payments, is unemployed, and holds a Gold credit card, yet shows no delinquency.

  - Customer_ID: CUST0010 missed 4 payments, has a low credit score (340), and high utilization (83%), yet is also not marked delinquent.

  - Customer_ID: CUST0014 is retired, missed 5 payments, but maintains a Standard card and shows no delinquency.
    These cases suggest potential **mislabeling**, **payment restructuring**, or **model-exempt behavior**. Further review by the data integrity or compliance team is recommended to clarify these exceptions.

➡ **Notable Insights:**

◆ **Credit Utilization and Credit Score** both show weak but positive associations with delinquency — suggesting **a multi-factor approach** is needed rather than relying on a single variable.

◆ **Loan Balance** shows a **slight negative** correlation, implying high-balance loans alone aren't directly predictive of delinquency — this may warrant deeper segmentation (e.g., secured vs. unsecured loans).

◆ **Age** is weakly positively correlated, but not a strong predictor on its own.

◆ **Missed_Payments** surprisingly showed a **negative correlation** with delinquency. This could be due to:

- Mislabelling or reverse coding
- Customers with low risk missing payments due to technical issues
- Need to re-inspect data quality or segment patterns

➡ **High-Risk Indicators:**

Indicator	**Explanation**
Low Credit Score	Customers with lower scores are generally more likely to be delinquent, signaling financial mismanagement or lack of credit history.
High Credit Utilization	A high percentage of credit usage can indicate financial stress, which increases default risk.
High Debt-to-Income Ratio	A high ratio means customers have less disposable income, raising their chances of delinquency.
Short Account Tenure	Newer accounts are typically riskier due to limited payment history and weaker loyalty indicators.
Missed Payments History	While weakly correlated overall, multiple missed payments still suggest chronic risk behavior.

5. AI & GenAI Usage

Generative AI tools (like ChatGPT, Gemini AI, Julius AI, were used) were used to summarize the dataset, impute missing data, and detect patterns. Here 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.'

- ‘Think as if you are data analyst and analyze the whole dataset provided and find missing values and provide how to counter those.’

- ‘Identify top 3 variables most likely to predict delinquency based on the provided dataset.’

- ‘Use the best imputation strategy for the missing values and what criteria were used to decide whether to impute, drop, or model missing values.’

6. Conclusion & Next Steps

The exploratory analysis of the delinquency dataset uncovered several meaningful patterns and actionable insights that directly inform the development of a robust predictive model. Core financial indicators such as credit utilization, debt-to-income ratio, and income emerged as consistent contributors to delinquency risk. Notably, behavioral trends—especially patterns in missed payments and credit score degradation—proved crucial in distinguishing high-risk customer profiles.

While the data is generally well-structured, key anomalies and missing entries highlight the importance of careful preprocessing. For instance, a subset of customers with repeated missed payments were not marked delinquent, signaling a need for data validation or policy clarification. Likewise, outliers in fields such as loan balance and debt ratios may skew model training if not properly capped or normalized.

The strategic use of Generative AI tools (such as ChatGPT and Julius AI) significantly accelerated the identification of missing data patterns and enhanced the quality of imputation decisions—particularly for income values where synthetic data was generated to maintain distribution integrity.

➡ Recommended Next Steps:

1. Data Quality Improvement:
- Audit anomalies (e.g., high missed payments without delinquency) to ensure label accuracy.
- Address outliers in Debt_to_Income_Ratio and Credit_Utilization using capping or transformation.
- Standardize and clean inconsistent entries in Employment_Status and other categorical fields.

2. Feature Engineering:
- Create aggregated behavioral scores (e.g., weighted payment history).
- Introduce interaction terms between income, utilization, and credit score to capture compound risk.

3. Modeling Phase:
- Begin baseline model training using logistic regression and tree-based algorithms (e.g., Random Forest, XGBoost).
- Apply SHAP or permutation-based feature importance to validate variable relevance.

4. Fairness & Bias Review:
- Assess models for bias across demographic segments (e.g., age or employment status).
- Ensure synthetic imputations do not disproportionately affect risk scoring for specific

groups.

5. Business Collaboration:

- Engage with credit and compliance teams to review edge cases and refine definitions of delinquency.
- Define actionable risk thresholds for use in real-time scoring applications.

In summary, this EDA lays a strong foundation for predictive modeling by surfacing both intuitive and unexpected drivers of delinquency. The insights gathered here not only support data science workflows but also offer strategic direction for credit risk management and operational decision-making.