**Exploratory Data Analysis (EDA)**

**Summary Report**

**1. Introduction**

The purpose of this report is to conduct a thorough exploratory data analysis (EDA) on Geldium's delinquency prediction dataset. The goal is to assess data quality, detect patterns and anomalies, and identify key variables that influence the likelihood of delinquency. This analysis will guide future predictive modelling efforts by uncovering early indicators of financial risk.

**2. Dataset Overview**

This dataset contains financial and behavioural data for a large pool of customers, with the goal of predicting the likelihood of loan delinquency. An initial review was conducted to understand the structure, content, and quality of the data.

**Key Dataset Attributes:**

* **Number of records:** 500
* **Key variables:**
  + Income: Monthly income of the customer
  + Loan Balance: Outstanding loan balance
  + Credit Utilization: Percentage of credit used relative to the limit
  + Credit Score: Customer’s credit score
  + Payment History: Number of missed payments in the last 12 months
  + Credit Inquiries: Number of hard credits pulls in the last 6 months
  + Credit History Length: Age of the oldest credit account
  + Employment Status: Current employment situation
  + Delinquent Account: Target variable (1 = delinquent, 0 = non-delinquent)
* **Data types:**
  + Numerical: Income, Loan Balance, Credit Utilization, Credit Score, Payment History, Credit Inquiries, Credit History Length
  + Categorical: Employment Status
  + Target Variable: Delinquent Account (Binary)

**Initial Anomalies & Inconsistencies:**

* No duplicate rows detected.
* Minor missing values in key numerical columns (e.g., Income, Loan Balance, Credit Score).
* No structural inconsistencies or type mismatches identified.

**3. Missing Data Analysis**

Identifying and addressing missing values is critical to ensuring the quality and fairness of any model developed.

**Key Missing Data Findings:**

* **Variables with missing values:**
  + Income: 39 missing entries
  + Loan Balance: 29 missing entries
  + Credit Score: 2 missing entries

**Missing Data Treatment:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Method** | **Justification** |
| Income | Median Imputation | Handles skewed distribution without distortion from outliers |
| Loan Balance | Mean Imputation | Symmetric distribution; mean preserves numeric integrity |
| Credit Score | Mean Imputation | Very few missing entries; minimal impact |

No variables were dropped due to missing data, and imputation ensures a complete dataset without introducing bias.

**4. Key Findings and Risk Indicators**

**Feature Relationships & Risk Factors**

|  |  |
| --- | --- |
| **Indicator** | **Why It Matters** |
| **Credit Utilization > 70%** | Indicates financial strain and increased risk of missed payments |
| **Past Missed Payments** | Strongest predictor of future delinquency |
| **Debt-to-Income Ratio > 36%** | Reflects reduced capacity to manage debt consistently |
| **Short Credit History (<2 years)** | Less reliable credit behaviour; more volatile financial patterns |
| **Multiple Credit Inquiries (5+)** | Suggests financial stress or overextension |

**Correlations with Delinquency**

* Strong positive correlation with high credit utilization and payment history inconsistencies.
* Lower income levels are associated with increased delinquency.
* Shorter credit histories and frequent credit inquiries also show elevated risk patterns.

**Unexpected Insights**

* **Age**: No clear correlation with delinquency, counter to common assumptions.
* **High-Income Individuals**: Some still show delinquency due to poor utilization or financial mismanagement.
* **Geographic Variance**: Similar regions showed differing delinquency patterns—requires further regional study.
* **Credit Behaviour Nuance**: In some cases, rising credit scores paired with multiple inquiries suggest strategic—not risky—credit usage.

**5. AI & GenAI Usage**

Generative AI tools were used to streamline the data exploration process, enhance accuracy, and detect early signals of delinquency risk.

**Example AI Prompts Used:**

* "Summarize key patterns in the dataset and identify anomalies."
* "Identify missing values in this dataset and recommend the best imputation strategy based on industry best practices."
* "Analyse the correlation between customer income and delinquency risk, summarizing key findings in simple terms."
* "Highlight early indicators of delinquency risk using payment and credit behaviour data."

AI-assisted insights helped confirm industry patterns and surface unexpected correlations worth deeper investigation.

**6. Conclusion & Next Steps**

**Summary of Key Findings:**

* Strongest predictors of delinquency include high credit utilization, past missed payments, short credit history, and high debt-to-income ratios.
* Income, while important, is not a standalone indicator.
* No major data quality issues remain after imputation.
* Unexpected regional and behavioural patterns suggest room for segmentation and tailored risk modelling.

**Recommended Next Steps:**

1. Decide the best way to deal with missing income and loan balance values, ensuring that the chosen method does not introduce bias.
2. Double-check whether high credit utilization and missed payments remain the strongest indicators of delinquency across different customer groups.
3. Look into records where customers have high income but low credit scores to see if there are reporting errors or other explanations.