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

# Introduction

The purpose of this report is to conduct an exploratory data analysis (EDA) on Geldium’s dataset in order to assess its completeness, accuracy, and consistency. The goal is to uncover key patterns, detect anomalies, and identify data quality issues that could impact delinquency prediction models. By analyzing missing values, outliers, and early indicators of credit risk, this report aims to provide a clear understanding of the dataset’s strengths and limitations. These insights will guide the development of a more reliable, data-driven delinquency risk model and support Geldium in designing effective customer 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.

**Key dataset attributes:**

* **Number of records:** Based on the row numbers visible, the dataset contains at least 32,849 records.
* **Key variables:**
  + id: A unique identifier for each customer/record.
  + gender: The gender of the customer (e.g., Male, Female).
  + area: The geographical area or region of the customer.
  + age: The age of the customer.
  + age\_in\_yrs: A calculated age in years (appears to be a numerical representation of age).
  + state: The state or province of the customer.
  + channel: The acquisition channel (e.g., 'Agent', 'Direct', 'Online').
  + A/c Type: The type of account (e.g., 'Normal', 'VIP').
  + customer\_segment: A segment assigned to the customer (e.g., 'High-Value', 'Low-Value').
  + months\_on\_network: The duration the customer has been on the network, in months.
  + payment\_status: The status of recent payments.
  + amount\_paid\_last\_month: The amount of the last payment.
  + payment\_history: The history of payments, possibly a categorical or coded variable.
  + last\_payment\_date: The date of the last payment.
  + pending\_balance: The outstanding balance.
  + reconnection\_fee: A fee that may be applied for reconnection.
  + total\_amount\_due: The total amount currently owed.
  + delinquent\_flag: This is likely the target variable, indicating whether an account is delinquent (1) or not (0).
* **Data types:**
  + **Categorical:** gender, area, state, channel, A/c Type, customer\_segment, payment\_status, payment\_history, delinquent\_flag.
  + **Numerical:** id, age, age\_in\_yrs, months\_on\_network, amount\_paid\_last\_month, pending\_balance, reconnection\_fee, total\_amount\_due.
  + **Date/Time:** last\_payment\_date.

# 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:**
  + The screenshots show NA values in the amount\_paid\_last\_month column.
  + pending\_balance also shows NA values.
  + reconnection\_fee has NA values.
  + total\_amount\_due has NA values.
  + last\_payment\_date has NA values.
  + age\_in\_yrs seems to be missing values in some rows.
* **Missing data treatment:**
  + **Imputation for numerical variables:** For numerical columns like amount\_paid\_last\_month, pending\_balance, reconnection\_fee, and total\_amount\_due, a suitable approach would be to impute the missing values.
    - **Method:** Impute with the mean or median. The median is a more robust choice as it is less sensitive to outliers.
    - **Justification:** Deleting records with missing values might lead to a significant loss of data, especially if many columns have missing values. Imputation allows us to retain these records and utilize the other information they contain. For total\_amount\_due, a logical imputation might be 0 if pending\_balance and reconnection\_fee are also NA, as this could indicate a non-delinquent or inactive account.
  + **Imputation or Deletion for last\_payment\_date:**
    - **Method:** For last\_payment\_date, if the missing value is a significant portion of the data and a critical feature, it might be challenging to impute accurately. A possible approach is to create a new categorical feature, "has\_last\_payment\_date", with values "Yes" and "No", and drop the original column.
    - **Justification:** This method captures the information that the date is missing, which could be a risk indicator in itself, without requiring complex date imputation.
  + **Handling age and age\_in\_yrs:** It seems age\_in\_yrs is a cleaner, numerical version of the age column, which may contain strings like '30-40'. The best approach is to drop the age column and use age\_in\_yrs as the primary feature, imputing any remaining missing values in age\_in\_yrs with the median.

# 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:**
  + A high correlation is expected between pending\_balance, reconnection\_fee, and total\_amount\_due. The total\_amount\_due is likely the sum of the others. The presence of a non-zero reconnection\_fee is a strong indicator of a past or current delinquent state.
  + There is likely a strong correlation between a high pending\_balance and delinquent\_flag = 1.
  + The payment\_status and payment\_history are likely the most direct indicators of a delinquent account and will be highly correlated with delinquent\_flag. A payment status of 'Late' or 'Pending' would be a strong risk indicator.
  + months\_on\_network could show an interesting relationship. Newer customers might be more likely to become delinquent, or very long-term customers might have higher delinquency rates if they have built up significant debt.
  + The channel and customer\_segment could also be predictive. For example, a certain channel ('Agent') or customer segment ('Low-Value') might have a higher delinquency rate.
* **Unexpected anomalies:**
  + The presence of NA values in core financial columns like amount\_paid\_last\_month and total\_amount\_due is an anomaly that requires further investigation. It could mean the data was not recorded, or it could be a proxy for a value of zero (e.g., zero amount paid, zero balance).
  + The age column contains non-numerical values like '30-40', which needs to be handled by using the age\_in\_yrs column or converting these string ranges to a single numerical value (e.g., the median of the range).
  + Check for duplicates in the id column to ensure each row represents a unique customer. The number of records seems large, so this is a crucial data cleaning step.

# 5. AI & GenAI Usage

### Generative AI Tools and Prompts

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 the results.

The analysis of the delinquency prediction dataset was facilitated by a generative AI model that performed tasks such as data summarization, missing value imputation, and pattern detection. The insights derived from these tasks were used to populate the initial sections of this report.

#### Example AI prompts used:

The following prompts, or similar variations, were used to guide the generative AI model in its analysis of the dataset:

* **For the Dataset Overview:**
  + "Summarize this dataset for a project report. Include the number of records, a list of key variables and their data types (categorical, numerical, etc.), and highlight any initial anomalies."
* **For Missing Data Analysis:**
  + "Analyze the attached dataset and identify all columns with missing values. Describe the type of data in these columns and suggest an appropriate imputation strategy for each, justifying your chosen method."
* **For Key Findings and Risk Indicators:**
  + "Summarize key patterns in the dataset and identify anomalies. Based on the provided variables, which features are most likely to be correlated with the 'delinquent\_flag'? Provide a summary of these correlations and highlight any unexpected or unusual data points that require further investigation."
* **For Code Generation (as a supporting task):**
  + "Write Python code to calculate the correlation matrix for all numerical variables in the dataset and visualize it as a heatmap."
  + "Generate Python code using the pandas library to fill in missing 'pending\_balance' values with the median of the column."

# 6. Conclusion & Next Steps

### 6. Conclusion & Next Steps

Based on the initial analysis of the delinquency prediction dataset, we've identified key findings and laid the groundwork for a predictive model. The dataset, while rich in detail with over 32,000 records, requires careful preprocessing, especially concerning missing values and inconsistent data types. The core of the analysis revealed that payment-related variables, such as total\_amount\_due, pending\_balance, and payment\_status, are likely the strongest indicators of delinquency.

**Key Findings:**

* The delinquent\_flag is a clear binary target variable, making this a supervised classification problem.
* **Missing data** is a significant issue in key numerical columns like amount\_paid\_last\_month and pending\_balance, which will require a robust imputation strategy.
* The age variable needs to be cleaned and converted to a numerical format, utilizing the more consistent age\_in\_yrs column.
* Strong correlations are anticipated between the delinquent\_flag and financial variables like total\_amount\_due and payment\_status.

**Recommended Next Steps:**

1. **Data Cleaning and Preprocessing:** Implement the suggested imputation strategy, using the median to fill in missing numerical values and cleaning the age column. Convert all categorical variables into a numerical format suitable for machine learning models (e.g., one-hot encoding).
2. **Feature Engineering:** Explore creating new features that might improve the model's predictive power. For instance, a new variable like months\_since\_last\_payment could be calculated from last\_payment\_date.
3. **Model Selection & Training:** Begin with a baseline model (e.g., Logistic Regression or a Decision Tree) to establish performance metrics. Then, experiment with more advanced models like Gradient Boosting Machines (e.g., XGBoost) or a Random Forest, which are well-suited for this type of classification task.
4. **Model Evaluation:** Use standard metrics such as **Accuracy, Precision, Recall, and the F1-Score** to evaluate the model's performance. Given the importance of identifying delinquent accounts, **Precision** and **Recall** will be critical metrics to monitor.