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

# 1. Introduction

This document contains an exploratory data analysis of Geldium’s dataset, which is aimed at evaluating data integrity, uncovering valuable insights, and identifying factors that contribute to the risk of credit default. The primary objective is to prepare the data for accurate predictive modeling and risk evaluation.

# 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, Income, Credit Score, Credit Utilization, Missed Payments, Debt-to-Income Ratio

- Data types:

* Customer\_ID: string/ID
* Age, Missed\_Payments, Account\_Tenure: integer
* Income, Credit\_Score, Credit\_Utilization, Loan\_Balance, Debt\_to\_Income\_Ratio: float/continuous
* Delinquent\_Account: binary/categorical
* Employment\_Status, Credit\_Card\_Type, Location, Month\_1…Month\_6: categorical

# 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 methods.

**Key missing data findings:**

* **Variables with missing values:**
* *Loan\_Balance* (20 records, 4% missing)
* *Debt\_to\_Income\_Ratio* (15 records, 3% missing)
* *Credit\_Utilization* (8 records, 1.6% missing)

**Missing data treatment:**

* **Loan\_Balance**
  + Method: Imputation with median by *Credit\_Card\_Type*
  + Justification: Preserves distribution differences across card products and mitigates skew from large balances.
* **Debt\_to\_Income\_Ratio**
  + Method: Imputation with overall median
  + Justification: Ratio is fairly stable across population; median fill avoids bias from extreme values.
* **Credit\_Utilization**
* Method: Cap outliers at 1.0, then impute missing with median by *Credit\_Card\_Type*
* Justification: Enforces realistic utilization bounds and respects segment-level usage patterns.

**4. Key Findings and Risk Indicators**

 **Missed\_Payments**: Direct measure of past payment behavior—more misses → higher future default risk.

* **Credit\_Utilization**: Utilization ≥80% signals overextension, strongly tied to delinquency.
* **Debt\_to\_Income\_Ratio**: High DTI (>50%) shows limited repayment capacity, boosting risk.
* **Delinquent\_Account**: A history of delinquency is the single best predictor of future default.
* **Account\_Tenure**: Short tenures (

Top Predictive Variables for Delinquency

1. **Missed\_Payments** – Strong indicator of payment behavior and credit discipline.
2. **Credit\_Utilization** – High utilization rates are often linked to higher default risk.
3. **Delinquent\_Account** – Whether the customer has defaulted in the past is a leading signal for future delinquency.

- Unexpected anomalies:

 Some customers have **zero account tenure** or appear to hold **multiple credit cards categorized as “Student” or “Business”** inconsistently.

* **Repeated late or missed payments** in the last 6 months are present even among customers with high income or high credit scores, suggesting unexpected behavioral variance.

# AI & GenAI Usage

 Leveraged a GenAI assistant (e.g., ChatGPT) to auto-summarize dataset structure, flag anomalies, and detect missing-value hotspots.

* Used AI-driven prompt engineering to identify top predictive features and suggest correlation checks.
* Consulted GenAI for best-practice imputation strategies (median fill, outlier capping) and category standardization.
* Employed AI to generate sample synthetic income values and validate distributional assumptions.

# 6. Conclusion & Next Steps

Our review of 500 customer records uncovered minor missingness (<5%) in Loan\_Balance, DTI, and Utilization, all of which have been imputed with segment-aware medians and capped at realistic bounds. Key risk drivers—Missed\_Payments, high Credit\_Utilization, elevated Debt\_to\_Income\_Ratio, and prior Delinquent\_Account—show strong correlations with future delinquency, while lower Credit\_Score and Income also contribute modestly.

**Next Steps**

1. Feature Engineering
   * Create binary flags for extreme utilization (>0.8) and DTI (>0.5)
   * One-hot encode standardized Employment\_Status and Card\_Type
2. Model Development
   * Train baseline classifiers (logistic regression, decision trees)
   * Advance to ensemble methods (random forest, XGBoost)
3. Validation & Tuning
   * Use stratified k-fold cross-validation
   * Optimize hyperparameters for precision-recall trade-off
4. Deployment & Monitoring

* Establish model monitoring for data drift
* Schedule periodic retraining with new data batches