

Exploratory Data Analysis (EDA) – Geldium Delinquency Dataset

This report summarizes an Exploratory Data Analysis (EDA) and data quality review of Geldium’s delinquency prediction dataset. The objective is to assess dataset structure, completeness, and inconsistencies that could affect delinquency risk modeling, and to identify early risk indicators to guide downstream feature engineering and predictive modeling.

2. Dataset Overview

Number of records: 500
Columns (19): Customer_ID, Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Delinquent_Account, Loan_Balance, Debt_to_Income_Ratio, Employment_Status, Account_Tenure, Credit_Card_Type, Location, Month_1, Month_2, Month_3, Month_4, Month_5, Month_6, Missed_in_6m, Late_in_6m
Target variable: Delinquent_Account (1=Delinquent). Delinquency rate: 16.0%.

Step 1 – Key Insights (Data Quality + Early Risk Signals)

- Notable missing data: Income (7.8%), Loan_Balance (5.8%), Credit_Score (0.4%).
- Key inconsistencies: Employment_Status category standardization needed (EMP/employed/retired variants).
- Key anomalies: Credit_Utilization values above 1.0 (max 1.026) — treat as over-limit or data issue.
- Early risk indicators to validate: high Debt_to_Income_Ratio, Employment_Status (Unemployed), Credit_Card_Type (Business/Student), payment behavior across Month_1–Month_6.

Initial data quality summary:

Initial review indicates the dataset is mostly complete, with missing values concentrated in Income (7.8%) and Loan_Balance (5.8%), while Credit_Score has minimal missingness (0.4%). No duplicate customer records were found, which is positive for modeling integrity. However, Employment_Status contains inconsistent category labels (e.g., EMP vs employed vs retired) and Credit_Utilization includes values above 1.0, both of which require cleaning. Delinquency prevalence is ~16%, which is suitable for supervised modeling but requires careful handling of class imbalance in later stages.

3. Missing Data Analysis

Income: 7.8% missing (critical affordability feature).; Loan_Balance: 5.8% missing (exposure/obligation feature).; Credit_Score: 0.4% missing (small but relevant).

Missing data issue	Handling method	Justification
Income (7.8% missing)	Impute (median by Employment_Status) + missing indicator	Preserves information for affordability risk; segment
Loan_Balance (5.8% missing)	Impute (median by Credit_Card_Type)	Maintains exposure feature without losing rows
Credit_Utilization > 1.0 (anomaly)	Cap at 1.0 + anomaly flag	Prevents unrealistic leverage effects while reta

4. Key Findings and Risk Indicators

- Overall delinquency rate: 16.0%.
- Debt_to_Income_Ratio shows higher delinquency in the high-DTI segment (>0.40), suggesting affordability stress as an indicator.
- Employment segment differences observed: Unemployed customers show higher delinquency rate than other groups.
- Credit_Card_Type differences observed: Business and Student card segments show higher delinquency than Platinum/Standard.

High-risk indicators (to prioritize in modeling):

- **High Debt-to-Income Ratio:** Higher delinquency observed in DTI > 0.40 group, indicating affordability stress.
- **Unemployment / unstable employment:** Unemployed segment shows higher delinquency rate vs others.
- **Business or Student card type:** These card segments have higher delinquency vs Platinum/Standard, may reflect usage profile.
- **Utilization anomalies / over-limit behavior:** Utilization > 1.0 suggests over-limit behavior or data issues; both are risk-relevant.
- **Recent payment irregularity (Month_1–Month_6):** Patterns of Late/Missed payments should be engineered into rolling behavioral features.

5. AI & GenAI Usage

GenAI was used only on aggregated, non-sensitive statistics (no PII or account-level financial details were shared) to accelerate insight generation and document drafting. Prompts included: (1) 'Summarize key patterns, outliers, and missing values in this dataset and highlight modeling risks.' (2) 'Suggest best-practice imputation strategies for missing income and loan balance values.' (3) 'List likely delinquency risk indicators given these variables (DTI, utilization, missed payments, tenure, employment).'

6. Conclusion & Next Steps

The dataset is structurally sound (500 records, 19 fields) with no duplicates, but has meaningful missingness in Income and Loan_Balance that should be imputed prior to modeling. Key standardization is required for Employment_Status categories and utilization values exceeding 1.0 should be handled (cap/winsorize and flag). Next steps: finalize imputation, add missing-value indicators, engineer behavior features from Month_1–Month_6, confirm delinquency label definition, and proceed to baseline model training with monitoring for segment bias.