# Geldium Delinquency Risk – EDA Summary Report (Updated)

Dataset: Delinquency_prediction_dataset.xlsx | Records: 500 | Variables: 22 | Delinquency rate: 16.0%

## Step 1: Key insights and data quality observations

• Dataset contains 500 customer records and 22 variables.
• Target variable: Delinquent_Account; delinquency prevalence = 16.0%.
• Top missingness: Income (7.8%), Loan_Balance (5.8%), Credit_Score (0.4%).
• Employment_Status has inconsistent categorical labels (EMP/employed/Employed, retired).
• Credit_Utilization contains 4 records > 1.0 (max ~1.026) – should be capped/flagged.
• No duplicate Customer_IDs detected; payment history columns Month_1–Month_6 are complete.

Initial data quality summary: The dataset is largely clean and suitable for modeling with minor remediation. Missingness is concentrated in Income and Loan_Balance, both important affordability/exposure drivers, and must be handled via imputation and/or segment-aware synthetic fills. Categorical standardization is required for Employment_Status to avoid feature fragmentation. A small number of Credit_Utilization values exceed 1.0 and should be capped at 1.0 or treated as data errors. Overall delinquency prevalence is ~16%, so class imbalance handling will be needed in later modeling.

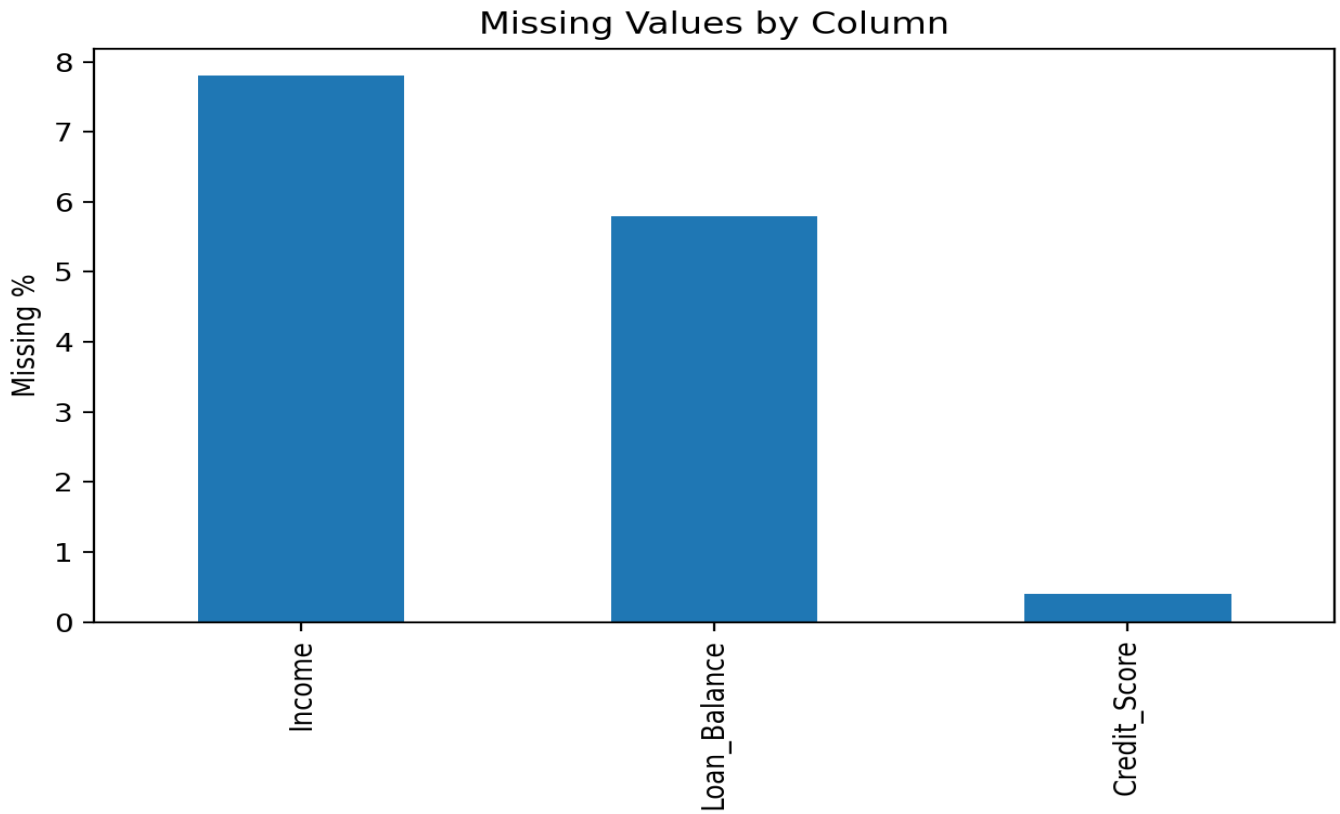
## Step 2: Missing data handling plan

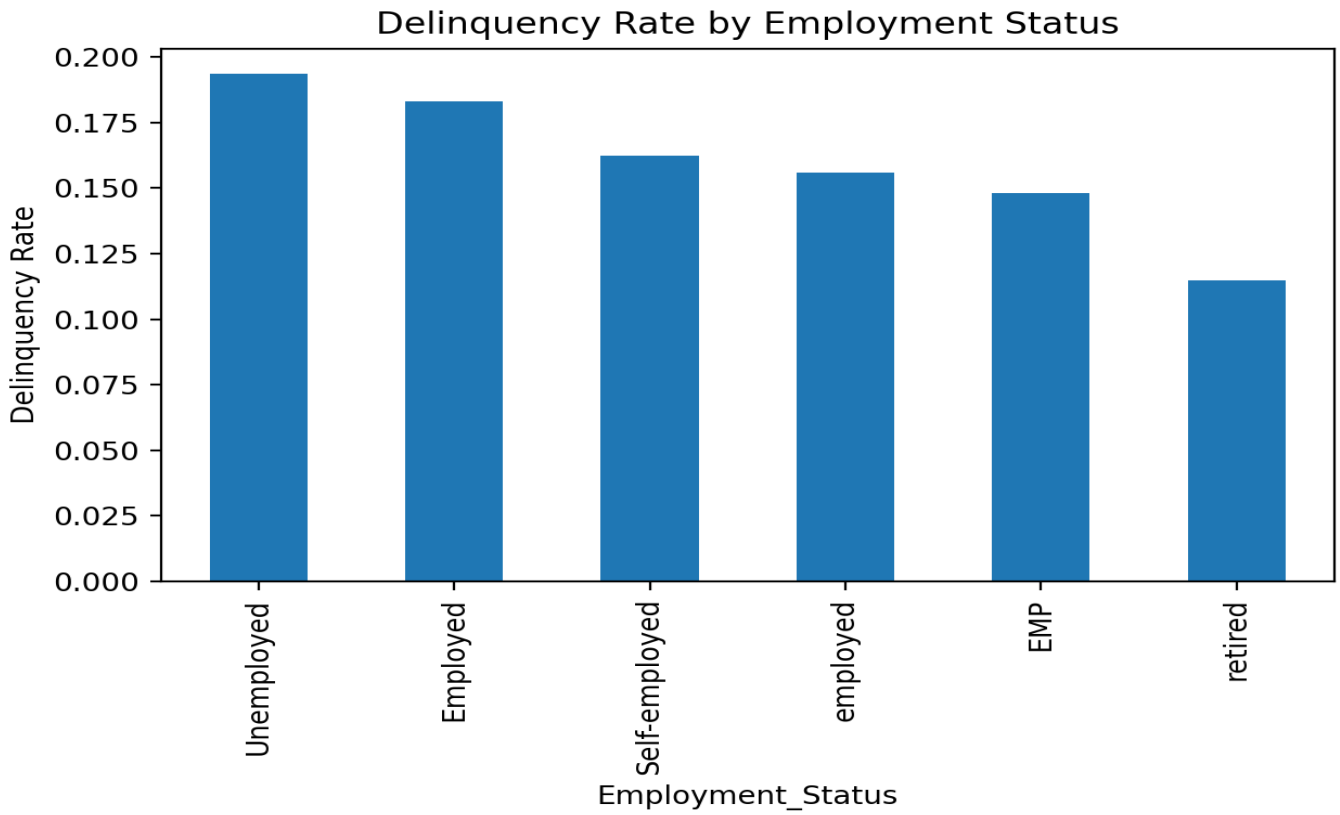
| Missing/data issue | Treatment approach | Justification |
|---|---|---|
| Income (7.8% missing) | Segment-aware imputation / synthetic generation + missing flag | Income is predictive for affordability; preserve distri |
| Loan_Balance (5.8% missing) | Median imputation within Credit_Card_Type + missing flag | Balance reflects exposure; median-by-segment is r |
| Credit_Utilization > 1.0 (4 rows) | Cap at 1.0 and add anomaly flag | Utilization logically bounded; flag preserves signal v |

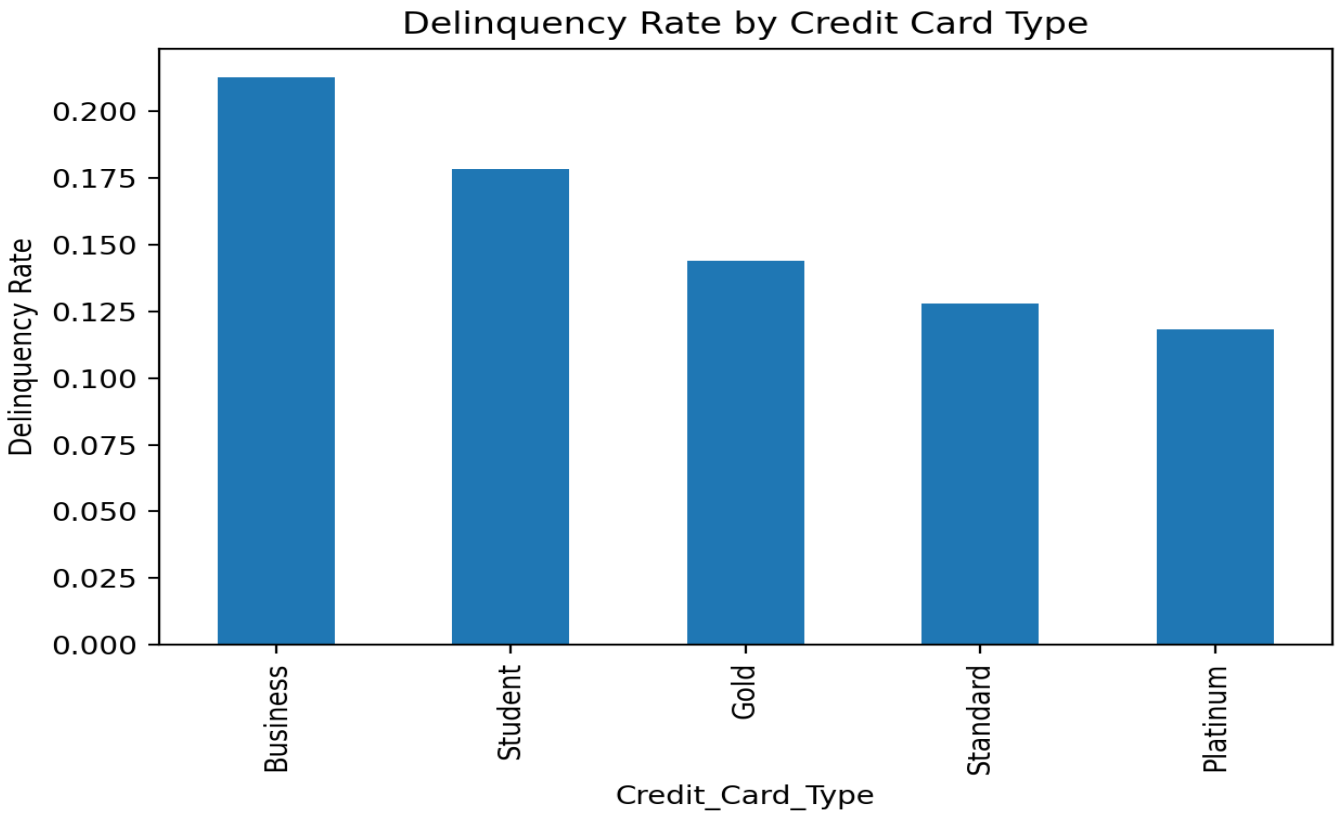
## Step 3: Risk indicators to consider in delinquency modeling

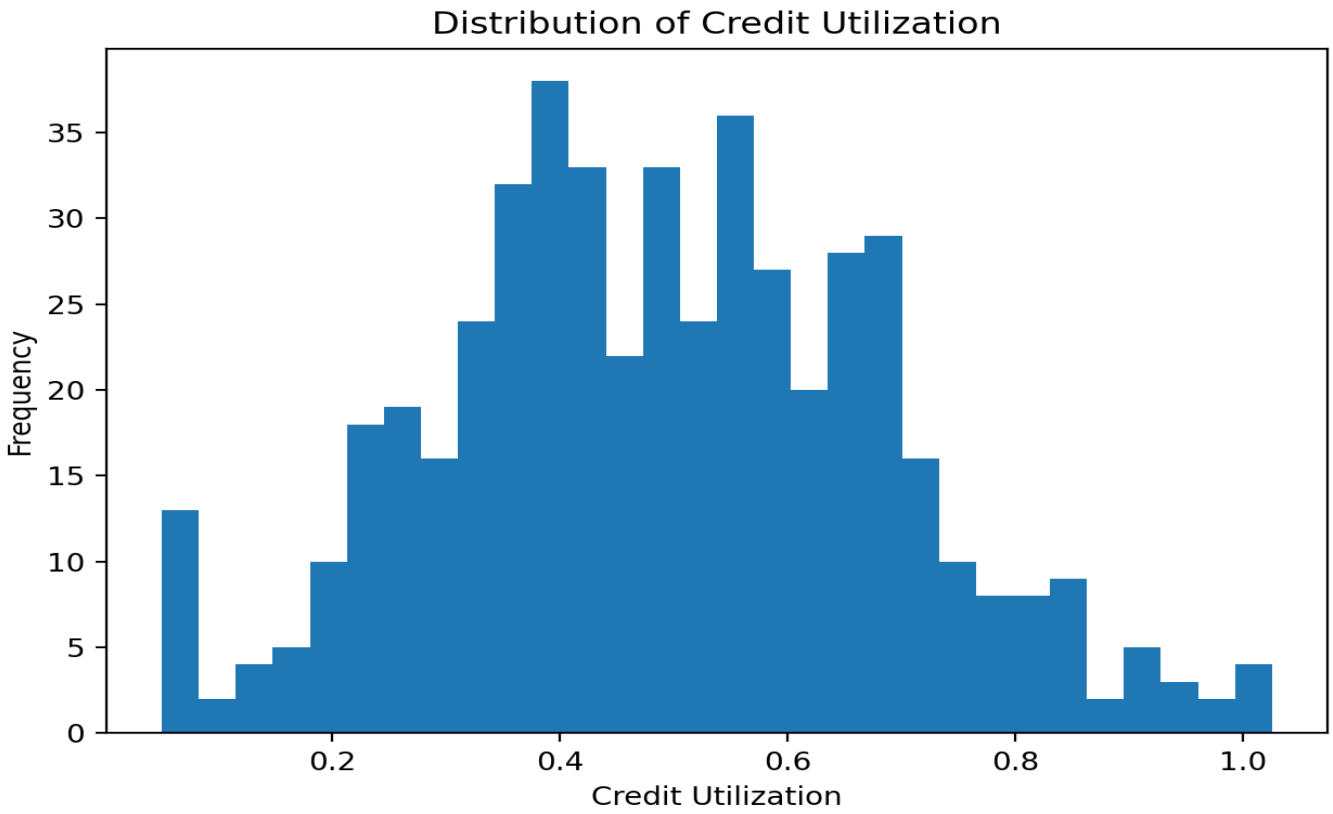
• Employment_Status: Unemployed has the highest delinquency rate (~19.4%).
  Why it matters: Indicates income instability and higher repayment risk.
• Credit_Card_Type: Business (~21.3%) and Student (~17.9%) show higher delinquency.
  Why it matters: Suggests product-level risk differences for collections segmentation.
• Location: Los Angeles (~19.6%) shows the highest delinquency among cities.
  Why it matters: Geographic effects may reflect economic conditions or portfolio mix.
• Debt_to_Income_Ratio and Credit_Utilization should be prioritized features.
  Why it matters: Both are standard affordability and financial-stress indicators for delinquency prediction.
• Payment history (Month_1–Month_6) should be engineered into streak/recency features.
  Why it matters: Patterns of repeated Late/Missed payments typically outperform raw counts in delinquency models.

## Supporting Visuals


Missing Values by Column
Missing %
0
1
2
3
4
5
6
7
8
Income
Loan_Balance
Credit_Score


Delinquency Rate by Employment Status
Delinquency Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Unemployed
Employed
Self-employed
employed
EMP
retired
Employment_Status


Delinquency Rate by Credit Card Type
Delinquency Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Business
Student
Gold
Standard
Platinum
Credit_Card_Type


Distribution of Credit Utilization
Frequency
0
5
10
15
20
25
30
35
0.2
0.4
0.6
0.8
1.0
Credit Utilization