Tata Gen AI Project

Geldium Delinquency Risk Assessment Strategy

Exploratory Data Analysis (EDA) Summary Report

1. Introduction.

This financial dataset focused on customer credit behavior, used for predicting delinquency (late or missed loan payments) based on financial and demographic features, including payment history.

2. Dataset Overview.

This dataset consists of 500 rows and 19 columns.Each row represents a unique customer, identified by a Customer_ID, with various attributes related to their financial profile and payment history over six months.

Key dataset attributes:
- Number of records: 500 records

- Key variables(Data types):
a. Demographic : Age (Integer), Income (float), Employment_Status (string), Location(string).

b. Financial : Credit_Score (Integer), Credit_Utilization (Float), Loan_Balance, Debt_to_Income_Ratio (Float).

c. Behavioral : Missed_Payments (Integer), Delinquent_Account (Binary Integer), last 6 payment statuses (string).

3. Data Quality Issues.

Missing Values :
Income, Loan_Balance, Credit score missing in some rows.

Inconsistent Data :
Employment status - Inconsistent capitalization (e.g., "Employed" vs. "employed" vs. "EMP") suggests a need for standardization. "EMP" and "employed" represent the same category.
Account Tenure - A value of 0 months may indicate new accounts or potential data entry errors.

Duplicates :
All Unique Customer ID's, No duplicate customers exist.

Outliers & Anomalies :
Credit_Utilization - A value 1.025 exceeds typical maximum of 1.0 for utilization ratio suggesting potential error.

Debt_to_Income_ratio - Minimum value 0.1 appears frequently, which is unusually low and indicates data entry error or default value for missing data.
Balances : Some customers have very low balances or missing balances, which may not align with realistic loan scenarios

4. Handling the Data Quality Issues.

Missing Values :
Identifying and addressing missing data is critical to ensuring model accuracy.

- **Variables with missing values** : Income, Loan Balance.
- **Missing data treatment** : Imputation.
Imputed missing values in Income and Loan_Balance using the Median.

Inconsistent Data :
- Column affected : Employment_status
- Inconsistent Data Standardized : Normalized "Employed," "employed," and "EMP" to a single category "Employed" in Employment_status column.

Outliers & Anomalies :
- Columns Affected : Credit_Utilization, Debt_to_Income_ratio
- Investigated Outliers : Validated Credit_Utilization columns values above 1.0 and Debt_to_Income_Ratio column values of 0.1.

Takeaways from the Data Cleaning :
This dataset is cleaned within Excel, addressing missing values, inconsistencies, and Outliers while maintaining data integrity.

a) **Median Imputation:** Used median for Income and Loan_Balance as it's robust to Outliers.

b) **Outliers Thresholds**: Credit_Utilization > 1.0 and Debt_to_Income_Ratio = 0.1 were flagged as Outliers based on domain knowledge.

c) **Consistency Check:** Loan-to-Income ratio > 1.

d) **Account_Tenure**: Zero tenure may be valid for new accounts.

5. Key Findings and Risk Indicators.

A. Age_Bin and Delinquency_Risk

1) **Pattern**-

- The 61-74 age group has the highest "High Risk" percentage (17.60%), followed by 41-50 (16.80%), suggesting that older customers are more delinquency-prone, possibly due to fixed incomes, added family responsibilities and health-related financial strain.
- Older age groups may face reduced earning capacity, increasing reliance on loans and risk of default.

2) **Unexpected findings -** The 18-30 age group (15.60% High Risk) is the moderately risky, contradicting the common assumption that younger customers are riskier, possibly due to lower loan balances or parental support.

3) **Action -** Focus risk assessments on 51-74 age groups; investigate why 18-30 shows lower risk despite potential financial inexperience.

B. Income_Bin and Delinquency_Risk

1) **Pattern -**

- The 100001-150000 income range leads with 24.60% "High Risk," followed by 50001-100000 (18.80%) and 150001-200000 (18.60%), showing that middle-to-high income groups are most vulnerable, potentially due to larger loans commitments or lifestyle spending., despite higher earnings.
- Higher income does not guarantee financial stability, especially if paired with significant debt, making this a key risk indicator.

2) **Unexpected findings-** The 0-50000 group (11.60% High Risk) is the lowest, and its "Low Risk" (4.40%) is the smallest, suggesting very low-income customers either avoid large debts or limited loan access, contrary to expectations of higher vulnerability.

3) **Action -** Target the 100001-150000 income range for risk management; explore debt-to-income ratios for low-income, low-risk cases.

C. Account_Tenure_Bin and Delinquency_Risk

1) **Pattern** -

- The 0-5 month tenure bin has the highest "High Risk" (22.40%), reinforcing that new accounts are a significant risk factor, are particularly vulnerable to delinquency, likely due to untested credit behavior.
- Short account tenure suggests a lack of established payment history, making it a critical risk factor.

2) **Unexpected findings-** The 11-15 month bin (16.60% High Risk) has the highest "Low Risk" (7.60%), suggesting a stabilization period where risk decreases, which contrasts with the persistent risk in 16-19 months (17.60%).

3) **Action -** Target new accounts (0-5 months) for closer monitoring; explore additional factors (e.g., income) influencing longer-tenure risks.

D. Loan Balance and Delinquency_Risk

1) **Pattern** -

- The 25001-50000 loan balance range remains the highest "High Risk" (23.40%),suggesting that moderate loan balances pose a significant delinquency risk, possibly due to manageable but burdensome debt levels.
- Customers with moderate loan balances may overextend their finances, increasing the likelihood of missed payments.

2) **Unexpected findings-** The 0-25000 range (17.80% High Risk) is close to 75001-100000 (18.20%), and its "Low Risk" (8.00%) is the highest, suggesting small loans carry balanced risk and resilience.

3) **Action -** Focus on the 25001-50000 range for risk mitigation; investigate why low and high balances also show elevated risk.

E. Delinquency Risk by Employment.

1) **Pattern -**

- The "Employed" category has the highest "High Risk" percentage (35.20%), significantly outpacing "Unemployed" (13.60%), "retired" (13.20%), and "Self-employed" (11.60%), suggesting that employed individuals are the most delinquency-prone group.

2) **Unexpected findings-** "Employed" also has the highest "Low Risk" (12.80%), indicating a bimodal distribution where employment status correlates with both high risk and stability, possibly due to varied income levels or loan commitments.

5. AI & GenAI Usage

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns.

AI prompts used:

- 'Summarized key patterns in the dataset and identify anomalies.'
- 'Suggested an imputation strategy for missing income values based on industry best practices.'

6. Conclusion.

- The consistent 73.60% "High Risk" across bins suggests a dataset skewed toward risk, possibly due to sampling or criteria overlap (e.g., Delinquent_Account = 1 driving the total).

Identified Anomalies:

Age_Bin Anomaly:
Low Risk in 18-30

Observation: The 18-30 age group has 15.60% "High Risk" but the highest "Low Risk" (5.00%), indicating younger customers are less risky than expected.

Possible Explanation: Lower loan balances or external support (e.g., family) may protect this group, contradicting typical assumptions of youthful financial inexperience.

Action: Investigate Income_Imputed and Loan_Balance_Imputed for 18-30 to confirm protective factors.

Income_Bin Anomaly:
Low Risk in 0-50000

Observation: The 0-50000 income group shows the lowest "High Risk" (11.60%) and "Low Risk" (4.40%), suggesting limited risk exposure despite low income.

Possible Explanation: These customers may avoid large loans or receive assistance, reducing delinquency risk, which is unusual for low-income profiles.

Action: Validate income data and check for external support or smaller loan sizes.

Account_Tenure_Bin Anomaly:
Persistent Risk in 16-19 Months

Observation: The 16-19 month tenure bin (17.60% High Risk) is only slightly lower than 0-5 months (22.40%), despite longer tenure.

Possible Explanation: Established accounts may face new financial pressures (e.g., loan renewals), or the dataset includes chronic defaulters.

Action: Analyze payment history trends for 16-19 month accounts to identify risk triggers.

Loan_Balance_Bin Anomaly:
Similar Risk in 0-25000

Observation: The 0-25000 range (17.80% High Risk) is close to 75001-100000 (18.20%), despite the loan size difference.

Possible Explanation: Small loans may be issued to high-risk borrowers, or low balances reflect partial repayments by defaulters.

Action: Review borrower profiles for 0-25000 loans to assess risk segmentation.