Exploratory Data Analysis (EDA) Summary Report Template

1. Introduction

This report provides a preliminary analysis of Geldium's customer dataset to support Tata iQ's analytics team in understanding the data's structure and identifying early risk indicators. The findings will inform enhancements to the delinquency risk model and targeted intervention strategies.

2. Dataset Overview

Number of records: 500

Key variables:

- Income: Annual income of the account holder
- Loan_Balance: Current loan balance
- Credit_Score: A numerical credit rating
- Credit_Utilization: Proportion of credit used
- Missed_Payments: Count of missed payments
- Delinquent_Account: Binary indicator (1 if delinquent)
- Debt_to_Income_Ratio: Ratio of debt payments to income
- Account_Tenure: Age of the account in years
- Month_1 to Month_6: Monthly repayment status

Data types:

- Numerical: Income, Loan_Balance, Credit_Score, Debt_to_Income_Ratio, etc.
- Categorical: Month_1 to Month_6 (values include "On-time", "Late", "Missed")
- Binary: Delinquent_Account

Notable anomalies:

- Several accounts have Account_Tenure of 0, which may indicate newly opened accounts or data entry errors.
- Some entries have very low Income (< €5,000)—potential outliers or possible data issues.

3. Missing Data Analysis

- Variables with missing values:
- Income: 39 missing values
- Loan_Balance: 29 missing values
- Credit_Score: 2 missing values
- Missing data treatment (proposed):

Column	Treatment	Justification
Income	Imputation	Income is critical; use median imputation to reduce skew from high incomes.
Loan_Balance	Imputation	Use median to retain distribution shape and avoid removing valuable data.
Credit_Score	Imputation	Only 2 missing values; mean imputation is acceptable here.

4. Key Findings and Risk Indicators

Correlations with delinquency (top 5 predictors):

Variable	Correlation with Delinquent_Account
Income	0.045
Credit_Score	0.035
Debt_to_Income_Ratio	0.034
Credit_Utilization	0.034
Age	0.023

Early risk indicators:

- Low income correlates slightly with delinquency.
- High credit utilization and high debt-to-income ratios appear to be weak but consistent indicators.
- Account_Tenure = 0 occurs more often in delinquent accounts (possible risk factor).
- Missed monthly payments are intuitively and likely strongly linked to future delinquency, though not yet numerically analyzed.

Unexpected anomalies:

- Some entries show Loan_Balance under €1,000 despite substantial income—possibly suggesting unusual financial behavior or data error.
- Several delinquent accounts show Account_Tenure = 0 — worth investigating whether early defaults are common.

5. AI & GenAI Usage

Generative AI tools (ChatGPT) were used to perform rapid EDA, highlight issues, and suggest imputation strategies.

Example AI prompts used:

- *"Summarize key patterns, outliers, and missing values in this dataset. Highlight any fields that might present problems for modeling delinquency."*
- *"Identify the top 3 variables most likely to predict delinquency based on this dataset. Provide brief reasoning."*
- *"Suggest an imputation strategy for missing income values based on industry best practices."*

6. Conclusion & Next Steps

This initial analysis highlights several mild correlations with delinquency, especially related to income, credit score, and account tenure. The dataset contains some missing values and anomalies that require careful preprocessing. Next steps include:

- Performing imputation based on strategies outlined above.
- Encoding categorical payment history columns for modeling.
- Conducting feature importance analysis with tree-based models.
- Evaluating if early tenure and sequential payment behavior predict default.