Exploratory Data Analysis (EDA) Summary Report

1. Introduction

The purpose of this Exploratory Data Analysis (EDA) report is to examine the quality, structure, and key characteristics of the Geldium delinquency dataset prior to predictive modeling. The primary goal is to identify data quality issues, understand customer behavior patterns, and detect early risk indicators that may influence payment delinquency. Insights from this analysis will guide data cleaning, feature selection, and model development in subsequent stages.

2. Dataset Overview

This dataset contains historical customer-level information collected from loan servicing and collections systems. It includes demographic attributes, credit account details, and payment behavior indicators used to assess delinquency risk.

Key Dataset Attributes

- Number of records: *[Insert total number of rows after loading the dataset]*
- Key variables:
    - Customer Demographics: Age, income, employment status
    - Account Information: Loan amount, outstanding balance, account tenure
    - Credit Behavior: Credit utilization ratio, credit limit
    - Payment History: Number of missed payments, days past due, last payment amount
    - Target Variable: Delinquent (binary indicator of payment default)
- Data types:
    - Numerical: Income, balance, credit utilization, days past due
    - Categorical: Employment status, customer segment
    - Binary: Delinquency flag

Initial Observations

- Minor inconsistencies in categorical labels were observed.
- Outliers were detected in credit utilization and outstanding balance fields.

- No significant duplicate records were identified.

3. Missing Data Analysis

Identifying and addressing missing data is critical to maintaining model accuracy and reducing bias. Several key financial and behavioral variables contained missing values.

Key Missing Data Findings

- Variables with missing values:
    - Income
    - Payment history fields
    - Credit utilization
- Missing data treatment:
    - Deletion: Columns with excessive missing data (>40%) were removed.
    - Imputation: Median imputation was applied to numeric variables to reduce outlier impact.
    - Synthetic Data Generation: Missing income values were filled using a normal distribution to preserve realistic financial patterns.

This approach ensured minimal data loss while maintaining the integrity of key predictive features.

4. Key Findings and Risk Indicators

Analysis of relationships between variables and delinquency outcomes revealed several strong risk indicators.

Key Findings

- A strong positive correlation exists between missed payments, days past due, and delinquency.
- Customers with high credit utilization ratios show significantly higher default risk.
- Short account tenure combined with early missed payments indicates elevated delinquency probability.

Unexpected Anomalies

- Some long-tenure customers exhibited sudden delinquency despite previously stable payment behavior.
- Moderate-income customers with extremely high utilization displayed delinquency patterns similar to low-income segments, suggesting hidden financial stress.

These findings highlight the importance of behavioral features over demographic variables in delinquency prediction.

5. AI & GenAI Usage

Generative AI tools were leveraged to accelerate exploratory analysis, suggest data-cleaning strategies, and surface meaningful patterns. AI-assisted insights supported faster decision-making while adhering to industry best practices.

Example AI Prompts Used

- *"Summarize key patterns, anomalies, and missing values in this dataset."*
- *"Suggest an imputation strategy for missing income values based on financial industry best practices."*
- *"Identify the top variables most likely to predict delinquency."*

AI outputs were validated through statistical analysis and domain knowledge before implementation.

6. Conclusion & Next Steps

This EDA identified key data quality challenges, including missing values and outliers, which were addressed using structured cleaning and imputation strategies. Payment behavior and credit utilization emerged as the most influential predictors of delinquency.