Geldium’s — EDA Report

1. Dataset Structure and Context

The dataset contains customer-level financial and behavior features that influence delinquency risk. Numerical fields such as Income, Credit Score, Debt-to-Income Ratio, and Credit Utilization play a critical role in understanding a user’s financial stability. Categorical fields like Employment Status, Credit Card Type, and Location help segment customer behavior patterns. Understanding the structure of these variables lays the foundation for EDA and model-building.

2. Missing Values — Deep Dive

Missing data is one of the most important challenges in financial datasets. Critical variables such as Income, Credit Utilization, and Monthly Payment History often contain missing entries due to customer non-disclosure, system recording errors, or lack of recent financial activity. Each missing pattern needs careful treatment to avoid bias in model predictions. For example, missing Income values may correlate with unstable employment, while missing payment history might itself be a risk indicator. Appropriate imputation strategies prevent data loss while maintaining fairness and model stability.

3. Data Quality Observations

We observed anomalies such as Credit Utilization values exceeding 100%, which is not practically possible. Extreme outliers in Income and Loan Balance may also affect model training by pulling distributions away from natural patterns. Detecting and resolving such anomalies early ensures that the model represents realistic financial behavior and strengthens prediction accuracy.

4. Early Risk Indicators — Detailed Interpretation

High Credit Utilization, low Credit Score, and frequent missed payments show strong correlation with delinquency. A high Debt-to-Income Ratio indicates that customers are financially over-leveraged, increasing the probability of default. Employment Status differentiates risk levels—unemployed or unstable workers face higher delinquency likelihood. These insights are crucial for feature engineering and model design. Furthermore, categorical delinquency spread shows how certain customer segments are statistically more vulnerable, helping stakeholders design targeted interventions.

5. Visual Patterns and Insights

Distribution patterns for Age, Loan Balance, and Credit Utilization highlight potential clusters and segments within the dataset. For example, younger customers may have lower income and higher utilization, while older customers may show more stable patterns. Visualizing these distributions helps contextualize numerical summaries and gives the analytics team a more intuitive understanding of customer behavior.

6. Recommendations Before Modeling

Before training predictive models, it is recommended to:

- Impute missing values using robust strategies (median, mode, or predictive imputation).
- Cap extreme outliers in Income, Debt-to-Income Ratio, and Credit Utilization.
- Create additional engineered features such as Utilization Flags, Payment Stability Scores, and Regional Risk Factors.
- Normalize skewed variables such as Income to improve model convergence.
- Validate fairness across demographic/categorical groups to prevent biased predictions.

7. Summary

This dataset provides a strong foundation for building a delinquency prediction model, but requires structured cleaning and preparation. Missing values, anomalies, and risk patterns must be addressed before modeling. With the recommended cleaning steps and insights captured, the dataset will be ready for feature engineering and predictive modeling in the next phase.