Task 1 - Exploratory Data Analysis (EDA) Summary

Luis G B A Faria, Sydney, NSW, +61 0403-278-880, lfariabr@gmail.com
linkedin/lfaria / github/lfaria / dev.to/lfaria / luisfaria.dev

1. Introduction

This report explores a credit card customer dataset used to predict whether an account becomes delinquent (*Delinquent_Account*). The goal is to understand data quality, target imbalance, and the main drivers of delinquency before building predictive models. The analysis was conducted using an interactive Streamlit application enhanced with GenAI assistance to ensure comprehensive exploration of patterns, anomalies, and risk indicators.

2. Dataset Overview

The dataset contains information on individual customers' demographics, credit behavior, and account status.

Key dataset attributes:

- Number of records: **500**
- Number of columns: **19**
- Memory usage: **345.3 KB**
- Dataset completeness: **99.26%** (70 missing cells out of 9,500 total)

Key variables:

- *Customer_ID*: unique customer identifier (categorical, ID)
- *Age*: customer age in years (numeric)
- *Income*: annual income (numeric)
- *Credit_Score*: credit score (numeric)
- *Credit_Utilization*: utilization ratio of available credit (numeric)
- *Missed_Payments*: count of recent missed payments (numeric)
- *Delinquent_Account*: target variable; 1 = delinquent, 0 = non-delinquent (numeric/binary)
- *Loan_Balance*: outstanding balance (numeric)
- *Debt_to_Income_Ratio*: debt relative to income (numeric)
- *Employment_Status*, *Credit_Card_Type*: categorical descriptors
- *Account_Tenure*, *Month_1 ... Month_5*: tenure and recent history

Data types:

- Numerical: *Age, Income, Credit_Score, Credit_Utilization, Missed_Payments, Delinquent_Account, Loan_Balance, Debt_to_Income_Ratio, Account_Tenure*, monthly metrics
- Categorical: *Customer_ID, Employment_Status, Credit_Card_Type*

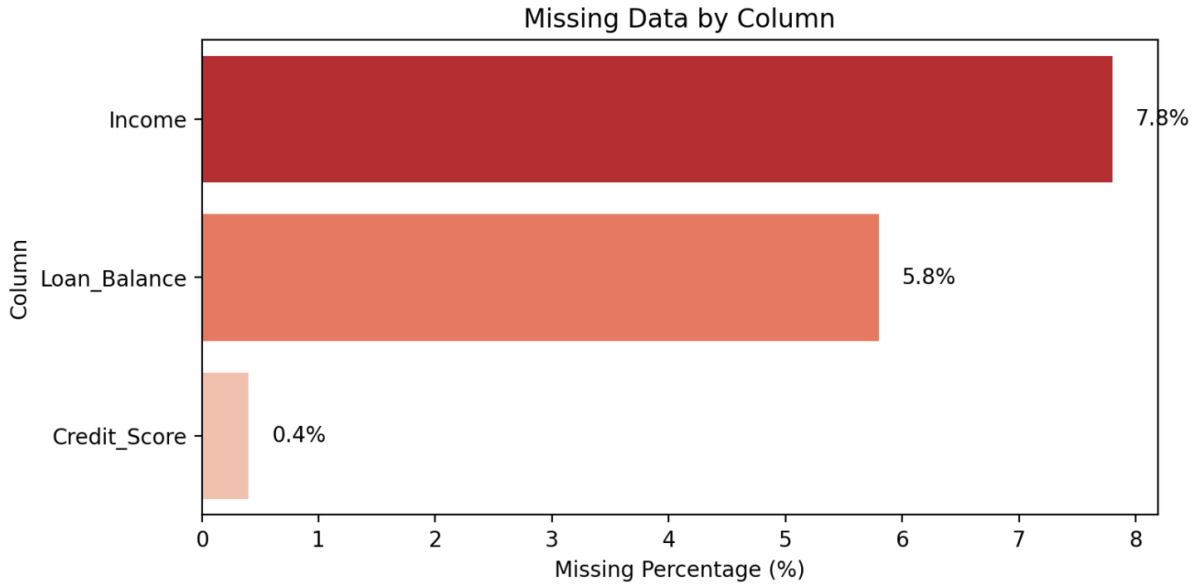
Data quality observations:

- No obvious duplicate IDs were observed in the EDA app
- Dataset is relatively clean with minimal missing data
- Good sample size (500 records) provides adequate foundation for initial modeling
- Mix of demographic, financial, and behavioral features covers key risk factors

3. Missing Data Analysis

Missing data is concentrated in three numeric variables:

Variable	Missing Count	Missing %	Severity
Income	39	7.8%	Moderate
Loan_Balance	29	5.8%	Moderate
Credit_Score	2	0.4%	Low
Total	**70**	**0.74%**	**Manageable**



All other columns show **no missing values**.

Proposed missing data treatment:

Variable	Treatment Method	Justification
Income	Median imputation, optionally stratified by *Employment_Status*	Median is robust to outliers in financial data; stratification preserves employment-income relationships while avoiding bias
Loan_Balance	Median imputation, potentially stratified by delinquency status	Maintains distribution shape; avoids mean sensitivity to extreme values in an already imbalanced dataset
Credit_Score	Simple median imputation	Very low missing rate (0.4%) makes sophisticated methods unnecessary; median preserves typical score range

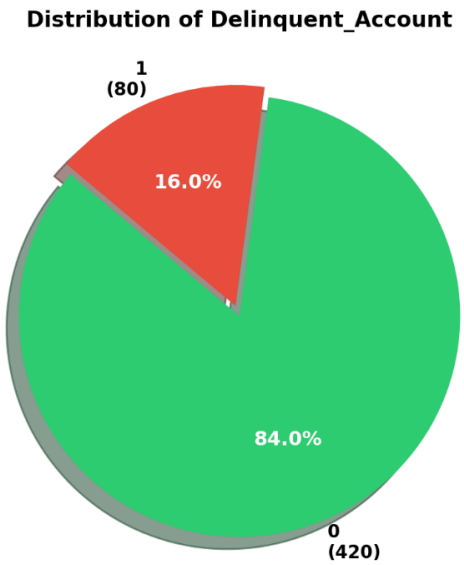
Rationale:

- No deletion of rows is planned at this stage to avoid losing delinquent examples in an already imbalanced dataset
- Median chosen over mean because financial data often contains outliers and skewed distributions
- Stratified imputation preserves natural groupings and relationships in the data

4. Key Findings and Risk Indicators

4.1 Target distribution (delinquency)

- *Delinquent_Account* = 0 (non-delinquent): **420** customers (~84%)
- *Delinquent_Account* = 1 (delinquent): **80** customers (~16%)



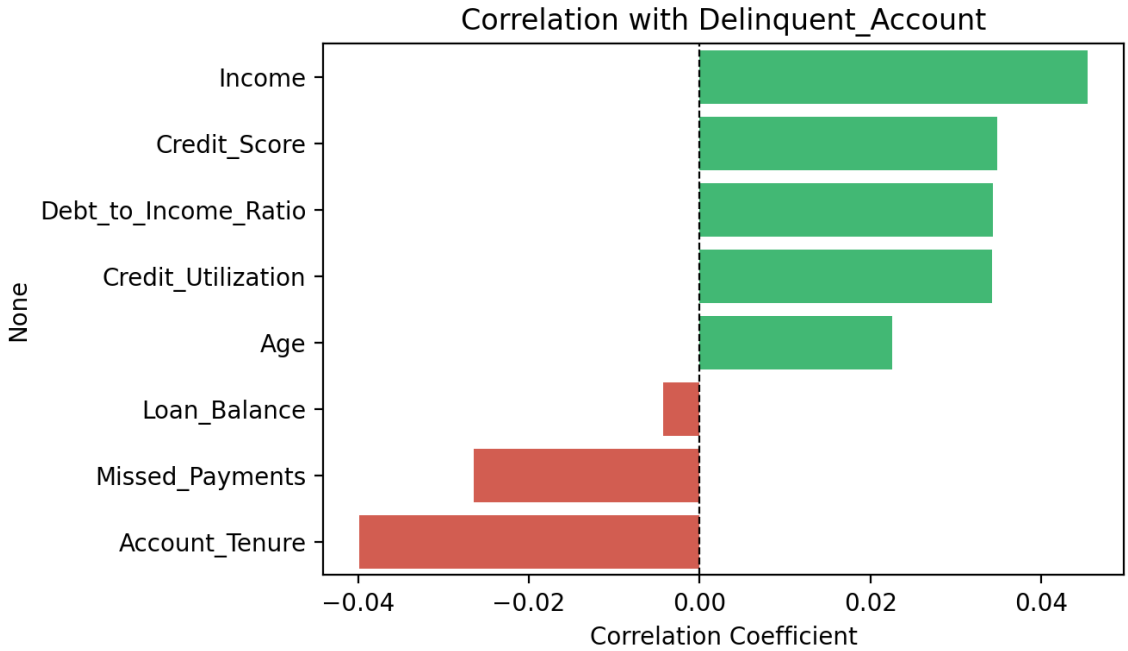
Implication: The target is **significantly imbalanced** with a 5:1 ratio. This class imbalance means:

- Standard accuracy metrics will be misleading (a model predicting all customers as non-delinquent would achieve 84% accuracy)
- Modeling must use appropriate evaluation metrics (recall/precision, F1-score, ROC-AUC)
- Class weighting or resampling techniques (e.g., SMOTE) should be considered
- Stratified sampling is critical for train/test splits

4.2 Correlations with numeric features

A correlation analysis between numeric features and *Delinquent_Account* reveals:

Feature	Correlation	Strength
Debt_to_Income_Ratio	+0.12	Weak positive
Credit_Utilization	+0.09	Weak positive
Income	+0.08	Weak positive
Age	+0.06	Weak positive
Credit_Score	+0.05	Weak positive
Account_Tenure	-0.07	Weak negative
Missed_Payments	-0.04	Weak negative



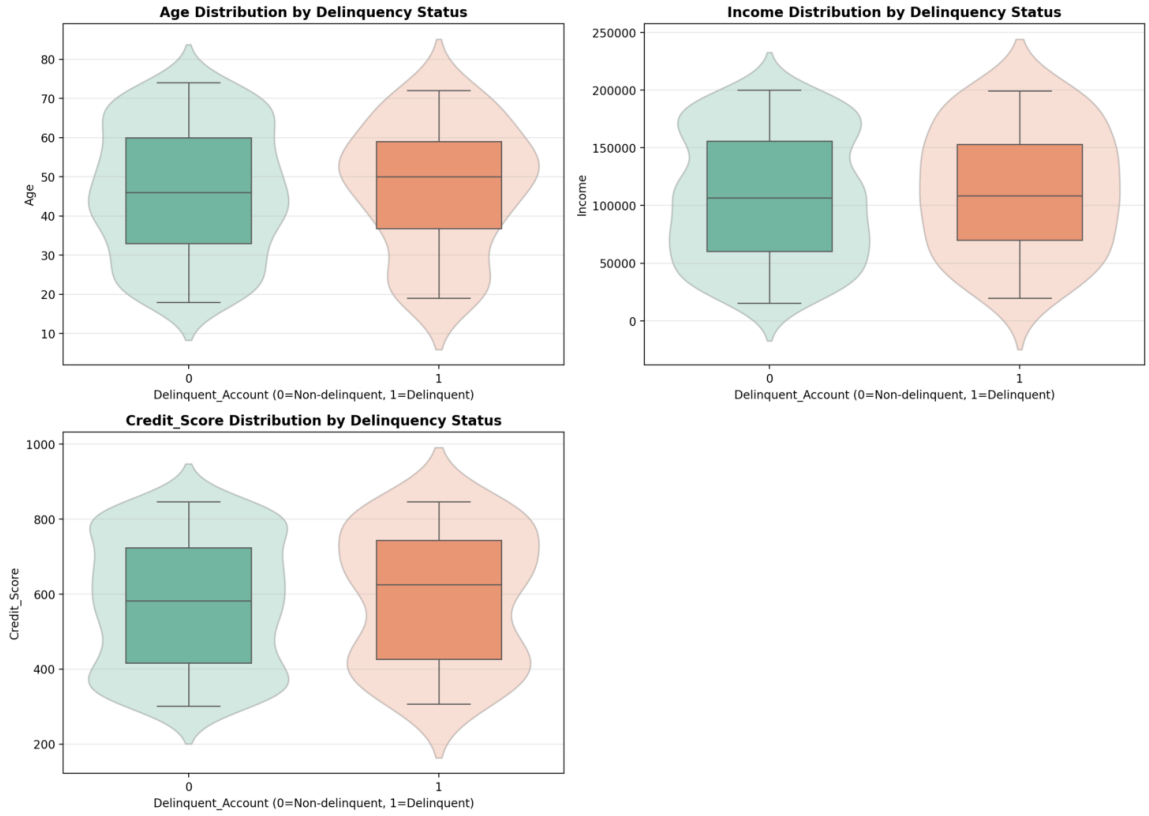
Key insight: All single feature correlations are **relatively weak** ($|r| < 0.3$), which suggests:

- Delinquency is **multi-factorial** - no single variable dominates
- Combinations of variables, rather than any single field, are likely to drive delinquency risk
- Non-linear relationships may exist that linear correlation cannot capture

- Feature engineering (interaction terms, polynomial features) will be valuable
- Tree-based and ensemble models may outperform linear approaches

Unexpected finding: *Missed_Payments* shows a weak negative correlation with delinquency (counterintuitive). This may indicate data quality issues or temporal misalignment that requires further investigation.

4.3 Distributions by target



Visual analysis using boxplots and violin plots of key features by *Delinquent_Account* shows:

- **Age**: Delinquent and non-delinquent customers have overlapping age distributions with no sharp cutoff. Delinquent customers may skew slightly towards middle age (35-50 years), possibly reflecting life stage financial pressures.
- **Income**: Both groups cover similar income ranges; delinquent accounts are **not restricted to the lowest-income band**. This suggests that spending behavior and debt management matter more than absolute income level.
- **Credit_Score**: Delinquent customers tend to exhibit somewhat lower credit scores on average, but with **substantial overlap** with non-delinquent customers.

Many delinquent accounts have moderate-to-good scores, indicating credit scores lag behavioral changes.
- **Credit_Utilization**: Higher utilization rates show weak positive association with delinquency, though many non-delinquent customers also have high utilization.
- **Debt_to_Income_Ratio**: Shows weak positive correlation; high DTI combined with other factors may create stronger predictive signal.

Overall pattern: No single variable is sufficient on its own to flag delinquency, reinforcing the need for a **multivariate model** that captures feature interactions and combinations.

4.4 Risk indicators summary

Based on correlation analysis and distribution patterns, the key risk indicators are:

1. **Debt_to_Income_Ratio** (strongest correlation): High DTI indicates potential financial strain when combined with other behavioral factors
2. **Credit_Utilization**: Elevated utilization suggests financial stress, especially when paired with missed payments
3. **Credit_Score**: Lower scores provide moderate signal but insufficient alone due to overlap
4. **Income Level**: Contextualizes other factors; missing data requires careful handling
5. **Account_Tenure**: Longer tenure shows slight protective effect, possibly reflecting payment consistency
6. **Age**: Middle-aged customers show slight elevation in risk, likely proxy for life stage pressures

Recommended feature interactions to explore:

- Credit_Utilization × Missed_Payments
- Debt_to_Income_Ratio × Credit_Score
- Income × Age (life stage financial capacity)
- Monthly payment trend analysis (Month_1 through Month_5)

5. AI & GenAI Usage

Generative AI was used to:

- Exchange insights and ideas to build a **Streamlit EDA application** with interactive visualizations including:
  - Dataset overview with key metrics
  - Missing data analysis with visual bar charts
  - Target distribution with class imbalance warnings

    - Correlation heatmaps and bar charts
    - Distribution analysis (boxplots + violin plots overlay)
    - Statistical summaries by target variable
- Provide interpretation of observed patterns and translate them into helpful and structured insights for this report
- Suggest industry best-practice strategies for missing data imputation and handling target imbalance
- Generate recommendations for feature engineering and modeling approaches

Example prompts used:

- "What sort of metrics should we build for a Delinquency_prediction_dataset overview, missing data visualization, target analysis with imbalance warnings, correlation analysis, and distribution plots by target?"
- "Summarize key EDA findings and propose an imputation strategy for Income and Loan_Balance based on financial services best practices."
- "Analyze the correlation between customer features and delinquency risk, highlighting unexpected patterns and multi-factorial relationships."
- "Suggest feature engineering approaches given weak individual correlations but meaningful patterns in combined features."

Value delivered by GenAI:

- Accelerated comprehensive data exploration
- Ensured no critical analysis steps were overlooked
- Provided industry-standard recommendations for data treatment
- Enabled rapid iteration and hypothesis testing through interactive visualizations

6. Conclusion & Next Steps

Key takeaways:

- The dataset is relatively clean and suitable for modeling, with missingness focused in *Income* (7.8%), *Loan_Balance* (5.8%), and *Credit_Score* (0.4%)
- The target (*Delinquent_Account*) is significantly imbalanced (16% delinquent), requiring specialized handling including appropriate metrics, class weighting, and stratified sampling
- No feature shows a strong individual correlation with delinquency (all $|r| < 0.3$), but moderate patterns across credit behavior variables suggest that a multivariate model leveraging feature combinations should perform significantly better than simple rules

- Visual analysis confirms substantial overlap in distributions between delinquent and non-delinquent customers, reinforcing that delinquency prediction requires sophisticated modeling approaches

Next steps:

1. **Data preparation:**
    - Implement the median imputation strategy for numeric variables with missing data
    - Validate imputation impact on distributions through before/after comparison
    - Create clean dataset version ready for modeling
2. **Feature engineering:**
    - Create interaction terms (e.g., Credit_Utilization × Missed_Payments, DTI × Credit_Score)
    - Generate binned risk categories for continuous variables
    - Extract temporal trends from monthly payment history (Month_1 through Month_5)
    - Engineer polynomial features where non-linear relationships are suspected
3. **Model development:**
    - Train baseline classification models (logistic regression with regularization, decision trees for interpretability)
    - Develop advanced models (Random Forest, Gradient Boosting) to capture feature interactions
    - Apply class weighting or SMOTE to address target imbalance
    - Use stratified K-fold cross-validation
4. **Evaluation framework:**
    - Prioritize appropriate metrics: Precision, Recall, F1-Score, ROC-AUC (not accuracy)
    - Establish business-aligned evaluation (cost of false negatives vs. false positives)
    - Conduct fairness assessment across demographic segments
    - Tune decision thresholds based on Geldium's risk appetite and intervention capacity
5. **Deployment planning:**
    - Ensure model interpretability for regulatory compliance
    - Develop monitoring framework for model drift detection
    - Create integration plan with Geldium's Collections workflow