Dataset Analysis Report

1. Executive Summary

The analysis of the provided dataset (500 records) reveals significant anomalies in the relationships between traditional risk factors and the delinquency outcome. Contrary to standard financial behaviour, variables like Credit Score and Missed Payments show weak or counter-intuitive correlations with Delinquency. This suggests potential issues with data quality, synthetic data generation, or label definitions that require immediate investigation before modeling.

2. ETL & Missing Data Analysis

- **Data Quality Checks**:
  - **Standardization**: The Employment_status column contained inconsistent values (EMP, employed, Employed), which were standardized to Employed.
  - **Completeness**: No duplicates were found.
- **Missing Values Detected**:
  - Income: 39 missing values (7.8%).
  - Loan_Balance: 29 missing values (5.8%).
  - Credit_Score: 2 missing values (0.4%).

3. Imputation Strategy & Execution

The following industry-standard strategies were applied:

- **Income (Synthetic Generation)**:
  - *Strategy*: To preserve the natural variance of the data, missing values were generated using a **Normal Distribution** assumption based on the observed non-null data (μ ≈ $108k, σ ≈ $53.6k).
  - *Result*: 39 realistic income entries were synthesized and filled.
- **Loan Balance & Credit Score**:
  - *Strategy*: **Median Imputation** was used. This is the best practice for financial variables to avoid skewing the data with outliers (unlike the Mean).
  - *Result*: Missing entries were filled with the median values ($45,776 for Loan Balance; 586 for Credit Score).

- **Credit Utilization (Best Practice Proposal)**:
  - Although this dataset had 0 missing values for Credit_Utilization, the best practice for handling missing data in this specific feature is **Regression Imputation** or **Median Imputation**. Because utilization is strictly bounded (0-100% usually) and often skewed, the Mean can be misleading.

4. Key Findings & Risk Indicators (Anomalies)

The GenAI-driven analysis highlights several "Red Flags" where the data contradicts expected financial patterns.

- **Paradox 1: The "Good" Credit Score Risk**
  - *Finding*: Delinquent accounts have a *higher* average Credit Score (591) than non-delinquent accounts (575).
  - *Implication*: The Credit Score variable may be unreliable or inversely labeled in this dataset.
- **Paradox 2: Missed Payments Inverse Relationship**
  - *Finding*: Customers flagged as Delinquent averaged *fewer* missed payments (2.85) than good customers (2.99).
  - *Implication*: The Missed_Payments count is effectively useless or broken for predicting Deliquent_Account in this specific batch.
- **Weak Signals**:
  - Correlation analysis shows all major features (Credit_Utilization, Income, Debt_to_Income) have near-zero correlation (< 0.05) with the target variable.

5. Actionable Insights & Next Steps

- **Critical Data Review**: The analytics team must verify the source of the Deliquent_Account label. The current patterns suggest the label may be randomly assigned or derived from a variable not present in this dataset.
- **Feature Engineering**: Given the weak linear relationships, explore non-linear models (e.g., Random Forest) or interaction features (e.g., Debt_to_Income * Utilization) to see if hidden patterns exist.
- **Modeling Recommendation**: Do not proceed to production modeling with this specific dataset until the inverse relationships (High Score = High Delinquency) are explained.

6. Final Data State

The dataset has been cleaned, standardized, and fully imputed. It is ready for further experimental modeling, though caution is advised regarding the validity of the target variable relationships.