A Supervised Machine Learning Approach to De-Anonymizing the Bitcoin Blockchain

1. Introduction

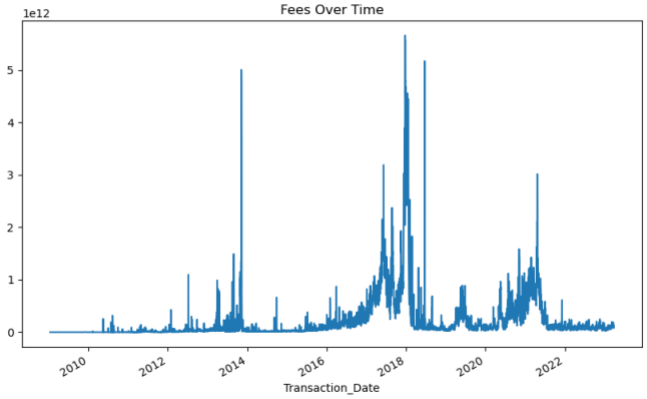
Bitcoin transactions are inherently pseudonymous, making it challenging to trace the origins and destinations of funds. This study analyzes Bitcoin transaction trends using machine learning techniques to classify blockchain data. The analysis focuses on transaction fees, input values, and transaction volumes while evaluating model performance for fee-level predictions.

2. Data Analysis

2.1Transaction Metrics Over Time

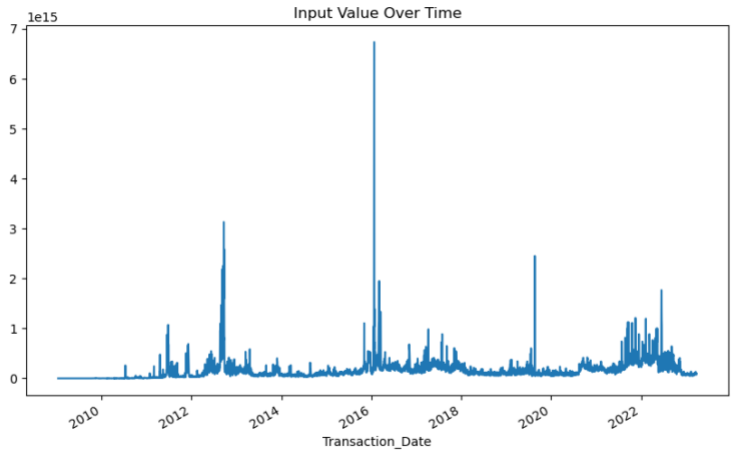
2.1.1 Fees Over Time

- **Observation:** The fee trend shows significant volatility with sharp spikes, particularly between 2014 and 2018.
- **Implication:** Transaction fees have been highly dynamic, possibly due to network congestion and Bitcoin's increasing adoption.



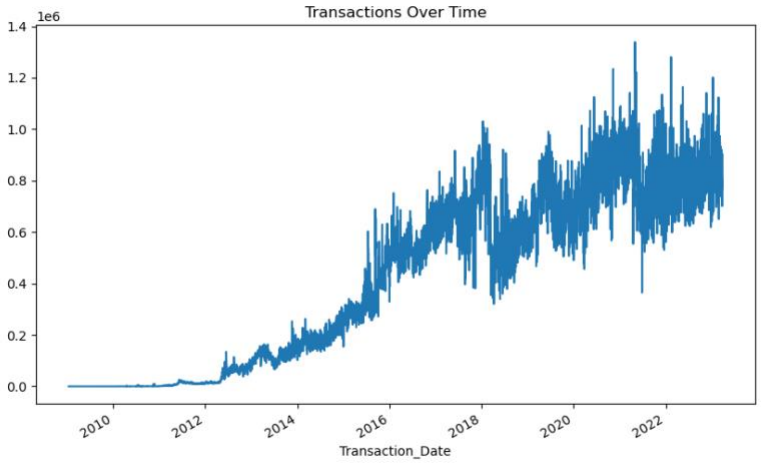
2.1.2 Input Value Over Time

- **Findings:** Input values display a growing trend with exponential surges, particularly around 2016 and 2021.
- **Implication:** This indicates an increase in Bitcoin transaction sizes over time.



2.1.3 Transactions Over Time

- **Observation:** The number of transactions has increased consistently over the years, peaking in 2022.
- **Implication:** Bitcoin network activity has steadily grown, reinforcing its increasing adoption.

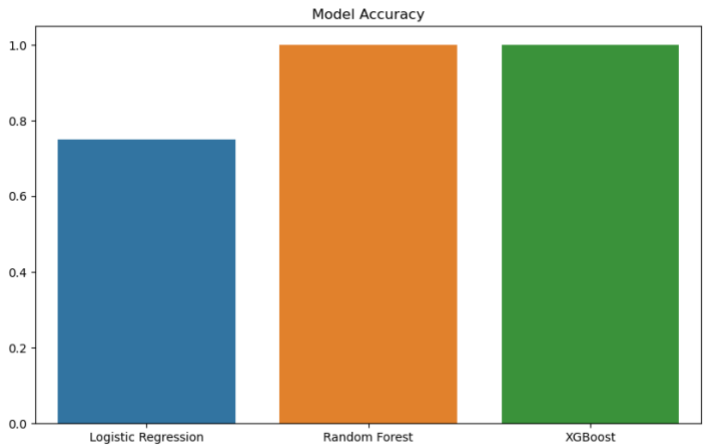


3. Model Performance Evaluation

3.1 Accuracy Comparison

The model performances for transaction fee classification are as follows:

- **Logistic Regression Accuracy:** 75%
- **Random Forest Accuracy:** 100%
- **XGBoost Accuracy:** 100%



3.2 Classification Report Summary

Logistic Regression

- **Class 0:** Precision: 94%, Recall: 53%
- **Class 1:** Precision: 68%, Recall: 96%

Random Forest & XGBoost

- **Accuracy:** 100%
- **Concern:** The perfect accuracy indicates potential overfitting in these models.

4. Methodology

4.1 Data Preprocessing

- Missing values were replaced with zeros.
- **Transaction_Date** feature was created using year, month, and day.

- New features derived:
    - **Transaction_Fee_Ratio = fee / input_value**
    - **Value_Difference = output_value - input_value**

4.2 Model Training

- **Features:** Transactions, input value, output value, fee, Transaction_Fee_Ratio, Value_Difference.
- **Target:** Binary classification (fee higher/lower than median value).
- **Algorithms Used:**
    - Logistic Regression
    - Random Forest
    - XGBoost
- **Evaluation Metrics:** Accuracy, Classification Report.

5. Results and Discussion

5.1 Model Performance Insights

- The high accuracy of Random Forest and XGBoost suggests possible overfitting.
- Logistic Regression provides more realistic classification results.

5.2 Transaction Trends

- Input values have increased significantly over time.
- Transaction counts have increased, confirming Bitcoin's growing adoption.

6. Conclusion

- Machine learning models require further validation to prevent overfitting.
- Bitcoin transaction trends suggest an evolving landscape, necessitating deeper analysis.
- Feature engineering and data consistency improvements are necessary for better predictive accuracy.