**Bitcoin price prediction using machine learning: An approach**

**to sample dimension engineering**

**1. Summarize of the Research**

**Objective**

To improve the accuracy of Bitcoin price predictions using machine learning techniques while addressing the role of sample size and feature engineering. The study explores both daily and high-frequency (5-minute interval) price predictions.

**Methodology**

* Based on granularity, the prediction was divided into two categories: **daily data** (low-frequency) and **5-minute interval data** (high-frequency).
* High-dimensional feature engineering was used, combining metrics such as trading volume, gold spot prices, and Google Trends.
* Evaluate statistical models (e.g., logistic regression, linear regression analysis) and machine learning algorithms (e.g., random forest, XGBoost, support vector machine, LSTM).
* Data sets used from CoinMarketCap (daily data) and Binance (5-minute time interval data).

**Findings**

* For daily predictions, less complex statistical algorithms performed better, with logistic regression having an accuracy of 66%.
* For predicting the 5-minute time interval, the complex machine learning algorithm performed best, with LSTM having an accuracy of 67.2%.
* High-dimensional feature sets make up for less complex models for daily prediction, but low-dimensional datasets gain with complex machine learning algorithms.

**Conclusions**

The study emphasizes that model complexity and data granularity and feature dimensions should go hand in hand. Statistics works best with low-frequency, high-dimensional data and machine learning with high-frequency, low-dimensional data.

**Key contributions**

* Sample dimension engineering was introduced to Bitcoin price forecasting.
* It was shown how feature granularity affects the performance of machine learning techniques.
* It provided a framework that balances simplicity and complexity for different forecasting scenarios.

**Limitations**

* Exploration of other granularities beyond daily and 5-minute intervals is limited.
* Not all machine learning models (e.g. ARIMA or extended RNN models) were evaluated.
* Data sources and features were limited, and potential influences such as broad sentiment or macroeconomic indicators were omitted.

**2. Identify Research gaps**

**Identified Research gaps**

* **Granularity Limitation:** The analysis focused only on two granularities (daily and 5-minutes interval), not including intermediate or any other intervals.
* **Limited Feature Space:** Despite the use of high-dimensional features, the analysis did not include additional, advanced sentiment analysis (e.g., social media) or macroeconomics.
* **Cross-Market Analysis:** Bitcoin alone was analyzed. Including relations with additional cryptocurrencies (e.g., Ethereum, stocks) could make predictions even better.
* **Model Diversity:** Excluded other potential predictive models such as advanced ARIMA, GRU, or transformer-based time series models.

**Opportunities for further studies**

* **Granularity Exploration**: Expand to other granularities (e.g., hourly, 15-minutes) to see how granularity affects predictive accuracy.
* **Sentiment Analysis:** Apply text mining to social media, news, and forums for real-time sentiment information to enhance feature engineering.
* **Comparative Analysis:** Compare predictive performance of new machine learning approaches such as transformers.
* **Multi-Asset Framework:** Detect cross-market relations and their impact on Bitcoin price determination.