Predicting BTC, ETH, and SOl Prices using a multivariate data set: A Machoine Learning Approach

Eren Muller

July 9, 2024

Contents

Chapter 1

Introduction

Chapter 2

Literature Review

Predicting stock prices is not only multifaceted but also regarded as one of the most challenging applications of time-series analysis (Kara et al., 2011). Despite the extensive empirical research on predicting stock prices, the literature specifically addressing crypto stock price predictions remains significantly less developed. Accurate and robust stock price predictions are of utmost importance for developing effective trading strategies (Leung, Daouk, & Chen, 2000).

This literature review will focus on past works and attempts at predicting both stock and cryptocurrency prices. We will also delve into the theoretical backgrounds regarding the feasibility of predicting stock prices, examining past attempts by analyzing their methodologies and the machine learning (ML) and deep learning (DL) methods employed. These papers utilized nuanced feature sets, including various market variables. Therefore, we will review literature that explores which feature sets were used and why. Additionally, this review aims to shed light on the impact of marketing-related variables on the predictive power of stock and crypto markets. In particular, we will investigate the influence of social media sentiment, such as Reddit discussions, on cryptocurrency prices. This comprehensive review seeks to provide a deeper understanding of the methodologies and variables that contribute to the predictability of financial markets.

2.1 Efficient market hypothesis

Eugene Fama's Efficient Market Hypothesis (EMH) asserts that stock prices reflect all available information, implying that they always trade at their fair value. Fama (1970) argues that since all new information is immediately incorporated into stock prices, consistently predicting market movements or outperforming the market through traditional stock-picking is essentially impossible. This underpins the argument for passive index fund investing, which aims to match market returns rather than exceeding them.

Behavioral finance challenges Eugene Fama's Efficient Market Hypothesis by arguing that psychological factors and irrational behavior of investors can lead to market inefficiencies. Scholars like Daniel Kahneman and Amos Tversky, who develop prospect theory, demonstrate that cognitive biases such as overconfidence and loss aversion significantly influence investor decisions, often leading to predictable and systematic errors. Kahneman and Tversky (1979) mention that these biases cause stock prices to deviate from their true values, creating opportunities for superior returns through strategic trading, contrary to EMH's assertion that such opportunities are fleeting or non-existent. Behavioral finance thus provides a framework to understand why and how markets might

not be entirely efficient.

In 2022, Ho-Jun Kang and his colleagues conduct research to investigate the presence of the Efficient Market Hypothesis (EMH) in the cryptocurrency market. Their study involves testing 893 cryptocurrencies, and the results reveal that only a small fraction of these currencies adhere to the EMH. Specifically, Kang et al. (2022) find that only 54 cryptocurrencies (6%) follow the weak-form EMH, and just 24 (3%) adhere to the semi-strong-form EMH. These findings suggest that the cryptocurrency market demonstrates limited efficiency in information processing. Moreover, the study concludes that most cryptocurrencies do not incorporate past prices or new information into their market prices.

In 2020, Vu Le Tran authors a paper examining the Efficient Market Hypothesis (EMH) within the cryptocurrency market. Tran (2020) concludes that market efficiency is highly variable over time, particularly noting significant inefficiencies before 2017. Tran observes that, over time, the cryptocurrency market is becoming increasingly efficient. Among the cryptocurrencies tested, Litecoin emerges as the most efficient, while Ripple is identified as the least efficient.

2.2 Past attempts at predicting the stock/crypto currency market

In the study conducted by Kara, an artificial neural network (ANN) and support vector machine (SVM) are employed to predict stock price movements on the Istanbul Stock Exchange. The independent variable in this research is a binary indicator reflecting whether the stock price will move up or down the following day. Kara (2011) finds that the SVM, particularly with a polynomial activation function, outperforms all other algorithms, including ANN and backpropagation network (BPN), achieving an accuracy of 71.5%. The feature set for this study comprises various technical analysis (TA) indicators such as the Moving Average Convergence Divergence (MACD), Moving Average (MA), and the stochastic oscillator %K (K%).

In another study focusing on trend deterministic data for stock price prediction, a classification model is used to forecast the up or down movement of stock prices. Patel (2015) mentions that this research incorporates a feature set consisting of binary variables indicating whether a technical indicator suggests an upward or downward trend. The highest performing model in this study is the random forest, which achieves an accuracy of 83.5%. However, a noted limitation of this approach is the binary nature of the technical indicators. The study suggests that incorporating additional levels to represent the degree of movement, such as 'slightly up', 'slightly down', and 'barely down', could enhance the model's accuracy.

In 2020, Chen conducts a study to predict Bitcoin prices using various machine learning methods, including logistic regression and long short-term memory (LSTM) networks. Chen (2020) finds that by utilizing 5-minute interval price data, the model achieves an accuracy of 66%, outperforming more complex neural network models. The feature set in this study is comprehensive, incorporating not only Bitcoin price data but also external factors such as gold spot prices, property and network data, as well as trading and market information.

In another study, Weng (2018) attempts to predict short-term stock prices using ensemble methods. The feature set in this research is diverse, comprising historical stock

prices, well-known technical indicators, sentiment scores derived from published newspaper articles, trends in Google searches, and the number of visits to Wikipedia pages. The study demonstrates impressive results, predicting the next day's stock prices with a mean absolute percentage error (MAPE) of less than 1.5%. The best-performing algorithms in this research are boosted decision trees, including XGBoost and AdaBoost.

Usami et al. conduct a study to predict the Karachi Stock Exchange (KSE) using various machine learning algorithms. They employ a classification model to forecast whether the market will go up or down. The feature set for this study is extensive, including oil rates, gold and silver rates, interest rates, foreign exchange (FEX) rates, news and social media feeds, simple moving averages (SMA), and autoregressive integrated moving average (ARIMA) data. Usami et al. (Year) find that the best performing model is the multilayer perceptron (MLP), a type of artificial neural network (ANN), alongside support vector regression (SVR).

In 2022, Mailagaha Kumbure et al. conduct a comprehensive literature review on the application of machine learning and data used for stock market forecasting. This review examines a total of 138 articles related to machine learning in stock markets, providing a detailed overview of the models, markets, and feature sets used in these studies. Mailagaha Kumbure et al. (2022) highlight that the most used machine learning methods are neural networks, support vector machines/support vector regression (SVM/SVR), and fuzzy theories. Additionally, they note that most of these papers incorporate technical indicators in their feature sets.

2.3 Social Media and Sentiment Analysis in the Role of Predicting Stock/Crypto Prices

The influence of social media on Bitcoin prices has been a topic of significant interest in recent research. Feng Mai's 2018 study employs textual analysis and vector error corrections to demonstrate a clear link between social media sentiment and Bitcoin price movements. Mai (2018) shows that bullish posts on social media platforms are associated with higher future Bitcoin prices. This suggests that social media sentiment is a valuable predictor of Bitcoin price fluctuations, highlighting the impact of public opinion and social discourse on cryptocurrency markets.

In addition to social media, online search activity also correlates with Bitcoin price movements. Kristoufek (2013) analyzes Google Trends and Wikipedia page visits, finding strong correlations between these data points and Bitcoin prices. This research suggests that increased online searches and Wikipedia activity, reflecting public interest and awareness, can significantly influence Bitcoin market trends. Complementing these findings, Wesley S. Chan's 2003 study on stock market prediction through news sentiment reveals that positive newspaper headlines often lead to overvaluation of stocks, while negative headlines result in undervaluation. Chan (2003) further notes that this sentiment effect is more pronounced in smaller market capitalization stocks and that investors typically react slowly to sentiment changes. Together, these studies underscore the significant role of public sentiment, whether expressed through social media, search activity, or news headlines, in influencing financial markets.

The predictive power of social media sentiment on cryptocurrency prices has been further explored in recent studies. Olivier Kraaijeveld's 2020 research focuses on the influence of Twitter sentiment on the returns of major cryptocurrencies. Kraaijeveld

(2020) concludes that Twitter sentiments indeed have predictive power over cryptocurrency prices, utilizing a lexicon-based sentiment analysis. The study highlights that news disseminated through Twitter can rapidly alter investor sentiments, leading to immediate and significant price movements. This finding emphasizes the crucial role of real-time sentiment analysis in anticipating market trends and price fluctuations in the volatile cryptocurrency market.

Similarly, news sentiment shows a notable impact on Bitcoin prices. Lavinia Rognone's 2020 study analyzes the effect of unscheduled news on Bitcoin compared to traditional currencies using intra-day data from January 2012 to November 2018. Rognone (2020) finds that Bitcoin often reacts positively to news, whether positive or negative, indicating a high level of enthusiasm among investors towards Bitcoin, unlike traditional stock markets. However, specific negative news, such as reports of fraud and cyber-attacks, have adverse effects on Bitcoin prices. The study utilizes RavenPack's real-time news data and employs a Vector Auto-Regressive Exogenous (VARX) model for the analysis. In parallel, Wasit Khan's 2020 research combines social media and news sentiment to predict stock market movements, using a dataset from Twitter and Yahoo Finance. Khan (2020) demonstrates that their predictive model achieves an accuracy of 80% after filtering out spam tweets, underscoring the significant impact of integrated sentiment analysis on market predictions. These studies collectively highlight the importance of sentiment analysis in understanding and forecasting market dynamics across various financial assets.

2.4 Conclusion

The literature review underscores the multifaceted and challenging nature of predicting stock and cryptocurrency prices, emphasizing the importance of robust predictions for effective trading strategies. Research on stock price prediction is extensive, utilizing various machine learning (ML) and deep learning (DL) methods. Studies such as those by Kara et al. (2011) and Patel (2015) highlight the effectiveness of SVM and random forest models, respectively, in forecasting stock prices using technical indicators. Similarly, Chen (2020) and Weng (2018) demonstrate the predictive power of logistic regression, LSTM networks, and ensemble methods for Bitcoin and stock prices, leveraging comprehensive feature sets that include market variables and sentiment scores. The review also notes the evolving efficiency of cryptocurrency markets, with studies like those by Kang et al. (2022) and Tran (2020) revealing limited adherence to the Efficient Market Hypothesis (EMH), indicating significant information processing inefficiencies. The impact of sentiment analysis on market predictions emerges as a critical theme. Research by Mai (2018) and Kraaijeveld (2020) establishes the predictive power of social media sentiment on cryptocurrency prices, while Kristoufek (2013) and Chan (2003) demonstrate similar effects for online search activity and news sentiment on Bitcoin and stock markets. Studies such as Rognone (2020) and Khan (2020) further validate the significant influence of real-time sentiment, integrating social media and news data to achieve high prediction accuracy. These findings collectively highlight the importance of incorporating diverse feature sets, including sentiment analysis, to enhance the predictability of financial markets and challenge traditional notions of market efficiency. article amsmath booktabs longtable

geometry a4paper, margin=1in

2.5 Datasets

This section details the datasets and technical indicators used for the machine learning algorithms to predict the next day's classification of BTC, ETH, and SOL prices. The datasets include historic prices, Google Trends, Wikipedia page views, and top 100 Reddit posts for sentiment analysis and topic modeling. Additionally, several technical indicators are engineered from the main price data tables.

2.5.1 Historic Prices

For BTC, ETH, and SOL, we collected historic prices with the following attributes:

- **Price:** The closing price of the cryptocurrency.
- **Open:** The opening price.
- **High:** The highest price during the trading period.
- **Low:** The lowest price during the trading period.
- **Close:** The closing price.
- **Volume:** The trading volume.
- **Pct Change:** The percentage change in price.

2.5.2 Google Trends

Google Trends data is collected for BTC, ETH, and SOL to understand the relative search interest over time. This data helps gauge public interest and potential market movements.

2.5.3 Wikipedia Page Views

Wikipedia page views for BTC, ETH, and SOL are used to measure the general public's interest in these cryptocurrencies. This data serves as a proxy for market sentiment and public awareness.

2.5.4 Reddit Posts

Top 100 Reddit posts per respective cryptocurrency subreddit (BTC, ETH, SOL) are collected. The data includes:

- **Title:** The title of the post.
- **Description:** A brief description of the post.
- **Date Posted:** The date the post was published.
- **Number of Likes:** The number of likes the post received.
- **Number of Comments:** The number of comments on the post.

- **Overall Score:** The overall score of the post.

This data will be used for sentiment analysis and topic modeling to gain insights into community sentiment and trending topics.

2.5.5 TVL (Total Value Locked)

Total Value Locked (TVL) represents the total capital held within a specific protocol, encompassing all the assets staked, loaned, or otherwise utilized. This metric provides insightful analytics on market confidence and the overall health of blockchain ecosystems. Higher TVL figures often correlate with increased media coverage and user interest, enhancing the platform's marketing appeal and perceived stability.

2.5.6 Gold, S&P 500, and VIX Historic Data

- **Gold Data:** Includes historical data on gold prices, including the opening, high, low, and closing prices. Gold is often considered a safe-haven investment.
- **S&P 500 (SNP) Data:** Includes historical data for the S&P 500 index, covering opening, high, low, and closing prices, as well as trading volume. It is a key indicator of the U.S. stock market's overall performance.
- **VIX Data:** The VIX measures market expectations of near-term volatility conveyed by S&P 500 stock index option prices. It reflects investor uncertainty and market sentiment.

2.5.7 Technical Indicators and Trading Signals

This project includes various technical analysis indicators calculated using pandas. Below are the formulas used for each indicator, formatted for clarity.

Indicator **Trading Signal Explanation**	**Formula**
10-Day Moving Average (10D MA) A price above the 10D MA suggests upward momentum, signaling a potential buy opportunity.	$\text{10D MA} = \frac{1}{10} \sum_{i=0}^{9} \text{Price}_{t-i}$
30-Day Moving Average (30D MA) When the price is above the 30D MA, it indicates a medium-term upward trend, suggesting a buy signal.	$\text{30D MA} = \frac{1}{30} \sum_{i=0}^{29} \text{Price}_{t-i}$

Indicator Trading Signal Explanation	**Formula**
Stochastic Oscillator %K An upward %K signal indicates buying pressure, suggesting a bullish trend.	$\%K = 100 \times \frac{\text{Close} - \text{Low}_{14}}{\text{High}_{14} - \text{Low}_{14}}$
Stochastic Oscillator %D An upward %D signal shows continued buying interest, confirming the bullish trend.	$\%D = \frac{1}{3}(\%K_t + \%K_{t-1} + \%K_{t-2})$
Relative Strength Index (RSI) RSI above 70 indicates overbought conditions (sell signal), below 30 indicates oversold (buy signal).	$\text{RSI} = 100 - \frac{100}{1 + \frac{\text{Average Gain}}{\text{Average Loss}}}$
Momentum Positive momentum indicates a strong upward trend, suggesting a buy signal.	$\text{Momentum} = \text{Price}_t - \text{Price}_{t-n}$

Table 2.1: Technical Indicators and Formulas (Part 1)

Indicator Trading Signal Explanation	**Formula**
MACD An upward MACD signal indicates bullish momentum, suggesting a buy opportunity.	MACD = 12D EMA − 26D EMA
Commodity Channel Index (CCI)	$\text{CCI} = \frac{\text{Price} - \text{MA}}{0.015 \times \text{MAD}}$

Indicator Trading Signal Explanation	**Formula**
CCI above 100 indicates overbought (sell), below -100 indicates oversold (buy).	
Bollinger Bands A price above the upper band suggests overbought conditions (sell), below the lower band suggests oversold (buy).	Upper Band = MA + 2 × STD, Lower Band = MA − 2 × STD
Fibonacci Retracement Used to identify potential reversal levels in the price movement.	Levels at 23.6%, 38.2%, 50%, 61.8%, and 100%
Average True Range (ATR) Higher ATR values indicate higher market volatility.	$\text{ATR} = \frac{1}{n}\sum_{i=0}^{n-1}\text{TR}_i$, $\text{TR} = \max(\text{High} - \text{Low}, \lvert\text{High} - \text{Close}_{prev}\rvert, \lvert\text{Low} - \text{Close}_{prev}\rvert)$

Table 2.2: Technical Indicators and Formulas (Part 2)

2.6 Final Dataset Overview

The final dataset combines historical prices, Google Trends, Wikipedia page views, Reddit posts, TVL, and various technical indicators. These diverse data sources provide a comprehensive view of market dynamics and sentiment, enhancing the predictive power of the models.