Intra-day Bitcoin Price Prediction with Machine Learning and Ensemble Methods

by

John Tsang[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

Fall 2021 – MAT 4900 – Undergraduate Research Project

Supervisor: Professor Aaron Smith

Abstract

This project compares the performance among five basic machine learning models to predict the intra-day five-minute average Bitcoin prices from August 5th to August 31st, 2021 on Coinbase and attempts to use ensemble methods (1) stacking, (2) bagging and (3) random forest to improve forecasting performance. The forecast evaluation criteria entail forecast accuracy, directional accuracy and profitability. The results find that the random walk model outperforms other models and ensemble methods in the forecast and directional accuracy. However, ensemble methods are helpful to increase the profitability of basic machine learning models.

[1]Email: wtsan094@uottawa.ca

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Bitcoin, the first decentralized digital currency in the world, emerged in October 2008. Since then, Bitcoin has received lots of media attention in the global financial market. Hence, some of the largest investment banks and companies have probed the Bitcoin market, and explore Bitcoin's value as a commodity for investment. The most notable of such entities include JP Morgan and Tesla.

More than ten years after the creation of the first Bitcoin, the Bitcoin market is becoming mature. Along with more individuals, exchanges and institutions participating in the market, more variants of Bitcoin become available in the market such as Ethereum and Dogecoin. Moreover, on September 7th, 2021, El Salvador in Central American became the first country to recognize Bitcoin as its national currency.

While the prices are stable for most commodities whose markets are mature, Bitcoin price is very volatile. For instance, the average daily close Nasdaq Bitcoin Reference Prices (NQBTC) in August 2021 is almost four times as large as that in August 2020 (USD $45,604.63 vs. USD $11,650.86.) The range of the daily close NQBTC in August 2021 is also almost four times as large as the average daily close NQBTC in August 2021 (USD $45,604.63 vs. USD $12,171.70.)

This significant daily volatility attracted market participants to join the Bitcoin day-trade market. Indeed, during the COVID-19 pandemic, many young adults took part in the Bitcoin market for fear of missing out on the gains [Menton, 2021]. Some of them even became full-time Bitcoin day-traders to participate in a large volume of Bitcoin transactions in a day. As a result, top crypto-exchanges recorded a 200-500 per cent jump in the number of day-traders during the pandemic [Dave, 2021].

Because of the popularity of day-trading Bitcoins, this project aims to predict the average Bitcoin prices on Coinbase, one of the most popular Bitcoin exchange platforms, from August 5$^{\text{th}}$, 2021 UTC 12:00:00[2] to August 31$^{\text{st}}$, 2021 UTC 23:55:00 for every five-minute interval. These average intra-day prices are helpful to investors interested in day-trading Bitcoin and Bitcoin's financial derivatives by signifying Bitcoin price trends within a day.

This project starts with a literature review in section 2. This literature review identifies useful predictors of Bitcoin prices, including variables reflecting market expectations (such as Google Trends) and traditional demand- and supply-side factors (such as transaction volume). Another finding is the possibility of using ensemble methods to improve Bitcoin price forecasts. A brief investigation into inter-day Bitcoin price prediction in section 3 supports these two findings.

The prediction of intra-day Bitcoin prices every five minutes on Coinbase consists of two stages. The first stage uses the following five models to predict average Bitcoin prices every five minutes.

1. the random walk (RW) model
2. the linear regression (LR) model
3. the support vector regression (SVR)
4. the k-nearest neighbour regression (k-NN)
5. the decision tree regression (DT)

Because section 2 identifies using ensemble methods to improve forecasting performance as a possible research direction, the second stage attempts to use three ensemble methods (1) stacking, (2) bagging and (3) random forest, to combine models 3 to 6 to improve the forecasting performance. At both stages,

[2]The delay from August 1$^{\text{st}}$ is to allow for sufficient data points to make the first few predictions.

this project uses the rolling window strategy. Specifically, this project estimates and generates forecasts as if the forecasting process had run in the past (*pseudo-out-of-sample forecast*) in a fixed-sized window walking forward. At the end of these two stages, this project evaluates the forecasting performance by forecast accuracy, directional accuracy and profitability. Section 4 describes procedures at each stage in detail.

This project identifies 20 predictors from reviewing related literature in section 2 and consider these predictors for the first and the second stages. These 20 predictors capture the demand- and supply-side aspects of the Bitcoin market and how social media affects the market at different time intervals. On the one hand, the demand-side and supply-side factors include:

- moving averages of Bitcoin prices,
- moving averages of total Bitcoin transaction volume and
- percentages of the buy transaction volume

spanning different periods. These three groups of variables come from a transaction-by-transaction data set by Kaiko Digital Assets Data Provider[3] from France. The five-minute moving average Bitcoin price this project predicts also comes from this data set. Another demand-side and supply-side factor is the change in daily close Nasdaq Bitcoin Reference Price (NQBTC). On the other hand, social media-related predictors capture public interest in Bitcoin on Reddit, Twitter, Wikipedia, Google searches and YouTube searches.

Section 6 describes data sets this project uses, explains how this project derives each variable from these data sets and identifies data characteristics. Visual inspection finds out that the five-minute moving average Bitcoin price to be predicted had trend changes at least twice throughout August 2021 (the 153$^{\text{rd}}$ five-minute interval on August 5$^{\text{th}}$ and the 166$^{\text{th}}$ five-minute interval on August 19$^{\text{th}}$). Most predictors also have breakpoints in trends at similar points in time.

Neither a single model nor an ensemble method outperforms the other in terms of all three forecast evaluation criteria (section 7). At Stage 1, the RW model has the best forecast and directional accuracy. The average annual rates of return from all five base models are very close to zero. Therefore, all models suffer losses if including the one-per-cent trading cost on Coinbase. Results from Stage 2 reveal that ensemble methods help improve forecasting performance, especially profitability. Unfortunately, the improvement from ensemble methods does not suffice to cover the one-per-cent transaction cost.

Section 8 focuses on three further issues pertinent to the forecasting exercise in section 7 and discusses how the forecasting performance changes when considering each of these three issues. The three issues are trend changes, using subsets of predictors and replacing the standard bootstrap with a modified version. On the one hand, the exploration finds that considering model trend changes can improve forecasting performance. On the other hand, moving averages of Bitcoin prices for different time frames are the most relevant predictor to forecast accuracy and profitability. In contrast, volume-related predictors are more relevant to directional accuracy. Furthermore, using a modified version of the standard bootstrap that has corrected for time series serial dependence can improve forecasting performance compared with the standard bootstrap.

Section 9 discusses the computational challenges this project faces. Given the number of observations for this project, it takes a long time to cross-validate all rolling windows for each model. Therefore, this

[3] Many thanks for Kaiko Digital Assets Data Provider to fulfil the order for this project at a discounted price (€99) under minimal conditions.

project uses Google Cloud service[4] for better hardware and designs computer programs with parallel programming techniques. The combination of better hardware with multiple central processing units (CPUs) and specially designed computer programs allow multiple CPUs to run simultaneously to reduce the total runtime.

The last section briefly discusses three limitations in this project and how this project can develop with regard to these three issues. The first problem is the missed opportunity to collect free and valuable data on real-time Bitcoin Tweets. The second one is about the use of sequential bootstrap instead of the modified version of the standard bootstrap this project uses. The last problem addresses the computational and financial limitations of this project.

None of the models in this project can beat the random walk model in the forecast and directional accuracy. No models and methods have annual rates of return higher than the trading cost on Coinbase (one per cent) either. This inferior performance is due to the use of coarse tools for forecasting a well-studied market. However, this project can serve as a starting point and background study of developing more complicated and more accurate intra-day Bitcoin price forecasting models.

2 Literature Review

The number of publications pertinent to Bitcoin rose as Bitcoin prices roses from 2017. As Bitcoin prices increased 19 times from US$1,000 to US$20,000 in 2017 and rose to around US$30,000 in 2020, the number of Bitcoin-related research in 2020 on Scopus, an abstract and citation database, is close to five times more than that in 2016 (229 vs. 1311) [Aysan et al., 2021].

These research publications examine Bitcoin prices in different time periods and from different databases. Among these papers, there are two recurring themes related to this research project:

1. Bitcoin price determinants and

2. Bitcoin price prediction models

Other unrelated themes include Bitcoin price volatility, market efficiency, technological aspects of Bitcoin, whether Bitcoin is a commodity or a currency and associated legal issues [Kayal and Rohilla, 2019, Aysan et al., 2021].

2.1 Bitcoin Price Determinants

Researchers have identified three types of determinants of Bitcoin prices: (1) the traditional demand- and supply-side factors, (2) macroeconomic factors and (3) factors that affect market expectations of the future value of Bitcoin. Table 1 summarizes these variables.

Traditional Demand- and Supply-side Factors Bitcoins have neither intrinsic values nor the support from government decree (*i.e.* fiat money.) The traditional model of demand and supply remains common to assess determinants of Bitcoin prices. In general, demand-side factors play a more important role in Bitcoin prices than the supply-side because the Bitcoin generation algorithm has already fixed the future supply of Bitcoin. Therefore, any expected change in prices has already incorporated this knowledge [Polasik et al., 2015, Kristoufek, 2015]. Examples of demand-side factors include transaction volumes [Polasik et al., 2015], real interest rates (interest rates adjusted for inflation,) tax burden and investment freedom [Viglione, 2015].

[4] Many thanks for Google to provide CAD$383 credits for running programs with its cloud service.

Table 1: Summary of Significant Determinants of Bitcoin Prices

Class	Determinant	
Traditional Demand- and Supply-side Factors	• Transaction volumes • Real interest rates	• Tax Burden • Investment freedom
Market Expectations	• Google trends • Twitter Posts	• Reddit Posts

Macroeconomic Factors [Baek and Elbeck, 2015] find that the returns of Bitcoin are independent of monthly changes in most traditional macroeconomic factors in the U.S. These variables they investigate are (1) the consumer price index (CPI), (2) industrial production, (3) real personal consumption expenditures, (4) the S&P 500 index, (5) 10-year Treasury note, (6) euro exchange rate and (7) the national average unemployment rate. They, therefore, conclude that market participants drive fluctuations in the Bitcoin market. Findings from [Vaddepalli and Antoney, 2018] and [Vieira, 2017] also support the irrelevance of macroeconomic factors to Bitcoin prices.[5]

Market Expectations: the Effects of Social Media [Athey et al., 2016] account for Bitcoin price fluctuations by market participants' belief in Bitcoin's future value because they find that the most common use for Bitcoin is investment. Traditional factors mentioned above are, on the contrary, less important. [Ciaian et al., 2016] support the insignificance of the above traditional factors by emphasizing Bitcoin's speculative nature.

Many recent publications discuss the influence of social media on Bitcoin prices through affecting market participants' expected future Bitcoin prices. [Kristoufek, 2013] and [Aalborg et al., 2019] find that a Google search of the word "Bitcoin" may drive the prices of a crypto-currency. Moreover, [Dastgir et al., 2019] use the Copula-based Granger Causality in Distribution (CGCD) test to discover a bi-directional relationship between the Google Trends search queries and Bitcoin returns. With the use of wavelet coherence, [Phillips and Gorse, 2018] find that online factors such as Google search volume, Wikipedia views and the number of new posts per day on the `/r/bitcoin` subreddit are positively related to Bitcoin prices in the medium and the long term.

2.2 Bitcoin Price Prediction Models

In general, researchers use (1) traditional statistical methods and (2) machine learning methods for predicting Bitcoin prices. The first groups include time-series and econometric models while the latter group includes machine learning, reinforcement learning and deep learning models. According to a literature survey including 145 research publications during the 2010-2020 period [Khedr et al., 2021], machine learning is slightly more popular than traditional statistical (and econometric) methods for predicting Bitcoin prices (46% vs. 54%.) Table 2 summarizes some base models from these two groups. With these base models, researchers, then, attempt to improve model performance by augmenting these models with Bayesian statistics, regime switching and Markov-switching (e.g., [Kodama et al., 2017, Ramadani and Devianto, 2020, Wu, 2021, Jang and Lee, 2017].)

Because Bitcoin price prediction is still a developing topic, there is still no consensus on which base

[5] However, [Kjærland et al., 2018] find that the S&P 500 index is important for explaining Bitcoin prices.

Table 2: **Summary of Some Base Bitcoin Price Prediction Models**

Group	Models
Traditional Statistical Models	linear regression, autoregressive integrated moving average (ARIMA), autoregressive conditional heteroscedasticity (GARCH), wavelet coherence analysis, vector autoregressive process (VAR), vector error correction (VEC), autoregressive distributed lag (ARDL)
Machine Learning Models	support vector regression (SVR), regression tree (RT), artificial neural network (ANN), random forest (RF), logistic regression, gradient boosting, long short-term memory (LSTM), recurrent neural network (RNN), multilayer perceptron (MLP)

model performs better. However, the following reasons motivate researchers to deviate from using traditional statistical methods to predict Bitcoin prices.

- For commodity price forecasts, machine learning models generally deliver more accurate predictions than traditional statistical models.

- Assumptions of machine learning models are not as strong as traditional statistical models.

At this stage of the development of Bitcoin price prediction, relevant research publications provide some lessons related to this project [Khedr et al., 2021, Mezquita et al., 2021].

- **Support Vector Regression vs. Random Forest**: Random forest usually outperforms support vector regression for prediction because feature selection from random forest uses predictors with the highest information efficiency and random forest is more robust to outliers.

- **Ensemble methods**: Not many research publications adopt ensemble methods for Bitcoin price prediction because most of these publications are small scale. However, these publications yield satisfactory results [Mallqui and Fernandes, 2019, Chowdhury et al., 2020].

- **The role of social media on Bitcoin Price**: Social media is an important determinant of Bitcoin price, as discussed in section 2.1.

Another closely related theme is to compare the forecasting performance of different statistical and machine learning models. [Valencia et al., 2019] compare the performance of neural network, support vector regression and random forest to predict Bitcoin prices with market data and social media data from Twitter and find that neural networks outperformed the other three models. Moreover, [McNally et al., 2018] compare a variant of the recurrent neural network, the long short term memory (LSTM) model and the ARIMA model and find that the LSTM model performs the best.

Overall, Bitcoin is a unique financial commodity. Although macroeconomic factors significantly affect exchange rates and prices of most other financial commodities such as gold, these factors are not crucial to Bitcoin prices. On the contrary, variables reflecting market expectations, such as Google Trends, influence Bitcoin prices. Traditional demand- and supply-side factors such as transaction volume also affect Bitcoin prices. Moreover, machine learning models are popular among researchers to predict Bitcoin prices because forecasts from these models are generally more accurate. Although there is no consensus on which model delivers more accurate Bitcoin price forecasts, a possible research direction is to use ensemble methods to improve Bitcoin price forecasts.

3 Background: Inter-day Bitcoin Price Prediction

Section 2 finds out that there is no consensus as to which model provides the best Bitcoin price forecasts. Therefore, this section uses the inter-day Bitcoin price to develop an initial understanding of Bitcoin price prediction, despite the primary goal of intra-day Bitcoin price prediction. Specifically, the remainder of this section is an exercise of predicting daily Bitcoin prices with the five models mentioned in section 1. Evaluating these predictions help form expectations on the forecasting performance of intra-day Bitcoin prices.

This exercise uses the same methodology as Stage 1 of the intra-day Bitcoin price forecasting outlined in section 4. Forecast evaluation process and evaluation criteria are also the same. The criteria are mean absolute percentage error (MAPE), percentage of forecasts with the correct direction (directional accuracy) and the average rates of return (profitability). The following are some key points of the setup of this exercise.

Data The Bitcoin prices to be predicted are the open, close, low and high Nasdaq Bitcoin Reference Prices (NQBTC) from May 24th, 2019 to August 31st, 2021. Section 6.2 has a detailed description of the NQBTC.

Rolling window size The rolling window size of this exercise is 14 days. In other words, this exercise estimates each model except for the random walk (RW) model with observations from the previous 14 days and predicts the Bitcoin price next day.

Predictor For simplicity, this exercise uses the NQBTC of the last seven days (the first seven lags) as predictors. For instance, predictors corresponding to the one-day ahead close NQBTC prediction are the first seven lags of the close NQBTC.

Tables 3, 4 and 5 summarize the annual MAPE, directional accuracy and average rates of return.

Table 3: Annual Mean Absolute Percentage Prediction Error, One-day ahead Nasdaq Bitcoin Reference Prices, By Model, May 24th, 2019 to August 31st, 2021, per cent

(a) Open Prices

Year	RW	KNN	DT	SVR	LR
2019	3.0	3.9	3.2	6.4	3.5
2020	2.3	3.5	2.7	6.4	2.8
2021	3.4	4.9	4.1	8.1	3.6

(b) Close Prices

Year	RW	KNN	DT	SVR	LR
2019	2.8	3.8	3.0	6.4	3.3
2020	2.4	3.6	2.8	6.5	2.9
2021	3.5	4.9	4.1	8.1	3.7

(c) High Prices

Year	RW	KNN	DT	SVR	LR
2019	2.8	3.8	3.3	6.3	3.3
2020	2.3	3.6	2.7	6.4	2.9
2021	3.3	4.6	3.8	8.1	3.7

(d) Low Prices

Year	RW	KNN	DT	SVR	LR
2019	2.8	4.0	3.2	6.4	3.4
2020	2.6	3.7	3.0	6.6	3.0
2021	3.7	5.0	4.4	8.3	4.1

Table 4: Annual Directional Accuracy, One-day ahead Nasdaq Bitcoin Reference Prices, By Model, May 24th, 2019 to August 31st, 2021, per cent

(a) Open Prices

Year	RW	KNN	DT	SVR	LR
2019	41.5	54.0	54.5	51.0	48.5
2020	47.0	50.5	52.7	46.7	47.3
2021	46.9	49.8	51.9	52.3	58.0

(b) Close Prices

Year	RW	KNN	DT	SVR	LR
2019	46.5	50.0	58.0	46.5	46.5
2020	47.0	49.2	47.5	47.8	47.3
2021	44.4	54.3	49.4	54.7	58.8

(c) High Prices

Year	RW	KNN	DT	SVR	LR
2019	42.5	51.5	52.5	45.0	51.5
2020	42.6	52.2	51.1	50.0	48.4
2021	46.5	49.8	49.0	48.6	53.5

(d) Low Prices

Year	RW	KNN	DT	SVR	LR
2019	43.5	52.5	56.0	50.5	47.5
2020	42.6	50.5	48.1	48.1	49.2
2021	42.0	53.9	54.3	51.9	51.9

Table 5: Annual Average Annual Rates of Return, One-day ahead Nasdaq Bitcoin Reference Prices, By Model, May 24th, 2019 to August 31st, 2021, per cent

(a) Open Prices

Year	RW	KNN	DT	SVR	LR
2019	0.07	-0.06	-0.14	-0.02	0.28
2020	0.44	0.42	0.13	0.39	0.32
2021	0.31	0.30	0.07	0.35	0.36

(b) Close Prices

Year	RW	KNN	DT	SVR	LR
2019	0.03	0.10	-0.49	0.03	0.17
2020	0.46	0.51	0.50	0.38	0.39
2021	0.31	0.38	-0.17	0.32	0.35

(c) High Prices

Year	RW	KNN	DT	SVR	LR
2019	0.03	0.04	-0.33	-0.03	0.25
2020	0.46	0.46	0.19	0.40	0.32
2021	0.29	0.23	0.18	0.35	0.35

(d) Low Prices

Year	RW	KNN	DT	SVR	LR
2019	0.03	0.04	-0.08	-0.01	0.15
2020	0.50	0.31	0.35	0.37	0.43
2021	0.33	0.25	0.55	0.36	0.10

Some observations:

- No models consistently outperform other models each year for all three forecast evaluation criteria.
- Performance in one criteria does not imply the performance of another criteria. For example, although the RW model has the highest forecast accuracy for all four prices in all three years, this model does not have the best directional accuracy compared with other models.
- **Forecast accuracy (MAPE)**: The random walk model outperforms all other models in all cases. This characteristic is common in other financial time series, such as exchange rates.
- **Directional accuracy**: The best model in each year can only predict approximately 50 per cent to less than 60 per cent of directional changes right. This performance is slightly better than a wild guess.
- **Profitability**: Each transaction only takes place if the magnitude of the predicted percentage change is higher than one per cent. Among the four prices in the three years, most models only achieve average annual rates of return less than one per cent. On average, profits from transactions based purely on these predictions are very limited and much lower than the predicted threshold one per cent. Therefore, transaction cost can easily lead to losses.

The above observations establish the following expectations on intra-day Bitcoin price forecasts and motivate how this project predicts intra-day Bitcoin prices.

- The RW model outperforms all other models in terms of forecast accuracy.
- Using predictors additional to Bitcoin price lags may improve directional accuracy and profitability.
- Because no one single model outperforms other models in terms of all three criteria, combining predictions from each model may yield better predictions.

4 Methodology

This section outlines how this project predicts Bitcoin prices one step ahead. With the rolling window strategy and a window size of two days (576 observations), each model generates each one-step-ahead prediction with only observations from the previous two days. The forecasting consists of two stages:

- **Stage 1:** Prediction with Base Models
- **Stage 2:** Prediction with Ensemble Methods

The description of each stage also includes the program library for each method's implementation and a list of candidate hyperparameters for searches. At the end of each stage, this project evaluates each method's forecast accuracy, directional accuracy, and profitability and compares the forecasting performance at both stages.

4.1 Stage 1: Prediction with Base Models

Stage 1 aims to prepare forecasts from each base model for ensemble methods at Stage 2. This project uses the following five base models to predict the Bitcoin price of the next period with predictors from section 5.

1. the random walk (RW) model
2. the linear regression (LR) model
3. the support vector regression (SVR)
4. the k-nearest neighbour regression (k-NN)
5. the decision tree regression (DT)

Each of these five models uses the rolling window strategy with a fixed window size of two days to perform out-of-sample forecasts as if observations following the window were unknown. This project shifts the rolling window one observation forward after each prediction. Figure 1 illustrates the rolling window strategy with window size w and T total observations.

Figure 1: Illustration of the Rolling Window Strategy with Window Size w and T Total Observations

Within each rolling window, this project

1. identifies the model's optimal hyperparameter. The optimal hyperparameter is associated with the lowest absolute percentage forecast error[6] by cross-validation. The cross-validation procedure is as follows.

[6] Section 4.3 discusses absolute percentage forecast error.

(i) This project creates a sub-window. This sub-window starts from the first observation in the rolling window and has the size 90% of the rolling window.

(ii) Within the sub-window, this project estimates the model with each hyperparameter from a pre-specified list and then predicts the observation immediately following the sub-window.

(iii) This project calculates the absolute percentage error of the prediction from step 1(ii).

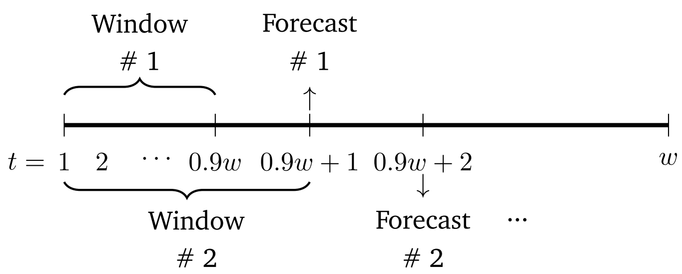
(iv) This project expands the sub-window created in step 1(i) to include the observation immediately following the sub-window. Repeat steps 1(ii) to 1(iii) until the size of the sub-window is the same as that of the rolling window. Figure 2 illustrates the first expanding sub-window in the first rolling window.

(v) The optimal hyperparameter is the one with the lowest average absolute percentage error.

2. estimates the model with the hyperparameter selected from step 1(v) and samples in the corresponding rolling window.

3. uses the estimated model to generate the one-step-ahead forecast.

Figure 2: Illustration of the First Expanding Sub-window in the First Size w Rolling Window



Algorithm 1 summarizes the prediction generation procedure for each model at Stage 1.
The remainder of this section briefly discusses each base model and its list of hyperparameters.

Random Walk (RW) Model The RW model (without drift) uses the last available Bitcoin price as the prediction. This model does not involve any parameter estimation. This project considers this simple model because of its superiority in forecast accuracy in many financial time-series, such as exchange rates.[7] As for the direction of change, this project extends the RW model to use the direction of change from the previous period to predict the direction of change in the next period.[8]

Linear Regression (LR) Model The linear regression model assumes a linear relationship between predictors and Bitcoin price and the error term following a normal distribution. To estimate this model, this project minimizes the objective function:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

where $\boldsymbol{y}$ is a vector of Bitcoin prices, $\boldsymbol{X}$ is the design matrix of all predictors, $\boldsymbol{\beta}$ is the vector of coefficients and $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ is the squared ℓ_2-norm of the error. This project uses module `LinearRegression` from Python library `scikit-learn` to estimate this model.

[7] This paradoxical situation is the Meese–Rogoff puzzle. [Moosa and Burns, 2014] provide a detailed account for this puzzle.

[8] This predicted direction of change contradicts the predicted prices. Specifically, the direction of change consistent with predicted prices is constant. However, predicting the Bitcoin price not changing for all periods is nonsensical. Because the RW model primarily serves as a benchmark, this project keeps this inconsistency.

Algorithm 1 Model Estimation and Forecast Generation Algorithm for Each Base Model at Stage 1

Input: $w_R > 0$ ▷ Rolling window size w_R
Input: `k-list` $\neq \emptyset$ ▷ List of hyperparameters `k-list`
Input: `model` $\in$ [LR,SVR,k-NN,DT] ▷ See page 8 for the full name of each model
Output: `prediction` ▷ `prediction` is a list for storing each prediction

```
Initialize prediction as a list to store each prediction
for each rolling window R = (x, y) of size w_R do
    Initialize dictionary error_dict[] as a list for each key k in k-list
    for each training set T_R and test set t_R = (x_{t_R}, y_{t_R}) in R do
        for each k in k-list do
            fitted-model ← the model fitted with T_R and k
            y_k ← prediction from fitted-model using x_{t_R}
            abs-pct-error ← |(y_k − y_{t_R}) / y_{t_R}|
            Append the list error_dict[k] with abs-pct-error
        end for
    end for
    optimal-k ← the key k where min_k average(error_dict[k])
    fitted-model ← the model fitted with R and optimal-k
    y_optimal-k ← prediction from fitted-model using the last observation's x
    Append the list prediction with y_optimal-k
end for
```

k-nearest Neighbour (k-NN) Regression Forecasts from this model are the simple averages Bitcoin price of the closest k observations. This model contains a hyperparameter k to be selected from the list $\{2, 3, 4, 5, 6\}$. This project uses module `KNeighborsRegressor` from Python library `scikit-learn` to estimate this model.

Support Vector Regression This model only accounts for certain observations outside a band for prediction. This project uses the polynomial kernel and selects the hyperparameter (the number of degrees) from $\{1, 2, 3\}$. This project uses module `SVR` from Python library `scikit-learn` to estimate this model.[9]

Decision Tree Regression This model generates a forecast by checking if it fulfills a series of conditions. The same sequence of conditions gives the same prediction. Therefore, the decision tree imitates human decision-making. The hyperparameter for the maximum depth of the tree to be selected comes from the list $\{1, 2, 3, 4, 5, 6, 7\}$. This project uses module `DecisionTreeRegressor` from Python library `scikit-learn` to estimate this model.

4.2 Stage 2: Prediction with Ensemble Methods

Motivated by section 3, the second stage aims to examine if ensemble methods can improve forecasting performance. Ensemble methods combine predictions from base models to generate a forecast. Through quantifying uncertainty of predictions from base models at Stage 1, ensemble methods may lead to better forecasting performance. This project considers three ensemble methods:

- Bagging
- Random Forest
- Stacking

[9]The convergence of a solution to the SVR model can be time-consuming (more than two hours) for some cases. In these cases, this project sets the maximum number of iterations (`max_iter`) as 100000.

Ensemble method On the one hand, bagging, random forest, and stacking in this project are parallel ensemble methods whose base model generation is parallel. Therefore, this project uses parallel computing to accelerate the training process. Section 9 discusses the use of parallel computing further.

Bagging Bagging, also known as bootstrap aggregation, improves prediction by addressing overfitting and reducing variance in forecasts. For each base model, instead of estimating the model with every sample in each rolling window, bagging

(1) creates B bootstrap samples[10]

(2) estimates the model with each bootstrap sample and generate a one-step-ahead prediction from each bootstrap sample

(3) generates the final prediction for the window by averaging predictions from each bootstrap sample.

This project sets the number of bootstrap samples to be $B = 100$ and applies bagging to models 2 to 5 outlined in section 4.1. This project uses the same set of hyperparameters for each model while bagging.

The bagged LR model is prone to unreasonable predictions if the domains in training and testing sets are different. Therefore, this project prevents this situation by applying the least absolute shrinkage and selection operator (LASSO) to regularize estimates of coefficients while bagging. The cross-validation process selects from a list of shrinkage parameters $\{0, 2, 4, 6\}$.[11]

Random Forest Random forest uses only decision trees. This method is the same as bagging except that only a subset of randomly selected predictors will be used at each split when estimating each tree in the second step. The restriction at each split decorrelates each tree to improve forecasting performance further. At each split, this project cross-validates the maximum number of predictors for each tree from a list $\{1, 3, 6, 11, 15, 20\}$. This project uses the module `RandomForestRegressor` from Python library `scikit-learn` to implement random forest. Following suggestions from [De Prado, 2018], this project also sets `min_weight_fraction_leaf` to 5 per cent.

Stacking Stacking generates each prediction from 2 layers of models. The first layer of models in this project consists of the five base models outlined in Section 4.1. Models in the first layer generate forecasts in the same way as Stage 1, and therefore, this project directly uses forecasts from Stage 1. The second layer contains one meta-model that learns how to use predictions from all base models to predict the one-step-ahead Bitcoin price. Candidate meta-models are models two to five outlined in section 4.1. This project also uses the same procedure as Stage 1 to search for the optimal hyperparameter in each window for each candidate meta-model.

4.3 Forecast Evaluation Criteria

This project uses the walk-forward approach to evaluate the out-of-sample forecasting performance of each model. Specifically, the evaluation for each forecast took place as if the prediction had run in the past. The following examines each of the three forecast evaluation criteria for this project.

[10] This project first uses the standard bootstrap that assumes independent and identically distributed data for simplicity. It is notable that Bitcoin prices are serially correlated, so this assumption is invalid. Therefore, this project attempts to rectify this violation in section 8.3.

[11] It is notable that the the LASSO linear regression is equivalent to the ordinary linear regression when the shrinkage parameter is 0.

Forecast Accuracy [12] This project uses the mean absolute percentage error (MAPE) to quantify forecast accuracy. The MAPE is the simple arithmetic average of the absolute value of the percentage deviation from the actual value. In other words, the MAPE averages the magnitudes of all prediction errors relative to the respective actual value. Mathematically, the MAPE for a model with T predictions is

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\text{Predicted Value}_t - \text{Actual Value}_t}{\text{Actual Value }_t}\right|$$

The higher the MAPE, the less accurate the forecasting model is. Conversely, the lower the MAPE, the more accurate the forecasting model is. Same as other forecast error measures, the MAPE captures the magnitude but not the direction of change.

Directional Accuracy Directional accuracy is the relative frequency (in %) of correctly predicting the direction of changes in Bitcoin prices. The higher the directional accuracy, the more accurate the prediction model is. Conversely, the lower the directional accuracy, the less accurate the model is. This measure captures only the direction of change but not the magnitude.

Profitability A profitability measure aims to capture both the magnitude and the direction of change at the same time. To measure each model's profitability, this project assumes abundant market liquidity (i.e. enough market marker orders). Every transaction each model triggered is a price-taker transaction. Therefore, Coinbase completes each transaction immediately in the same period.

This project implements a simple "buy low, sell high" strategy to measure each method's profitability. At period t, each method predicts Bitcoin price at period $(t+1)$, if the magnitude of the predicted Bitcoin price change is more than 1.0%, there will be a transaction. Otherwise, there will be no transactions. This 1.0% threshold is due to Coinbase's 0.5% fee for each transaction.[13].

- If the method predicts a higher Bitcoin price at period $t+1$ compared with period t, there will be a purchase of 1 Bitcoin in period t. In the period immediately followed (*i.e.* $t+1$), there will be a sale of a Bitcoin in the market to realize the profit (or loss.)

- If the method predicts a lower Bitcoin price at period $t+1$ compared with period t, there will be a sale of 1 Bitcoin in period t. In the period immediately followed (*i.e.* $t+1$), there will be a purchase of a Bitcoin to realize the profit (or loss.)

Each transaction at period $t+1$ yield a realized rate of return r_{t+1}. At the end of the backtest period, the profitability of each model is its simple arithmetic average of realized rates of return. The higher the average rate of return, the better the model is.

5 Predictors

As discussed in section 2.1, research shows that the predictive power of traditional demand- and supply-side factors and variables reflecting market expectations on Bitcoin prices are significant. Therefore, this project uses these two types of predictors to forecast Bitcoin prices. This section examines each predictor and briefly justifies the correlation between Bitcoin prices and each of these predictors.

[12] This project sometimes uses forecast error to refer to forecast accuracy when appropriate.

[13] This webpage outlines the transaction fee schedule of Coinbase: https://help.coinbase.com/en/pro/trading-and-funding/trading-rules-and-fees/fees

5.1 Traditional Demand-side and Supply-side Factors

This subsection examines (1) moving averages of Bitcoin prices and total Bitcoin transaction volume, (2) the percentage of the volume of buy transactions and (3) the change in daily close Nasdaq Bitcoin Reference Price as traditional demand- and supply-side Bitcoin price predictors for this project. The main difference between these three types of predictors is that the first two types allow for intra-day differences in forecasts. In contrast, the remaining one allows only for inter-day forecast differences because the value of the variable updates only every day.

Moving Averages of Bitcoin Prices and Total Bitcoin Transaction Volume In general, more than one Bitcoin transaction[14] takes place every second on Coinbase, 24 hours every day. Therefore, we use a simple moving average to summarize prices and total transaction volume during a period. This project uses moving averages of Bitcoin prices in the previous 5 minutes, 10 minutes, 30 minutes and 60 minutes as predictors.

- **Bitcoin Price**: Bitcoin price is the price in USD per unit of Bitcoin exchanged in a transaction. This price is the price when the exchange matched the buyer and the seller for a transaction. The random walk hypothesis establishes a correlation between a Bitcoin price and its lags.

- **Total Bitcoin Transaction Volume**: Bitcoin transaction volume is the amount of Bitcoin exchanged in a transaction. The Law of Demand and the Law of Supply establish the existence of a correlation between Bitcoin transaction volume and price.

Percentage of the Volume of Buy Transactions For each of the four periods to compute moving averages, the percentage of the volume of buy transactions is the percentage of transaction volume corresponding to a price taker's buy position. As discussed in subsection 2.1, market expectations significantly affect Bitcoin prices. Therefore, a higher percentage signifies the public's higher confidence in Bitcoin and hence a correlation with Bitcoin price.

Change in Daily Close Nasdaq Bitcoin Reference Price The Nasdaq Bitcoin Reference Price is a daily weighted average Bitcoin price in core exchanges, adjusted for price volatility. Therefore, the change in the daily close reference price reflects price fluctuations in the entire Bitcoin market. Unlike the above predictors, this predictor is an inter-day factor inducing daily Bitcoin price differences in forecasting models.

5.2 Social Media-related Variables

Following findings from recent research publications, as discussed in subsection 2.1, this project uses variables reflecting public interest in Bitcoin on social media to predict Bitcoin prices in Coinbase. The seven predictors below cover the changes in the public's attention to Bitcoin on Google, YouTube, Reddit, Twitter and Wikipedia. The data frequency of each of these variables is daily. Therefore, in each forecasting model, these variables only reflect inter-day forecast differences.

Google Trends Indices in Web Search, News and YouTube Google Trends provides indices to show the evolution of the world's interest in Bitcoin based on Google web searches, Google news searches and YouTube Searches every day. Higher public interest translates to higher demand for Bitcoin and hence a correlation between Bitcoin prices and each of these three indices.

[14] Section 6.1.1 provides a brief definition of a transaction.

The Numbers of New Post and New Comments on the Subreddit `r/Bitcoin` Reddit is one of the most popular online forums globally where people can discuss different topics in each relevant subreddit. As more posts and comments reflect higher public interest in Bitcoin, the numbers of new posts and new comments in the subreddit `r/Bitcoin` correlate with Bitcoin prices.

Number of Bitcoin Tweets Twitter is another worldwide online platform where people share their thoughts. Therefore, through the same mechanism justified above, the number of new tweets each day about Bitcoin correlates with Bitcoin prices.

Wikipedia Page View Wikipedia page of a topic is usually the first page people visit when they would like to gain a quick and brief understanding of a topic. Therefore, like the above predictors, the number page view each day correlates with Bitcoin prices.

In summary, among all predictors discussed, only moving averages of Bitcoin prices, moving averages of total Bitcoin transaction volume, and the percentage of the volume of buy transactions fluctuate within a day. Therefore, social media-related predictors can only drive inter-day differences among predictions. On the contrary, traditional demand- and supply-side factors can cause both intra-day and inter-day differences among all predictions. Table 6 summarizes all predictors in this project.

Table 6: Summary of All Predictors

(a) Traditional Demand- and Supply-side Factors

	Variables	Frequency
1.	Moving Averages of Bitcoin Prices	5, 10, 30 and 60 minutes
2.	Moving Averages of Total Bitcoin Transaction Volume	5, 10, 30 and 60 minutes
3.	Percentage of the Volume of Buy Transactions	5, 10, 30 and 60 minutes
4.	Change in Daily Close Nasdaq Bitcoin Reference Price	Daily

(b) Social Media-related Variables

	Variables	Frequency
1.	Google Trends for Web Search, News and YouTube	Daily
2.	Numbers of New Posts and New Comments on the Subreddit `r/Bitcoin`	Daily
3.	Number of Bitcoin Tweets	Daily
4.	Wikipedia Page View	Daily

6 Data

The data set for this project consists of three parts: (1) Bitcoin trade data on Coinbase, (2) the Nasdaq Bitcoin Reference Price and (3) social media data. This section examines each of these three components, how this project derives its data set from these three types of sources and explores data characteristics of each predictor.

6.1 Bitcoin Trade Data on Coinbase

Bitcoin trade data on Coinbase used in this project comes from Kaiko Digital Assets Data Provider. This data set contains every transaction on Coinbase that took place in August 2021. This project derives moving averages of Bitcoin prices and total Bitcoin transaction volume and the percentage of the volume of buy transactions within different time intervals from this data set. The general trend changed at least twice within the five-minute moving average Bitcoin price this project predicts.

6.1.1 Data Set Description

The data set purchased from Kaiko Digital Assets Data Provider contains the time, price, volume and trade direction of every executed transaction from the first millisecond on August 1st, 2021 to the last millisecond on August 31st, Universal Time Coordinated (UTC) on Coinbase, one of the most popular Bitcoin exchanges in the world. Table 7 summarizes the definitions of every column of this data set, followed by elaboration of some key terms in definitions of each column.

Table 7: Definition of Each Column of the Tick-by-tick Bitcoin Trade Data in Coinbase

Column	Definition
ID	Unique trade ID (unique to the exchange).
Date	Epoch timestamp in milliseconds.
Price	In One Bitcoin per US dollar.
Amount	Quantity of Bitcoin bought or sold.
Sell	True or False, referring to the trade direction. A trade marked as "true" means that a price taker placed a market sell order. Otherwise, a price taker placed a market buy order.

Source: https://www.kaiko.com/pages/cryptocurrency-data-types#trades.

Transaction When a market participant places an order in Coinbase, the exchange first finds a matching and unfulfilled order of the opposite position. If there is no such order, the market participant becomes a market maker, waiting for other market participants to match his order. On the contrary, if there is a matching order, the market participant becomes a price taker fulfilling an order placed earlier by a market maker. When the exchange matches a price taker and a market maker and executes the order, a transaction is complete.[15] The time for each transaction is the time when a price taker fulfills the order.

Trade Direction A trade direction of "buy" refers to a price taker placing an order to buy Bitcoin (column sell being False in the dataset). In contrast, a trade direction of "sell" refers to a price taker placing an order to sell Bitcoin (column sell being True in the dataset).

Epoch Timestamp in Milliseconds Epoch timestamp (also know as Unix timestamp) in this dataset is a 13-digit integer that represents the year, month, date, hour, minute, second and millisecond when a transaction is complete. A value of zero refers to January 1st, 1970 at 00:00:00.000.

6.1.2 Derived Variables: Summary and Characteristics

The moving averages of Bitcoin prices and total Bitcoin transaction volume and the percentage of the volume of buy transactions come from this data set.

- **Moving averages of Bitcoin prices and total Bitcoin transaction volume**: For each time interval, each data point of the moving averages is the simple arithmetic average of all prices from all transactions during the period.
- **Percentage of the volume of buy transactions**: It is the number of buy transactions in the interval divided by the total number of transactions in the interval.

[15] This description is a simplified version of how a transaction takes place in Coinbase. [Medalie, 2019] from the data provider wrote an article with a more detailed explanation.

Table 8 summarizes moving averages of Bitcoin prices, moving averages of total Bitcoin transaction volume and percentages of the volume of buy transactions on Coinbase every 5, 10, 30 and 60 minutes.

Table 8: Data Summary of Moving Averages of Bitcoin Prices, Moving Averages of Total Bitcoin Transaction Volume and Percentages of the Volume of Buy Transactions on Coinbase, Every 5, 10, 30 and 60 minutes, August 1$^{\text{st}}$ 00:05 (UTC) - August 31$^{\text{st}}$, 2021 23:59 (UTC)

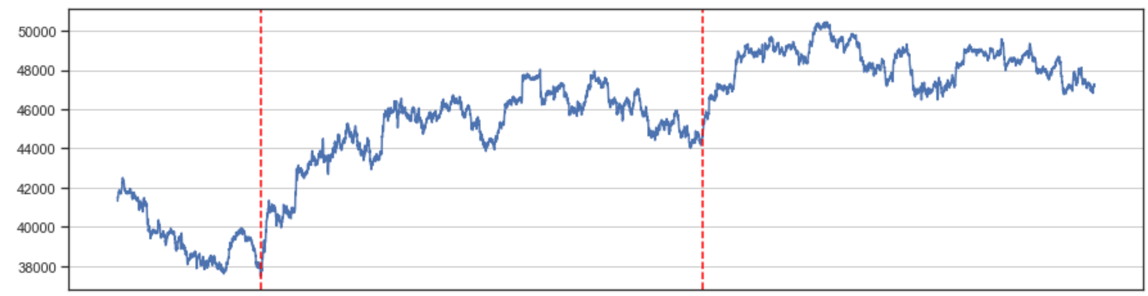
	Count	Mean	Std	Min	25%	50%	75%	Max
(a) Moving Average Bitcoin Prices								
5 minutes	8927	45,558.7	3,275.6	37,465.9	44,143.0	46,347.6	48,123.0	50,452.1
10 minutes	4464	45,558.9	3,275.6	37,539.2	44,138.3	46,346.4	48,126.0	50,450.7
30 minutes	1488	45,558.6	3,276.9	37,654.1	44,160.0	46,344.1	48,130.9	50,391.1
60 minutes	744	45,557.6	3,277.5	37,723.0	44,115.6	46,339.7	48,132.5	50,352.0
(b) Moving Average Total Bitcoin Transaction Volume								
5 minutes	8927	0.0850	0.0380	0.0125	0.0578	0.0776	0.1039	0.4127
10 minutes	4464	0.0865	0.0347	0.0230	0.0614	0.0803	0.1045	0.3315
30 minutes	1488	0.0882	0.0301	0.0307	0.0659	0.0832	0.1041	0.2298
60 minutes	744	0.0891	0.0278	0.0363	0.0694	0.0850	0.1040	0.2068
(c) Percentages of the Volume of Buy Transactions								
5 minutes	8927	0.51	0.15	0.06	0.39	0.51	0.62	0.95
10 minutes	4464	0.50	0.13	0.08	0.41	0.50	0.60	0.87
30 minutes	1488	0.50	0.10	0.18	0.43	0.50	0.57	0.80
60 minutes	744	0.50	0.09	0.24	0.44	0.50	0.56	0.78

Moving averages of Bitcoin prices The general price trend of the five-minute moving averages of Bitcoin prices changed at least twice (*structural breaks*) by visual inspection of the five-minute moving average Bitcoin prices (Figure 3). Precisely, this time series contains at least three episodes (with red vertical lines on Figure 3 denoting the two breakpoints):[16]

- A downward trend from the beginning to the 153$^{\text{rd}}$ five-minute interval on August 5$^{\text{th}}$
- An upward trend from the 153$^{\text{rd}}$ five-minute interval on August 5$^{\text{th}}$ to the 166$^{\text{th}}$ five-minute interval on August 19$^{\text{th}}$
- A milder increasing trend from the 166$^{\text{th}}$ five-minute interval on August 19$^{\text{th}}$ to the end

Moving averages of Bitcoin prices for other intervals show similar patterns.

Figure 3: Five-minute Moving Average Bitcoin Prices on Coinbase, August 1$^{\text{st}}$ 00:05 (UTC) - August 31$^{\text{st}}$, 2021 23:59 (UTC), US Dollar



Note: The two vertical dotted lines denote the two structural breakpoints at the 153$^{\text{rd}}$ five-minute interval on August 5$^{\text{th}}$ and the 166$^{\text{th}}$ five-minute interval on August 19$^{\text{th}}$.

[16]Using a Wald Test for a known structural break with time being the only feature in the regression, there is very strong evidence for a structural break in the 153$^{\text{rd}}$ five-minute interval on August 5$^{\text{th}}$ ($\chi^2(2) = 6441.6$, p-value=0) and very strong evidence for a structural break in the 166$^{\text{th}}$ five-minute interval on August 19$^{\text{th}}$ to the end ($\chi^2(2) = 4026.5$, p-value=0).

Table 8 and Figure 3 also show moving averages of Bitcoin prices and total Bitcoin transaction volumes and the percentages of the volume of buy transactions during the four intervals fluctuate. While the former two are left-skewed, the percentages have roughly symmetric distributions.

6.2 The Nasdaq Bitcoin Reference Price

This project uses the daily Nasdaq Bitcoin Reference Price (NQBTC) in August 2021 to develop an initial understanding of Bitcoin price prediction and to derive the daily change in close NQBTC as a predictor for five-minute moving average Bitcoin prices. The general trend of the NQBTC matches the five-minute moving average. Trend changes in the five-minute moving average Bitcoin price are similar to the NQBTC. However, daily changes in the close NQBTC did not have trend changes in August 2021.

6.2.1 Data Set Description

The NQBTC is a daily weighted average of Bitcoin price (in US dollar) in five core Bitcoin exchanges[17], adjusted for price volatility.[18] This data set of NQBTC contains close (at 20:05 UTC), open, high and low prices spanning from August 1st, 2021 to August 31st, 2021. Table 9 shows that the distributions of the four prices are similar. Moreover, these four distributions are similar to the distribution of the five-minute moving average Bitcoin prices to be predicted (Table 8).

Table 9: Data Summary of Daily Open, Close, High and Low Nasdaq Bitcoin Reference Prices, August 1st - August 31st, 2021

	Count	Mean	Std	Min	25%	50%	75%	Max
Open	31	45,423.55	3,327.69	37,972.10	43,893.15	46,393.30	47,755.15	50,263.4
Close	31	45,604.63	3,312.49	38,054.80	44,242.50	46,867.70	47,955.85	50,226.5
High	31	46,152.82	3,334.37	38,555.50	44,880.80	47,204.90	49025.05	50329.40
Low	31	45,218.96	3,297.72	37,904.40	43,964.60	46,533.10	47,746.25	49,099.00

Moreover, Figure 4 shows that the four NQBTC Bitcoin prices follow a similar trend. This trend is also comparable to the five-minute moving average Bitcoin prices. More importantly, the movement of these five prices experienced the same changes in trends on August 5th and August 19th, 2021.

6.2.2 Derived Variables: Summary and Characteristics

As discussed in section 5.1, this project uses daily changes in the close NQBTC as a predictor. Table 10 shows that the distribution of the daily changes is right-skewed. Figure 4 demonstrates that patterns of the daily changes in the close NQBTC remain unchanged in August 2021.

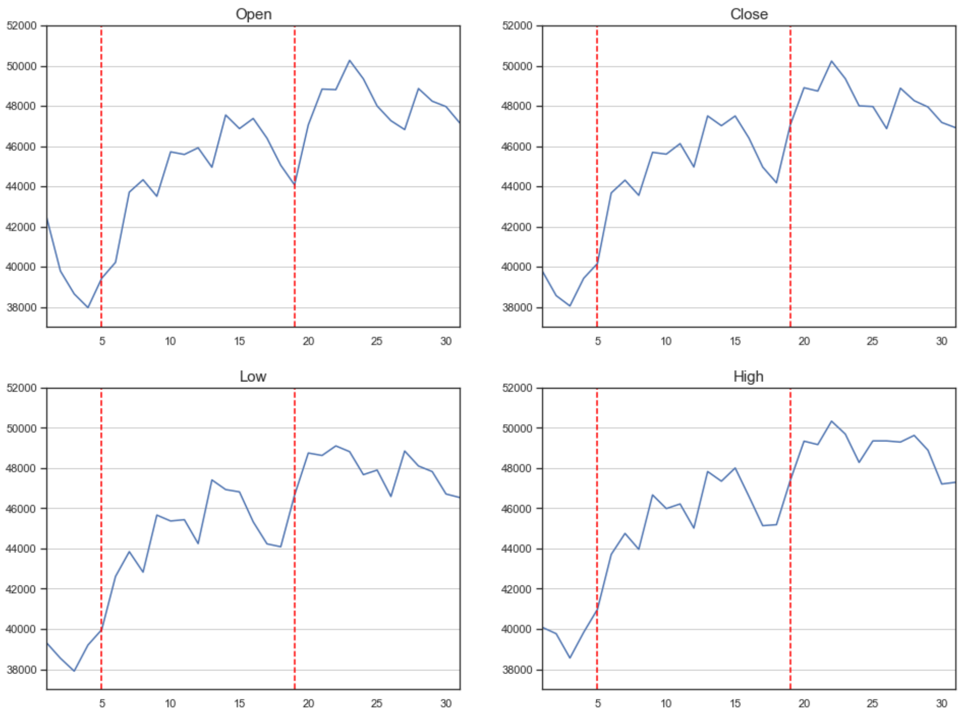
Table 10: Data Summary of Daily Changes in Close Nasdaq Bitcoin Reference Prices, August 1st - August 31st, 2021

	Count	Mean	Std	Min	25%	50%	75%	Max
Daily Changes in Close NQBTC	31	149.56	1,448.17	-2,499.3	-827.9	-258.8	1,055.9	3,506.3

[17]The five core exchanges are BitStamp, Coinbase, Kraken, Gemini and itBit.

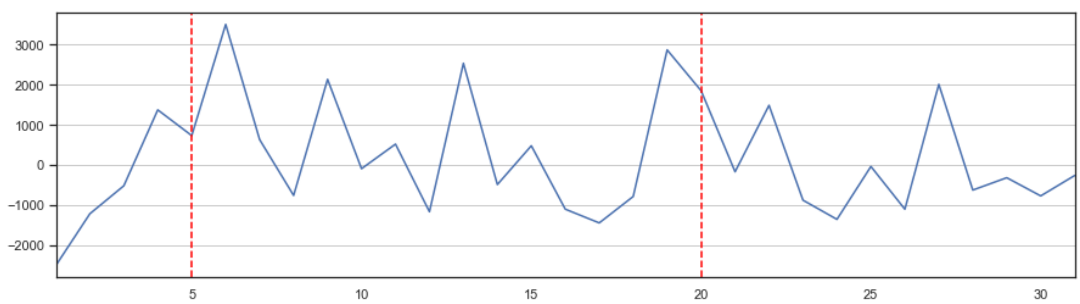
[18]This methodological document contains more details about the Nasdaq Bitcoin Reference Prices: https://hdx-prd-web-cms-upload-bucket.s3.amazonaws.com/Anexo_V_Hashdex_Bitcoin_ETF_Metodologia_Ingl_s_cd1e343dad.pdf.

Figure 4: The Nasdaq Bitcoin Open, Close, Low and High Prices, August 1st, 2021 - August 31st, 2021, US Dollar



Note: The two vertical dotted lines on each graph denote the two structural breakpoints identified in Figure 3. These two points are August 5th and August 19th.

Figure 5: Daily Changes in the Close Nasdaq Bitcoin Prices, August 1st, 2021 - August 31st, 2021, US Dollar



Note: The two vertical dotted lines on each graph denote the two structural breakpoints identified in Figure 3. These two points are August 5th and August 19th.

6.3 Social Media Data

This project collects data about the public's interest in Bitcoin on mainstream social media. These social media websites are Google, YouTube, Reddit, Twitter and Wikipedia. Each of these variables is available in daily frequency and spans from August 1^{st}, 2021 to August 31^{st}, 2021. Some of these variables had the same trend changes as the five-minute moving average Bitcoin price.

6.3.1 Data Set Description

This data set consists of seven variables reflecting the popularity and public interest in Bitcoin on five mainstream social media websites. Google Trends provides three of the seven variables that indicate the popularity of Bitcoin in each of its searches: web search, news search and YouTube search. Two of the remaining four variables reflect the intensity of the discussion about Bitcoin on a popular online forum Reddit. One other variable captures the degree of the public's attention to Bitcoin. The last variable shows people's interest in more in-depth information about Bitcoin from the online encyclopedia Wikipedia.

Google Trends Google Trends (`https://trends.google.com/trends/?geo=US`) provides daily indices for the number of web searches, YouTube searches and news searches for keyword "Bitcoin" from August 1^{st}, 2021 to August 31^{st}, 2021 in the world. Each index ranges from 0 to 100, with 0 indicating the lowest popularity and 100 indicating the highest popularity during the period.

Daily Number of New Posts and New Comments in Subreddit `r/Bitcoin` The website Subreddit Stats - statistics for every subreddit (`https://subredditstats.com/r/Bitcoin`) provides daily counts of new posts and new comments at 20:00 UTC in the subreddit `r/Bitcoin` from August 1^{st}, 2021 to August 31^{st}, 2021.

Daily New Bitcoin Tweets The website BitInfoCharts (`https://bitinfocharts.com/comparison/tweets-btc.html`) provides the daily number (in thousands) of new Bitcoin tweets on Twitter from August 1^{st}, 2021 to August 31^{st}, 2021.

Daily Wikipedia Page Views The website Pageviews Analysis (`https://pageviews.toolforge.org/?project=en.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2021-08-01&end=2021-08-31&pages=Bitcoin`) provides the daily number of page views of the Wikipedia page of Bitcoin. This data series spans from August 1^{st}, 2021 to August 31^{st}, 2021.

6.3.2 Variables: Summary and Characteristics

Table 11 summarizes the data set of all social media related variables and shows that most variables are approximately symmetric, except that the number of new posts on `r/Bitcoin` is roughly right-skewed. Figure 6 shows time series plots of all social media related variables. Although each of these variables follow different patterns, most of them, except the number of new posts on `r/Bitcoin` and the number of new tweets, had trend changes on the two points when the five-minute moving averages of Bitcoin price trend changed.

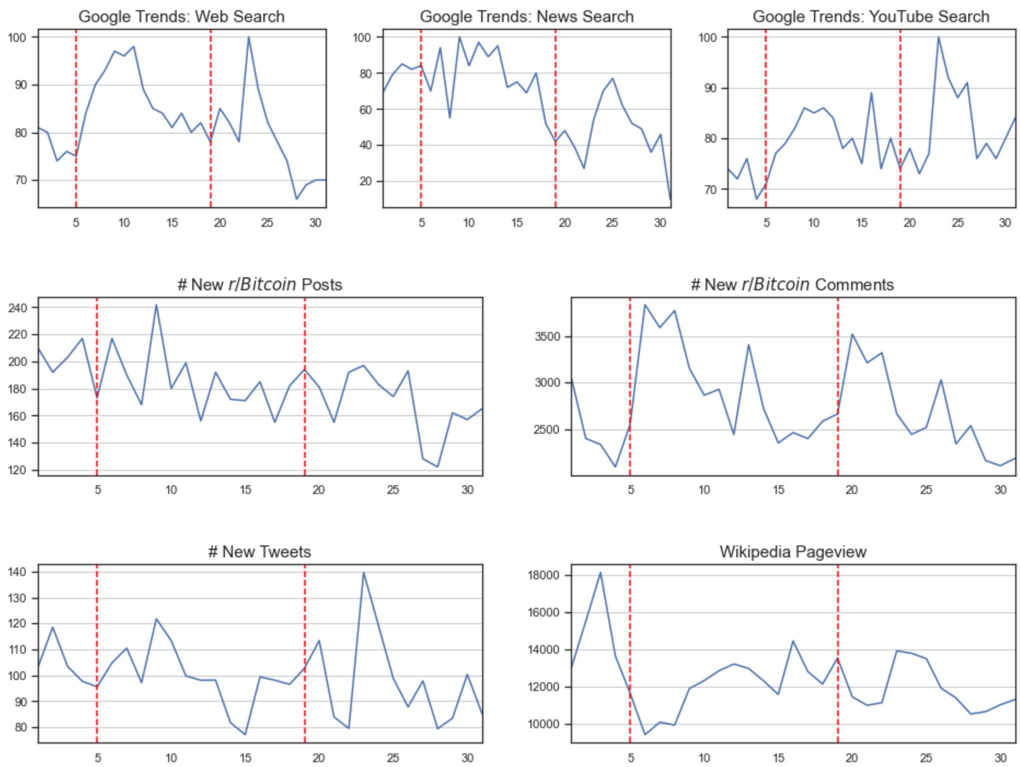
6.4 Correlation Analysis

Examining the evolution of the sample correlation between each predictor and the moving average Bitcoin Price to be predicted in each rolling window helps ensure that each predictor will be helpful for prediction. Although the correlations between each predictor and the Bitcoin price to be predicted

Table 11: Data Summary of Social Media Related Variables, August 1st - August 31st, 2021

	Count	Mean	Std	Min	25%	50%	75%	Max
(a) Google Trends								
Web Search Index	31	82.3	8.7	66.0	77.0	82.0	87.0	100.0
News Search Index	31	65.9	22.2	10.0	50.5	70.0	83.0	100.0
YouTube Search Index	31	80.1	7.0	68.0	75.5	79.0	84.5	100.0
(b) Reddit								
# New Posts	31	180.9	25.1	122.0	166.5	182.0	193.5	242.0
# New Comments	31	2762.7	504.4	2091.0	2397.0	2584.0	3110.0	3840.0
(c) Twitter								
# New Tweets	31	99.5	14.1	77.1	91.7	98.1	104.1	139.7
(d) Wikipedia								
# Pageview	31	12340.8	1765.6	9410.0	11210.5	12125.0	13340.5	18110.0

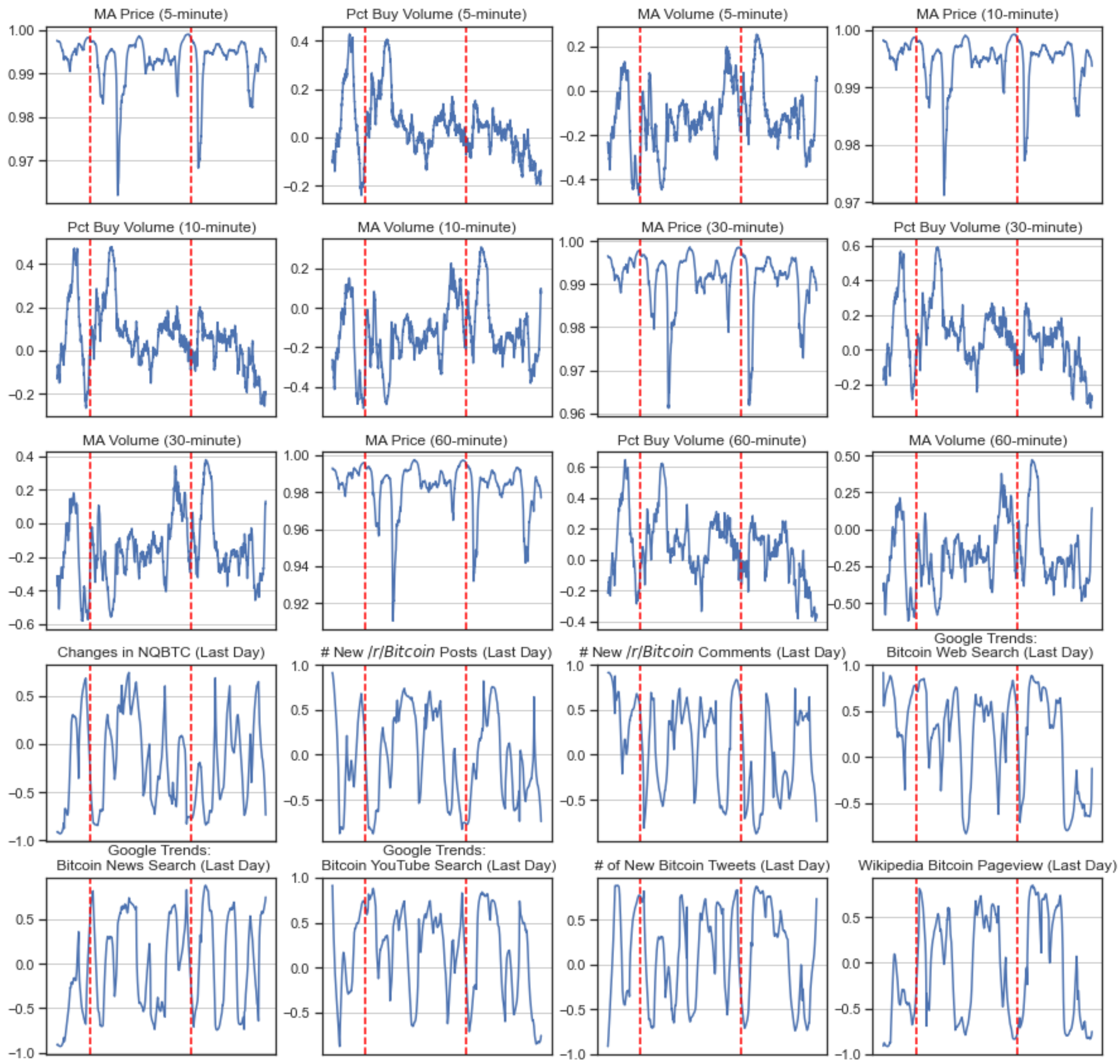
Figure 6: Social Media Related Variables, August 1st, 2021 to August 31st, 2021



Note: The two vertical dotted lines on each graph denote the two structural breakpoints identified in Figure 3. These two points are August 5th and August 19th.

fluctuate over time, at least one variable correlates with the Bitcoin price to be predicted every period (Figure 7). Moreover, among all predictors, it is not surprising that lagged moving averages of Bitcoin prices over different time intervals are most correlated with the five-minute moving average Bitcoin Price to be predicted. Therefore, predictors listed in section 5 should provide value for predicting the five-minute moving average Bitcoin price.

Figure 7: Rolling Pearson Correlation Coefficients between the Moving Average Bitcoin Price Every Five Minutes and Predictors



•MA Price: The moving average of Bitcoin prices; Pct Buy Volume: The percentage of the volume of buy transactions; MA Volume: The moving average of total Bitcoin transaction volume

•The number of minutes in brackets refers to a lag of the number of minutes.

• Note: The two vertical dotted lines on each graph denote the two structural breakpoints identified in Figure 3. These two points are August 5th and August 19th.

6.5 Data Pre-processing

The magnitude of the moving averages of Bitcoin prices and Wikipedia page views is much larger than other variables. Avoiding the dominance of predictors is particularly important for the SVR model and the k-NN model. Therefore, this project standardizes each predictor before prediction. Specifically, for each predictor x_t at time t, the standardized predictor $\tilde{x}_t$:

$$\tilde{x}_t = \frac{x_t - \bar{x}_t}{s_{x_t}}$$

where $\bar{x}_t := \frac{1}{W} \sum_{i=t-W+1}^{t} x_i$, W is the size of the rolling window and $s_{x_t} := \sqrt{\frac{1}{W-1} \sum_{i=t-W+1}^{t} (x_i - \bar{x}_t)^2}$. For consistency, all models use standardized predictors for estimation and prediction.

7 Forecast Evaluation Results

This section reports forecasting evaluation results from Stages 1 and 2 using all predictors mentioned in section 5. At Stage 1, the RW model performs the best forecast and directional accuracy but has the second-worst profitability. Indeed, none of the five base models outperforms the other with regard to all three forecast evaluation criteria. At Stage 2, while there were no noticeable changes in forecast accuracy with ensemble methods, ensemble methods deliver worse directional accuracy in some instances. However, profits from ensemble methods are higher in most cases.

7.1 Evaluation of Forecasts from Stage 1

The rankings of forecast accuracy and directional accuracy are similar, while profitability ranking is different. For instance, the DT model performs the second worst for forecast accuracy (MAPE: 0.77 per cent) and the worst for directional accuracy (49.02 per cent). However, the DT performs the best for profitability (0.011 per cent). Figure 8 ranks all five base models by each evaluation criterion, with the leftmost model the best and the rightmost model the worst.

Forecast error There is a considerable gap between the MAPEs of the top three models and the two worst models (the DT model and the SVR model). A possible justification for this gap is the linearly inseparable nature of the data. The DT model and the SVR model linearly partition the data set in the rolling window while the other four models do not partition the data set. However, this gap does not appear in directional accuracy and profitability.

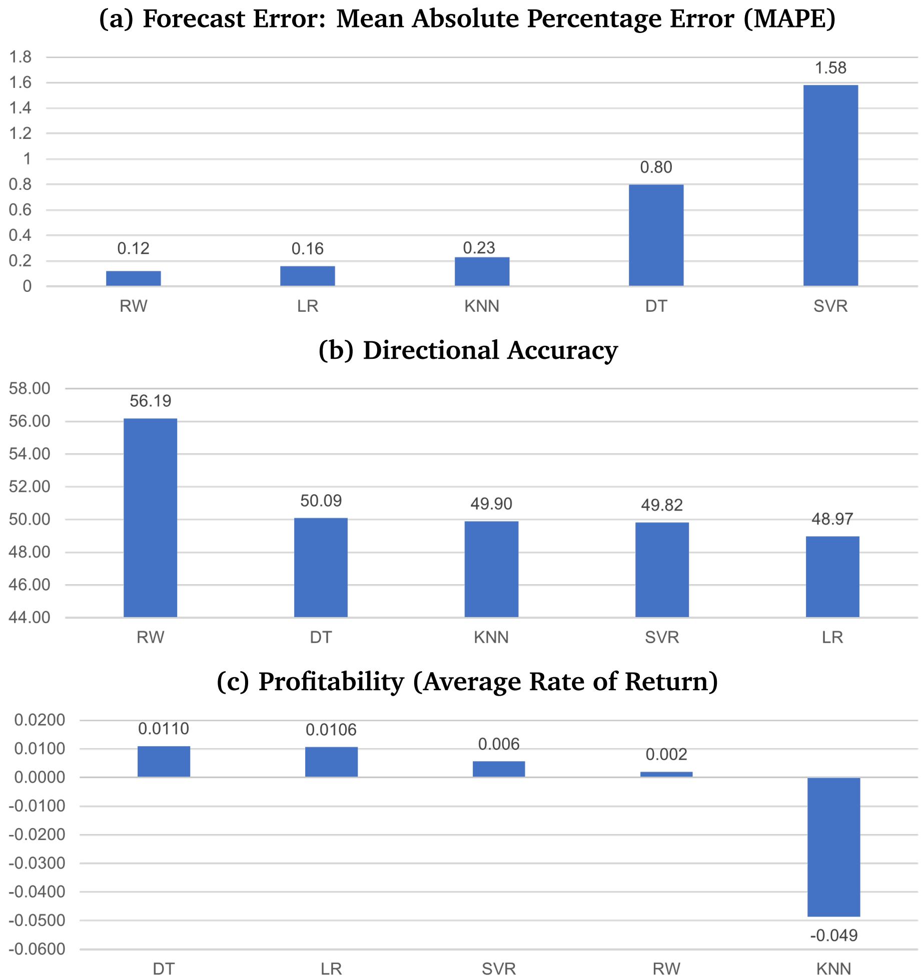
Directional accuracy All five models can only correctly predict about 50 to 56 per cent of the Bitcoin price movement direction. The performance for the RW model is slightly better than coin-flipping, while that for all four other models is roughly the same as coin-flipping. This result is consistent with the expectation discussed in section 3 and the results from a recent research article [Ibrahim et al., 2021][19].

Profitability The average rates of return (or loss) from all models are low. Indeed, including the one-per-cent transaction fee charged by Coinbase, on average, all five models suffer losses with the strategy of instant profit realization. These minimal profits (or losses) are due to the popularity of Bitcoins. With a large number of transactions within a very short period, any deviation from the market equilibrium disappears instantly and hence the minimal magnitude of profits (or losses).

Other remarks:

[19] [Ibrahim et al., 2021] compare the directional accuracy of Bitcoin price movement in the next five minutes among the autoregressive integrated moving average (ARIMA), the prophet (by Facebook), the random forest, the random forest lagged-Auto-Regression, and multi-layer perceptron (MLP) neural network. With feature engineering and various data transformation, they find that the MLP model performs the best and achieves 54 per cent directional accuracy.

Figure 8: Rankings of all Five Base Models at Stage 1, by (a) Forecast Error, (b) Directional Accuracy and (c) Profitability, August 5th - August 31st 2021, per cent



- RW: random walk model, LR: linear regression, KNN: k-nearest neighbour regression, DT: decision tree regression, SVR: support vector regression
- On each of the three panels, the leftmost model performs the best and the rightmost model performs the worst in terms of the panel's criterion.

- The outperformance of the RW model over other models in terms of forecast accuracy and directional accuracy is the same as expected in section 3. Indeed, this phenomenon is common in financial time-series, as mentioned in section 4.

- None of these base models outperforms the other in all three forecast evaluation criteria. For instance, while forecasts from the RW model are the most accurate with regard to forecast and directional accuracy, it is the second-worst model in terms of profitability between two consecutive five-minute periods.

7.2 Evaluation of Forecasts from Stage 2

As mentioned in section 4.2, the purpose of Stage 2 is to examine if ensemble methods can improve forecasting performance. In general, ensemble methods do not significantly improve the forecast accuracy and have mixed effects on directional accuracy. However, ensemble methods improve profitability in most cases. Table 12 summarizes forecast evaluation results from Stages 1 and 2.

Table 12: Forecast Error, Direction Accuracy and Profitability of Predictions, Stage 1 and Stage 2, All 20 Predictors, August 5th – August 31st 2021, per cent

	Forecast Error (MAPE)	**Directional Accuracy**	**Profitability**
Stage 1:			
The Random Walk (RW)	0.1	55.97	0.0024
The k-nearest Neighbour (k-NN) Regression	0.2	49.80	-0.0468
The Decision Tree (DT) Regression	0.8	49.93	0.0108
The Support Vector Regression (SVR)	1.6	49.68	0.0051
The Linear Regression (LR)	0.2	49.02	0.0106
Stage 2:			
(a) *Stacking*:			
The k-nearest Neighbour (k-NN) Regression	0.2	46.71	0.1500
The Decision Tree (DT) Regression	0.8	49.68	0.0113
The Support Vector Regression (SVR)	1.5	49.67	-0.0026
The Linear Regression (LR)	0.2	43.84	0.0289
(b) *Bagging*:			
The k-nearest Neighbour (k-NN) Regression	0.2	50.5	-0.022
The Decision Tree (DT) Regression	0.8	49.8	0.013
The Support Vector Regression (SVR)	1.6	49.95	0.0055
The Linear Regression (LR) with LASSO	0.2	46.64	-0.1428
(c) Random Forest	0.3	49.81	0.0376

Stacking Using each of the five base models from Stage 1 as a meta-model to combine each model's predictions at Stage 1 does not have notable effects on each model's forecast accuracy. The directional accuracy of most models remains roughly the same, except that the direction accuracy of the LR model drops from 49.0 per cent o 43.8 per cent. One possible cause of this reduction is the presence of outlying base-model predictions in rolling windows. Indeed, the forecast error of the SVR model is considerably higher than four other models (Figure 8).

On the contrary, using the k-NN, the DT and the LR models as meta-models to combine predictions increases profits. For instance, the average rates of return from the k-NN model turned from losses to profits (-0.06 per cent vs. 0.15 per cent). However, the average rates of return of the SVR turned from positive to negative (0.005 per cent to -0.003 per cent).

Bagging Bagging improves the forecasting performance of each base model at Stage 1 by lowering the variance of predictions through averaging predictions from different bootstrap samples. This project finds that bagging yields mixed results in the three forecast evaluation criteria model-wise. For example, bagging increases the forecast error of k-NN model slightly (0.23 per cent vs 0.25 per cent) but increases the directional accuracy (49.8 per cent vs 50.5 per cent). It also increases profits by 0.03 percentage points (-0.0486 per cent vs -0.0215 per cent), although still suffering losses. Moreover, for the DT model, it is notable that bagging improves forecast accuracy (0.02 percentage points) and profitability (0.002 percentage points) mildly, despite a slight reduction in the directional accuracy (50.1 per cent vs 49.7 per cent).

Random forest Random forest quantifies uncertainty in predictions from the DT model by bootstrap and a random choice of 6 predictors among the 20 predictors. Using the standard bootstrap reduces the forecast error from 0.8 per cent at Stage 1 to 0.3 per cent with a slight decrease in the directional accuracy (49.93 per cent vs 49.81 per cent) and a three-fold increase in profitability (0.01 per cent vs 0.04 per cent). The success of the RF model compared with bagging decision trees is due to the random choice of predictors. Specifically, section 6 identifies a strong correlation between the five-minute moving average Bitcoin price to be predicted and its first lag (Figure 7). The random choice of predictors decorrelates tress and leads to lower variance than bagging decision trees.

Other remarks:

- The outperformance of the RW model over other base models and other models at Stage 2 in forecast accuracy and directional accuracy is the same as expected in section 3. Indeed, this phenomenon is common in financial time-series, as mentioned in section 4.
- No models at Stages 1 and 2 outperform the other in all three forecast evaluation criteria. This outcome is reasonable because tools this project consider are coarse.
- In most cases, ensemble methods can help improve forecast profitability. Unfortunately, on average, the improvement is not significant enough to help cover the one-per cent transaction cost charged by Coinbase.

8 Further Discussion

Based on findings from evaluating forecast results in section 7, this section explores three further issues related to this project:

1. trend changes in the five-minute moving average Bitcoin price,
2. using subsets of predictors and
3. replacing standard bootstrap

Due to limited computational resources, this project examines the second issue with bagging. Moreover, this project explores the third issue only with bagging the k-NN model and the DT model and the first lag of the five-minute moving average Bitcoin price as the only predictor. The discussion shows how the forecasting performance changes when considering each of the above three issues.

8.1 Trend Changes

Section 6 identifies that the trend of the five-minute moving average Bitcoin price to be predicted has a breakpoint on August 5$^{\text{th}}$ and another on August 19$^{\text{th}}$, 2021. If a trend changes, forecasts from models whose training data include a breakpoint should be worse than those that do not. This lower performance is due to the modelling of two stochastic processes as a single process instead of two. Therefore, this project compares forecasting performance between two sets of periods: (1) the entire timeframe of this project (August 5$^{\text{th}}$ – August 31$^{\text{st}}$, 2021) and (2) two sub-periods from August 9$^{\text{th}}$ to August 18$^{\text{th}}$ and August 22$^{\text{nd}}$ to August 31$^{\text{st}}$, 2021.[20] Table 13 confirms this intuitive argument by showing that column (2) has better forecast, directional accuracy and profitability than column (1) in most cases.

[20] Every rolling window for predicting the Bitcoin price in these two intervals should not contain the two breakpoints. The reason for choosing August 9$^{\text{th}}$ and 18$^{\text{th}}$ is that starting on a new day is simpler.

Table 13: Forecasting Performance Comparison between the Periods (1) August 5th – August 31st 2021 and (2) August 9th – August 18th 2021 and August 22nd – August 31st 2021, All 20 Predictors, per cent

	August 5th – August 31st, 2021 (1)			August 9th – August 18th, 2021 August 22nd – August 31st, 2021 (2)			Differences (2) - (1)		
	Forecast Error	Directional Error	Profitability	Forecast Error	Directional Error	Profitability	Forecast Error	Directional Error	Profitability
Stage 1									
KNN	0.2	49.9	-0.0486	0.2	50.0	0.1982	-0.0	0.1	0.2468
DT	0.8	50.1	0.0110	0.6	50.3	0.0110	-0.2	0.2	0.0000
SVR	1.6	49.8	0.0056	1.2	50.1	0.0016	-0.4	0.3	-0.0040
LR	0.2	49.0	0.0106	0.1	48.7	-0.0944	-0.0	-0.3	-0.1050
Stage 2									
(a) *Stacking*									
KNN	0.2	46.7	0.1500	0.2	46.9	0.1982	-0.0	0.1	0.0482
DT	0.8	49.7	0.0113	0.6	49.8	0.0110	-0.2	0.1	-0.0003
SVR	1.5	49.7	-0.0026	1.2	49.8	0.0016	-0.2	0.2	0.0042
LR	0.2	43.8	0.0289	0.2	43.8	-0.0944	-0.0	-0.1	-0.1233
(b) *Bagging*									
k-NN	0.2	50.5	-0.0215	0.2	51.0	0.1071	0.0	0.5	0.1286
DT	0.8	49.7	0.0128	0.5	49.8	0.0083	-0.3	0.1	-0.0045
SVR	1.6	50.0	0.0055	1.2	50.3	0.0015	-0.4	0.3	0.0040
LR with LASSO	0.2	46.6	-0.1428	0.2	46.6	-0.3231	0.0	0.0	-0.1803
(c) Random Forest	0.3	49.8	0.0376	0.2	49.7	0.0376	0.0	-0.1	0.0000

8.2 Using Subsets of Predictors

Not all 20 predictors listed in section 5 are equally useful to predict the Bitcoin price. Indeed, section 6.4 shows that the correlation between the five-minute moving average Bitcoin price and each of the 20 predictors is different and fluctuates. Therefore, this project performs the forecasting exercise outlined in section 4 again with each of the following three groups of predictors:

- **Group 1** (Bitcoin prices only): the last 5-minute, 10-minute, 30-minute and 60-minute moving average Bitcoin prices and the daily change in the close Nasdaq Bitcoin Reference price (NQBTC)
- **Group 2**: the last 5-minute moving average Bitcoin prices
- **Group 3** (volume-related variables only): moving average total Bitcoin transaction volumes and percentages of the volume of buy transaction for the last 5-minute, 10-minute, 30-minute and 60-minute intervals

Due to limited computational resources, this exercise applies only base models, stacking and random forest. Table 14 summarizes the forecasting performance.

Forecast Accuracy Using predictors from group 1 or group 2 instead of all predictors can yield lower forecast errors. On the contrary, this exercise obtains higher forecast errors for all models if using group 3 alone. This difference reflects that volume-related predictors are not crucial to the forecast compared with moving averages of past Bitcoin prices. Moreover, columns (1) and (2) of Table 14 panel (a) show that including more past moving average Bitcoin prices for different time intervals can lead to slightly greater improvement in forecast accuracy.

Directional Accuracy Comparing columns (1) and (2) of Table 14 panel (b) shows that including moving average Bitcoin prices from more previous intervals results in higher directional accuracy in most cases. In addition, using predictors from group 3 alone instead of all predictors delivers higher directional accuracy most of the time. Therefore, the role of volume in forecasting the change of direction of five-minute moving average Bitcoin prices correctly is more important than the prices. This result is consistent with the Law of Demand from economics.

Table 14: Forecast Error, Directional Accuracy and Profitability of Predictions, Stage 1 and Stage 2 (Except Bagging), Three Groups of Predictors, August 5th – August 31st, 2021, per cent

	All Predictors	Group 1	Group 2	Group 3	(1)	(2)	(3)
(a) Forecast Error (MAPE)							
k-NN	0.227	0.160	0.168	1.136	-0.067	-0.059	0.909
DT	0.797	0.796	0.800	1.196	-0.001	0.003	0.399
SVR	1.578	1.197	1.533	1.964	-0.381	-0.045	0.386
LR	0.154	0.156	0.182	1.832	0.002	0.028	1.678
Stacking:							
k-NN	0.178	0.175	0.183	0.207	-0.003	0.005	0.029
DT	0.793	0.792	0.793	0.793	-0.001	0.000	0.000
SVR	1.461	1.468	1.417	1.517	0.007	-0.044	0.056
LR	0.180	0.182	0.181	0.181	0.002	0.001	0.001
Random Forest	0.251	0.253	0.319	1.407	0.002	0.068	1.156
(b) Directional Accuracy							
k-NN	49.90	48.14	45.56	50.17	-1.76	-4.34	0.27
DT	50.09	50.01	49.90	49.74	-0.08	-0.19	-0.35
SVR	49.82	50.23	50.05	49.90	0.41	0.23	0.08
LR	48.97	49.58	43.60	50.05	0.61	-5.37	1.08
Stacking:							
k-NN	46.71	46.91	47.36	48.83	0.20	0.65	2.12
DT	49.68	49.40	49.50	49.90	-0.28	-0.18	0.22
SVR	49.67	49.82	49.49	49.63	0.15	-0.18	-0.04
LR	43.84	43.61	43.63	43.90	-0.23	-0.21	0.06
Random Forest	49.81	50.58	49.34	50.40	0.77	-0.47	0.59
(c) Profitability							
k-NN	-0.05	0.44	0.15	-0.10	0.49	0.20	-0.05
DT	0.01	0.01	0.02	-0.07	0.00	0.01	-0.08
SVR	0.01	0.01	-0.02	0.00	0.00	-0.03	-0.01
LR	0.01	0.02	0.06	0.00	0.01	0.05	-0.01
Stacking:							
k-NN	0.15	0.10	0.19	-0.17	-0.05	0.04	-0.32
DT	0.01	0.01	0.01	0.01	0.00	0.00	0.00
SVR	-0.00	-0.01	0.00	0.00	-0.01	0.00	0.00
LR	0.03	-0.14	0.17	0.04	-0.17	0.14	0.01
Random Forest	0.04	0.04	0.01	-0.00	0.00	-0.03	-0.04

• Group 1 consists of the last 5-minute, 10-minute, 30-minute and 60-minute moving average Bitcoin prices and the daily change in the close Nasdaq Bitcoin Reference price (NQBTC).

• Group 2 consists of the last 5-minute moving average Bitcoin prices.

• Group 3 consists of moving average total Bitcoin transaction volumes and percentages of the volume of buy transaction for the last 5-minute, 10-minute, 30-minute and 60-minute intervals.

• (1): Group 1 - All Predictors; (2): Group 2 - All Predictors; (3): Group 3 - All Predictors

Profitability The higher values in column (2) of Table 14 panel (c) than columns (1) and (3) in most cases reflects that using the last 5-minute moving average Bitcoin prices alone instead of other predictors results in approximately the same or higher average profits in most cases. Considering moving average prices from more distant past yields lower average profits. Furthermore, using group 3 predictors alone leads to lower average profits in most cases. It is also notable that using only predictors from groups 1, 2 and 3 does not lead to average profits that can cover the one-per-cent transaction cost on Coinbase.

Overall, if a Bitcoin price forecasting model focuses on only one of the three criteria, the model does not need to use all 20 predictors:

- If prediction focuses on forecast accuracy and profitability, the most relevant predictors are moving averages of Bitcoin prices for different time intervals.
- Predictors pertinent to Bitcoin transaction volumes within different time intervals are the most relevant if prediction focuses on directional accuracy.

Moreover, in the face of computational constraints, using only the last available Bitcoin price to be predicted instead of other predictors can achieve good forecast accuracy and profitability at the expense of directional accuracy. This observation is consistent with the forecasting performance of the random walk model in section 7.

8.3 Replacing the Standard Bootstrap with a Modified Version

Predictions from bagging and random forest reported in section 7 come from the standard bootstrap. The standard bootstrap assumes that every data point is independently and identically distributed. However, the five-minute moving average Bitcoin prices predicted and all predictors are time series. In other words, each observation in each of these time series is serially correlated. Therefore, the assumption of standard bootstrap does not fit. Algorithm 2 attempts to correct this violation by combining two methods to generate bootstrap samples:[21]

- the seasonal-trend decomposition using locally estimated scatterplot smoothing (STL decomposition)[22] (function STL_Decomposition $(D,\, x,\, p)$ in Algorithm 2).
- the moving block bootstrap (function Moving_Bootstrap_Indices $(D,\, p)$ in Algorithm 2) and

This variant of bootstrap aims to resolve the serial dependence issue by bootstrapping residuals from the STL decomposition instead of the entire serially correlated time series. Using the moving block bootstrap instead of the standard bootstrap is to capture the temporal structure and enough variation of the time series. This project uses Python libraries `arch` for moving block bootstrap and `statsmodel` for the STL decomposition.

[21] This algorithm is a simplified version of a bootstrap method from [Bergmeir et al., 2016].

[22] The STL decomposition uses locally estimated scatterplot smoothing to additively decompose a time series into three components: (1) the trend component, (2) the seasonal component and (3) the residual component.

Algorithm 2 A Modified Bootstrap Algorithm for Time Series

Input: D ▷ D: the data set to be bootstrapped
Input: $p = 12$ ▷ p: the periodicity of the sequence for STL decomposition
Input: $B = 100$ ▷ B: the bootstrap sample size
Input: $M = 50$ ▷ M: Size of a moving block
Input: function Moving_Bootstrap_Indices (D, p) ▷ Note 1 below
Input: function STL_Decomposition (D, x, p) ▷ Note 2 below
Output: D_B-`lst` = [] ▷ D_B-`lst` = []: a list of bootstrapped data sets

```
function MODIFIED_BOOTSTRAP(D, p, B, M)
    Initialize list D_{x_B}-lst = [] to save each bootstrapped time series
    bootstrap-index-lst ← Moving_Bootstrap_Indices (D, p)
    for each set of indices I in bootstrap-index-lst do
        for each variable x in the data set D do
            {T, S, R} ← STL_Decomposition (D, x, p)
            R_B ← A time series created from each observation of R in the order of indices I
            S_{x_B} ← T + S + R_B    ▷ S_{x_B} is the bootstrapped time series of column x of data set D
            Append list D_{x_B}-lst = [] by S_{x_B}
        end for
        Append list D_B-lst = [] by D_{x_B}-lst = []
    end for
    return D_B-lst
end function
```

• Note 1: Moving_Bootstrap_Indices (D, p): a function that returns a list whose members are indices of observations in D corresponding to each bootstrap sample
• Note 2: STL_Decomposition (D, x, p): a function that returns a collection of the trend (T), the seasonal (S) and the residual (R) component time series of column x of D decomposed by the STL decomposition with periodicity p. The sum of T, S and R equals the time series of column x of D.

Table 15 compares forecasting performance between using the standard bootstrap and Algorithm 2 for bagged k-NN models and bagged DT models during the two sub-periods unaffected by trend changes mentioned in section 8.1.[23] Due to limited computational resources and results from section 8.1, this comparison uses only the first lag of the five-minute moving average Bitcoin price as the predictor.

Taking the serial dependence into account, this exercise finds that

- there is significant improvement in profitability with worse performance in forecast accuracy and directional accuracy.
- This bootstrap method can further improve forecasting performance by cross-validating parameters p and M of Algorithm 2.

[23]This exercise uses the two sub-periods without the breakpoints identified in section 6 instead of the entire time frame because the moving block bootstrap component of Algorithm 2 preserves some temporal structures. Therefore, the bootstrap may create several trend breaks in the bootstrap sample of a rolling window from resampling. This bootstrap sample does not reflect the distribution of the original time series in the rolling window.

Table 15: Forecast Error, Directional Accuracy and Profitability of Predictions, Bagging with Standard Vs. Modified Bootstrap, the First Lag of Five-minute Moving Average Bitcoin Prices as the Only Predictor, August 9th – August 18th and August 22nd – August 31st, 2021, per cent

	Standard Bootstrap (1)	Modified Bootstrap (2)	Difference (2) - (1)
(a) Forecast Error			
The k-nearest Neighbour Regression (k-NN)	0.18	2.58	2.40
The Decision Tree Regression (DT)	1.01	1.79	0.78
(b) Directional Accuracy			
The k-nearest Neighbour Regression (k-NN)	45.31	51.74	6.43
The Decision Tree Regression (DT)	50.90	50.52	-0.38
(c) Profitability			
The k-nearest Neighbour Regression (k-NN)	-0.0696	0.0034	0.037
The Decision Tree Regression (DT)	-0.0100	0.0028	0.0128

9 Computational Challenges and Solutions

The generation of each forecast is computationally intensive. Indeed, it is very time-consuming to cross-validate all rolling windows for each model, given the number of observations for this project. Using a standard laptop to generate all predictions is not feasible. Therefore, this project uses parallel programming to accelerate prediction generation from each model and Google Cloud service for better hardware.

Parallel Programming The prediction generation process in each rolling window is independent of the other. Specifically, the prediction generation of one rolling window does not affect the prediction generation process of another rolling window. Therefore, each central processing unit can fit a model and predict from the fitted model in each rolling window simultaneously in any order. This lack of dependency makes the forecast generation process *perfectly parallel*. Given this perfect parallelism, this project uses the Python library `joblib` to accelerate model estimation and prediction. Specifically, this Python library helps allocate tasks of each rolling window to a central processing unit (CPU). With multiple CPUs running together, the total runtime is much shorter.[24]

Google Cloud Service Google Cloud Service provides hardware that enables the exploitation of the perfectly parallel nature of this project. Specifically, most computer programs for model estimation and prediction for this project run on Google's virtual machines with 16 N1 standard CPUs and 60GB of system memory. As a result, programs that take days to run on a laptop now take only a few hours. Except for the SVR model, it took roughly an hour to estimate every base model at Stage 1 and generate predictions. The SVR model required about three hours due to the time needed for convergence. This hardware specification took very long to run programs for each ensemble method. For example, the random forest took more than 12 hours.

Forecasting problems after Stage 1 of this project are more computationally complicated than Stage 1. Therefore, towards the end of this project, the hardware specifications for Stage 1 is not sufficient. This project uses two virtual environments for Stage 2. The first environment has 32 CPUs and a system

[24]This project does not consider multi-threading to accelerate runtime because functions provided by libraries to estimate each model use multi-threading also. Using multi-threading in addition to the multi-threading operations by program libraries without complete understanding may delay runtime. Therefore, for simplicity, this project resorts only to parallel programming.

memory size of 120 GB, while the other one has 64 CPUs and a system memory size of 240 GB. Despite more CPUs and larger system memory size, the latter virtual machine still took about five hours to create all bootstrap samples for each rolling window with Algorithm 2.

10 Limitation and Future Development

This section briefly discusses three limitations encountered in this project and how this project can develop with regard to these three issues. The first problem is the missed opportunity to collect free and valuable data on Bitcoin Tweets. The second one is about using sequential bootstrap instead of the standard bootstrap. The last problem addresses the computational and financial limitations of this project.

Bitcoin Tweets Although section 2 identifies the importance of social media-related variables to Bitcoin price forecasting, this project cannot fully use these variables because they are not available with the same frequency as the intra-day Bitcoin price to be predicted. One data source of social media-related variables is Twitter. Through Twitter's application programming interface (API), one can retrieve all tweets with the keyword "Bitcoin" in real-time. Data companies sell these data sets at a high price, so this project cannot use this variable. Including the number of tweets with the word "Bitcoin" could have tremendous value to this project.

Sequential bootstrap for Bagging and Random Forest [De Prado, 2018] suggest using sequential bootstrap instead of the standard bootstrap to improve forecasting performance because of the independently and identically distributed assumption of the standard bootstrap. Unfortunately, to my knowledge, there are no publicly available program libraries in `R` and Python that can efficiently implement sequential bootstrap.[25] Developing computer programs for efficient sequential bootstrap is not easy[26] and is outside the scope of this project. In the future, a part of this project can include the development of programs for sequential bootstrap.

Limited Computational Resources Running all programs for model estimation and forecast generation on Google Cloud Service not only takes a long time, but also costly. The computational cost of this project is approximately CAD$600. Fortunately, Google Cloud Service provides a CAD$383.73 credit for first-time users. With this financial and computational constraint, this project cannot run all models in section 8. In the future, if there is enough financial resource, this project can complete the analysis in section 8.

11 Conclusion

In conclusion, this project studies the use of five basic machine learning forecasting models to predict intra-day moving average Bitcoin prices every five minutes on Coinbase from August 5th to August 31st, 2021. The forecast evaluation criteria are:

- Forecast accuracy (mean absolute percentage error, MAPE),
- Directional accuracy and
- Profitability (average rates of return).

[25] Hudson and Thames Quantitative Research in London, the U.K. used to provide a publicly available Python library `mlfinlab` which contains an efficient implementation of the sequential bootstrapping discussed in [De Prado, 2018]. However, since February 2019, they changed the business model from optional Patreon support to a monthly subscription. The monthly subscription fee of this library is £100 per month.

[26] Developing an efficient program for constructing the initial design matrix requires complicated multi-threading.

Using a set of 20 predictors reflecting the Bitcoin market and public interest in Bitcoin, this project finds that no single model and ensemble method can produce forecasts superior to others for all three criteria. A further investigation identifies that taking trend changes in Bitcoin prices and using bootstrap that accounts for serial dependence can improve forecasting performance. This report ends with a discussion of the computational challenges this project faces, solutions, and other limitations.

12 Bibliography

[Aalborg et al., 2019] Aalborg, H. A., Molnár, P., and de Vries, J. E. (2019). What can explain the price, volatility and trading volume of bitcoin? *Finance Research Letters*, 29:255–265.

[Athey et al., 2016] Athey, S., Parashkevov, I., Sarukkai, V., and Xia, J. (2016). Bitcoin pricing, adoption, and usage: Theory and evidence.

[Aysan et al., 2021] Aysan, A. F., Demirtaş, H. B., and Saraç, M. (2021). The ascent of bitcoin: Bibliometric analysis of bitcoin research. *Journal of Risk and Financial Management*, 14(9):427.

[Baek and Elbeck, 2015] Baek, C. and Elbeck, M. (2015). Bitcoins as an investment or speculative vehicle? a first look. *Applied Economics Letters*, 22(1):30–34.

[Bergmeir et al., 2016] Bergmeir, C., Hyndman, R. J., and Benítez, J. M. (2016). Bagging exponential smoothing methods using stl decomposition and box–cox transformation. *International journal of forecasting*, 32(2):303–312.

[Chowdhury et al., 2020] Chowdhury, R., Rahman, M. A., Rahman, M. S., and Mahdy, M. (2020). An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. *Physica A: Statistical Mechanics and its Applications*, 551:124569.

[Ciaian et al., 2016] Ciaian, P., Rajcaniova, M., and Kancs, d. (2016). The economics of bitcoin price formation. *Applied Economics*, 48(19):1799–1815.

[Dastgir et al., 2019] Dastgir, S., Demir, E., Downing, G., Gozgor, G., and Lau, C. K. M. (2019). The causal relationship between bitcoin attention and bitcoin returns: Evidence from the copula-based granger causality test. *Finance Research Letters*, 28:160–164.

[Dave, 2021] Dave, S. (2021). Day trading in crypto keeps young indians busy at night.

[De Prado, 2018] De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

[Ibrahim et al., 2021] Ibrahim, A., Kashef, R., and Corrigan, L. (2021). Predicting market movement direction for bitcoin: A comparison of time series modeling methods. *Computers & Electrical Engineering*, 89:106905.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*, volume 112. Springer.

[Jang and Lee, 2017] Jang, H. and Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *Ieee Access*, 6:5427–5437.

[Kaiko - Digital Assets Data Provider, 2021] Kaiko - Digital Assets Data Provider (2021). Tick-by-tick Bitcoin transaction data on Coinbase, August 1, 2021 to August 31, 2021. data retrieved from Amazon Web Services.

[Kayal and Rohilla, 2019] Kayal, P. and Rohilla, P. (2019). Bitcoin in the literature of economics and finance: a survey. *Available at SSRN 3441652*.

[Khedr et al., 2021] Khedr, A. M., Arif, I., P V, P. R., El-Bannany, M., Alhashmi, S. M., and Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *Intelligent Systems in Accounting, Finance and Management*, 28(1):3–34.

[Kjærland et al., 2018] Kjærland, F., Khazal, A., Krogstad, E. A., Nordstrøm, F. B., and Oust, A. (2018). An analysis of bitcoin's price dynamics. *Journal of Risk and Financial Management*, 11(4):63.

[Kodama et al., 2017] Kodama, O., Pichl, L., and Kaizoji, T. (2017). Regime change and trend prediction for bitcoin time series data. In *CBU International Conference Proceedings*, volume 5, pages 384–388.

[Kristoufek, 2013] Kristoufek, L. (2013). Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, 3(1):1–7.

[Kristoufek, 2015] Kristoufek, L. (2015). What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PloS one*, 10(4):e0123923.

[Mallqui and Fernandes, 2019] Mallqui, D. C. and Fernandes, R. A. (2019). Predicting the direction, maximum, minimum and closing prices of daily bitcoin exchange rate using machine learning techniques. *Applied Soft Computing*, 75:596–606.

[McNally et al., 2018] McNally, S., Roche, J., and Caton, S. (2018). Predicting the price of bitcoin using machine learning. In *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*, pages 339–343. IEEE.

[Medalie, 2019] Medalie, C. (2019). Data deep dive: Why normalizing cryptocurrency trade data is so complex. *Medium*. https://blog.kaiko.com/why-normalizing-cryptocurrency-trade-data-is-so-difficult-1ac1c388f3e2.

[Menton, 2021] Menton, J. (2021). Millennials are quitting jobs to become crypto day traders. here's the risk, reward. *USA Today*.

[Mezquita et al., 2021] Mezquita, Y., Gil-González, A. B., Prieto, J., and Corchado, J. M. (2021). Cryptocurrencies and price prediction: A survey. In *International Congress on Blockchain and Applications*, pages 339–346. Springer.

[Moosa and Burns, 2014] Moosa, I. and Burns, K. (2014). A reappraisal of the meese–rogoff puzzle. *Applied Economics*, 46(1):30–40.

[Nasdaq Inc., 2021] Nasdaq Inc. (2021). Daily Nasdaq Bitcoin Reference Price, August 1, 2021 to August 31, 2021. data retrieved from https://www.nasdaq.com/market-activity/cryptocurrency/btc.

[Phillips and Gorse, 2018] Phillips, R. C. and Gorse, D. (2018). Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PloS one*, 13(4):e0195200.

[Polasik et al., 2015] Polasik, M., Piotrowska, A. I., Wisniewski, T. P., Kotkowski, R., and Lightfoot, G. (2015). Price fluctuations and the use of bitcoin: An empirical inquiry. *International Journal of Electronic Commerce*, 20(1):9–49.

[Ramadani and Devianto, 2020] Ramadani, K. and Devianto, D. (2020). The forecasting model of bitcoin price with fuzzy time series markov chain and chen logical method. In *AIP Conference Proceedings*, volume 2296, page 020095. AIP Publishing LLC.

[Uras et al., 2020] Uras, N., Marchesi, L., Marchesi, M., and Tonelli, R. (2020). Forecasting bitcoin closing price series using linear regression and neural networks models. *PeerJ Computer Science*, 6:e279.

[Vaddepalli and Antoney, 2018] Vaddepalli, S. and Antoney, L. (2018). Are economic factors driving bitcoin transactions? an analysis of select economies. *Finance Research Letters*, 163(12):106–109.

[Valencia et al., 2019] Valencia, F., Gómez-Espinosa, A., and Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, 21(6):589.

[Vieira, 2017] Vieira, P. J. M. (2017). Price analysis of bitcoin: Volatility, key drivers and evolution.

[Viglione, 2015] Viglione, R. (2015). Does governance have a role in pricing? cross-country evidence from bitcoin markets. *Cross-Country Evidence from Bitcoin Markets (September 25, 2015)*.

[Wu, 2021] Wu, C. (2021). Window effect with markov-switching garch model in cryptocurrency market. *Chaos, Solitons & Fractals*, 146:110902.

[Zhou, 2019] Zhou, Z.-H. (2019). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.