TEXAS TECH UNIVERSITY™

Edward Whitacre Jr. College of Engineering

Sentiment Analysis and Bitcoin Price Prediction

Toyosi Bamidele

ENGR 6330 – Comprehensive Masters Evaluation

Date: 07/06/2021

Table of Contents

Abstract

The rise of cryptocurrency, i.e., Bitcoin, as decentralized means of money transfer has given rise to the need for efficient methods for Bitcoin price forecasting. This paper assesses the influence of Bitcoin-related Google web search volume and Tweet sentiment data in the prediction of Bitcoin prices during a short time window. As social media and web search usage increases, opinion mining and machine learning integration provide insights on periods whereby sentiment data and search volume relate to Bitcoin prices. An efficiently mapped data analysis system to ensure data quality and model efficiency is imperative. The paper explores a machine learning framework from Data Sourcing, Data Preprocessing and Transformation, Exploratory Data Analysis, Correlation Analysis, Feature Selection, and Modeling. The system provided insights into areas for model improvement. Feature selection methods employed include Correlation Analysis, F-regression and Shapley values, and Long Short-Term Memory (LSTM) model for Bitcoin close price prediction. Overall, Google web search data proved to be a more robust feature vs. Tweet sentiment, as it improved model goodness of fit and reduced Root Mean Square Error (RMSE) alongside the historical close price. Furthermore, the results indicate that Google trends for the period examined influence Bitcoin's close prediction.

Further studies should explore other web search engines such as Bing for Bitcoin and BTC search volume. Additionally, analyzing and scraping online news headlines relating to Bitcoin for the sentiment of comment entries and the volume of likes or shares can provide insight for future studies. The primary limitations to this study revealed that data quality issues of publicly available tweet data and accessibility issues for the scraping of historical tweet data.

Introduction

The growth of cryptocurrencies within the last decade has opened avenues for price prediction within Machine learning. For example, the Bitcoin price took parity with the U.S. dollar in February 2011 and rose to $5000 in September 2017. It hit its highest price at $60,000 in April 2021 and traded at $35,000 as of July 2021[1]. Cryptocurrency has seen its growth due to the democratized nature of digital assets. This study focuses on Bitcoin prices, Bitcoin Google search data, and Tweet data with hashtag BTC and Bitcoin for a short window between February 5th, 2021, and February 11th, 2021, up to 134 Observations. Specifically, the study focuses on understanding the influence of historical Bitcoin prices, Google search data, and Tweet Sentiment on the Bitcoin close price prediction. It is essential to glean insights into how previous time steps of prices and additional features influence the performance of machine learning algorithms. As a result, bitcoin traders and trading enthusiasts can gain insights into purchasing or selling decisions to maximize investment opportunities. Machine learning techniques provide a method to predict prices alongside time step features. The critical contribution of this study is the introduction of a data mining framework from problem definition to modeling when exploring trend data, opinion mining, and time-series data.

Additionally, Deep learning techniques have shown great promise for Bitcoin price prediction, as illustrated by Mittal et al. [5], specifically the Long Short Term Memory (LSTM) model. The model provides additional long-term dependencies versus traditional neural networks and recurrent neural networks (RNN) to address the vanishing gradient problem. Therefore, the research question involves utilizing LSTM as a base model, including high, low close, and open historical prices, vs. an updated model with additional features towards improving model performance and error minimization.

Keywords: Correlation Analysis, F-regression and Shapley values, Long Short-Term Memory (LSTM), Google web search data, Tweet Sentiment, Bitcoin Price Prediction

Literature Review

I. Machine Learning and Deep Learning Frameworks

Nguyen et al. and Polyzotis et al. [8] discusses the overall machine learning approach utilized in the data community. First, Nguyen et al. present the Cross-industry Standard process (CRISP) for Data mining[8]. The six stages include business understanding, data understanding, data preparation, modeling, evaluation, and model deployment; it denotes the first 5 phases of an iterative approach [8]. Additionally, it touches on selecting the best machine learning algorithm based on different use cases, including exploring the data dimension and domain, computational time, task urgency, desired prediction ranges, and loss minimization to determine the limits based on the machine learning problem. Nguyen et al. furthermore discuss model evaluation techniques such as error metrics and accuracy metrics based on classification, regression, and clustering problems. Finally, Polyzotis et al. [9] also discusses a machine learning framework. However, it introduces the notion of sanity checks by critically challenging the data to ensure that feature ranges, distributions, and labels are sensible using data visualization plots and descriptive statistics.

II. Sentiment Analysis

Angiani et al. [7] discuss opinion mining and its associated data preprocessing steps. It explores polarity and subjectivity analysis and its relation in translating text data to numerical sentiment. It discusses steps for text preprocessing, including essential cleansing, stop word removal, emoticon processing, dictionary usage, negation handling, and stemming. The results revealed the dictionary usage did not improve model performance with the primary text cleansing methods. Beyond that, analyzing the polarity and subjectivity of tweet texts, Ibrahim [2] and Mohapatra et al. [12] utilize Vader scoring, a rule-based technique with the capability to analyze emoticons frequently used on social media platforms (frequently used on social media sites such as Facebook, Twitter, and Instagram). Typical cases would involve the removal of all symbols, including emoticons. However, Vader scoring with the analysis of emoticons further adds to the understanding of sentiment alongside text and provides more value for natural language understanding.

III. **Correlation Analysis and Feature Selection**

Ji [16] proposes correlation analysis as a method for feature selection by exploring the Pearson and Spearman correlations, which depict the strength of the relationship between predictor and target variables. The correlation coefficient computes values to a scale of -1 to + 1 where 1 constitutes a strong negative correlation, 0 constitutes no correlation, and + 1 constitutes a strong positive correlation [16]. A strong correlation depicts a robust linear relationship between highly correlated variables. The highly correlated variables to the target variable reveal higher performance and a lower mean square error (MSE) utilizing the Long Short term memory model in predicting the number of transactions on a bitcoin block. Pirbazari et al. [3] provide insights on various feature selection techniques for energy load forecasting. It introduces feature selection methods such as F- regression, Recursive Feature Elimination, MElastic Net, and Mutual Information alongside an ensemble model for modeling. Results revealed F- regression highly ranked features with the lowest error metric for various clusters analyzed. Lundberg and Lee [12] introduce the concept of Shapley values for model explainability, also being translation as a feature selection method before significant modeling on large datasets. Shapley values based on game theory is an additive attribution method used to rank feature importance based on the prediction of the target variable.

IV. Machine Learning algorithms

Ray [14] discusses frequently used Machine learning algorithms and notes the associate strengths and constraints. It explores the various machine learning use cases, including supervised, unsupervised, semi-supervised, and reinforcement learning. It touches on various industries problems such as stock price prediction, natural processing, understanding, pattern recognition, and recommender systems. Finally, it covers the reviews on several machine learning algorithms such as regression models, decision trees, super vector machines, clustering algorithms, and Bayesian learning to provide readers with an appropriate overview for model selection.

V. Deep learning

Albariqi and Winarko [10] explore neural network models for the prediction of short and long ranges of bitcoin price changes using the Recurrent Neural network (RNN) and Multilayer perceptron (MLP) model. It examines 1300 time-series observations from August 2010 to October 2017 on a 2-day rate up to 60 days. Results revealed that the MLP model performed better on a 60-day range vs. the RNN model on a 56 day. Overall, the models performed better on long-term day ranges vs. shorter day ranges and recommend future work exploring including feature selection and hyperparameter tuning of machine learning using sequence models with varying architectures. Mittal et al. [5] provide insights on Regression, Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) models and assess the impact of sentiment data on bitcoin price fluctuations. Results revealed that LSTM and RNN models realized a more robust performance versus traditional regression techniques. The overall findings revealed tweet sentiment as a weak feature for model improvement for bitcoin price prediction.

Further work suggested include sentiment analysis on other social media sites such as Facebook and news outlets. Siami et al. [15] also provides insights on the use of LSTM for time series forecasting alongside the exploration of the traditional Auto-Regressive Integrated Moving Average (ARIMA) model. The LSTM model provided better model results on the test data vs. the ARIMA model.

VI. **Web Search Data**

Philippas et al. [6] examines tweets and google search trends of Bitcoin and BTC related to Bitcoin price predictions. The analysis shows solid indications for social media data towards driving bitcoin prices into bear or bull positions. The data source compromises Bitcoin prices from January 2016 to May 2019. Features include daily prices of bitcoin, bitcoin Google search trend data, and tweet volume data with "BTC" and "bitcoin" hashtags and keywords. The paper focuses on periods with significant economic and market volatility. The results reveal that the Google search data as a feature is a significant source for volatility for Bitcoin price while tweets only influence periodically. Furthermore, the Google trend data is a more vital feature for price prediction vs. tweet volume. Finally, the author recommends analyzing other digital currencies such as Ethereum, Ripple, Litecoin, and Dash to validate the relationship between social media attention and price prediction.

VII. Ethics and Data Governance

Ethical and Data Governance are prominent issues for various organizations in the current business climate; data privacy is key to building customers and client trust to solve business problems. Jethin et al. [1] touch on ethics concerning tweet data and sentiment analysis; for cryptocurrency price prediction alongside opinion mining, tweet data usage is critical. It carries a great responsibility on their researcher, data analyst, or machine learning engineer to ensure ethical data usage. In addition, users must show intentionality in data storage procedures and data analysis to ensure a concealed user's identity about the text data.

VIII. Cryptocurrency Background

Limba et al. [17] discusses the concept of cryptocurrency and blockchain technologies from a theoretical perspective by exploring social and legal constructs and principles relating to the utilization of digital currencies. First, it explores the history of Bitcoin, starting off with the Nakamoto payment model, a money transfer analogy without the need of a financial institution, where transactions are quick with minimal cost. Second, it discusses the World Economic forum's view on bitcoin anchored on blockchain technology as one of the Top 10 emerging technologies to watch alongside autonomous vehicles, Artificial intelligence (A.I.), and the internet of things (IoT) future of financial platforms. Third, it examines the legal constructs of cryptocurrency, such as lack of regulation, including the threat to consumer protection due to anonymity and traceability concerns for transactions that result in theft. Secondly, the lack of government support for digital currencies implies that many governments are still less supportive of the globalization of cryptocurrency and seen it as a financial opening and not globally regarded as a means of monetary exchange. Overall, future political and government mandates will determine the scope of utilizing digital currencies as a disruptive technology. Although the paper touches on the strengths of cryptocurrency, it focuses on legal constraints. As a result, it lacks depth on the benefit of cryptocurrency benefits. Benefits include the clarity of the audit trails by "removing the middleman" robust security due to the blockchain backbone, whereby transfers cannot be overturned in a situation due to robust encryption techniques.

Topic Motivation

Cryptocurrency, specifically Bitcoin, regarded as the world's first digital currency, is a digital currency that transacts with a peer-to-peer system without the need for a financial institution [17]. It is a beneficial tool based on blockchain technology that democratizes financial services for banking customers by eliminating or minimizing the cost of transactions and accelerating remittance speed to recipients. As of June 26th, 2021, Bitcoin represents a market capitalization of 639 billion dollars and periodic volatile price swings on daily and long-term prices [11]. For example, in 2021, the value of Bitcoin rose to $63,000 on April 13th and fell to $30,000 on May 19th. The Bitcoin price change in the 36 days results in a 52.4% drop in price [11]. The volatility of bitcoin prices introduces a significant level of uncertainty for investors. To further close the gap regarding price forecasts, text data and trend data provide additional insight into the current state of the market. Social media websites and Search engine sites such as Twitter and Google used as information sources show interesting trends regarding price influences due to the popularity of the digital asset. Social media users share their opinions with the public. Their views released as tweets, hashtags, and search trends can influence decision-making, resulting in buy/sell trades based on the sentiment and search data. Critically analyzing the historical bitcoin price directions with sentiment and trend data can be advantageous to Bitcoin traders and trading enthusiasts in purchase decisions.

Furthermore, employing data analytics and machine learning provides individuals with cryptocurrency interests to analyze and predict price changes using additional text data and trend data to make a much more informed decision. Therefore, the primary motivation for this paper is to provide a holistic framework in the use of analyzing time-series data alongside text data by introducing a framework from data collection to data cleansing, data transformation, text data preprocessing, exploratory data analysis, feature engineering, feature selection, and machine learning modeling. Additionally, this paper introduces a framework for analyzing numerical data alongside text data, which is helpful in text or opinion mining problems.

Research problem

This paper intends to provide insight on the extent of the influence of sentiment and search data, on Bitcoin prices, specifically for predicting time series-based data such as the Close daily price. It explores the correlation of Bitcoin prices, including Open, Close, High, Low, and additional features such as feature engineered data using natural language methodologies, sentiment analysis, and Web search trend data.

The dataset of interest is hourly Bitcoin prices from February 5th, 2021, to June 10th, 2021, alongside 800,165 tweet text records with hashtag #Bitcoin and #btc aggregated to an hourly basis for the same period collected via the Twitter API available on Kaggle.com. The Bitcoin prices and tweet data alongside hourly Google search trends data with hashtag #Bitcoin and #btc prefaces the process for analysis and machine learning modeling. Text mining of tweets includes implementing natural language processing, ensuring clean data necessary for sentiment and analysis, and modeling. In this study, Sentiment Analysis explores the social media opinions from tweets to provide an overall picture of users' sentiment for the period. Key sentiment indicators include subjectivity, polarity for text data, and Vader scoring for emoticon processing alongside text data. Ultimately, this study aims to examine hourly bitcoin price predictions using Machine learning models, precisely Long -Short Term Memory (LSTM) and Ensemble Decision Tree models like Random Forest for the Feature selection phase. Furthermore, by assessing error benchmarks such as Root Mean Square Error (RMSE) and evaluation metrics such as R-squared, the impact of additional features alongside price data is measured to determine features with more substantial predictive power for Bitcoin prices.

Methodology used.

This Bitcoin price prediction study introduces a framework to standardize machine learning procedures involving numerical and text data through an iterative approach of 8 main selective phases. The phases included Problem Definition, Data Sourcing, Data Preprocessing and Transformation, Exploratory Data Analysis, Correlation Analysis, Feature Selection, Modeling, and Model Optimization (see Figure 1.1) [8],[9]. A machine learning framework is an iterative approach that post-modeling. It involves redefining the problem definition, data sourcing, or data augmentation based on modeling results and analysis outcomes to improve model performance [9].

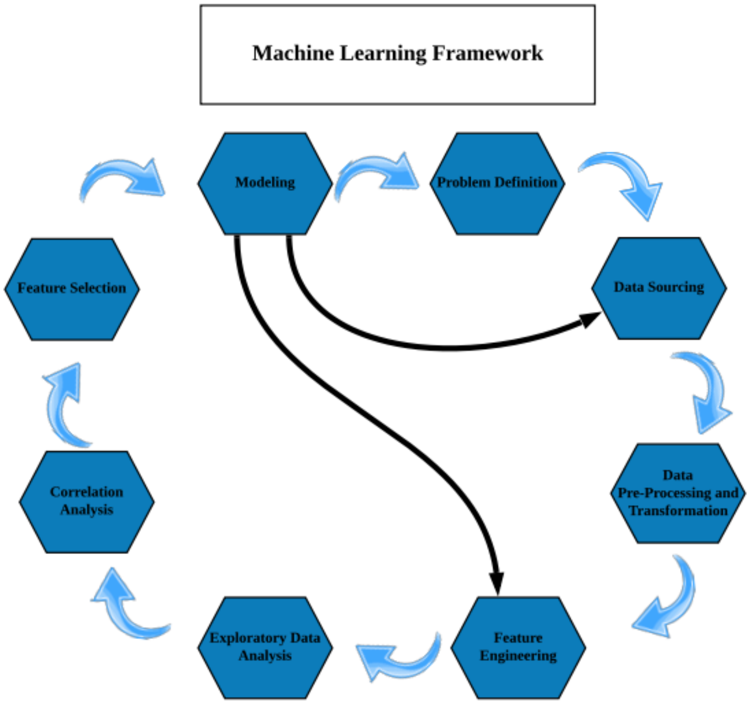


Fig 1.1 Machine Learning Framework

Problem Definition

This study's purpose is to investigate the influence of web search data and tweets to determine its influence on the prediction of Bitcoin Close price alongside historical Close, Open, High, and Low prices to determine the best features for price prediction.

Data Sourcing

Overall, Data sources for this study include publicly available data and data extracted from Application Programming Interface (API) available with integration in Python.

I. Bitcoin Prices

The Bitcoin data is extracted using the Historic-Crypto library, an open-source and interacts with CoinBase Pro API. The historical data retrieved includes Open, Close, High, Low, and Volume of hourly Bitcoin prices from February 5th, 2021, to June 20th, 2021.

II. Tweets

Tweet data is collected from Kaggle.com using the opendatasets library; data is extracted using the Tweepy Python library using #bitcoin and #BTC hashtags. Data collected include the username, user location, user description, account creation date, number of user's followers, number of user's friends, user favorites, date and time of the tweet, hashtags, utility used, and retweet indicator. For this study, the date and text columns ranging from February 5th, 2021, to June 20th, 2021, are used exclusively for analysis and modeling. The tweet data is on a1-min interval timestamps for the tweet text data.

III. Google Trends

Google Search Trends Data for Bitcoin and BTC is extracted using the pytrends API available in Python using the Historical Hourly Interest method of the keywords indicated from February 5th, 2021, to June 20th, 2021.

Data Preprocessing and Transformation

In preprocessing the tweet text data, fundamental cleansing and natural language processes such as morphological analysis ensure cleansed text before further transformation and analysis.

Fundamental Cleansing

This process involves the removal of insignificant symbols and elements to retain clean tweets. The steps include removing URLs, hashtags such as #BTC, #bitcoin, user mentions, and punctuations using regular expressions [7]. Regular expressions enable automated search and replacement of text data, making text analysis more efficient, available as the regex or re library in Python.

Morphological Analysis

Morphological analysis occurs using the nltk library in Python, supplying a series of methods for natural language processing, including Stop word Removal, Word Tokenization and Lemmatization.

Stopword Removal

Stopwords are common English language words filtered during natural language processing. For example, common words include 'the,' 'a',' 'have,' 'is,' do not provide additional meaning or significance to the text or sentence context [7]. Although these words are frequent in text documents, they may hinder information comprehension for algorithms, and removal is necessary to boost word relevancy and effectiveness for analysis and modeling[7].

Word Tokenizing

This process involves splitting individual text observations into smaller chunks called tokens, such as breaking down a sentence such as 'The cat is female' to ['This,' 'is,' 'a', 'cat'][7].

Lemmatization

Lemmatization enables word normalization by finding the word's root meaning called "lemma" contingent on the sentence context; it operates with a dictionary lookup to ensure word normalization to its base form [8]. Examples include such as core word extraction, such as "from finding to find."

Following the Fundamental Cleansing and Morphological Analysis step, the clean text observations are changed to lower case to ensure normalization across the text data. Sanity checks include checking for missing values revealing limited null values in the text data.

Feature Engineering

Feature Engineering is a system of extracting features as predictors from source data concerning the data's subject area[26]. This procedure includes creating new features such as tweet volume, sentiment data, and aggregation procedures.

Tweet volume

The tweet volume feature is extracted by aggregating the number of entries to tweets for each day per hour to provide the number of tweets relating to #BTC and #Bitcoin per hour for the specified date range.

Tweet sentiment

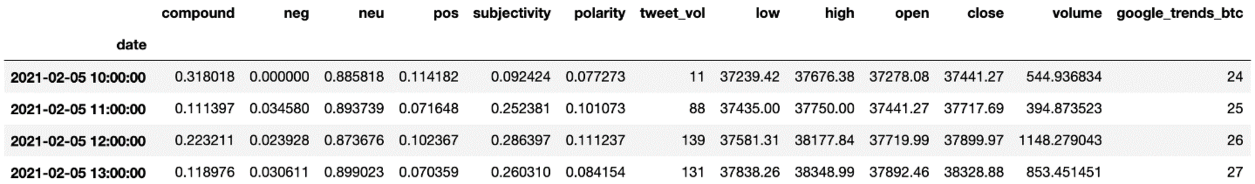
Tweet sentiment provides information on the subjectivity or polarity of tweets to examine attitudes and opinions towards Bitcoin. Public perception of Bitcoin provides information on investor's attitudes. Tweet polarity categorizes tweets as positive, negative, and neutral; positive and negative polarity shows attitudes ranges while neutral polarity shows the negligible impact on decision making regarding Bitcoin purchase or sale [7]. The Textblob library in Python possesses two functions, polarity and subjectivity; when applied to text data returns numerical values. Polarity is on a scale from [-1 to 1], indicating -1 as an opposing opinion and +1 indicates positive opinion [7]. Subjectivity is on a scale [0 to 1] which shows the closer to 1 indicates more subjective text related to less factual information [7]. Additionally, sentiment analysis utilizes Vader scoring, a rule-based tool that deals with words, slangs, and emojis commonly used by social media accounts, split into positive-negative, neutral, and compound polarities [2],[13]. Vader scoring adds an extra level of analysis by identifying emoticons and emojis and translating the attitudes to numerical data vs. plain text data [13].

Data Merging

To get all the sourced data to the same level of granularity, aggregation by mean occurs to transform sentiment data from 1-min intervals to hourly intervals, tweet data, bitcoin prices, and google trends data is merged by using the date column as the primary key.

Exploratory Data Analysis

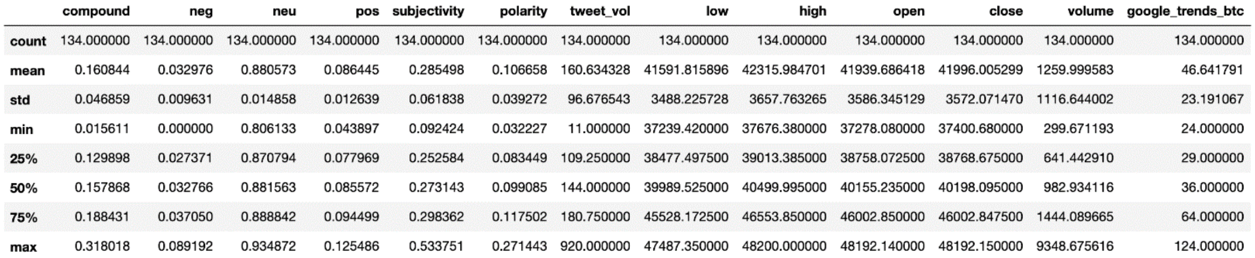
The Data Preprocessing provides the merged data, as seen in Table 1.1, with a snapshot of the first four observations. The processed features include sentiment data (subjectivity, polarity, and Vader scoring attributes), tweet volume, bitcoin prices (low, high, open, close), bitcoin volume, and Google search trends by the hour.



	compound	neg	neu	pos	subjectivity	polarity	tweet_vol	low	high	open	close	volume	google_trends_btc
date													
2021-02-05 10:00:00	0.318018	0.000000	0.885818	0.114182	0.092424	0.077273	11	37239.42	37676.38	37278.08	37441.27	544.936834	24
2021-02-05 11:00:00	0.111397	0.034580	0.893739	0.071648	0.252381	0.101073	88	37435.00	37750.00	37441.27	37717.69	394.873523	25
2021-02-05 12:00:00	0.223211	0.023928	0.873676	0.102367	0.286397	0.111237	139	37581.31	38177.84	37719.99	37899.97	1148.279043	26
2021-02-05 13:00:00	0.118976	0.030611	0.899023	0.070359	0.260310	0.084154	131	37838.26	38348.99	37892.46	38328.88	853.451451	27

Table 1.1 The first four observations of the processed dataset

Descriptive statistics provides a statistical summary of the dataset. The total number of observations selected for modeling is 134 due to data completeness and availability, ranging from February 5th to February 10th, 2021, as shown in Table 1.2. The target value' close' has a maximum, minimum, and average Bitcoin close price of $48,192, $37,400, and $41,996, respectively.



	compound	neg	neu	pos	subjectivity	polarity	tweet_vol	low	high	open	close	volume	google_trends_btc
count	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000
mean	0.160844	0.032976	0.880573	0.086445	0.285498	0.106658	160.634328	41591.815896	42315.984701	41939.686418	41996.005299	1259.999583	46.641791
std	0.046859	0.009631	0.014858	0.012639	0.061838	0.039272	96.676543	3488.225728	3657.763265	3586.345129	3572.071470	1116.644002	23.191067
min	0.015611	0.000000	0.806133	0.043897	0.092424	0.032227	11.000000	37239.420000	37676.380000	37278.080000	37400.680000	299.671193	24.000000
25%	0.129898	0.027371	0.870794	0.077969	0.252584	0.083449	109.250000	38477.497500	39013.385000	38758.072500	38768.675000	641.442910	29.000000
50%	0.157868	0.032766	0.881563	0.085572	0.273143	0.099085	144.000000	39989.525000	40499.995000	40155.235000	40198.095000	982.934116	36.000000
75%	0.188431	0.037050	0.888842	0.094499	0.298362	0.117502	180.750000	45528.172500	46553.850000	46002.850000	46002.847500	1444.089665	64.000000
max	0.318018	0.089192	0.934872	0.125486	0.533751	0.271443	920.000000	47487.350000	48200.000000	48192.140000	48192.150000	9348.675616	124.000000

Table 1.2 Summary Statistics

Data Visualization

Utilizing Data subsets for visualization by grouping Bitcoin prices and volume, google search and tweet volume, and subjectivity and polarity for plotting. The Bitcoin prices (open, close, high, and low) follow a linear relationship with a multimodal distribution due to the presence of more than a singular peak, as shown in Figure 1.2. In addition, the Bitcoin volume distribution shows a right skew with an outlier point at 9,348. Subjectivity and Polarity data trend along the same line. Polarity data is more left-skewed than subjectivity data. Polarity data shows mild positivity with values greater than 0, and with most subjectivity, data less than 0.5 indicates overall opinions and attitudes are more subjective than objective information. Google trends data and tweet volume is left-skewed and show a nonlinear relationship between variables., as seen in Fig 1.3.

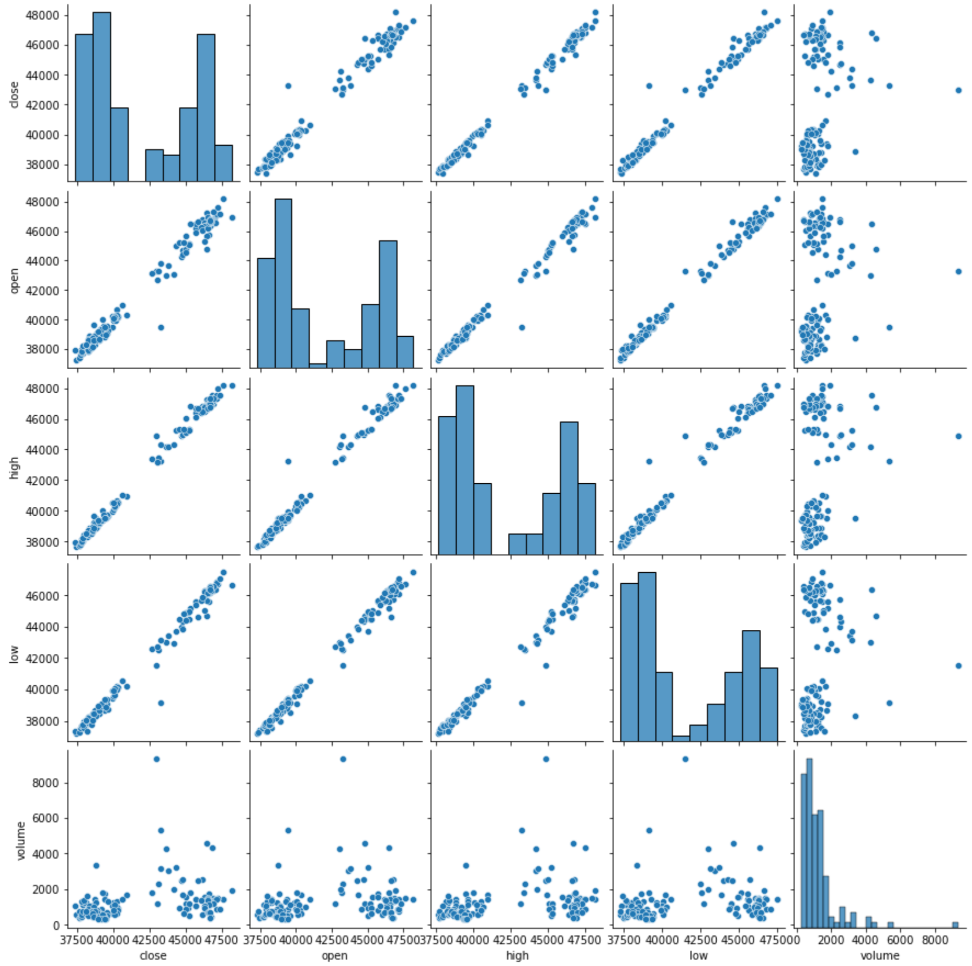


Figure 1.2 Histogram Bitcoin Prices and Volume

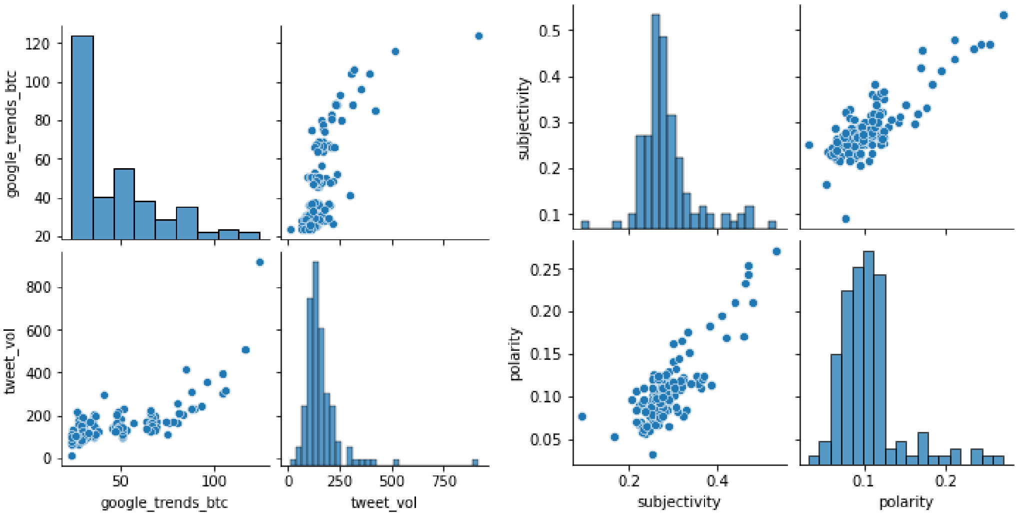


Figure 1.3 Histogram of search volumes and sentiment data

Correlation Analysis

As seen in (1)[16], the Pearson correlation equation shows the correlation between variables, and correlation ranges from -1 to 1, where -1 reflects a strong negative correlation, 0 no correlation, and +1 a strong positive correlation[16].

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (1)$$

$\mathrm{r} = correlation\ coefficient$

$x_i = x - variable\ values$

$\bar{x} = mean\ of\ x\ variables$

$y_i = y - variable\ values$

$\bar{y} = mean\ o\ y\ variables$

Feature Selection

Feature selection enables an increase in computational speed and efficiency by focusing on features that have the most predictive power towards the target, reducing the risk of model overfitting [3]. In addition, other techniques such as the f regression and Shapley values can provide insights into the best features for predicting Bitcoin Prices to validate the correlation analysis further [12]. The study explores feature selection techniques, which are filter and wrapper techniques. Filter techniques subset features and rank them by applying a statistical scoring method. Wrapper methods operate by utilizing a machine learning algorithm evaluation on a series of features and scores the combination of features based on the predictive outcome [8].

F-test

F-test is a filtering approach that uses statistical approaches such as f-scores and p-values, which specify the significance of attributes ranked with the most significant p values [3]. The results indicate the individual effect of each feature on the model's performance. The f_regression function based on sklearn computes the F-statistic as seen in (3)[3] by obtaining the correlation between features matrix X and given target variable y, as seen in (2)[3].

$$\rho_i = \frac{\left(X[:,i] - mean(X[:,i])\right) * (y - mean(y))}{std(X[:,i]) * std(y)} \tag{2}$$

Using the correlation values, the F-statistics:

$$F_i = \frac{\rho_i^2}{1-\rho_i^2} * (n-2) \tag{3}$$

Furthermore, n is the number of observations in the dataset

The SelectKBest function ranks the F-score from highest to lowest. It returns the most significant features, ranked from the highest correlated to the least correlated features to the target variable.

SHAP Values

SHAP values, known as Shapley Additive exPlanations, are additive attributions methods based on game theory [12]. For example, the Shapley values created by Llyod Shapley indicate that the assumption that "if a coalition c collaborates to produce a value **v**, how much did each member contribute to the final values?"[12]. The determination of a fair contribution occurs by sampling coalitions that include all members(baseline) and coalitions that excludes an individual member to determine the member's marginal contribution by finding the difference between the baseline value and the value with the member excluded [12].

The process occurs over a series of permutations of each member group. The mean marginal contribution of the member represents the Shapley value, which is the average amount of contribution that a particular member makes to the coalition value [12]. Regarding machine learning models, Shapley values explain the additive contribution of the predictor variable in predicting the individual target predictions [12].

The model explainability outcome operates by sub-setting the predictor variables while leaving out a predictor variable Xi, followed by computing the effectiveness of Xi's addition to the subset. The Shapley value of feature Xi is defined as the weighted average difference between subsets that include feature Xi and subsets that exclude feature Xi. Overall, the Shapley value measures the features contributing to the machine learning model, as seen in (4)[12].

$$\emptyset_i(f,x) = \sum_{z^i \subseteq x'} \frac{|z'|(M-|z'|-1)!}{M!} \left[f_x(z') - f_x(z' \backslash i)\right] \tag{4}$$

Where:

$\emptyset_i(f,x)$ represents the Shapley value for feature i, given a black box model f, and input data x

z' represents the subset of input data x

M represents the total number of features.

$f_x(z')$ represents a model for the subset with the feature.

$f_x(z'\backslash i)$ represents a model for subset without feature.

The SHAP algorithm in Python introduces Tree SHAP for Decision tree-based models, such as Ensemble models: Random Forest and Gradient boosted trees [23]. The samples feature subset and fit the data-based, and outputs are deemed "approximated Shapley values." Tree SHAP provides arguments for approximate Shapley values rather than computing all permutations, which is computationally expensive.

Algorithm for Feature Selection

The tree algorithms known as CART Classification and Regression trees utilized for the feature selection process are Random Forest, an ensemble tree method, where prediction values are aggregated based on the value of several trees [2],[14]. Based on the base model Decision trees, tree-based models are a supervised machine learning technique that operates by recursively partitioning data based on certain predictor variables to maximize homogeneity. For example, for a regression problem, the recursive partitioning steps occur by selecting a predictor variable x_i(a root node) and picking a value of x_i denoted as s_i(leaf nodes). Following the initial branch, the step divides the training data into equal portions to ensure maximum purity; this process occurs over a series of predictor variables denoted as decision nodes and selected values until the point of reduced impurity called terminal "leaf" nodes.

Random Forest

The Random forest model, a bagging technique as seen in Fig 1.4, is a parallel approach, whereby trees are extended to the maximum extent, unless otherwise specified, given a set of features and labels, ((x1, y1) ……………(xn,yn)), where n represents the number of observations to form a dataset denoted as D as seen in (5)[19].

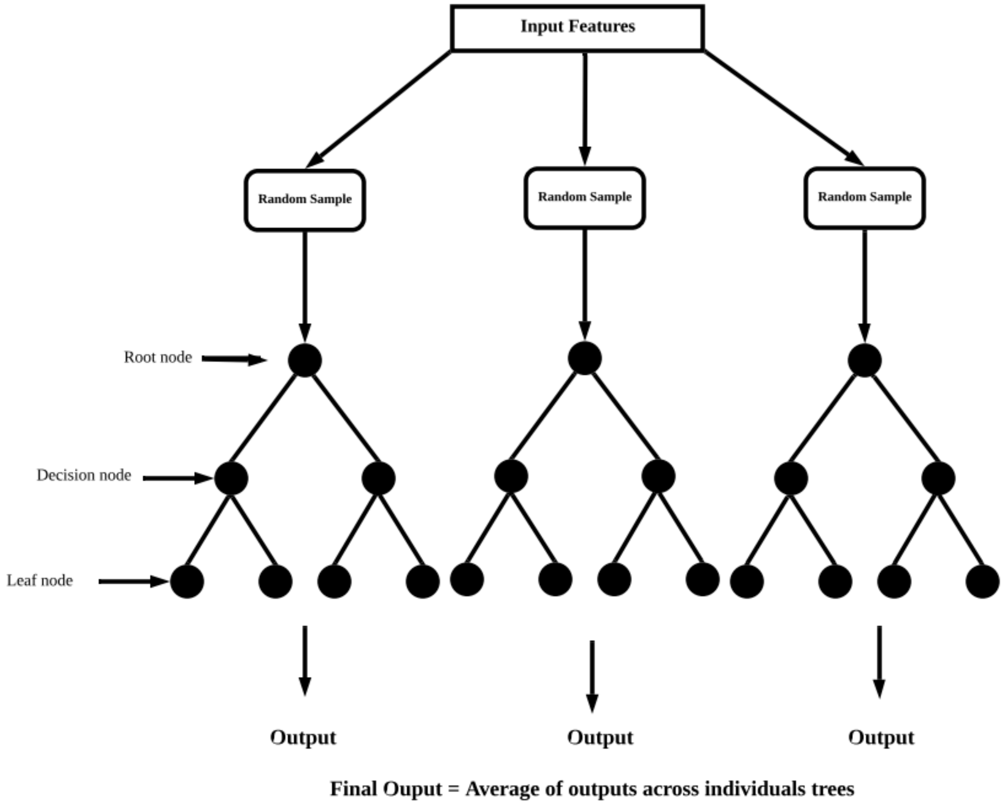


Fig 1.4 Random Forest general architecture

For b = 1 to B:

- B is the total number of trees [22]
- Choose a bootstrap sample of Di (Random Sample) from D (Input Features)
- Construct tree Ti using Di: s.t
    - A random subset of predictor variable are selected at each node
    - Splits will only be considered on feature subsets.
- Regression Predictions are based on averaging outputs across all trees in parallel T_b[19]

$$f_{rf}^{B} = \frac{1}{B}\sum\nolimits_{b=1}^{B} T_b(x) \qquad (5)$$

Modeling

Long Short-Term Memory (LSTM) networks serves as an updated form of Recurrent Neural Networks (RNN) long-term dependencies learning capabilities [15]. Before applying LSTM, it is vital to explore neural networks, specifically Artificial Neural Network (ANN) and Recurrent Neural Network (RNN).

Artificial Neural Networks (ANN)

Artificial Neural Networks, a Feed-Forward network, is an algorithm based on the brain's activity and how neurons are connected to enable learning [16]. Denoted as a black box, it typically possesses three layers with a set of nodes within each layer, which includes:

i. Input layer which accepts the predictors through individual nodes

ii. The hidden layer, the transformation process/ layer using the activation function.

iii. The output layer is the layer that produces the predicted target values.

It operated using forward and backward propagation, as seen in Figure 1.5:

i. The forward pass of predictors goes through the neural network through each node in the input layer, and transformation occurs using initialized weights and bias terms. Then, through to the hidden layer and a selected activation function is used to transform the output of the hidden layer node and produce output results.

ii. The error associated with each output node, the loss function, determines the backward propagation process, specifically the gradient for each node. Therefore, conducted to update weights to reduce the error with each forward pass until a necessary accuracy defined is reached and the error associated with the weights is negligible.

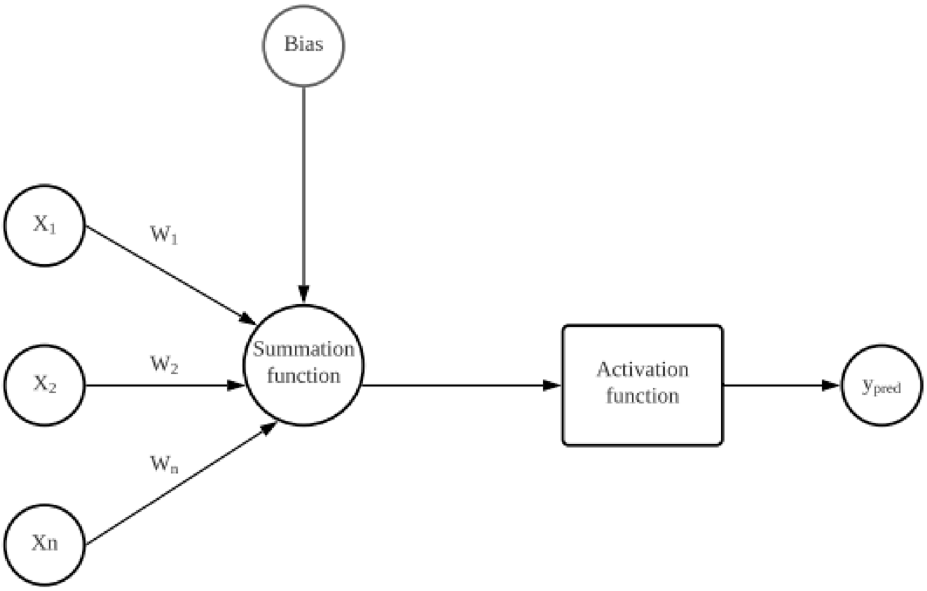


Fig 1.5 Artificial Neural Network (ANN) Architecture [16]

y_{pred} = f(x) = $\sum x_i \, w_i$

w = weghts

B = bias

X= input

y = prediction

Recurrent Neural Network (RNN)

RNN models benefit sequence or time-series data. It works similarly to the ANN but possesses a hidden state representing information from previous steps, as seen in Figure 1.6 with whereby 'A' represents the hidden state. It is a recurrent connection, where the final output from the sequence is passed to the feed layer to produce the predicted values. Short-term memory is a drawback of RNN as more backpropagation steps occur [20]. It has trouble retaining information from the previous step due to the vanishing gradient problem, which is the nature of backpropagation used to optimize a neural network. The gradient will exponentially shrink as it goes through the backpropagation process. A small gradient means minor adjustments causing poor learning for early layers, failing to learn long-range dependencies across timestamps.

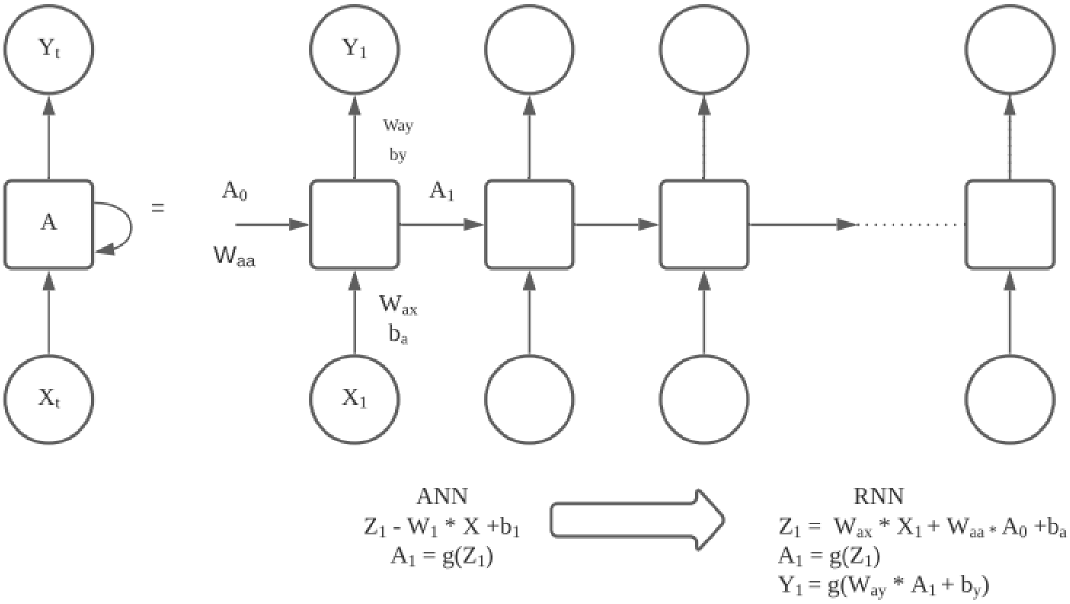


Fig 1.6 Recurrent Neural Network (RNN) Architecture [20]

Long Short-Term Memory (LSTM)

LSTMs represent an updated recurrent neural network (RNN) with the capability of learning long-term sequence dependencies using "gates."[15]. Gates can learn what information to add or remove to a state and eliminate short-term memory problems[24]. The LSTM network possesses three gates. The cell state acts as a conduit that transmits previous data along the sequence, regarded as the memory of the network [24]. It carries data from the processing of the sequence, including data from previous time steps, leading to a reduction in short-term memory problems. As it progresses through the sequence chain, information gets integrated or eliminated to the cell state through the various gates[24]. The network gates determine information permitted to the cell state. The gates compute relevant outputs which are kept or ignored during modeling. Gates contain the outcome of sigmoid transformations between 0 and 1, which helps in updates, 0 may be forgotten, and values multiplied by 1 stay the same [24].

Three gates that regulate information flow through an LSTM illustrated in Figure 1.7 include:

- Forget gate: The outcome of the computation through the forget gate reveals data for elimination or retention through the sigmoid activation function, values, where 1 or 0 are returned, closer to 0 indicates forget, closer to 1 indicates keep [24].
- Input gate: For cell state updates, the input gate process involves the utilization of prior hidden state, and the current input transformed through the activation function sigmoid function transforming outputs within 0, and 1 range, sequentially the hidden state and current state are passed to the tanh activation function to transform values within -1, and 1 range, the transformation outcomes from the tanh and sigmoid activation functions are multiplied [24]. The sigmoid update will determine the data kept from the prior output.
- Output gate decides the next hidden state.

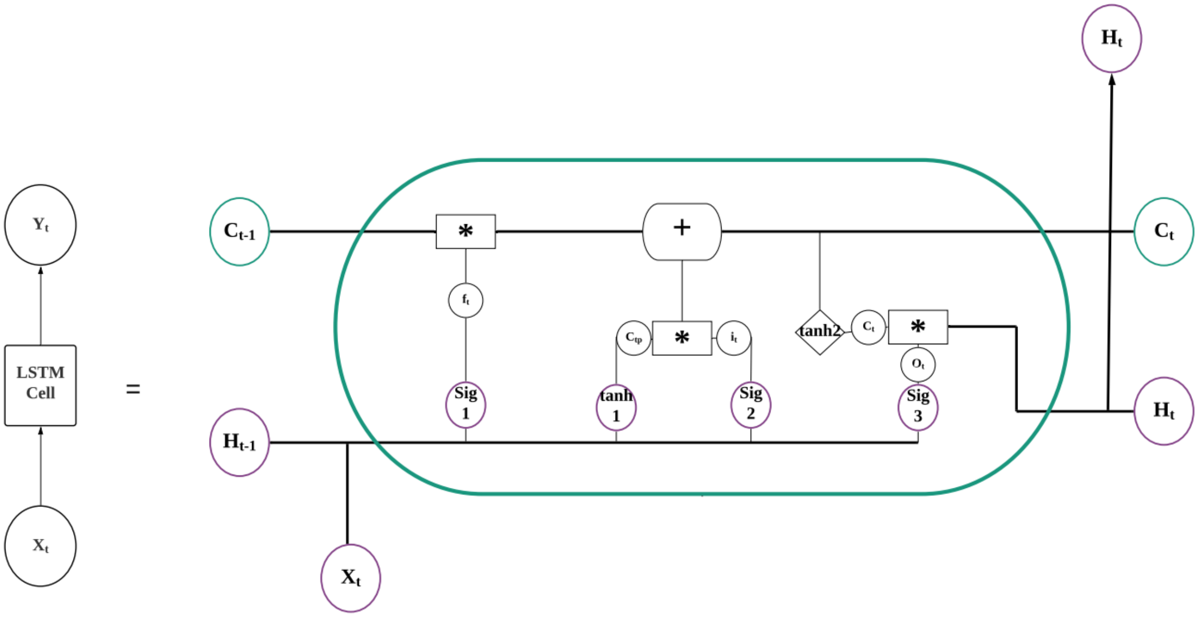


Fig 1.7 LSTM architecture [24]

Process flow steps [26]:

- **Forget gate.**
    - H_{t-1}+ X_t -> sigmoid function 1-> f_t (0,1)
        - Closer to 0 – forget.
        - Closer to1 - keep.
- **Input gate**
    - H_{t-1}+ X_t -> sigmoid function 1-> I_t (0,1)
    - H_{t-1}+ X_t -> tanh function 1 -> C_{tp} (-1,1) * regulate network.
- **Cell state Update**
    - $C_t = f_t * C_{t-1} + I_t + C_{tp}$
- **Output gate**
    - H_{t-1} + X_t -> sigmoid function 3 -> O_t (0,1)

- **Hidden State update**
    - C_t -> tanh function -> C tanh output
    - C tanh output * O_t -> H_t

C_t and H_t are carried over to the new time step.

Whereby variables are:

Inputs

- X_t represents the current input.
- C_{t-1} represents memory from previous LSTM unit.
- H_{t-1} represents the output from the last LSTM unit.

Computations

- f_t represents the forget gate output.
- I_t input gate output.
- C_{tp} – network regulation step

Outputs

- O_t output gate outcome
- C_t represents new updated memory.
- H_t represents final output/ next hidden state.

Activation Functions

- Sig – sigmoid function
- Tan h – tan h function

Vector operations

- * Scaling information
- + adding information.

Results

Trend Analysis

The trend analysis reveals across google searches, tweet volume, and Bitcoin volume an upward spike on February 8th, 2021, as seen in Figure 2.1, which trends upwards with bitcoin prices as shown in Figure 1.5. With online news search, this spike reflects the news release of Telsa's purchase of $1.5 billion in bitcoin and its acceptance as a form of payment for its offerings[21].

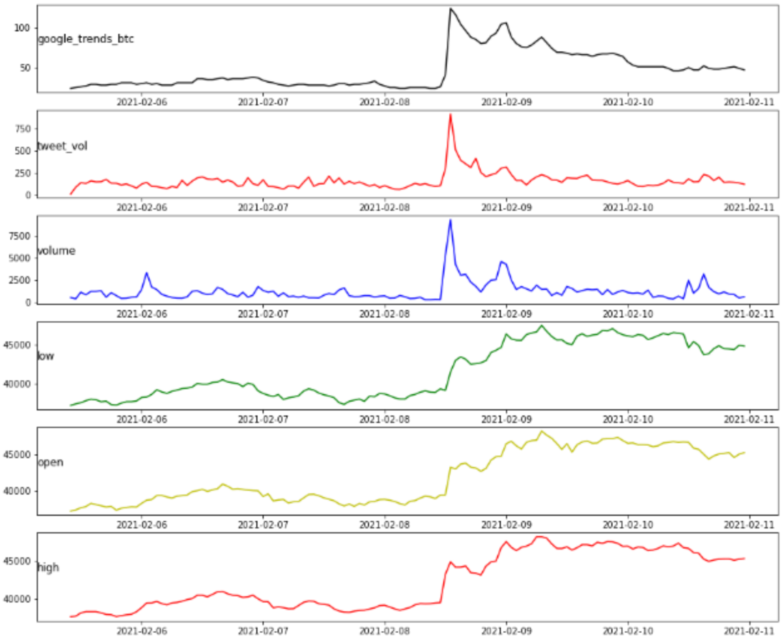


Figure 2.1 Trend plot (Bitcoin price attributes and search and tweet volumes)

Correlation Analysis

The Pearson correlation coefficient of that data shows a strong correlation among the Bitcoin price attributes between 0.99 and 1, and google trends data show a positive correlation of 0.76. In contrast, tweet volume shows a low correlation of 0.3 concerning the close price. Overall sentiment data, as seen in Figure 2.2 attributes, reveal a weak correlation concerning close price, which is on par with Ji[16] results, as shown in Figure 1.

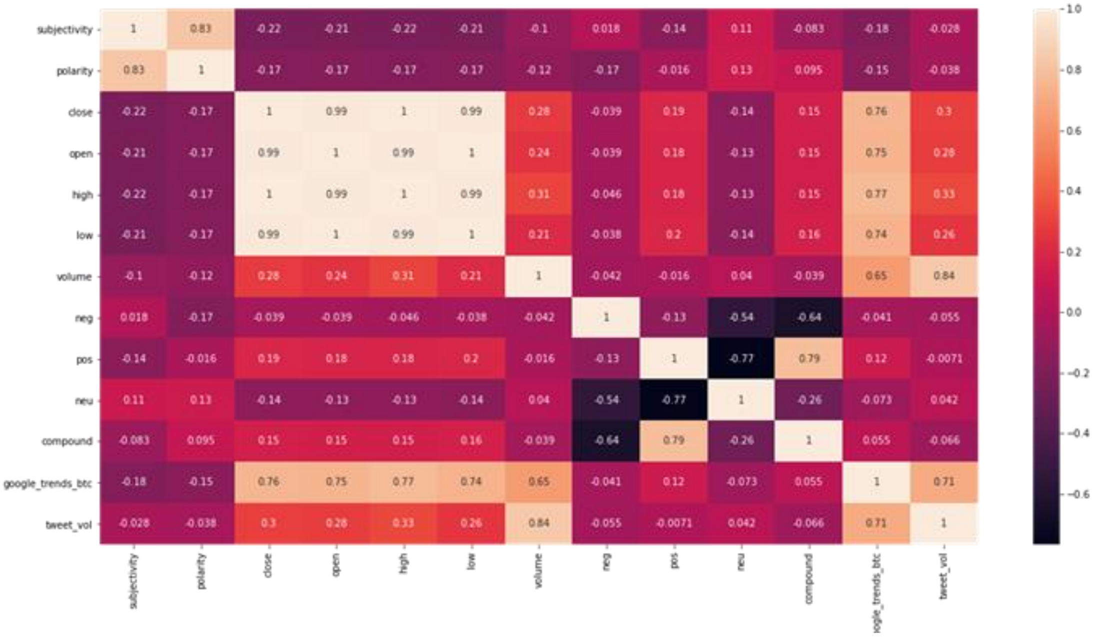


Figure 2.2 Correlation matrix of variables

FRegression

As seen in Figure 2.3, the f regression computation output reveals high, low, open, close, and google trends and tweet volume as the top five predictors.

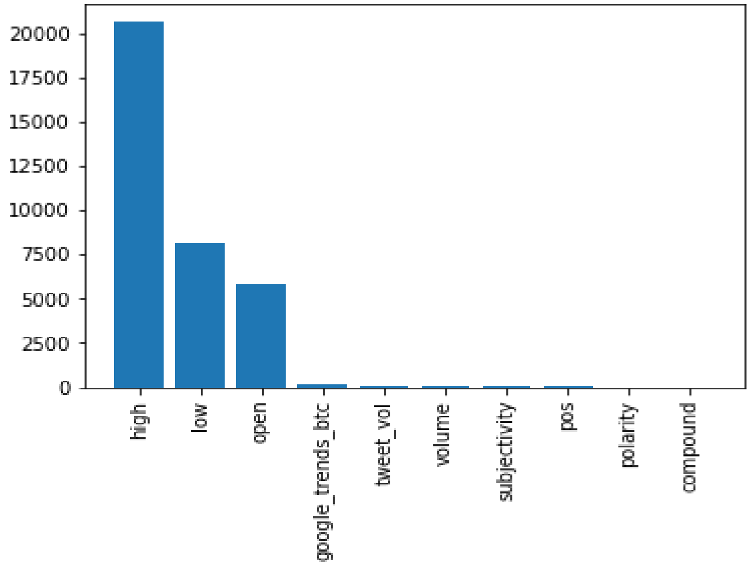


Figure 2.3 Results of F-test showing Top 10 predictor variables

Shapley values

As seen in Figure 2.4 coupled with the Random Forest, the output of the Shapley values algorithm reveals the top predictors as follows high, google trend, open, low.

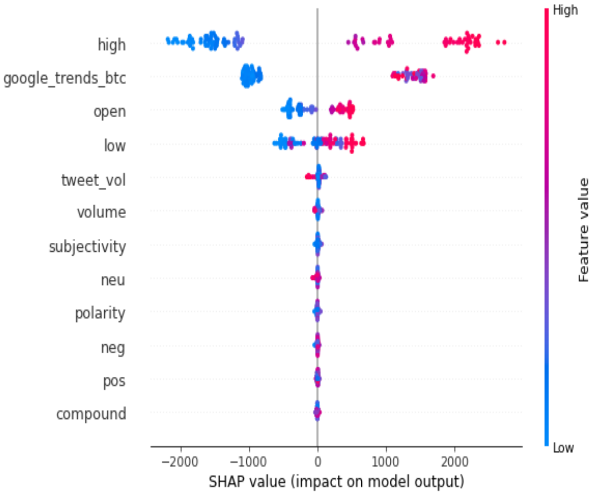


Figure 2.4 Results of F-test showing Top 10 predictor variables.

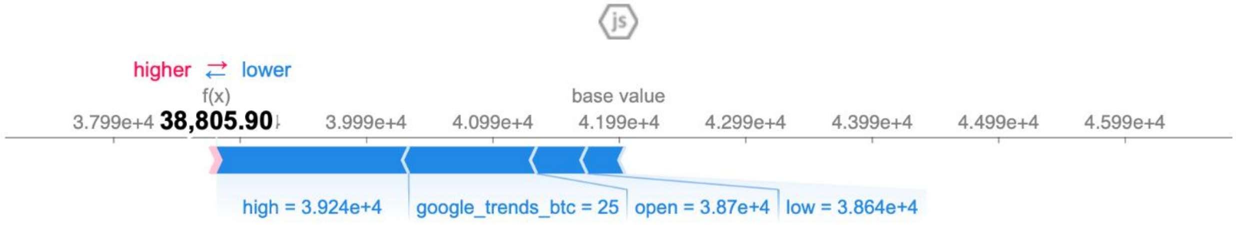
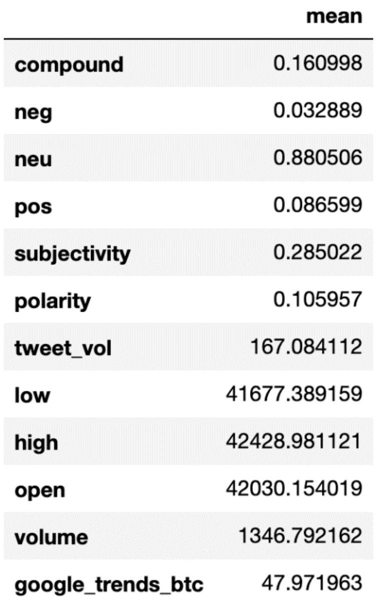


Figure 2.5 Shapley value output for the first target variable

Examining the prediction of the first observation based on Shapley values and Random Forest operation. The output value f(x) represents the prediction for the close price at 38,805.90. The base value, which is the mean value across the target values y, the features depicted with blue, push the value lower from the base value. Overall high, google trends search, open and close impact the outcome prediction, As seen in Figure 2.5.



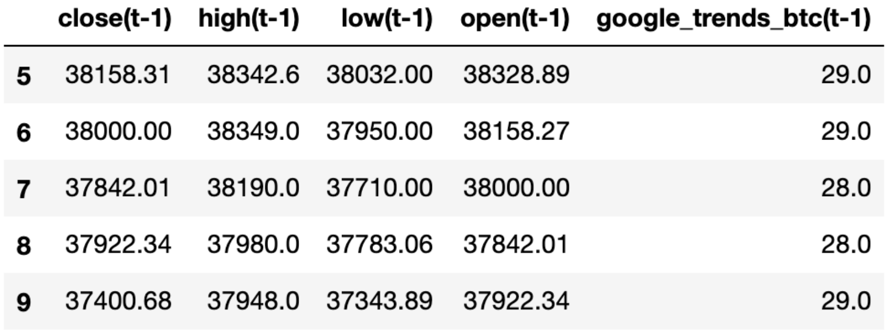
	mean
compound	0.160998
neg	0.032889
neu	0.880506
pos	0.086599
subjectivity	0.285022
polarity	0.105957
tweet_vol	167.084112
low	41677.389159
high	42428.981121
open	42030.154019
volume	1346.792162
google_trends_btc	47.971963

Table 2.1 Average values of variables in the test set

Given the base value of $41,990, the individual features are lower than their respective mean values, as seen in Table 2.1, thus pushing the prediction to the left from the base value.

Modeling Results

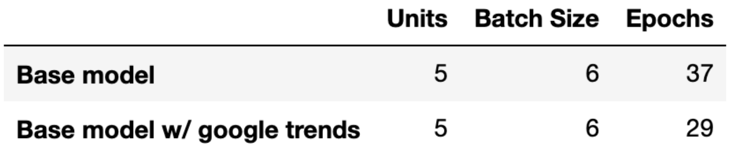
The LSTM model was trained using an 80% split on train data and a 20% split on test data. In addition, hyperparameters such as units, batch size, and epochs were set for the model, as seen in Table 2.3. The LSTM model was used to test the impact of the addition of the google trends features alongside base features: close, high, low, and open in model performance. Additionally, the time series dataset for the features was framed to a pair of input and target sequences with five time steps across all scaled features for modeling; see output for one step on Table 2.2 (unscaled data).



	close(t-1)	high(t-1)	low(t-1)	open(t-1)	google_trends_btc(t-1)
5	38158.31	38342.6	38032.00	38328.89	29.0
6	38000.00	38349.0	37950.00	38158.27	29.0
7	37842.01	38190.0	37710.00	38000.00	28.0
8	37922.34	37980.0	37783.06	37842.01	28.0
9	37400.68	37948.0	37343.89	37922.34	29.0

Table 2.2 Top 5 observations for time step 1 of the reframed dataset

Long Short-Term Model Parameters

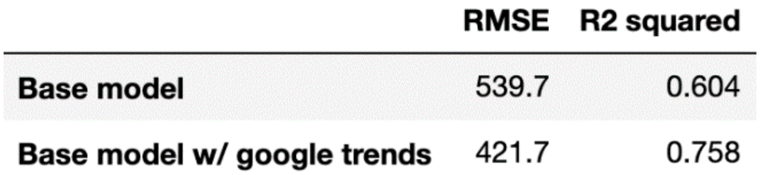


	Units	Batch Size	Epochs
Base model	5	6	37
Base model w/ google trends	5	6	29

Table 2.3 Hyperparameters for LSTM model *Early stopping utilized to combat overfitting.

The modeling results, as seen in Table 2.4 for the base model and base model features, including google trends, yield 539.7 and 421.7 for the Root mean square error and R2 squared values of 0.604 and 0.758, respectively.

Long Short-Term Results



	RMSE	R2 squared
Base model	539.7	0.604
Base model w/ google trends	421.7	0.758

Table 2.4 LSTM error metric and goodness of fit results

Base model

The base model and updated model produce a good fit with a slight gap between the training and test loss curves. The training and test loss decreased and stabilized to the same level as Figures 2.6 and 2.7, indicating a good fit for the data model. The actual and predicted time series for the base and base models with google trends data are depicted in Figures 2.8 and 2.9.

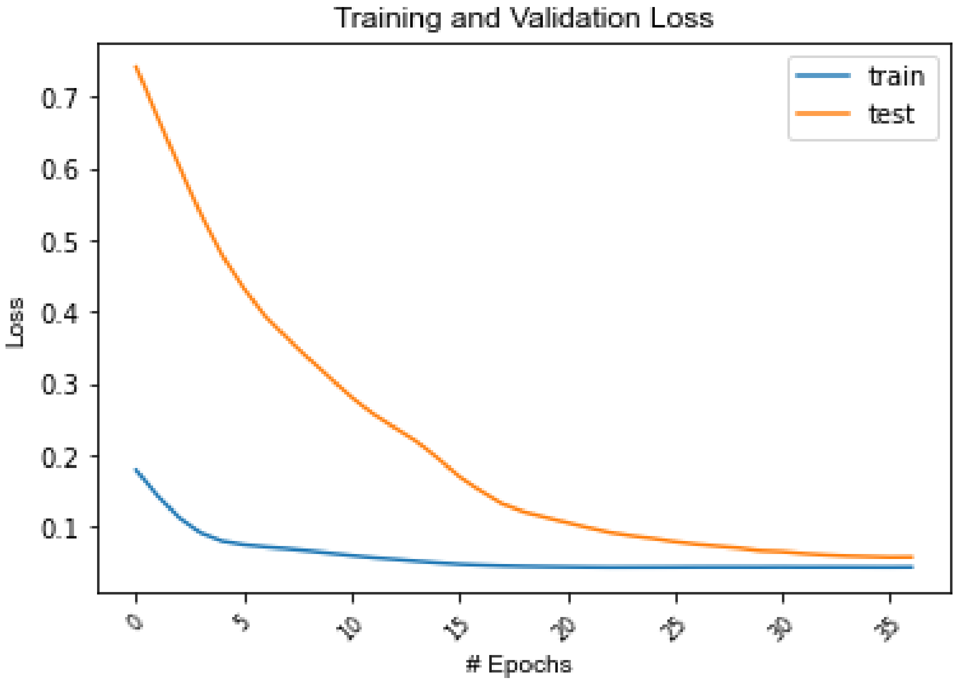


Fig 2.6 Base model Training and Validation loss curve

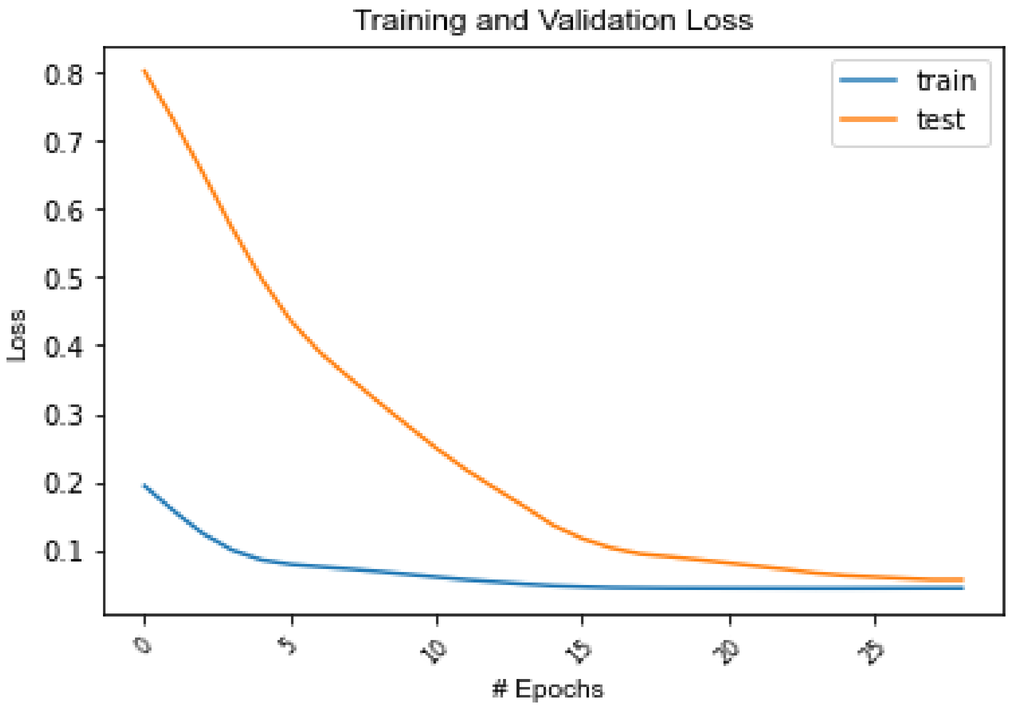


Fig 2.7 Base Model with google trends: Training and Validation loss curve

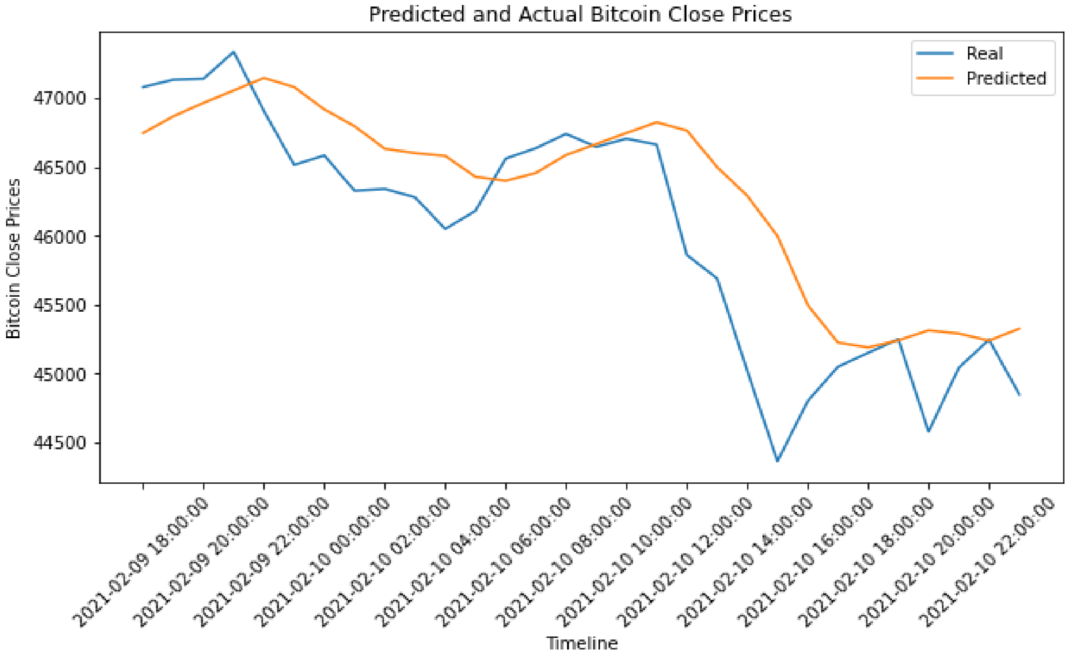


Fig 2.8 Base model: Predicted vs. Actual Bitcoin Close Price on the test set

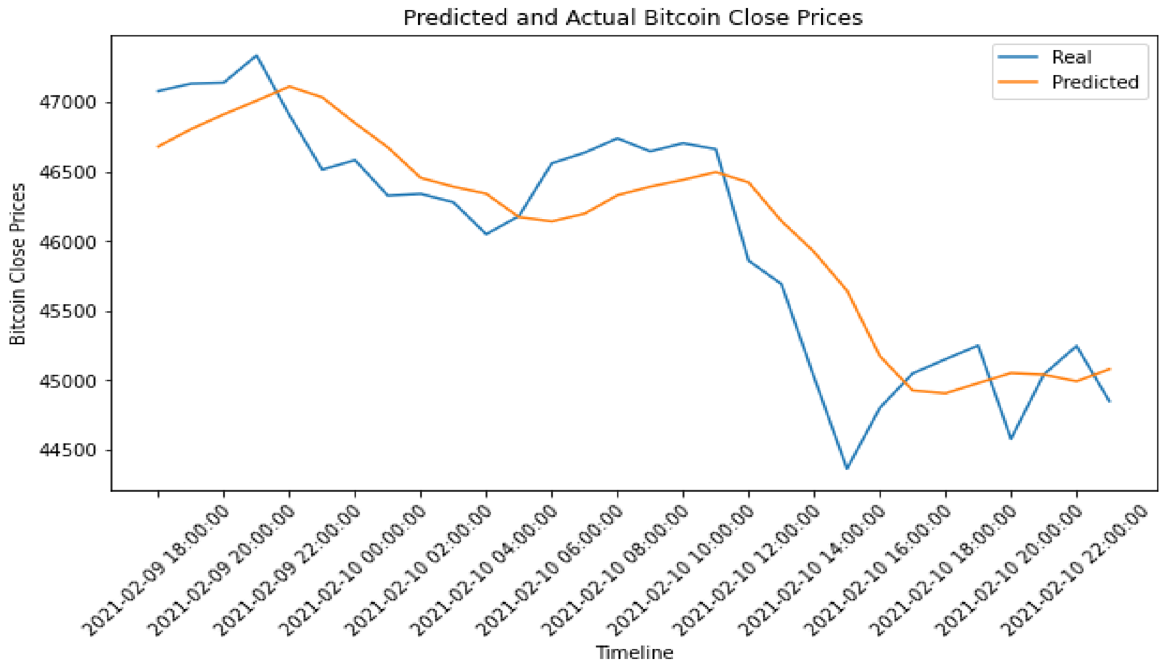


Fig 2.9 Base model w/ google trends: Predicted vs. Actual Bitcoin Close Price on the test set

Discussion

The correlation analysis, F-statistic test, and Shapley values all ranked the tweet sentiment attributes lower than google search data and tweet volume. The results indicate that search data is a more robust indicator of price vs. subjectivity and polarity, which is in line with the findings of Mittal et al. [4] and Albariqi and Winarko [10]. In addition, the results indicate that the correlation analysis, F-statistic test, and Shapley values expansibility provide value for feature selection. The top 5 correlation values, F-statistic values, and shapely values revealed similar results. However, Shapely values ranked google trends higher as a strong indicator for close price prediction vs. high, low, and open prices.

The model results indicate that the inclusion of the google trends features increased the model's goodness of fit by 15% compared to the base model and lowered the RMSE value by $118. The modeling and analysis results provide insight into using time historical price data alongside correlated features for prediction purposes. The reframed supervised problem shows that additional features aside from the previous day's price can provide an increased performance of modeling on the test dataset; additionally, employing early stopping can provide the appropriate number of epochs to reduces issues such as model overfitting.

The main limitation of this study related to data quality is that high data quality is imperative for training machine learning models rather than hyperparameter optimization. A subset of the data excludes the text input feature, which would have provided a larger dataset for the machine learning process, hence post the cleansing step, only the first 137 observations were utilized to ensure an efficient and robust analysis, as these data points were the most complete. Further research is required to establish if varying or more extended periods show different results with tweet sentiment, google trends data, and historical price data. Furthermore, this study focused on a short time window. The 137 hours examined have a peculiar pattern mainly related to the Tesla announcement by Elon Musk, and exposing the model to more extended time-series observations with more volatility might yield alternative insights and provide opportunities for hyperparameter optimization.

Conclusion

The research aimed to identify the effectiveness of additional features such as sentiment data and google trends search alongside the previous bitcoin "close," "high," "low," and "open" price in predicting the Bitcoin close price. Based on the quantitative analysis for feature selection, it can be concluded that the correlation analysis results can be further supported with the F-statistic test and Shapley explainability for feature selection rather than employing all engineered features, which could have lower performance on the model and decreased computational speed. Furthermore, this research clearly illustrates that google trends for the period examined proved an increase in the goodness of fit and decrease in RMSE compared to the base model. Based on these conclusions, bitcoin trading should consider inclusion on web search data for price modeling. Additionally, this technique can be explored on other cryptocurrencies such as Ethereum. For a better understanding of the implications of these results, additional data collection for tweet sentiment for more observation will prove valuable to the model optimization. Currently, the Twitter API provides only seven days' worth of historical data and requires a Twitter developer license before data can be accessed, so data accessibility is also a core issue; otherwise, publicly available data (i.e., Kaggle.com) must be scrutinized to ensure completeness before analysis and modeling. Future studies can examine the search data related to bitcoin and BTC on search engines such as Bing, Baidu, and Duck Duck Go.

Additionally, examining the top news sites with headlines relating to BTC and Bitcoin by scraping comment entries and analyzing the number of likes, shares, and volume of comment entries for additional inference for bitcoin price prediction. Ethical considerations for future studies for analyzing tweet data should ensure data privacy of tweet user id during sentiment analysis publications and discussion. The main benefit of Twitter accessibility involves the progression of machine learning procedures for improving various cases of tweet volume and sentiment data and not the exploitation of user's free speech.

Appendix(es)

Page Requirements

Reference

[1] A. Jethin, H. Daniel, N John, and I. Juan "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," *SMU Data Science Review*: Vol. 1: No. 3, Article 1. (2019)

[2] A. Ibrahim, "Forecasting the Early Market Movement in Bitcoin Using Twitter's Sentiment Analysis: An Ensemble-based Prediction Model," *2021 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021.

[3] A. M. Pirbazari, A. Chakravorty, and C. Rong, "Evaluating Feature Selection Methods for Short-Term Load Forecasting," *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2019.

[4] A. Mittal, V. Dhiman, A. Singh, and C. Prakash, "Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and Web Search Data," *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 2019.

[5] A. Mittal, V. Dhiman, A. Singh, and C. Prakash, "Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and Web Search Data," *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 2019.

[6] D. lippas, H. Rjiba, K. Guesmi, and S. Goutte, "Media attention and Bitcoin prices," *Finance Research Letters*, vol. 30, pp. 37–43, 2019.

[7] G. Angiani, L. Ferrari, T. Fontanini, P. Fornacciari, E. Lotti, F. Magliani, S. Manicardi," A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter". *KDWeb,* (2016).

[8] G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. López García, I. Heredia, P. Malík, and L. Hluchý, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, 2019.

[9]N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data Lifecycle Challenges in Production Machine Learning," *ACM SIGMOD Record*, vol. 47, no. 2, pp. 17–28, 2018.

[10] R. Albariqi and E. Winarko, "Prediction of Bitcoin Price Change using Neural Networks," *2020 International Conference on Smart Technology and Applications (ICoSTA)*, 2020.

[11] R. de Best , "Bitcoin price from October 2013 to July 5th, 2021," *Statista*. [Online]. Available: https://www.statista.com/. [Accessed: 05-Jul-2021].

[12] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions.," *CoRR*, vol. abs/1705.07874, 2017.

[13] S. Mohapatra, N. Ahmed, and P. Alencar, "KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments," *2019 IEEE International Conference on Big Data (Big Data)*, 2019.

[14] S. Ray, "A Quick Review of Machine Learning Algorithms," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon),* 2019.

[15] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018.

[16] S.-H. Ji, U.-J. Baek, M.-G. Shin, Y.-H. Goo, J.-S. Park, and M.-S. Kim, "Best Feature Selection using Correlation Analysis for Prediction of Bitcoin Transaction Count," *2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2019.

[17] T. Limba, A. Stankevičius, and A. Andrulevičius, "Cryptocurrency as disruptive technology: theoretical insights," *Entrepreneurship and Sustainability Issues*, vol. 6, no. 4, pp. 2068–2080, 2019.

[18] T. Korenius, J. Laurikkala, K. J�rvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," *Proceedings of the Thirteenth ACM conference on information and knowledge management - CIKM '04*, 2004.

[19] R. Zou and M. Schonlau, "Applications of Random Forest Algorithm," *https://www.stata.com/*. [Online]. Available: https://www.stata.com/meeting/canada18/slides/canada18_Zou.pdf. [Accessed: 05-Jul-2021].

[20] W. Feng, N. Guan, Y. Li, X. Zhang and Z. Luo, "Audio visual speech recognition with multimodal recurrent neural networks," *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 681-688, doi: 10.1109/IJCNN.2017.7965918.

[21] Stevekovach, "Tesla buys $1.5 billion in bitcoin, plans to accept it as payment," *CNBC*, 08-Feb-2021. [Online]. Available: https://www.cnbc.com/2021/02/08/tesla-buys-1point5-billion-in-bitcoin.html. [Accessed: 06-Jul-2021].

[22] Y. Mao and A. Monahan, "Linear and nonlinear regression prediction of surface wind components," *Climate Dynamics*, vol. 51, no. 9-10, pp. 3291–3309, 2018.

[23] T. Tan, "Back to Basics: Assumptions of Common Machine Learning Models," Medium, 08-Jun-2020. [Online]. Available: https://towardsdatascience.com/back-to-basics-assumptions-of-common-machine-learning-models-e43c02325535. [Accessed: 02-Jul-2021].

[24] M. Phi, "Illustrated Guide to LSTM's and GRU's: A step by step explanation," Medium, 28-Jun-2020. [Online]. Available: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21. [Accessed: 01-Jul-2021]

[25] C. Olah, "Understanding LSTM Networks," Understanding LSTM Networks . [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed: 03-Jul-2021].

[26] J, WeiWei. "Applications of deep learning in stock market prediction: recent progress." *ArXiv* abs/2003.01859 (2021)