Price forecasting of Bitcoin using Sentimental Analysis on Bitcoin Related Tweets and Bitcoin Related Tweets by Elon Musk

BY :- NAGARJUNA KOCHARLA and SAHITHI NALLANI

Abstract

We performed price forecasting based on bitcoin tweets from 2019-2022 using VADER for sentimental Analysis and Elon Musk's bitcoin tweets from the same time period. We used Kaggle to obtain the tweets in a CSV file. After that, we cleaned the data, pre-processed it, filtered only bitcoin-related tweets by running the dataset through a list of keywords, and got the sentiment scores based on the polarity using Vader. We also did an ADF test to make the data stationary since bitcoin data is a project trend. We then selected features for the models to perform the prediction on our two datasets. While using random forest regression we noticed an MAE of 16 for the bitcoin dataset and around 720 for the Elon musk dataset(which is only 132 data points) and for linear regression the error was around 1375 for the bitcoin dataset and around 1500 for Elon musk.

1. Introduction

Given the impulsive nature of cryptocurrency trading, platforms such as twitter and reddit have become the hotbed for all things stocks in recent times. In our project,we aim to study the impact and forecast bitcoin (BTC) price fluctuation for Bitcoin related tweets from 2019-2022 and for SpaceX and Tesla Founder Elon Musk's tweets for the same time frame using sentiment analysis. The bitcoin data we obtained was on a day to day basis, where features were chosen based on the effect any particular feature has on price metrics. We did a lot of data preprocessing considering the noisy dataset of both the tweets as well as the bitcoin data The sentimental analysis we are using is VADER .Understanding the correlation between tweet polarity and actual price might help the common trader to make a more informed decision.We will be using two machine learning methods for the price forecasting 1. Random Forest Classifier 2. Linear Regression.

2. Related Work

Given the depth and volume of data and features available for cryptocurrency data, many studies and experiments were carried to accurately predict the price of any cryptocurrency based on various metrics. One such study was conducted on forecasting the price of bitcoin using sentiment analysis. The researchers who wrote the paper performed sentiment analysis on bitcoin tweets from 12th of March 2018 to the 12th of May 2018. In total the researchers calculated the polarity(a score from +1 to -1 to tweets) of 92550 tweets. For bitcoin historic data, python's quandl API was used. A random forest regression model was implemented with features as polarity scores, historic bitcoin prices and predictor as close price.

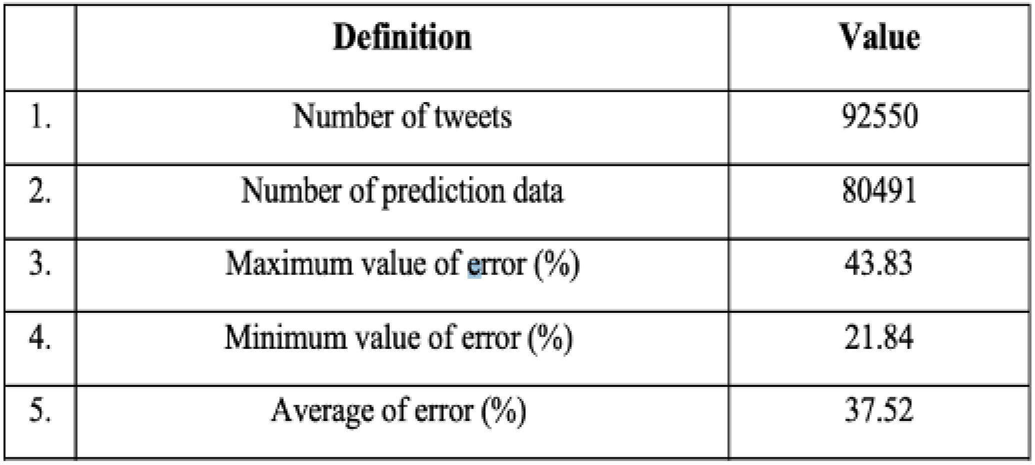
Strengths

Their model achieved an average absolute mean error of 37.52 with a minimum error of 21.84. These metrics are very good for financial data, where the scope for error is high. I believe their usage of VADER(Valence Aware Dictionary and Sentiment Reasoner) enabled them to calculate a continuous value of polarity. For instance, Bitcoin is a must buy might have a polarity score of 0.9 but a tweet like Bitcoin could be a good buy would have a polarity of 0.5. Giving a continuous value as a feature point to the model, in my view, yielded better results.

Limitations and how we improved our data

In cleaning the tweets, the researchers removed @ mentions and did not consider the user favorites to filter out less relevant tweets. Filtering the tweets based on the likes of tweet might have provided them with a richer dataset, we filtered the tweets on these metrics (explained in data and evaluation).Below is a quick review of their models performance

TABLE II. STATISTICS OF ERRORS ON PREDICTION



	Definition	Value
1.	Number of tweets	92550
2.	Number of prediction data	80491
3.	Maximum value of error (%)	43.83
4.	Minimum value of error (%)	21.84
5.	Average of error (%)	37.52

3. Method

We have developed 2 methods to perform the forecasting/predicting the price of bitcoin

1. Random Forest Regression
2. Linear Regression

Random Forest Regression

I chose random Forest regression primarily due to the fact that it is an ensemble learning method which means it combines multiple ML algorithms for prediction, and outputs the best performing method. Since the dataset is large enough, I opted for random Forrest's default split of 75:25. Also, since bitcoin data is continuous, a regression model was chosen. While there are other models like LSTM and ARIMA, Since the data was pre-processed and cleaned, I believe the Machine learning model would do a good job in predicting the price.

Theoretical and Experimental Characteristics

The model was implemented on two datasets.

1. Overall Bitcoin related tweets from 2019-2022
2. Elon Musk Bitcoin related tweets from 2019-2022

For the random Forest model implemented on Dataset 1, I initially give 10 estimator trees (same as in the paper mentioned above), with feature set as tweet polarity score,tweet_user_followers, tweet_user_favourites, bitcoin day open price and trading volume, and the value to be predicted is the day close price.

For the random Forest model implemented on Dataset 2, I initially give 10 estimator trees (same as in the paper mentioned above), with feature set as tweet polarity score , bitcoin day open price and trading volume, and the value to be predicted is the day close price. Although I initially started with 10 as my estimator, after using the gridsearchcv method from sklearn, I found out that an estimator tree value of 50 (I changed this to 90 as 90 was giving the lowest error) and max_depth of 6 gives the lowest mean absolute error. I trained my model using these parameters for a lower error. I did try to use cross validation since the dataset for Elon musk tweets is small but the difference in error was negligible, so I ended up sticking to my initial split of 75:25.

Linear Regression

One of the models I am using for bitcoin forecasting is Linear Regression. It is basically used for relating or correlating variables and forecasting. The reason I chose Linear regression for bitcoin price forecasting is because using regression analysis is widely used and widely applicable. This analysis will give you a better understanding of the statistical inference overall. To begin with we are importing the dataset from bitcoin_tweets_2021_prices_final (1).csv.The dataset contains 6 features which are data,text,user_followers,user_favourites,compound,open,close and volume. These features were obtained by merging the bitcoin data and the tweets. I then perform data analysis which is exploratory and then describe the data which is seen as count,mean,std and min. I then check for null values in the feature dataset but thankfully due to data-preprocessing there were no null values present in the features. Then I move on to correlation , the reason we do correlation is to see the relation on how one or more features are basically related ,I do so by using the corr() command for seeing the correlation of all the features. After that we are going to visualize the correlated features using both corr() and heatmap() of the seaborn library. Now I will implement correlation() basically take the data and the threshold value with basically 0.81 value and display the returned features .Here our target feature is ‘Close’ .So now the correlation of the target feature with column is ‘Close’ and ‘Open’. Now we need to separate the independent and dependent features .Splitting the data into training and testing and for that we use the train_test_spilt function, checking the shapes of all the four features .Then I apply feature scaling basically scales the data in a range for applying the normalization technique to normalize data. Then I test the model using test data. Then evaluate the model that it is performing .Checking the mean absolute error for finding out the difference in the price range and the error came out to be 1375.2581805400414. Similar process was done for Elon Musk’s Tweets . Here the dataset and the tweets are really less and merging the datasets to the price forecasting did not produce much results because of the quantity of tweets . The error however is less comparatively but it's still high , one can say looking at this that the price fall from day to day is less in correlation to how often Elon Musk tweets and how people get influenced by it . But it is not significant. Mean Absolute Error = 1511.5784879089504. We also calculated the R scores for both the dataset and for the bitcoin dataset 0.9822261339260688 and for elon musk 0.9625261339268658.

4. Data and Evaluation

We downloaded the raw tweet datasets from Kaggle page (https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets) for bitcoin tweets and raw tweets of Elon musk (https://www.kaggle.com/datasets/ayhmrba/elon-musk-tweets-2010-2021) .

We initially used tweepy to extract tweets but due to the size of our dataset and scale, we decided to use Kaggle’s datasets, which were also scraped using tweepy.

Uniqueness of our data

Although we used datasets from Kaggle, we had to perform pre-processing, to make the data set ready for model train

Prepossessing steps

Elon Musk dataset

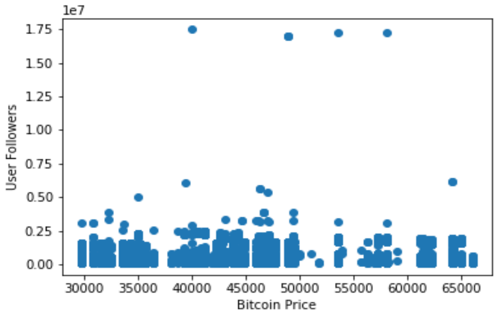
1. Concatenate 2019,2020,2021,2022 datasets for elon musk tweets
2. Create a new dataframe of only columns data,tweet for elon musk dataset
3. Remove urls, hashtags, irrelevant numbers, characters and retweets, we decided to keep the @mentions, because we observed that musk usually mentions a coins official page without any details about sentiment, but this still influences the trading day.
4. Run the cleaned dataset through a keyword list manually collected from the internet, after searching for the most popular crypto hashtags and keywords and filter out non bitcoin tweets this way
5. Merge the dataset to bitcoin historic dataset

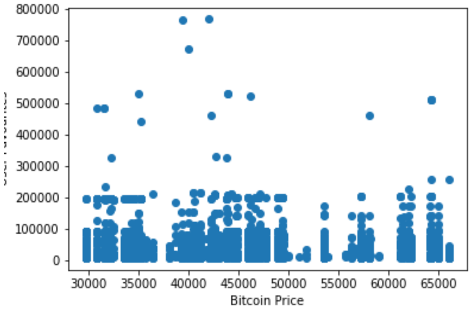
Bitcoin Dataset

1. Similar process was followed for bitcoin data, to remove URLs, hashtags etc but @mentions for this dataset were removed.
2. Merged the dataset to bitcoin historic dataset

Interesting and useful data statistics

While training the model, we include only 'user_ followers', 'user_favourites', 'Compound', 'Open', 'Close ', 'Volume' columns in our bitcoin dataset. After the first run on the test/validation set, we got an MAE of 16.02597987046904. We plotted the relation between user_followers, user_favourites vs bitcoin price, to check if these actually have an effect on the price, and found that they do not, as seen in the plots below





We believe from the above plots that, apart from some outliers, the User_followers and user_favorites do not have a massive effect on price. So on our second pass, we excluded these columns from the feature set.

For the elon musk dataset, all fields, sentiment scores, open and volume have an effect on price

Evaluation Approach

After the above revelations, we cut down our feature set to [‘Sentiment Scores’, ‘Open’,’ Volume’] for bitcoin data, and [‘Sentiment Score’, ‘Open’,’ Volume’] for the Elon Musk dataset. And the value to be predicted is the close price of bitcoin on any particular day. As mentioned before 75 percent of data was used to train and 25 percent was left for validation or testing. Although the dataset of elon musk was small, we did not find a massive difference in error between k-fold validation and a traditional 75:25 split so we stuck to the 75:25 split. To stationarize the dataset we used the Augmented Dickey Fuller test .

Evaluation Metric

To evaluate the model, we used Mean Absolute Error (MAE) as the metric along with the R2 score. We chose MAE because, after looking at a plethora of research papers on price and financial data forecasting online, many researchers are using MAE as an evaluation metric. And R2 score is a standard for any regression model. Using MAE and R2 gives us a way to measure our results from a financial standpoint and also from the general Machine learning standpoint

5. **Results and Insights**

We ran both the models on two datasets (bitcoin overall, elon musk) to predict the price of bitcoin. A total of 4249 data points for bitcoin set, and 132 data points for elon musk dataset.

Mean Absolute Errors

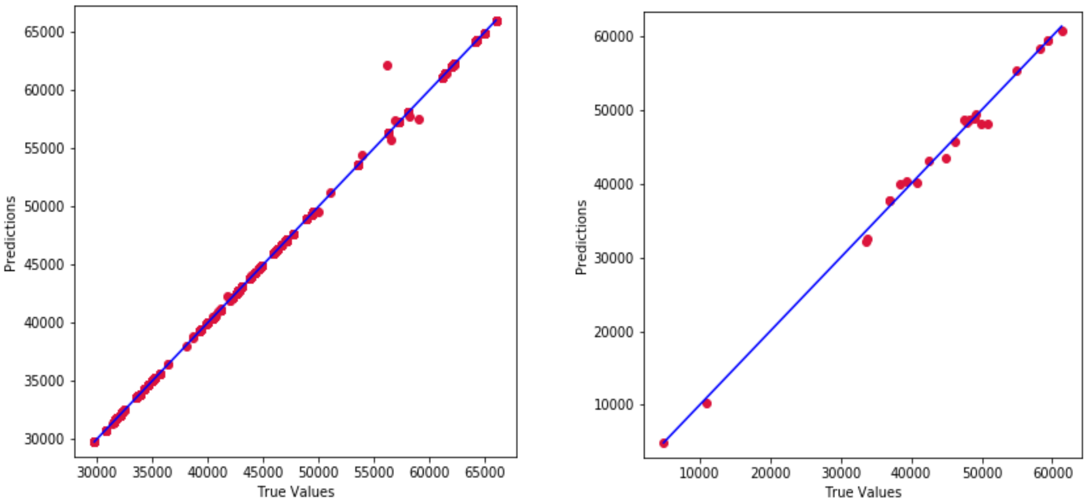
Dataset Name	Random Forest Model	Linear Regression Model
Bitcoin	10.48463847893668	1375.2581805400414
Elon Musk	741.6235132323729	1511.5784879089574

R2 Score

Dataset Name	Random Forest Model	Linear Regression Model
Bitcoin	0.9996616189541517	0.9822261339260688
Elon Musk	0.9945283139354377	0.9625261339268658

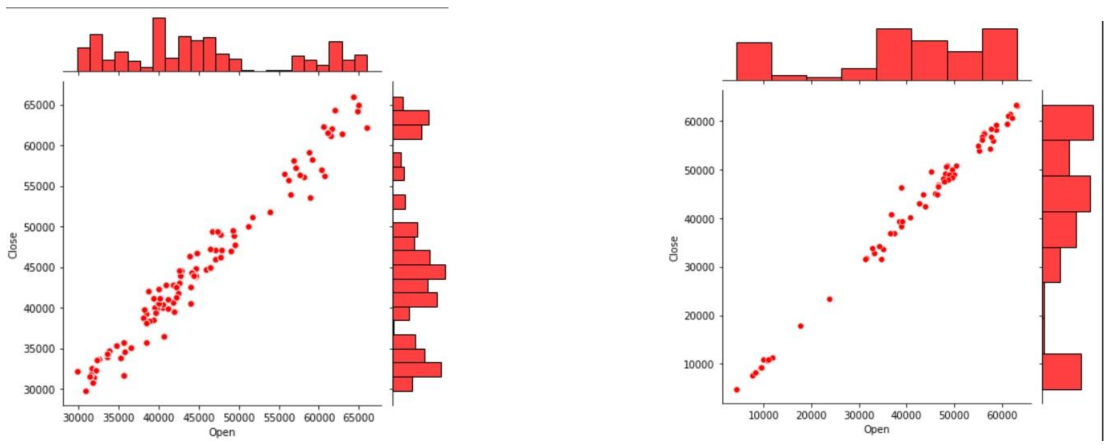
Given that our bitcoin dataset had more volume, the test results were better. Only having 132 data points for Elon musk dataset gave us an average error

Regression Plots for Random Forest Model 1.Bitcoin Data 2. Elon Musk Data



The regression line fits almost perfectly for the bitcoin dataset, which means the true values and the predicted values were similar using the test phase.

Regression Plots for Linear Regression Model 1.Bitcoin Data 2. Elon Musk Data



Insights from the results

After evaluating the performance of our models, and analyzing the MAE, we realized, the volume of tweets has more influence on the price of bitcoin more often than popularity of the user tweeting it. Elon’s tweet, albeit influential, only seems to have a short term effect on the price and the price usually corrects itself.We can see for the correlation data for the target features open and close for the bitcoin data it's more because of the number of tweets.

6. Conclusion and Future work

People’s perspective on buying bitcoin or any cryptocurrency can be completely biased , Elon Musk’s twitter platform surely speaks of this but upon using machine learning models and preprocessing the data we realized that since the amount of his tweets are few and while merging it with the bitcoin price data while using sentiment analysis the change in the price value when he tweets about Bitcoin is not that much and is not a huge influence in the market.But upon the completion of this project we came to know that Elon Musk bought Twitter for $44 Billion dollars. After this we can see how his influence now can change the crypto value.

7. Contribution Chart

Complete the following Table to clearly report the contributions that each team member made to the final project. This is required for individual projects as well.

Task/Sub-task	Student ID	Commentary on contribution
Dataset Cleaning/Preprocessing	01970377	Extracted dataset from kaggle, concatenated elon musk tweets from 2019-2022 and bitcoins tweets from the same time frame. Clean the dataset tweets, run it through a comprehensive list of keywords for bitcoin related tweets and merge the now clean tweet dataset to bitcoin historic price dataset on date.
Sentiment Analysis (polarity scores) Bitcoin tweets	01970377	Used VADER to predict sentiment scores for tweets
Sentiment Analysis (polarity scores) Elon Musk tweets	01996604	Used VADER to predict sentiment scores for tweets
Random Forest Regression Model	01970377	I initially give 10 estimator trees (same as in the paper mentioned above), with feature set as tweet polarity score,tweet_user_followers ,tweet_user_favourites, bitcoin day open price and trading volume, and the value to be predicted is the day close price.
Linear Regression Linear Regression	01996604 01996604	The dataset contains 6 features which are data,text,user_followers,user_favourites,compound,open,close and volume.These features were obtained by merging the bitcoin data and the tweets.I then

		perform data analysis which is exploratory and then describe the data.I move on to correlation , the reason we do correlation is to see the relation on how one or more features are basically related ,I do so by using the corr() command for seeing the correlation of all the features.I also obtained the bitcoin data by using yahoo from day to day analysis with the features.

8. References

[1]https://www.researchgate.net/publication/349469107_Forecasting_Bitcoin_Price_Fluctuation_by_Twitter_Sentiment_Analysis/link/6041fbd292851c077f18b 6d8/download.

[2]https://www.mlq.ai/price-prediction-with-linear-regression/

[3]https://www.mlq.ai/price-prediction-with-linear-regression/