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

COMP 5800. Social Computing

UMASS Lowell

By Nagarjuna Kocharla and Sahithi Nallani

Introduction & Motivation

- 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 SpaceX and Tesla Founder Elon Musk’s tweets using sentimental analysis. The bitcoin data we obtained was on a day to day basis from which we extracted a few features upon which we ran our algorithms .
- 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 and measure the error which is the price drop at the end of each day .The challenge we run into is stationarizing the bitcoin data which is done using ADF but the dataset is large and is quite polarizing.

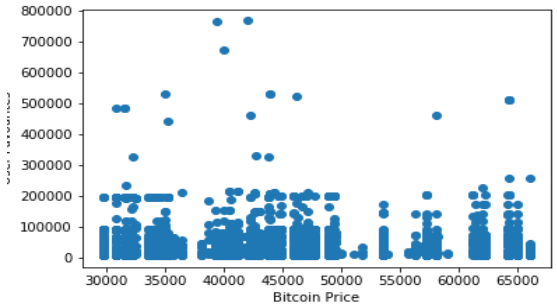
Dataset & Evaluation Method

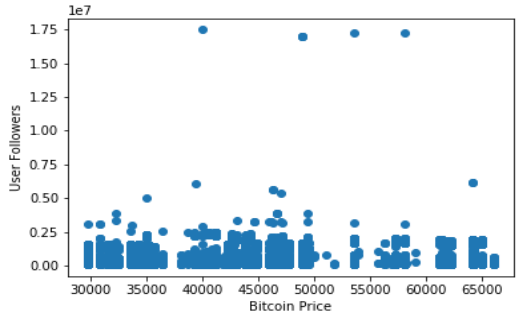
Uniqueness of Our Data

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

1. Concatenate 2019,2020,2021,2022 datasets for elon musk tweets
2. 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
3. Ran the datasets through a set of crypto keywords manually scraped.
4. Merged data with historic bitcoin prices

Useful Statistics





Method (Random Forest Regression)

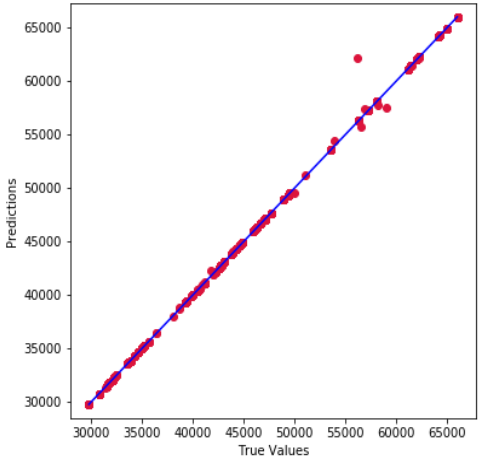
- Used Random Forest regression because it implements ensemble learning, where the prediction is done by combining ML algorithms and choosing the best performing model
- Used 10 estimator trees for Bitcoin data set, and 50-90 estimator trees for Elon Musk dataset with max_depth as 6
- Used a 75:25 split for both datasets. I tried k-folding elon musk dataset as the dataset was small, but not a massive reduction in error.

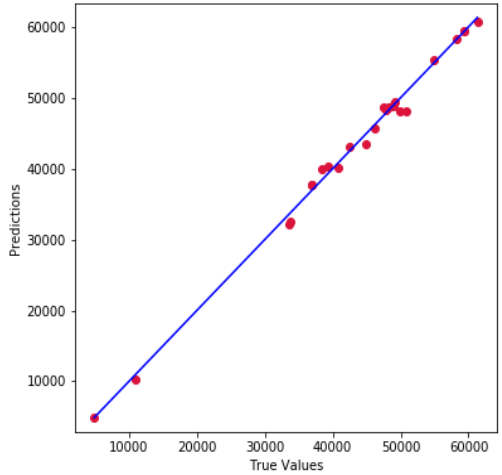
Methods(Linear Regression)

- 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.
- 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.
- 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()
- 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’
- 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.
- 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.

Results (Random Forest)

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



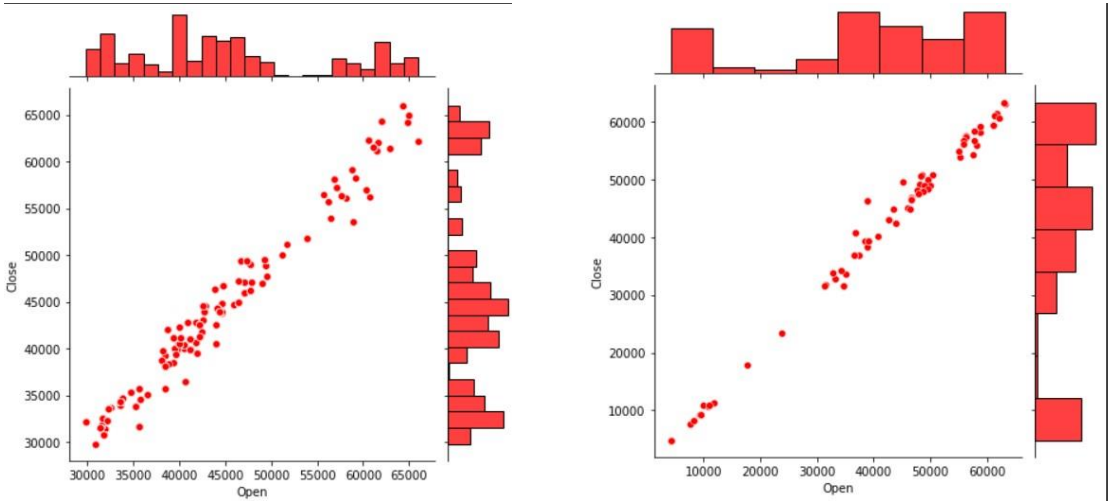


Dataset Name	Random Forest Model	Linear Regression Model
Bitcoin	10.48463847893668	1375.2581805400414
Elon Musk	741.6235132323729	1511.5784879089574

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

Results (Linear Regression)

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



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

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

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.
- After evaluating the performance of our models, and analysing the MAE, we realized, the volume of tweets has more influence on the price of bitcoin more often and popularity of the user tweet.
- 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 whereas for Elon’s data its very sparse.

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/