CRYPTO WEB

SENITMENT & PRICE ANALYSIS

UOFT SCS FinTech Boot Camp | Toronto
Amira Ali | Nadeem Hassan | Jeff Zhang
June 20th, 2022

SUMMARY

INITIAL OBJECTIVE

Analyze tweets regarding Ethereum to determine if the various sentiments had an affect on Ethereum prices.

QUESTIONS TO ANSWER

- Does twitter activity affect Ethereum prices?
- If so, how strong is the correlation?
- If not, what is highly correlated with prices?
- Are we able to predict future prices by applying machine learning models?

QUESTIONS

Does twitter activity affect Ethereum prices?

- Need twitter data, long enough to determine trends
- Need minute-minute price data on Ethereum

If so, how strong is the correlation?

- Sentiment scores?

If not, what is highly correlated with prices?

- Analyze other features in the data set , which ever has a higher correlation will be included in the model

Are we able to predict future prices by applying machine learning modes?

- Linear regression, RandomForest, XGBoost

TECHNOLOGIES USED

CoinGecko API

Alpaca API

Kaggle

Jupyter Notebook

Python

Google Finance

Google Colab

DATA COLLECTION

TWITTER

- Instead of using twitter's API, mainly due to time constraints, we used pre-existing data from Kaggle. Data ranging from 2013-2021.
- Bulk CSV file that was cleaned, dropped all null values, removed punctuation to prepare the data for analysis

created_at	tweet	cleaned_tweets
2013-01-30 17:07:29	Attention ye who have not children: Chooseth w...	attention ye who have not children chooseth wi...
2013-09-05 19:35:25	Unite knows something we dont... something bet...	unite knows something we dont something better...
2013-10-02 23:17:18	ok i have to ask... why do people favorite twe...	ok i have to ask why do people favorite tweets...
2013-10-13 16:37:17	M(eth)iley C(rack)yrus	methiley crackyrus
2013-10-23 20:24:45	get a sled and ride it out homie	get a sled and ride it out homie

DATA COLLECTION cont.

ETHEREUM DATA

- Price Data – using Alpaca API. Minute interval data
- S&P 500 & Nasdaq historical data retrieved from Google Finance
- Ethereum market cap data retrieved from Coin Gecko API

```
# Set tickers
ticker = ["ETHUSD"]

# Set timeframe to '1Minute'
timeframe = "1Min"

# Set start and end datetimes.
start_date = pd.Timestamp("2017-01-01", tz="America/New_York").isoformat()
end_date = pd.Timestamp("2017-06-30", tz="America/New_York").isoformat()

# Get 1 year's
df_ticker = alpaca.get_crypto_bars(
    ticker,
    timeframe,
    start=start_date,
    end=end_date,
).df

# Display sample data
df_ticker.head(10)
```

DATA COLLECTION Limitations

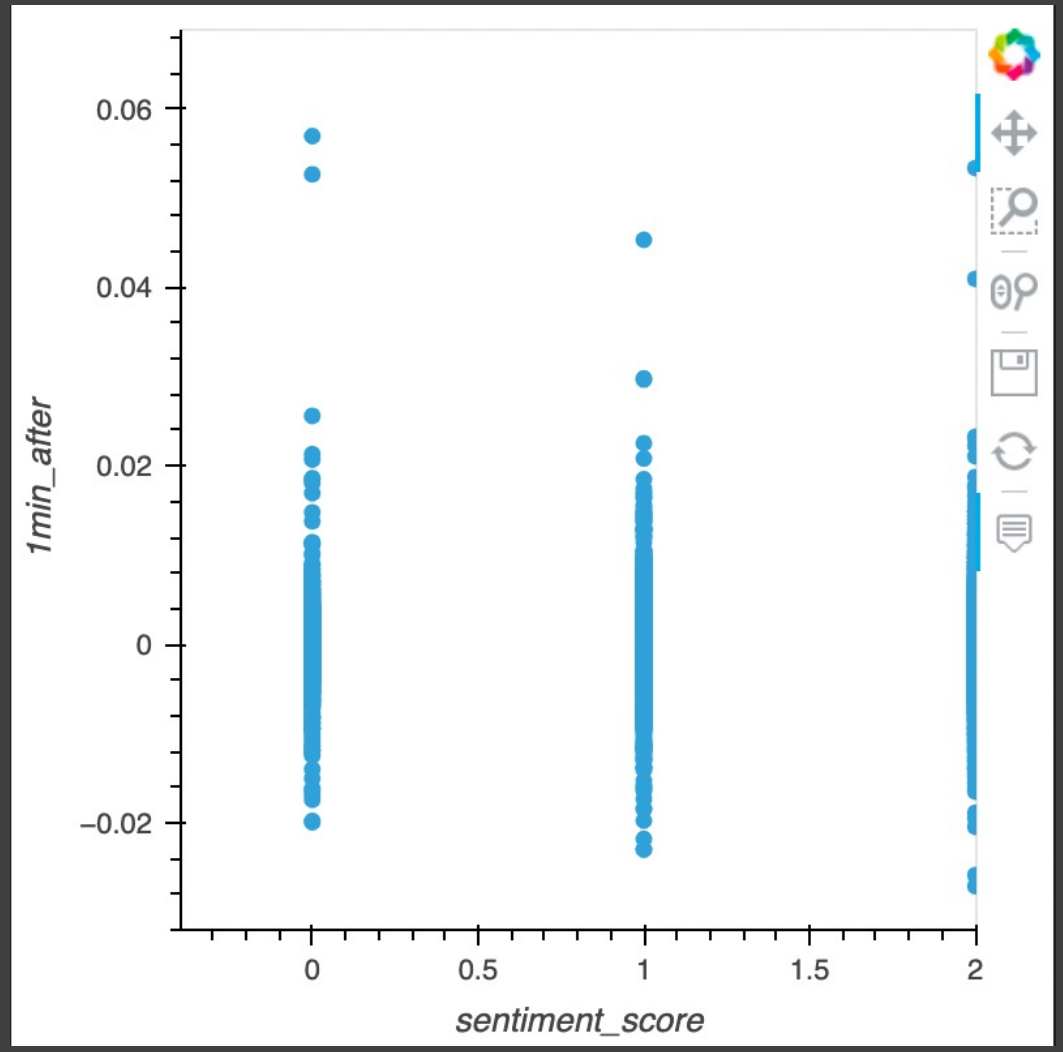
- **Twitter**
  - Initially wanted to use twitter's API, confusion surrounding essential vs elevated plans
- **Alpaca API limits**
  - Ethereum price data had to be run using semi-annual periods due to the large amounts of data being called
- **Different time frames**
  - Kaggle twitter data date range : Jan 30th , 2013 - Feb 10th, 2021
  - Ethereum price data date range May 17th, 2016 – onwards
  - Forced to drop 3 years of data

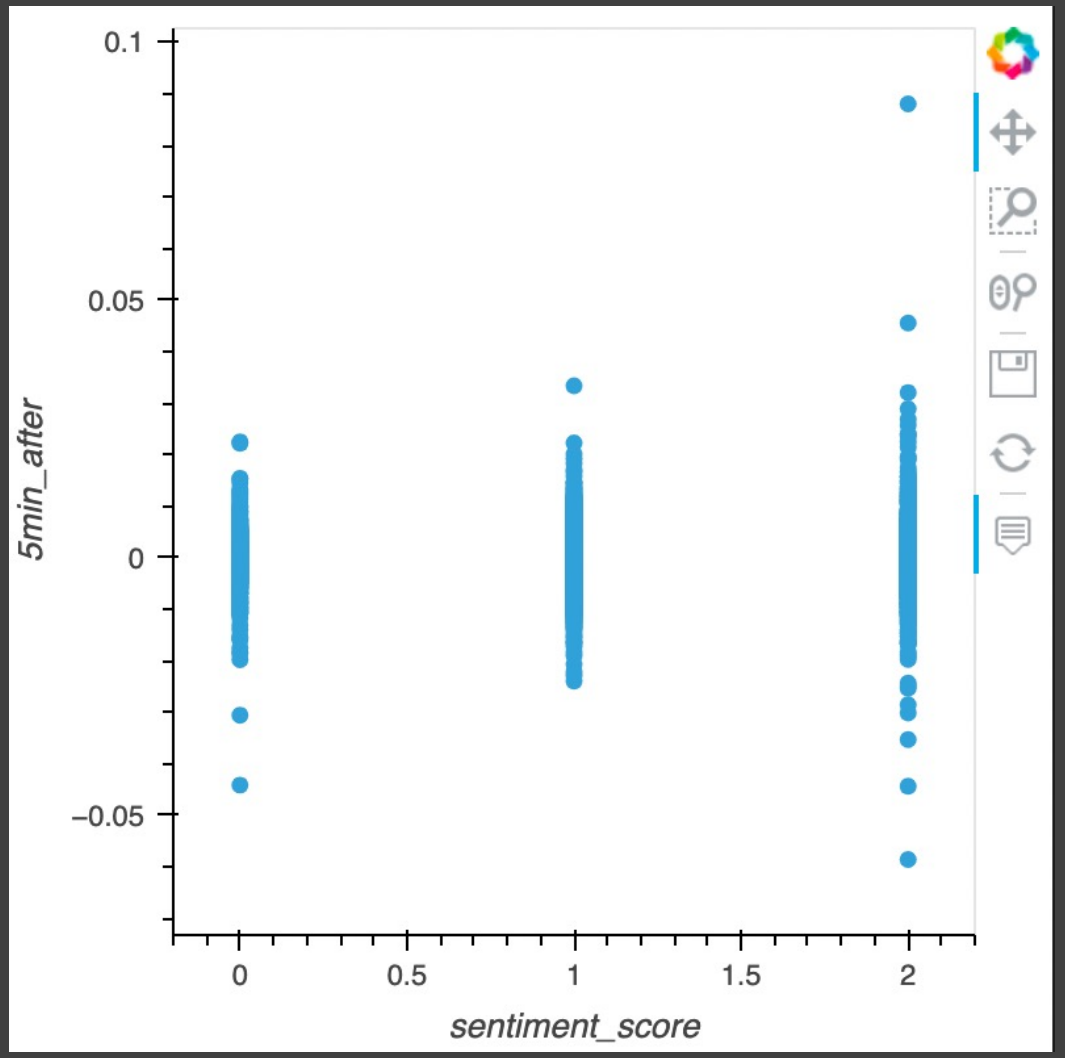
DATA ANALYSIS I

created_at	tweet	cleaned_tweets	subjectivity	polarity	sentiment	sentiment_score
2013-01-30 17:07:29	Attention ye who have not children: Chooseth w...	attention ye who have not children chooseth wi...	0.900000	0.7000	Positive	2
2013-09-05 19:35:25	Unite knows something we dont... something bet...	unite knows something we dont something better...	0.416667	0.3750	Positive	2
2013-10-02 23:17:18	ok i have to ask... why do people favorite twe...	ok i have to ask why do people favorite tweets...	0.750000	0.1875	Positive	2
2013-10-13 16:37:17	M(eth)iley C(rack)yrus	methiley crackyrus	0.000000	0.0000	Neutral	1
2013-10-23 20:24:45	get a sled and ride it out homie	get a sled and ride it out homie	0.000000	0.0000	Neutral	1

created_at	tweet	cleaned_tweets	subjectivity	polarity	sentiment	sentiment_score
2021-02-10 15:35:45	Someone has a buy order right now for 800K+ $D...	someone has a buy order right now for 800k drg...	0.535714	0.285714	Positive	2
2021-02-10 15:44:25	LAST public service announcement I'll make bef...	last public service announcement ill make befo...	0.505556	-0.187500	Negative	0
2021-02-10 15:44:25	It is ~1/100th the value of $eth. I won't ment...	it is 1100th the value of eth i wont mention i...	0.533333	0.000000	Neutral	1
2021-02-10 15:46:05	Complaints about $AVAX, a 6 month(!!) old proj...	complaints about avax a 6 month old project ha...	0.531250	0.062500	Positive	2
2021-02-10 15:48:27	One of the reasons that $ORN is a magnificent ...	one of the reasons that orn is a magnificent p...	0.758333	0.616667	Positive	2

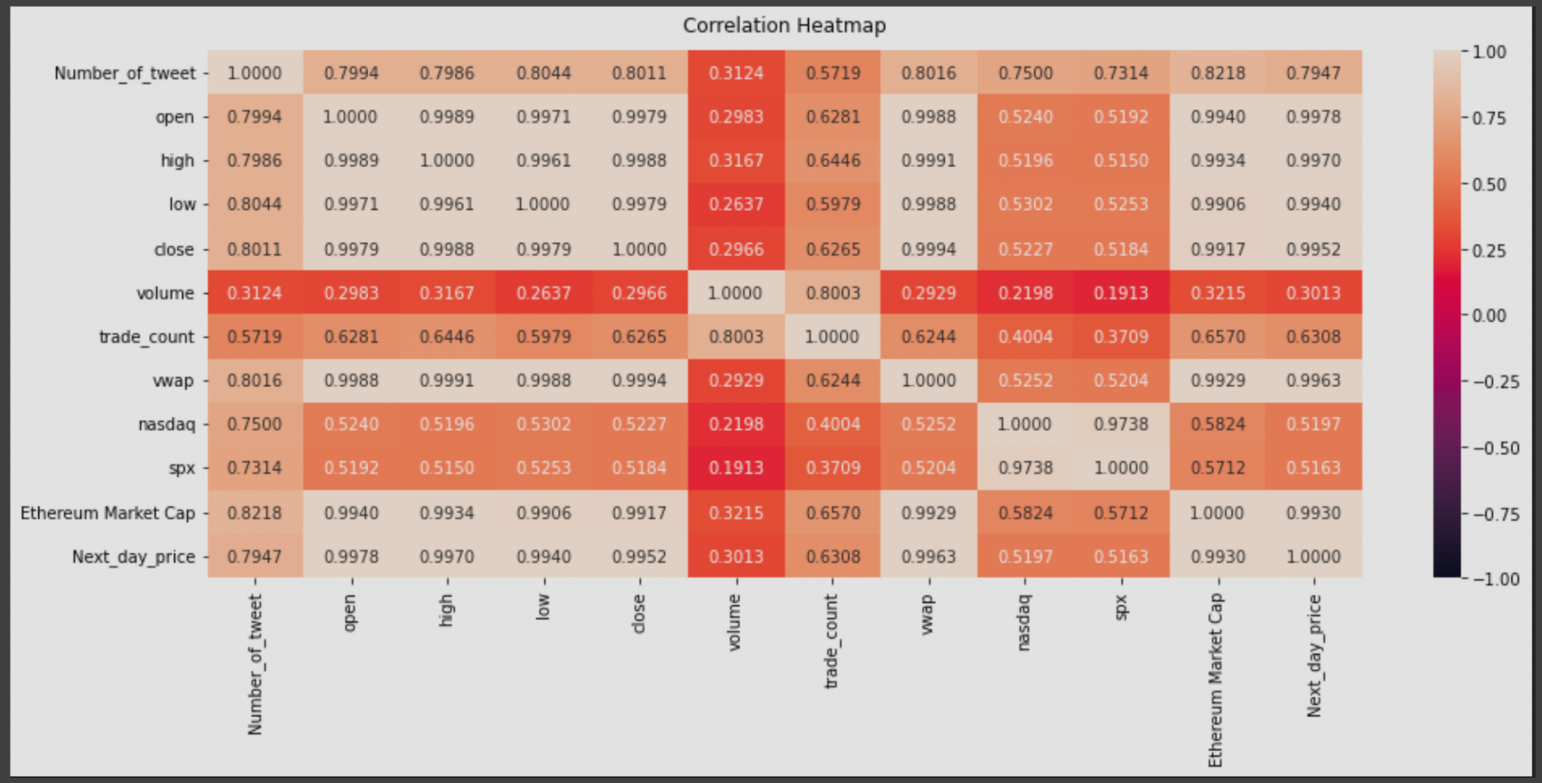
DATA ANALYSIS II





SECOND OBJECTIVE

DATA ANALYSIS III

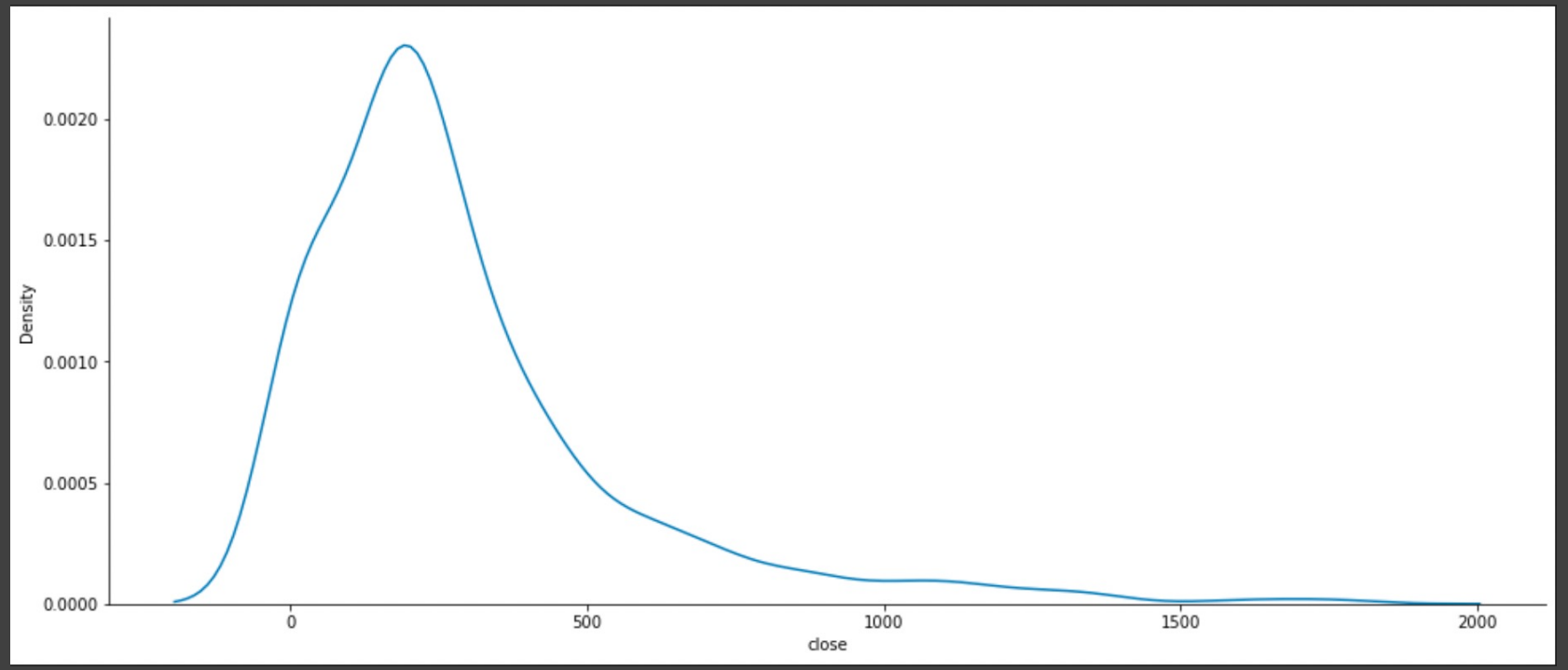


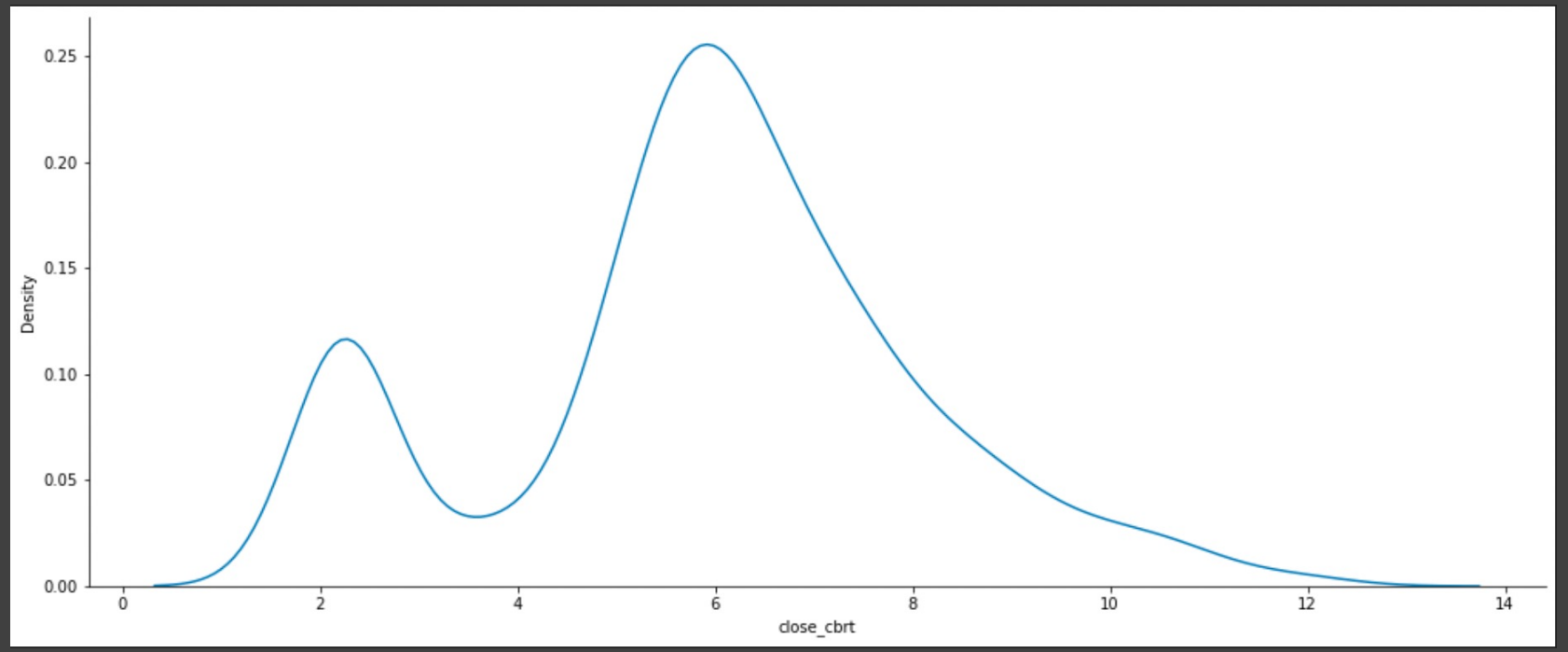

```
close                  1.000000
vwap                   0.999428
high                   0.998839
open                   0.997869
low                    0.997855
Next_day_price         0.995150
Ethereum Market Cap    0.991722
Number_of_tweet        0.801051
trade_count            0.626473
nasdaq                 0.522720
spx                    0.518441
volume                 0.296583
Name: close, dtype: float64
```

```python
X = eth2_df[['Date', 'vwap_cbrt', 'Ethereum Market Cap_cbrt', 'Number_of_tweet_cbrt', 'trade_count_cbrt', 'close_cbrt',
             'lag_1', 'lag_2', 'lag_3', 'lag_4', 'lag_5', 'lag_6', 'lag_7',
             'rolling_3_mean', 'rolling_4_mean', 'rolling_5_mean', 'rolling_6_mean', 'rolling_7_mean',
             'expanding_2_mean', 'expanding_3_mean', 'expanding_4_mean']
            ]

y = eth2_df[['Date', 'tomorrow_close_cbrt']]
```

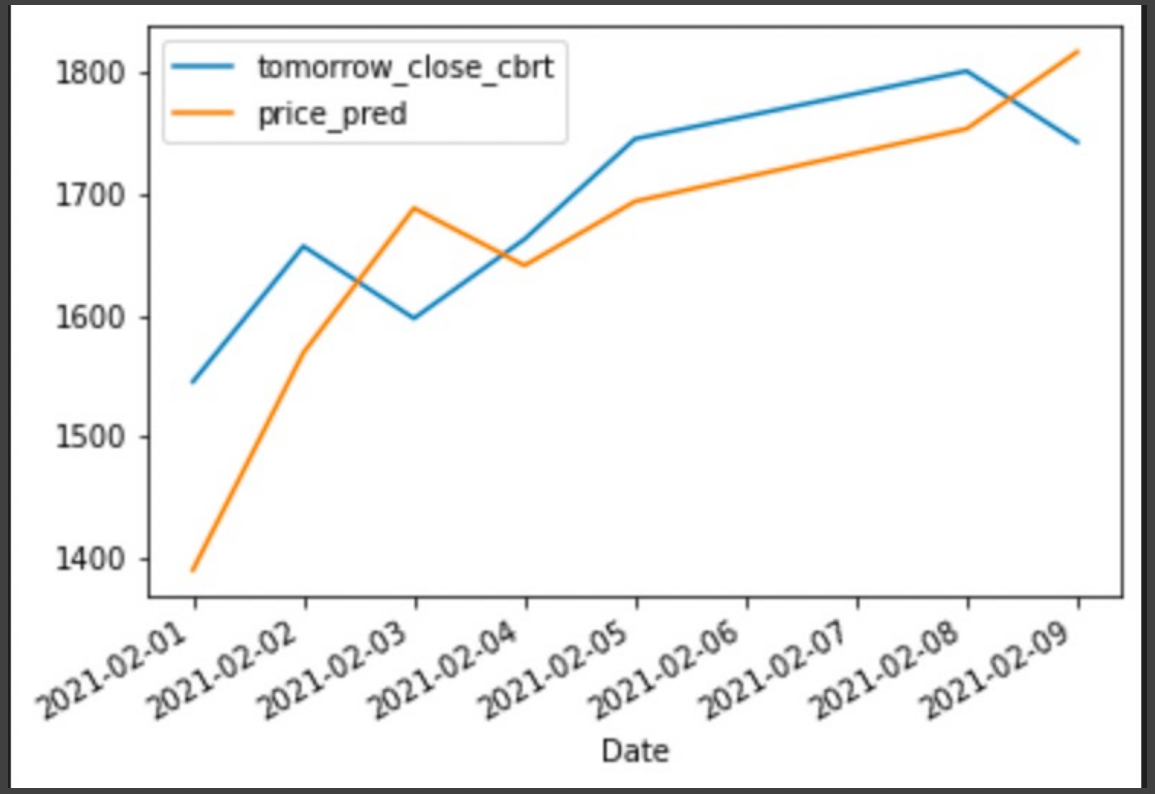
Python


Density
0.0020
0.0015
0.0010
0.0005
0.0000
0
500
1000
1500
2000
close


Density
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
14
close_cbrt

DATA ANALYSIS IV

Linear Regression



`MAPE` - Mean absolute percentage error. It measures accuracy of a forecast model as a percentage. The higher the number, the less accurate the results are.

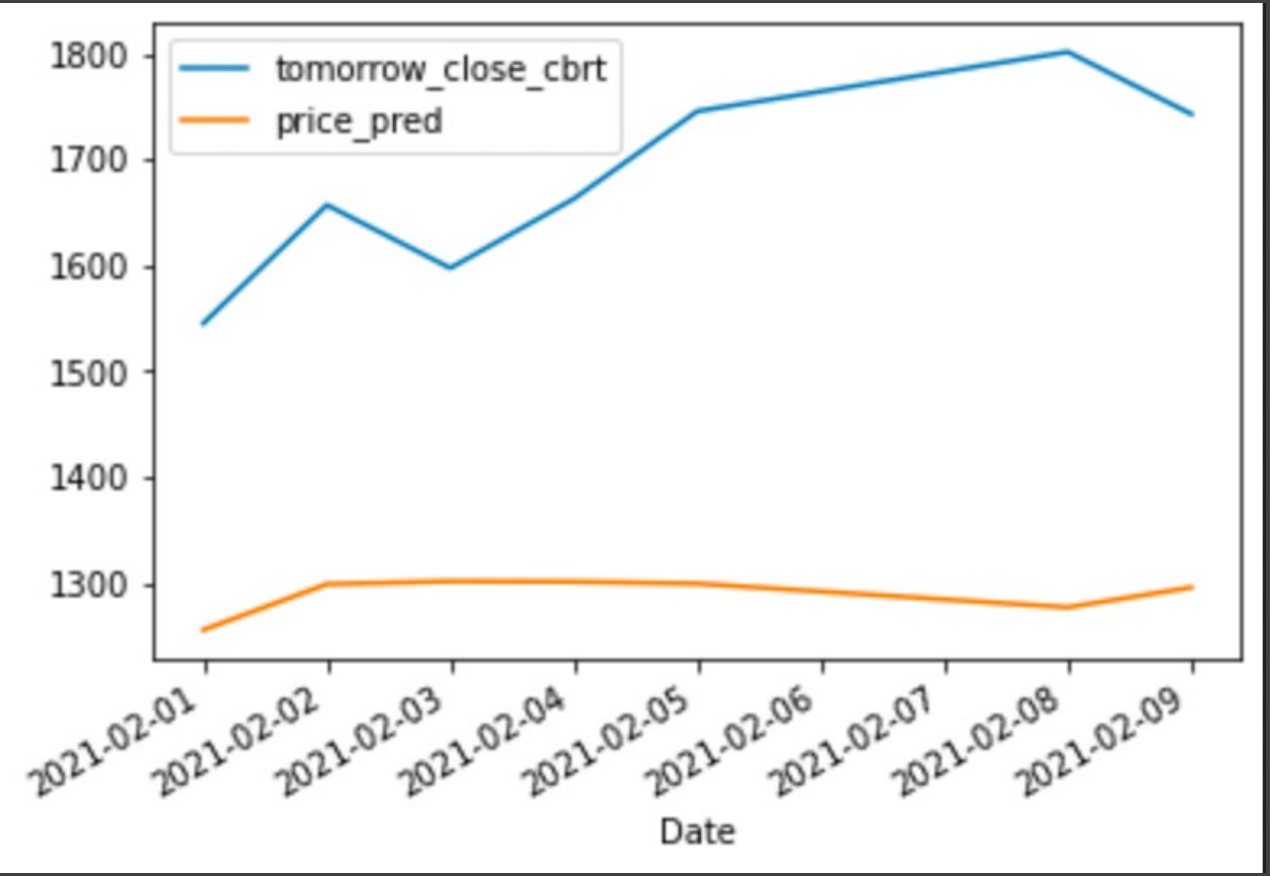
`MAE` - Mean absolute error. This score tells us the mean difference between the actual and predicted values. The lower the better.

`RMSE` - Root mean square deviation. This is the standard deviation of the prediction errors. Lower RMSE indicates a better model.

MAPE	1.5510641286721347
MAE	0.18294694577561568
RMSE	0.21073956591219212

DATA ANALYSIS IV

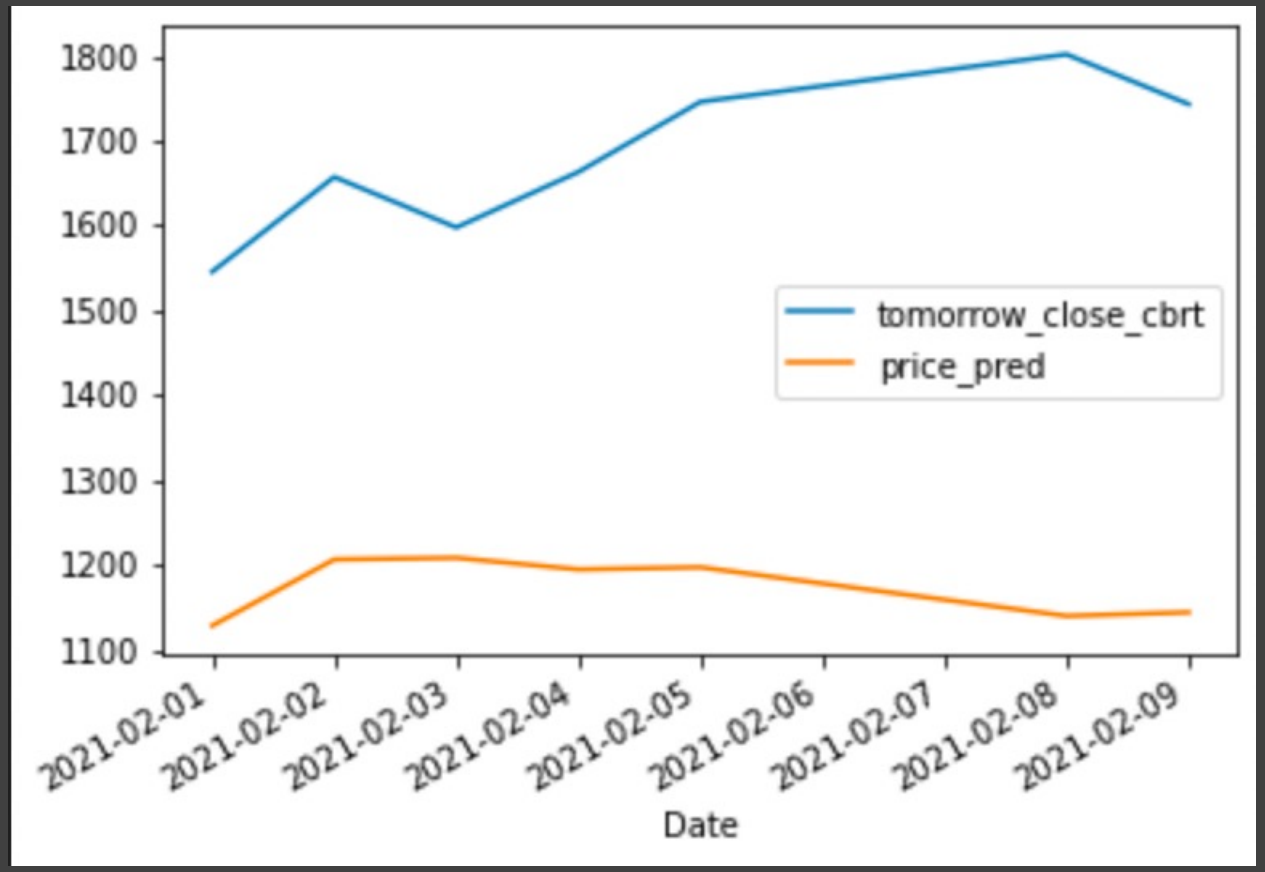
Random Forest



MAPE	8.356436911356955
MAE	0.9959573732637854
RMSE	1.0137439458189803

DATA ANALYSIS IV

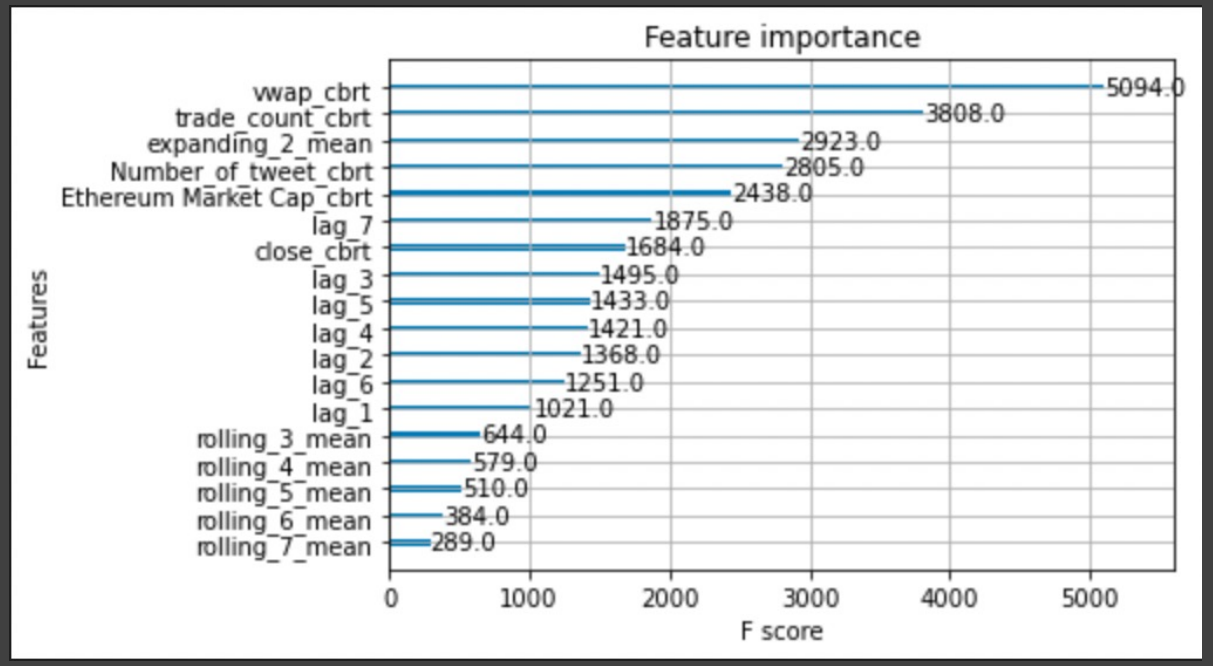
XGBOOST

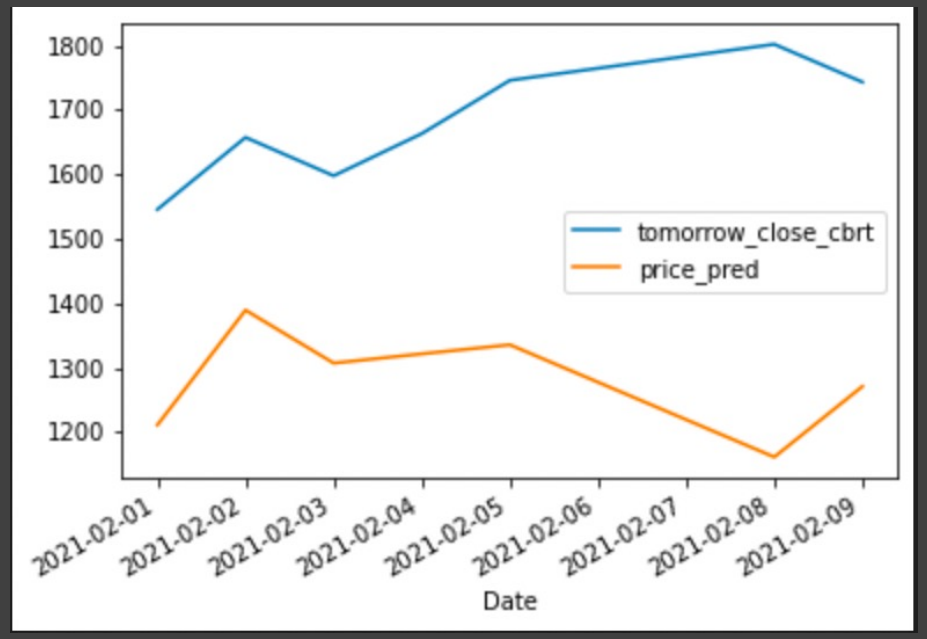


MAPE	11.204562037825198
MAE	1.3342855037791528
RMSE	1.3538536562324184

DATA ANALYSIS IV

XGBOOST – 2






```
xgb_mape = mape(y_test_subset, y_pred)
xgb_mae = mae(y_test_subset, y_pred)
xgb_rmse = rmse(y_test_subset, y_pred)
print(xgb_mape)
print(xgb_mae)
print(xgb_rmse)
```

```
8.510349190022433
1.0142405775444987
1.05985627245193
```

RESULTS/ANALYSIS

- We agreed the Linear Regression model was the most accurate in predicting Ethereum prices. Despite our efforts to re-train the XG Boost model, the scores were still lower than we had hoped.
- XG boost better for short term analysis
- Linear Regression has a lag in the prediction

TO CONSIDER NEXT TIME

- Incorporate more data to feed in the model
- Determine how we can predict future prices
- Involve more currencies in the analysis

QUESTIONS?