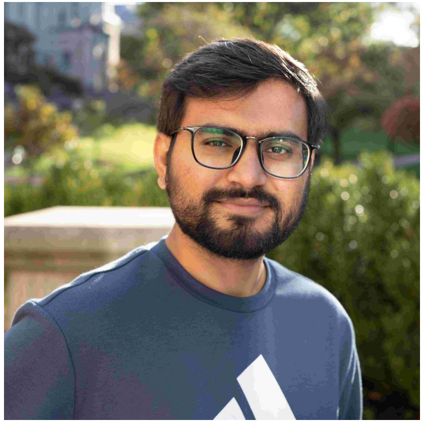
BitCoin Tweets Data Mining

Group Members:



Ashutosh Hathidara
(ashuhath@iu.edu)

Gaurav Atavale
(gatavale@iu.edu)

Suyash Chaudhary
(suschaud@iu.edu)

Exploratory Data Analysis (EDA)

	user	timestamp	replies	likes	retweets	text
0	KamdemAbdiel	2019-05-27 11:49:14+00	0.0	0.0	0.0	È appena uscito un nuovo video! LES CRYPTOMONN...
1	bitcointe	2019-05-27 11:49:18+00	0.0	0.0	0.0	Cardano: Digitize Currencies; EOS https://t.co...
2	3eyedbran	2019-05-27 11:49:06+00	0.0	2.0	1.0	Another Test tweet that wasn't caught in the s...
3	DetroitCrypto	2019-05-27 11:49:22+00	0.0	0.0	0.0	Current Crypto Prices! \n\nBTC: $8721.99 USD\n...
4	mmursaleen72	2019-05-27 11:49:23+00	0.0	0.0	0.0	Spiv (Nosar Baz): BITCOIN Is An Asset & NO...

Initial total Number of tweets: ~20 million
Initial size of dataset: ~4 GB

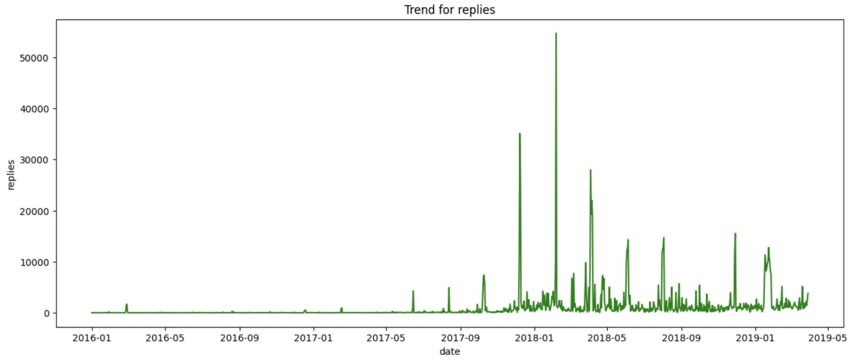
Exploratory Data Analysis (EDA) - Language Detect and Data Filtering

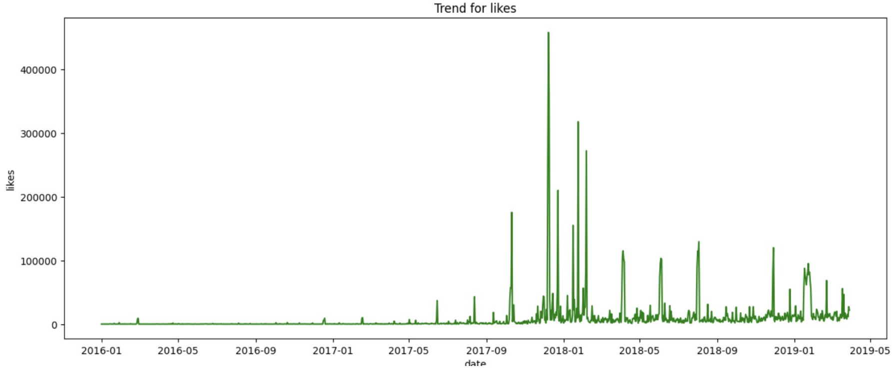
Detecting language

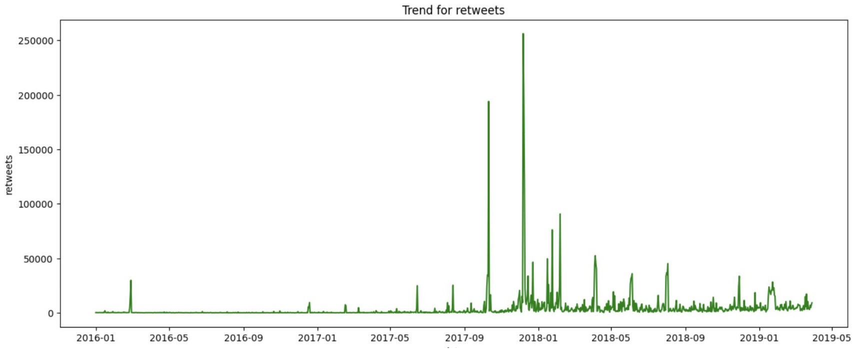
	user	timestamp	replies	likes	retweets	text	tweet_lang
0	KamdemAbdiel	2019-05-27 11:49:14+00	0.0	0.0	0.0	È appena uscito un nuovo video! LES CRYPTOMONN...	it
1	bitcointe	2019-05-27 11:49:18+00	0.0	0.0	0.0	Cardano: Digitize Currencies; EOS https://t.co...	en
2	3eyedbran	2019-05-27 11:49:06+00	0.0	2.0	1.0	Another Test tweet that wasn't caught in the s...	en
3	DetroitCrypto	2019-05-27 11:49:22+00	0.0	0.0	0.0	Current Crypto Prices! \n\nBTC: $8721.99 USD\n...	en
4	mmursaleen72	2019-05-27 11:49:23+00	0.0	0.0	0.0	Spiv (Nosar Baz): BITCOIN Is An Asset & NO...	en

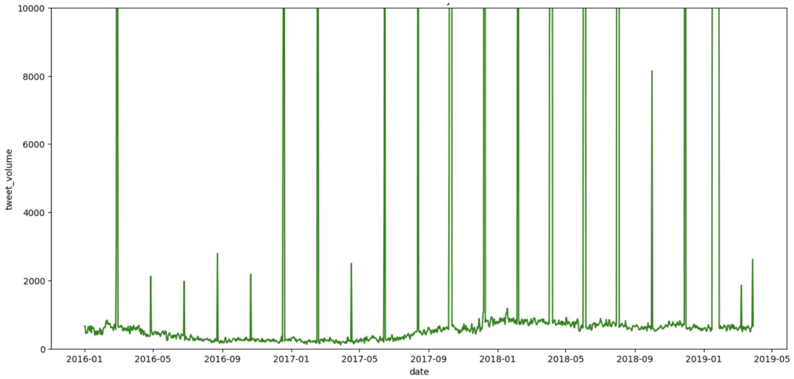
- Language detection was done using python package langdetect (https://pypi.org/project/langdetect/)
- For flexibility reasons, we filtered out only English (‘en’) language tweets from the data for further analysis.
- We also filtered data between 2016-01-01 and 2019-03-29 for further analysis as mentioned on the Kaggle. (There were outlier tweets outside this range)
- The final number of tweets for analysis = ~2.3M

Exploratory Data Analysis (EDA) - Day-wise trend for Tweets/Likes/Retweets/Replies

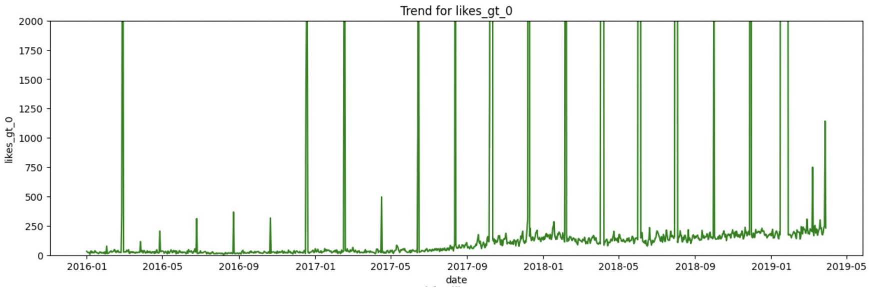


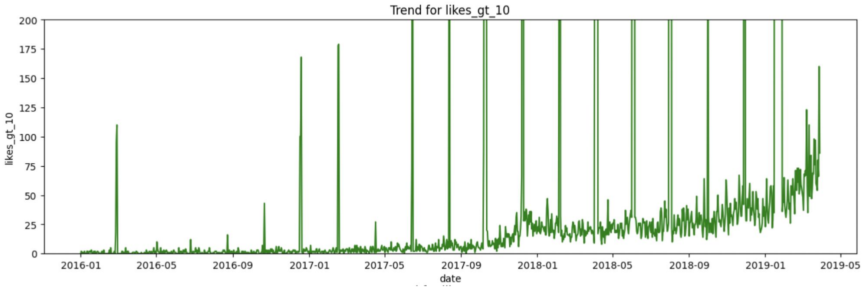


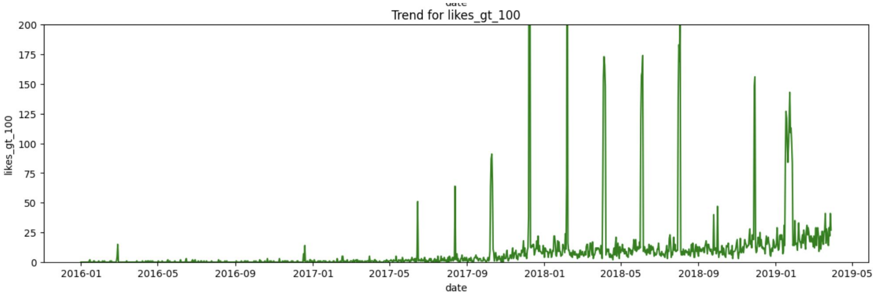


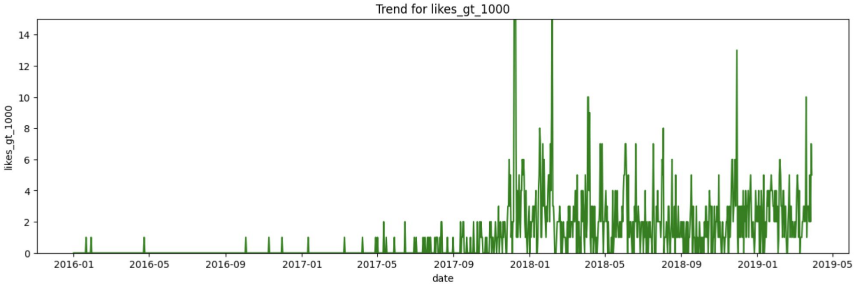


Exploratory Data Analysis (EDA) - Day-wise trend for number of tweets with specific number of likes (>0, >10, >100, >1000)

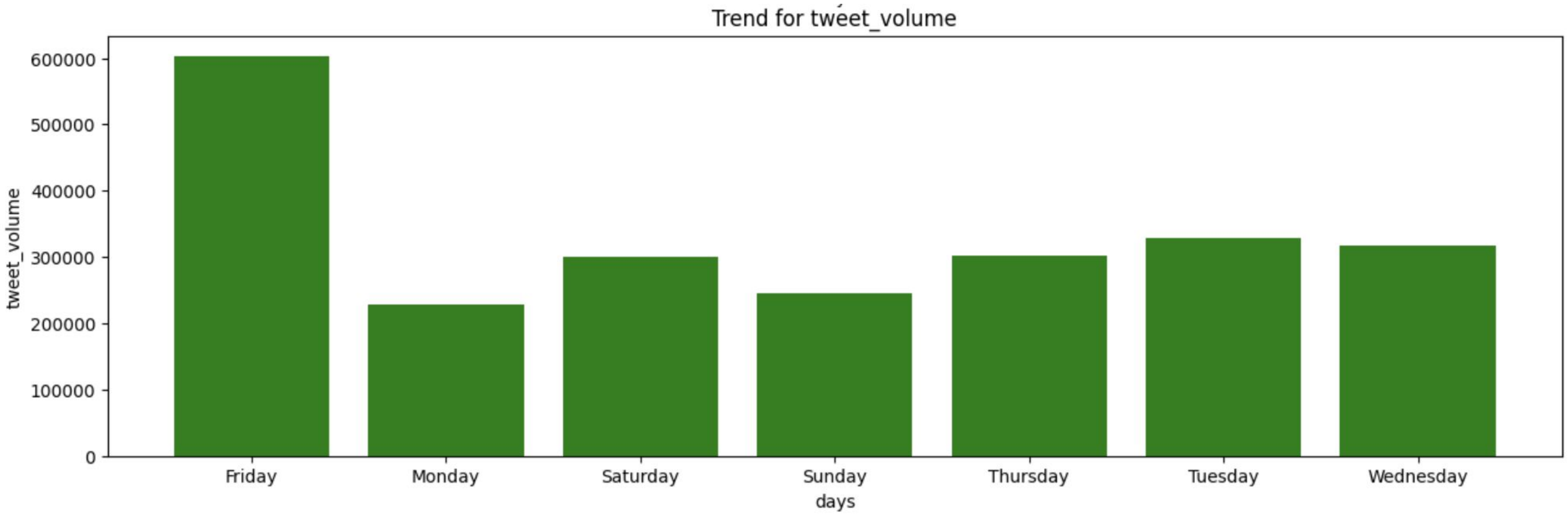




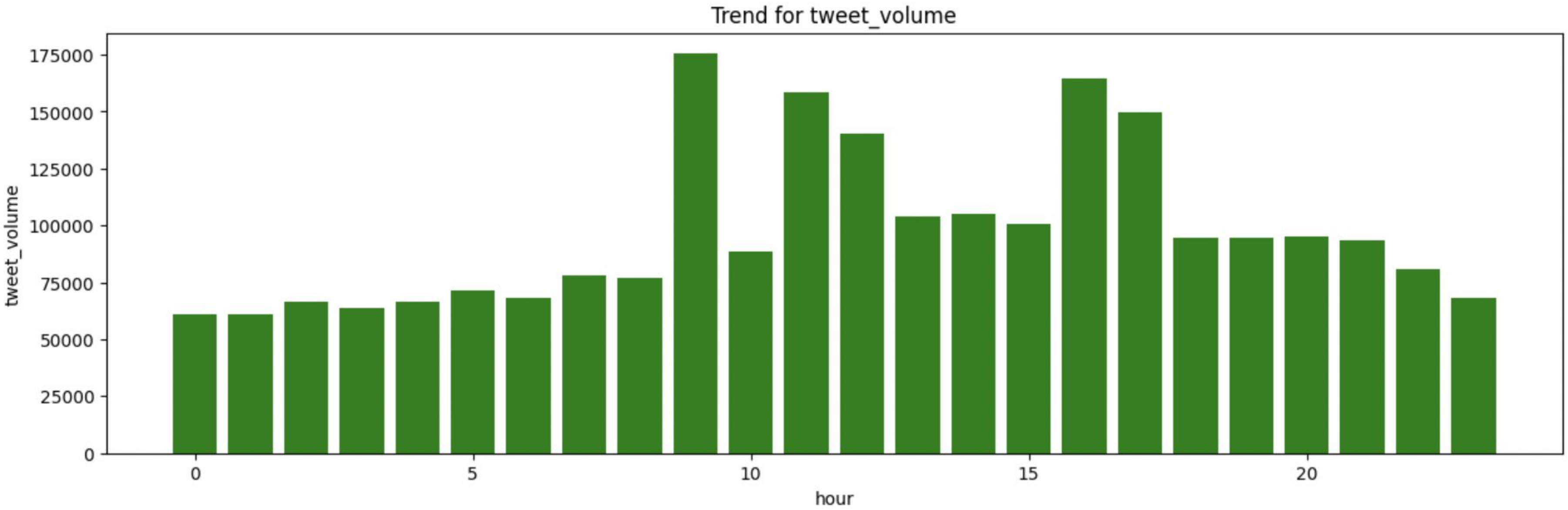




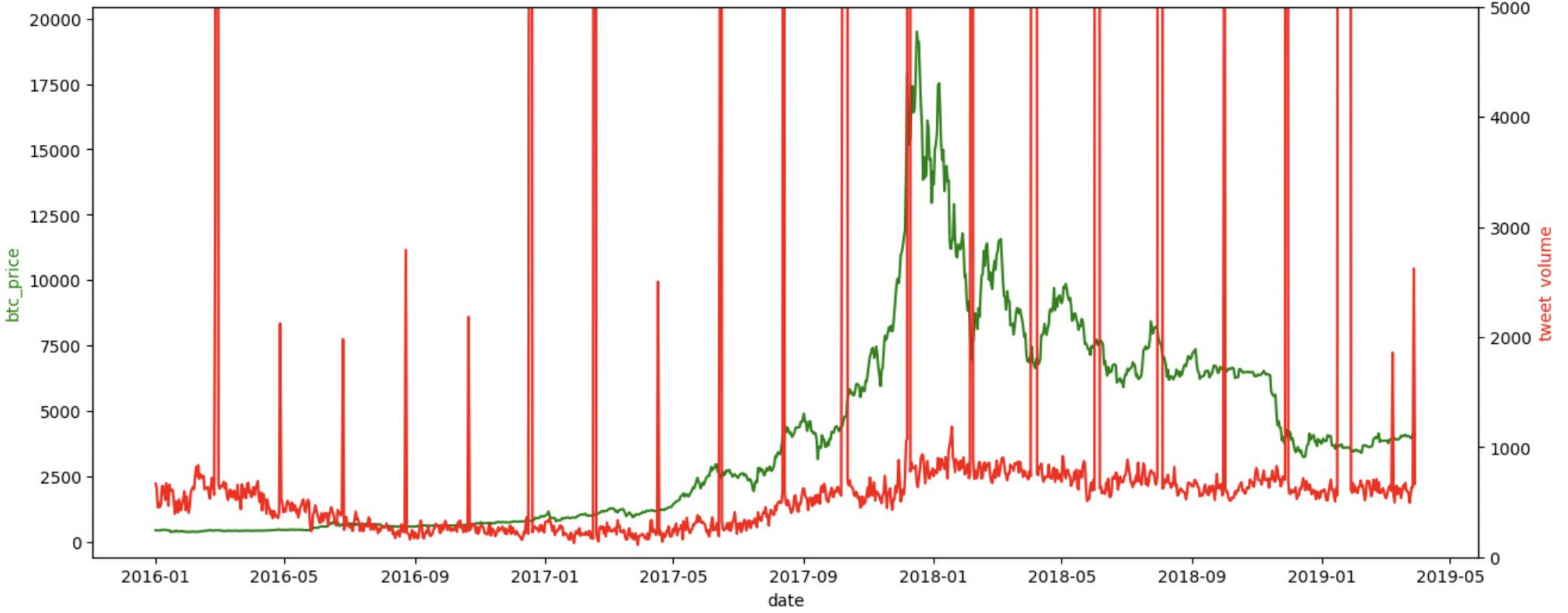
Exploratory Data Analysis (EDA) - Day of the week trend



Exploratory Data Analysis (EDA) - Hour of the day



Exploratory Data Analysis (EDA) - Comparison of tweet trend to BitCoin price

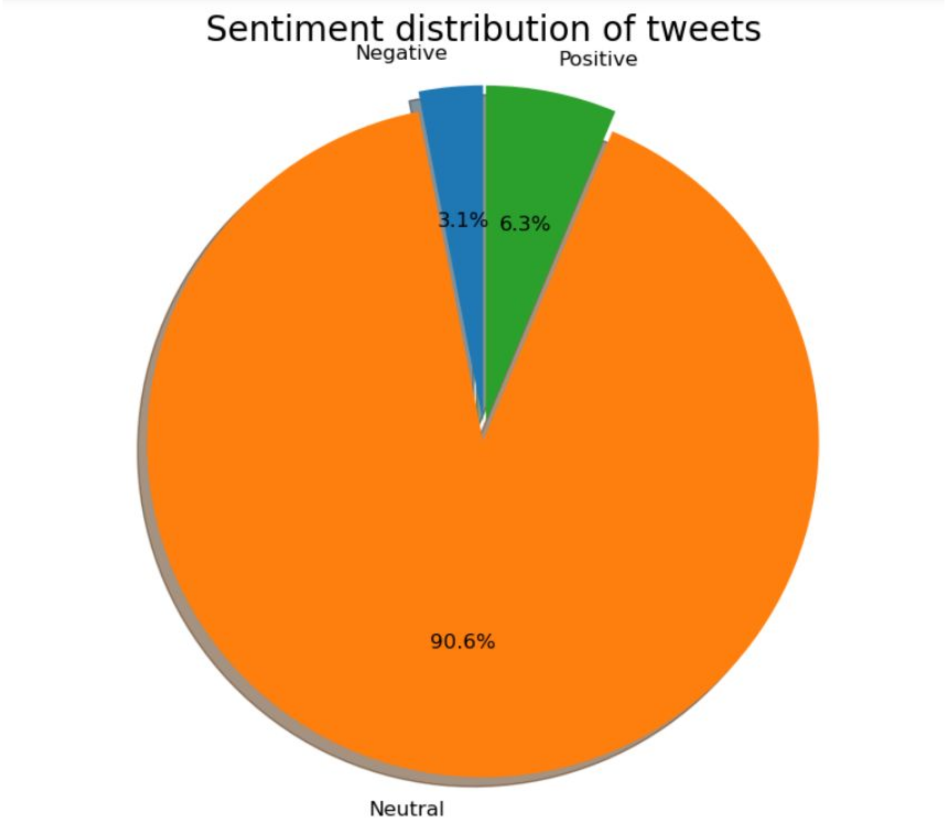


Final data after EDA for prediction purpose

	date	btc_price	tweet_volume	likes_gt_0	likes_gt_10	likes_gt_100	likes_gt_1000	retweets_gt_0	retweets_gt_1000	pos_sent	...
0	2016-01-01	433.437988	666	31	0	0	0	212	0	157	...
1	2016-01-02	430.010986	625	25	2	0	0	216	0	159	...
2	2016-01-03	433.091003	451	20	0	0	0	162	0	149	...
3	2016-01-04	431.959991	493	22	1	0	0	143	0	151	...
4	2016-01-05	429.105011	455	22	0	0	0	115	0	146	...

```
Index(['date', 'btc_price', 'tweet_volume', 'likes_gt_0', 'likes_gt_10',
       'likes_gt_100', 'likes_gt_1000', 'retweets_gt_0', 'retweets_gt_1000',
       'pos_sent', 'neg_sent', 'neu_sent', 'user', 'hr_0_6', 'hr_6_12',
       'hr_12_18', 'hr_18_24', 'Friday', 'Monday', 'Saturday', 'Sunday',
       'Thursday', 'Tuesday', 'Wednesday', 'btc_cur_price'],
      dtype='object')
```

Sentiment Analysis - Tweet sentiment distribution



Correlation between Bitcoin price and tweet sentiment

Positive Sentiment	0.13
Negative Sentiment	0.15
Neutral Sentiment	0.07

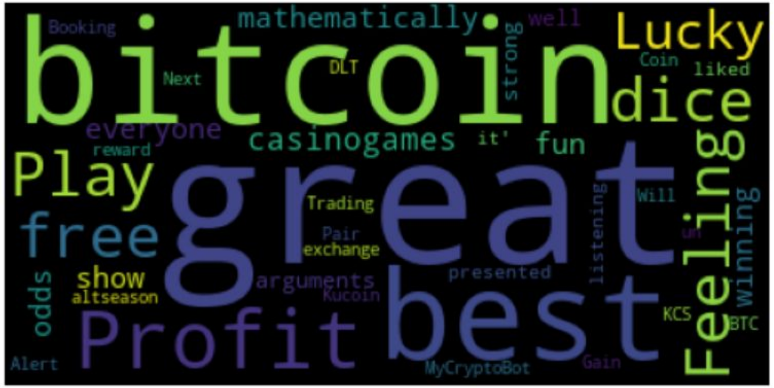
Correlation b/w Bitcoin price fluctuation (Today-Yesterday) and tweet sentiment

Positive Sentiment	-0.05
Negative Sentiment	-0.06
Neutral Sentiment	-0.04

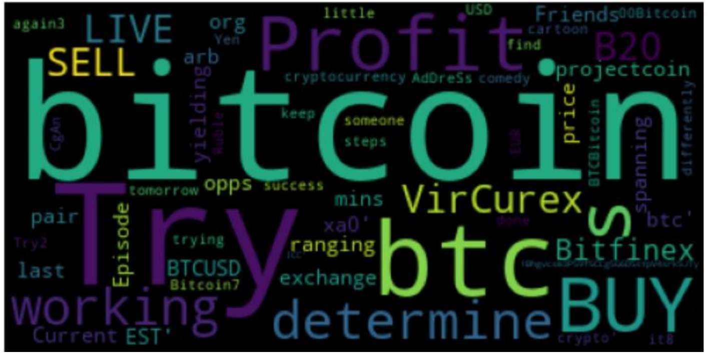
Sentiment Analysis: Word cloud of extracted tweet sentiments

Negative Tweets

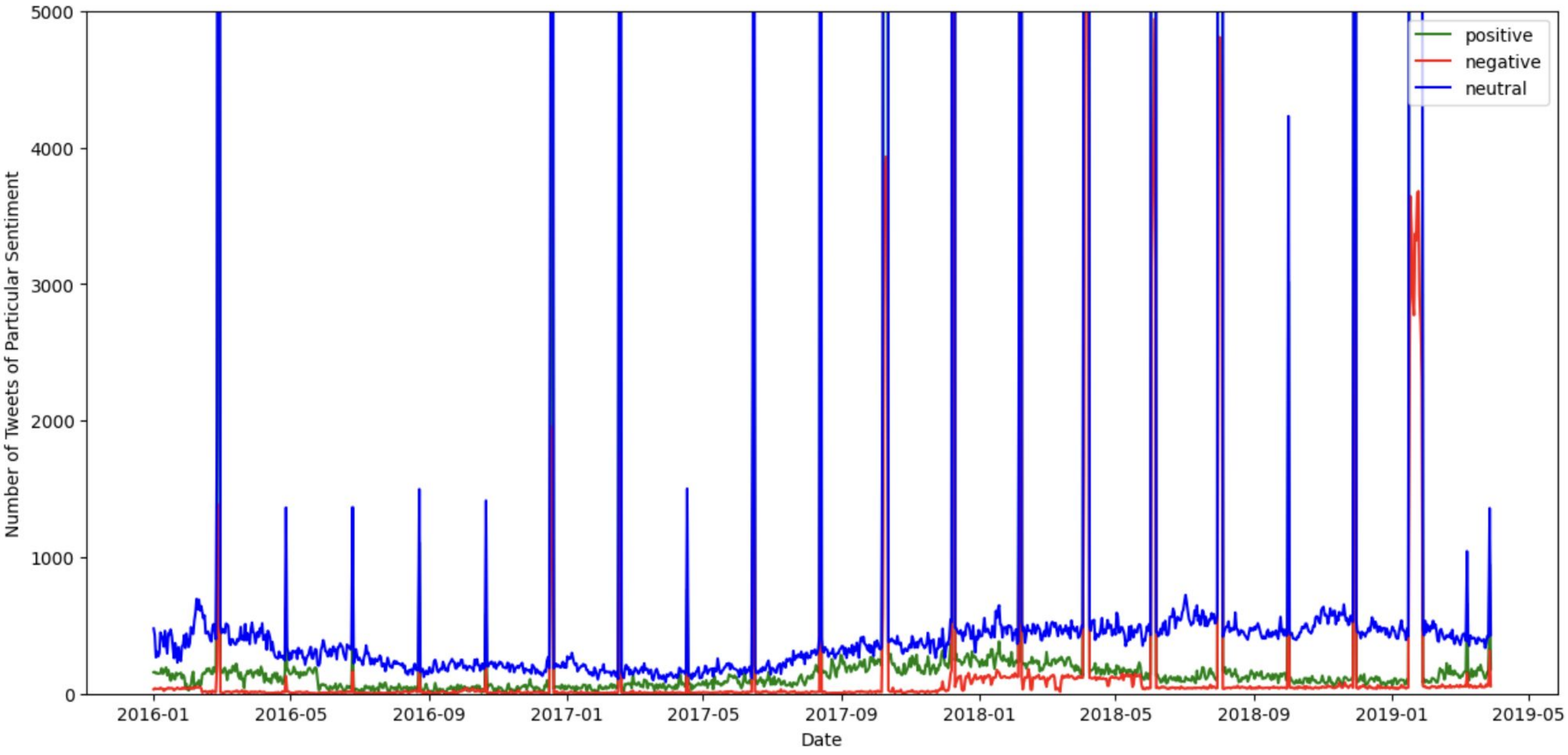
Positive Tweets



Neutral Tweets

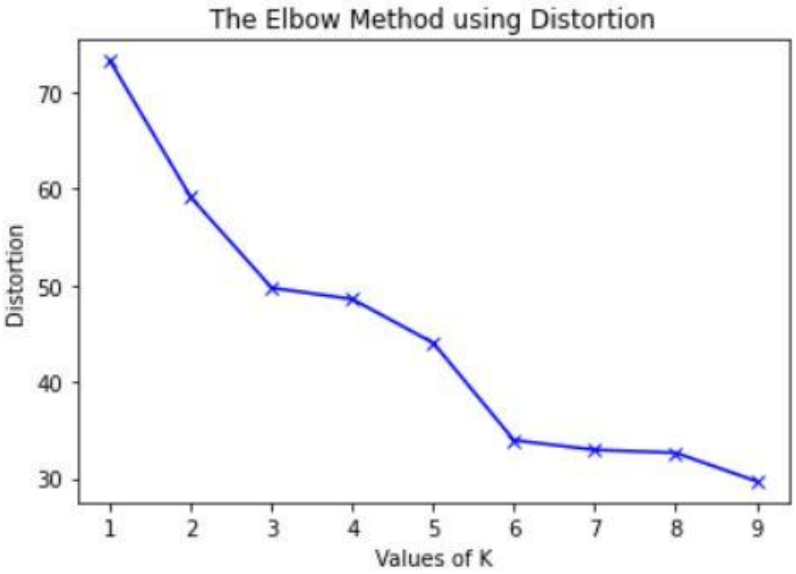


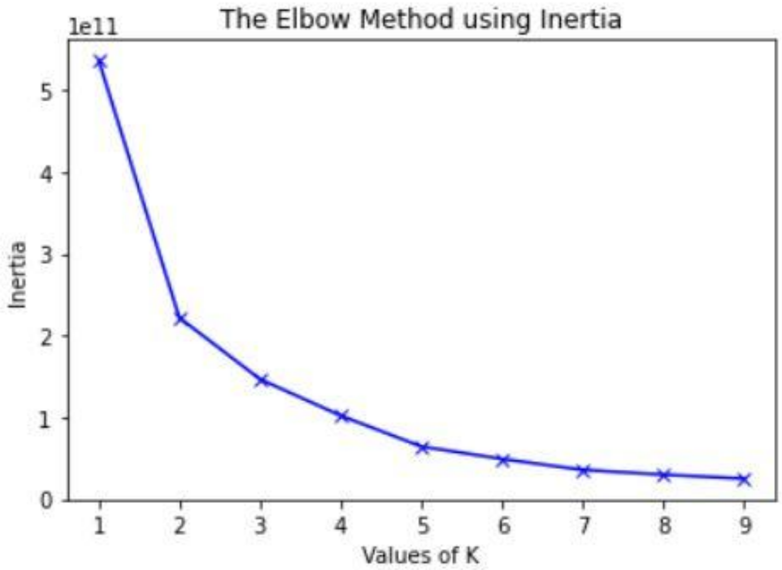
Sentiment Analysis: Sentiment Trend



Clustering

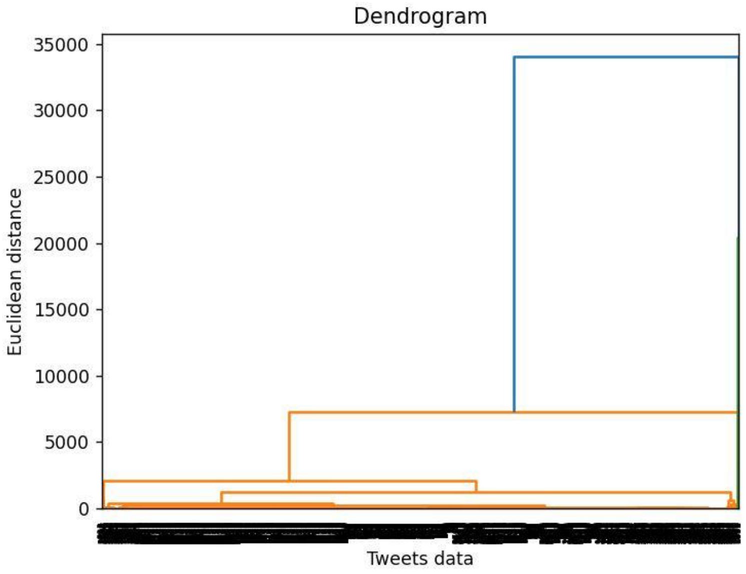
K-means Clustering

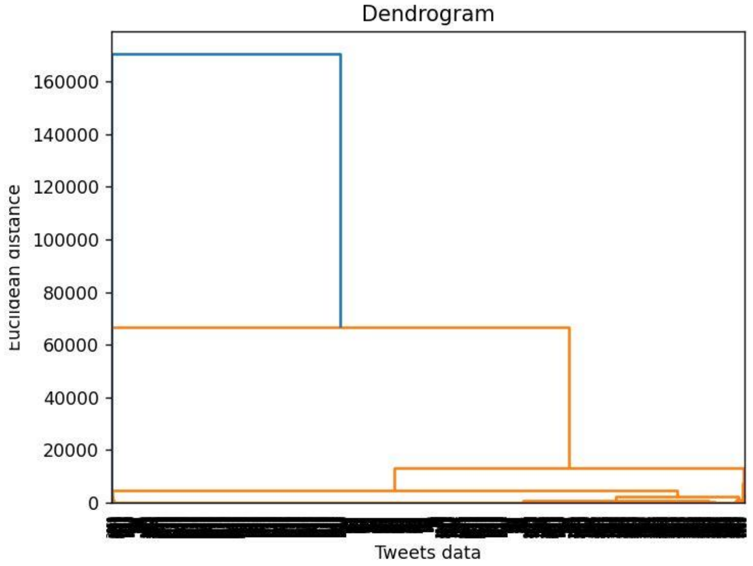




Clustering

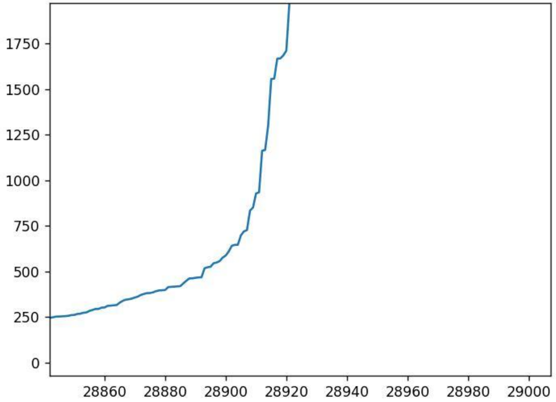
Hierarchical Clustering



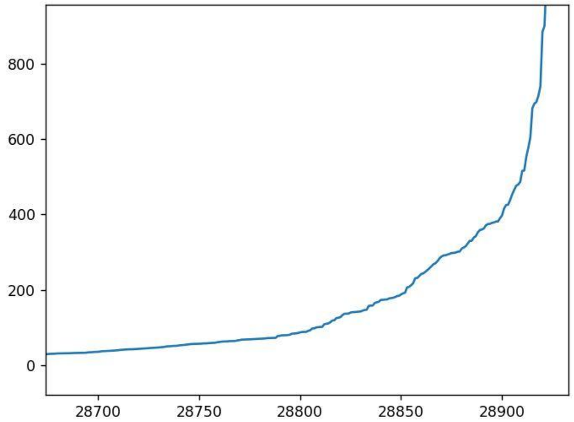


Clustering

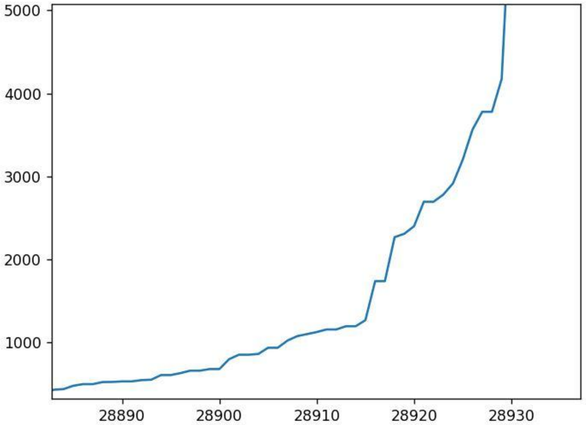
DBSCAN Clustering



Epsilon 825, K=1



Epsilon 725, K=1



Epsilon 1375, k=1

Regression: Predicting Tomorrow’s BTC Price Using Today’s Price

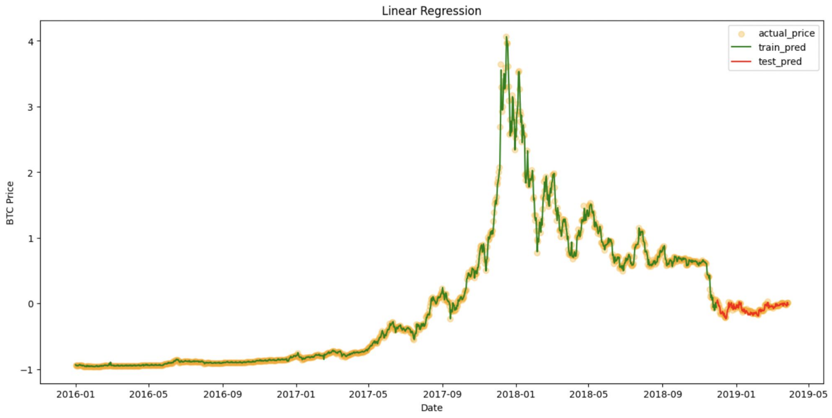
Output Label
(Next day price)

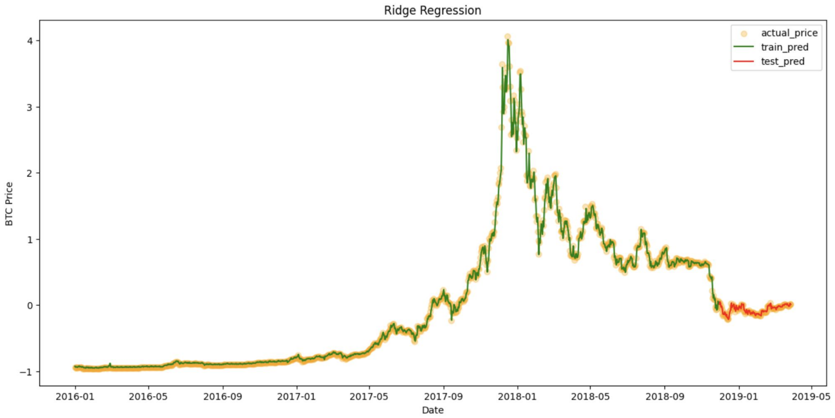
```
Index(['date', 'btc_price', 'tweet_volume', 'likes_gt_0', 'likes_gt_10',
       'likes_gt_100', 'likes_gt_1000', 'retweets_gt_0', 'retweets_gt_1000',
       'pos_sent', 'neg_sent', 'neu_sent', 'user', 'hr_0_6', 'hr_6_12',
       'hr_12_18', 'hr_18_24', 'Friday', 'Monday', 'Saturday', 'Sunday',
       'Thursday', 'Tuesday', 'Wednesday', 'btc_cur_price'],
      dtype='object')
```

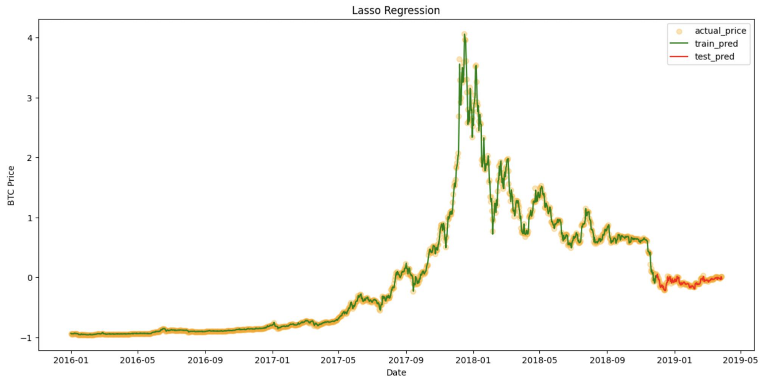
Current day price

- We are discarding the ‘date’ feature since it is not numeric and not adding anything to the data.
- We have also standardized the data using **StandardScaler()** from sklearn.

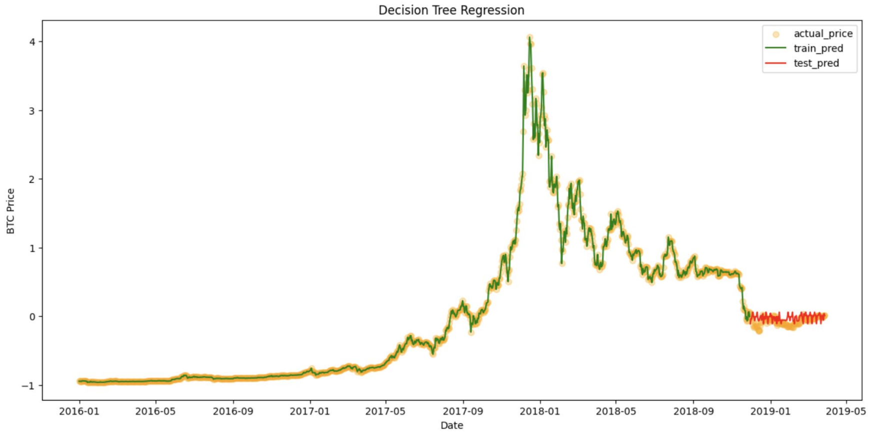
Regression: Regression Models

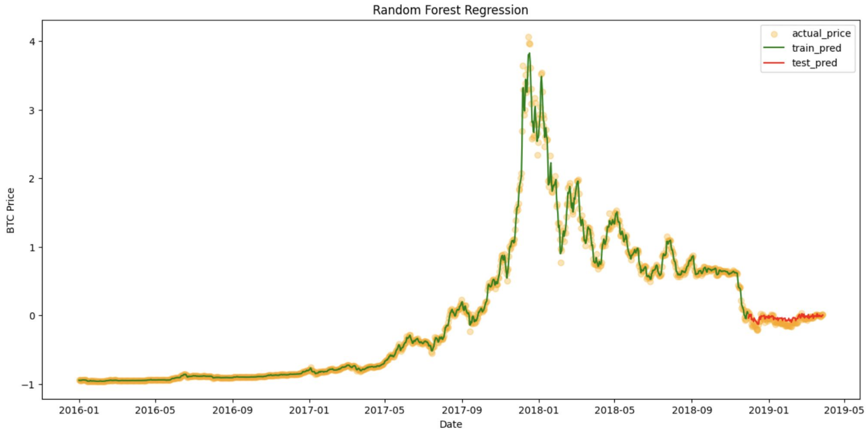




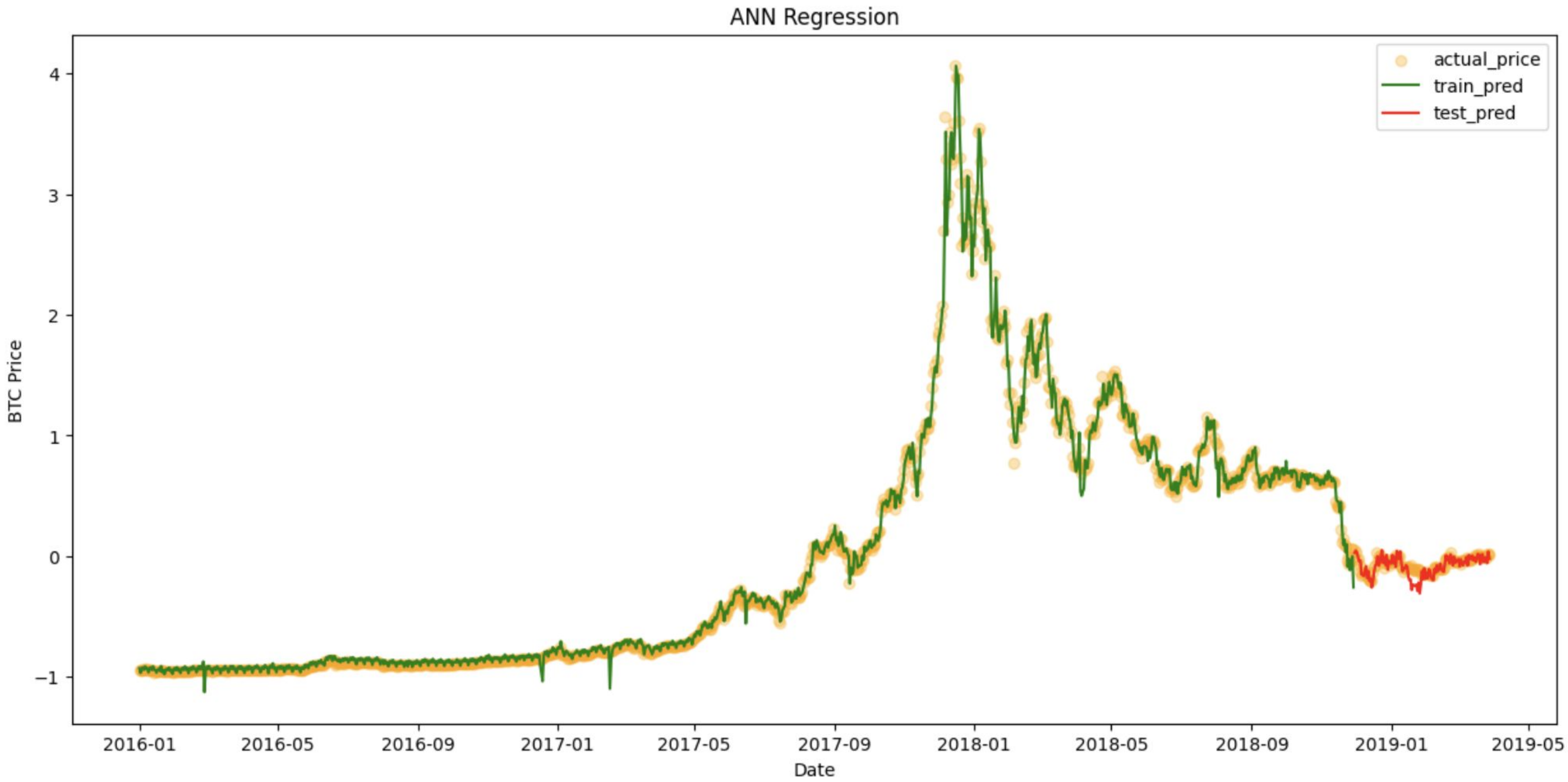


Regression: Tree-based Models

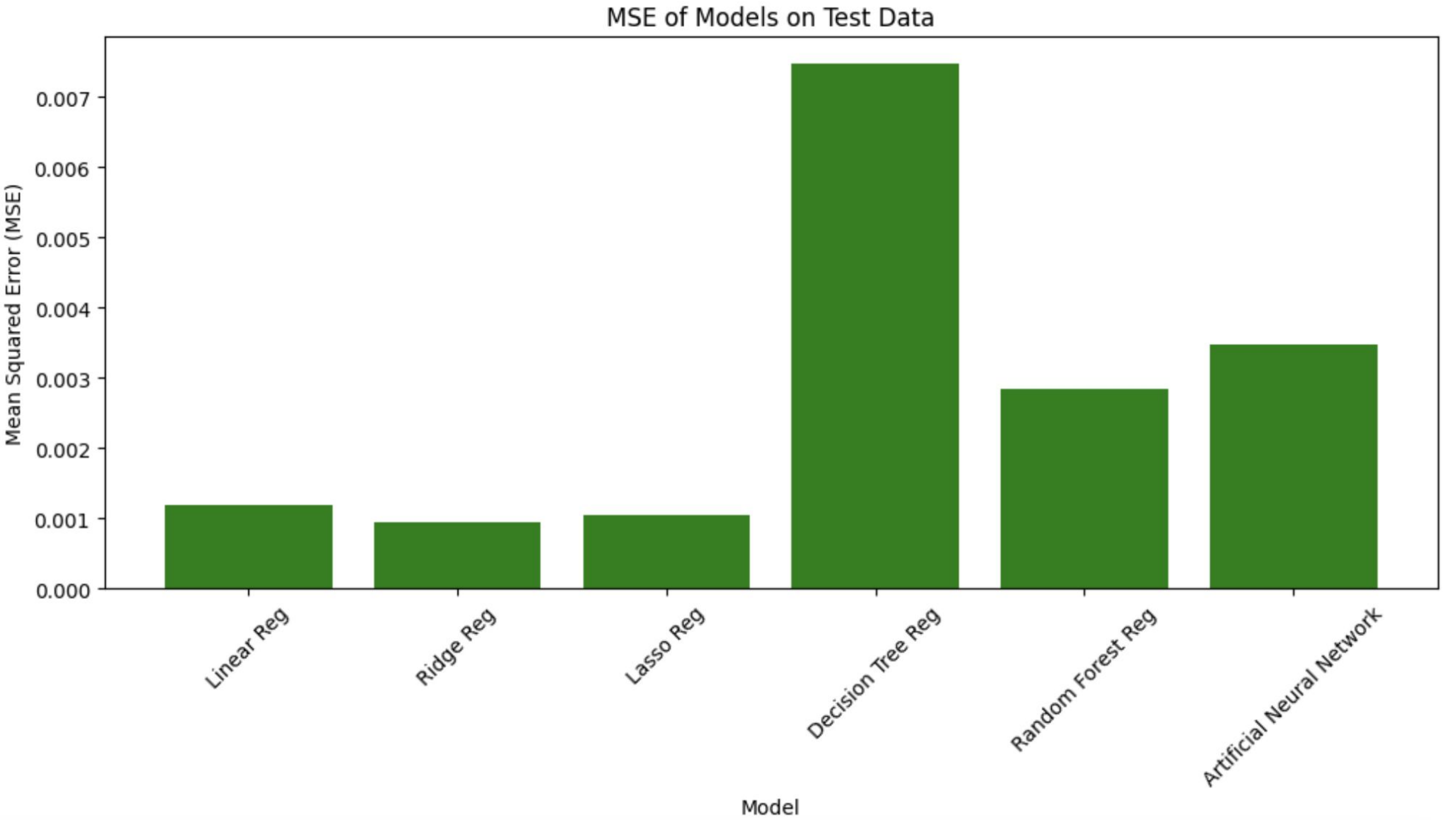




Regression: Artificial Neural Network



Regression: Comparison between Models

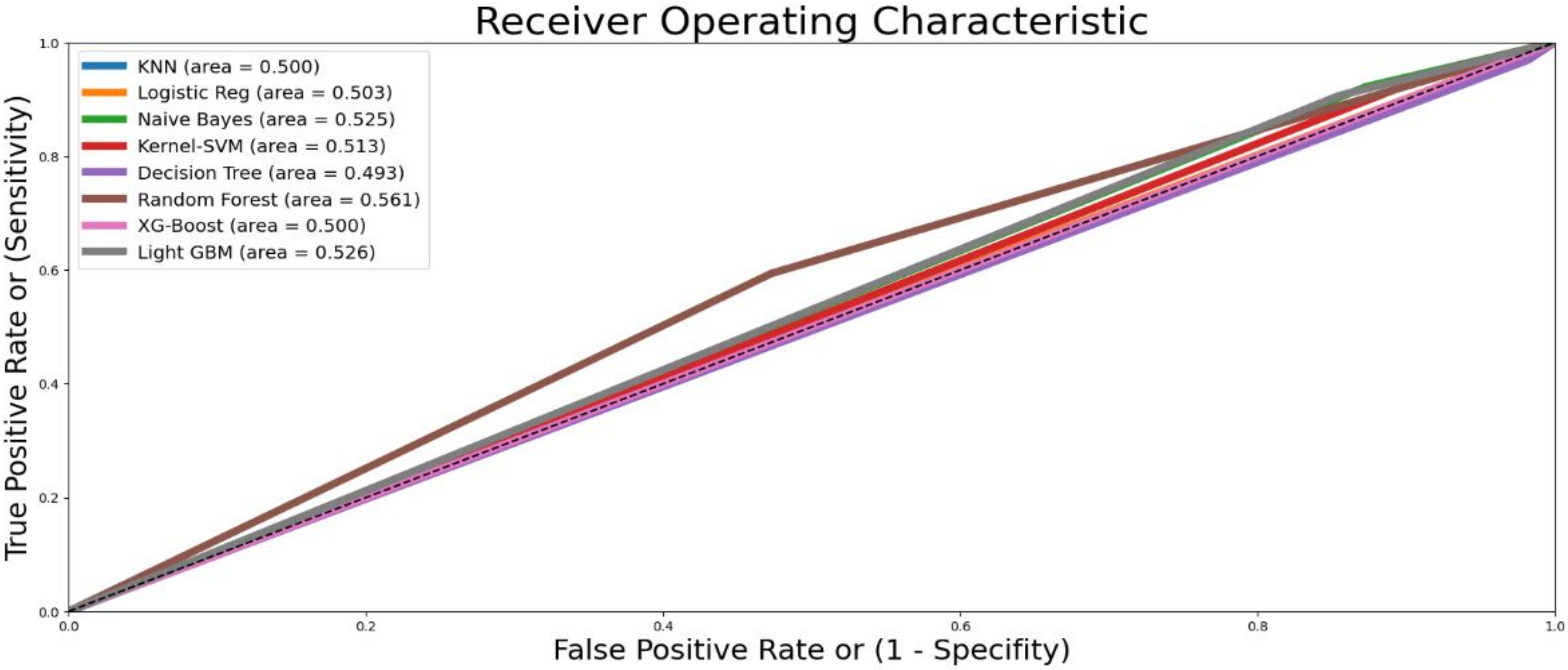


Classification: Models developed and results

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.49	0.54	0.41	0.46
KNN	0.54	0.54	1	0.7
Naive Bayes	0.56	0.55	0.92	0.69
Kernel SVM	0.55	0.54	0.95	0.69
Decision Tree	0.53	0.53	0.97	0.69
Random Forest	0.62	0.64	0.86	0.75
XGBoost	0.54	0.54	1	0.7
Light GBM	0.55	0.55	0.91	0.69

Target variable: BTC_price_movement_direction (Next day price-Current day price)
1 - if price goes up
0- if price dips

Classification: Area under ROC curve for each of the models



We can see that Random Forest classifier has the highest area under the curve