Bitcoin Value Prediction using Twitter Sentiment and Statistical Analysis

Nuttaset Pattanadee
Kevin Chakornsiri
Chayisara Sakunkoo
Julian Kikuchi

QSS 20. Final Class Status Update. 03.07.2022

Outline

- Motivation
- Research questions
- Data
- Methods
  - Twitter Sentiment Extraction
  - "Money" Tweets Extraction
  - Machine Learning
- Results
- Exploratory Data Analysis
- Classical Linear Regression (Econometric approach)
- Limitations/Next steps

Motivation

- The value of Bitcoin is very volatile. To capture the trend on it using statistical analysis may not be enough to paint the picture on which direction Bitcoin market value will go to.
- There hasn't been a research that combines the Twitter Sentiment into the statistical data to create a model.

Research Questions

- Our overall research question are based on the finding the answer parameters.
  - **Prediction Model:** How accurate is our machine learning model using statistical analysis?
  - **Twitter Sentiment:** Which type of users (verified vs unverified) have more impact to the value of Bitcoin? Is there a trend between Twitter sentiment and the differential value of Bitcoin? Can the text regarding 'money' in Tweets be analysed to predict the value of Bitcoin?
  - **Combined Model:** Does adding Twitter sentiment improves the performance of the model? Does adding the 'money' improves the performance of the model?

Data Sources

The following table shows the data sources used in the project.

	Bitcoin Data	**Twitter Data**
Dataset Name	Historical data, Block-chain data	Scraped Tweets
Data Interval	2020 to 2021	
Unit of Analysis	Day	Tweet
Fields/Variables	open_bitcoin, close_bitcoin, value_number_transaction, value_hash_rate	id, tweet, metrics

Methods: Data Acquisition

- Acquiring Twitter data using Twint scraping tool
  - Tweet data, e.g., Tweet and its metrics
- Acquiring Bitcoin data using TvDatafeed API interface
  - Historical data, e.g., Open Price and Close Price of Bitcoin
  - Block-chain data, e.g., Number of Transactions in Bitcoin network and Total Hash Rate

Methods: Feature Selection for Tweet Data

- Sentiment Data from tweets
  - Clean the tweet data
  - Analyze the sentiment in each tweet
- Bitcoin Price Opinions from tweets
  - Extract money(number) from each tweet
  - Normalize the money with the open price on that day

Methods: Feature Selection for Tweet Data

- **Sentiment Data from tweets**
  - Clean the tweet data
  - Analyze the sentiment in each tweet
- Bitcoin Price Opinions from tweets
  - Extract money(number) from each tweet
  - Normalize the money with the open price on that day

Methods: Sentiment Data - Data Cleaning

- Drop the duplicated tweets (ADs)
- Filter only EN language
- Convert to lower case
- Remove URLs, mentions, non-word characters, numbers, and stopwords
- Reduce character sequences more than 3 to 3 and 2 or more spaces to a single space.

Methods: Sentiment Data - Data Cleaning

Before Cleaning	After Cleaning
Bitcoin the worst decision i made this decade	worst decision made decade
#bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017 https://t.co/W3ghBhpfMX	rally begun reached peak
Bitcoin faces uncertain 2022 after record year https://t.co/76aqPx4l1Q https://t.co/76aqPx4l1Q	faces uncertain record year
@Mike_Rooker I don't give investment advice here bud, but never gamble what you can't afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	give investment advice bud never gamble afford lose opinion highly speculative load
Bitcoin up 2.58% over the last 24 hours https://t.co/iuY90oolUI	up last hours

Methods: Sentiment Data - Sentiment Analysis

- From vaderSentiment.vaderSentiment libraly, we use SentimentIntensityAnalyzer to analyze the cleaned tweet data
- Add words about Cryptocurrency in lexicon (ex. up, down, green, red, buy, and sell)
- Get the compound score for each tweet

Methods: Sentiment Data - Sentiment Analysis

ID	Tweet	Sentiment Score
12122391436877414 40	Bitcoin the worst decision i made this decade	-0.6249
12122673167899525 12	#bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017 https://t.co/W3ghBhpfMX	0.1027
14771383189827256 33	Bitcoin faces uncertain 2022 after record year https://t.co/76aqPx4l1Q https://t.co/76aqPx4l1Q	-0.2960
12135019270587924 50	@Mike_Rooker I don't give investment advice here bud, but never gamble what you can't afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	-0.3634
12130857715671080 97	Bitcoin up 2.58% over the last 24 hours https://t.co/iuY90oolUI	0.0

Methods: Feature Selection for Tweet Data

- Sentiment Data from tweets
  - Clean the tweet data
  - Analyze the sentiment in each tweet
- **Bitcoin Price Opinions from tweets**
  - Extract money(number) from each tweet
  - Normalize the money with the open price on that day

Methods: Bitcoin Price Opinions - Data Extracting

- Extract money(number) from each tweet

Tweet	Bitcoin Price Opinions
Bitcoin the worst decision i made this decade	[]
#bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017	[2013, 20000, 2017]
Bitcoin faces uncertain 2022 after record year	[2022]
@Mike_Rooker I don't give investment advice here bud, but never gamble what you can't afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	[]
Bitcoin up 2.58% over the last 24 hours	[2.58, 24]

Methods: Bitcoin Price Opinions - Data Extracting

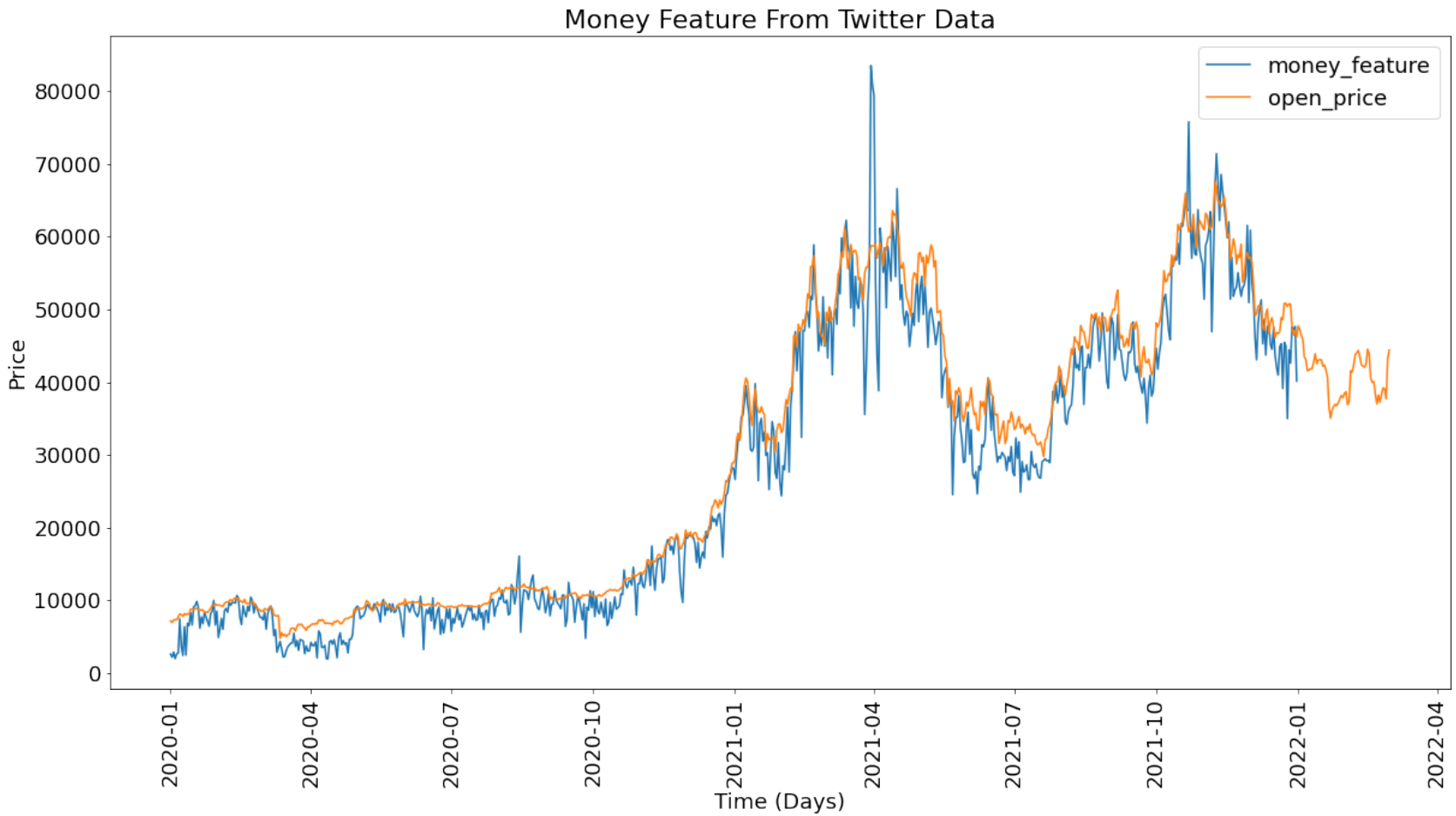
- Filter the number to the range of open price on each day

  $(0.25 * \mathit{OpenPrice} < \mathit{BitcoinPriceOpinions} < 1.75 * \mathit{OpenPrice})$

Tweet	Bitcoin Price Opinions
Bitcoin the worst decision i made this decade	[]
#bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017 https://t.co/W3ghBhpfMX	[20000]
Bitcoin faces uncertain 2022 after record year https://t.co/76aqPx4l1Q https://t.co/76aqPx4l1Q	[]
@Mike_Rooker I don't give investment advice here bud, but never gamble what you can't afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	[]
Bitcoin up 2.58% over the last 24 hours https://t.co/iuY90oolUI	[]

Methods: Bitcoin Price Opinions - Data Extracting

- Graph between average bitcoin price opinion (money feature) on each day and bitcoin open price on each day

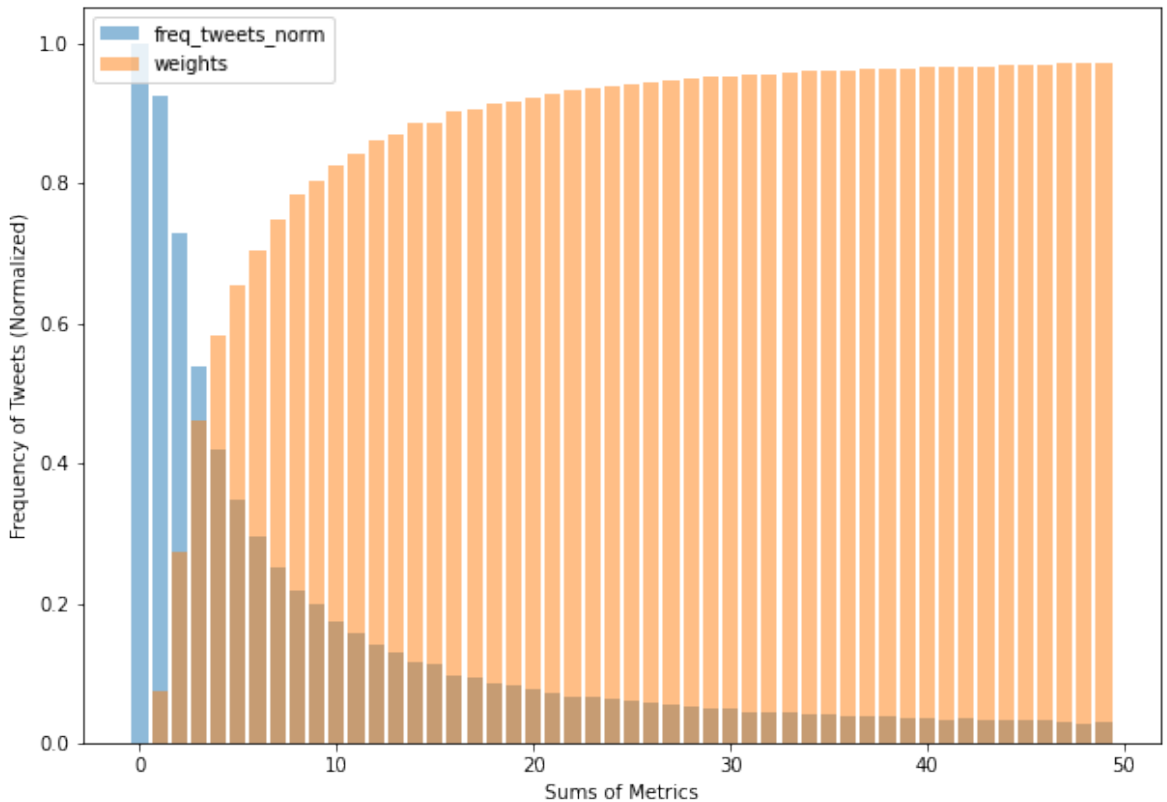


Methods: Feature Weighting

- Each Tweets need to be weighted according to its key metrics:
  - Replied counts
  - Like counts
  - Retweets counts
- The metrics will be used to weigh the sentiment of that Tweets
  - The weights will be calculated from the distribution of 'Sum of Metrics'
  - The 'Frequency of Tweets' will be normalized to indicate the importance of Tweet

Methods: Feature Weighting

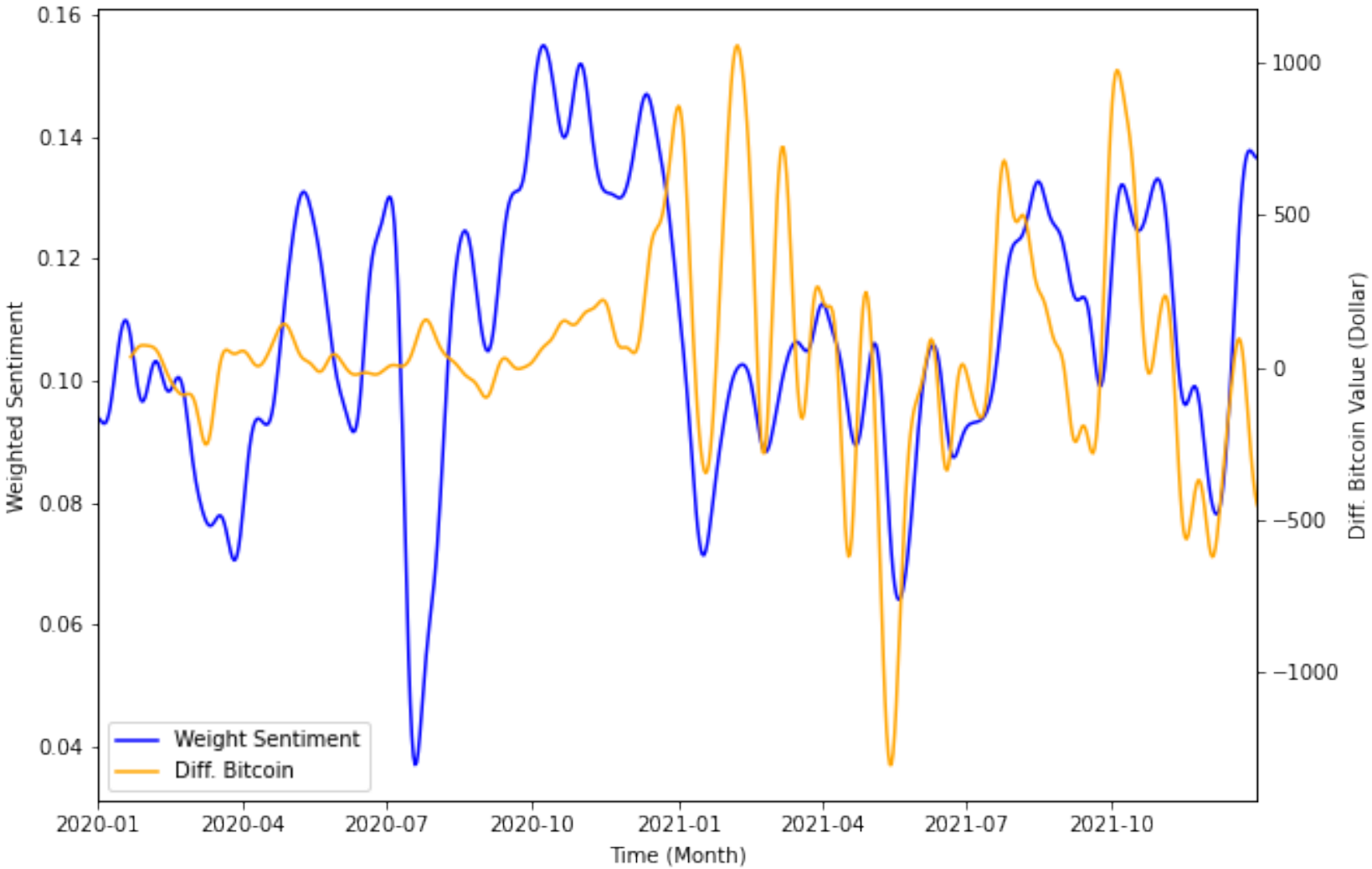
- The figure below shows the relationship between 'Sum of Metrics' and 'Frequency of Tweets'



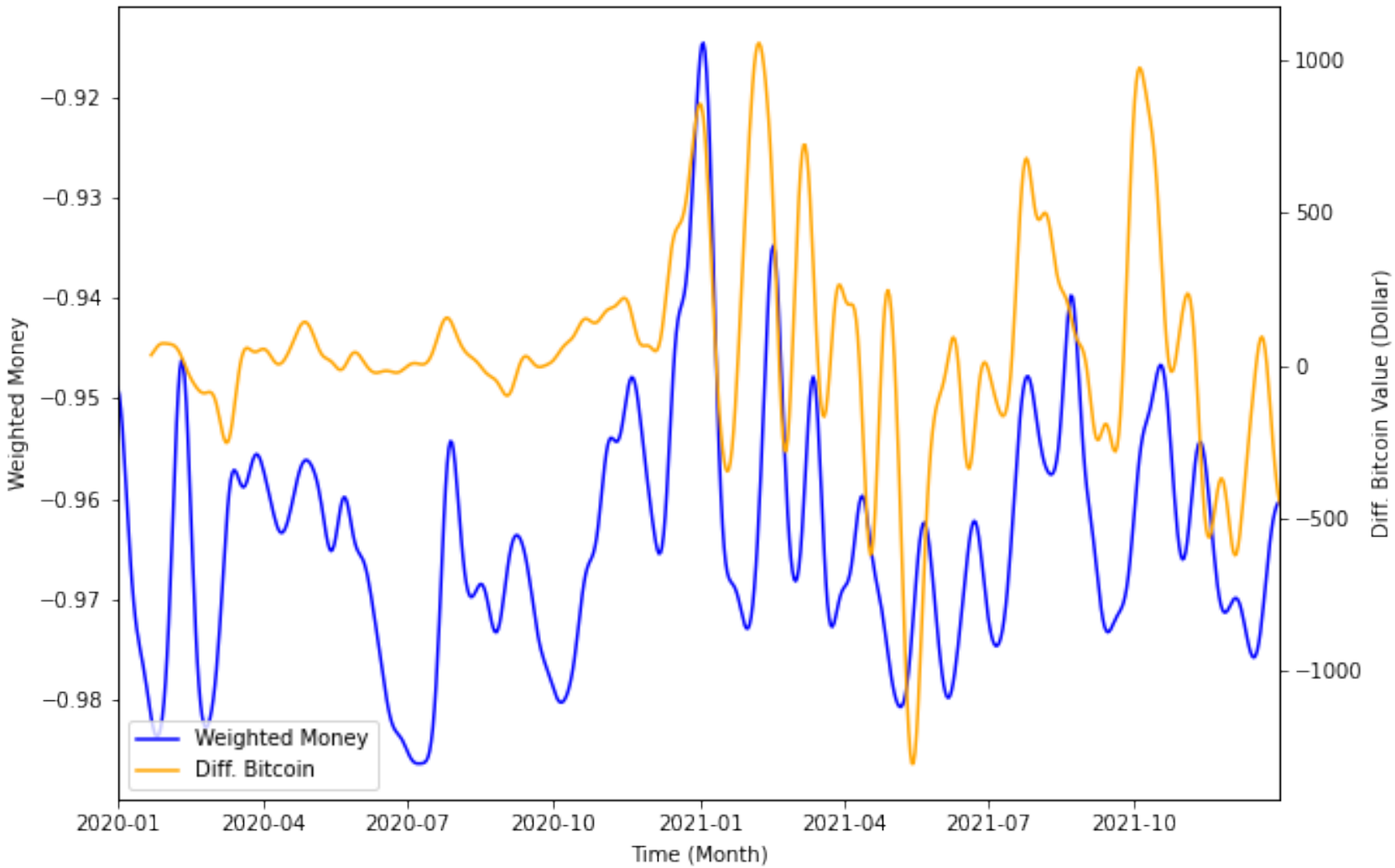
Methods: Analysis or Visualization

- To compare how both features might somehow be related to the value of Bitcoin, the following plots are analysed:
  - Sentiment Value vs. Differential of Bitcoin value
  - Normalized "Money"vs. Differential of Bitcoin value
- The reason why we use the **Differential** of Bitcoin value is that so we can analyse the direction of the price in correspond to the feature values.

Methods: Analysis or Visualization



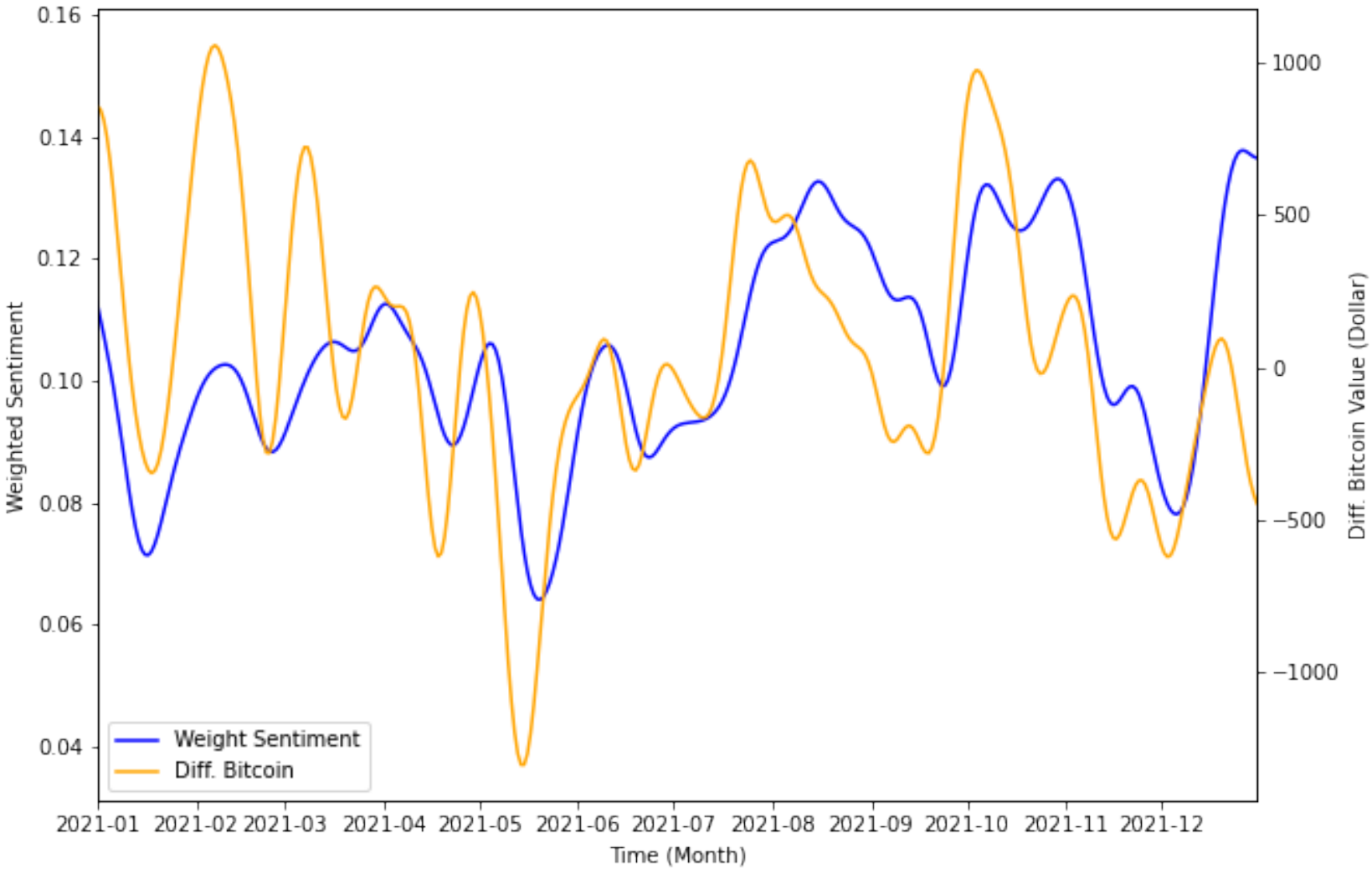
Methods: Analysis or Visualization



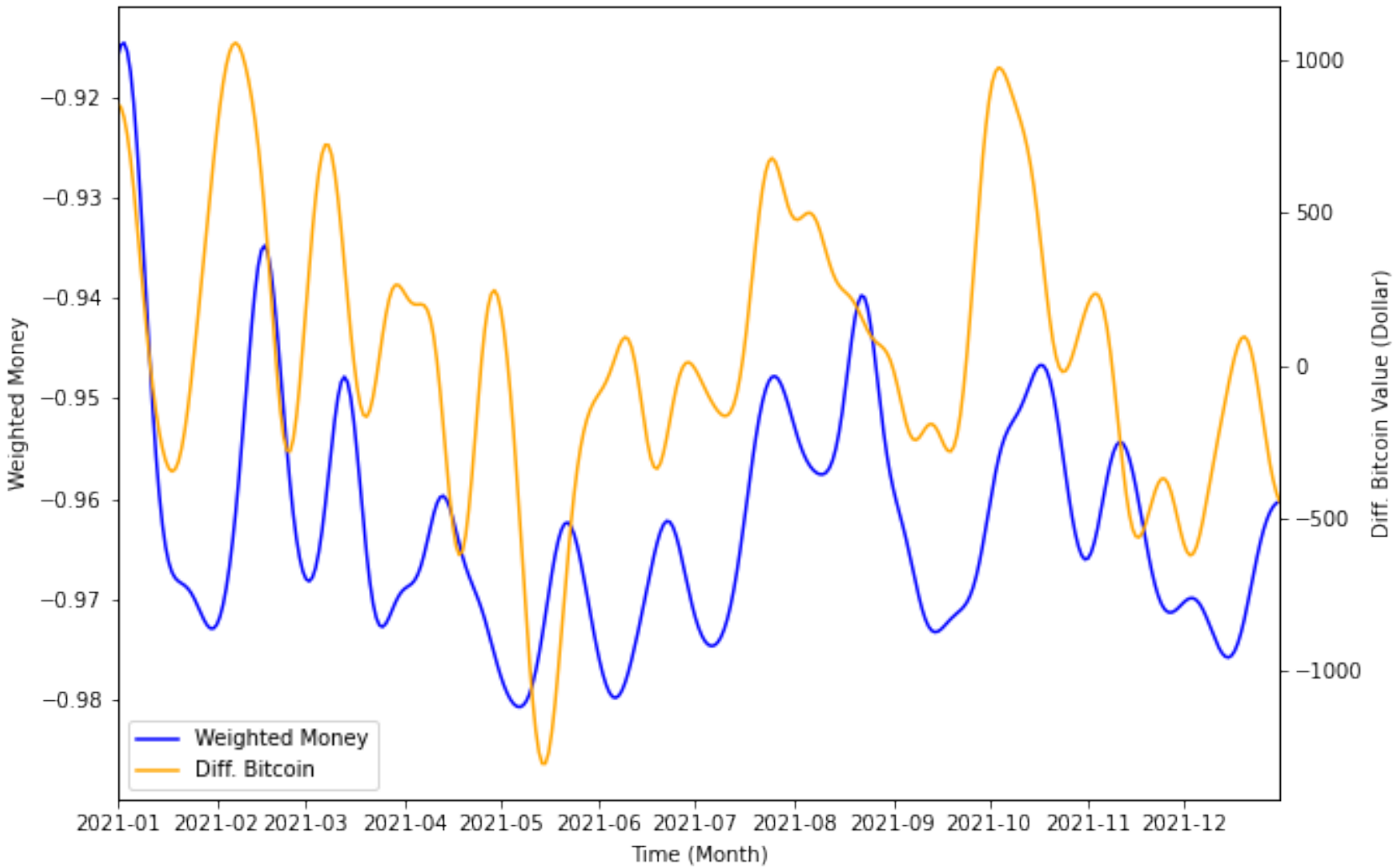
Methods: Analysis or Visualization

- Key insights that we've found:
  - In 2020, the feature values do not have much correlation to the value of Bitcoin
  - 2021 is the year where Bitcoin saw an uproar in price, causing the market value to be very volatile.
- To further analyse the data, we filter to the plot to only 2021 to see a clearer picture

Methods: Analysis or Visualization



Methods: Analysis or Visualization



Methods: Analysis or Visualization

- Key insights that we've found:
  - We can see a very interesting correlation between the two lines inside both plots. Both lines seems to follow the trend of the Bitcoin value
  - Although sometimes lag behind a bit, the value can theoretically be a good indication of whether the value would go up or down
- In conclusion: Adding these two features to the model may result in a higher accuracy in prediction.

Methods: Machine Learning Ridge Regression

Ridge regression is ordinary least squares regression with an L2 penalty term on the weights in the cost function.

- Merge the historical data and the block-chain data

	datetime	open_bitcoin	close_bitcoin	value_hash_rate
0	2020-01-01	7200.77	6965.71	96717718.3123778
1	2020-01-02	6965.49	7344.96	115924073.721928
2	2020-01-03	7345.0	7354.11	115238132.457301

- Apply linear interpolating to get rid of null value

```
open_bitcoin                0
close_bitcoin               0
open_eth                    0
open_bnb                    0
open_ada                    0
value_number_transaction    0
value_number_address        0
value_transaction_second    0
value_total_bitcoin         0
value_hash_rate             0
datetime                    0
```

```
datetime                    0
open_bitcoin                0
close_bitcoin               0
open_eth                    0
open_bnb                    0
open_ada                    0
value_number_transaction    0
value_number_address        2
value_transaction_second    0
value_total_bitcoin        35
value_hash_rate             0
```

Methods: Machine Learning Ridge Regression

- Separate data into x and y dataframes
  - Dataframe x has 9 features as columns which are open_bitcoin,open_eth,open_bnb,open_ada, value_number_transaction,value_number_address, value_transaction_second, value_total_bitcoin,value_hash_rate
  - Dataframe y is the output that we want to predict which is close_bitcoin
- Split data into 60% training, 20% validation, and 20% testing set chronologically
- Standardize the x_train, x_validate, and x_test dataframes

Methods: Machine Learning Ridge Regression

- Create a function to tune hyperparameter which is alpha in the L2 penalty term

```
def tune_alpha(x_train,x_validate,y_train,y_validate):
    alp =np.logspace(-2, 5, 1000).tolist()
    column_names = ["alpha", "RMSE","MAE"]
    data =[]
    for i in alp:
        model = Ridge(alpha=i)
        model.fit(x_train, y_train)
        y_validate_predict = model.predict(x_validate)
        rmse=mean_squared_error(y_validate, y_validate_predict, squared=False)
        mae=mean_absolute_error(y_validate, y_validate_predict)
        data.append([i,rmse,mae])
    df = pd.DataFrame(data, columns=column_names)
    return df
```

Methods: Machine Learning Ridge Regression

- Tune the alpha using the created function to find the alpha that provide the model with the lowest root mean square error and mean absolute error
- Create a Ridge Regression model with the learned alpha which is 0.5646
- Create a dataframe containing parameter of the model along with its coefficient
- Predict the x_test dataframe with the model and store the prediction in y_pred
- Evaluate the performance of the model by create RMSE, MAE, and R2 score.

Results: Feature and its coefficient

	Feature	Coefficients
0	open_bitcoin	3243.935371
1	open_eth	-133.801654
2	open_bnb	35.630904
3	open_ada	23.872388
4	value_number_transaction	-74.558126
5	value_number_address	106.299081
6	value_transaction_second	-31.505036
7	value_total_bitcoin	-5.033275
8	value_hash_rate	-14.366322

It can be clearly seen that the dominating parameter is the open_bitcoin parameter. It affects our prediction the most. The more magnitude the coefficient has, the more impact it has to our model. For example, if the open value of bitcoin is very high, the prediction from our model tends to be high as well. The least important factor in our model is the total bitcoin that has been mined.

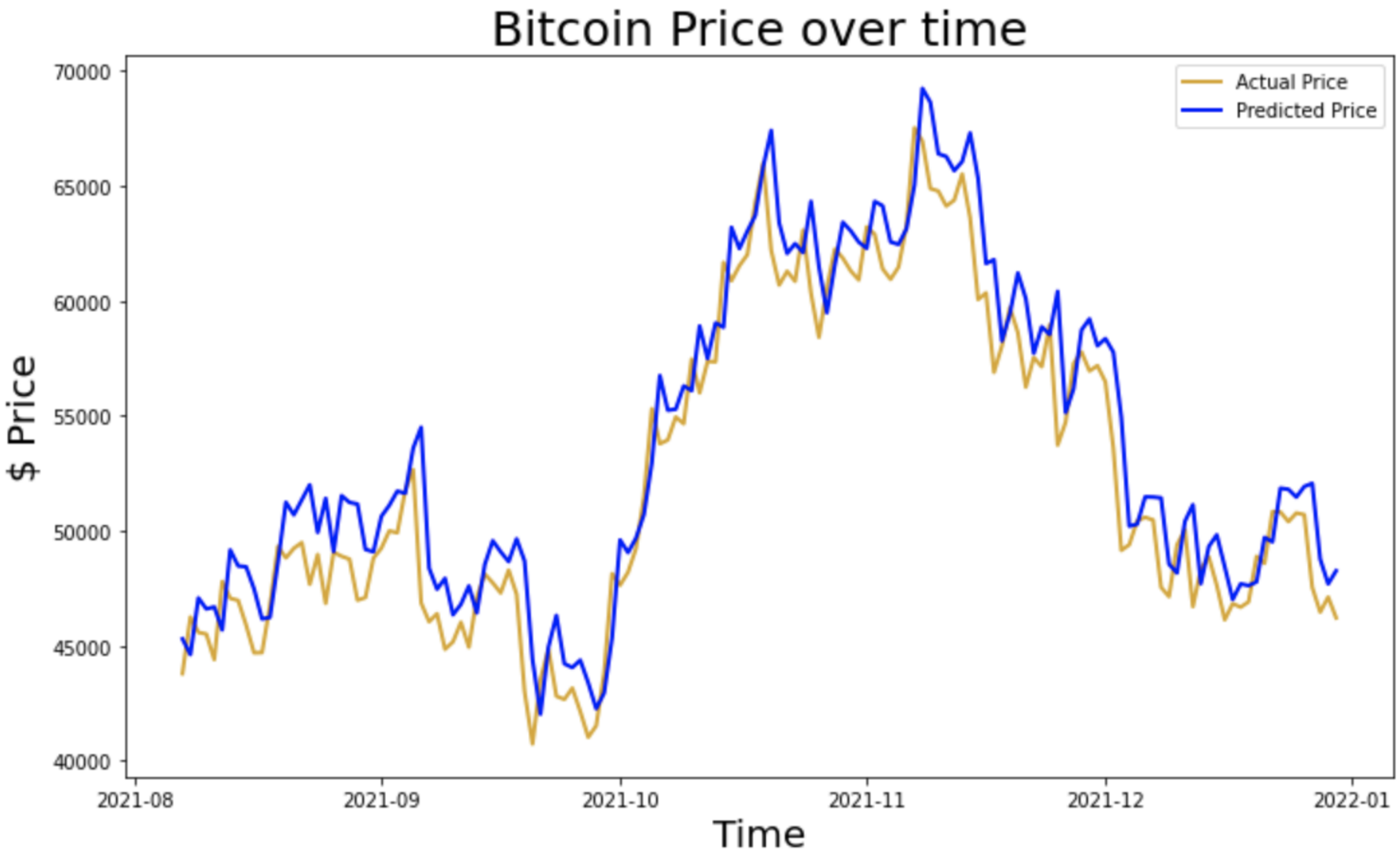
Results: RMSE,MAE, R2 score

```
Root Mean Squared Error of the train set: 383.74402876131956
Root Mean Squared Error of the test set: 2316.536366444953
Mean Absolute Error of the train set: 243.72157813269274
Mean Absolute Error of the test set: 1832.112972223676
R2 score: 0.8871355399380723
```

R2 score tells us how well our model performs. The higher R2 means the predictions fits well with the true values. If the R2 is low, the prediction and the actual value would be really far away from each other. According to our R2 score, our model performance is significantly high. For the

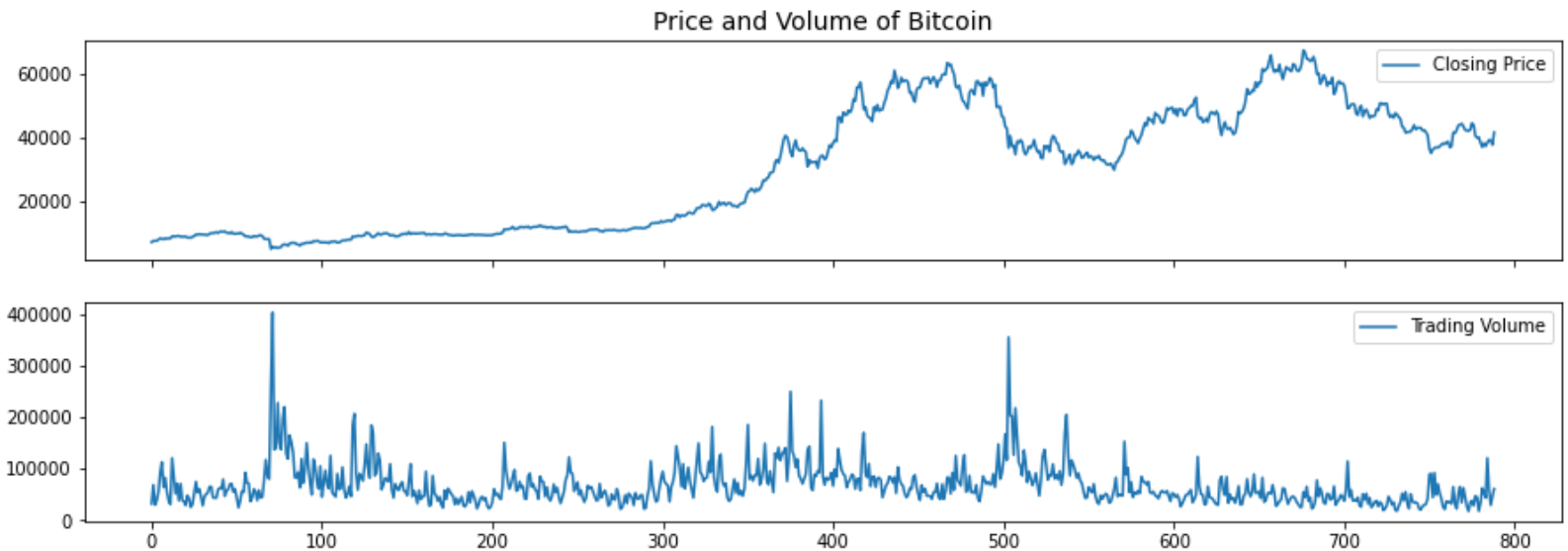
RMSE and MAE, they are used to measure the error of our data. If we had not added the L2 penalty term and not tuned its alpha, the differences between training error and testing error would be much higher than this due to overfitting.

Results: Bitcoin price over time



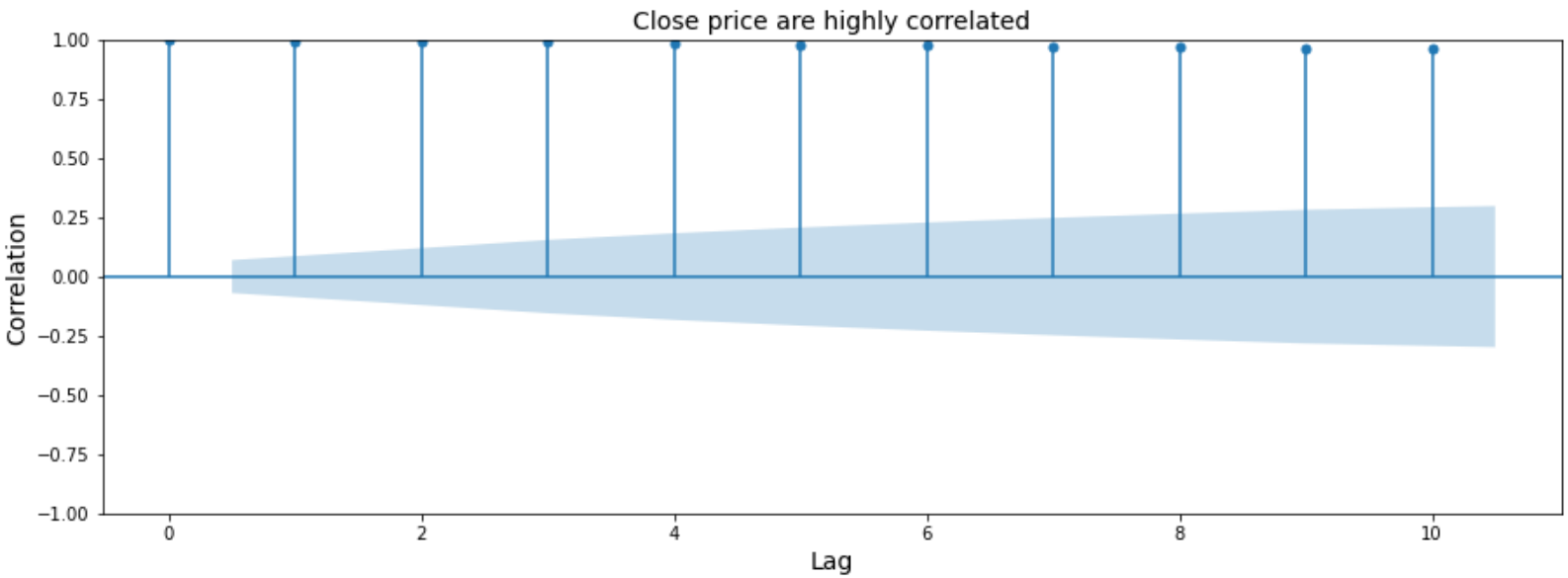
Exploratory Data Analysis

- Price of Bitcoin rising over time
- Price of Bitcoin is not independent - today's price is related to yesterday's etc
  - Problem in running a classical regression to estimate coefficient



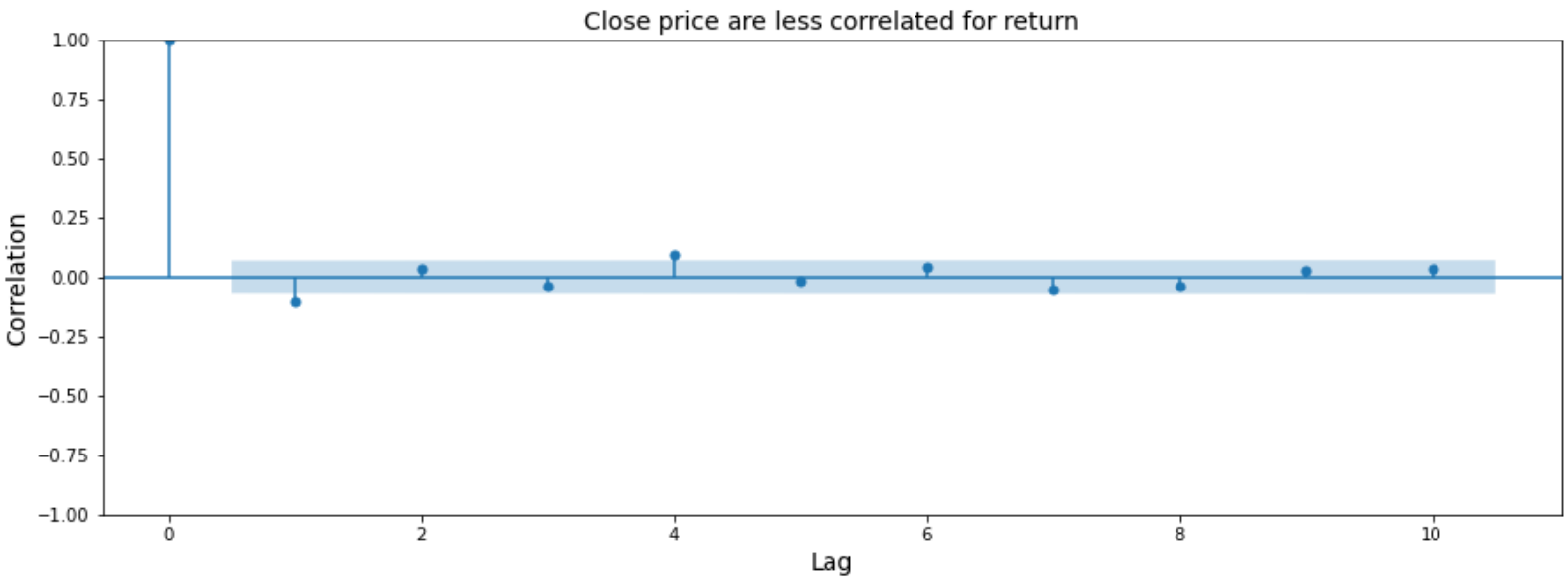
Exploratory Data Analysis

- Price is indeed highly correlated in all lags...
  - Yesterday's price is almost the same as today's price



Exploratory Data Analysis

- 1 - day Return is not correlated as much
- Better to use return instead of price for classical regressions to estimate the correlation between sentiment and return
  - Return: $\frac{Price_{t+1} - Price_t}{Price_t}$ where t represents time



Classical Linear Regression

- Instead of focusing on estimating the price, we try to estimate the correlation between sentiment and return
- Question:
  - Does it make more sense to calculate the correlation between return and change in sentiment or just sentiment value itself?
  - I think it makes more sense to see whether an increase/decrease in positive sentiment is correlated with increase/decrease in return

Classical Linear Regression

95% confidence interval contains zero

$$y_{t+1} = \alpha + \frac{\Delta s_{t-1}}{s_{t-2}}$$

where y_{t+1} represents 1-day return of bitcoin at time t+1 and $\frac{\Delta s_{t-1}}{s_{t-2}}$ represents a percentage change in sentiment at time t-1. We have time t+1 for independent variable because at day t, given change in sentiment on day t - 1, we want to see if the sentiment can be predictive of return.

Classical Linear Regression

Dep. Variable:	close_return	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.001
Method:	Least Squares	**F-statistic:**	0.3893
Date:	Sun, 06 Mar 2022	**Prob (F-statistic):**	0.533
Time:	17:33:34	**Log-Likelihood:**	1412.2
No. Observations:	792	**AIC:**	-2820.
Df Residuals:	790	**BIC:**	-2811.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.0030	0.001	2.074	0.038	0.000	0.006
weighted_sent_change	8.427e-05	0.000	0.624	0.533	-0.000	0.000

Omnibus:	248.908	**Durbin-Watson:**	2.204
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5510.965
Skew:	-0.870	**Prob(JB):**	0.00
Kurtosis:	15.805	**Cond. No.**	10.7

Classical Linear Regression

95% confidence interval contains zero

$$y_{t+1} = \alpha + s_{t-1}$$

where s_{t-1} is the sentiment at time t-1.

Classical Linear Regression

Dep. Variable:	close_return	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.001
Method:	Least Squares	**F-statistic:**	0.1001
Date:	Sun, 06 Mar 2022	**Prob (F-statistic):**	0.752
Time:	17:33:41	**Log-Likelihood:**	1299.1
No. Observations:	731	**AIC:**	-2594.
Df Residuals:	729	**BIC:**	-2585.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.0021	0.004	0.470	0.639	-0.007	0.011
weighted_sentiment	0.0123	0.039	0.316	0.752	-0.064	0.088

Omnibus:	248.061	**Durbin-Watson:**	2.239
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5525.713
Skew:	-0.974	**Prob(JB):**	0.00
Kurtosis:	16.328	**Cond. No.**	25.9

Next steps

- Combining the result from sentiment analysis with the result from statistical analysis and compare the performance of both models.
- Improve our model using the exploratory about return and changes in sentiment