UCD Michael Smurfit
Graduate Business School

Financial Data Science Project

[redacted]

Project Submission

Group 5

Assessment Submission Form

Student Name	**Student Number**
Ross Kearney	[redacted]
Mark [redacted]	[redacted]
Joseph [redacted]	[redacted]

Introduction

With $3.5 Trillion in assets under management by hedge funds, predicting the stock markets is a huge industry. Despite the widespread belief that stock prices follow a random walk pattern, several studies have found evidence that human emotions and psychology have an impact on the stock market. According to the efficient market hypothesis, all aspects of the stock market are already price adjusted for all news, current events, and product releases, and there is no way to predict future prices (Pagolu, 2016).

With the rise of social media, information about popular sentiment has been readily available. Social media is evolving into an ideal medium for sharing public sentiment on any topic and has a big impact on public opinion in general. Every day, about 140 million tweets are shared by over a million individuals. As a result, Twitter can be a useful investment tool. The information gleaned from tweets has the potential to be very beneficial for stock market prediction.

For stock market analysts, there is a need to develop new tools that, when used in conjunction with traditional prediction models, can fine-tune predictions by factoring in factors that are not directly related to the asset. Such factors include the general public's perception of the market and opinions of the studied asset in particular. Complex patterns in data can be recognized with the help of Machine Learning. This report develops and tests a Machine Learning model in order to predict stock values based on social sentiment.

Section 1: Data Set Collection

The essence of our project is to determine public sentiment around an asset based on tweets or recent news stories surrounding them in 'The Financial Times'. From this sentiment, we want to determine whether or not an accurate prediction of future price can be deduced and used within a trading strategy. The assets chosen were Tesla, Bitcoin and the S&P 500 (^GSPC). These three assets represent a good blend of institutional/retail ownership and controversial public opinion, enabling an interesting analysis.

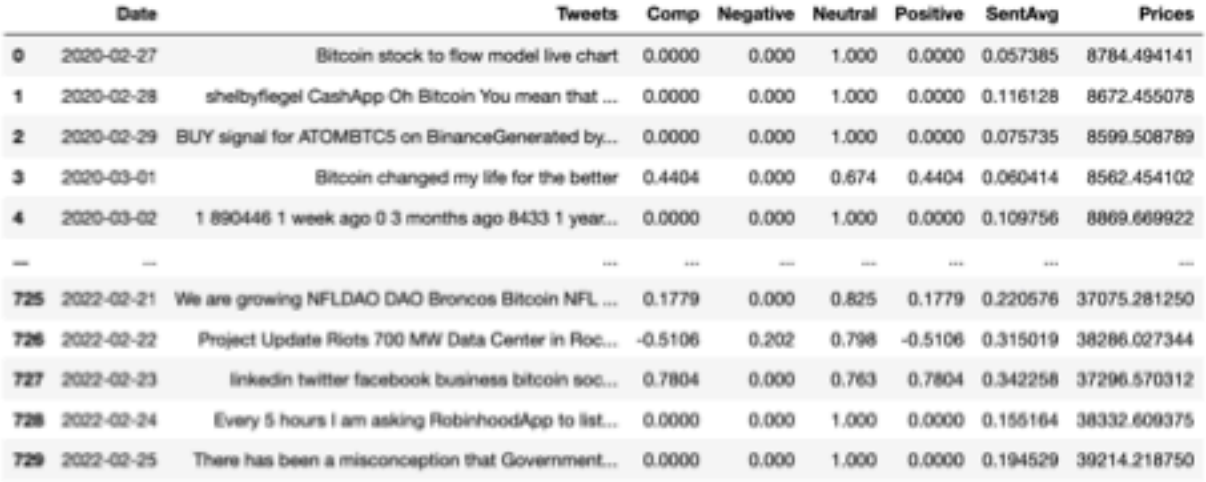
We obtained access to the Academic Research Twitter API in order to scrape tweets[3]. This allows the Python application to access the data and interact with external software components making the web scraping process easier. In order to pull tweets for the selected time period, a loop was created to collect 15 tweets per day for any given search parameter, along with the dates of publication[7].

Following this, we compiled a data set consisting of headlines from Financial Times articles in order to determine their sentiment as a representation from a more mainstream media source to add another layer to our research. In this instance, we ran a loop which scraped headlines relating to a chosen search query from the first 20 pages of the financial times along with the date they were published. Extracting them and converting them into 'lxml' format using the programme 'BeautifulSoup'.[25]

Finally, we imported price data from Yahoo finance using the 'yfinance' program for each of our assets[10]. This was an important component as the daily closing prices were used to compare and make future price predictions based on sentiment.

Section 2: Database creation and querying

Once the necessary data was scraped from Twitter and the Financial Times, we compiled it into dataframes using the 'pandas' package[8]. We added five additional columns to our dataframe: 'Comp', 'Negative', 'Neutral', 'Positive', 'SentAvg' and 'Prices',[13] in order to facilitate sentiment analysis. Once completed, this then enabled us to assign the polarity to each tweet [13 & 14]. Considering 15 tweets were pulled per day, the sentiment of these tweets were averaged to give a representation of the sentiment for that day. The corresponding asset price for each date was added into the 'Prices' column. This served as our database with which we conducted the sentiment analysis.



	Date	Tweets	Comp	Negative	Neutral	Positive	SentAvg	Prices
0	2020-02-27	Bitcoin stock to flow model live chart	0.0000	0.000	1.000	0.0000	0.057385	8784.494141
1	2020-02-28	shelbyfiegel CashApp Oh Bitcoin You mean that ...	0.0000	0.000	1.000	0.0000	0.116128	8672.455078
2	2020-02-29	BUY signal for ATOMBTC5 on BinanceGenerated by...	0.0000	0.000	1.000	0.0000	0.075735	8599.508789
3	2020-03-01	Bitcoin changed my life for the better	0.4404	0.000	0.674	0.4404	0.060414	8562.454102
4	2020-03-02	1 890446 1 week ago 0 3 months ago 8433 1 year...	0.0000	0.000	1.000	0.0000	0.109756	8869.669922
...	...	...	...	...	...	...	...	...
725	2022-02-21	We are growing NFLDAO DAO Broncos Bitcoin NFL ...	0.1779	0.000	0.825	0.1779	0.220576	37075.281250
726	2022-02-22	Project Update Riots 700 MW Data Center in Roc...	-0.5106	0.202	0.798	-0.5106	0.315019	38286.027344
727	2022-02-23	linkedin twitter facebook business bitcoin soc...	0.7804	0.000	0.763	0.7804	0.342258	37296.570312
728	2022-02-24	Every 5 hours I am asking RobinhoodApp to list...	0.0000	0.000	1.000	0.0000	0.155164	38332.609375
729	2022-02-25	There has been a misconception that Government...	0.0000	0.000	1.000	0.0000	0.194529	39214.218750

Figure 1: Full dataframe of 'bitcoin' tweets with sentiment scores and 'BTC-USD' prices each day

Section 3: Cleaning and organisation of the data

The data did not require a substantial amount of cleaning and organisation once scraped from Twitter. One step which was necessary was removing the URL's and special characters from each tweet [2]. Additionally, we encountered an issue surrounding some blank entries for prices in the dataframe mainly due to holidays and weekends[12]. This was overcome by inserting the previous close price into any empty cells.

The only cleaning and organisation that was needed for the Financial Times articles was when in the rare case where two articles would mention the search term in one day, instead of taking both entries for individual entries, we got the average sentiment of the two and used that figure instead.

Section 4: Data Visualisation

Data collected was analysed at a high level, firstly to identify the percentage of positive vs negative sentiments for the chosen assets.

Twitter search: tsla
From: 2021-02-27
To: 2022-02-26
Ticker search: TSLA

% positive = 93.13
% negative = 6.86

Financial Times search: tesla
Produces articles From: 2019-12-30
To: 2022-04-15
Results produced from first 20 pages

% of positive titles= 31.11 *
% of negative titles= 16.67 *

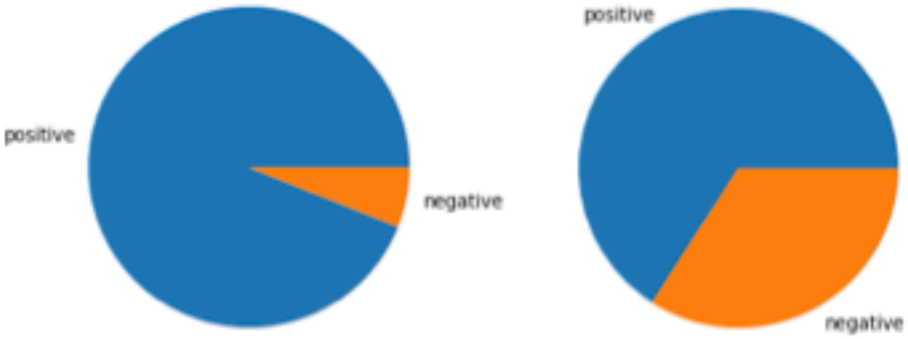


Figure 1: Positive and negative sentiment pie chart for 'Tesla' search on twitter (left) and Financial Times (right)

* It is notable that nearly 50% of articles did not have a positive or negative sentiment score, solidifying the fact that the Financial Times is not sensationalised media.

Twitter search: sp500
From: 2021-02-27
To: 2022-02-26
Ticker search: ^GSPC

% positive = 91.67
% negative = 8.33

Financial Times search: Dow Jones *
Produces articles From: 2012-12-30
To: 2022-04-15
Results produced from first 20 pages

% of positive titles= 38.68
% of negative titles= 23.51

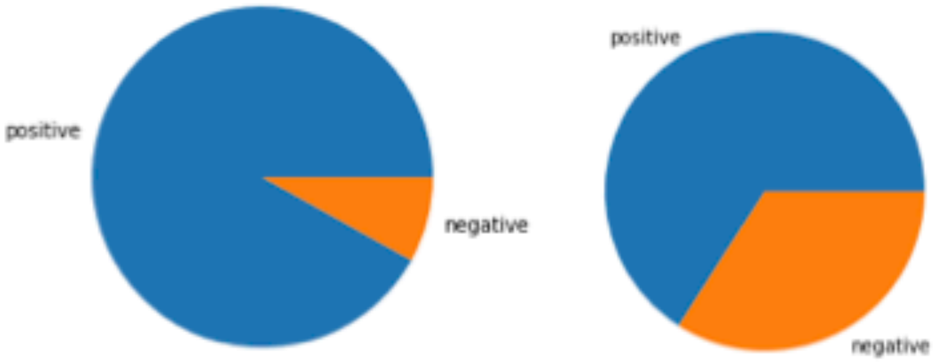


Figure 2: Positive and negative sentiment pie chart for sp500 search on twitter (left) and Financial Times (right)

* Surprisingly, very few financial times articles mentioned the S&P 500, therefore 'Dow Jones' was used to convey the sentiment of the market as a whole.

Twitter search: bitcoin

From: 2021-02-27

To: 2022-02-26

Ticker search: BTC-USD

% positive = 99.17

% negative = 0.55

Financial Times search: bitcoin

Produces articles From: 2019-10-30

To: 2022-04-15

Results produced from first 20 pages

% of positive titles= 18.50

% of negative titles= 26.37

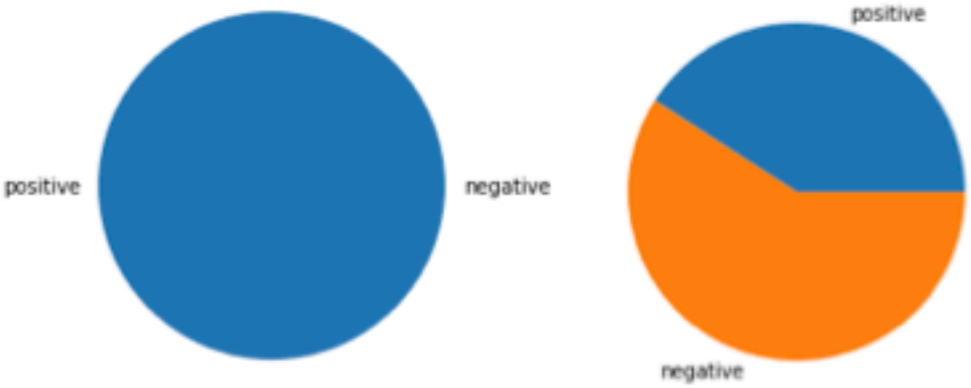


Figure 3: Positive and negative sentiment pie chart for bitcoin search on twitter (left) and Financial Times (right)

Section 5: Information signal from textual analysis

The basis of this project is an accurate curation of public sentiment based on tweets and headlines from the Financial Times. Tweets and articles were scored between 0 and 1 for positive, negative, neutral and compound sentiments. Compound indicates whether the statement is overall negative or positive. If it has negative value then it is negative, if it has positive value then it is positive. If it has value 0, then it is neutral [14]. Once the sentiment scores of each of the tweets each day were calculated, these scores were averaged and compiled to give a daily average sentiment score (see figure 1 above). Similarly, for the Financial Times, if any given day had more than one article published about the chosen asset, the sentiment scores for that day were averaged to present a daily average sentiment score. These average sentiment scores were used in conjunction with the asset prices for each day, which were presented in a column in the dataframe.

Section 6: Machine Learning modelling - methodology, analysis and results.

The machine learning modelling, a random forest regression model, utilised a dataframe composed of eight column headers: 'Date', Tweets, 'Comp', 'Negative', 'Neutral', 'Positive', 'SentAvg', 'Prices'. The data set was split between a training dataset, used to inform the model, and a testing dataset, used to test the predictions made by the model on data not 'seen' during training[17]. Each of these subsets contained the 'Positive' and 'Negative' columns for their respective data. A second pair of datasets was created with the price of the chosen commodity/stock for the duration of the training and testing datasets.

The first supervised learning algorithm used was the random forest algorithm[19]. Inputting the training dataset of the positive and negative sentiments and the prices allowed the model to construct a multitude of decision trees based on the training data to output mean/mode prediction based on the individual trees.

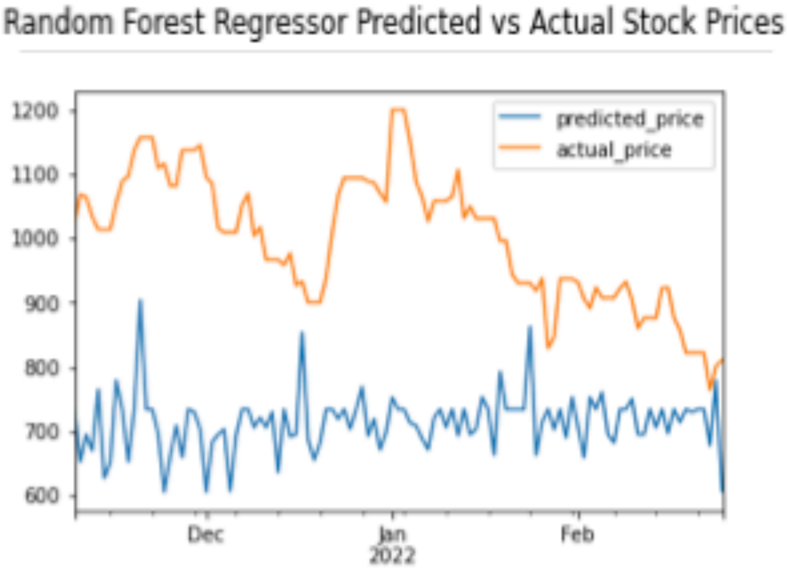


Figure 4: Random forest model predicting TSLA price based on twitter sentiment

It could be said that the overall sentiment from the tweets regarding Tesla, deemed the fair value share price to be in the $650 - $750 range. The predicted value, trained on unseen data, remains roughly within this range, and the actual share price drops steadily over the time frame to within $100 dollars of the predicted value. Although the regression model produces a score of only 0.297, suggesting that only ~30% of the data fit the regression model, the convergence of the predicted and actual price is hard to ignore.

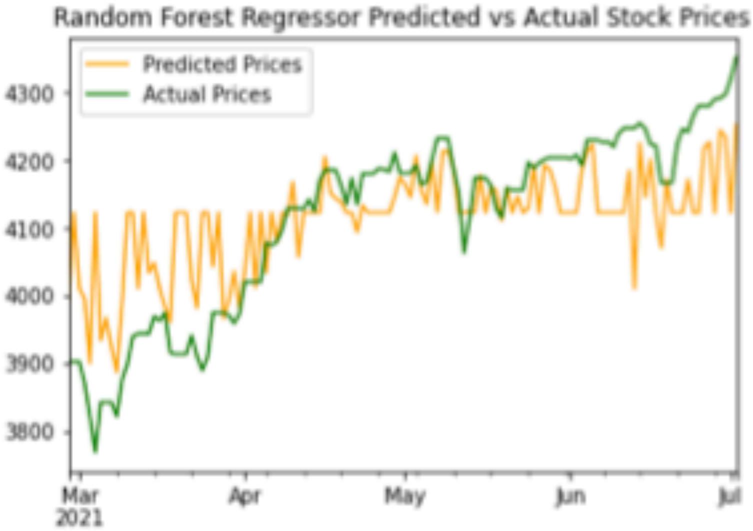


Figure 5: Random forest model predicting S&P500 price based on twitter sentiment

Predicting the S&P 500 (^GSPC) produced the highest R^2 score of our analysis, 0.437. Our hypothesis is that, given the S&P 500 is primarily held by institutions, and by its very essence not a single company, it is less susceptible to the influence of individuals and therefore more easily predictable.

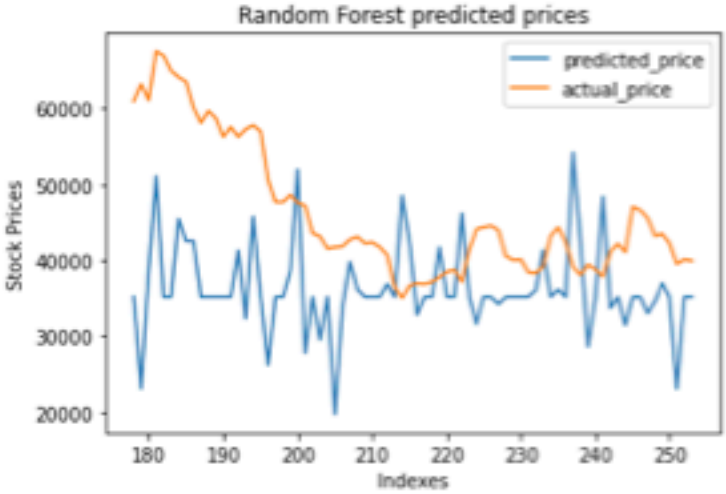


Figure 6: Random forest model predicting BTC-USD price based on Financial Times sentiment

Taking the sentiment of bitcoin from the financial times produced figure 4 above. Although the financial times would not be classified as a sensationalised news outlet, it has a primarily negative sentiment when it comes to the topic of bitcoin. This negative sentiment may be justified when compared to the produced graph of the predicted bitcoin (BTC-USD) price. The bitcoin value predicted using the Financial Times' sentiment consistently has a value of ~$35000, consistently throughout the time period analysed. The actual value can seemingly be seen to begin to converge towards this price prediction.

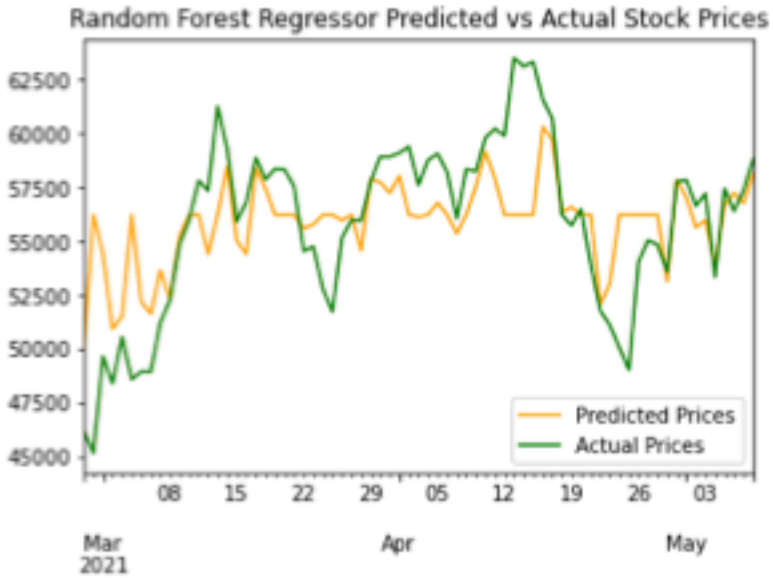


Figure 7: Random forest model predicting BTC-USD price based on Twitter sentiment

The same sentiment analysis for bitcoin was conducted using Twitter sentiment. The price prediction (figure 5) using twitter sentiment is slightly more varied than the Financial Times, however it seems to equally predict the price of bitcoin over the course of a few months. This regression analysis produced a score of 0.345.

Finally, we conducted a prediction using an elementary analysis, graphing the 30-day daily rolling average sentiment and graphing it against the share/commodity price. The aim with this was to step away from the technical analyses and attempt to correlate pure public sentiment to price movements of different asset classes. The 30-day rolling average sentiment for Tesla on Twitter was graphed against its share price for a year long period. This simplistic analysis produced the following graph.

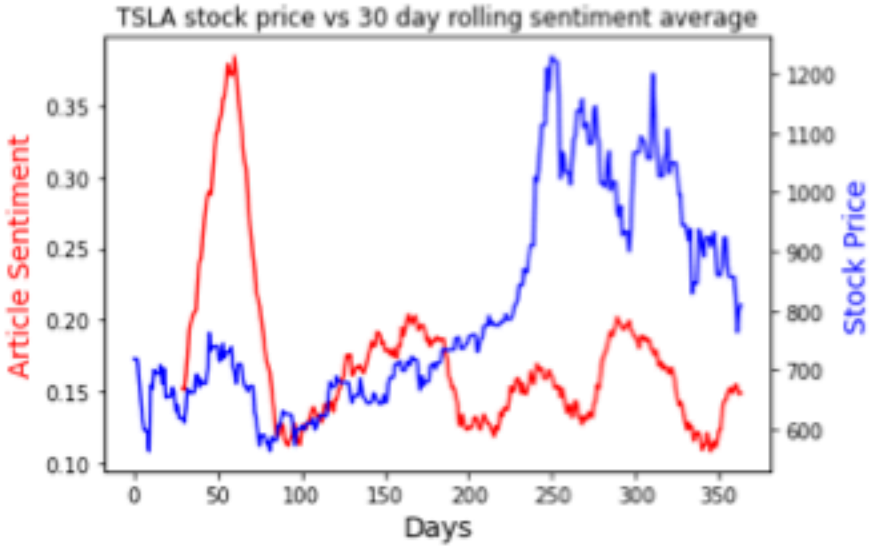


Figure 8: Twitter sentiment for Tesla (TSLA) plotted against its stock price

It is noteworthy that this graph is not a stock prediction graph, simply a graph of the average sentiment score from tweets about Tesla.

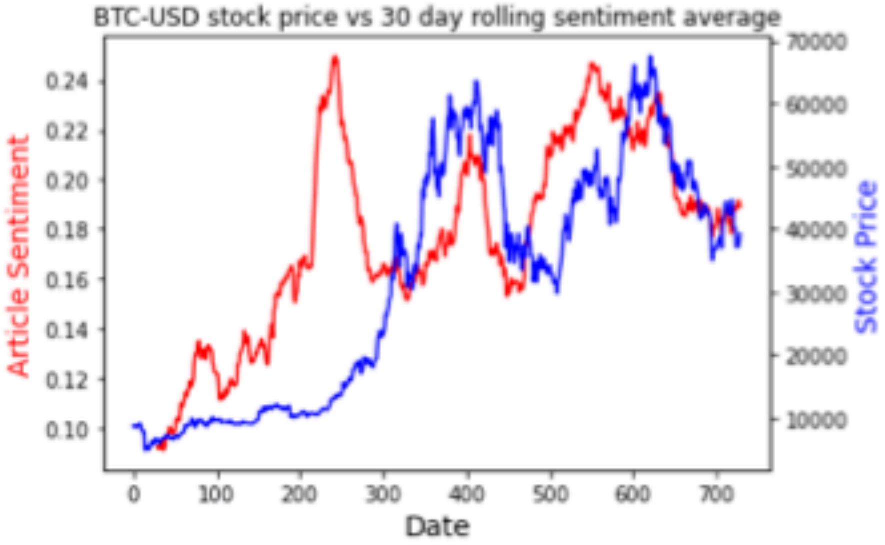


Figure 9: Twitter sentiment for bitcoin (BTC-USD) plotted against its futures price

It is clear to see that at times sentiment seems to drive price, and at other times price seems to drive sentiment. There are some notable peaks within the graph that suggest a rally in public sentiment precedes a rally in the price of bitcoin, which would not be outside of the realm of possibility considering bitcoin is primarily held by retail investors, and many prolific bitcoin holders are outspoken on Twitter too.

Section 7: Business Analysis - what is the impact of your findings

We have demonstrated in this report that a correlation exists between the rise/fall in an asset's price and public sentiments about any given asset on Twitter. The implementation of a sentiment analysis - that can judge the sentiment existing in a tweet - is the key contribution of our study. We mentioned at the outset that positive public sentiment on Twitter about an asset could reflect in its price, our hypothesis is supported by the findings.

As noted in the results above, the S&P 500 scored highest in the Random Forest Regression. One possible reason for this is its lower volatility than Tesla and/or Bitcoin due to its instionalisation. Additionally, the Asset Price vs 30 day rolling average sentiment also proved to be a useful area of research, particularly in Bitcoin's case due to its substantially lower institutional ownership. Indicating that sentiment analysis about an asset popular primarily amongst retail investors is better suited to predicting said asset, e.g. bitcoin, than it is at predicting assets popular with large institutional investors, e.g. S&P 500.

Trading strategies developed based on this analysis could prove profitable and would be an interesting area of research to delve deeper into in future work.

Bibliogrpahy

Briggs, J., 2020. *Sentiment Analysis for Stock Price Prediction in Python.* [online] Medium. Available at: https://towardsdatascience.com/sentiment-analysis-for-stock-price-prediction-in-python-bed40c65d178 [Accessed 24 March 2022].

Pagolu, V., 2016. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. Available at: https://arxiv.org/pdf/1610.09225.pdf [Accessed 4 April 2022].

STOCK-PRICE-PREDICTION-USING-TWITTER-SENTIMENT-ANALYSIS/STOCK PREDICTION USING TWITTER SENTIMENT ANALYSIS PROJECT (FINAL) - Updated.ipynb at master · anubhavanand12qw/STOCK-PRICE-PREDICTION-USING-TWITTER-SENTIMENT-ANALYSIS. [online] GitHub. Available at: https://github.com/anubhavanand12qw/STOCK-PRICE-PREDICTION-USING-TWITTER-SENTIMENT-ANALYSIS/blob/master/STOCK%20PREDICTION%20USING%20TWITTER%20SENTIMENT%20ANALYSIS%20PROJECT%20(FINAL)%20-%20Updated.ipynb [Accessed 10 April 2022].