CRYPTO-CURRENCY PRICE ANALYSIS

GROUP MEMBERS:

DANIIL VOLOVIK (300313470),
DO MAN UYEN NGUYEN (300318626),
FABIO TURAZZI (300318010)

TABLE OF CONTENT

1. Statement of problem & Project Goal

Cryptocurrency has been growing in popularity and relevance, and Twitter can be accredited as one of the most active mediums for crypto-enthusiasts to communicate.

Our goal for this project is to examine the relationship between tweets’, reflecting public opinion, and the daily price of popular e-coins, applying Machine Learning techniques to predict price fluctuations.

2. Project tasks

1

Scrape tweets

- Scrape Twitter searching for relevant keywords and hashtags;
  - Tool: Twint library.

2

Scrape coins' prices

- Scrape coins' daily prices from relevant cryptocurrencies:
  - Tool: CoinGecko API.

3

Perform Sentiment Analysis on Tweets

- Process/clean tweets and perform sentiment analysis
  - Tools: Twitter-preprocessor and VADER libraries

4

Build predict models

- Train Machine Learning and Deep Learning models to predict prices using tweets
- Establish the best model for prediction.

3.1 Datasets Used

Tweets: contains tweets searched by keywords using cryptocurrency names or abbreviations and the VADER Sentiment Analysis scores

date	time	replies_count	retweets_count	likes_count	coin_name	tweet	neg	neu	pos	compound
2013-03-29	13:05:03	0	12	18	cryptocurrency	Bitcoin: The Cyberpunk Cryptocurrency http://...	0.0	1.000	0.000	0.0000
2013-04-15	10:37:11	3	34	17	cryptocurrency	Bitcoin Isn't the Only Cryptocurrency in Town ...	0.0	1.000	0.000	0.0000
2013-04-14	18:34:04	15	153	61	cryptocurrency	#Bitcoin, a "cryptocurrency", went on a tear l...	0.0	1.000	0.000	0.0000
2013-04-18	14:04:24	2	27	26	cryptocurrency	I'm going to make my OWN crypto-currency and e...	0.0	0.885	0.115	0.3810
2013-05-09	15:14:19	4	66	30	cryptocurrency	Your momma's cryptocurrency is so virtual, she...	0.0	1.000	0.000	0.0000
...	...	...	...	...	...	...	...	...	...	...
2019-12-16	9:18:10	27	14	21	ripple	...برو #linkup پیامرسان ها درآمد سرشاری دارند اما	0.0	0.886	0.114	0.5972
2020-10-04	18:36:31	7	3	50	ripple	Hi 25 cent #xrp 🤑	0.0	0.727	0.273	0.4588
2020-10-27	15:30:23	2	3	47	ripple	$ocean #ALLTHEBANKS $Ocean $ewt $dot $qnt $xr...	0.0	1.000	0.000	0.0000
2020-11-21	8:28:39	2	0	75	ripple	Hi .42 cent #xrp 🤘🧡🤛✅🍺				
2021-01-27	14:22:44	3	0	25	ripple	@xrp_mami @BloombergAsia An entire continent...				

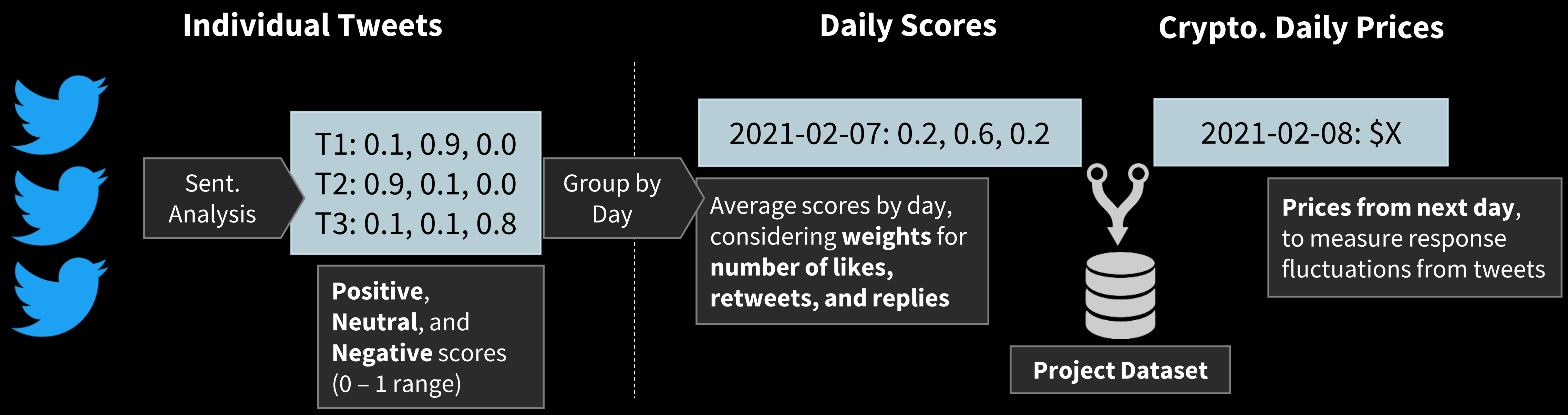
S.A. scores

	coin_name	price	market_cap	total_vol	date
0	cardano	0.659472	2.091005e+10	8.483795e+09	08-02-2021
1	cardano	0.626357	1.985498e+10	6.225004e+09	07-02-2021
2	cardano	0.538552	1.720209e+10	5.138348e+09	06-02-2021
3	cardano	0.441599	1.412487e+10	2.526990e+09	05-02-2021
4	cardano	0.441216	1.404678e+10	2.963647e+09	04-02-2021
...	...	...	...	...	...
12812	yearn-finance	3793.033675	1.144837e+08	8.185937e+06	04-08-2020
12813	yearn-finance	4063.531281	1.212525e+08	9.198283e+06	03-08-2020
12814	yearn-finance	3863.416015	1.156429e+08	9.230661e+06	02-08-2020
12815	yearn-finance	4128.207821	1.235146e+08	1.945850e+07	01-08-2020
12816	yearn-finance	4367.882143	1.307421e+08	1.688721e+07	31-07-2020

Cryptocurrency prices: contains daily prices, market cap, and total volume from coins, scraped using CoinGecko API.

- Only daily data found for public access, so we combined different coins for enough data to train more robust models.
- Using hourly data to train individual models for each currency would ideal for practical use in price prediction.

3.2 Combining the Datasets



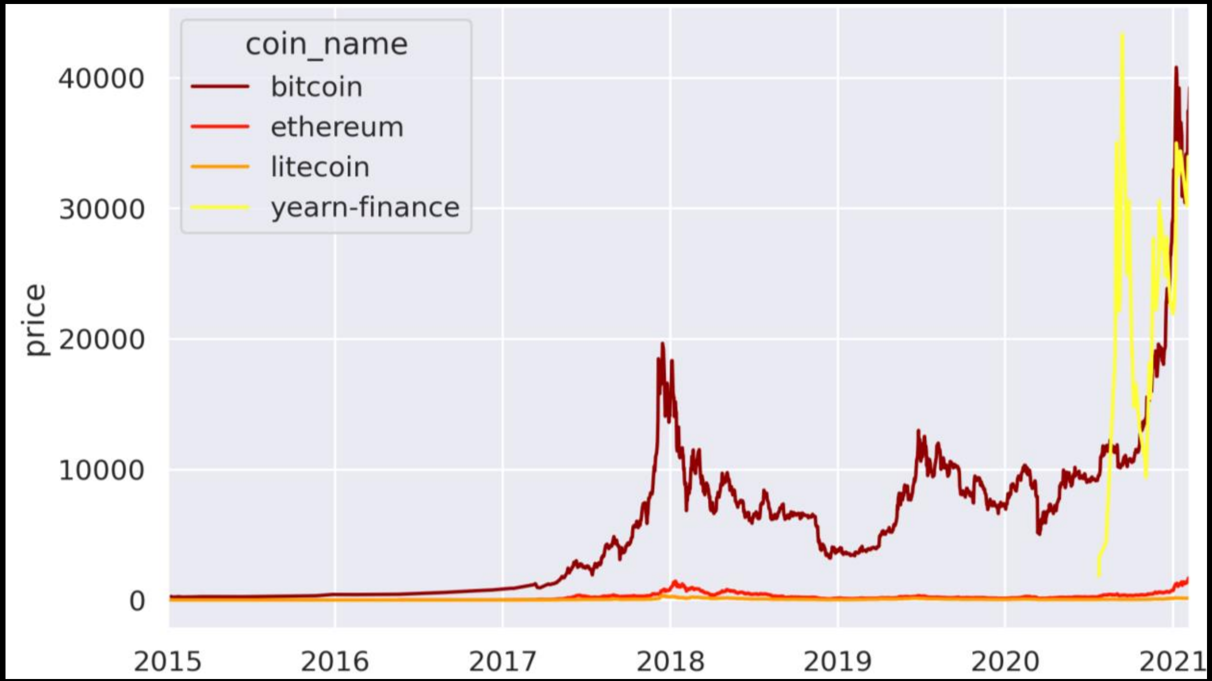
- Tweets cleaned (**ftfy** and **preprocessor** libraries) and used for Sentiment Analysis (**VADER** library).
  - Assigned a '**positive**', a '**negative**', and a '**neutral**' score ranging from **0 to 1**.
  - Scores were aggregated daily, considering **weights** for **number of likes (0.2), retweets (0.7),** and **replies (0.1).**
- Aggregated tweets were **joined** with **next day cryptocurrency prices** to create the project's dataset.

3.3 Dataset Features

- The combined dataset displayed below contains the following features:
  - **total_vol**: total volume of coin in the market
  - **date**: date of the information
  - **price**: price observed on the day following the listed date
  - **negative, neutral, positive**: Sentiment Analysis scores
  - **total_tweets**: number of tweets collected for the coin is a given date
  - **coin_name**: corresponding cryptocurrency (encoded)

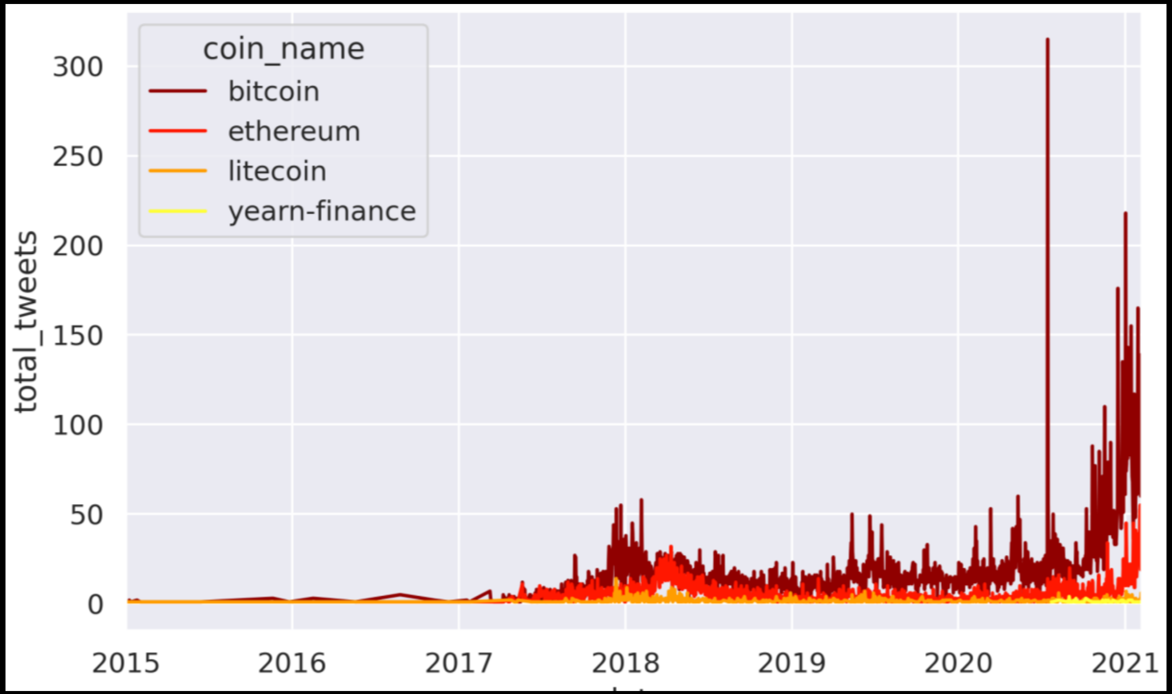
	total_vol	date	price	positive	negative	neutral	total_tweets	bitcoin	litecoin	yearn-finance
0	4.744626e+10	737827	39279.41287	0.084205	0.066617	0.849145	87.0	1	0	0
1	5.449481e+10	737826	38007.83223	0.081577	0.075804	0.842591	139.0	1	0	0
2	4.976214e+10	737825	36816.50808	0.065362	0.010455	0.924175	60.0	1	0	0
3	5.073070e+10	737824	37494.71762	0.102012	0.023837	0.874128	80.0	1	0	0
4	4.926886e+10	737823	35485.98593	0.185786	0.009167	0.805051	77.0	1	0	0

3.4 Coins' Prices Overview



- The line graph: prices over time for the four coins.
- All coin prices seem to have a positive growth trend over time;
- BTC and YFI prices are significantly higher than others, with the latter presenting a very steep increase in recent years.
- We understand that this discrepancy in behavior may impact our models' predictive capability, but this issue could only be fully solved by gaining access to more granular data to run individual models for each coin.

3.5 Coins’ Number of Total Tweets

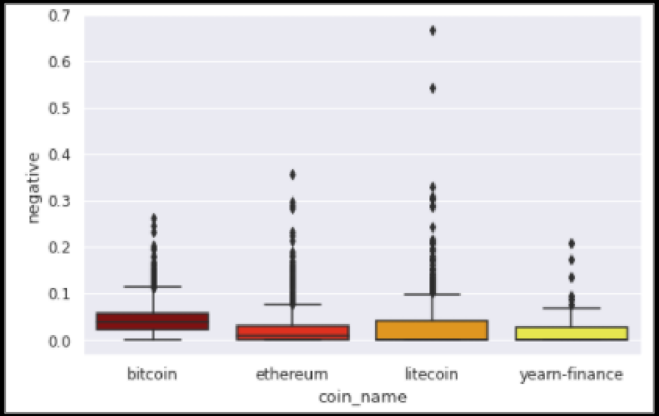


- Line graph: number of tweets over time for the four coins.
- Cryptocurrency number of tweets also present a positive trend, with Bitcoin’s popularity again being predominant among other coins.
- For the analysis, we experimented with using a minimum threshold of tweet’s likes filter out noise, selecting a value that optimized our model results.
  - Threshold: minimum of 200 likes

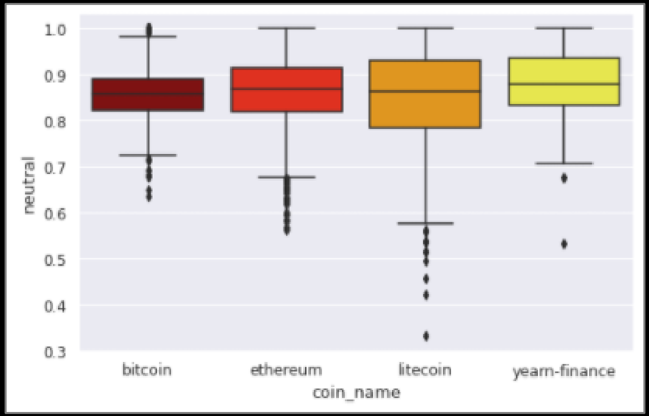
3.6 Coins’ Sentiment Score Distribution

- Graphs: Distribution of coins’ sentiment scores.
- Tweets generally received higher **Neutral** scores, with most of the sample presenting a score above 0.7.
- **Positive** scores presented were generally higher than **Negative** ones, indicating that positive tweets were more frequent.

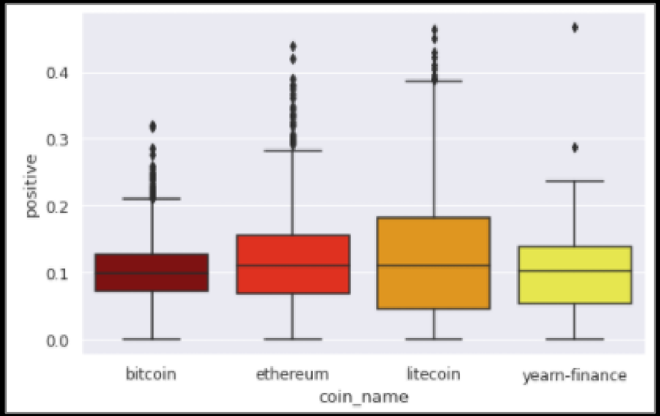
Negative



Neutral



Positive



4.1 Data Preparation

- In addition to the initial dataset preparation, the following actions were performed to prepare for modelling:

Create Dummy Variables

Separate categorical variable coin_name into corresponding Dummies for each coin.

Ethereum Dummy was dropped to avoid redundancy.

Rescale Data

Scale the data using MinMax Scaler and Standard Scaler.

Performance on the models was later compared for both scaled and unscaled versions of the dataset.

Split Data

Split data into training and testing sets with a 75:25 ratio.

4.2 Model Definition

- The following Machine Learning models were experimented with to find the best prediction performance:
  - **Linear Regression** (with and without **Lasso** regularization)
  - **Polynomial Regression** (with and without **Lasso** regularization)
  - **Neural Network** (Keras Sequential model)
  - **Decision Tree** and **Random Forest Regressors**
- All models had hyperparameters tuned using 5-fold cross-validation before applied on the testing set to validate the performance.

Features:
Unscaled,
MinMax Scaled,
Standard Scaled

5-fold CV
(tuning)

Tuned Models:
Linear & Polynomial Regression (w/ Lasso)
Neural Network – Keras Sequential model
Decision Tree and Random Forest
Regressors

5. Model Performance Table

Model Performance

	Model	R2_Test	R2_Train	RMSE	Scaling	Params
0	RandomForest Regressor	0.988675	0.994798	708.980808	standard_scaler	max_depth = 9
1	RandomForest Regressor	0.987454	0.994859	746.232236	no_scaling	max_depth = 9
2	RandomForest Regressor	0.985835	0.994147	792.912607	min_max_scaler	max_depth = 9
3	Decision Tree Regressor	0.985795	0.995127	794.034003	no_scaling	max_depth = 9
4	Decision Tree Regressor	0.985509	0.995127	801.978243	standard_scaler	max_depth = 9
5	Decision Tree Regressor	0.984582	0.991914	827.226309	min_max_scaler	max_depth = 8
6	Polynomial Regression (Lasso)	0.879234	0.884052	2315.189461	no_scaling	Alpha = 0.1; Degree = 2
7	Polynomial Regression (Lasso)	0.878033	0.889425	2326.670995	min_max_scaler	Alpha = 0.1; Degree = 2
8	Polynomial Regression (Lasso)	0.878033	0.889425	2326.670995	min_max_scaler	Alpha = 0.1; Degree = 2
9	Polynomial Regression (Linear)	0.872786	0.894575	2376.189204	standard_scaler	Degree = 2
10	Polynomial Regression (Linear)	0.872786	0.894575	2376.191137	min_max_scaler	Degree = 2
11	Polynomial Regression (Lasso)	0.868759	0.893497	2413.511008	standard_scaler	Alpha = 0.1; Degree = 2
12	Polynomial Regression (Lasso)	0.868759	0.893497	2413.511008	standard_scaler	Alpha = 0.1; Degree = 2
13	Polynomial Regression (Linear)	0.851140	0.811729	2570.417580	no_scaling	Degree = 2
20	Neural Network	0.840174	0.814507	2663.408036	standard_scaler	See NN topology
14	Lasso Regression	0.829831	0.814503	2748.241563	standard_scaler	Alpha = 10.0
15	Lasso Regression	0.829817	0.814526	2748.353635	no_scaling	Alpha = 0.0
16	Linear Regression	0.829807	0.814555	2748.432678	min_max_scaler	NaN
17	Linear Regression	0.829807	0.814555	2748.432678	standard_scaler	NaN
18	Linear Regression	0.829807	0.814555	2748.432678	no_scaling	NaN
19	Lasso Regression	0.824452	0.812382	2791.336831	min_max_scaler	Alpha = 10.0
21	Neural Network	0.823079	0.824801	2802.228394	min_max_scaler	See NN topology

- Model Performance table containing the following information:
  - Model & Scaling method used;
  - Tuned hyperparameters;
  - R2 performance on Testing/Training sets;
  - RMSE obtained.

- **Best model**: **Random Forest Regressor** using **Standard Scaler** and setting tree max depth to 9.
  - Approximate accuracy of **98.9%** on the testing set, with an **RMSE** of **708.98.**

6. Price Prediction Demonstration

- Random Forest Regressor does not output meaningful regression coefficients for us to break down price, but the model can be demonstrated by making predictions.
- We evaluated two instances of contrasting scenarios, using Bitcoin with fixed volumes and dates:

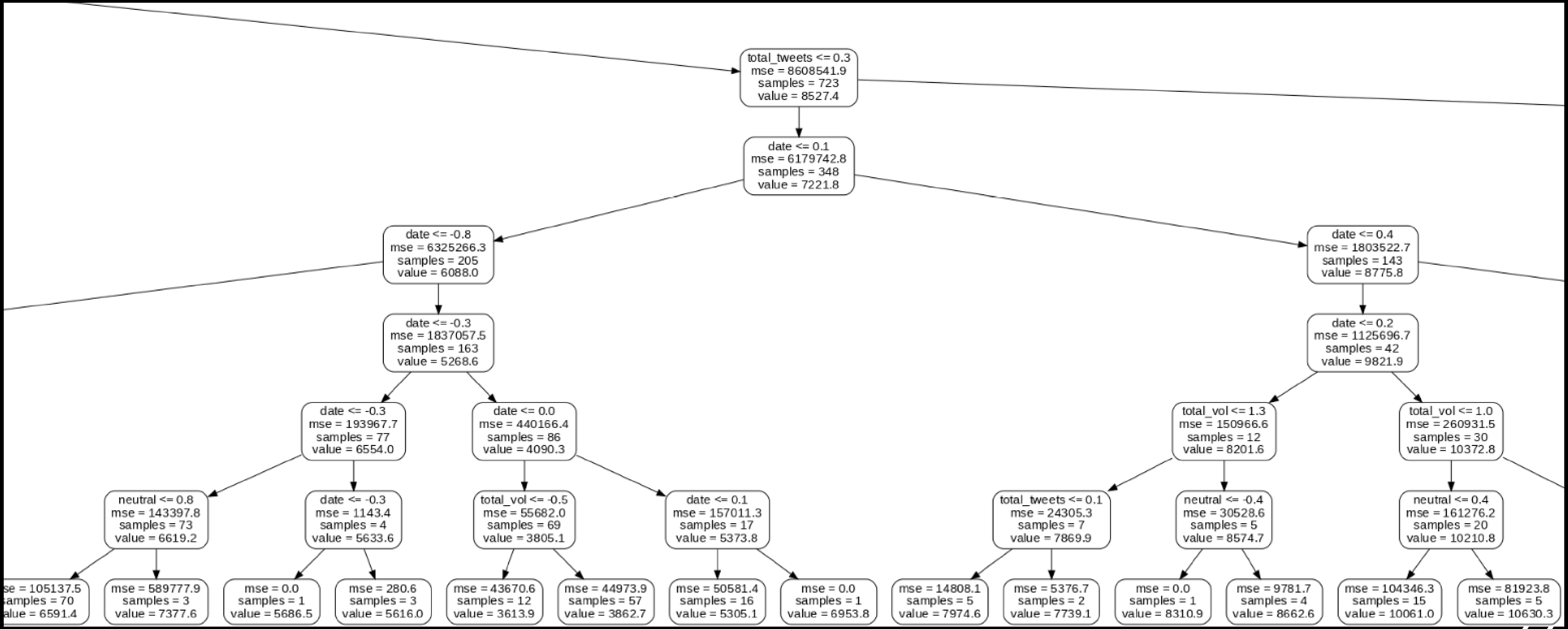
```
# Creating a dataframe for new predictions
data_new = [
    # Positive tweets scenario
    [150000000000, '08-02-2021', 0.9, 0, 0.1, 100, 1, 0, 0],
    # Neutral tweets scenario
    [150000000000, '08-02-2021', 0, 0, 1, 100, 1, 0, 0],
    # Negative tweets scenario
    [150000000000, '08-02-2021', 0, 0.9, 0.1, 100, 1, 0, 0],
    # Large number of tweets
    [150000000000, '08-02-2021', 0, 0, 1, 500, 1, 0, 0],
    # Small number of tweets
    [150000000000, '08-02-2021', 0, 0, 1, 10, 1, 0, 0]
          ]
```

Predictions

Positive Tweets	Neutral Tweets	Negative Tweets	Many Tweets	Few Tweets
36745.2	36410.6	36343.9	36726	35827.2

- **Tweets sentiment**
  - Positive Scenario – 0.9 pos, 0.1 neu, 0.0 neg
  - Neutral Scenario – 0.0 pos, 1.0 neu, 0.0 neg
  - Negative Scenario – 0.0 pos, 0.1 neu, 0.9 neg
- **Tweets quantity**
  - Large number – 500 daily tweets
  - Small number – 100 daily tweets
- **Predictions**
  - All else equal, a day with larger positive tweet score rendered higher price predictions than a negative one.
  - Large volume of daily tweets also affected price positively

6.1 Decision Tree Example



7. Summary & Final Remarks

Project Summary:

- In this project, we successfully implemented Sentiment Analysis to extract public opinions regarding cryptocurrency and used Machine Learning to determine how they impact cryptocurrency price fluctuations.
  - The best model predicted prices with a 98.9% accuracy in our testing dataset.
- We made hypothetical predictions showing that our model correlates positive tweets and high number of mentions with positive price fluctuations.

Additional Considerations:

- Although our model has performed well overall, the predicted prices are still slightly offset from Bitcoin’s current price level.
  - We believe that this limitation for the models is being caused by our necessity to combine different currencies to generate sufficient data for the ML models;
  - To further improve predictions for practical use, we believe that access to more granular price data would be beneficial, rendering enough datapoints to run separate models for each coin.