MGM7205o Assignment

Cryptocurrency Forecasting from a Reliable Social Media Network

Manoj Kumar

Contents

HAT based Prioritization Approach

GREEN HAT				Rank
Field	**Use Case**	**Pros**	**Cons**	
Media	Infodemic of COVID-19	Epidemic Analysis Identify reliable and questionable sources and its spread on social media.	Very sensitive use case and needs high accuracy. To acheive high accuracy, need large data collection	
	Removal of BOTS in Social Media Network	Helps to remove untrusted information which helps in multiple use cases which consumes data from social media	Labels for training model	2
Blockchain	Twitter sentiment analysis for Crypto price prediction indicator	Quick summary of hot topics and market news.	Crypto news phishing and bot accounts	1
	Crypto price forecasting based on algorithmic trading	Automated portfolio management with reduced human emotions effect	Highly volatile	4
E-Commerce	AR Recommendation System	Drives the customer with more decision power	Privacy, ethical and complex	3
	Price Comparison model for ecommerce platforms	Provide quick overview of product ranges and prices at single click	Data availability and integration is complex	
	AI chatbot to personalise the customer needs.	Customer can place an order based on their choices	Procurement, supply chain	

Introduction

Cryptocurrencies have become a very popular topic recently, primarily due to their disruptive potential and reports of unprecedented returns.

CRYPTOCURRENCY IS THE NEW ASSET

- If you have missed the internet revolution to make money, then buy crypto.

CHOOSE THE RIGHT CRYPTOCURRENCY

- Cryptocurrency can make someone rich or lose his savings in very short interval. Very volatile market.

FACTORS DRIVING CRYPTO PRICES

- Each cryptocurrency depends on factors like application, transparency, accessibility, reputation of the founder, networks trustworthiness etc

SOCIAL MEDIA IS THE CRYPTO BAZAAR

- Every other person is talking about crypto by posting “Buy that, Buy this, Sell this”. Ex, Elon Musk (Bull of the crypto market)

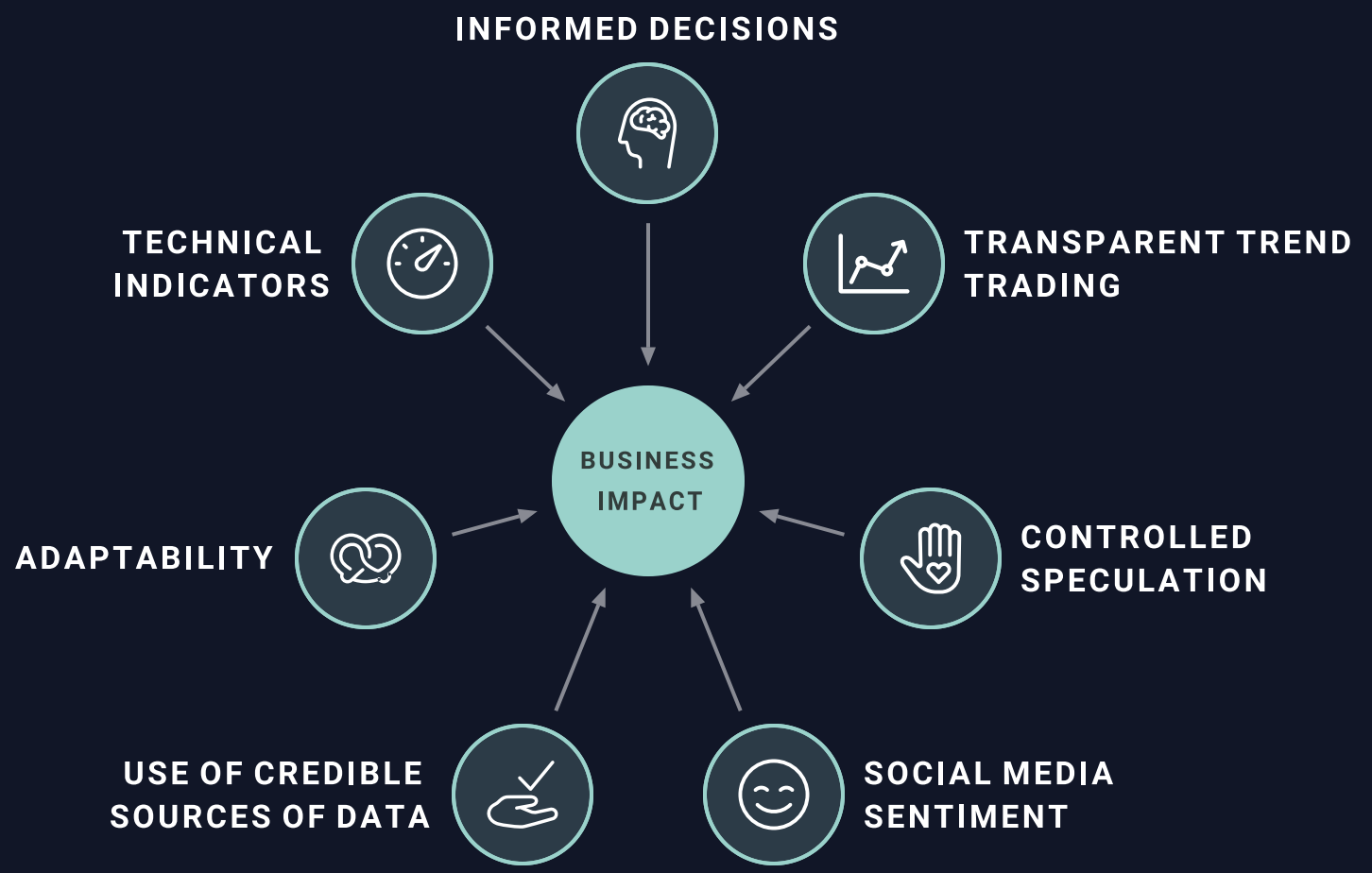
FAKE NEWS & BOTS ON SOCIAL MEDIA

- An analysis showed about 14-15% of tweets posted by bots.*

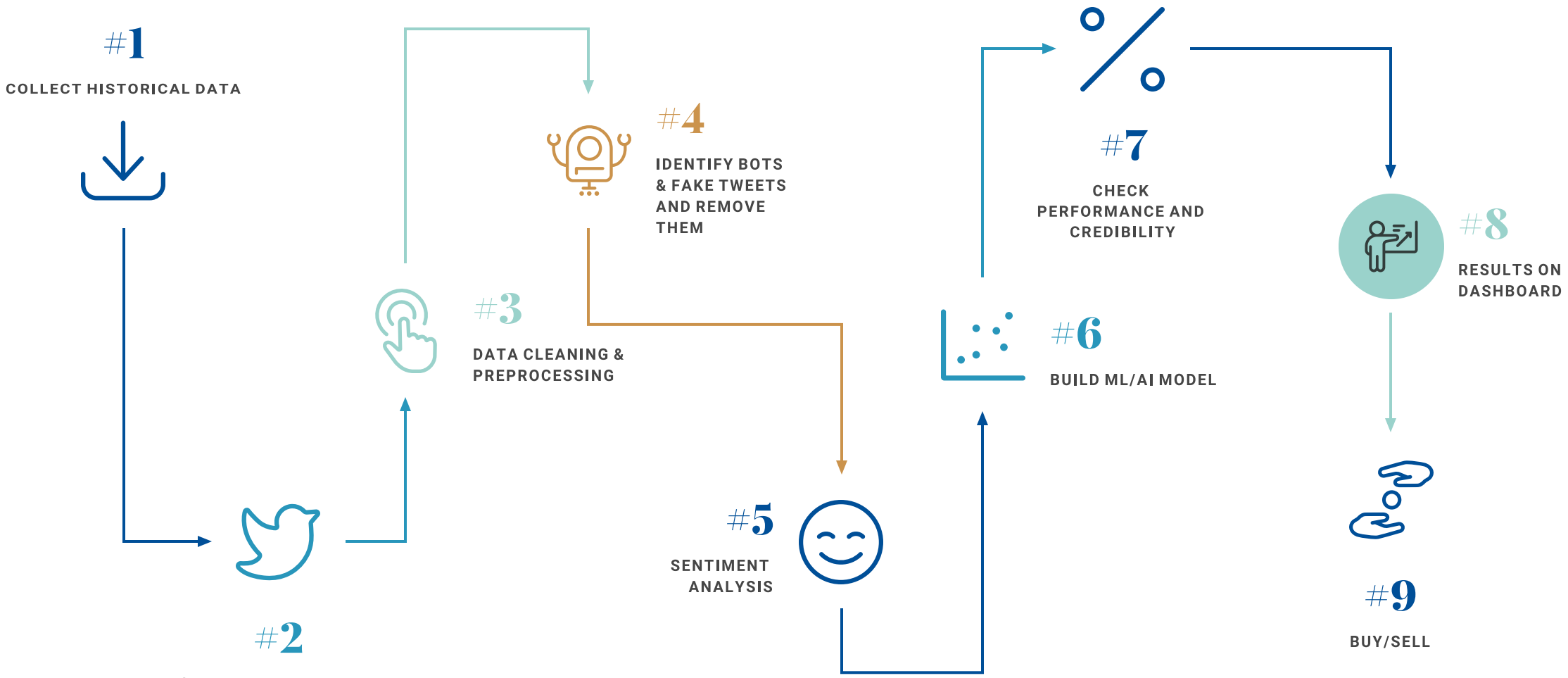
LEVERAGING ML/AI TO FORECAST

- AI/ML has the power to use both historical finance data and unstructured text data to forecast exact prices or buy/sell signal.

Business Impact

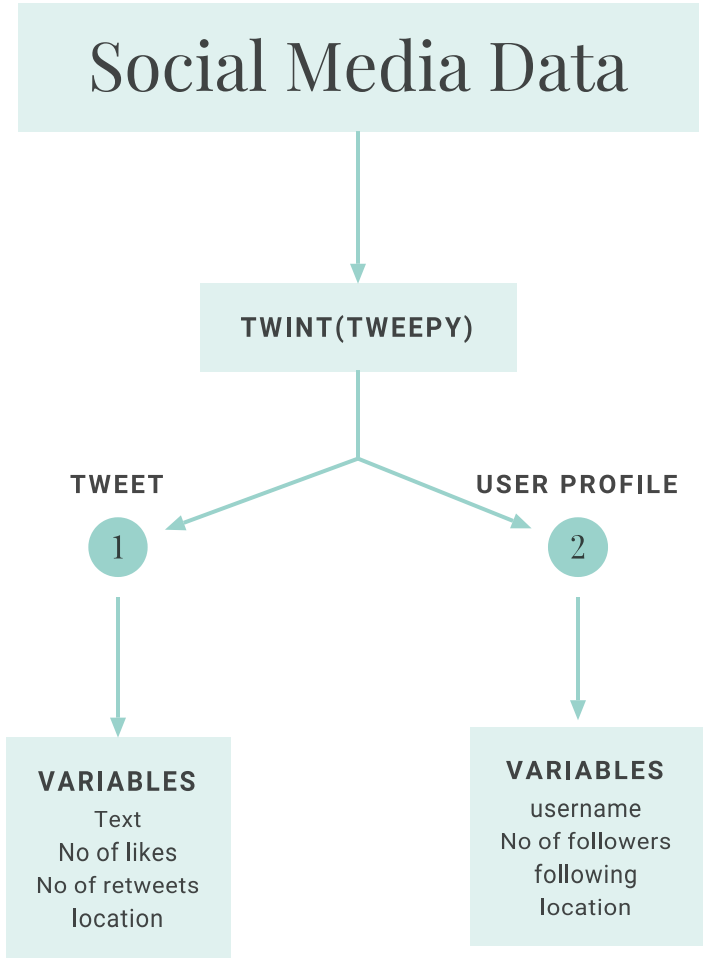


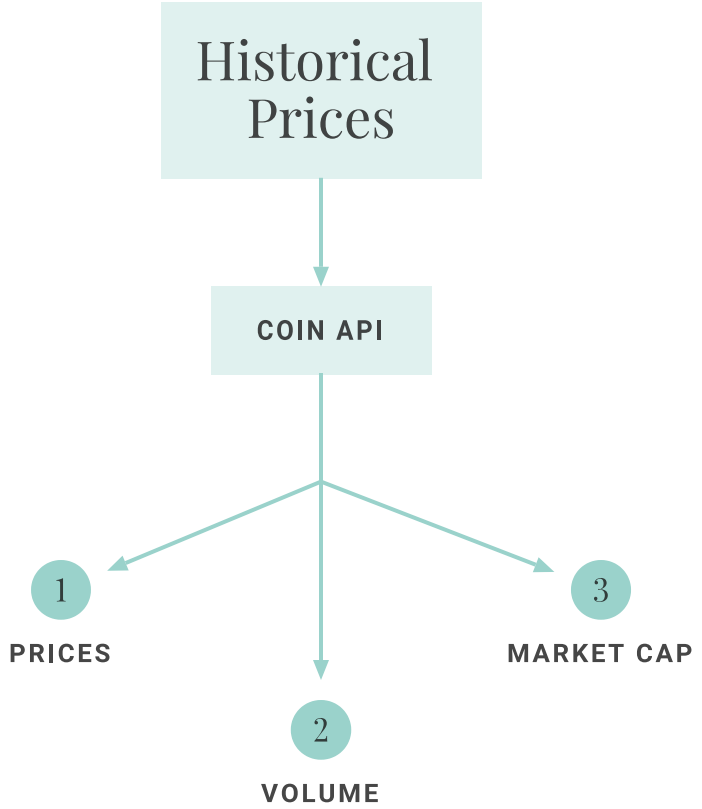
Solution Methodology



Data Collection

This shows how and what data is collected ?





Data Cleaning

- Remove "RT"(Re-tweet) if present
- Remove URLs, excess (white) space and mentions
- Reduce character sequences >3 to 3
- Apply case-folding
- Remove Tweet if number of tokens < 4
- Remove hashtags
- Expand contractions
- Handle slang and/or acronyms
- Remove ticker symbols
- Remove tokens with numerical characters
- Apply WordNet lemmatisation

  Remove stopwords and custom list

Removal of Bot Accounts & Fake tweets

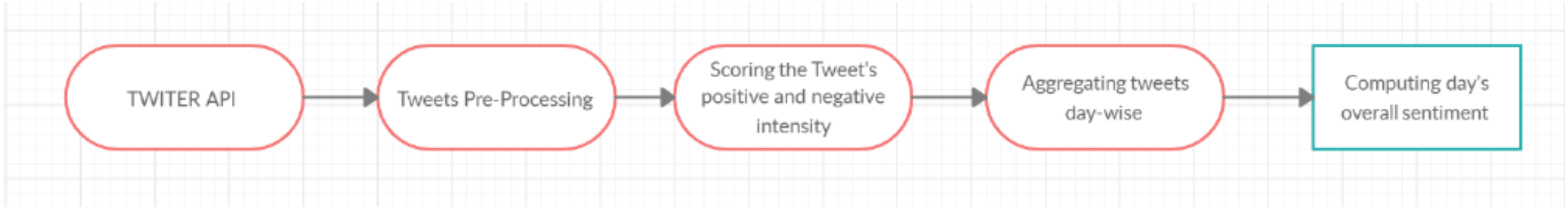
- Manually label sample of data with equal bots and real users.
- Feature engineering of user and tweets for further classification.
- Below features will be extracted:
  1. Use of red list words like "FREE, GIVE AWAY, PROFIT, 100%"
  2. Tweets containing more than 14 hashtags
  3. Tweets containing more than 14 ticker symbols
  4. Ratio between followed and follower
  5. Platform source of the tweet
  6. Location of tweet

  Tree based ML models, Naive Bayes Approach or Deep Learning Models

Sentiment Analysis:

- **Semantic orientation (SO)** is a measure of subjectivity and opinion in the text. It helps in capturing an evaluative factor and strength of a subject topic or idea. The evaluative factor is whether the opinion is positive or negative, and the strength is the degree to which the word or sentence, in the opinion, is positive or negative. When we use semantic orientation in the analysis of public opinion, like online articles, social media posts, it helps in measuring the popularity and success of that particular topic.

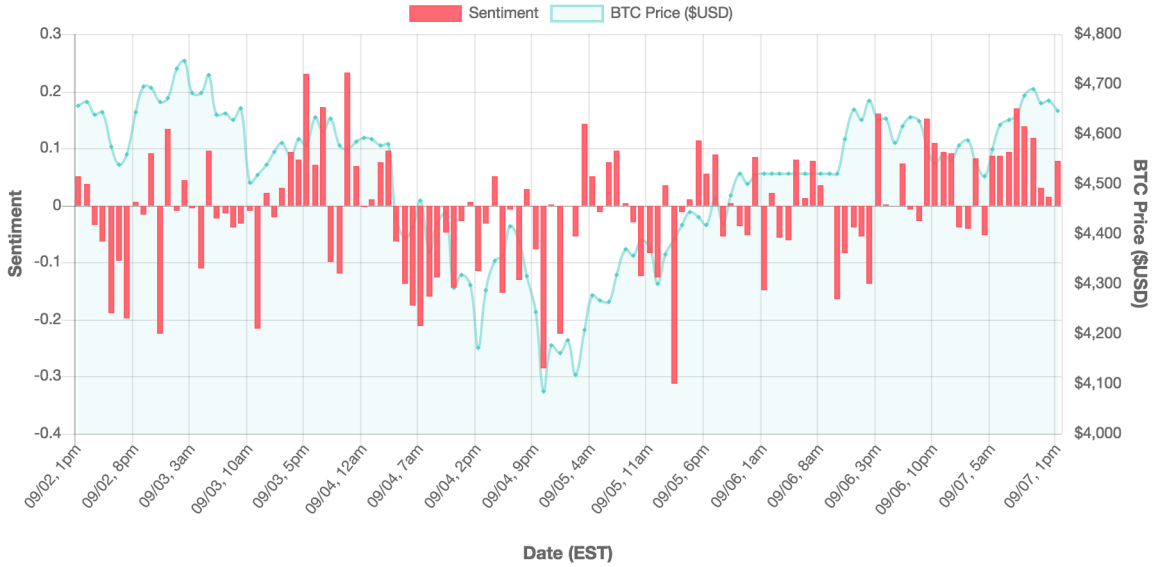
1. **Bag-of-Words (BOW)** : is a representation of the text that describes the occurrence of words within a document.
2. **Word2Vec Approach :** a word can have different meanings when used in a different context; word2vec tries to capture the different contexts of words based on its position in the sentence**.**
3. **Recurrent Neural Network**: RNN uses the information learned in previous states and propagates it forward into the network.
4. **Bidirectional Encoder Receiver from Transformers:** It considers both left and right directions making it better understand the context while embedding the word.

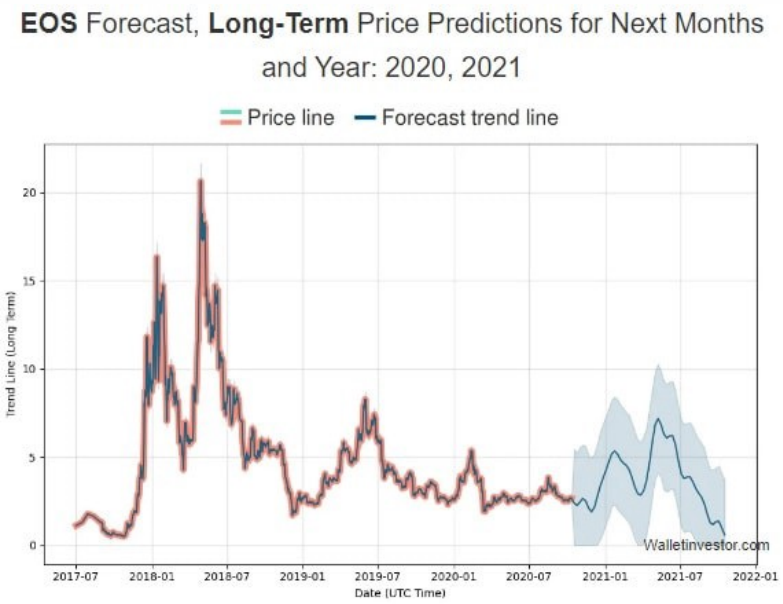


Forecasting Models

- Get top 10 cryptocurrencies based on market capitalization
- Training, Testing & Validation based on backtesting period
- To capture technical indicators we use lag values, moving averages, volume traded etc.
- Below M.L/A.I models to be applied and an ensemble of the best would be used to forecast :
  1. Traditional forecasting models like moving averages etc
  2. RNN based Models
  3. Tree based Regression Models

Dashboard





References

- https://www.sciencedirect.com/science/article/pii/S104244312030072X
- https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1911&context=etd_projects
- https://medium.com/@SamuelCouch/understanding-cryptocurrencies-with-sentiment-analysis-5fc4cf66ec28
- https://economictimes.indiatimes.com/markets/stocks/news/moving-beyond-bitcoin-to-the-next-crypto-revolution-in-2021/articleshow/81773999.cms
- https://www.sciencedirect.com/science/article/pii/S1544612319314199