

CMPT 733: Big Data Programming II (Spring 2019) – Final Project

TradeSpade - Price Signal Forecast for Financial Assets

Anurag Bejju, Rishabh Singh, Nikitha Ravi, Manan Parasher
abejju@sfu.ca, rishabhs@sfu.ca, nravi@sfu.ca, mparashe@sfu.ca

1. Introduction

The decision to *buy* or *sell* stocks and cryptocurrencies is an interesting challenge faced by day traders in today's financial market. Such choices are being made daily in the markets across the globe ^[1]. The whole idea whether the movement of financial assets can be predicted has kept economists, researchers and investors very occupied for decades. There have been two distinct trading philosophies for financial asset market prediction: *fundamental analysis* and *technical analysis*. Technical analysis ^[2] emphasizes more on the study of market actions through the use of charts, fundamental analysis ^[3] concentrates on the economic forces of supply and demand that cause the price to move higher, lower, or remain the same.

Based on these concepts, researchers have most commonly used *Efficient Market Hypothesis (EMH)* ^[4] and *Random Walk Theory (RWT)* ^[4] theories as a base for their research. According to EMH, an asset's prices fully reflect all available information. In other words, market prices should only react to new information or changes instead of the past or present prices. In case of a random walk theory, it is nearly impossible to predict price with an accuracy of more than 50%.

At this stage, we need to understand the assumptions made as part of these theories. The first being - investors are seemed as rational entities who make decisions to buy or sell financial assets. Secondly, it is assumed that all investors have complete information about the market to make an informed choice. But in reality, the two assumptions do not always hold much weight as witnessed during the late 2000 financial crisis. This lead on to scientists taking a more practical and sighted approach, which helped them to gauge financial markets in a more informed manner. They started using news analytics and sentiment analysis to predict the stock and cryptocurrencies prices. With a huge amount of behavioral data available on social media platforms, they were able to use it as a varying metric while predicting price.

With so much research being done by financial analysts and data scientists to recognize patterns in for various financial assets, it propelled us to dwell into the nuances of stock and cryptocurrency markets. As Financial Times reported, profits for financial asset managers rose to \$102bn making it one of the most lucrative market to be in. With more and more people diversifying their asset portfolio and with the crypto market projected to reach US\$6702.1 mn in 2025, *Our product intends to provide day traders assistance with intraday trading by predicting Buy and Sell signals in order to maximize profits and make optimized decisions.*

We would like to target both the *traditional* and *exploratory traders* by providing a robust application that can help them make data-driven decisions as well as actively support novice traders by providing intuitive financial predictions based on historical and contextual information collected for the last one year. We would also like to depict the influence of social media and everyday news on market fluctuations.

2. Motivation and Background

There has been a lot of research to predict price signals for stocks and cryptocurrency assets. Predominantly, historical data of these financial markets are the popular source to build predictive financial models. In a paper published by Shen et al. ^[5], historical index data was transformed into various technical indicators which had great influence to China's stock market. Then they developed a model which considers past information rather than the latest information which has significant impacts on the behavior of the stock market.

Later, researchers like Schumaker and Chen ^[6] started using Financial news articles to propose a predictive machine learning approach for textual analysis using several different textual representations. However, there are some disadvantages for just using financial news as a parameter to predict price. There are a lot of none financial related events that can have an impact on the moment of the price. Then the focus shifted to sentiment analysis of the textual information with the assumption that the sentiment or emotion is one of the vital factors that can influence the stock market. Johan Bollen et al. ^[7] found that sentiments, such as calm, happiness and Anxiety Index, have a predictive power to inform the broad direction of stock market in the future.

3. Problem Statement

We wanted to provide solutions to the following 3 queries through our project:

- a) *Can we design a model that helps with Stock and Crypto Currency forecast based on features other than just OHLCV?*
- b) *Can we find the impact of global factors on market volatility and derive the correlation between them?*
- c) *Which technical indicators are most important for market direction analysis?*

As we have seen in the background section of this paper, a lot of researchers were either just using financial news or OHLCV or social media for predicting the direction of financial assets. Our team wanted to combine all the above factors that have some evidence of having an effect on the price and find out if a good predictive model can be made. We also wanted to see how *qualitative* (financial and general news) and *quantitative* (Twitter and Reddit) metrics impact the volatility of centralized and decentralized markets. Lastly, we wanted to find the most important indicators that have a great influence in predicting the direction of the market.

4. Data Science Pipeline

The Data Pipeline for our project consists of *5 stages*. Each stage has been briefly explained in the below-mentioned sections.

4.1. Data Collection

The initial stage was also one of the most tedious stages of our project. We gathered almost 8 batches of data from 6 different sources. In the end, we managed to collect a massive *9.71 million records (around 52.4 GB)* worth of data that had to be aggregated, cleaned and processed to build our model. Another key issue that required attention was to not allow the source server to be overloaded with multiple get requests at a time. It could result in blocking of destination IP and stop further requests from being processed. In order to achieve this, we wrote customized scripts for most of the source websites that asynchronously request data. A delay of 30 seconds was added after each request to avoid server being hit with high frequency.

Table 1: Data Collection (1 st March 2018 – 19 th Feb 2019)								(Size > 50 GB)
Category	Stocks				Crypto Currency			
Assets Used	Walmart Inc., Alphabet Apple Inc., Chevron, Exxon Mobil, Microsoft, Coca-Cola, Home Depot, Wells Fargo				Binance Coin, Bitcoin , Bitcoin Cash ,EOS, Ethereum , Litecoin, Stellar, TRON, XRP			
Data Type	Financial	General News	Financial News	Reddit		Twitter		Financial
Data Collected	OHLCV	News, Title, Publisher, Links	Title, Sub Title, Date, Publisher	Score, Subreddit, Title, Comments		Tweet, Comments, Re-Tweets, Likes		OHLCV
Collection Source	finam.ru	Custom News API	Financial Times	Pushshift API Metrics		Twitter API + Custom API for older Tweets.		Crypto-compare
Data Count	Hourly Data	6,875,000 News Articles	25,000 Financial News Articles	1,284,023 Reddits (Crypto)	557,391 Reddits (Stocks)	468,888 Tweets (Stocks)	328,704 Tweets - Crypto	Hourly Data

Here's a brief overview of our data collection strategy. We chose the most volume centralized (Stocks) and decentralized (Cryptocurrencies) financial assets for our experiment.

a) Stocks - OHLCV:

We collected hourly data for the period or almost one year (1st March 2018 – 1st Feb 2019) for 9 stocks in total. The stocks we have analyzed are: *Walmart Inc., Alphabet Apple Inc., Chevron, Exxon Mobil, Microsoft, Coca-Cola, Home Depot, Wells Fargo*. This data was collected by using a python script that collects Open, High, Low, Close and Volume data from *finam.ru*

b) Cryptocurrency - OHCLV:

Similarly, We collected hourly data for the same period 1st March 2018 – 1st Feb 2019 (1 Year) for 9 cryptocurrencies in total. The cryptocurrencies we have analyzed are: Binance Coin, Bitcoin, Bitcoin Cash, EOS, Ethereum, Litecoin, Stellar, TRON, XRP. This data was formed by using cryptocompare api that collected Open, High, Low, Close and Volume data.

c) General News:

To collect general news, we created a python script based on a Custom News API that asynchronously downloads news in batches of 10000. Since the API always gets 10000 articles from the end date specified, the script dynamically updates this value till it reaches 1st March 2018. Using this, close to 6,875,000 News Articles were recorded and parameters like title, subtitle, publisher name, source link, score were collected for the one-year span.

url	counts
medium.com	25352
github.com	19149
www.youtube.com	11251
www.nytimes.com	8121
techcrunch.com	5166
www.bloomberg.com	5026
arstechnica.com	4834
www.theguardian.com	4745
www.theverge.com	4288
en.wikipedia.org	3743
hackernoon.com	3679
twitter.com	3446

Fig 1: Top General News Sources

d) Financial News:

Similarly, for financial news, we used financial times API to collect news articles in batches of 1000. Using this, close to 25,000 News Articles were recorded and parameters like title, sub-title, date, publisher were collected for the one year span.

Negative Financial News			Positive Financial News		
	neg_title	neg_compound		pos_title	pos_compound
	Russia failed to prevent terror attack, Strasbourg rules	-0.9538		China rich help drive global billionaires' wealth — UBS	0.9468
	Iran accuses Saudi Arabia of 'promoting terrorist groups'	-0.9538		Equity investors appear relaxed at Fed hike prospect	0.9460
	Death toll rises in St Petersburg, high and dry in India and the missing women in finance	-0.9485		Investors who bet on Macy's shares enjoy a great ride	0.9413
	Egypt strikes mosque attack suspects as death toll rises	-0.9468		Never mind the sell-off, Eagles' Super Bowl win points to positive year	0.9403
	Dirty air: how India became the most polluted country on earth	-0.9450		Companies with strong ESG credentials make better investments	0.9392
	Pressure on rupee exposes Pakistan's economic weakness	-0.9432		China gives helping hand to credit card industry	0.9382
	Tesco fraud trial collapse puts deferred prosecution deals in the dock	-0.9423		Euro enjoys renaissance as investors pivot to continent	0.9313
	Hedge funds trimmed bullish bets before US crude fell below \$50	-0.9382		Love Island adds to attraction of ITV ad hopes	0.9246
	Fears over future market crash stalk Norway's \$1tn oil fund	-0.9348		The FT's 10 fintech firms to watch	0.9246
	Assad blamed for chemical attack, Kirchner's family indictment and Elkan's gamble	-0.9337		Tap creative industries to boost Africa's economic growth	0.9246

Fig 2: Positive and Negative Financial News Articles

e) Twitter:

Social Media metrics are a key parameter to assess the trader's sentiment against a stock or cryptocurrency. We individually gathered these metrics for each of the 18 financial assets in consideration. Twitter data was gathered using a custom script that uses Twitter API. We were successfully able to gather 468,888 tweets for stocks and 328,704 tweets for cryptocurrencies for a timespan of one year. Quantitative and textual information like comments, re-tweets, likes were gathered to perform further analysis.

f) Reddit:

Likewise, Reddit is also another Social Media platform that has a dedicated subreddit group for each of the financial asset in consideration. We individually gathered these metrics for each asset using a custom script that uses Pushshift API. We were successfully able to gather: 557,391 Reddits for stocks and 1,284,023 Reddits for cryptocurrencies for a timespan of one year. Quantitative and textual information like score, subreddit, title, comments were gathered to perform further analysis.

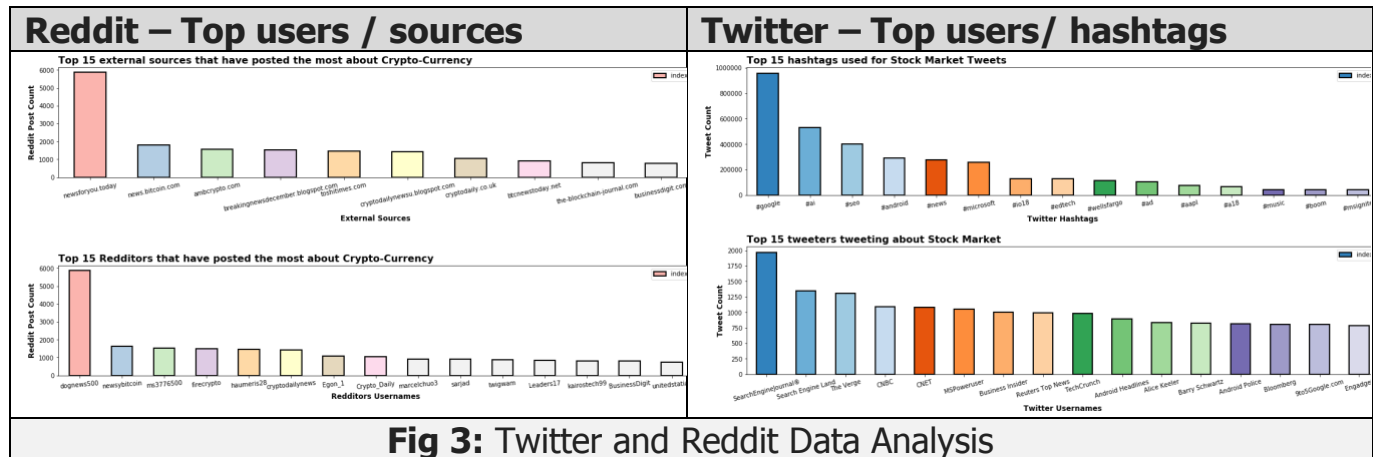


Fig 3: Twitter and Reddit Data Analysis

4.2. Data Pre-Processing - ETL

Real-world data at its earliest stages can often be very unstructured and unclear in format. Since the data collected was from 6 different sources, it brought in significant challenges with it. In order to make our data more concise and be able to perform a more accurate analysis, we had clean, reshape, modify it. Here are some ETL tasks which had to be performed to make the data usable.

a) Grouping:

Since news and social media data was collected on a minute by minute scale and OHCLV data was hourly, we had to transform our qualitative and quantitative metrics into hourly buckets and assign a sentiment score for it.

b) Natural Language Processing:

We had huge amounts of textual information like news title, tweets, comments, Reddits, sub titles from 4 different sources. Challenges like spam and irrelevant data were key and had to be addressed in this part. We devised a mechanism using EDA where we only consider top news sources and top users information for our sentiment analysis. This way we were able to keep only the useful data and exclude the rest.

For sentiment analysis, we used VADER (Valence Aware Dictionary and sEntiment Reasoner), which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews. This is because VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

c) Aggregation:

We combined all the OHCLV Data with other news and social media metrics in order to design our predictive model. The data was joined based on the asset name and hourly stamp present in each of the 6 sources.


symbol	asset_name	created_utc	open	high	low	close	volumefrom	volumeto	reddit_compound	news_compound	fin_compound	tweet_compound
BNB	Binance Coin	2018-01-22 08:00:00	14.14	14.44	14.00	14.10	105978.95	1507209.31	0.110769	0.060700	0.242300	0.4443
BNB	Binance Coin	2018-01-22 09:00:00	14.10	14.35	13.85	14.20	83496.47	1184968.02	0.104713	0.096200	0.359200	0.1647
BNB	Binance Coin	2018-01-22 10:00:00	14.20	14.24	14.03	14.05	33415.74	471694.25	0.104555	0.128156	-0.101133	0.6590
BNB	Binance Coin	2018-01-22 11:00:00	14.05	14.11	13.41	13.43	110293.99	1515195.81	0.256573	0.089106	-0.299700	0.0000

Fig 4: Example of aggregated data

4.3. Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. Primarily, EDA is performed to visualize what the data can tell us beyond the formal modeling or hypothesis testing task. This stage was crucial to understand relationships and formalize parameters that would be used in our model. Here are just a few questions we have tried to answer using EDA.

EDA	Inference
<p>Sentiment Distribution for each CryptoCurrency</p>	<p>From these pie charts, we can see that the sentiment scores are satisfactorily distributed. Also, we can tell that Ethereum followed by Bitcoin Cash has a good overall sentiment score among Redditors. Stellar followed by Binance Coin have the highest negative sentiment score among redditors.</p>
	<p>Form these word clouds, we got a good understanding of what words are most prominent in our data sources. Words like crypto, bitcoin, XRP are what we have expected.</p>
<p>Popularity Count per Stock</p>	<p>From these graphs we can tell that Alphabet followed by Microsoft were the two most talked stocks among tweeters. There is very less discussion happening about Exxon Mobil or Wells Fargo.</p>

EDA	Inference
	<p>Premise: Public opinion and news have a positive or negative influence on the price in most cases.</p> <p>Testing Procedure: In order to prove the above statement, we have considered the most relevant reddit posts that have extreme cases of a sentiment score. Then we have plotted these events with the closing price of cryptocurrency.</p> <p>Result: There was a very close relationship between them</p>

4.4. Model Training

4.4.1. Methodology

Initially, we have tried various machine learning algorithms to generate predictions and recorded their accuracies. These include *logistic regression* (accuracy:69.05%), *multilayer perceptron classifier* (accuracy: 51.54), and *artificial neural network* built with keras (accuracy:51.54%). Out of all these model, the best performance was obtained by the baseline model of XGBoost. Tree boosting is a highly effective machine learning method and XGBoost is one of the most popular systems recognized in a number of machine learning and data mining challenges. As per the paper on XGBoost ^[8], among the 29 challenge winning solutions 3 published at Kaggle's blog during 2015, 17 solutions used XGBoost. The reasons for choosing XGBoost, apart from it outperforming other models significantly in our own experiments, were quite similar to the reasons stated in the paper itself:

"Among the machine learning methods used in practice, gradient tree boosting is one technique that shines in many applications. Tree boosting has been shown to give state-of-the-art results on many standard classification benchmarks. LambdaMART, a variant of tree boosting for ranking, achieves a state-of-the-art result for ranking problems. Besides being used as a stand-alone predictor, it is also incorporated into real-world production pipelines for ad click through rate prediction. Finally, it is the defacto choice of ensemble method and is used in challenges such as the Netflix prize."

In addition, XGBoost is a scalable algorithm which can be very useful when dealing with big data. Initially, a baseline XGBoost model was created with only the four basic features i.e. *Open, High, Low and Close*. Then we added, a commonly used technical indicator in financial analysis, Moving Average, which is mostly used by intra-day traders. The following two types of moving averages to obtain an accurate description of the markets,

- a) *Simple Moving Averages*
- b) *Exponential Moving Averages*

Simple Moving Average: A simple moving average (SMA) is arithmetic moving average calculated by adding recent closing prices and then dividing that by the number of time periods in the calculation average. [9]

The formula for SMA is

$$SMA = \frac{A_1 + A_2 + \dots + A_n}{n}$$

where

A_n = the price of an asset at period n

n = the number of total periods

Exponential Moving Average: An exponential moving average (EMA) is a type of moving average (MA) that places a greater weight and significance on the most recent data points. [10]

The formula for SMA is

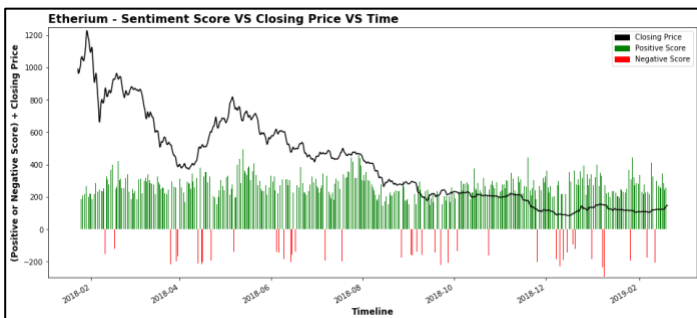
$$EMA_{Today} = \left(Value_{Today} * \left(\frac{Smoothing}{1 + days} \right) \right) + EMA_{Yesterday} * \left(1 - \left(\frac{Smoothing}{1 + days} \right) \right)$$

The following features were added in addition to the original OHLCV feature set:

1. SMA n : Simple Moving Average over n periods($n=20$ and 40 were chosen)
2. EMA n : Exponential Moving Average over n periods($n=20$ and 40)
3. SMA/EMA increments: The difference between current and previous SMA/EMA
4. SMA and EMA of next periods
5. Relative volume increments
6. News and Social Media features: four features - one from each (i.e *financial news*, *general news*, *twitter data* and *reddit data*) were obtained, after performing sentiment analysis and a combination of the positive, negative and neutral sentiments obtained.

4.4.2. Evaluation:

4.4.2.1. EDA in News:



We plotted a chart depicting the closing price of the asset with the sentiment predominant for that time period. We can infer from the time series chart, that a relationship can be established between the fluctuation of sentiment score with closing price. As you can see, whenever there is an increase in positive score, the

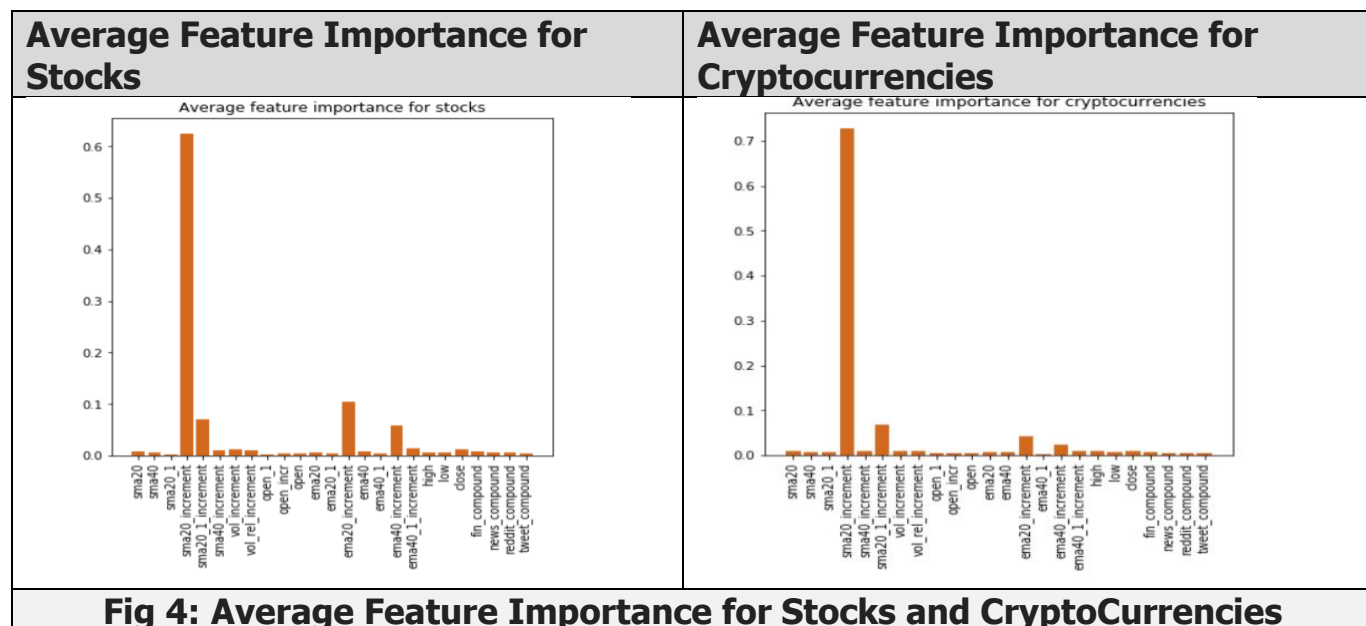
stock price increased and vice-versa.

4.4.2.2. Accuracies obtained from XGBoost:

The most important metric for the evaluation of our model was accuracy. Given below are the accuracy tables for both stocks and cryptocurrencies.

Table 2: Average accuracies for Stocks and Cryptocurrencies					
Name of Stock	Accuracy With news	Accuracy without news features	Name of Coin	Accuracy With news	Accuracy without news features
Walmart	93.44	92.81	Binance Coin	89.41	89.35
Microsoft	90.96	90.18	Bitcoin	86.91	87.12
Home Depot	88.12	79.64	EOS	89.13	89.03
Alphabet	91.49	91.23	Litecoin	89.96	89.96
Apple	90.96	80.00	Stellar	90.26	90.26
Wells Fargo	89.18	88.25	Tron	88.84	88.88
Chevron	91.49	91.05	Ripple	88.26	88.30
Coca Cola	91.32	83.85	Bitcoin Cash	85.53	85.52
Exxon Mobil	92.20	91.22			
Average	91.02	87.58	Average	88.54	88.55

As we can see, after tuning the model was able to obtain sufficiently high accuracies of 91.02% for stocks and 88.54% for cryptocurrencies. Also, as we can see that news effects some assets more than others. For instance, incase of stocks, both Apple and Coca-Cola see significant improvement with the news features added, while the rest see marginal improvements.



Also, as we can see from the graphs above, the news features were given lesser importance in both the asset classes in comparison to the other features. In addition, the increment of ema-20 and ema-40 was given more importance for stocks than cryptocurrencies.

A. Feature importance for stocks:

Table 3: Feature importance for stocks			
Five Most important features		Five least important features	
Feature	Percentage Importance (%)	Feature	Percentage Importance(%)
<i>sma20_increment</i>	62.40	<i>open_1</i>	0.29
<i>ema20_increment</i>	10.37	<i>sma20_1</i>	0.30
<i>sma20_1_increment</i>	7.09	<i>ema40_1</i>	0.34
<i>ema40_increment</i>	5.74	<i>ema20_1</i>	0.42
<i>ema40_1_increment</i>	1.50	<i>open</i>	0.43

B. Feature importance for cryptocurrencies:




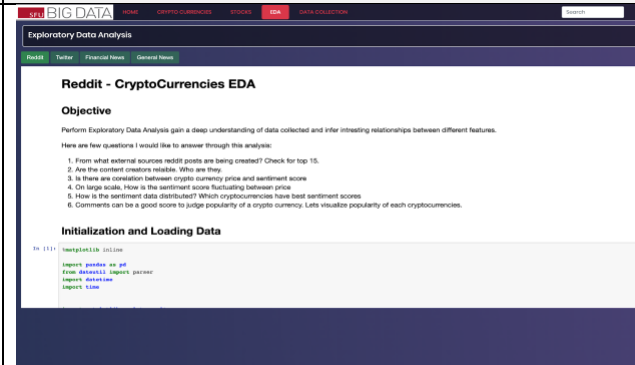
Feature importance for cryptocurrencies			
Five Most important features		Five least important features	
Feature	Percentage Importance (%)	Feature	Percentage Importance (%)
<i>sma20_increment</i>	72.68	<i>ema40_1</i>	0.26
<i>sma20_1_increment</i>	6.71	<i>reddit_compound</i>	0.45
<i>ema20_increment</i>	4.18	<i>open_1</i>	0.45
<i>ema40_increment</i>	2.34	<i>tweet_compound</i>	0.48
<i>ema40_1_increment</i>	1.05	<i>open</i>	0.52

As we can see from the above tables, two out of four news features are the least used features for cryptocurrencies, hence the impact of news features is lesser in predictions for cryptocurrencies.

4.5. Data Product

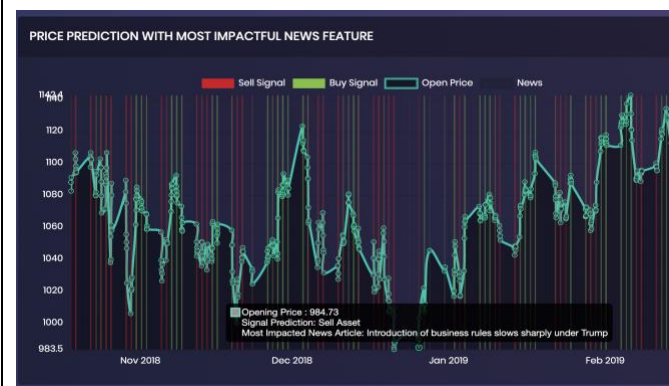

The data product is an end to end application that obtains the data, processes it and outputs buy and sell signals with reasonably good results. It is fairly easy to use for by the day traders and does not require complex knowledge of Machine Learning Algorithms which are running behind the easy to use front end. With an intuitive UI, it is easy to navigate and provides an option to explore each stage of our data pipeline. In addition, it provides many useful widgets such as the historical price checker, the news ticker, and the price vs sentiment score bar graph.

One of the key highlights of the product is the "on hover" news feature. This feature displays the headline of the news influencing the decision of the model at the time. This can be useful for not only the day traders while making their own decisions, but also to the layman who is studying the market Another one is the display of the key features used for making the prediction. This can enable the trader to make their own decision, just considering the most important key indicators used by the model. Tools like *Flask, JQuery, Nginx, Unicorn, Supervisor, ChartJS, ApexChartJS and HTML* where used. The application itself was hosted on SFU Cloud.

Overview of our Data Product	
Home Page	Dynamic Crypto Dashboard
	
Dynamic Stocks Dashboard	EDA
	

4.5.1. Available Widgets:

Each Widget in each Dashboard gets dynamically updated for each stock and cryptocurrency

Widgets	Inference
	<p><i>Price Prediction with most Impactful News Feature:</i></p> <p>This widget not only suggests the price signal but also links to the news its decision is based on. As you can see in the picture, there has been a steep dip in the price when trump introduced new business rules in Jan 2019. This gives one such news article per hour that influenced price of the asset.</p>
	<p><i>Price Vs Sentiment Score:</i></p> <p>This widget combines all the sentiment scores from different sources and plots it against the opening price of an asset per day.</p>

LATEST NEWS

- China's housing glut casts pall over the economy
2019-02-19 23:30:24+00:00
- Fast Asia Open: Japan trade data, Australia wages
2019-02-19 23:12:12+00:00
- US stocks await latest on trade talks
2019-02-19 22:44:09+00:00
- Walmart earnings, trade optimism drive US stocks higher
2019-02-19 21:33:51+00:00
- Gold glitters on dovish signals
2019-02-19 21:33:34+00:00

Latest News:

This widget top 100 news that impacted the price of the asset.

HISTORICAL PRICE CHECKER



Historical Price Checker:

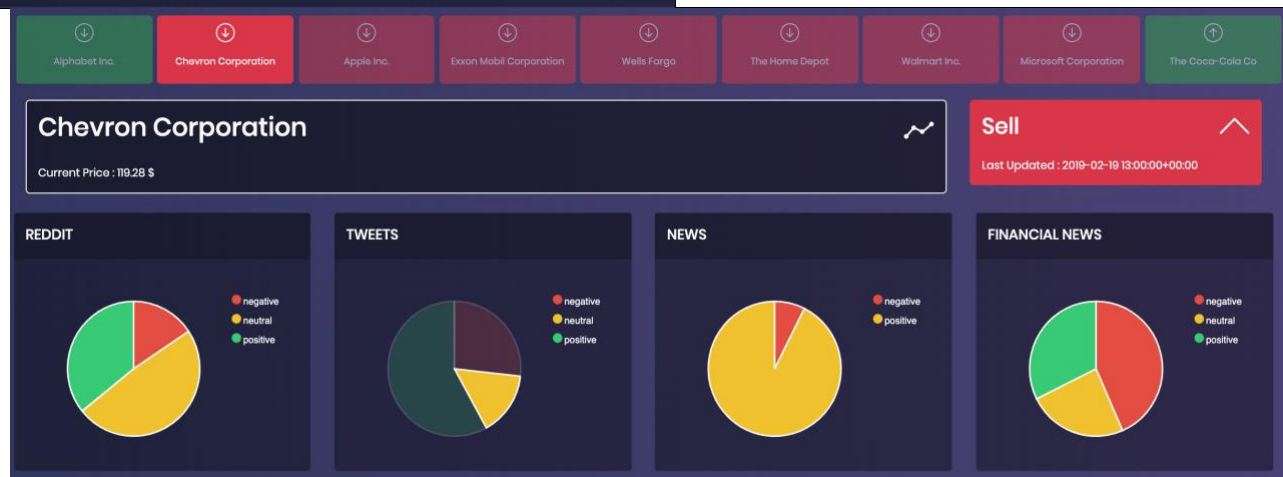
With this widget, you can track the historical price of a financial asset for the last one year.

HISTORICAL DATA WITH PREDICTIONS

Date	Open	High	Low	Close	Volume	Value	Price Signal
2019-02-19 13:00:00+00:00	119.28	119.43	119.11	119.40	22175	Sell	Sell
2019-02-19 12:00:00+00:00	119.38	119.47	119.22	119.30	32863	Sell	Sell
2019-02-19 11:00:00+00:00	118.55	119.39	118.55	119.39	39751	Buy	Sell
2019-02-19 09:00:00+00:00	118.92	119.34	118.91	119.07	28326	Buy	Buy
2019-02-15 15:00:00+00:00	118.94	119.34	118.82	119.27	62348	Buy	Buy
2019-02-15 14:00:00+00:00	118.71	119.02	118.70	118.96	39244	Buy	Buy
2019-02-15 13:00:00+00:00	118.98	119.00	118.80	118.72	38012	Sell	Buy
2019-02-15 12:00:00+00:00	119.19	119.19	118.80	118.95	32485	Sell	Sell
2019-02-15 11:00:00+00:00	119.15	119.26	118.90	119.20	44727	Sell	Sell
Date	Open	High	Low	Close	Volume	Value	Price Signal

Historical Data with Predictions:

This widget provides the Open, High, Low, Close and Volume information with the predicted price signal for the last 10 hours.



This widget is a dynamic way to gauge the overall sentiment from 4 sources. It also gives the latest price signal for each financial asset. It is color coded with red indicating sell and green indicating a buy signal.

5. Lessons Learnt:

This project has been a great learning ground to implement all the skills and techniques gained through *CMPT 732* and *733* courses. From its very beginning, we had to first find an interesting problem and come up with a workflow that could provide solutions to it. We learned different data collection strategies and used ETL to clean, transform and modify it according to our needs.

We performed exploratory data analysis through which we learned some really helpful patterns and information from our data. This also helped us with our model training phase. We learned that XGBoost is one of the most useful algorithms which can give good results. Finally, we were successful in assembling a complete data product which can be used by day traders to make optimized buy or sell decisions. Overall, it prepared us to showcase technical and interpersonal skills which are crucial to be a successful data scientist.

6. Conclusions:

In conclusion, we were successfully able to create a product that can output buy and sell signals which can be used to maximize profit by intraday traders. We saw both from the EDA and the final model results - *that news and public sentiment does play a role in the market*, and how these can be exploited successfully for market direction prediction. This effect appears to be more pronounced in the traditional equity markets, maybe because the traditional markets are in themselves constrained from many other factors that this is sufficient to produce a marked change, or maybe because the cryptocurrency market is so volatile that this volatility overshadows the impact of the news and other public sentiments. The most important feature and technical indicator was the difference between the current and the previous Simple Moving Average over 20 periods. The top 5 important technical indicators are also outputted and can be used even for trading manually if one does not wish to use the predictions given by our model.

There can be a lot of future work that can be possible in this. In addition to just the SMA and EMA, we can also include other market indicators such as support and resistance, relative strength indicators and other breadth indicators, just to name a few. We can also enable the current implementation to work in live data so as to provide real-time predictions. And finally, we can also predict the actual prices of the assets rather than just the direction the asset is going which can be more useful, and see how accurate that model would be.

7. References

- [1] Patel, H., Parikh, S., & Darji, D. (2016). Prediction model for stock market using news based different Classification, Regression and Statistical Techniques: (PMSMN). 2016 International Conference on ICT in Business Industry & Government (ICTBIG), 1-5.
- [2] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603-9611, 2015.
- [3] J. J. Murphy, *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin
- [4] Li, X., Wang, C., Dong, J., Zhu, F., Hameurlain, S., Liddle, A., . . . Zhou, Xiaofang. (2011). Improving Stock Market Prediction by Integrating Both Market News and Stock Prices. In *Database and Expert Systems Applications: 22nd International Conference, DEXA 2011, Toulouse, France, August 29 - September 2, 2011, Proceedings, Part II* (Vol. 6861, *Lecture Notes in Computer Science*, pp. 279-293). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [5] Shen, W., Guo, X., Wu, C., Wu, D. Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm. *Knowledge-Based Systems*, 24 (3): 378-385, 2011.
- [6] Schumaker, R.P., Chen, H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, 27 (2): 12, 2011.
- [7] Bollen, J., Mao, H., Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2 (1): 1-8, 2011.
- [8] Chen, Tianqi and Guestrin, Carlos, XGBoost, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016.
- [9] <https://www.investopedia.com/terms/s/sma.asp>
- [10] <https://www.investopedia.com/terms/e/ema.asp>

8. Appendix I

Links to our product:	
Page	URL
Home Page	http://nml-cloud-58.cs.sfu.ca/tradespade
Dynamic Crypto Dashboard	http://nml-cloud-58.cs.sfu.ca/cryptocurrencies
Dynamic Stocks Dashboard	http://nml-cloud-58.cs.sfu.ca/stocks
Exploratory Data Analysis	http://nml-cloud-58.cs.sfu.ca/eda
Data Collection Notebooks	http://nml-cloud-58.cs.sfu.ca/data_collection

9. Appendix II

Link to code repository: <https://csil-git1.cs.surrey.sfu.ca/cmpt-733-tradespade/tradespade.git>