Stock Market forecasting analysis using different probabilistic and machine learning Models

Abdullah Mousselli
Oklahoma University, ECE Department
Amousselli@ou.edu

Mohamed Afify
Oklahoma University, ECE Department
Mafify@ou.edu

Abstract:

Stock Market is a challenging field when it comes for future prediction and Due diligence work by investors. Most of financial institutes prefers to do analysis of the companies instead of data prediction as it is a hard tool to use in the market currently. Using fundamental and technical data to predict the market direction also leave a lot of uncertainty as the current economic analysis and the political news on the other side.

In our research project we are proposing a data analysis using the probabilistic and Machine learning model using all the tools as fundamental, technical, and economical analysis. Our models show promising data and good accuracy prediction of the market movement. Our model is working on 10 famous stock tickers in 2020 and have a high return with an average of 542% over one year. Models that are tested in that paper are Bayesian, Hidden Markov, time series and Machine learning models.

Introduction:

The stock price is valued by the company value as their revenue, number of share floats and their catalysts which derive the stock value. This is based on the theoretical work, but in the real-life stock price is getting affected by many factors. These factors can be sectored in many directions as fundamental as what we mentioned company EPS which is reported every quarter by the company and the number of contracts they report by each year. Other factor is the technical and that is based on 4 values of the stock on a daily average, it's the opening price, closing price, highest, and lowest. These values will give us many statistical values based on the trading volume of the stock. Volume is the analytical value that shows the trend direction of the stock if it's going up or down. This will give us indication of the moving average of the SMA and DMA which will shows the resistance and support levels of the value of stock. Also, indication of breaking can get higher volume and get new highs or new lows, therefore it is very important to have these analyses in the stock model.

As I mentioned the volume is the factor of the uptrend in any stock value. The volume increases because of good news or bad news that company reports. For example, SPI energy stock surged 3,100% on a EV venture announcement [1]. Because of that announcement there was over 348 million of trading volume on the stock. There are other technical attributes that are strong indication as the RSI which shows if the stock is oversold or overbought which means it can start going up or it's time the stock will go down. Also shoring volume can be an indication of a high

risk on the stock and they can manipulate the price to go down. Therefore, it's a strong indication to have it in the model to enhance the accuracy of prediction.

Last factor is the economy status of the country, as unemployment and price index move, investors will move as they are the main gear of the market index movement. In April 23, the unemployment rate hits 20% after the COVID-19 pandemic hits US [2]. This dropped the S&P 500 index points by 471 points. Also Buy index shows if the citizens are buying or saving the money, that can be a good indicator to see a future rise of the stock or there is a huge correction coming up. Therefore, we add these two indicators to the market to check how the model accuracy will get better based on the gathered information.

This project explored non-probabilistic and probabilistic models for stock trend prediction, different Machine Learning algorithms were applied such as: Bayesian model, Hidden Markov Model, Logistic Regression, Nearest Neighbor (KNN), Support Vector Machine (Support Vector Machine), Gradient Boosting Classifier (XGB), Decision Tree, Random Forest, Multi-layer Perceptron classifier and Naive Bayes. All attributes were used to classify the trend as up or down.

Deep Learning Time-Series models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) were also applied to forecast stock strategy based on predicting the price. A times series with different look-back values was used.

Data Preprocessing:

The data we are using are mainly focusing on the Stock prices (HLOC) which states the highest price it reaches through the trading period, the lowest it did reach, what the price opening and the closed price. The range can be in 1 minute till month. In our data we are using the one-day range on a 5-year period. The stocks used in our analyzing scheme are 10 strong stocks, some of them belongs to the Dow Jones Industrial average which contains the most powerful 30 companies in the US industry. The other stocks are popular and hyped by the investors as the vaccine company Pfizer and tesla the EV company. The 10 companies' names are Apple, Tesla, Walmart, Boeing, Disney land, JP Morgan, Microsoft, Nvidia, Pfizer and AstraZeneca.

We also collected the stock indicators as the volume of trading, MACD, RSI, Volume delta, SMA close at 50 days period, DMA, and Bollinger bands. These indicators can give the model a strong indication of the stock direction.

Date	CI Price Index	UI	open	high	low	close	volume	MACD	boll	RSI	Class
1/7/2015	42.80	5.70	26.80	27.05	26.67	26.94	1.60E+08	-0.01	26.85	26.85	up
1/8/2015	42.80	5.70	27.31	28.04	27.17	27.97	2.37E+08	0.04	27.07	27.07	up
1/9/2015	42.80	5.70	28.17	28.31	27.55	28.00	2.15E+08	0.07	27.23	27.23	up
1/12/2015	42.80	5.70	28.15	28.16	27.20	27.31	1.99E+08	0.05	27.24	27.24	down
1/13/2015	42.80	5.70	27.86	28.20	27.23	27.56	2.68E+08	0.06	27.28	27.28	up
1/14/2015	42.80	5.70	27.26	27.62	27.13	27.45	1.96E+08	0.05	27.30	27.30	down

1/15/2015	42.80	5.70	27.50	27.51	26.67	26.70	2.40E+08	0.01	27.24	27.24	down
1/16/2015	42.80	5.70	26.76	26.90	26.30	26.50	3.14E+08	-0.03	27.17	27.17	down
1/20/2015	42.80	5.70	26.96	27.24	26.63	27.18	2.00E+08	-0.03	27.17	27.17	up
1/21/2015	42.80	5.70	27.24	27.76	27.07	27.39	1.94E+08	-0.01	27.19	27.19	up
1/22/2015	42.80	5.70	27.57	28.12	27.43	28.10	2.15E+08	0.04	27.25	27.25	up
1/23/2015	42.80	5.70	28.08	28.44	27.88	28.25	1.86E+08	0.09	27.32	27.32	up
1/26/2015	42.80	5.70	28.43	28.59	28.20	28.27	2.22E+08	0.13	27.38	27.38	up
1/27/2015	42.80	5.70	28.10	28.12	27.26	27.28	3.82E+08	0.10	27.37	27.37	down
1/28/2015	42.80	5.70	29.41	29.53	28.83	28.83	5.86E+08	0.17	27.46	27.46	up

This project also utilized a stock news indicator for real-time stock trend prediction. This was achieved by applying sentiment analyzer on each news and scoring the sentence on a scale from [-1 to 1] where 0 represents a neutral news, positive scores indicates positive news for the company which implies a price increase in the near future and negative scores that indicates a soon drop in the stock value. Taking the daily average score for each ticker provided valuable and reliable information for the overall sentiment of a stock.

We also included economic indicators such as Unemployment Rate and Price index per Month as we believe these indicators affects the stock market, especially on the long term.

The proposed ideas split the data range in 3 to 2 years as for training and validating. Multiple models were investigated in this project in terms of average holding per stock or trade strategy in terms of buying and selling. The trading strategy is classified into two categories which is up and down based on the difference of the closing price between today and yesterday, if it is negative it is determined as down and vice versa.

The probabilistic models had a different data preprocessing scheme, as all attributes were classified as “0” or “1” based on the direction of the indicator between the day and one before and then compared with the target which the stock index movement.

Stock Sentiment Analysis:

Mainly there are two methods for forecasting stock market trends. The first one is Technical Analysis which considers indicators such as past stock price and volume to determine the future trend of a certain stock. The other method is Fundamental Analysis which depends on Analysis data about a certain company for example news to get insights about the future trends of a stock indicator. However, achieving efficiency from Technical Analysis or Fundamental Analysis is disputed by efficient-market hypothesis which states that the stock market is prices are essentially unpredictable.

In this project we tested a Fundamental Analysis technique to forecast future trends of a certain stock indicator by using news headlines. Each headline is classified as either Negative (Bad), Zero (Neutral) or Positive (Good). To achieve this, we used news headline from FINFIZ [3]. A famous website for stock market financial data and live updated news. The free account provides access

to the last 100 news headline of each selected stocks. While premium account and other paid API services provides wider historical data in this part of the project, we determined that last 100 news headlines are sufficient for a 5-days based study.

Each headline was then analyzed using VADER (Valence Aware Dictionary for Sentiment Reasoning) [4]. which is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text. One big difference between this library and other NLP (Natural Language Processing) libraries is that VADER dictionary contains words for Stock news [5] and it is easy to update the library with new words as we needed to include words especially for Bio-Stocks such as AZN and PFE where words like EUA, sti-1499 and many other determine good news for bio vaccine companies.

We found that the model performed well specially for companies that run on news (specially bio companies with small market capital).

The following graph shows the stock news analysis for DIS (on the left and Walmart on the right.

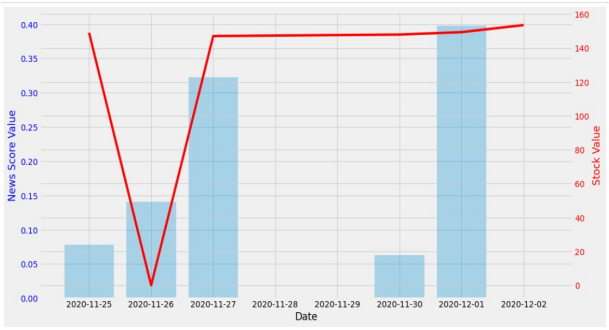

Figure 1. DIS stock against news graph

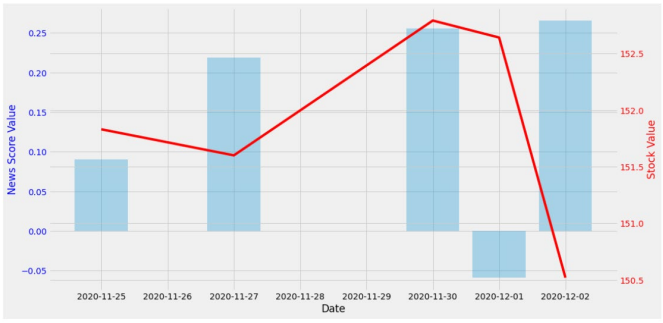

Figure 2. WMT stock against news graph

As we see in the graphs above the red line represents the stock value (price) and the blue bars represents the news analysis for that day. For any news that come on a weekend or a holiday we summed up the intensity of those news to the following weekday.

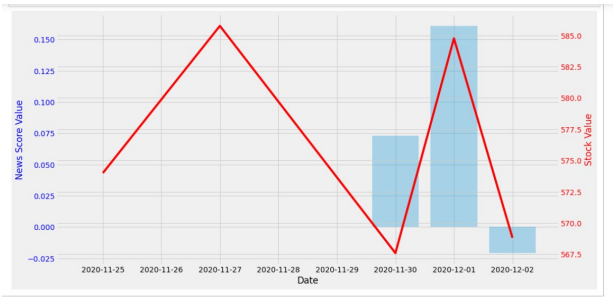

Figure 3. stock against news TSLA

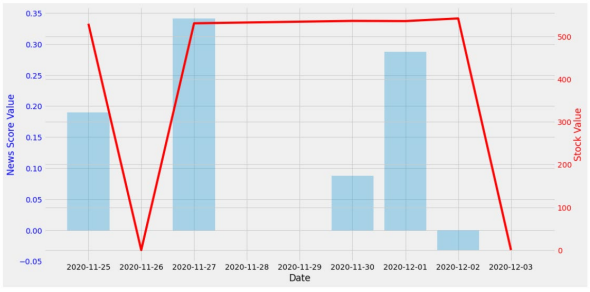

Figure 4. stock against news NVDA

When applying the model on other tickers (other than the ones suggested in this study) we found that is model runs well on stocks that are known to run on news such as small capital vaccine (bio) stocks. For example, the following graph showcases a news sentiment analysis applied on the last 100 news for SRNE Sorrento Therapeutics (a small capital bio company).

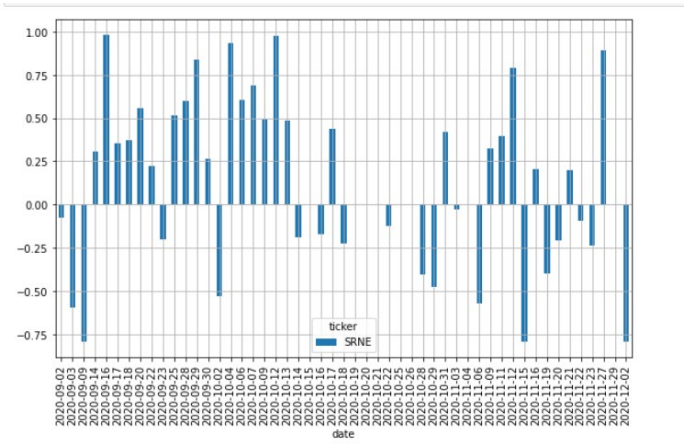

Figure 5. news analysis for SRNE

This model performed poorly on stocks that have little to no daily news (usually stocks with small cap and low volume). The graph on the right demonstrates news sentiment analysis for GEVO which is a Biofuel and Low-Carbon Chemicals company. We can see that there is 10-day gap with no significant news (positive or negative) about this company.

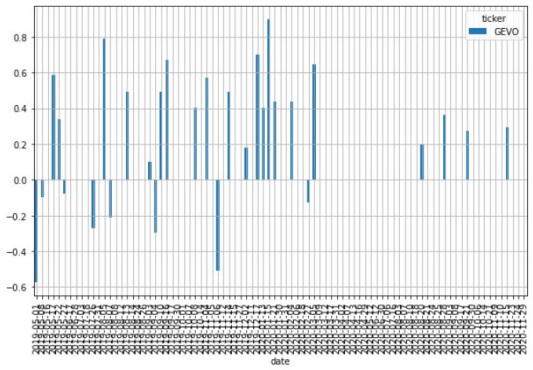

Figure 6. news analysis for GEVO

Investment Portfolio Analyzation and Optimization

In this project we also conducted experiments on portfolio optimization. By analyzing stock data for the last two years (starting from January 2018 up to December 2020) to determine the optimal percentage for each ticker in the investment portfolio. We created a portfolio consisting of the 10 suggested companies (tickers) with equal percentages for each ticker and equals to 10%. After that we calculated Expected Return for this portfolio which is a measure of the center of the distribution of the random variable that is the return. We also calculated portfolio volatility which is the degree of variation of a trading price series over time. Sharpe Ratio is another indicator we calculated

which is used to help understand the return of an investment compared to its risk. The ratio is the average return earned in excess of the risk-free rate per unit of volatility or total risk.

The following table shows the results for the 734-days investment period (starting from January 2018 up to December 2020).

Table 1 ER vol and SR for initial portfolio

Expected return	0.911
Volatility	0.462
Sharpe Ratio	1.797

To further analyze the investment portfolio, we calculated Skewness which sis a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. As well as Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers

Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case. "When both skewness and kurtosis are zero, the pattern of responses is considered a normal distribution. A general guideline for skewness is that if the number is greater than +1 or lower than –1, this is an indication of a substantially skewed distribution. For kurtosis, the general guideline is that if the number is greater than +1, the distribution is too peaked. Likewise, a kurtosis of less than –1 indicates a distribution that is too flat. Distributions exhibiting skewness and/or kurtosis that exceed these guidelines are considered nonnormal." (Hair et al., 2017, p. 61).

The following tables shows the Skewness of each ticker in our portfolio on the left and the kurtosis on the right.

Table 2. Skewness and Kurtosis for each ticker

AZN	0.459728
PFE	-0.157431
AAPL	1.26387
NVDA	1.367187
MSFT	0.632709
TSLA	1.943015
JPM	1.005540
BA	-0.968987
WMT	0.449713
DIS	0.279234

AZN	-1.106396
PFE	-0.958792
AAPL	0.480822
NVDA	0.993529
MSFT	-0.858963
TSLA	2.599985
JPM	0.920201
BA	-0.553470
WMT	-0.644347
DIS	-1.129722

We also analyzed the buy-sell signals for each ticker based on the moving averages (5-day, 10-day, 50-day, 100-day).

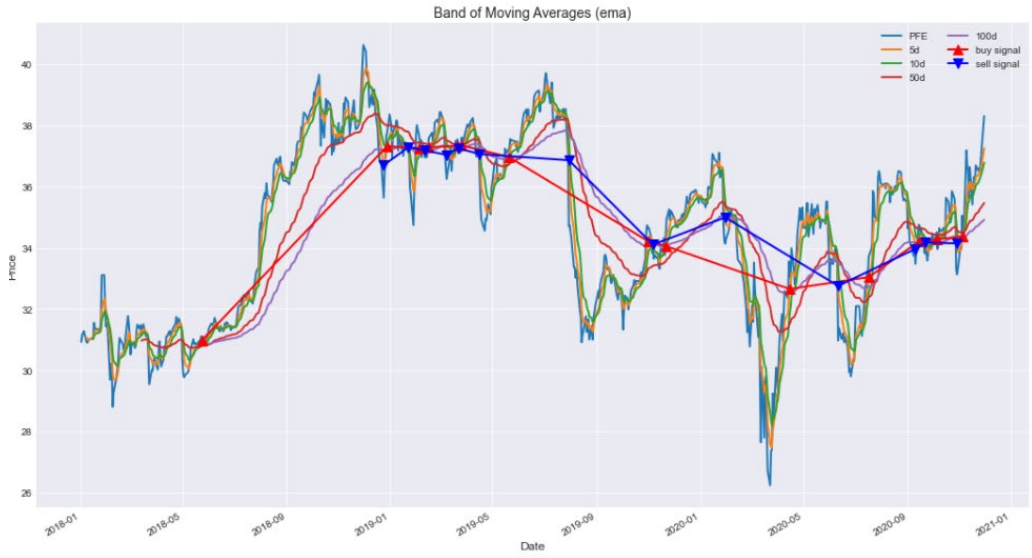

Figure 7. buy-sell signals for PFE

For portfolio optimization we explored two methods. One is Efficient Frontier and the second one is Monte-Carlo simulation. The efficient frontier is the set of optimal portfolios that offer the highest expected return for a defined level of risk or the lowest risk for a given level of expected return. Portfolios that lie below the efficient frontier are sub-optimal because they do not provide enough return for the level of risk. Portfolios that cluster to the right of the efficient frontier are sub-optimal because they have a higher level of risk for the defined rate of return.

Efficient Frontier was calculated based on two optimization preferences. The first one is minimum Volatility and the second one is maximum Sharpe Ratio

Table 2 Optimized portfolio for min Volatility

Time window/frequency	734
Risk free rate	0.08
Expected annual Return	0.514
Annual Volatility	0.332
Sharpe Ratio	1.308

Table 3 Optimized portfolio for Max Sharpe Ratio

Time window/frequency	734
Risk free rate	0.08
Expected annual Return	1.943
Annual Volatility	0.713
Sharpe Ratio	1.615

Table 5 Optimal Weights min vol

AZN	0.272357
PFE	0.236112
AAPL	0.0
NVDA	3.469447e-18
MSFT	5.854692e-18
TSLA	0.000958
JPM	0.0
BA	1.322727e-17
WMT	0.372915
DIS	0.117658

Table 4 Optimal Weights min vol

AZN	0.006855
PFE	0.0
AAPL	0.395786
NVDA	8.147779e-17
MSFT	0.019397
TSLA	0.475793
JPM	0.0
BA	0.0
WMT	0.102168
DIS	0.0

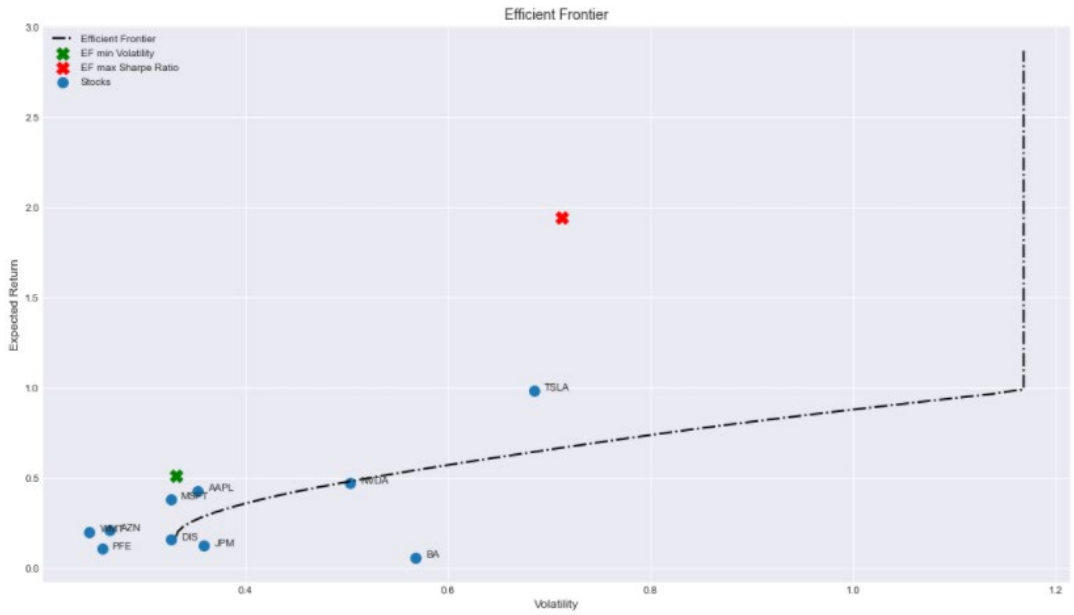

Figure 8. EF

The second method to optimize the portfolio was Monte-Carlo simulation. In this method we loop through 50000 possible portfolio and pick the one most suitable to the optimization preference (Minimum Volatility or Maximum Sharpe Ratio).

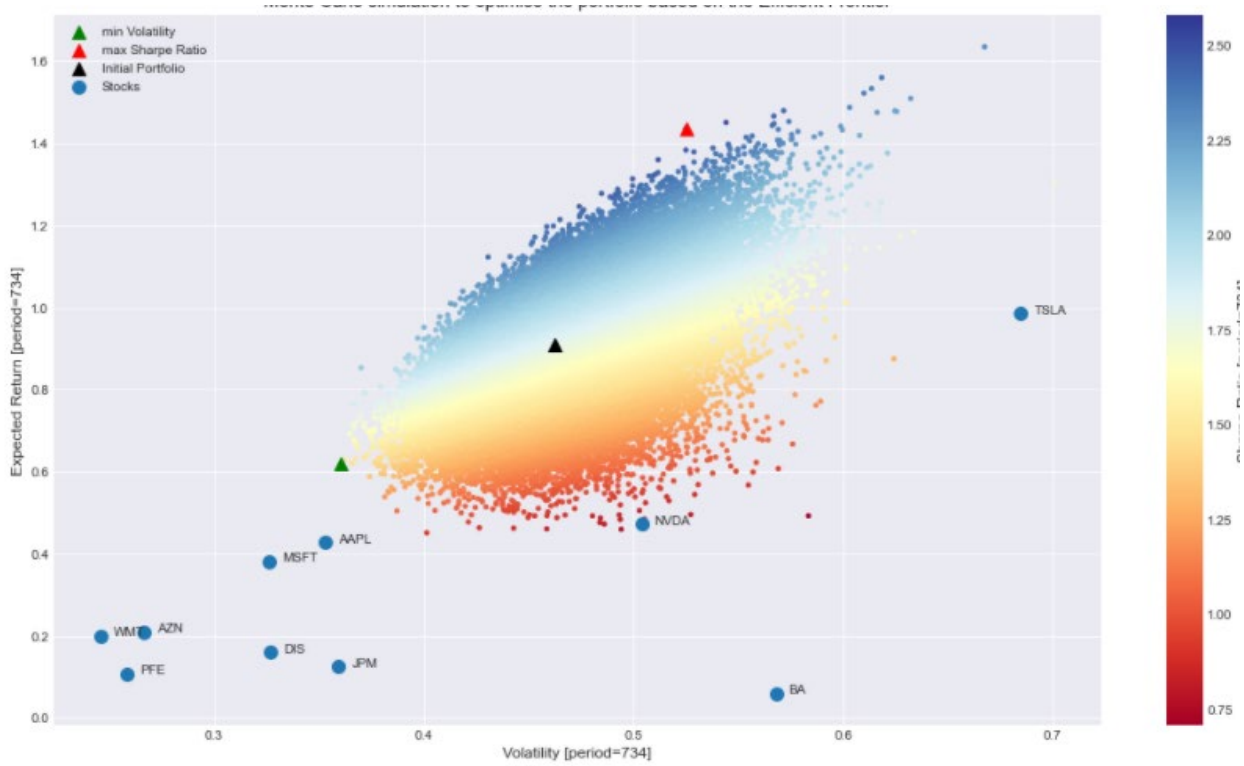

Figure 9. Monte-Carlo portfolio optimization

Stock Prediction Using Machine Learning Algorithms

To compare the performance of our probabilistic based prediction models we used the data for each ticker that has 21 different features such as volume, MACD, boll-up and other indicators that we produced in the Data preprocessing section to predict the trend of the stock as up "1" or down "0" using different Machine Learning classification methods. The methods we tried are: Naïve Bayes, Support Vector Machine, Logistic Regression, Neural Network, Nearest Neighbors, Gradient Boosting Classifier, Decision Tree and Random Forest.

Before feeding the data to the prediction it is important to check the correlation between the features. The following graphic represents the Confusion Matrix for the 21- features.

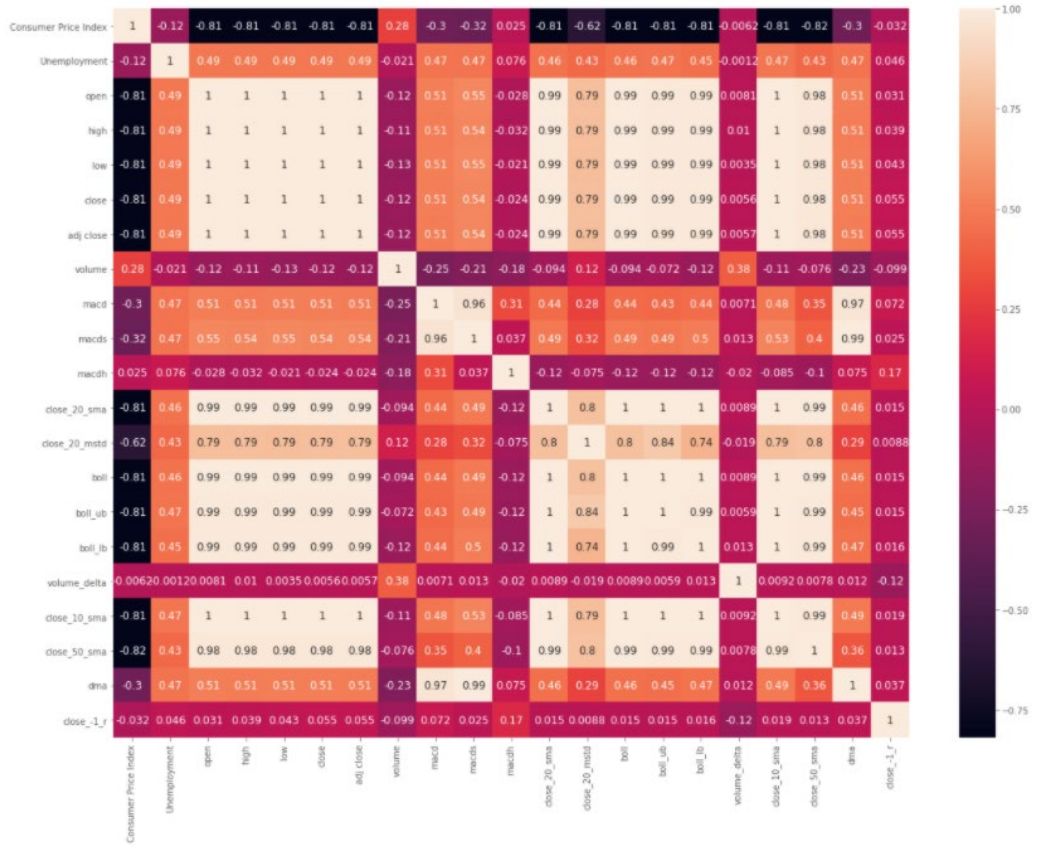

Figure 10. features correlation Confusion Matrix

The following table shows the scores for each classifier with cross validation.

Classifier	**Cross Validation Score**
SVM	51%
LR	48%
NN	46%
Naïve Bayes	46%
Decision Tree	46%
Random Forest	45%
Gradient Boosting Classifier	44%
KNN	43%

As we can see from the table above all classifiers performed poorly with SVM being the best with 51% score. This can be due to the fact that some of the features used are much less important (very smaller coefficient value compared to others). A study should be conducted to analyze the effects of each feature and determine the weights for each feature or maybe some features were not correlated correctly further investigation for improvement is required.

The following graphic shows feature importance analysis based on the Logistic Regression Classifier:

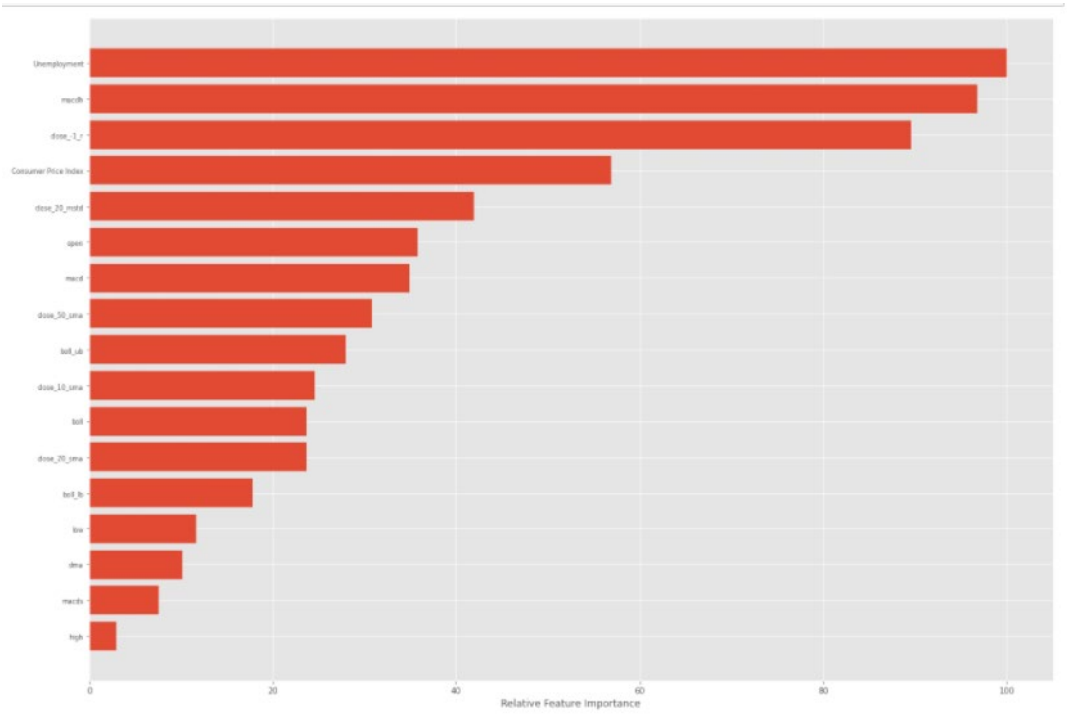

Figure 11. Feature Importance Analysis with LR

Stock Prediction Analysis using LSTM and GRU

As the Machine Learning Classification Methods, we tried failed to provide us with efficient results we were inclined to treat the prediction as a time-series problem where we look back to a period of time in our case, we tried many and settled with the best result (60 days). Each time series has 18 features we dropped the (close, adj close, volume) to avoid overfitting where the model basically looks at the value (Data leakage) a problem that most of the LSTM stock prediction models online suffers from (please check the notebook on Github for more details like model structure etc). Using this approach, we were able to achieve accuracy around 70% for all tickers and was the most 76% for LSTM model on Apple stock. While we applied cross validation to check for overfitting and data leakage it is premature to say that the accuracy is 100% correct as more investigation is required to confirm this percentage. According to the results we got we can see that treating the problem as a time-series improved the accuracy than a trend-up trend-down classification problem for Machine Learning Methods.

FThe following graph shows the training and validation loss for LSTM and GRU models trained on the Apple dataset.

Figure 12. train and Val loss for Apple (LSTM, GRU

Bayesian model:

The Bayesian model taking all the factors that affect the Price value of the Stock which is the "Adjusted Price". The price action is Up and Down and based on that two classes the investor will take the decision to buy or sell the stock.

To have a dataset that satisfies the model which will only take "0" and "1" and decide the node route based on the value. Therefore, every row was deducted from the previous one and based on the value whether it is negative or positive it will assign the positive value to 1 and the negative to 0. Then the columns of the attributes then are shifted one day backwards to equal the current price direction so the model can detect the next day price action based on the data.

The model is predicting each stock based on the parameters, each node will have a conditional probability which is trained on the historical data from 2015 to 2017. As shown in the following figure the mechanism of the Bayesian network and how it predicts the market based on the given data.

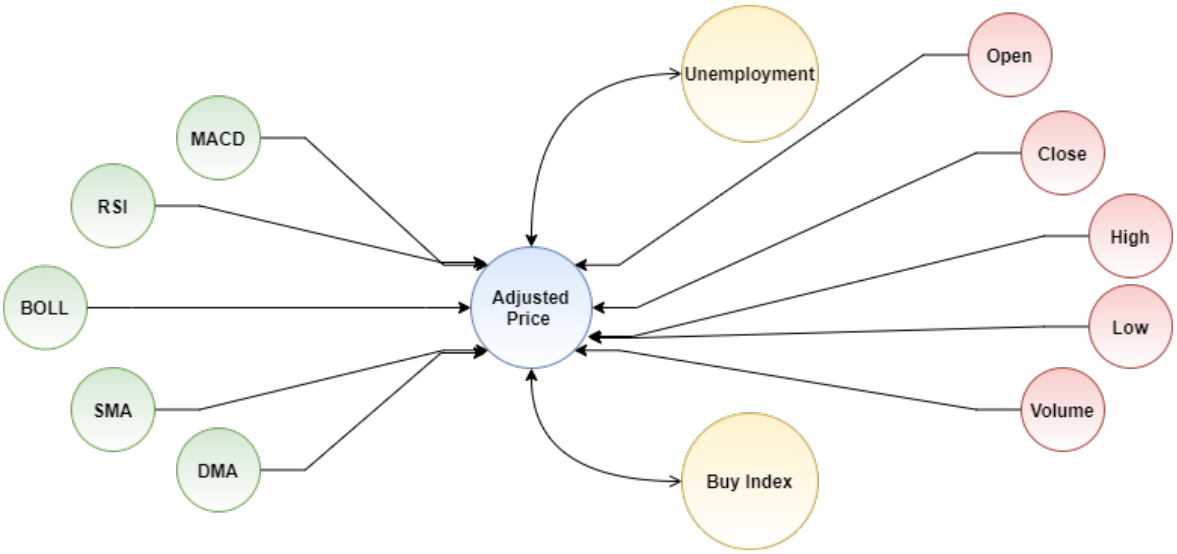

Figure 13. Bayesian Network Model

The following tables shows all the probability of price going down and up based on the direction of each attribute, the first table shows the probabilities when the price direction is down which means sell, while the second table for the Up direction. The calculations are based on the historical data of Apple stock.

Table 7. Example set of Apple stock on down-trend learned from historical data

Features	Down	UP
Price Index	0.82	0.39
Unemployment	0.71	0.46
Open	0.27	0.45
Low	0.36	0.48
High	0.43	0.49
Close	0.54	0.50
Volume	0.47	0.50
MACD	0.35	0.48
MACDH	0.41	0.49
MACDS	0.36	0.48
SMA	0.56	0.50
BOLL	0.46	0.50
BOLL UP	0.56	0.50
BOLL DOWN	0.51	0.50
MSTD	0.57	0.50
Volume delta	0.51	0.50
SMA 50	0.51	0.50
SMA 10	0.65	0.48
DMA	0.42	0.49

Table 8. Example set of Apple stock on up-trend learned from historical data

Features	Down	UP
Price Index	0.83	0.37
Unemployment	0.72	0.45
Open	0.79	0.41
Low	0.68	0.47
High	0.71	0.45
Close	0.51	0.50
Volume	0.48	0.50
MACD	0.64	0.48
MACDH	0.61	0.49
MACDS	0.60	0.49
SMA	0.71	0.46
BOLL	0.52	0.50
BOLL UP	0.71	0.46
BOLL DOWN	0.63	0.48
MSTD	0.59	0.49
Volume delta	0.58	0.49
SMA 50	0.69	0.46
SMA 10	0.72	0.45
DMA	0.60	0.49

We ran the model on the 10 picked stocks, all of them had an average of 70% accuracy overall, best stock prediction was Apple with an accuracy of 72%. To see the stock direction prediction versus the real price movement we plotted a graph to show the accuracy. In the following figure the Red color is a prediction to sell and green color a prediction to buy. We see an accurate prediction scheme except when there is a high volatility as shown in the last month. The model can't see a clear direction of the price.

Table 9. Bayesian Model Prediction for the 10 Picked Stocks

Stock Name	Accuracy
APPLE	0.721704
AstraZeneca	0.625832
Disney	0.704394
Boeing	0.695073
JP Morgen	0.69241
Microsoft	0.696405
NEVDIA	0.707057
Pfizer	0.701731
Tesla	0.713715

Walmart	0.715047

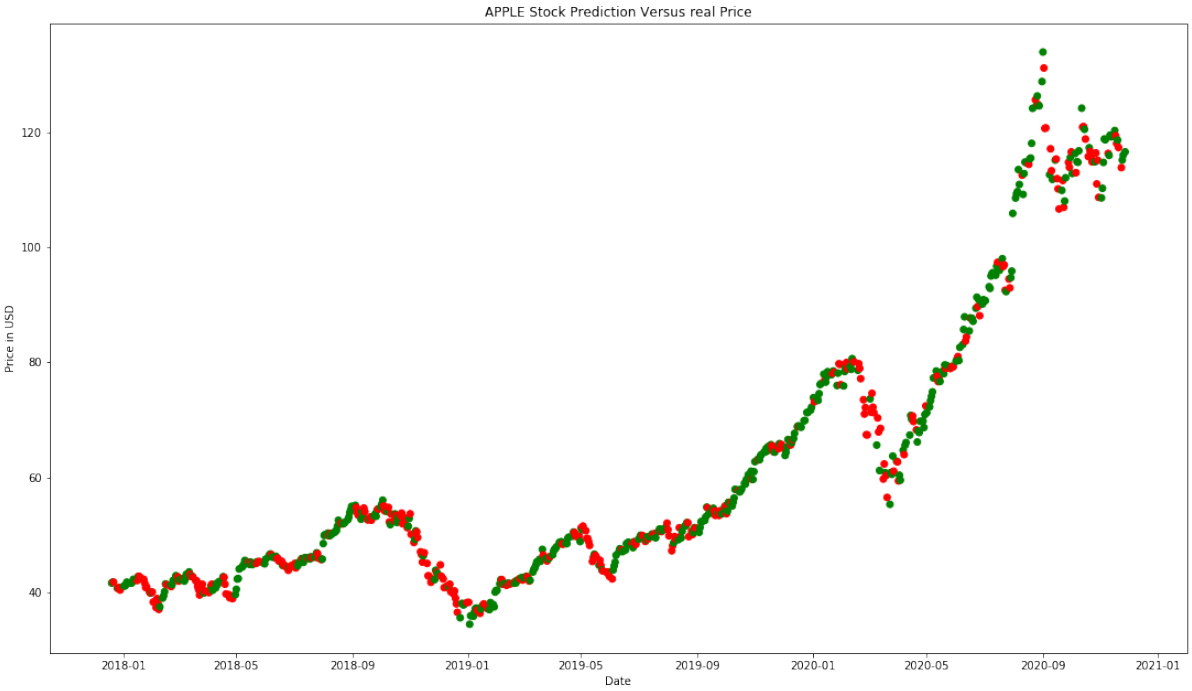

Figure 14. Apple stock prediction movement versus the real price action

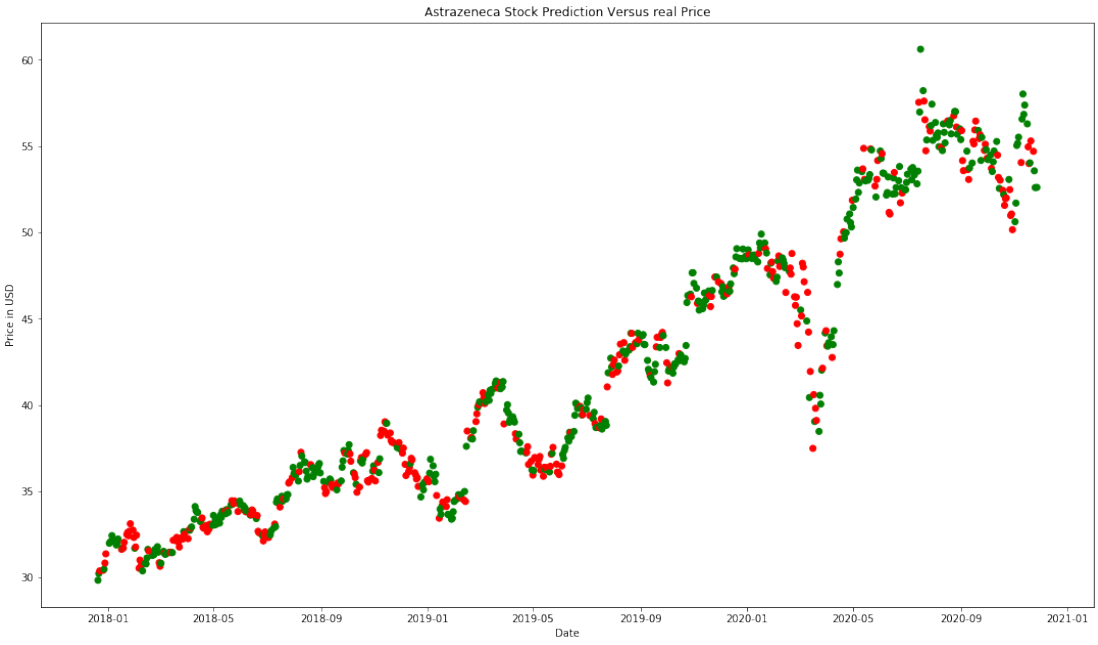

Figure 15. AstraZeneca stock prediction movement versus the real price action

Hidden Markov model:

The Markov chain rule is a complicated model to find the hidden state chain of the market, in the market we have mainly three hidden states which is the Bear, Bull, and nonvolatile movement when there is no price action. Trader can win money in the three states by doing option trading in a short period term. The model will detect it when it moves from one state to another through a segment of data.

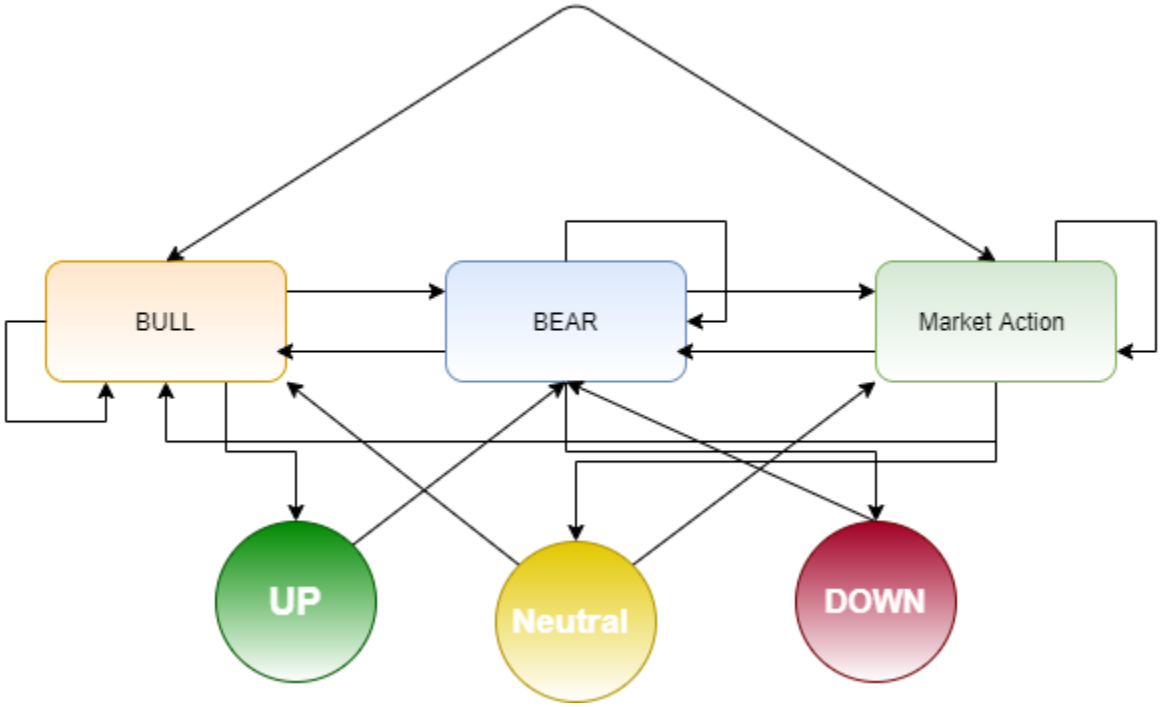

Figure 16. Diagram of Stock market three hidden states in HMM model

The observed states in the model is the class value of the close price, which is showing if it's equal to previous, going up or going down. Therefore, we need to know the initial market state probability to analyze the market, so we can train the historical data and see the price movement. Out hidden states are discrete and has high volatility depending on the market movement. I used the collected data with the actual values collected from Apple and tesla index. I then trained used the Gaussian Mixture model with 3 hidden states to check if it can detect the three states.

The model did not catch the side way movement of the stock as predict in the results. It also did poorly in the up and down movement of the stock. Therefore, it's not a good model to use for the hidden state parameters. A lot of other papers had really good indicators using them, therefore, it must need more data preprocessing to reach that optimum accuracy.

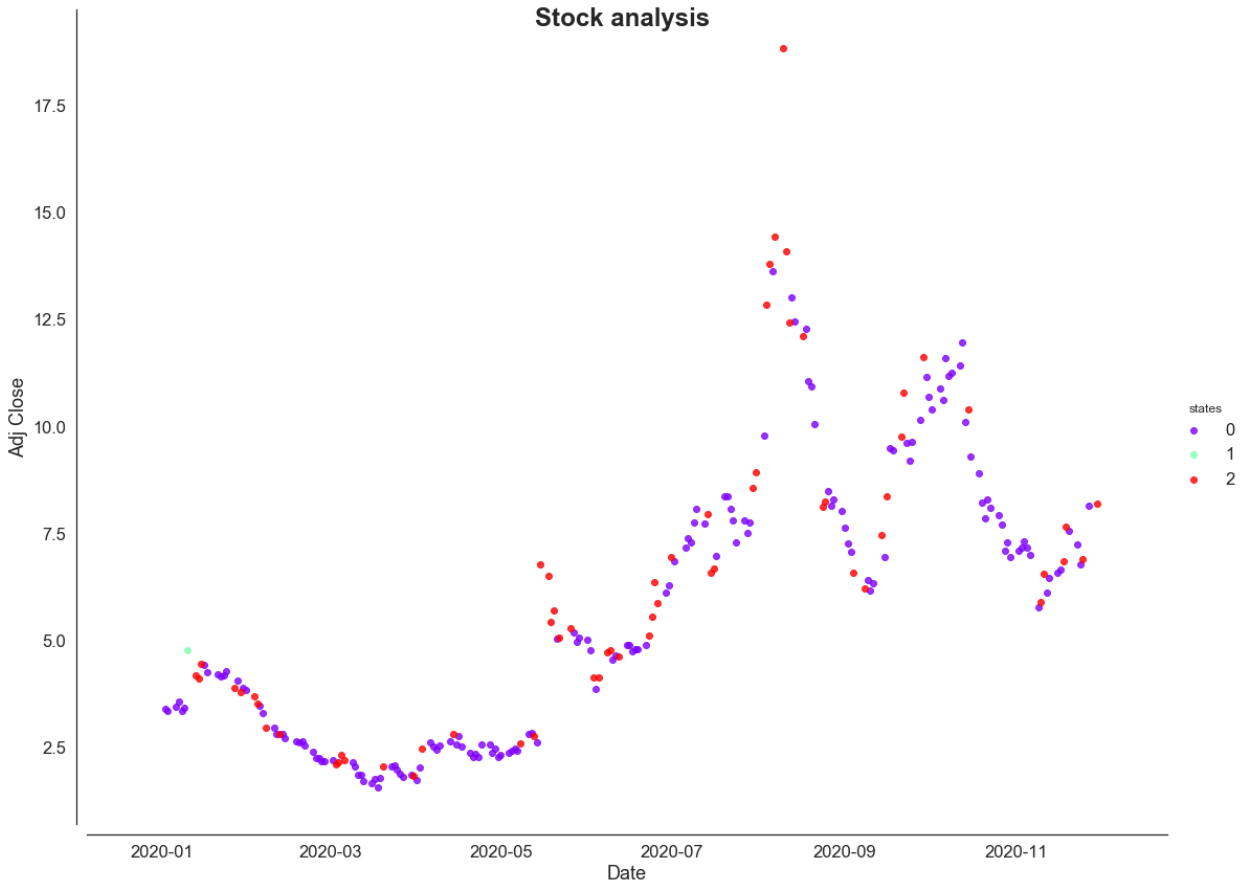

Figure 17. Hidden States versus the price action.

Conclusion:

The results we got from our real-time stock news sentiment trend analysis model showed that the model performed good on mostly all the suggested tickers and preformed significantly better for companies that run on news (specially bio companies with small market capital). The model performed bad on companies that have low amount of daily news which is understandable as the news headlines are the only feature the model is basing its predictions on. This project also showcased the importance of analyzing and optimizing the investment portfolio using optimization methods like Efficient Frontier and Monte-Carlo simulation to pick the best investment portfolios based on historical information. The results we got from classifying the stock trend using Machine learning methods were significantly worse than using the probabilistic modeling techniques this can be either due to feature correlation as pre-training weights needs to be assigned or it can mean that these algorithms are not suitable for stock trend classification. As we approached this problem as a time-series model with LSTM and GRU we were able to significantly improve the results.

However, it is premature to determine that this improvement was accurate only based on the experiments we conducted.

Citations:

1. DeCambre, Mark. "Stock of Tiny SPI Energy Skyrockets 3,100% on Wednesday as It Announces EV Venture." *MarketWatch*, MarketWatch, 23 Sept. 2020, www.marketwatch.com/story/stock-of-spi-energy-skyrockets-3100-on-wednesday-as-it-announces-ev-venture-2020-09-23.
2. Lambert, Lance. "Real Unemployment Rate Soars Past 20%-and the U.S. Has Now Lost 26.5 Million Jobs." *Fortune*, Fortune, 23 Apr. 2020, fortune.com/2020/04/23/us-unemployment-rate-numbers-claims-this-week-total-job-losses-april-23-2020-benefits-claims/.
3. *FINVIZ.com - Stock Screener*, finviz.com/.
4. Cjhutto. "Cjhutto/VaderSentiment." *GitHub*, github.com/cjhutto/vaderSentiment.
5. Cjhutto. "Added Some Words and Phrases for Stock and Political News. by AkashSarda · Pull Request #34 · Cjhutto/VaderSentiment." *GitHub*, github.com/cjhutto/vaderSentiment/pull/34/files.