**Abstract:**

**Introduction:**

This project explored non-probabilistic models for stock trend prediction, different Machine Learning algorithms were applied such as: 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.

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 Finaviz [[Link]](https://finviz.com/). 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) [[Link]](https://github.com/cjhutto/vaderSentiment#citation-information). 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 [[Link]](https://github.com/cjhutto/vaderSentiment/pull/34/files) 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 preformed 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.

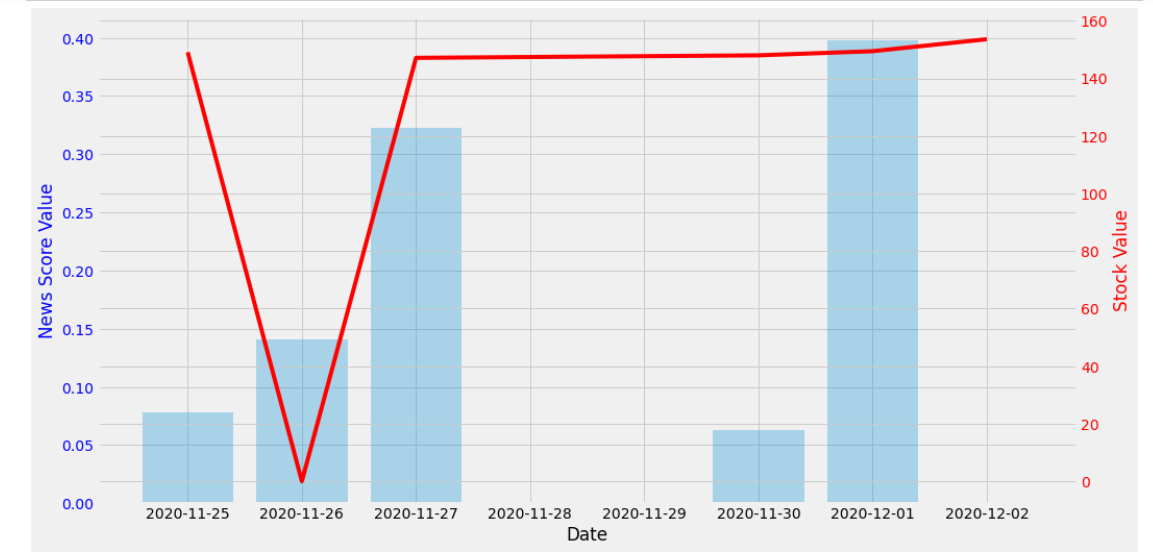
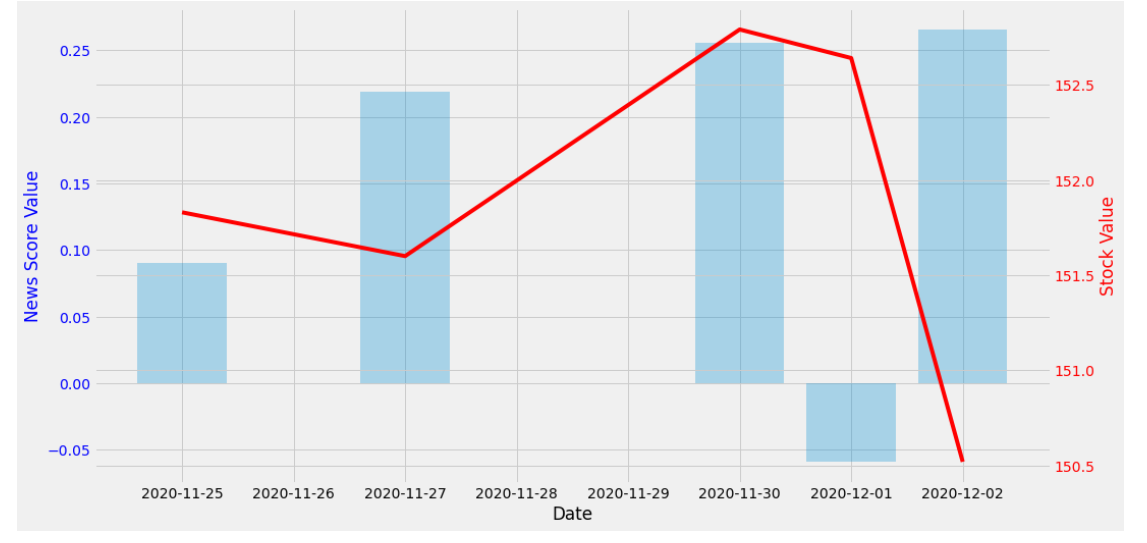


Figure 1 WMT stock against news graph

Figure 2 DIS 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.

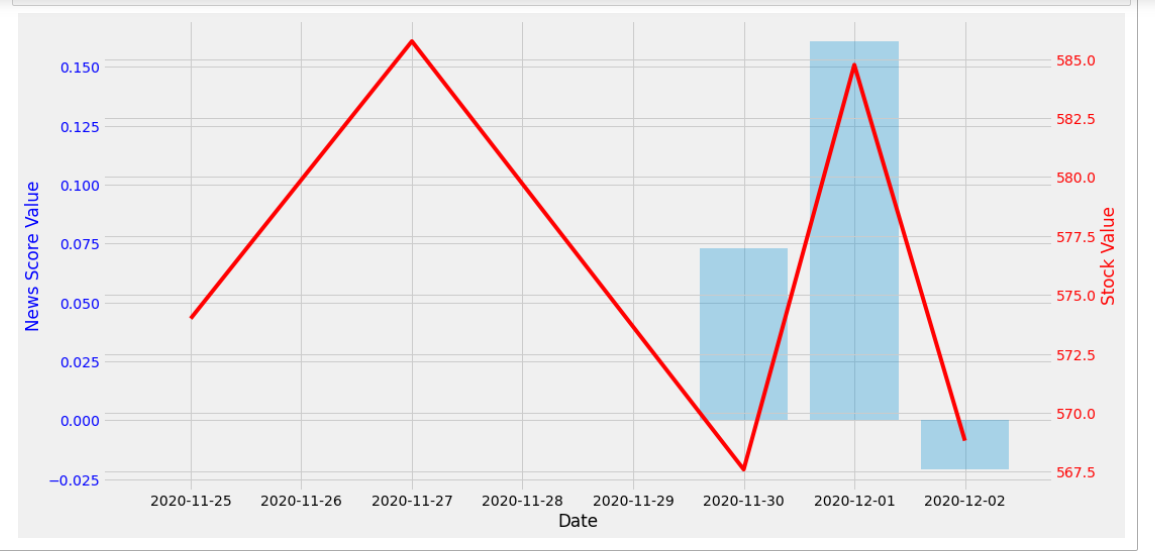
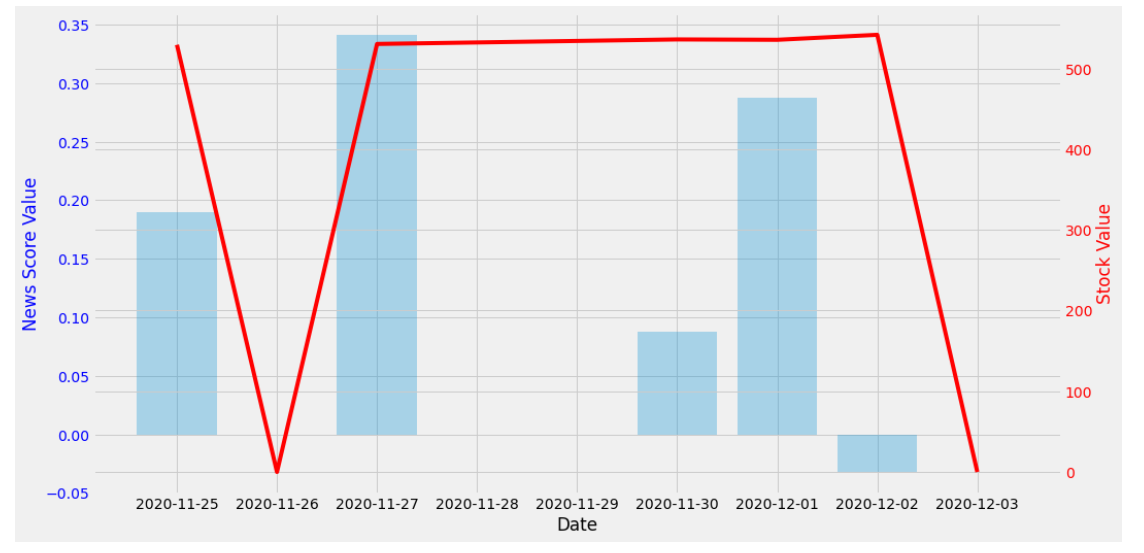


Figure 3 stock against news NVDA

Figure 4 stock against news TSLA

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).

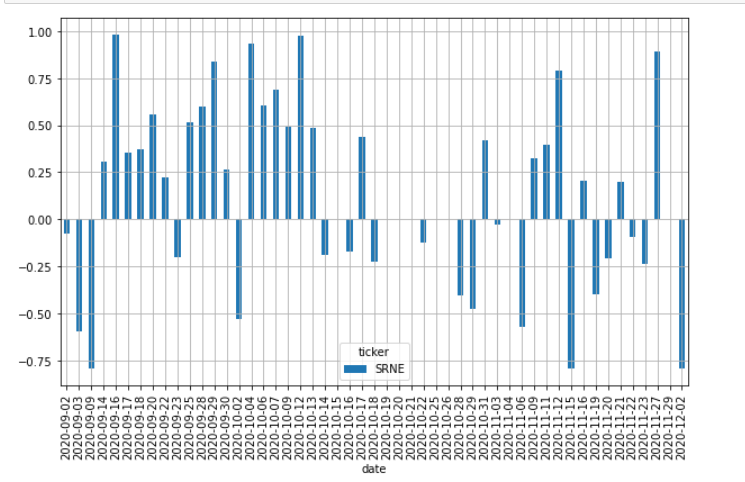
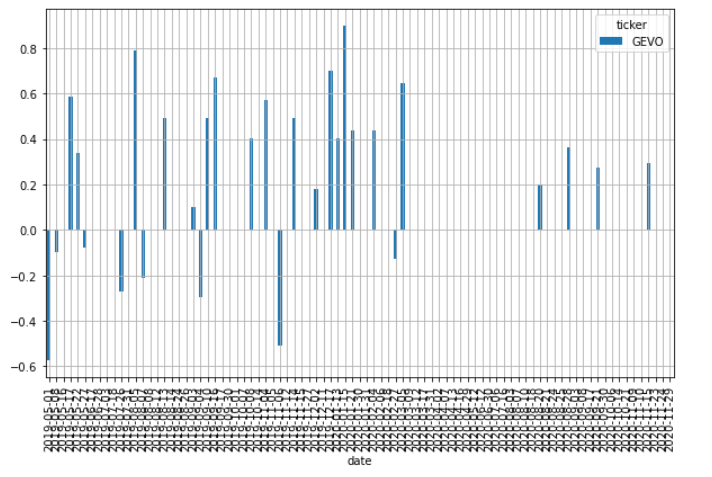


Figure 5 news analysis for SRNE



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 Bio-fuel 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 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).

|  |  |
| --- | --- |
| Expected return | 0.911 |
| Volatility | 0.462 |
| Sharpe Ratio | 1.797 |

Table ER vol and SR for initial portfolio

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.

|  |  |
| --- | --- |
| 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 |

|  |  |
| --- | --- |
| 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 |

Table 2 Skewness and Kurtosis for each ticker

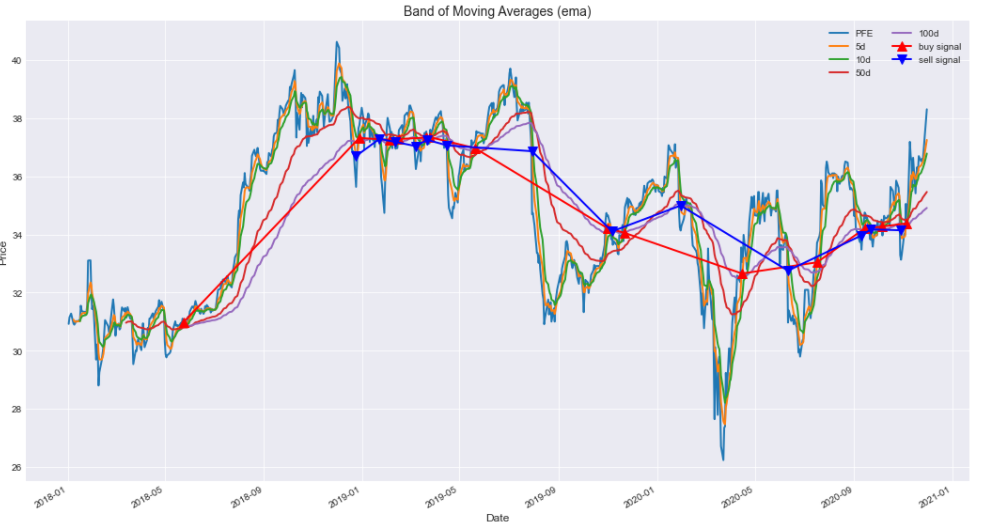
We also analyzed the buy-sell signals for each ticker based on the moving averages (5-day, 10-day, 50-day, 100-day).

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

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

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

Table Optimized portfolio for min Volatility

Table Optimized portfolio for Max Sharpe Ratio

|  |  |
| --- | --- |
| 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 |

|  |  |
| --- | --- |
| 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 Optimal Weights min vol

Table Optimal Weights min vol

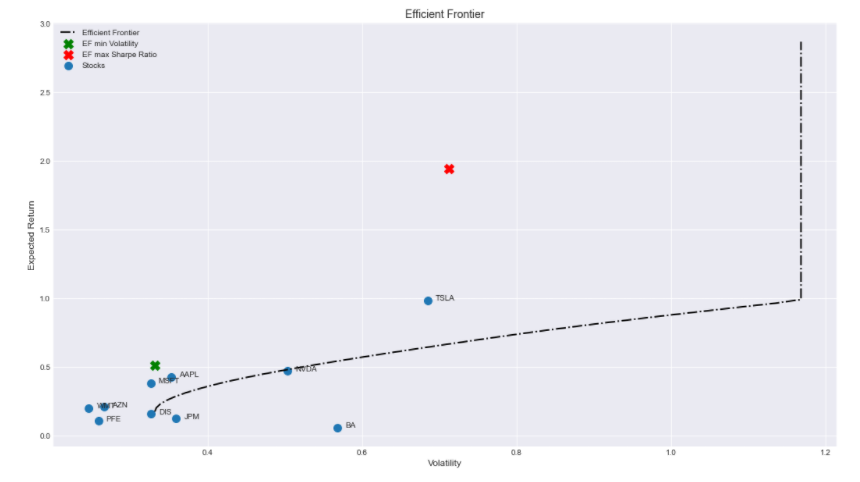


Figure 8 EF portfolio optimizer

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 Monte-Carlo portfolio optimization

**Stock Prediction Using Machine Learning Algorithms**

In an effort 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.

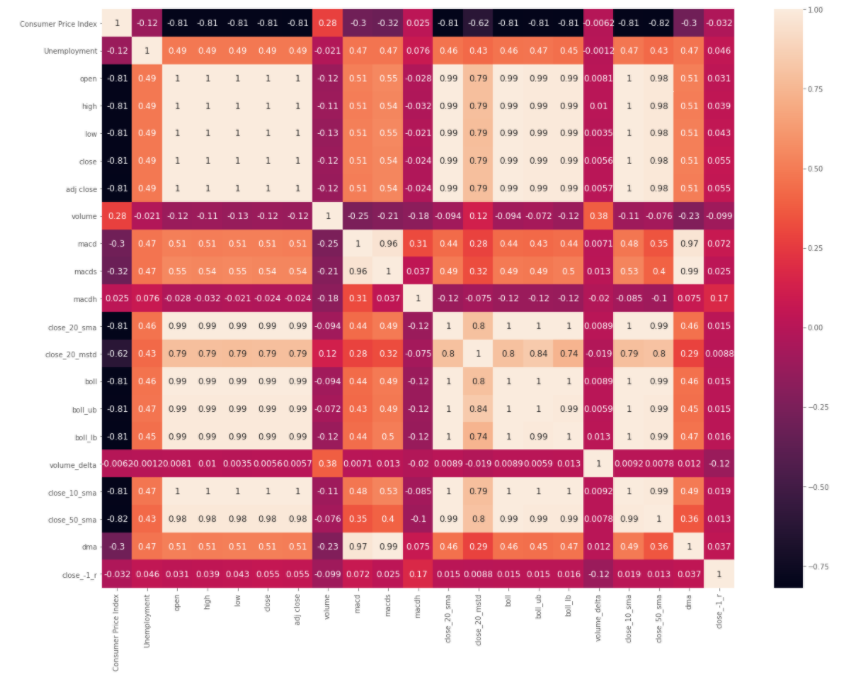


Figure 10 features correlation Confusion Matrix

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:

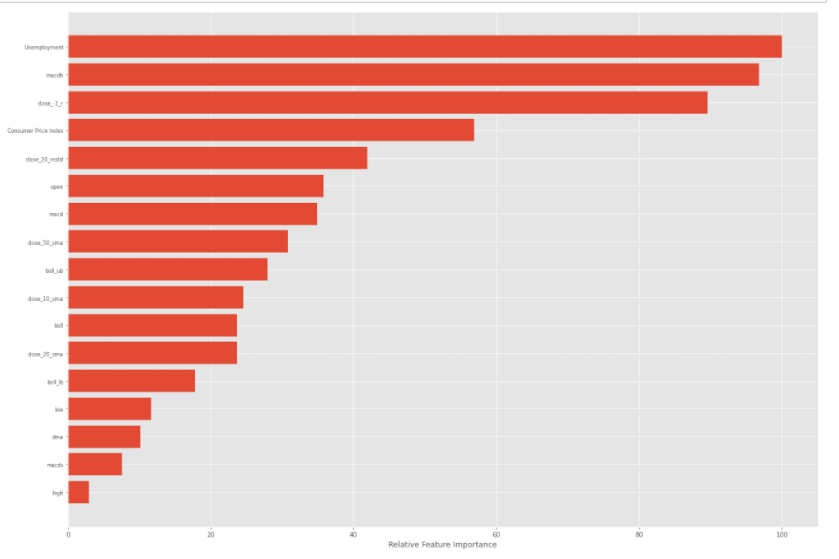


Figure 11 Feature Importance Analysis with LR

**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.

The 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.

A picture containing text, electronics

Description automatically generated

**Figure. 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. 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 |

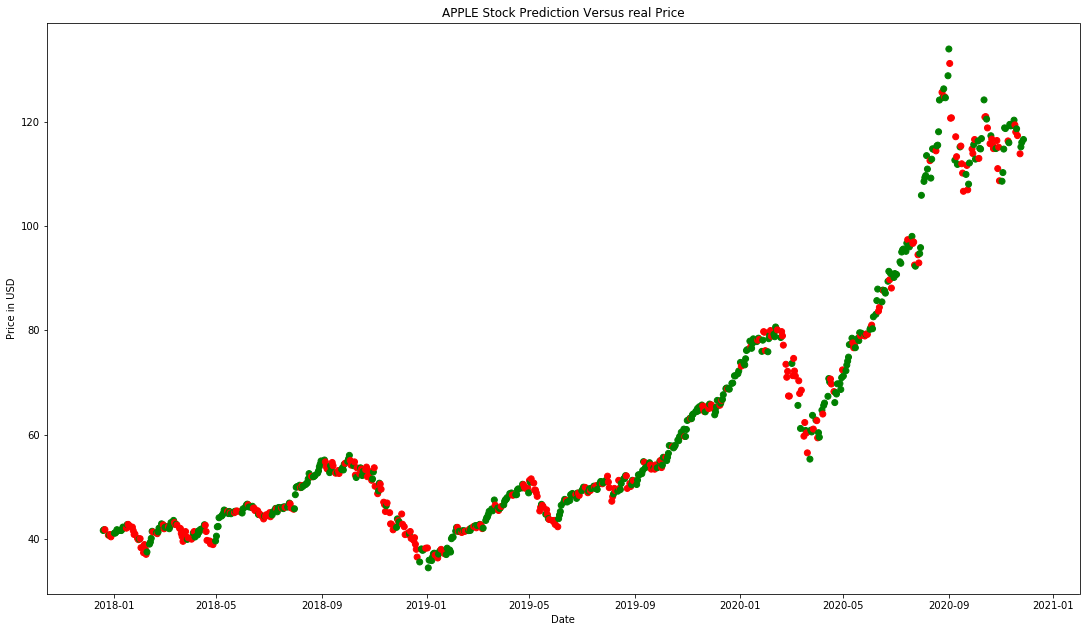
**Table. Example set of Apple stock on up-trend learned from historical data**

|  |  |  |
| --- | --- | --- |
| **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. 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. Apple stock prediction movement versus the real price action**

**Hidden Markov model:**

**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.