**Predicting Stock Prices using Top News Headlines**

Spandan Gupta

Spring 2020

**Motivation, Problem statement:** Considering the drastic changes in the economy currently, I worked on predicting the stock market depending on the daily news headlines. 1000 top new headlines, upvoted by Reddit users was chosen to tally it with DJIA historical stock data.

With DJIA being the index for blue-chip stocks in the USA and since the economy is equally going bad as many other nations, I decided to start this analysis by choosing on USA News from Reddit. [1]

**Literature review:** Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange [2]. The prediction methodologies are either fundamental analysis or technical analysis methods.

Fundamental analysis is a method of measuring a security's value by examining related economic and financial factors. Factors ranging from the state of the economy and industry conditions to microeconomic factors like the effectiveness of the company's management. [3]

Technical analysts focus on patterns of price movements, trading signals and various other analytical charting tools to evaluate a security's strength and weakness. This analysis is carried out on historical trading data which includes stocks, futures, commodities, fixed income, currencies, and other securities. [4] Futures [5] are derivative financial contracts that mandate the parties to transact an asset at a predetermined future date and price.

In this project we will be working with time-based data, which will involve working on time-based data (years, days, hours, minutes) to derive hidden insights to make informed decision making. Time series model are used when we have serially correlated data. Since this is my first project where I will be working with time series data, understanding all the basic concepts of this field is important.

Starting with stationary series. The mean of the series should be a constant. The variance of the series should not be a function of time too. This property is known as homoscedasticity. Variance is the average of the squared differences from the Mean. [6] Homoscedasticity , is a situation when the error term is the same across all values of the independent variables. [7] Covariance of the i th term and the (i+m)th term should not be a function of time. Covariance provides a measure of the strength of the correlation between two or more sets of random variates. [8]

Past works relating to prediction of stock prices are referenced to an article by Aishwarya Singh on Analytics Vidhya. She started out by calculating RSME value, then moved ahead by implementing linear regression which performed poorly. k-nearest neighbor as well did not perform well. Auto Regressive (AR) Intergrated(I) Moving Average(MA), this performed better than the past two algorihms. Facebook's time series forecasting library Prophet also gives an analysis from which it is evident that, stock market is dependent on what is currently going on in the market. Lastly, as my many people LSTM was used. Long Short Term Memory, worked best as compared to the ones used, but it is not said to be the best approach. [9]

**Exploratory Data Analysis and Implementation:** Reddit has hot, new, controversial, top and gilded sections of news. I chose the top 1000 news headlines. At this point, using Python I scraped this data from Reddit. I referenced Felippe Rodrigues article [10] to complete this step.

* These 1000 headlines were from 8-30-2014 to 04-05-2020. Thus, I extracted DJIA historical data also of the same timeline.
* Learning from Aaron7sun's [11] work I too combined the News headlines and the DJIA data by labeling it. The days when the Adj closing price decreased, labeled 0, the days when it increased labeled them 1 and the days for which the historical stock price was not available labeled those days 2.

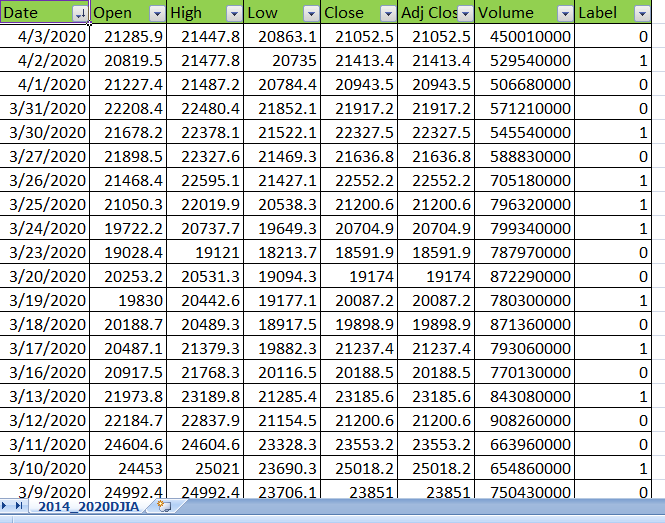


Figure 1: DJIA Historical Stock Data

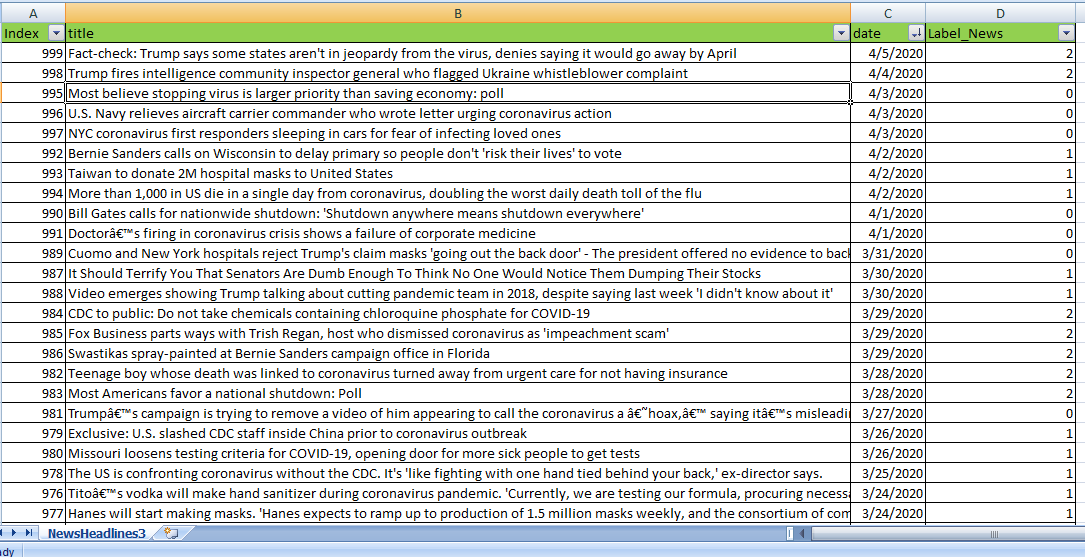


Figure 2: Top News headlines on Reddit

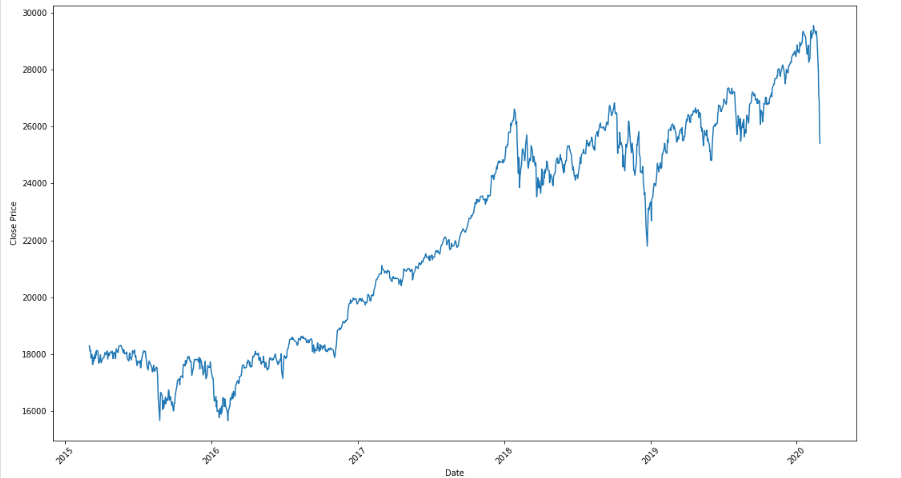


Figure 3: Shows the closing prices of DJIA varying over 2015- 2020

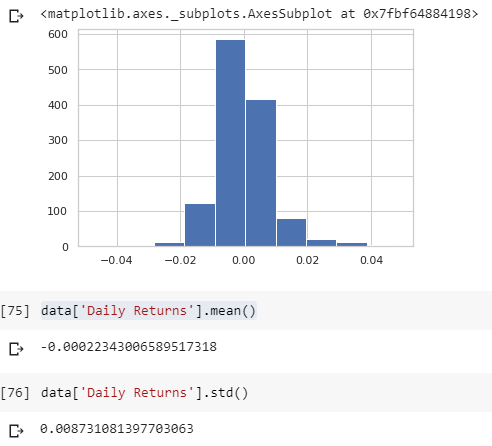


Figure 4: Shows the daily returns on stocks

* The train test split was done as 70:30 ratio.

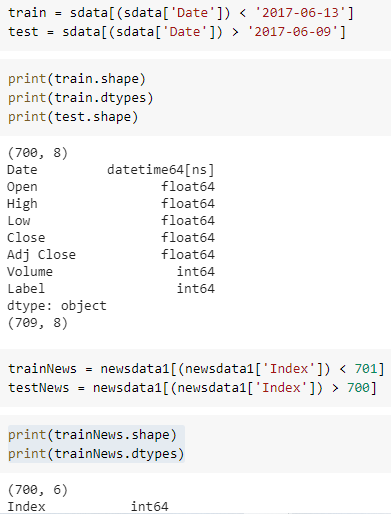


Figure 5: Train Test split

* Since here text was going to be analyzed, natural language processing came into picture. Thus, count vectorizer was first implemented. Count Vectorizer, counts the number of times the keyword appears.

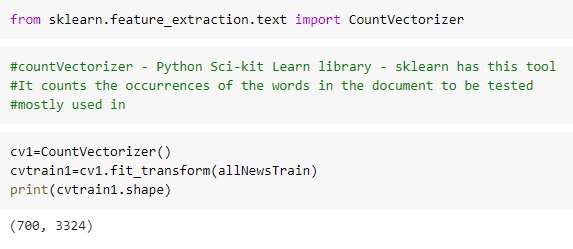


Figure 6: Count Vectorizer

* Next Logistic Regression accuracy score was 0.387 Logistic Regression was chosen since it is classification scenario.

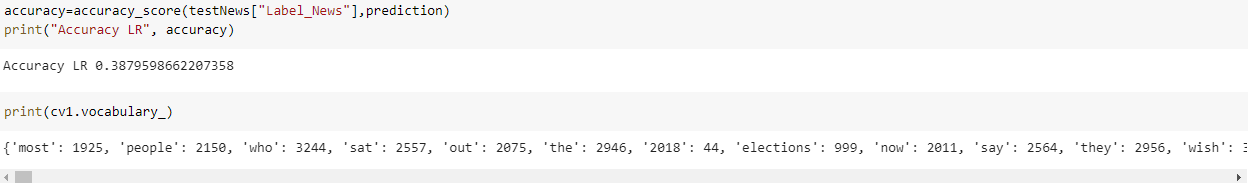


Figure 7: Accuracy score Logistic Regression

* Next again accuracy score was calculated after implementing TF-IDF : term frequency - inverse dense frequency.
* This was chosen because frequently used had to be omitted to avoid redundancy. TF-IDF function has parameters which modulated and filter words. Words having less than 0.01 frequency can be omitted and words having greater than 0.6 frequency can be omitted. The maximum count of keywords can also be set. I tried changing these parameters but surprisingly the accuracy score again came 0.411

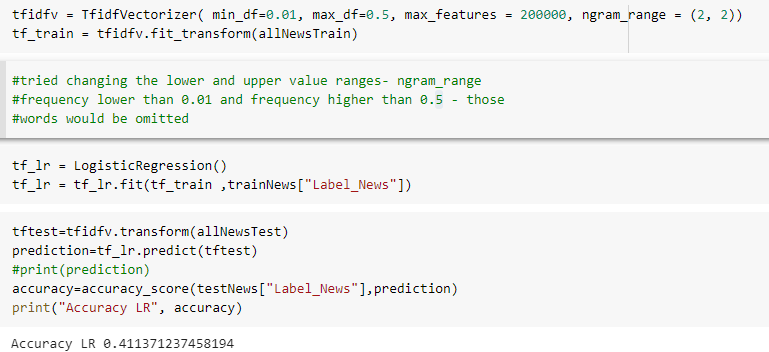


Figure 8: TF-IDF

* Next, I ran Naive Bayes on these datasets. It is also a supervised classification algorithm. This algorithm is built on Bayes theorem. The probability that a feature is dependent on another feature. The accuracy score stayed the same, 0.4113.



Figure 9: Naïve Bayes

* Random Forest Classifier was next implemented. The score decreased to 0.36

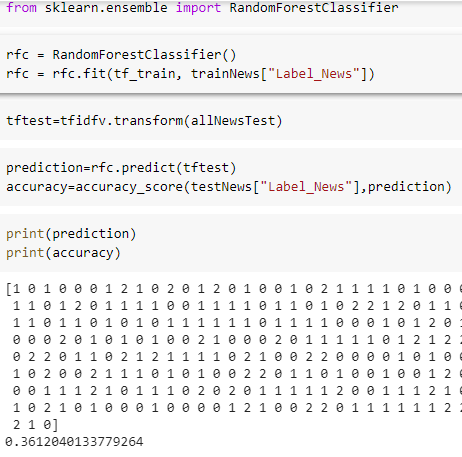


Figure 10: Random Forest Classifier

* Another try, by tweaking the parameters gave 0.411 score.
* Next, Gradient Boosting reaped as accuracy score 0.414 It seems to not aiding to a better performance even after trying out so many models.

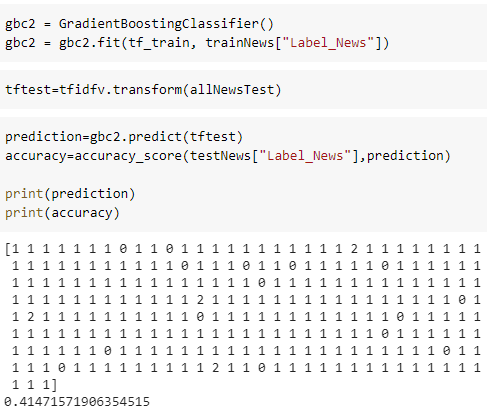


Figure 11: Gradient Boosting

* Thus, tried Multi-layer feed forward network. Deep Neural Network

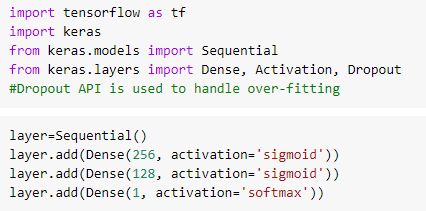


Figure 12: Sequential Multi-layer network

* A neuron accepts input signals via its dendrites, which pass the electrical signal down to the cell body. The axon carries the signal out to synapses, which are the connections of a cell’s axon to other cell’s dendrites.
* The feed forward neural network, backpropagation and activation and loss function relating to this scenario could have aided into a better prediction of test data. With this thought I proceeded towards its implementation.
* Used the Sigmoid activation function as we are expecting output as [0,1]. Also, since a probability distribution is concerned at the end layer Softmax function used.

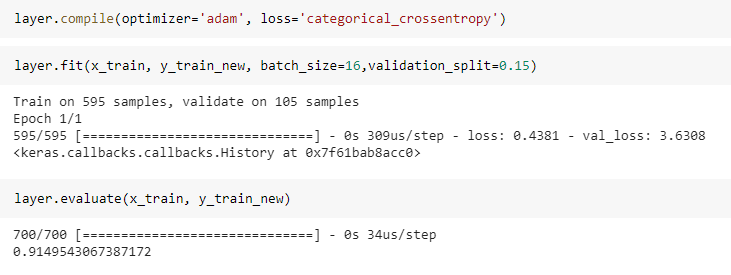


Figure 13: Evaluation of Feed Forward Neural Network

* A last try of LSTM.

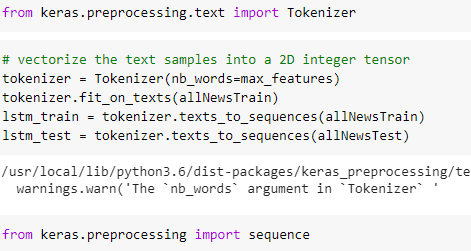


Figure 14:LSTM

* An accuracy of 0.408 was observed

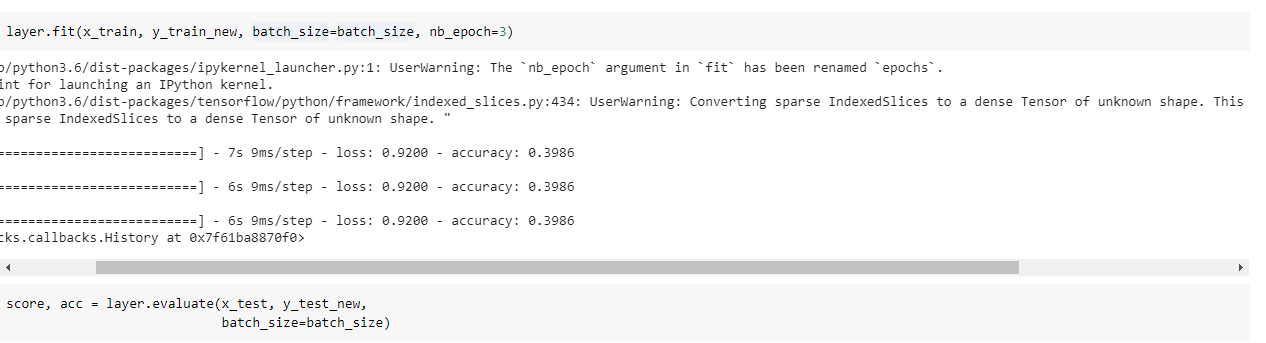


Figure 15: Fitting of LSTM

* Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. [9]
* More the complexity, more the features.
* During back propagation, recurrent neural networks suffer from the vanishing gradient problem. Gradients are values used to update a neural networks weights. The vanishing gradient problem is when the gradient shrinks as it back propagates through time. If a gradient value becomes extremely small, it doesn’t contribute too much learning. [10]

**Conclusion and Future Work:** As understood from the accuracy score achieved, it states the fact that stock prices are the most difficult to predict. Also probably,

The Future Work would be figure out other factors. Link it with a company’s background. Since now I have got a grasp of this, within a week or two I have plan on implementing the prediction of stock prices, based on backgrounds of company. Next step will be to answer questions like, what was related news of Apple company since it’s stock prices dropped. Similarly for Amazon and Facebook, if they are doing good these days, what are the relevant news headlines about them now?

**References:**

[1] Reddit USA News, <https://www.reddit.com/r/usanews/>

[2] Stock Market Prediction, <https://en.wikipedia.org/wiki/Stock_market_prediction>

[3] Fundamental Analysis, <https://www.investopedia.com/terms/f/fundamentalanalysis.asp>

[4] Technical Analysis, <https://www.investopedia.com/terms/t/technicalanalysis.asp>

[5] Futures, <https://www.investopedia.com/terms/f/futures.asp>

[6] Variance, <https://www.mathsisfun.com/data/standard-deviation.html>

[7] Homoscedasticity, <https://www.statisticssolutions.com/homoscedasticity/>

[8] Covariance, <http://mathworld.wolfram.com/Covariance.html>

[9] Aishwarya Sigh, Analytics Vidhya, <https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/>

[10] How to scrape Reddit with Python, <https://www.storybench.org/how-to-scrape-reddit-with-python/>

[11] Reddit Time Machine, <https://github.com/sjhddh/reddit_time_machine>

[12] LSTM, https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

[13] LSTM Guide, https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21