Comparison of Logistic Regression and Long Short Term Memory in Stock Market prediction

Andrew Garcia
Department of Computer Science
University of South Carolina Upstate
Spartanburg, South Carolina, USA
andrewag@email.uscupstate.edu

ABSTRACT

The New York Stock Exchange has existed since 1792 when it was created to set rules on how stocks could be traded and established. Although without in depth knowledge how does someone who has no experience in investing start? This is the purpose of this study, to help the everyday person who wants to gain financial literacy through investing. Over the past couple years machine learning has become a forefront in the computer science field, its used all over the world in many manners. Therefore I will be conducting a study to evaluate two of the most used algorithms in stock market prediction, logistic regression and long short term memory. I will measure the accuracy, consistency, and complexity to try to come to a conclusion on which machine learning algorithm is best for stock market prediction.

Keywords

Machine Learning, Logistic Regression, LR, Long Short Term Memory, LSTM, Multivariate, RMSE

1.INTRODUCTION

Since the creation of the stock market, it has become an integral part of the United States business financial capital, giving them the ability to build successful companies and through the companies a stable and wealthy country. As there is a deep relationship between the stock market and the economy. The economy directly impacts the stock market in different ways such as inflation, in that it effects share prices, and in the same aspect the stock markets change affects how much people and companies spend [6]. This relationship can be seen proven throughout history, such as the stock market crash of 1929 which caused the great depression, the largest and most impactful economic recession the United States has ever had [9]. Through the Stock Exchange the United States has been able to flourish financially and has become the country we live in today, but not without its hardships. Although the Stock market plays a big role in the economy, it does not only exist for the wellness of businesses and the financial wellness of the country, it also plays an important role in the wealth of the people of the United States.

The richest Americans throughout history own most of the stock market, but this doesn't mean nobody else can become wealthy through investing in stocks. With time more and more people across the country have begun to seek financial literacy in the stock market, since we are in the age of the internet the secrets of wealth in the stock market cant be kept a secret as it was in the past. Throughout history millions of people have become millionaires off of the stock market alone. In fact more than a million people just last year became millionaires from the stock market [8]. Through these facts who wouldn't want to invest in the stock market, as it is available for anyone in the United States to invest in, although it is not as easy as it sounds. The stock market is unpredictable without in depth knowledge of how stocks behave over time and economic fluctuations. Even with in depth knowledge and experience, without computers to analyze data through in depth algorithms the stocks are too unpredictable. As seen through the stock market crashes in the past, which have been caused by pure speculation and panic selling which is all human error.

Despite these facts, stock market trading is not only for those that have years of experience or someone to teach them the inns. This may have been the fact in the past but times have changed, the creation of the computer opened a door for millions of more people to enter stock trading. On February 8, 1971 the stock market became electronic which allowed for lower fees, more efficient markets, and more information for investors [7]. Not only did the internet benefit the stock market for experienced investors but it also benefits prediction for stock market crashes. With the algorithms created in recent years the stock market has become safer and easier for regular investors to get into. This study will attempt to cover the most optimal algorithm for stock market prediction by comparing two of the most used algorithms in stock market prediction today, Logistic Regression and Long Short Term Memory.

2.LITERATURE REVIEW

Given that the stock market is such an important topic for not only the United States economic stability but also any other countries economy, there is a plethora of stock market prediction papers involving machine learning from stock markets all over the world. Since I have a limited time to work on this project I will limit my research and data set to the United States alone. Even though machine learning is such a prevalent theme in stock market prediction, Logistic Regression and Long Short term memory isn't regularly compared against each other. More articles and journals highlight LSTM by itself or compared to other algorithms which could imply LSTM is opted for more than LR.

Logistic Regression is known as a multivariate analysis model, which is a model that needs two or more variable amounts[2]. This research is shown in [2], where it is discussed that different types of variables can be used in this model, not only normal ones. In the past research has shown that Logistic Regression is very reliable in predicting the downfall to a business and other financial downfalls as shown in [1]. However it is shown in [1] that in recent years more people have been trying to prove that the Logistic Regression model also can prove useful in regular stock prediction as well. As Logistic Regression is a model that is used to estimate the probability of a binary outcome given multiple values [10]. Thus with this idea, if the model if it is trained to analyze stock history it can predict stock change, with the binary factor being whether it will rise or fall as discussed in [10]. This is shown in [1] and [2] where their research discusses the value of the simplified logistic regression equation, which is estimated for maximum likelihood for stocks. Which is

$$Z = \log (p/1\text{-}p)$$

where p is the probability that the outcome is good. The authors of both [1] and [2] show in their research that the Logistic Regression equation proves useful. With [1] research showing that LR has 89.7% accuracy for predicting poor performance and 87.2% accuracy for predicting good performance with 88.4% overall accuracy. Although research on [2] shows a little lower accuracy with 81.55% overall accuracy. Despite the fact that [1] and [2] differ in percentages, the margin is not great and both with over 80% overall accuracy, which proves LRs usefulness/accuracy in the prediction of stocks.

LSTM is an entirely different machine learning algorithm, in that LR is referred to as a Multivariate algorithm and LSTM is referred to as a recurring neural network algorithm. Researchers in [3], [4], and [5] all discussed its key difference from other machine learning algorithms. They are known as recurrent neural networks that can lean order of dependence in sequential prediction problems [3]. [3] Also goes on to discuss LSTMs uses such as machine translation and speech recognition due to its features for complex problems. As LSTM has the ability to store past information, which helps with stock price prediction. This is seen in the structure of LSTM, as seen in Figure 1, which is discussed in [3] and [5]. It is comprised of four components: a cell, input, output, and forget gate. The cell retains information while the other three gates control the flow of information, both in and out.

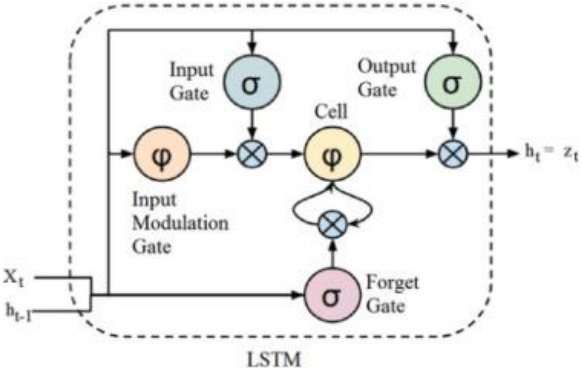


Figure 1. LSTM Structure [3]

The differences between LR and LSTM are clear, LSTM has a definitive advantage over LR, as LSTM is able to store data for a certain amount of time while other gates operate input and output of other information. Researchers in [4] make this difference clear as they put this theory into practice, by using the Root Mean Square Error(RMSE). Which is the square root of the mean of the square of all error. The use of RMSE is highly used within the prediction community, as it makes a great general purpose error metric for predicting numbers.

RMSE measures the average difference between the actual values and the values predicted by a model, in this case LSTM.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum\left(\hat{Y}_i - Y_i\right)^2}$$

The equation as seen above, which the researchers in [4] then trained with a dataset that consisted of National stock exchange with a window size of 22 days. This resulted in a training RMSE of 0.00983 and testing RMSE of 0.00859. These numbers prove the equation is highly effective as the closer the outcome of RMSE is to zero the more accurate it is, since zero means that the test is a perfect fit for the data. All research articles prove that both LR and LSTM have great benefits with each respective strengths, although to definitively say which is better for stock market prediction I will have to put the research into practice and experiment with my own datasets.

3.METHODOLOGY

3.1 Logistic Regression

As stated previously, Logistic Regression classifies its information using a binary method, it makes use of this method using a linear function based on the variables from provided data. Fundamentally LR estimates the probability of a 0 or 1, which in this instance, is the probability of the observed stocks going up or down. Our models variables will be the stocks date and closing price. The closing price is being used because it gives the model an accurate view of what a stock is being sold at on the individual day. The model evaluates the history of the stock data based on if it has gone up '1' or down '0' since logistic regression assumes a binary approach to its functionality. Although these results are not easily derived from the data, it is an estimated probability, the model selects a 0 or 1 by rounding based off the data. If the output is greater than 0.5 then the result will be a 1 and if it is less than 0.5 it will be a zero.

In the case of Logistic Regression, unlike other machine learning algorithms which predict variable y based on one or many predictor variables x, LR predicts the probability of y occurring, given values of x. Logistic in LR is derived from "log odds" which is the probability that is modelled. In other words LR forecasts the odds ratio of the outcome based on the dichotomous dependent variables. These dependent variables are known as log odd or logit, natural log of odds. Hence the odds ratio is used to determine the odds of success or failure, where a values less than on shows that the case is not likely attainable and a value greater than one show that it is. These equations are seen below.

$$\frac{\pi(x)}{\pi(1-x)} = [\exp(-X^{T}\ \beta]^{-1} = \text{odds}$$

This equation shows taking natural log on both sides.

$$\log\frac{\pi(x)}{\pi(1-x)} = [(-X^{T}\beta]^{-1}$$

The second equation is a transformation in log which is known as logistic transformation.

Although these equations prove quite useful for determining the odds of LR, they are not the only equations used for linear functions. The Sigmoid Function, which is a Logistic Function is unique in its ability to give results for classification which the odd equation could not. The graph and equation can be seen in Figure 3, in which z= ∞, $g(z)$ =1 and z= $-\infty$, $g(z) = 0$. This can be expressed with the function $-\infty < x < +\infty \Longrightarrow 0 < \sigma(x) < 1$.

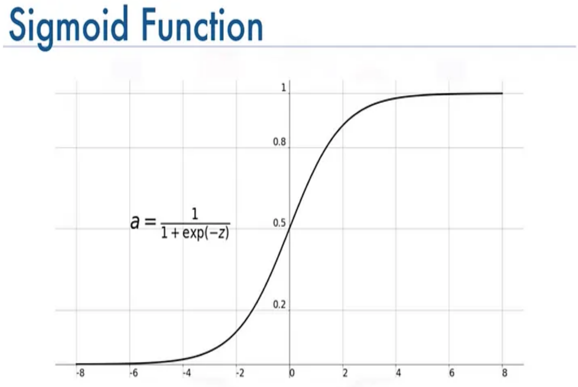


Figure 2. Sigmoid Function [12]

Figure 3 clearly illustrates that the upper and lower bounds of the Sigmoid Function equation are separated by the 0.5 rule of LR as discussed previously. Through this equation I am able to calculate the linear function of the variables such that it creates a curve best fit for the relationship between the independent and dependent variables. Which brings us to our third and final equation of the Logistic Model which is

$$zi = c + \beta1xi1 + \beta2xi2 + \cdots + \beta p\ xi$$

Where z = log(p/1-p) and p is the probability that the outcome is good. The final Logistic Regression equation is estimated by using the maximum likelihood estimation for classifying stock performance [1].

By definition, LR seems to have more of a disadvantage than other machine learning algorithms since its linear function is restricted by its output, being that its only possible results are 0 and 1. Although as described above, with the use of various equations and efficient use of variables, Logistic regression can be just as applicable if not more than other machine learning algorithms.

3.2 Long Short Term Memory

While Logistic Regression is known as a machine learning algorithm, LSTM is in a different league of its own. Rather than classifying as a machine learning algorithm, it is in a deeper subclass known as deep leaning, more specifically RNN deep learning algorithm. Numerous deep leaning algorithms have been created to deal with various problems along with dataset structures. However, the main issue with deep learning is its inflexibility, as information can only move in one direction, forward. Moreover, since each input is processed independently, information cannot be retained from previous steps, therefore these deep learning algorithms are ineffective for problems that involve sequential data, such as stock history [4]. However RNN is designed specifically for these problems. It utilizes a strategy called backpropagation in which adjusting parameters literately, from the outer layer all the way to the inner layer [4]. In other words it finds a gradient for all parameters. RNN is better than other neural networks when it comes to preserving information but it is not effective in long term dependents due to the vanishing gradient issue [4]. This is where LSTM can show its effectiveness, with its memory cells it is able to overcome the vanishing gradient issue. LSTMs model is illustrated in Figure 3 in which it is designed to model sequential input in time t.

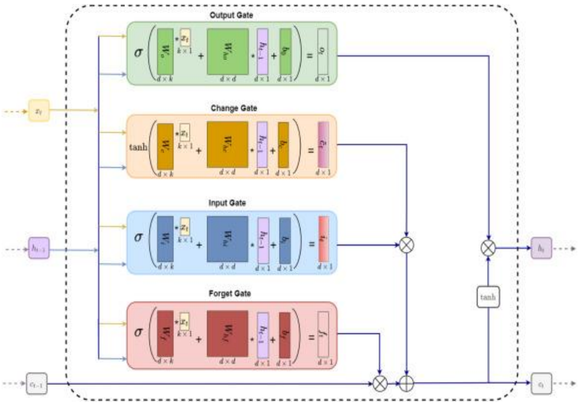


Figure 3. LSTM Architecture [4]

Where for a given input sequence at time t

$$\{x_1, x_2, \ldots, x_n\}, x_t \in \mathbb{R}^{k \times 1}$$

The memory cell ct updates the information using three different gates, input it, forget ft, and change ct gate. While the hidden state ht is updated using the output and memory cell gates ot and ct [4].

$$\begin{aligned}
i_t &= \sigma\left(W_i x_t + W_{hi} h_{t-1} + b_i\right), \\
f_t &= \sigma\left(W_f x_t + W_{hf} h_{t-1} + b_f\right), \\
o_t &= \sigma\left(W_o x_t + W_{ho} h_{t-1} + b_o\right), \\
\tilde{c_t} &= \tanh\left(W_c x_t + W_{hc} h_{t-1} + b_c\right), \\
c_t &= f_t \otimes c_{t-1} + i_t \otimes \tilde{c_t}, \\
h_t &= o_t \otimes \tanh\left(c_t\right)
\end{aligned}$$

LSTM cell inputs three different pieces of information, current input sequence xt, short term memory from previous cell ht-1, and long term memory from previous cell ct-1[4]. The forget gate takes in the information from xt and ht-1 and outputs a 0 or 1, a 0 being that it forgets all information from previous cell and 1 being that it stores all the information from previous cell ct-1.

Deep learning algorithms along with Recurrent Neural Networks have been developed to solve various problems but they still failed to solve long sequential issues such as language processing or musical analyzation. LSTM makes use of deep learning and RNN techniques in combination with its long chain rule. Through LSTMs long memorization architecture it is built specially for sequential problems such as stock prediction.

4.IMPLEMENTATION

The code for the implementation of both the Logistic Regression and Long Short Term Memory are written in Python 3.9 in Pycharm community edition. Python has various libraries in pycharm to aid with the implementation of these two complex algorithms. The following libraries are used for Logistic Regression implementation: Pandas for data analysis, Numpy for operations on metrics, Matplotlib for graphing the data and results, Sklearn for implementation of LR. LSTM has some slight differences in libraries but they both share Pandas, Numpy, and Matplotlib, and Sklearn, the only difference is tensorflow.keras. Since LSTM is much more complex, it utilizes more libraries than LR for implementation.

5.EXPIRIMENTAL SETUP

5.1 Dataset Analysis

The dataset I chose for this study is a combination of multiple datasets. Since we need multiple stock history datasets we will gather three to five large stock datasets such as Netflix, Apple, and Google to have a variety of data. I will analyze these datasets to reach a better understanding of the data that I am training my models with. The readings recorded differ between each dataset but each are within the ranges of 2005 and 2020 with some datasets spanning fifteen years and some spanning five to ten years. The datasets also differ in the amount of datapoints each dataset has, with some having as little as 1000 rows and others having as much as 3700 rows. All datasets of course have differing features which vary from 6 to 40 features but for this study we are only interested in selective features thus any extra are useless. Although the datasets may differ in these aspects the main features that every stock dataset has is Date, Open, High, Low, Close, and Volume.

Everything in this our datasets are numerical values as I am looking at stock information, the only thing that can be considered a string in our code can be the date, which I am changing into a float for the model to be able to interpret. A visualization of these features are seen in table 1, which were taken out of the Netflix dataset.

Table 1. NFLX Snippet [4]

Date	Open	High	Low	Close	Volume
2/5/2018	262	267.9	250.03	254.26	11896100
2/6/2018	247.7	266.7	245	265.72	12595800
2/7/2018	266.58	272.45	264.33	264.56	8981500
2/8/2018	267.08	267.62	250	250.1	9306700
2/9/2018	253.85	255.8	236.11	249.47	16906900

As seen from the snippet above which was taken from the Netflix dataset, our model requires minimal information for prediction. Since I am studying stock market data, there are no huge outliers within my data as each stock from day to day caries within reasonable values.

5.2 Evaluation Metrics

To evaluate my two models, Logistic Regression and Long Short Term Memory I am going to evaluate them based on three evaluation metrics. These metrics consist of accuracy, consistency, and root mean square error. With RMSE meaning the average difference between a models predicted values and the actual values of the data. These three were chosen as I felt they would be the best three metrics to measure how well the models work when put against each other.

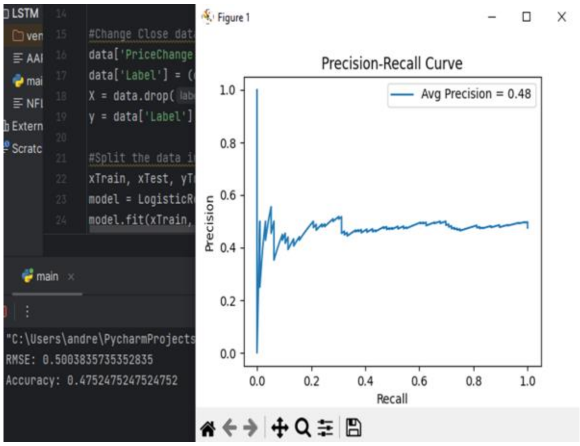
5.3 Experimental Steps

My study begins with my dataset, which is taken from several sources on Kaggle.com. These datasets were imported on pycharm community edition which is written all in python. After each dataset is imported into pycharm, our models fitted with the datasets. Since our two algorithms are different in their model process, they will require different means in plotting the data. Logistic Regression takes in the data and changes the values from numerical values to binary in order to fit its model. While LSTM takes in the data completely and models it accordingly.

6.RESULT ANALYSIS

As stated previously, the results of this study are based on several major companies stock history, and using the history I was able to predict the stock. Both Logistic Regression and Long Short Term Regression were trained and tested with the same stock data to give both machine learning algorithms a level playing field. I used three different stock data for the training and testing which included Netflix, Google, and Microsoft.

6.1 Logistic Regression Results

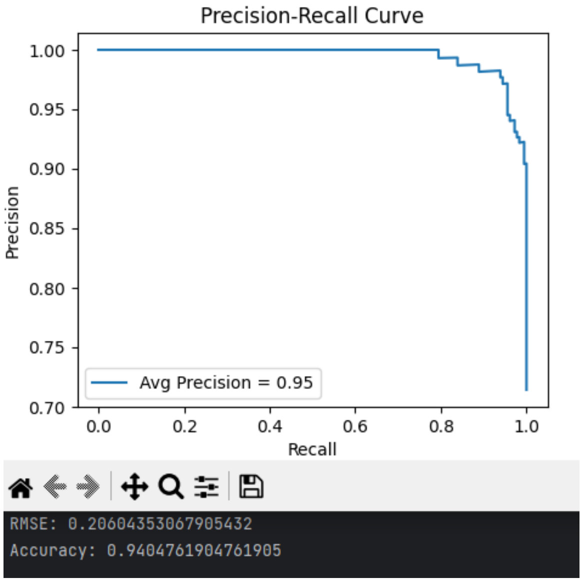


```
#Change Close dat
data['PriceChange
data['Label'] = (
X = data.drop( lab
y = data['Label']

#Split the data i
xTrain, xTest, yT
model = LogisticR
model.fit(xTrain,
```

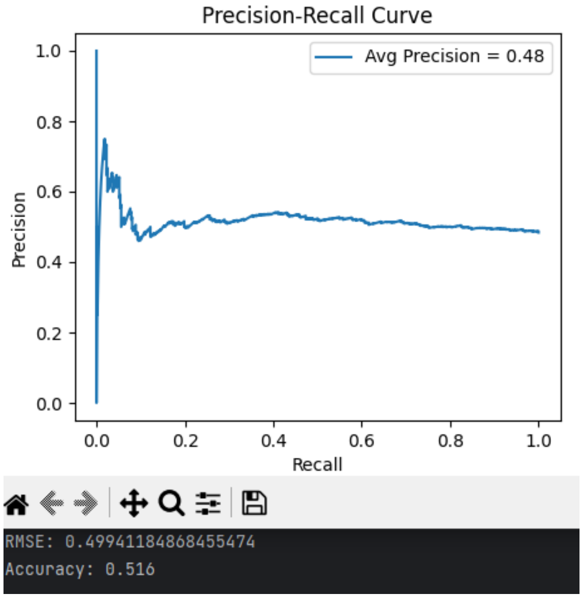
```
"C:\Users\andre\PycharmProjects
RMSE: 0.5003835735352835
Accuracy: 0.4752475247524752
```

Figure 4. LR NFLX Results



```
RMSE: 0.20604353067905432
Accuracy: 0.9404761904761905
```

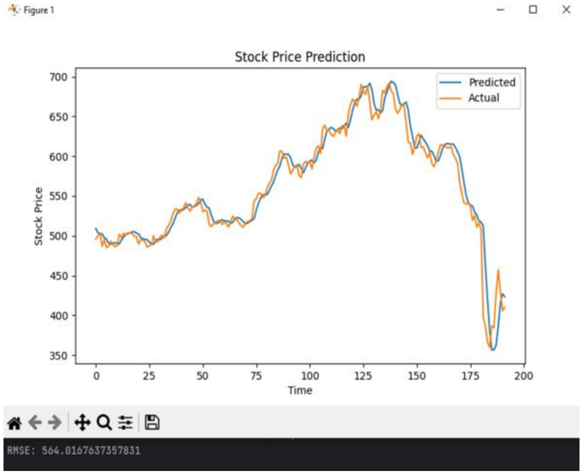
Figure 5. LR GOGL Results



```
RMSE: 0.49941184868455474
Accuracy: 0.516
```

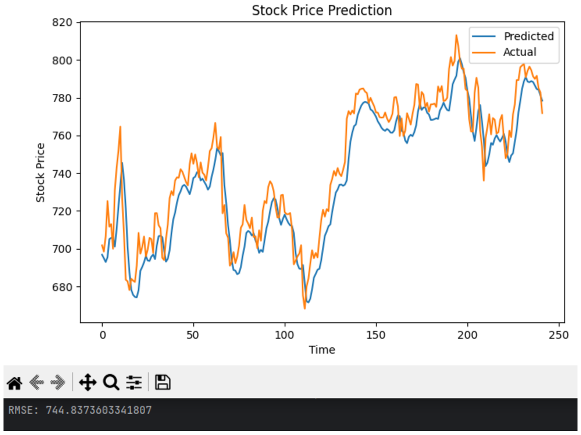
Figure 6. LR MSFT Results

6.2 LSTM Results



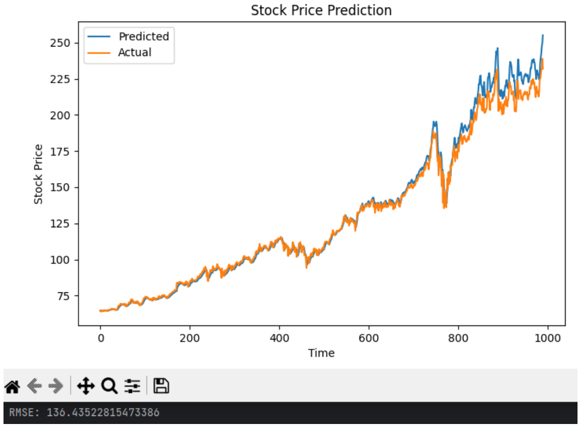
```
RMSE: 564.0167637357831
```

Figure 7. LSTM NFLX Results



```
RMSE: 744.8373603341807
```

Figure 8. LSTM MSFT Results



```
RMSE: 136.43522815473386
```

Figure 9. LSTM GOGL Results

7.CONCLSION

In conclusion my studies purpose was to determine which popular stock market prediction algorithm, LR or LSTM, was the best using various techniques and datasets. The answer to this question could help many people who don't have the financial literacy or time to research how to invest in stocks. Previous research in combination with my own conclude that LSTM has a large advantage over LR when it comes to stock market prediction. Although they are both classified as a machine learning algorithm, LSTM is in a league of its own as its also classified as an RNN deep learning algorithm. As stated previously, deep learning gives LSTM a large advantage over LR because of its ability to preserve and forget information. While LSTM is a complex deep learning algorithm LR is the opposite of this, its so simplistic that its data is measured in ones and zeros. Through this simplistic model it is able to predict whether the stock market will go up, 0, or down ,1. The complexity of the two models couldn't be more different but what algorithm is most suited for this study is what I want to decern.

LSTMs complexity advantage is evident in section 6 where we can see clearly how accurate the algorithms really are. Starting with section 6.1 LRs accuracy is measured using the precision recall curve, RMSE, and an accuracy score. Although the values are not too favorable, using the Netflix dataset the results show a 48 precent accuracy with an RMSE of .5. This is again shown with the Microsoft dataset having a 51 percent accuracy with a RMSE of .49. Despite these negative results, LR can be very accurate as seen in figure 5 with the Google dataset showing an accuracy of 95 percent and an RMSE of .2. LSTM on the other hand was not measured in precision recall curve, instead both the actual and predicted values were graphed to determine how accurate the model predicted compared to the actual values. As seen in section 6.2, LSTM excels with all the datasets used, some more than others but overall, it shows a consistent accurate prediction of each dataset. While both machine learning algorithms have the ability to predict stock data accurately, LSTM clearly has an advantage over LR when its comes to being consistently accurate. Both models have their respective strengths but for something as complex as predicting stock market direction, a more complex algorithm such as LSTM is more suited than a simple one such as LR.

8.REFERENCES

[1] Ali, S. S. ., M. Mubeen, I. . Lal, and A. . Hussain. “Prediction of Stock Performance by Using Logistic Regression Model: Evidence from Pakistan Stock Exchange (PSX)”. *Asian Journal of Empirical Research*, vol. 8, no. 7, July 2018, pp. 247-58, doi:10.18488/journal.1007/2018.8.7/1007.7.247.258

[2] Zaidi, M. Amirat, A. “Forecasting stock market trends by logistic regression and neural networks”. International journal of economics, commerce and management, Vol 4, no 6, June 2016, doi: http://ijecm.co.uk/issn2348-0386

[3] Parshv Chhajer, Manan Shah, Ameya Kshirsagar, ‘The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction”, Decision Analytics Journal, Vol 2, 2022, doi: https://doi.org/10.1016/j.dajour.2021.100015.

[4] Hum Nath Bhandari, Binod Rimal, Nawa Raj Pokhrel, Ramchandra Rimal, Keshab R. Dahal, Rajendra K.C. Khatri, “Predicting stock market index using LSTM”, Machine Learning with Applications, Volume 9, 2022, 100320, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2022.100320.

[5] Adil Moghar, Mhamed Hamiche, “Stock Market Prediction Using LSTM Recurrent Neural Network”, Procedia Computer Science, Volume 170, 2020,Pages 1168-1173, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.049.

[6] Liberto, Daniel. “How the Stock Market Affects the U.S. Economy.” *Investopedia*, 28 Feb. 2023, www.investopedia.com/how-stock-market-affects-economy-5296138#:~:text=Why%20Is%20the%20Stock%20Market,profit%20from%20their%20growth%20prospects.

[7] Furhmann, Ryan. “How the Internet Has Changed Investing.” *Investopedia*, 31 July 2022, www.investopedia.com/financial-edge/0212/how-the-internet-has-changed-investing.aspx#:~:text=When%20the%20internet%20arrived%2C%20it,information%20and%20transparency%20for%20investors.

[8] “Research Guides: Wall Street and the Stock Exchanges: Historical Resources: Stock Exchanges.” *Stock Exchanges - Wall Street and the Stock Exchanges: Historical Resources - Research Guides at Library of Congress*, guides.loc.gov/wall-street-history/exchanges#:~:text=American%20Stock%20Exchange&text=Founded%20by%20the%20National%20Association,trading%20for%20over%202%2C500%20securities. Accessed 15 Sept. 2023.

[9] Williams, Ward. “Timeline of U.S. Stock Market Crashes.” *Investopedia*, Investopedia, 15 Apr. 2023, www.investopedia.com/timeline-of-stock-market-crashes-5217820#:~:text=The%20term%20stock%20market%20crash,crisis%20or%20major%20catastrophic%20event.

[10] Godot, Joe. “Using Logistic Regression to Predict Stock Movements with R.” *Medium*, Artificial Intelligence in Plain English, 20 Mar. 2023, medium.com/ai-in-plain-english/using-logistic-regression-to-predict-stock-movements-with-r-6375b35479bb.

[11] Mrinalini Smita *"Logistic Regression Model For Predicting Performance of S&P BSE30 Company Using IBM SPSS"* International Journal of Mathematics Trends and Technology 67.7 (2021):118-134.

[12] Toprak, Mehmet. “Sigmoid Function”. Medium. https://miro.medium.com/v2/resize:fit:1400/format:webp/1*a04iKNbchayCAJ7-0QlesA.png

[13] Roondiwala, Murtaza & Patel, Harshal & Varma, Shraddha. “Predicting Stock Prices Using LSTM”. International Journal of Science and Research (IJSR). 2017. 10.21275/ART20172755.