Stock Prediction Using Machine Learning

SHUBHA SINGH, SREEDEVI GUTTA, AHMAD HADAEGH
Department of Computer Science and Information System
California State University San Marcos
333 Twin Oak valley Rd. San Marcos CA, 92009, USA

Abstract: The Trend of stock price prediction is becoming more popular than ever. Share market is difficult to predict due to its volatile nature. There are no rules to follow to predict what will happen with the stock in the future. To predict accurately is a huge challenge since the market trend always keep changing depending on many factors. The objective is to apply machine learning techniques to predict stocks and maximize the profit. In this work, we have shown that with the help of artificial intelligence and machine learning, the process of prediction can be improved. While doing the literature review, we realized that the most effective machine learning tool for this research include: Artificial Neural Network (ANN), Support Vector Machine (SVM), and Genetic Algorithms (GA). All categories have common and unique findings and limitations. We collected data for about 10 years and used Long Short-Term Memory (LSTM) Neural Network-based machine learning models to analyze and predict the stock price. The Recurrent Neural Network (RNN) is useful to preserve the time-series features for improving profits. The financial data High and Close are used as input for the model.

Key-Words: - Date Mining, Machine learning, Stock Prediction, Date Cleaning, Data Prediction, Recurrent Neural Network, Data Normalization, Support Vector Machine (SVM), Genetic Algorithms
Received: June 27, 2021. Revised: October 25, 2021. Accepted: November 28, 2021. Published: December 16, 2021.

1. Introduction

Stock market is dynamic, unpredictable due to nature of the volatile market. Predicting any stock value accurately is a huge challenge as there are so many factors to consider such as news, sentiments, economy, financial reports and much more. The strategies for investment in stock market is very complex and depends on tons of data. Profit always comes with the risk of losses. To minimize the risk of losing money and maximize the profit the techniques to predict stock value is highly useful. There are two main approaches that are being used for predicting stock values. In the first method which is Traditional Time Series method the prediction is based on the historical data of that particular stock. In this method the stock's closing price, opening price volume etc. has been used. The second method, that is qualitative, the prediction is based on factors like company profile, news articles, economy, social media, market sentiments etc.

For stock market the size of the data is quite huge and random so we need models that are efficient and can deal with the complexity of this huge amount of data. The stocks data are complicated and difficult to understand due to the hidden patterns. Machine learning techniques have potential to deal with the complexity and dig to solve the multilayered complicated patterns and come up with a good prediction.

The main purpose of this research is to predict future price of a particular stock. In this research I collected stock price data from Yahoo Finance to feed the data to machine learning algorithm model.

2. Related Work

In the work of Kim and Han [3], the researchers have used ANN and GA to predict future stock price. The data authors have used is from Korea stock price index (KOSPI). The sample data they have collected from KOSPI was about 10 years ranging from January 1989 to December 1998. They applied some required optimization and techniques to prepare the data to be used. They used GA to optimize ANN. There were 12 hidden layers that were not adjustable. Also, the author only focused on two factors in optimization while he agrees that GA has great potential for optimization. Similar work is also presented by Qiu and Song [6] where they are predicting the movement of Japanese stock by using ANN and GA. They named this algorithm GA-ANN model since it was a mixture of both.

Hassan and Nath applied Hidden Markov [7] Model (HMM) for predicting the stock prices for

four different Airlines. One of best thing in their research paper was that the approach they took do not need any expert to build the model. The problem with that research is that they used very less data for evaluation and the data was related to a specific industry so this may not fetch a good result in predicting. The authors have used data of around two years only which is very less for machine to understand and predict the trend.

Lee in [6] used SVM to predict the stock market trend. The author collected data from NASDAQ from Taiwan Economic Journal. Lee used the method supported sequential forward search (SSFS) for the feature selection. The author also created some procedure that will adjust parameter that have different values. The structure that they used was very clear for the feature selection of the model.

Thomas Fischer and Christopher Krauss in [8] used long short-term memory (LSTM) to predict the stock market trend. They collected dataset from S&P 500 index for Thomson Reuters from December 1989 to September 2015. After collecting the data, they converted the list into binary matrix. For optimization they used RMSprop. They used the latest technique to perform the task, but they did not have any background knowledge in financial domain. In this paper the author did not mention how they train the model with long-time dependencies.

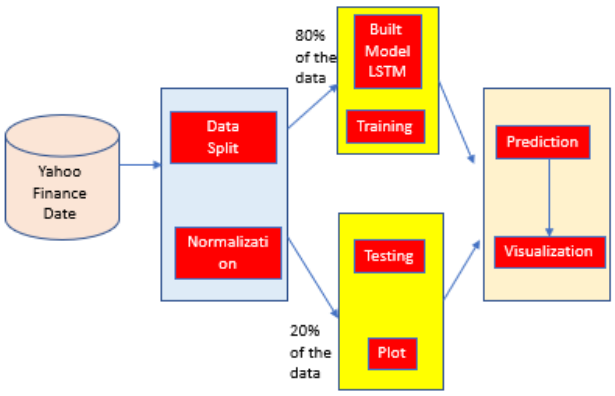


Figure 1: System Architecture

M. Roondiwala, H. Patel and S. Varma. in [10] has used RNN-LSTM model on NIFTY-50 stocks. They collected data for 5 years and RMSE to find out the error rate. The window size they are using to predict the price movement is of 21 days.

3. Methodology

3.1 Data Collection and Preparation

As shown in Figure 1, the data for three companies from different sectors for around 10 years from date 1/1/2012 to 10/10/2021 has been collected from Yahoo Finance. The dataset includes the data for Tech company, Banking sector, and a Food service to understand movement of stock prices in different sectors. The data contains information about the stock such as High, Low, Open, Close, Adjacent close, and Volume. From table 1, I have selected the Close column to train and test my model.

Date	High	Low	Open	Close	Volume	Adj Close
2012-01-03	14.732143	14.607143	14.621429	14.686786	302220800.0	12.610314
2012-01-04	14.810000	14.617143	14.642857	14.765714	260022000.0	12.678082
2012-01-05	14.948214	14.738214	14.819643	14.929643	271269600.0	12.818837
2012-01-06	15.098214	14.972143	14.991786	15.085714	318292800.0	12.952843
2012-01-09	15.276786	15.048214	15.196429	15.061786	394024400.0	12.932296
...	...	...	...	...	...	...
2021-08-09	146.699997	145.520004	146.199997	146.089996	48908700.0	146.089996
2021-08-10	147.710007	145.300003	146.440002	145.600006	69023100.0	145.600006
2021-08-11	146.720001	145.529999	146.050003	145.860001	48493500.0	145.860001
2021-08-12	149.050003	145.839996	146.190002	148.889999	72282600.0	148.889999
2021-08-13	149.440002	148.270004	148.970001	149.100006	59318800.0	149.100006

Table 1: Collected Data

The data has been divided into training and testing segment. Training set is the subsection of the original data that is divided to train the model and test data is the subsection of the original data that is divided to test the model by comparing with the actual value to evaluate the accuracy of the model. I have divided them into 80:20 ratio where 80% of the data has been used for training purpose and 20% of data is used for testing. The data has been scaled to normalize the data between a range so that it is easier for the algorithm to learn patterns from the data. The given range for the data is from 0 to 1.

After scaling the data, I created the training dataset where timestamp is 60. It is the batch size of the data that is processed by the model for each iteration for input and output. So, in the first iteration, it will use first 60 data as input and 61st as output. In the next iteration it will use from 2nd to 61st as input and 62nd as output and so on. In the next step, I build the LSTM model.

Epoch is the number of the passes of the entire training dataset the model completes. Here I am using epoch number 100 so the model will go through each dataset 100 times.

3.2 RNN

LSTM is a special type of Recurrent Neural Network (RNN). This uses data from previous records and based on that it predicts the future. RNN can remember our previous input in their memory when a huge dataset is given.

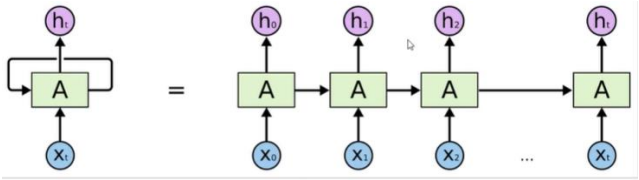


Figure 2: An unrolled RNN
(Aditi Mittal Oct 12, 2019
Understanding RNN and LSTM)

Figure 2 shows how unrolled RNN works and Figure 3 shows how Rolled RNN works. Here if we unloop the data, we can see that the first output is based on the first input and second output is based on the first and second then the third output is based on first, second and third and so on. RNN can be thought as multiple copies of the same neural network where each passing a message to the successor.

The problem with RNN is that it is long term dependent, so if we take a large set of RNN data then there could be two types of problem.

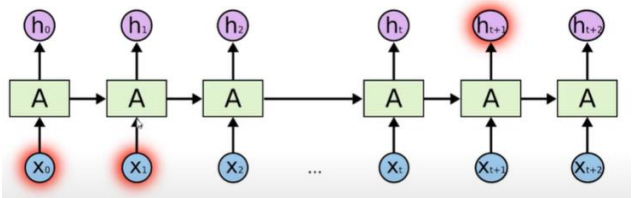


Figure 3: Rolled RNN

3.2.1 Vanishing Gradient

As we know Neural networks works on Backpropagation which means when we get the output, we compared it with actual output and if there is any error either large or small error it will update the weight of the neural network.

If the error is less than 1, it will be multiplied by learning rate and while reaching to the last cell it will be very less that is almost zero. That is why it is called vanishing gradient.

3.2.2 Exploding Gradient

When the algorithm assigns a high importance to the weights, it will explode while reaching to the end cell

4. Proposed System

As represented getting the historical data from market is mandatory step. Then we will have to extract the feature, which is required for data analysis, then divide it as testing and training data, training the algorithm to predict the price and the final step it to visualize the data. Figure 4 represents the Architecture of the proposed system.

- Step 1: Data Collection.
- Step 2: Data Preprocessing after getting the historic data from the market for a particular share and read the close price
- Step 3: Divide the data
- Step 4: Do a feature scaling on the data so that the data values will range between 0 and 1.
- Step 5: Building the RNN (Recurrent neural network) for data set and Initialize the RNN by using sequential regressor.
- Step 6: Building the LSTM model
- Step 7: Train the model by providing batch size and epochs.
- Step 8: Making the predictions and visualizing the results using plotting techniques.

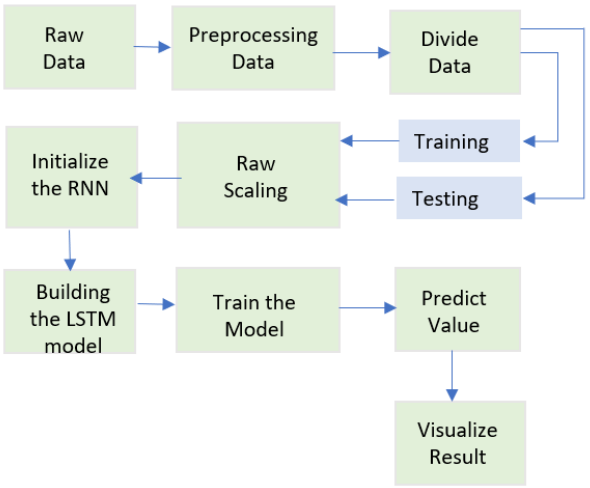


Figure 4: Proposed system

4.1 Long Short-Term Memory (LSTM)

To overcome from this problem, we come up with another version of RNN that is Long Short-Term Memory (LSTM). This is a special kind of RNN that is capable of learning long-term dependencies. This is specially designed to avoid long-term dependency problems and memorize information for long period of time. LSTM can process entire sequence of datasets and uses a supervised learning process.

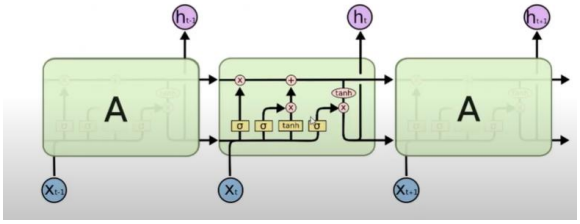


Figure 5: LSTM

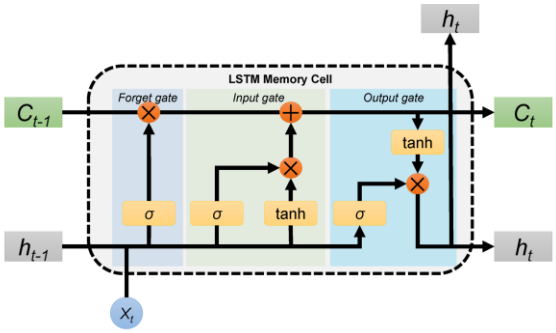


Figure 6: LSTM memory cell

In Figure 5 we can see a single cell is updated with controlling gates which controls the previous memory as well as the input. This is how they handle vanishing and exploding gradients. Figure 6 shows clear picture of how an input gate an output gate and a forget gate are working. The cell remembers values and the gates regulate the flow of information.

4.2 Building the LSTM model:

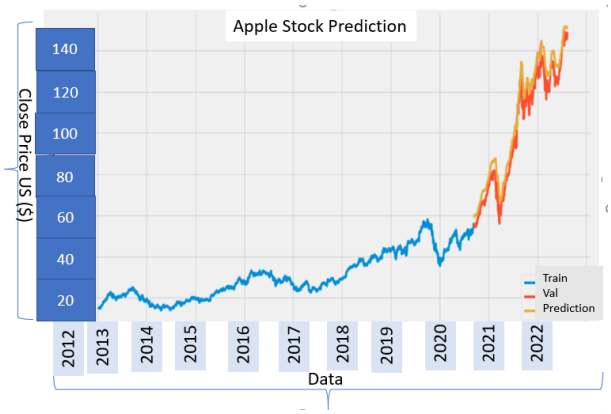
To build the LSTM model we will need to import Keras library and packages. Keras is basically tenser flow high-level API for building and training models. Here we are going to import some libraries.

Sequential – It is basically linear stack of layers through which we can create a sequential layer by passing the list by it.

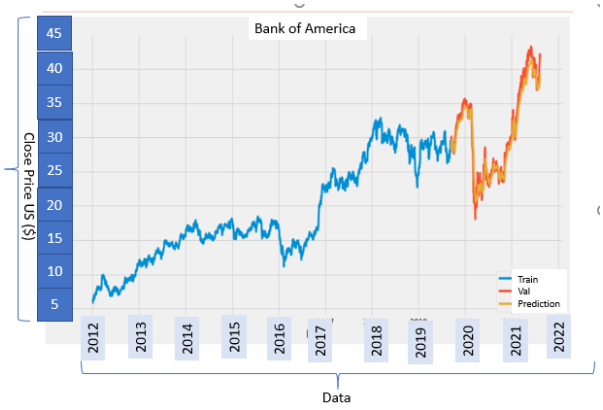
Dense - It is the regular deeply connected neural network layer. It is most commonly used layer to change the dimension of the output. It represents a metrics vector multiplication so the value which are trainable parameter get updated during back propagation.

Next, we are going to initialize RNN. For time series problems like this we are going to use regression model. The first step in this is to read the data which is a sequential data, and we are going to assign this to regressor.

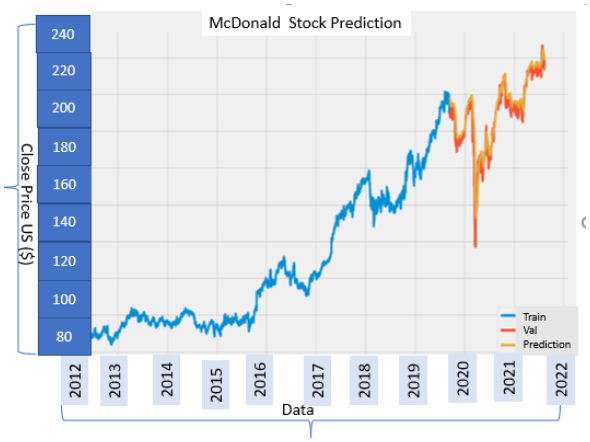
After this we are going to move to the next step that is the most important stage training the model. In this stage the data we collected is going to fed to the model and trained for prediction.



Graph 1: Apple Stock Prediction



Graph 2: Bank of America Stock Prediction



Graph 3: McDonald's Stock Prediction

This LSTM model is composed of a sequential input layer followed by three LSTM layers and a dense layer. In the model we are also using dropout. The dropout is a regularization technique for reducing overfitting, so it drops out units in neural networks.

After this we are going to compile the model. Here we are using optimizer. Optimizers are methods used to change the attributes of the neural network such as weights and learning rate in order to reduce the losses. The type of optimizer used can greatly affect how fast the algorithm converges to the minimum value. Here we are using Adam optimizer. The Adam optimizer combines the perk of two other optimizers ADAgrad and RMSprop.

Next step is to train my model where I have given the epoch as 100. Epoch is a term used to indicate how many times the algorithm has completed. Then I have created testing data from scaled values, converted the data to a NumPy array and reshaped the data.

The root mean squared error (RMSE) is used to evaluate the model. It is a good measure of how accurate the model predicts or response. The lower value of RMSE indicates the good prediction.

5. Results and Evaluation

After training the result of the prediction we need to show the result. The result could be shown in many ways but the most effective way to show the result in a concise view is through a Graph. In Graph we can show the comparison between the result of both actual values and prediction value using the data that were collected and processed.

In these graphs (Graphs 1, 2, and 3), the blue line (Train) is showing the data used in training, the red line (Val) is the actual value of the data and orange line (Prediction) is the graph for predicted data. We can see actual and predicted values are going in similar direction. In all the graphs we see that both lines are very close to each other.

Any model cannot predict stock value accurately for many reasons such as if there is any positive or negative news come from that company the stock prices will have an impact. Furthermore, if some sudden changes in market happens, we cannot predict that beforehand. There could be multiple reasons the stock price can fluctuate aberrant. In this project, I am trying to feed the pattern of the stock price for a particular company and based on that I am creating a model that will be trained with the historical data and based on the pattern it will apply an algorithm and will predict the future value. Thus, we cannot be completely sure in predicting the stock price. The analysis shows that to predict the value of such kind of activity is not a straight task because there are many factors and problems that should be considered before making a prediction.

Date		
2020-01-09	77.407501	76.052917
2020-01-10	77.582497	76.567558
2020-01-13	79.239998	77.097389
2020-01-14	78.169998	77.659370
2020-01-15	77.834999	78.246483
...	...	...
2021-10-25	148.639999	140.269745
2021-10-26	149.320007	140.796402
2021-10-27	148.850006	141.324753
2021-10-28	152.570007	141.824951
2021-10-29	149.800003	142.308578

Table 2a: Apple

To understand this better, I have also generated values in table format. This is easier to read and understand for any specific values. In the table, we can easily figure out what the actual value of the specified date is and what the model is predicting for the same date. (see tables 2a, 2b, and 2c)

Date		
2019-09-13	30.170000	28.353430
2019-09-16	30.129999	28.691261
2019-09-17	29.940001	28.995203
2019-09-18	30.000000	29.226727
2019-09-19	29.820000	29.400042
...	...	...
2021-08-09	40.669998	37.646336
2021-08-10	41.430000	37.988972
2021-08-11	41.950001	38.423378
2021-08-12	42.150002	38.908005
2021-08-13	41.630001	39.380836

Table 2b: Bank of America

Date		
2019-09-13	209.809998	218.407440
2019-09-16	207.399994	217.266922
2019-09-17	209.850006	215.960999
2019-09-18	210.429993	214.958710
2019-09-19	210.520004	214.264206
...	...	...
2021-08-09	234.679993	241.229477
2021-08-10	233.449997	240.345184
2021-08-11	235.550003	239.406204
2021-08-12	236.669998	238.754028
2021-08-13	238.820007	238.419830

Table 2c: McDonald's

	1 year	5 years	10 years
AAPL (RMSE)	5.6	3.2	0.43
BOFA (RMSE)	3.48	1.8	0.76
McDonald's RMSE	4.9	2.8	0.9

Table 3: RMSE for different years data

ARIMA (RMSE)	Linear Regression (RMSE)	SVM (RMSE)	LSTM (RMSE)
72.64	3.22	1.58	0.43

Table 4: RMSE Comparison with other works

6. Comparison:

In table 4 we can see the difference of RMSE between these models. This ARIMA RMSE result I have taken from DEVELOPERS CORNER [11]. The ARIMA model got a RMSE of 72.64 that is not good prediction whereas LSTM model got RMSE of 0.43 which is very good as compared to the other model. There could be multiple reasons for this some of which can be the ARIMA model is not working good because the author is using only about five years of data to train the model. The Linear Regression and SVM RMSE I have taken from Vaishnavi Gururaj, Shriya V R and Dr. Ashwini K [12] work. We can clearly see here that LSTM is giving the best results among these models.

For good prediction huge data is necessary in [10] M. Roondiwala, H. Patel and S. Varma have collected data only for 5 years if they would have collected more data, the prediction rate will be more accurate. I have collected data for 10 years to train my model so the accuracy label of my model will increase. We can see in the Table 3 how size of data impact models accuracy.

Hassan and Nath in [7] applied Hidden Markov Model (HMM) to predict the stock price but the problem is that they have chosen and collected data only for one sector that is Airline and they collected data for four different company, but all are from same sector. To see if our model is working well it is good to use data from diverse sectors. In my work I have used data from different sectors such as Tech company, Bank sector and Food services. The data Hassan and Nath collected is only for 2 years that is not sufficient for a model to train with good accuracy.

In the work of Kim and Han [3], they have used ANN to predict future stock price. The problem with the ANN is that in this each node that is hidden is simply a node with a single activation function. Whereas in my work I have used LSTM, and in this each node is a memory cell that can store contextual information. LSTM can remember specific function for longer time, and this helps the model to work better.

In the work of Lee in [6] used SVM to predict the stock market trend. The problem with SVM is that it is not suitable for huge data set and for predicting stock price we need tons of data. Sometime SVM will underperform if the number of features for each data point exceeds the number of training data samples, whereas LSTM performs better in these situations.

7. Conclusion

In conclusion we can say that the popularity of stock market trading is growing rapidly which is encouraging researchers to find out new methods for the prediction using new techniques.

The forecasting techniques is not only helpful to researchers, but it also helps investors or any person dealing with the stock market. To help predict stock indices or forecasting model with good accuracy is required.

Overall, we conclude that Stock prediction using Machine learning can increase the prediction accuracy. Through historical data we can find the pattern of the stock and how the stocks movement has been over years and learn through them. A model can learn the trends with huge amount of data provided and train themselves to predict the movement of the stock price. However, it is also important to understand the parameters that

influence the prediction in order to develop an efficient product with accuracy of results. We can conclude that historical data can be used as powerful tool that allows us to forecast the movement more accurately. However, our method might not generate the accurate result in this case for the number of reasons mentioned such as any negative or positive news or sudden change in market etc. Machine learning is the best way to recognize patterns and predict results rather than apply the analysis manually on large sets of data. There are some drawbacks that need to consider while applying the algorithm to predict the result. Nobody should blindly rely on these models and invest since there are number of reasons that the prediction can go in other direction.

8. Future Work:

The model used in my project has shown very promising result. From the various results shown through graphs and tables in the research can confirm that my model is capable of tracing the evolution of closing prices. For our future work we will try to find the best sets for about data length and number of training epochs that better suit our assets and maximize our prediction accuracy.

Reference:

[1] Adil Moghar and Mhamed Hamicheb. Stock Market Prediction Using LSTM Recurrent Neural Network, Procedia Computer Science. Volume 170, 2020, Pages 1168-1173

[2] Pramod and Mallikarjuna Shastry.Stock Price Prediction Using LSTM. January 2021. Test Engineering and Management 83 (May-June 2020):5246-5251.

[3] Kyoung-jaeKim and IngooHan. prediction of stock price index. Volume 19, Issue 2, August 2000, Pages 125-132

[4] Raghav Nandakumar1, Uttamraj K R2, Vishal R3, Y V Lokeswari4. Stock Price Prediction Using Long Short-Term Memory. Volume: 05 Issue: 03 |Mar-2018. International Research Journal of Engineering and Technology (IRJET)

[5] Shipra Saxena. Introduction to Long Short-Term Memory. March 16, 2021

[6] Mingyue Qiu and Yu Song. Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model May 19, 2016

[7] M.R. Hassan and B. Nath. Stock market forecasting using hidden Markov model. 23 January 2006

[8] Thomas Fischer and Christopher Krauss long short-term memory networks for financial market predictions Volume 270, Issue 2, 16 October 2018, Pages 654-669

[9] Ming-ChiLee. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Volume 36, Issue 8, October 2009, Pages 10896-10904

[10] M. Roondiwala, H. Patel and S. Varma, "Predicting stock prices using LSTM," International Journal of Science and Research (IJSR), vol. 6, no. 4, pp. 1754-1756, · April 2017

[11] https://analyticsindiamag.com/comparing-arima-model-and-lstm-rnn-model-in-time-series-forecasting/

[12] Vaishnavi Gururaj, Shriya V R and Dr. Ashwini K, Stock Market Prediction using Linear Regression and Support Vector Machines Volume 14, Number 8 2019

[13] Colah's blog August 27, 2015 Understanding LSTM Networks

[14] Hongxiang Fan, Mingliang Jiang, Ligang Xu, Hua Zhu, Junxiang Cheng,and Jiahu Jiang, Comparison of Long Short Term Memory Networks and the Hydrological Modelling Runoff Simulation

Author Contributions:

Shubha Singh: Graduate Student doing her thesis. Her role was doing the research and implementation of this paper. This paper was part of her master thesis.

Sreedevi Gutta: Advisor of Shubha. Dr. Gutta was the advisor of Shubha and provided valuable comments every week mainly on sections 3 and 4.

Ahmad Hadaegh: “co-advisor”. Dr. Hadaegh also directed Shubha in her work providing valuable feedback mainly on sections 4 and 5.