Stock Market Open Price Prediction Using ML

MUHAMMAD ADIL INAM(19030001), LUMS Pakistan

Stock market value prediction using NN is one of the hot topic in research community. Many publications is seen in recent years in the mentioned domain. Generally, talking about prediction/forecasting many variables comes to play. Super computers are often used in weather forecasting due to large amount of data and many variables governing the change. Similarly, stock market value is dependent on many values. Geo political factors, Company PR, Natural disasters, Climate, global economy etc. are some of the factors that govern the stock market value of a company

KEYWORDS

Stock Market, Machine learning, ML, NN

1 INTRODUCTION

When it comes to stock market, a lot of variable comes to play important role in stock market value prediction. Often it takes years to practice to become expert in stock market trading. Stock market value prediction is one most researched area in machine learning. A lot of research has been around the concept to design a system, which can successfully predict the stock market price. In order for the system to be efficient, the system should be able to correctly estimate increase or decrease in stock market value. In this project, we will explore ways to design a system, which can make an estimation of the stock market price.

Accurate stock market prediction is of great interest to investors; however, stock markets are driven by volatile factors such as microblogs and news that make it hard to predict stock market index based on merely the historical data. [1] Investments in stock market are purely based on prediction by the financial analysts. Many investors have teams of financial analysts performing the evaluation for the stocks for the investors. Becoming an expert in predicting the stock market takes years of experience and practice.

From the historical data available, it is clear that many factors are involved in stock market value. It has been seen that geo politics, company culture, natural disasters, climate and many other factors play an important role. During the covid pandemic stock market took a fall, similarly storms, earth-quakes, pandemics, fake news, elections etc. play a vital role in governing the stock market.[2]

2 MOTIVATION

Prediction of stock market is a challenging task and if some day we can create a system that can predict with accuracy stock market value it would be of a great deal and could help many investors and individuals in investing their money correctly. There has been a great deal of effort being put by the research community in designing new and improved machine learning models that can do the job accurately. SVM is a model being widely used for training of models. [3,4]. Many researchers have a theory that quarterly finance reports released by the company should reflect in the increase/decrease in stock market value.

3 LITERATURE REVIEW

SVM and LSTM based models are highly used by researchers in stock market prediction models. Jason and Jianguo by using SVM achieved 71.3% accuracy [3], Raut and Shinde evaluated SVM, Naïve Bayes algorithm and achieved 80% accuracy in SVM model [5]. Sai Krishna Lakshminarayanan, John McCrae compared LSTM

and SVM based models and reported that LSTM based models have better results as compared to SVM based models. [6]. Sachin, Sneh and Jay Shanker gave published a model using SVM that used not only historical finance data available but also they manually did sentiment analysis on company tweets to model stock market prediction. With collection of news, tweets and historical financial data, they were able to achieve 80% accuracy in their SVM model [7]

5 DATASET

Most of company's historical stock market data is available on the web. Many researchers have opted for crawling the data themselves. Yahoo fiancé is widely used for getting stock market data for training. In this project, we will be utilizing Microsoft historical stock market data available on the yahoo finance. [8]

Since many variables govern stock market value but unfortunately the data set available for training does not contains all those variables like political attributes, climate, natural disaster, pandemics etc. These variables are condition or decision variables for our dataset e.g. if there is no natural disaster and political state of country is normal then business will grow and stock market value will increase. During recent years when US president Donald trump cut off business with china stock market price of companies like huawai stock market price decreased. Usually stock market prices are steady but unforeseen and unfortunate incidents like geo political issues(US-China), pandemic (COVID-19) etc create an uncertainty in the stock market price and prices usually dip down unexpectedly.

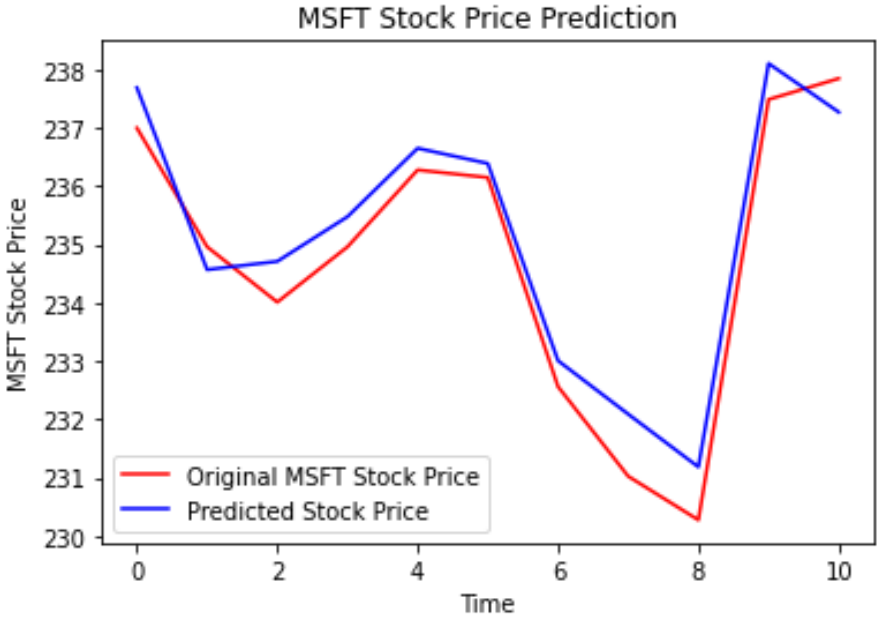
6 METHODOLOGIES.

For the project we will be implementing LSTM based RNN. From the literature reviews it's clear that LSTM based models have better results as compared to SVM based NN. The model will be trained on open price of stock market for MSFT finance data for 1 year. In our RNN we implemented 4 layers of neurons with 50 neurons in each layer. The data was cleansed and scaled to 0-1 for Neural network training. We used 7 to 1 approach for neural network training where we gave 7 days of data for training and next day as output for neural network.

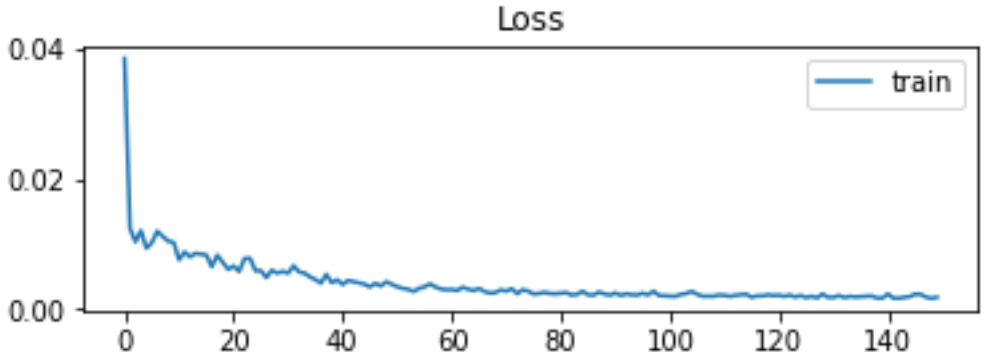
We tested our models using various optimizers, epochs and batch sizes. We concluded on using Adam as optimizer, mean squared error as loss function and MAPE as accuracy matrix. As our output we tasked the trained model to predict the price of stock market for the next few days and compared it with actual stock market value.

7 RESULTS of NN Models.

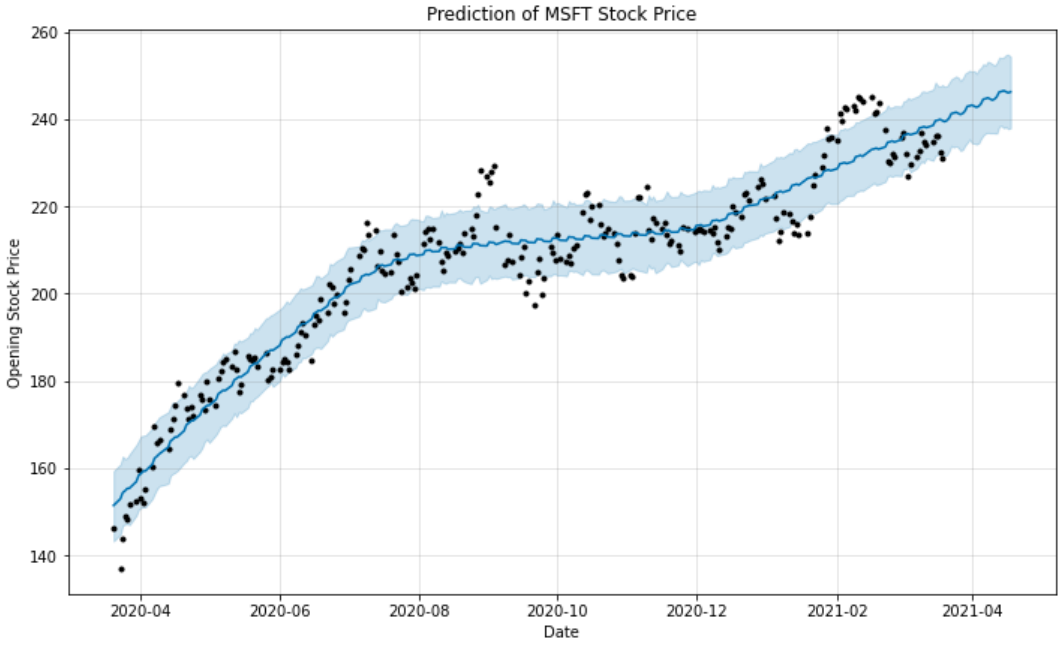
We used 1 years (2020-2021) of stock market data to train our RNN LSTM based model. We then tested our model with predicting the stock open value for the next few days and compared the results with the actual stock market value. We were able to achieve 95 % average accuracy with our system. Our system was able to make certain correct predictions.



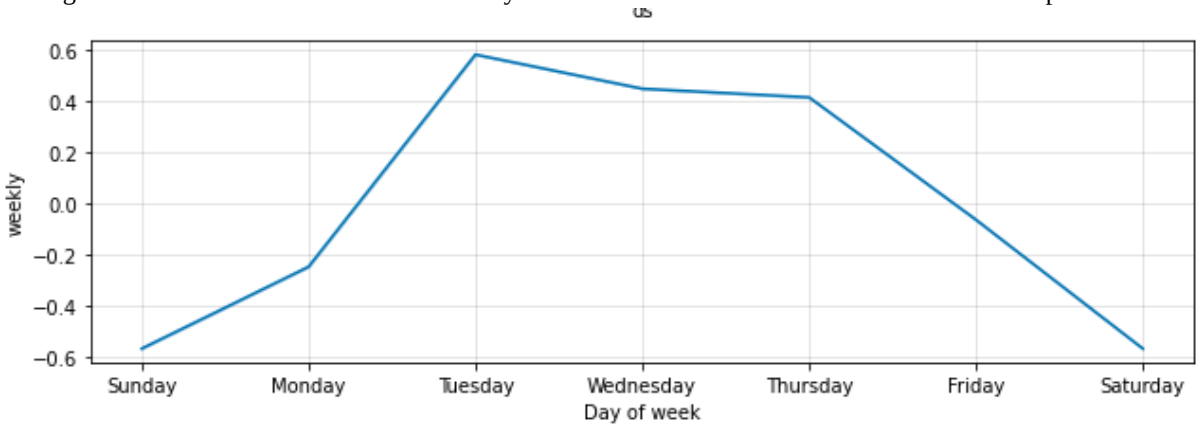
Mean square plot for Neural Network training is shown bellow



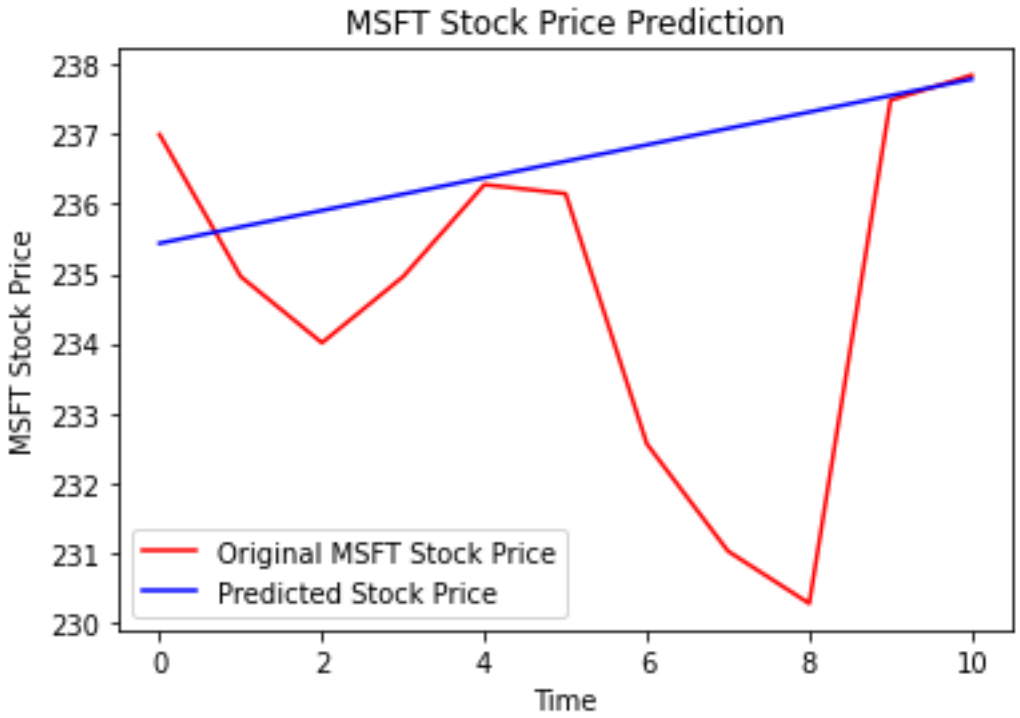
With few optimizations in place, Recall prediction period set to 2 week, with 150 epochs and 2-batch size the model was able to achieve 95% accuracy. The system showed very promising results. In comparison to RNN we used facebook fbprophet model implementation for stock market prediction the model was trained and evaluated on same dataset. The fbprophet gave us the trends of the stock for week, month and year. The model has a linear curve as compared to RNN model we created. The results of the model are as shown below.



The figure shows the linear curve with boundary error values where dark dots are the actual stock price.



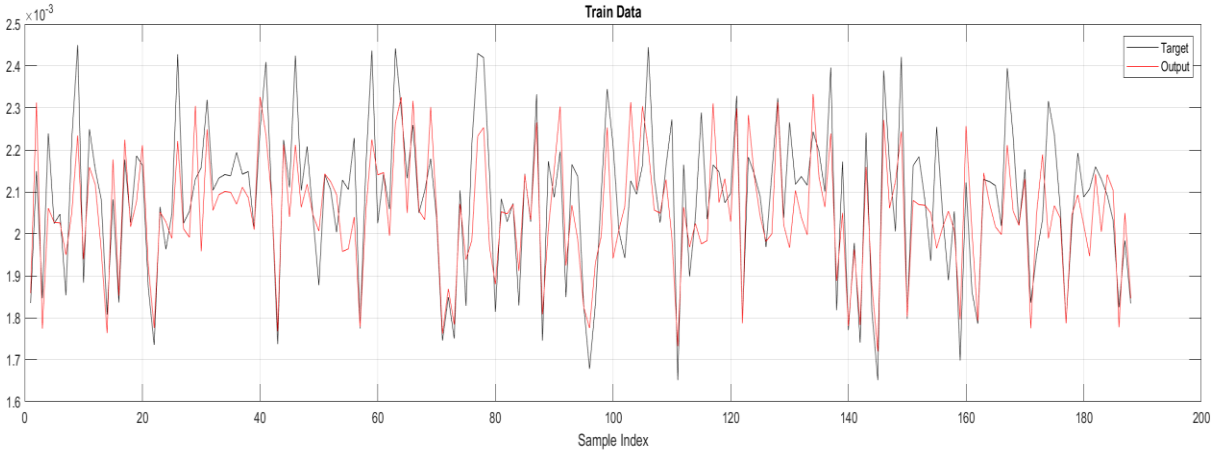
The above figure shows the trend of the stock value during the period of week. Usually the stock price of Microsoft is highest at Tuesday.



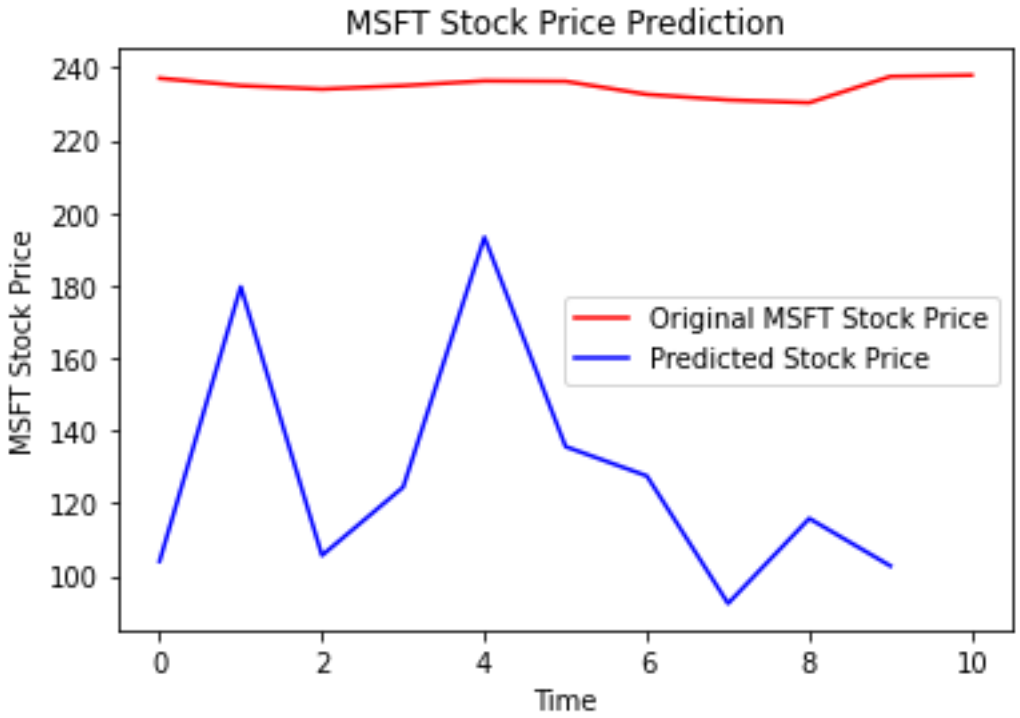
There is a lot of room for fine tuning fbprophet for better predictions.

8 ANFIS Implementation.

Adaptive neuro-fuzzy inference system is another way of implementing a model that we will use predict stock market price. Sugeno is the most commonly used FIS in research community [9]. The model implemented use data mining techniques to divide the domain space into different classes. We tested with several clusters of classes for training of anfis mode we found out, that best results were achieved using 5 clusters. Moreover, we set epochs to 200, used Mean square error for error calculation and backpropagation optimization method. The model is implemented using Matlab and was tested and trained on the same Microsoft dataset used previously in previous NN implementation.



Above figure shows train data for anfis model along with results of anfis



Above figure shows the test data of anfis model. The model accuracy is very high and very low MSE of 0.0001. Feature selection was done manually to figure out best outcome from model. We used stock price open, close, low and high price for training the model. Using less number of features resulted in lower accuracy. We used same 14 days lookup for predicting the next day.

9 Genetic Algorithm.

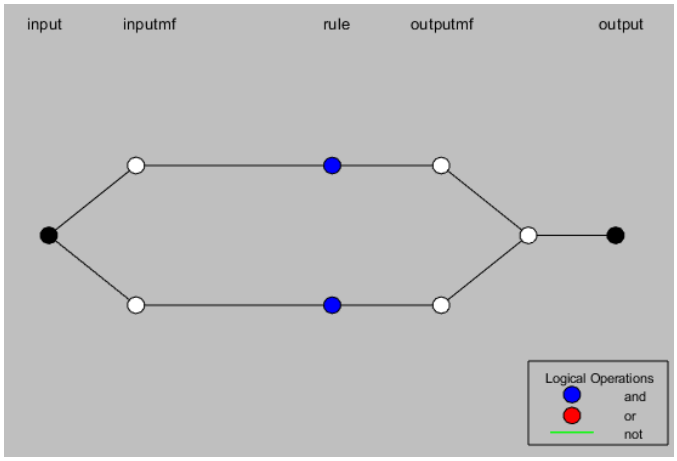
Genetic Algorithm (GA) is a search-based optimization technique based on the principles of Genetics and Natural Selection. Genetic algorithm is widely used for various optimization and feature selection for neural network training models. We opted to use genetic algorithm for feature selection. Our dataset has following features available in stock data Date, Open, Close, Low, Adj Close and Volume. Since We doing time series prediction on Open Price of Stock, We opted to drop Open and Date columns from the dataset and used label encoding with 10 buckets for other remaining columns. Our input data for genetic now has Close, Low, Adj Close and volume columns and output has one Open as column. We tested our model with various classifiers and opted to use Random Forest classifier since it gave better results and low error. We used n_estimator = 50 and Population size set to 100. The model gave us “High” to be used for Model training as best feature. To double check, our results we ran test several time. The model always gave high as best feature for model training. We also added noise in our dataset by introducing a random column with random values between 1-200. Our model never predicted that column to be used as feature for network training.

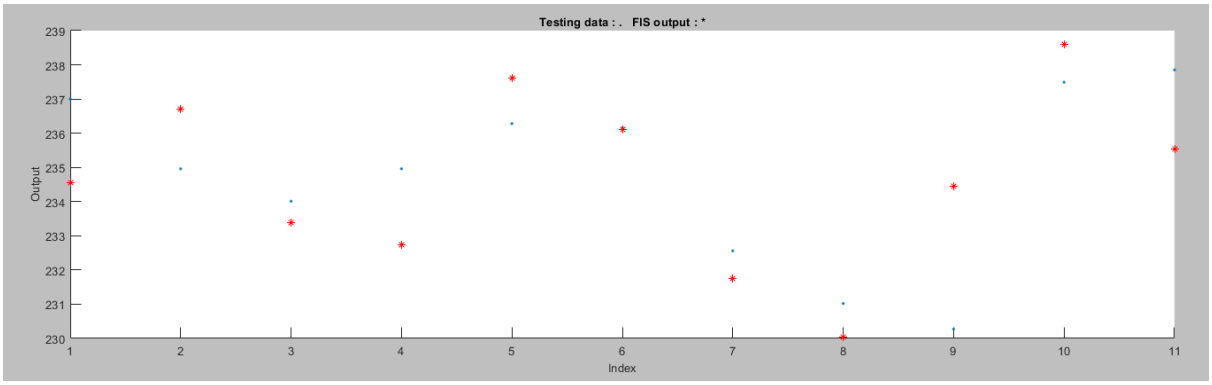
BEST FEATURE	**ACCURACY**
1ST BEST “HIGH”	0.892
2ND BEST “LOW”	0.865

2nd best “Low” was predicted as best feature when high was removed from dataset. Using this data from genetic algorithm, we opted to use this for training and optimization of our ANFIS model.

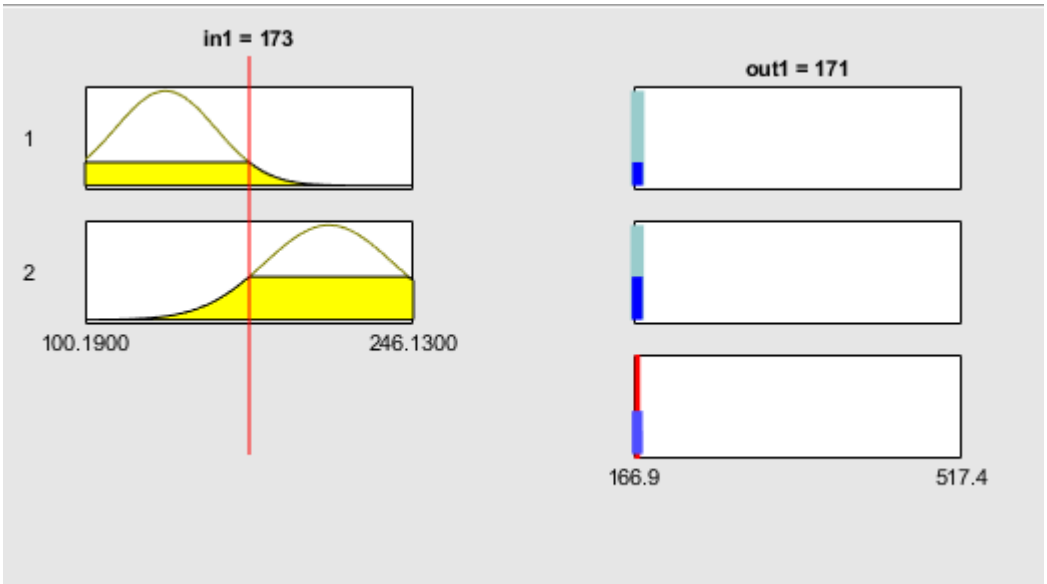
11 ANFIS Optimization.

After feature selection using genetic algorithm, we opted to optimize and train the ANFIS model. We create 33-3-models using various settings in MATLAB. One of our model was trained using 2 years of dataset; the model used sub clustering to generate sugeno model FIS and was training for 500 Epochs on hybrid optimization method. The model has Minimal training RMSE = 1.79379. The model structure and plot against a test dataset is shown bellow.





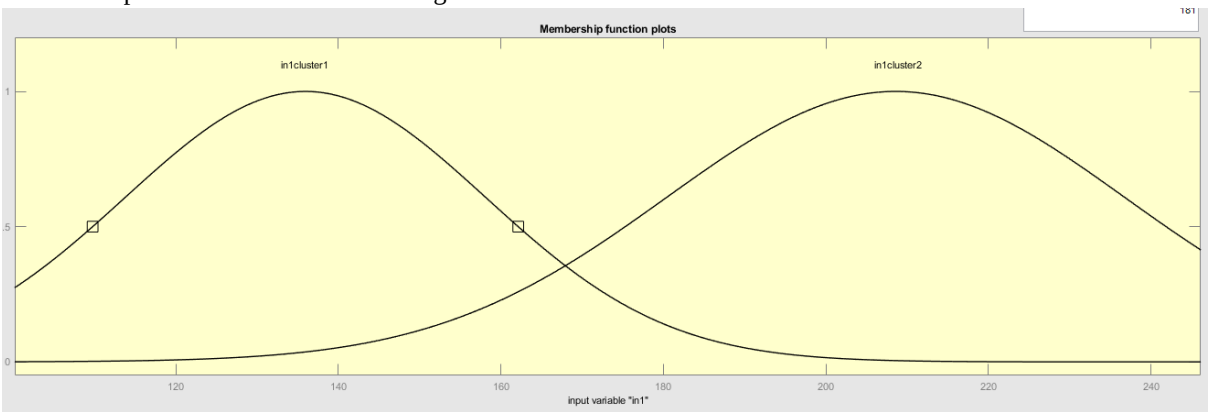
This model was trained with only the dataset generated by the feature selection model of GA. Two rules were automatically generated by Matlab for this model. The below figure shows rules visualization.



For input n1 and output out1, the description of the rules are:

- If n1 belongs to cluster1 then out1 belongs to cluster1
- If n1 belongs to cluster2 then out1 belongs to cluster2

Membership function for n1 and out1 is given below.



Another model was trained on complete dataset with all features and another with dataset for few months only. The accuracy of the model greatly depends on number of features and large datasets.

11 Future work.

There are lot of challenges in time series prediction; many companies are very interested in finding optimum models for time series prediction i.e. facebook came up with their own fbprophet model that use machine learning and neural network for time series prediction. A lot of research has been on going in area of time series prediction and in stock market prediction. If models can be trained to predict stock exchange accurately this will open a new

era of stock trading. Unfortunately the datasets available for machine learning contains only few attributes and since from out training and modeling we found out greater the number of attributes greater the accuracy of the models. Since stock market is effected by many external factors like politics, company cultures, Sales, Import export etc. there should be a way to add these features to dataset available. Many researchers have put forth a model for adding external factors like politics to dataset by rating the political favorable factors for stock value on a scale. This can be used for adding other factors like sales, natural disasters to datasets. These factors create an uncertainty in the stock market and prices fluctuate unpredictably which result in difficulty to train models with extra information. A similar time series model for weather prediction uses hundreds of variable for training and forecasting of weather. Prediction of value using anfis is highly optimal since it requires far less training and less time with high accuracy.

REFERENCES

[1] Khan, W., Ghazanfar, M.A., Azam, M.A. *et al*. Stock market prediction using machine learning classifiers and social media, news. *J Ambient Intell Human Comput* (2020). https://doi.org/10.1007/s12652-020-01839-w

[2] https://en.wikipedia.org/wiki/List_of_stock_market_crashes_and_bear_markets

[3] *Williams, P John; Barlex, David (2019). [Contemporary Issues in Technology Education] Explorations in Technology Education Research (Helping Teachers Develop Research Informed Practice) || Stock Market Data Prediction Using Machine Learning Techniques. , 10.1007/978-981-13-3010-0(Chapter 52), 539–547. doi:10.1007/978-3-030-11890-7_52*

[4] *on Jiang and Jianguo Liu. 2020. Predicting Stock Market N-Days Ahead Using SVM Optimized by Selective Thresholds. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing (ICMLC 2020). Association for Computing Machinery, New York, NY, USA, 11–16. DOI:https://doi.org/10.1145/3383972.3384010*

[5] Deepak, Raut Sushrut, Shinde Isha Uday, and D. Malathi. "Machine learning approach in stock market prediction." *International Journal of Pure and Applied Mathematics* 115.8 (2017): 71-77.

[6] Lakshminarayanan, Sai Krishna, and John McCrae. "A Comparative Study of SVM and LSTM Deep Learning Algorithms for Stock Market Prediction." *AICS*. 2019.

[7] Batra, Usha; Roy, Nihar Ranjan; Panda, Brajendra (2020). *[Communications in Computer and Information Science] Data Science and Analytics Volume 1229 (5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, Gurugram, India, November 15â€ “ 16, 2019, Revised Selected Papers, Part I) || . , 10.1007/978-981-15-5827-6(),* –. doi:10.1007/978-981-15-5827-6 (page58-67)

[8] https://finance.yahoo.com/quote/MSFT/history?period1=1609459200&period2=1616630400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

[9] https://doi.org/10.1016/j.knosys.2010.05.004.