International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June 2018 www.irjet.net p-ISSN: 2395-0072

DEMAND SENSING USING DEEP LEARNING

Rehna Kamil [1], Janisha A [2]

[1] *M.Tech in Computer Science and Engineering, LBS Institute of Technology for Women, Poojappura,Tvm.*
[2] *Assistant Professor, Department of Computer Science and Engineering, LBS Institute of Technology for Women, Poojappura, Tvm.*

Abstract - *The price of oil and gas is tied to major economic activities in all nations of the world, as a change in the price of oil and gas invariably affects the cost of other goods and services. This has made the prediction of oil and gas price a top priority for researchers and scientists alike. In this paper we present an intelligent system that predicts the price of oil and gas. This system is based on ARIMA and Support Vector Machines. The autoregressive integrated moving average (ARIMA) models have been explored in literature for time series prediction. ARIMA models are known to be robust and efficient in financial time series forecasting especially short-term prediction than even the most popular ANNs techniques. Traditionally, the autoregressive integrated moving average (ARIMA) model has been one of the most widely used linear models in time series forecasting series forecasting. However, the ARIMA model cannot easily capture the nonlinear patterns. Support vector machines (SVMs), have been successfully applied in solving nonlinear regression estimation problems. Various deep learning models along with other algorithms will be assessed throughout this work. Deep learning (LSTM) based method will be explored in finding it. Data for our system was obtained from the West Texas Intermediate (WTI) dataset spanning 10 years. To evaluate the performance of the model, the study employs two measures, MSE (regression loss function) and r2_score (regression score function). These are used to compare the performance of the proposed technique and that of ARIMA method for the most efficient in oil and gas price forecasting.*

***Key Words*: Price prediction,Time series, ARIMA,SVM,LSTM,Performance measures.**

1.INTRODUCTION

Time series modeling is a dynamic research area which has attracted attentions of researchers community over last few decades. The main aim of time series modeling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the past. Due to the indispensable importance of time series forecasting in numerous practical fields such as business, economics, finance, science and engineering, etc., proper care should be taken to fit an adequate model to the underlying time series. It is obvious that a successful time series forecasting depends on an appropriate model fitting. A lot of efforts have been done by researchers over many years for the development of efficient models to improve the forecasting accuracy. As a result, various important time series forecasting models have been evolved in literature. Time series forecasting-predicting future values of variables within temporal datasets-is largely an unsolved problem in complex or chaotic domains such as weather(e.g. humidity,temperature or wind speed) and economics(e.g. currency exchange rates or stock prices).Examining the problem with the traditional stochastic model-Autoregressive Integrated Moving Average (ARIMA) and machine learning model-Support Vector Regression(SVR) and with the latest models in the field of machine learning – Long Short Term Memory Network(LSTM) based Recurrent Neural Networks(RNNs)-give rise to new hope of being able to predict time series in these domains. This study focuses on the West Texas Intermediate (WTI) Dataset in United States of America. Fuel Oil Consumption price information is collected. After observation,month and weather are the two main factors that detemine the price has been found. Therefore we aggregate our data according to month.In order to perform forecasts, a model is used to learn from past data. In the literature there are linear models such as Autorregressive Integrated Moving Average (ARIMA) which is able to map linear patterns from the time series. Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are examples of nonlinear methods.

A time series may present trend or seasonal patterns which can be modeled by traditional statistical methods (ARIMA),however nonlinear patterns are not easily captured.Since real world processes often generate nonlinear time series, the employment of nonlinear models could be more appropriate to increase the accuracy of the system.Machine learning techniques and more importantly deep learning algorithms have introduced new approaches to prediction problems where the relationships between variables are modeled in deep and layered hierarchy. Machine learning-based techniques such as Support Vector Machines (SVM).The SVM models specifically,the Support Vector Regression (SVR) is the former approach of regression problems. The SVR is a learning algorithm based on statistical learning theory which employs the structural

risk minimization and is capable of performing generalizations based on past data and handling noise. Deep learning-based algorithms such as Long Short-Term Memory (LSTM) has gained lot of attention in recent years. Deep Learning methods are capable of identifying structure and pattern of data such as non-linearity and complexity in time series forecasting.

The main objective of this paper is to investigate which forecasting methods offer best predictions with respect to lower forecast errors and higher accuracy of forecasts.The key contributions of this paper are:

-Conduct an empirical study and analysis with the goal of investigating the performance of traditional forecasting techniques,machine learning-based techniques and deep learning-based algorithms.

-Compare the performance of ARIMA,SVR and LSTM with respect to minimization achieved in the error rates in prediction.

1.1 Motivation of Research

Modeling the price of oil is difficult because of the changing variability over time. When the demand for a commodity like oil exceeds supply, the price will rise extremely high, which is due to the fact that demand and supply are quite inelastic in the short run. Even though people will be shocked about higher oil prices, it takes time to adjust habits and consumption. Supply cannot be increased that fast either. If production is on the margin of production then adding more capacity is expensive and takes time. Over time people will adjust their oil consumption which will therefore restore the demand-supply balance. In a period of excess demand, oil producers will use their previously uneconomic wells and rigs, as demand adjusts producers will shut off the most costly wells and rigs first. The wide range of oil production cost results in increased medium-term volatility. Therefore shifts in demand will cause a much more drastic change in price than before. A price increase provides quite a challenge , such price movements might seem unpredictable, it is a challenge to find a model that performs relatively well, when being confronted with obstacles.

1.2 DATASET

Data for our system was obtained from the West Texas Intermediate (WTI) datase Fuel Oil Consumption data, spanning 16 years.The training data from January 2000 to December 2016 and predicted next years' data monthwise.The dataset is divided into two different time periods,one is from

January 2000 to December 2017 and one is from January 2004 to December 2017.The historical price of U.S. Fuel Oil Consumption is described in Fig-1.

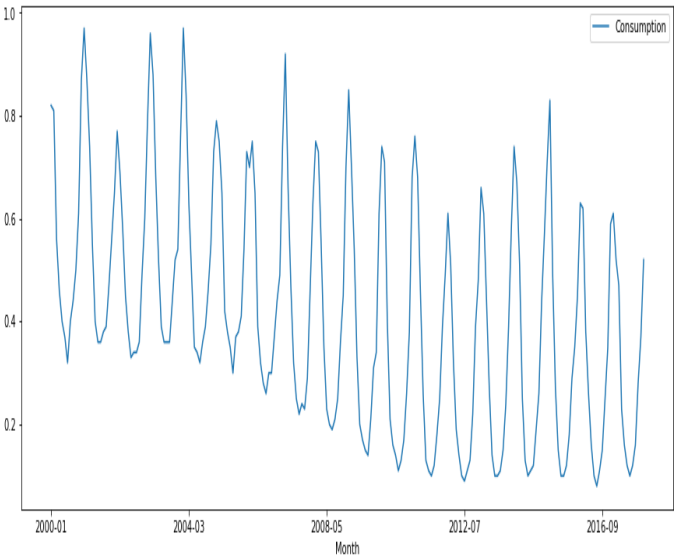

Fig-1: Historical price of U. S. Fuel Oil Consumption, WTI futures(USDOE, 2017).

2. DESIGN AND METHODOLOGY

2.1 Autoregressive Integrated Moving Average Model (ARIMA)

Statisticians George Box and Gwilym Jenkins developed a practical approach to build ARIMA model, which best fit to a given time series and also satisfy the parsimony principle.

ARIMA model is derived by general modification of an autoregressive moving average (ARMA) model. This model type is classified as ARIMA(p,d,q), where p denotes the autoregressive parts of the data set, d refers to integrated parts of the data set and q denotes moving average parts of the data set and p,d,q is all non-negative integers.

ARIMA models are generally used to analyze time series data for better understanding an forecasting. Initially, the appropriate ARIMA model has to be identified for the particular datasets and the parameters should have smallest possible values such that it can analyze the data properly and forecast accordingly.

Their concept has fundamental importance on the area of time series analysis and forecasting. The Box-Jenkins methodology does not assume any particular pattern in the historical data of the series to be forecasted. Rather, it uses a four step approach of model identification, parameter estimation diagnostic checking and output generation to determine the best model from a general class of ARIMA models. Then this model can be used for forecasting future values of the time series. The Box-Jenkins forecast method is schematically shown in Fig. 2.

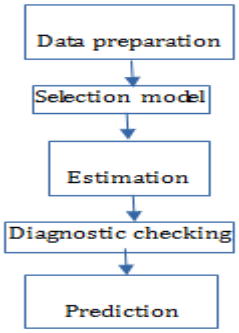

Fig -2: The Box-Jenkins forecast method.

A crucial step in an appropriate model selection is the determination of optimal model parameters. One criterion is that the sample Autocorrelation(ACF) and Partial autocorrelation(PACF), calculated from the training data should match with the corresponding theoretical or actual values .The partial autocorrelation function (PACF) identifies the appropriate lag p in an extended ARIMA(p,d,q) model. Both ACF and PACF are used to check whether the model selected by AIC criterion is appropriate.After selecting the best optimal model,it is used for prediction. The performance is also measured with the actual and predicted data.

2.2 Support Vector Regression(SVR)

The general scheme for SVR methodology is shown in Figure 3. In the first step, data is visualized and analysed.In the next step it is transformed into a form that makes it suitable to build SVR.Nonetheless, before using SVR to predict an outcome, one should select the suitable kernel and hyper parameters first. In the hyper parameter selection phase, Grid search analysis method is used, in order to identify the most suitable hyper parameters for each proposed kernel. The best optimal model with the best hyper parameters can be found. Afterwards, they will be used to train the machine, and the resulting model will be tested in the final test.

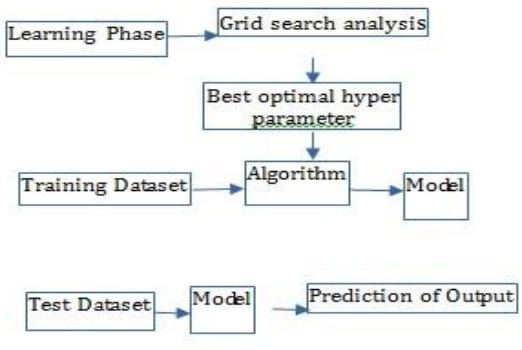

Fig-3: The general scheme of SVR methodolgy

Before choosing the right kernel, it is important to measure the nature of the data, i.e., the distribution of the dataset. The hyperparameter candidates for the three different kernels suggested by grid search and evolution strategy will be compared against each other. The best one, i.e.the one with the lowest error, will be taken as the hyper parameter setting for the final experiment. According to [AS03], if the data distribution is normal, then the Gaussian or RBF kernel is recommended to be used,otherwise the polynomial kernel. The whole samples will be used for training the machine, and the rest for testing the model.Finally, prediction is made and performance is measured with actual data and the predicted data.

2.3 Long Short Term Memory Network (LSTM)

Various types of neural networks can be developed by the combination of different factors like network topology, training method etc. For this experiment, we have considered Recurrent Neural Network and Long Short-Term Memory.

This section we will discuss the methodology of our system. Our system consists of several stages which are as follows:-

- Stage 1: Raw Data: In this stage, the historical stock data is collected from WTI dataset and this historical data is used for the prediction of future oil prices.

- Stage 2: Data Preprocessing: The pre-processing stage involves

a) Data discretization: Part of data reduction but with particular importance, especially for numerical data.

b) Data transformation: Normalization.

c) Data cleaning: Fill in missing values.

d) Data integration: Integration of data files.

After the dataset is transformed into a clean dataset, the dataset is divided into training and testing sets so as to evaluate.

- Stage 3: Feature Extraction: In this layer, only the features which are to be fed to the neural network are chosen. We will choose the feature from U.S Fuel Oil Consumption,Heating Degree U.S. Average, Cooling Degree U.S. Average.

- Stage 4: Training Neural Network: In this stage, the data is fed to the neural network and trained for prediction assigning random values. Our LSTM model is composed of a sequential input layer followed by LSTM layer and dropout and dense layer with ReLU activation.

- Stage 5: Output Generation: In this layer, the output value generated by the output layer of the RNN is compared with the target value. The error or the difference between the target and the obtained output value is minimized by using back propagation algorithm which adjusts the weights and the biases of the network.

The detailed steps of LSTM methodology is described in Fig-4.

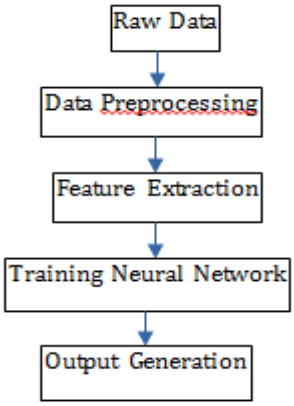

Fig-4: LSTM methodology.

3. EVALUATION AND ANALYSIS

Walk forward testing is an evaluation method used to determine how well a time series prediction model performs over a period of time. The idea is to take a segment of a given data set and partition it into two different parts. The first part of the segment is used as training data for the model, the second part is used for test data for measuring the performance. The segment is then evaluated by calculating the performance and error measurements for the test partition.

After gaining a reasonable knowledge about the series modeling and forecasting, it has to implement them on practical datasets.

3.1 Performation Metrics

The raw data is divided into two parts, viz. The 70% Training Set and 30% Test Set. To judge forecast performances of different methods, the measures Mean Squared Error(MSE) and Coeffiecient of Determination(r2_score) are considered.The Mean Squared Error (MSE) is a measure of average squared deviation of forecasted values.The formula for computing Mean Squared Error is as follows:

$$MSE = 1/n \sum_{i=1}^{n} (y_i - \ddot{y}_i)^2$$

where ÿ is a vector of n predictions generated from sample data on all variables, and y is the vector of observed values of the variable being predicted.

The Coefficient of determination (r2_score) estimates the accuracy of our model, based on a maximum of 1.0 score. It is computed as a value between 0 (0 percent) and 1 (100 percent).The Coefficient of detemination is an important tool in determining the degree of linear-correlation of variables ('goodness of fit') in regression analysis.The formula for computing Coefficient of Determination (r2_score) is as follows:

$$r2\_score = \{(1/n)*\sum[(x_i - x)*(y_i - \bar{y})]/(s_x - s_y)\}^2$$

where n is the number of observations used to fit the model,$\sum$ is the summation symbol, x_i is the x value for observation i, x is the mean x value,y_i is the y value for observation i, $\bar{y}$ is the mean y value, s_x is the standard deviation of x, and s_y is the standard deviation of y.

4. RESULTS

4.1 Data Analysis and Preprocessing

The initial and the most important step in series analysis is determination of whether a time series is stationary or not. From the practical point of view, the series non-stationarity is usually caused by the presence of the trend or seasonality components inside the series.

Very often, simple visual observation of mean and variance functions helps to make suggestion, whether the series is stationary or not .Analysis of ACF and PACF plots is another useful and very informative method for identifying series non-stationarity. It is known, that for a stationary series, the ACF drops to zero relatively quickly, while the ACF of non-stationary series decreases slowly.

The plot of ACF and PACF of series(Fig-5) demonstrates slowly decreasing progress with clear evidence of regularly repeating patterns. This corresponds to the non-stationarity of the series, with the presence of seasonality and trend components.

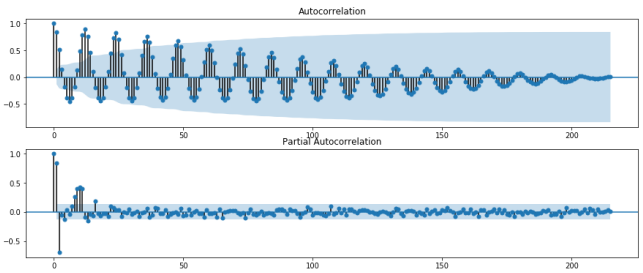

Fig-5: The ACF AND PACF plot of series.

Finally, there has been used an ADF test (Augmented Dickey-Fuller test) to confirm the previous assumptions about the series .This is one of the statistical unit root tests, that is frequently used for determination of series stationarity. The null-hypothesis for an ADF test is that the data are non-stationary. Usually 5% threshold is being used, what means, that null-hypothesis is rejected if the p − value is less than 0.05. The result of the ADF test for the raw data of the given data set confirms the assumption of time series non-stationarity.

Now, there is definitely no doubts, that the time series is non-stationary and that it requires some preprocessing steps to stationarize it. At the beginning, the log transformation has been performed. This helped to stabilize the variance of the series, but it wasn't enough to make it stationary. Another important transformation is differencing. It stabilizes the mean of the series and eliminates the trend and seasonality components. The trend and seasonality can be reduced using Differencing and Decomposing.

4.2 Prediction Results

ARIMA:-According to Box-Jenkins method, in ARIMA (p, d, q) the value of p and q should be 2 or less or total number of parameters should be less than 3 . Therefore, for checking Akaike Information Criteria(AIC) of the model we have only checked for p and q values 2 or less. Both ACF and PACF are used to check whether the model selected by AIC criterion is appropriate. The model with the least AIC value is selected.

In RIMA, the data includes the fuel oil consumption data from 2000 Jan to December 2016 is taken for study. Several iterations are carried out in prediction. In the first iteration, the data from 2000 Jan to Dec 2003 is taken as training set and predict the values for the year 2004.In the second iteration, the data from 2001 Jan to Dec 2004 is taken as training set and predict the values for the year 2005.In the third iteration,the data from 2002 Jan to Dec 2005 is taken as training set and predict the values for the year 2006 and so on until it predict the values for the year 2017. Finally will evaluate the prediction performance for the year 2017. The data of 2017 is reserved for comparison. The forecast diagram of ARIMA model is shown in Fig-6

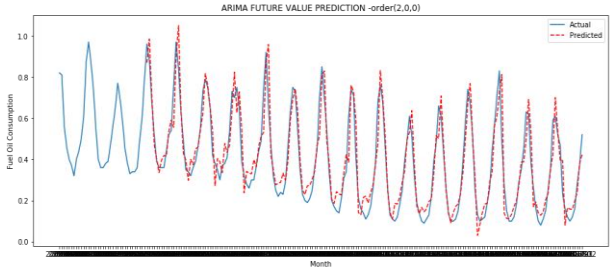

Fig-6 : The forecast diagram of ARIMA model.

The blue line indicates the actual data and the red dotted line indicates the predicted values.

SVR:-The steps involved in SVR methodology includes data visualization, data normalization, kernel and parameter selection and then finally, prediction and output generation. In the first step , data is visualised to check for stationarity. Then data is normalized if necessary, using scaling,log tansformation, differencing etc.In SVR, inorder to predict an outcome, one should select the suitable kernel and hyper parameters first.

The hyper parameter is selected using Grid search analysis method. The best one, i.e. the one with the lowest error, will be taken as the hyper parameter setting for the final experiment. The 70% of samples will be used for training the machine, and the rest 30% for testing the model.

The SVR prediction is divided in to SVR1 and SVR2. The SVR1 is the actual fuel oil consumption data with additional features such as weather factors and the ARIMA predicted output. The SVR2 is the actual data along with changes made in the additional features in weather factors(i.e. previous years' weather conditions are updated with actual weather conditions) and the ARIMA predicted output. The changes that are made in weather factors are to check whether it affects the prediction performance. The data is divided in x inputs such as $x_1,x_2,..x_n$ and a y output . The division of training set and test set are as follows. In SVR, for both SVR1 and SVR2 , the data from 2000 Jan to December 2016 is taken as training set and predict the values for the year 2017.Finally will evaluate the prediction performance for the year 2017. The data of 2017 is reserved for comparison.The forecast diagrams of both SVR1 and SVR2 is represented in Fig-7 and Fig-8.The blue line indicates the actual data, green dotted line indicates the predicted test data and orange line indicates actual test data.

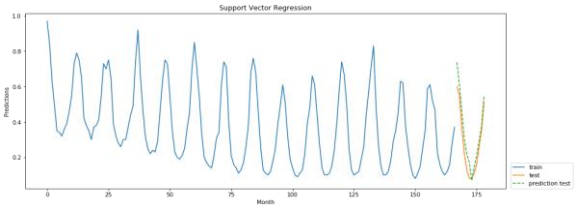

Fig-7: The forecast diagram of SVR1.

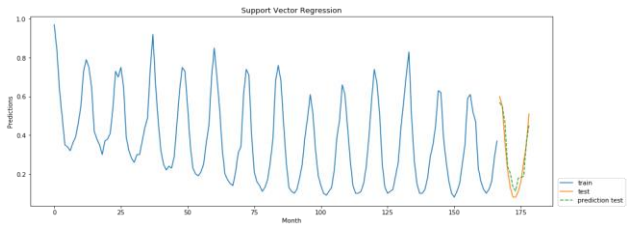

Fig-8: The forecast diagram of SVR2

LSTM:-LSTM can benefit from repeatedly being trained on the same data. Each repetition of training over the full training data set is called an epoch, and the LSTM models have been trained using multiple epochs. After each epoch, the performance of the LSTM is evaluated by measuring its prediction accuracy over the test data set.

The data is divided in x inputs such as $x_1,x_2,..x_n$ and a y output . The data includes the actual fuel oil consumption data with additional features such as weather factors and the ARIMA predicted output. The division of training set and test set are as follows. In LSTM, the data from 2000 Jan to December 2016 is taken as training set and predict the values for the year 2017.Finally will evaluate the prediction performance for the year 2017. The data of 2017 is reserved for comparison.The forecast diagram of LSTM are shown in Fig-9.

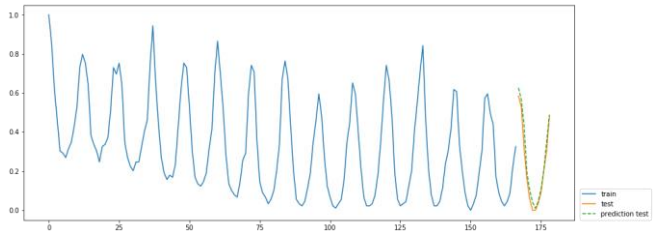

Fig-9: The forecast diagram of LSTM.

Afterwards, three different time series forecasting methods have been tested. The LSTM model presented the best forecasting performance, but all three methods demonstrated the remarkable results and confirmed, that they can be adequately used for time series forecasting.

The performance measures obtained for the series are shown in Table 1:

Table-1: Performance measures.

Models	Mean Squared Error (MSE)	r2_score
ARIMA	0.006	0.820
SVR1	0.004	0.847
SVR2	0.002	0.911
LSTM	0.002	0.949

The Mean Squared Error (MSE) using ARIMA, SVR1, SVR2 and LSTM model are 0.006,0.004,0.002,0.002 respectively.The R2 score of ARIMA, SVR1, SVR2 and LSTM are 0.820,0.847,0.911,0.949 respectively.The SVR1 and SVR2 are feature sets with same data but made differences in the additional features to check how it affects the performance of the model, or the impact it makes on the model performance.From the results, it is clear that its change improves the accuracy and thereby performance of the model. The MSE and R2 score values clearly indicates that LSTM model outperform ARIMA and SVR model (with a reduction in error rates).

It can be seen from the above table that the best performance is obtained by using LSTM model. The forecasting results of all the models presented are done. From the performance measures obtained for each dataset, one can have a relative idea about the effectiveness and accuracy of the fitted models.

5. CONCLUSION

With the recent advancement on developing sophisticated machine learning-based techniques and in particualr deep learning algorithms, these techniques are gaining popularities among researchers across diverse disciplines. The major question is then how accurate and powerful these newly introduced approaches are when compared with traditional methods. This paper compares the accuarcy of ARIMA, SVR and LSTM, as representative techniques when forecasting time series data. It has been seen that, the proper selection of the model orders (in case of ARIMA), the number of input, hidden, output neurons , epoch, batch size and iterations (in case of LSTM) and the hyper-parameters (in case of SVR) is extremely crucial for successful forecasting. We have discussed the important function, viz. AIC which is frequently used for ARIMA model selection. For selecting the number of appropriate neurons, batch size , epochs and iterations , random selection is done and for selecting hyper-parameters, Grid Search Analysis Method is carried out, as mentioned earlier.These techniques were implemented and applied on a set of data and the results indicated LSTM was superior to ARIMA and SVR.

ACKNOWLEDGEMENT

Authors wish to acknowledge deepest thanks to Dr. Jayamohan J, Principal of LBS Institute of Technology for Women and Mr. Manoj Kumar G, Professor and Head, Department of Computer Science and Engineering,LBS Institute of Technology for Women for providing us with all facilities for the completion of this work.

REFERENCES

[1] Adebiyi A.A., Adewumi A.O., (2014) "Stock Price Prediction Using the ARIMA Model", 16th International Conference on Computer Modelling and Simulation.

[2] Alonso A.M., Garc'ia-Martos C., (2012) Time Series Analysis- Forecasting with ARIMA , Universidad Carlos III de Madrid, Universidad Polit'ecnica de Madrid.

[3] Armstrong J.S., (2001) Evaluating Forecasting Methods, A Handbook for Researchers and Practitioners (Ed.J. Scott Armstrong).Kluwer.

[4] Brownlee J., (2017) , How to Create an ARIMA Model for Time Series Forecasting with Python, https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[5] Brownlee J., (2016) , Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[6] Colah's Blog (2015), Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[7] Ding X., Zhang Y., Liu T., Duan J., (2015) "Deep Learning for Event-Driven Stock Prediction", Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI) : 2327-2333

[8] Gamboa J., (2017), Deep Learning for Time-Series Analysis, University of Kaiserslautern, Kaiserslautern, Germany.

[9] Hyndman R.J. (2014), Variations on Rolling Forecasts. https://robjhyndman.com/hyndsight/rolling-forecasts/

[10] Schmidhuber J., (2015), Deep learning in neural networks: An overview, Neural Networks 61:85-117.

[11] Y.Bao, Z.Liu, (2006) Springer, "A fast grid search method in support vector regression forecasting time series", International Conferencce on Intelligent Data Engineering and Automated Learning, pages 504-511.

[12] J. Bergstra and Y. Bengio,(2012), Random search for hyper-parameter optimization ,Journal of Machine Learning Research, pages 281-305.

[13] P.Malhotra, L.Vig, G.Shroff, and P. Agarwal, (2015),"Long short term memory networks for anomaly detection in time series,Proceedings, page 89,Presses universitaires de Louvain.

[14] Roondiwala M. Patel H., Varma S., (2017) "Predicting Stock Prices Using LSTM", International Journal of Science and Research (IJSR) 6(4).

[15] Lee S.I., Seong Joon Yoo S.J., (2017), A Deep Efficient Frontier Method for Optimal Investments, Department of Computer Engineering, Sejong University, Seoul, 05006, Republic of Korea.

[16] Kane M.J., Price N., Scotch M., Rabinowitz P., (2014), Comparison of ARIMA and Random Forest Time Series Models for Prediction of Avian Influenza H5N1 Outbreaks, BMC Bioinformatics 15(1), 276.

[17] T. O. Ayodele, (2010), Machine learning overview, INTECH Open Access Publisher.

[18] M.T.Wiley, (2011), Machine learning for diabetes decision support (pp.55-72),PhD thesis, Ohio University.

BIOGRAPHIES

Author1-Mrs.Rehna Kamil : Recieved her Bachelor's Degree in Computer Science and Engineering from LBS Institute of Technology (LBSITW), Kerala, India. She is currently pursuing her master's degree in Computer Science and Engineering at LBS Institute of Technology, Kerala, India. Her area of interest are Machine Learning, Data Analytics, Image Processing and Computer Networks.

Author2-Mrs.Janisha A : Assistant Professor in Department of Computer Science and Engineering at LBS Institute of Technology for Women, Kerala, India. Her area of interest includes Machine Learning and Computer Networks.