International Journal of Advanced Scientific Inovation Volume 01 Issue 01, December 2020
ISSN:2582-8436 doi:10.5281/zenodo.4641697

Crude Oil Price Forecasting Using Machine Learning

Shambulingappa H S
[#]*Assistant Professor, Department of CSE*
SJMIT Chitradurga, Karnataka, India
shambu.hs13@gmail.com

Abstract **- Crude oil, also called black gold, is naturally available raw petroleum derivative made out of hydrocarbon stores in natural underground repositories. It can fluctuate in color to several shades of yellow and black based on its hydrocarbon blend and stays fluid at a temperature and climatic weight. Crude oil, also called raw petroleum, can be turned into usable petroleum derivatives like diesel, gasoline, several categories of petrochemicals. Trend and seasonality prediction in time series data deals with prediction of future movements of data from the previous analysis of the data. Analysis is based on the idea that what has happened in the past gives traders an idea of what will happen in the future. Data is collection of related information. Data mining is the practice of examining large pre- existing databases in-order to generate new information. Time series is a collection of observations of well-defined data items obtained through repeated measurements. The time series cab be classified into stock and flow. Trend is the slopping line added to relate the two time series or it is continued increase or decrease in series over time. Seasonality is a characteristic of time series in which the data experience regular and predictable changes that recur every calendar year. Any predictable change or pattern in a time series that recurs or repeats over a one-year period can be said to be seasonal. Trend analysis is a statistical technique that deals with time series data**.

Keywords : Time series, Forecasting, ARIMA, Machine Learning.

I. INTRODUCTION

Crude oil, additionally called dark gold, is a normally accessible crude oil subsidiary made out of hydrocarbon stores in common underground vaults. It can change in shading to a few shades of yellow and dark in view of its hydrocarbon mix and remains liquid at a temperature and climatic weight. Crude oil, also called raw petroleum, can be turned into usable petroleum derivatives like diesel, gasoline, several categories of petrochemicals.

Impact of crude oil price in India Indian is one of the world's biggest oil consumers. Among all industry lines, oil and natural gas (ONG) industry has a significant impact on the growth of the country's economy and Gross Domestic Product (GDP). It accounts for at least 15% of GDP. The industry operates in three segments: 1) upstream, 2) downstream, 3) midstream. The upstream segment covers all activities related to exploration and production. The midstream segment deals with stowage and transportation of natural gas and crude oil. The downstream segment deals with fabrication and refining of petroleum products, storage, transportation and marketing of commodities like natural gas and crude oil

Factors Impacting crude oil Prices:

Cost of Raw Petroleum: The Increment in raw petroleum costs in the worldwide market is one critical figure in charge of increment in petrol prices in Indian household market.

Increase in Demand: Petrol prices in India have elevated in response to healthy economics of the country and other emerging Asian countries.

Misalignment between Demand and Supply: Indian oil organizations confront issue to reach the petroleum demand with deficiency of supply and production from refinery centers. Petrol prices shall vary in case there is a misalignment between supply and demand.

Tax: The tax structure for petrol and petroleum products is governed by the Indian government. Oil companies in India will hike fuel prices to recoup losses when there is a jump in tax rate. When the state government increases the VAT on petrol, petrol rates shall increase. If the state government taxes are reduced, petrol rates shall decrease.

Impact of Dynamic Fuel Pricing in India: The Indian government has permitted oil marketing companies to determine the retail price of fuel based on currency exchange rate and fluctuations in international oil prices. OMCs are owned by states. Hence, they are not permitted to hike fuel prices during the election time since it harms consumers. The government instead permits them to price more even if global oil prices are falling. Revising fuel prices on a daily basis leaves no impact to consumers as global oil prices shall not change much on a daily basis. Fuel prices in India are largely affected if there is a major global event affecting crude oil price.

Countrywide Petrol Price Comparison: India has restricted petroleum surpluses. The country's oil consumption is mounting at a decent rate as per recent statistics. Other largest oil consumers are Japan, South Korea, Russia, Mexico, Germany, Canada, and China. Despite currencies, the factors influencing petrol pricing are more or less the same, but the

percentage of allocation will vary from one nation to another.

India is an agricultural-oriented nation and seeks to harness the economy through higher diesel subsidies. Due to which, diesel is priced lower than petrol. This may not be the case with some other economies. Moreover, every country has different importing patterns that will influence Petrol Price. A country with a lower import rate has a lower price charged for both diesel and petrol.

Oil marketing corporations have revised the additional tax remitted towards several state taxes. This has prompted costs of oil-based commodities expanding in states, for example, West Bengal, Gujarat, Karnataka, Maharashtra, and Mumbai, Assam. The Irrecoverable Taxes Compensation Scheme of 2002 was needed to be audited to mirror the costs in states where irreversible duties had seen decay.

Contribution: Distinguished the idea of the marvel spoke to by the arrangement of perceptions is and estimating (future estimations of the time arrangement variable is anticipated).

The example of watched time arrangement information is distinguished and pretty much formally depicted by both of these objectives. We can decipher and incorporate it with other information (that is utilized as a part of hypothesis of researched marvel, illustration regular ware costs) once the example is built up. We can extrapolate the recognized example to foresee future occasions paying little heed to the profundity of our comprehension and the legitimacy of our understanding of the marvel.

II. Related Work

Swati takiyar et al [1], While arranging load sharing, stack streams, stack shedding, stack exchanging, limit building by means of intensity infrastructural improvement and power age to give some examples, are the basic components considered called here and now stack gauges. Since the most recent two decades, this territory of research is presented by excusing the significance. The patterns in specialized arrangement defended through relative examination underpin well known STLF systems on which the sequential survey of the applicable writing is exhibited by an investigation. Single day approach, reenactment models and time arrangement models are the systems secured under this investigation. Relapse models, exponential smoothing, stingy stochastic models, bolster vector machines, master frameworks, man-made brainpower based modes, information mining and cross breed models are the time arrangement models. Due scope is given to significant sub-class models.

Electrical power makers require stack anticipating developing the economy continuously the significant measure of electrical vitality is given. Activity of electrical utilities, outlining and arranging has STLF as a vital part. Transitory estimations of framework's heap, top and vitality, the hourly, day by day, week by week and month to month expectations are worried by stack determining as per net and galiana. Here and now stack estimating (STLF) is up to 1-day, medium load determining (MTLF) is multi day to multiyear and long haul stack guaging (LTLF) is 1-10 years are the spans of arranging skyline which is the premise of load anticipating arrangement as per srinivasan and lee.

Yu Zhou et al [2]**,** In the present-day life, equipment's of hydro power station may occur at any time and public transport vehicles are the catastrophic failure of equipment are the preliminary problem. Operational managers record the necessary steps to avid failures. Open transport organization record the disappointment time, repair cost, down time et cetera. This is seen as arrangement of information after some time. An arrangement of time arrangement will be gotten if the information is organized by time arranges. The time arrangement show utilization is normal, to demonstrate the disappointment information and to estimate the disappointment number. The fundamental examinations are the premise of complex cycle attributes contained in disappointment drift.

To show the disappointment information, the suitable model is auxiliary time arrangement display. The productivity of auxiliary time arrangement demonstrates is represented by the use of true information.

Rodrigo N. Calheiros et al[3], The offered comparable administrations develops by the cloud suppliers through the opposition for end clients, the general population mists on which the cloud-based programming as an administrations applications are sent as organizations move from work area applications. Hazard losing their clients to rivals, cloud based organizations must accomplish great nature of administrations are the focused market where we have to survive. Workloads encounter variety after some time as it is a test of meeting the QOS with a savvy measure of assets. Once discharging them when they are required, propel portion and regarding assets to appraise the future need of utilizations, and proactive unique provisioning of assets are the techniques by which the issues are explained. In light of Autoregressive Integrated Moving Average (ARIMA) demonstrate, the acknowledgment of a cloud workload expectation module for suppliers are exhibited. Utilizing the genuine hints of solicitations to web servers, exactness of future workload is assessed and in light of ARIMA demonstrates the forecast is presented. As far as productivity in asset use and QOS, the effect of accomplished precision is assessed.[5][6].

III. PROPOSED SYSTEM

In present trend Auto-Regression Integrated Moving

Average (ARIMA) [4] is utilized to foresee the future developments in light of the past information. It gives interpretable and accurate results. It is very easy to implement. It is adaptive to changes in the trend and seasonal patterns. Business forecasting, weather forecasting etc. Evaluate current accomplishments of performance. Facilitates comparison.

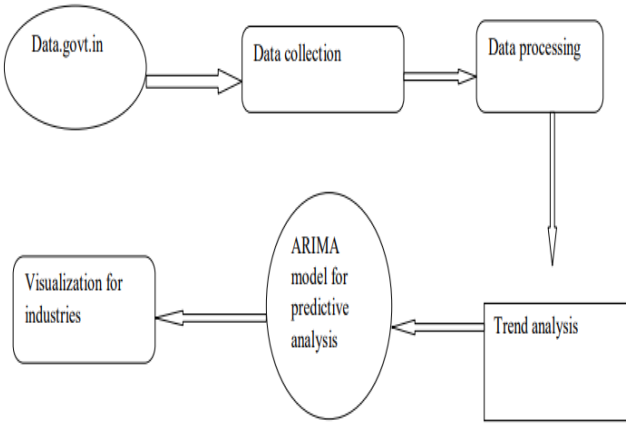


Fig 1: System Architecture

Data collection: it is the well-ordered method for collecting and measuring data from different fields to get approximate and full image of required area. Data collection leads to understand applicable questions and assess results and to predict coming trends and possibilities.

Data pre-processing: it is a data mining process that involves cleaning the needed data, mix the information, information change, and diminishment of chose information and discretization of dataset. The crude information contains numerous mistakes and is regularly fragmented, conflicting. Information pre-handling is a strategy for unraveling such blunders.

Classification: In machine learning, classification is a method in which prediction of the training dataset group is done and is used to recognize the class names on generated dataset. In data mining classification methods can be of following types: regulated learning, unsupervised learning, semi managed learning

Trend analysis: A pattern examination is a piece of specific examination that undertakings to predict the future advancement of stock in light of past data.

ARIMA model: Stochastic demonstrating approaches that can be utilized to ascertain the likelihood of a future esteem lying between two indicated limits. ARIMA (p, d, q): Elrazaz and Mazi created auto backward combination of moving normal. The stationary time series is the consequence of the change of the non-stationary time series and change is finished by using a distinction administrator [7].

Data collection is the well-ordered method for collecting and measuring data from different fields to get approximate and full image of required area. Data collection leads to understand applicable questions and assess results and to predict coming trends and possibilities. Information pre-preparing is an information mining process that includes cleaning the required information, mix the information, information change, and diminishment of chose information and discretization of dataset. The crude information contains numerous blunders and is regularly fragmented, conflicting. Information pre- handling is a strategy for tackling such mistakes. Data is collected from www.data.gov.in website. Collected data is pre-processed in various steps which includes data formatting, Data cleaning and data sampling.

Prediction Algorithm.

Crude oil is an ordinarily happening, foul oil based great made out of hydrocarbon stores and other normal materials. A kind of non-sustainable power source, crude oil can be refined to convey usable things, for instance, gas, diesel and distinctive sorts of petrochemicals [8]. It is a non-practical resource, which infers that it can't be supplanted ordinarily at the rate we eat up it and is in this way a confined resource. It is generally utilized as a part of enterprises; here we are anticipating the generation of couple of rough oils which are as often as possible utilized as a part of businesses. There are different techniques for expectation, which incorporates:

- AR model
- MA model
- ARIMA model

AR model : Auto regression is a period arrangement display that utilizations perceptions from past strides as contribution to a relapse condition to foresee the incentive at whenever step. It is extremely basic thought that can bring about exact figures on a scope of time arrangement issues. An auto relapse display mentions a presumption that the objective facts at past time steps are helpful to anticipate the incentive at whenever step. This connection between factors is called relationship[9].

On the off chance that the two factors alter in same course (eg. Go up together or down together), this is known as a positive connection. In the event that the factors move inverse way as esteem change (eg. One goes up and one goes down), at that point this is called negative relationship. We can utilize factual strategies to ascertain the relationship between's the yield factors and qualities at past time ventures at different

distinctive slacks. The more grounded the relationship between's the yield variable and a particular slacked variable, the more weight that auto relapse model can put on that factor when demonstrating. Once more, on the grounds that the relationship is figured between the variable and itself at past time steps, it is called an autocorrelation as a result of the sequenced structure of time arrangement information.

The connection measurements can likewise pick which slack factors will be valuable in a model and which won't. Curiously, if all slack factors demonstrate low or no connection with the yield variable, at that point it proposes that the time arrangement issue may not be unsurprising. The formation of an autoregressive model creates another indicator variable by utilizing the Y variable slacked at least 1 periods.

$$Yt = f(Yt\text{-}1, Yt\text{-}2, Yt\text{-}p, €t)$$

MA model : Moving normal smoothing is a credulous and successful procedure in time arrangement gauging. It can be utilized for information planning, future building, and even straightforwardly to make expectations. Smoothing is a framework associated with time course of action to clear the fine-grained assortment between steps.

The expectation of smoothing is to expel commotion and better uncover the flag of the hidden easygoing procedures. Moving normal are basic and basic kind of smoothing utilized as a part of time arrangement investigation and time arrangement anticipating. figuring a moving normal includes making another arrangement where the qualities are contained the normal of crude perceptions in the first run through arrangement. A moving normal requires that you indicate a window measure called the window width. This characterizes the quantity of crude perceptions used to figure the moving normal esteem.

The "moving" part in the moving typical implies the way that the window portrayed by the window width is slid along the time course of action to find out the ordinary characteristics in the new plan. There are two standard sorts of moving ordinary that are used central and trailing moving typical.

Centerd moving average: The incentive at time (t) is computed as the normal of crude perceptions at, previously, and after time (t).

For instance, an inside moving normal with a window of 3 would be figured as

Center_ma(t)=mean(obs(t ,obs(t),obs(t+1))

This technique requires information of future qualities, and all things considered is utilized on time arrangement investigation to better comprehend the dataset. An inside moving normal can be utilized as a genera technique to expel pattern and regular segments from a period arrangement, a strategy that we frequently can't utilize when determining.

IV. EXPERIMENTAL RESULT

Scenario 1 finding the trend : We have plotted the graph against price and date. The above graph shows the positive trend and we infer that the price of the crude oil hikes in the coming years.

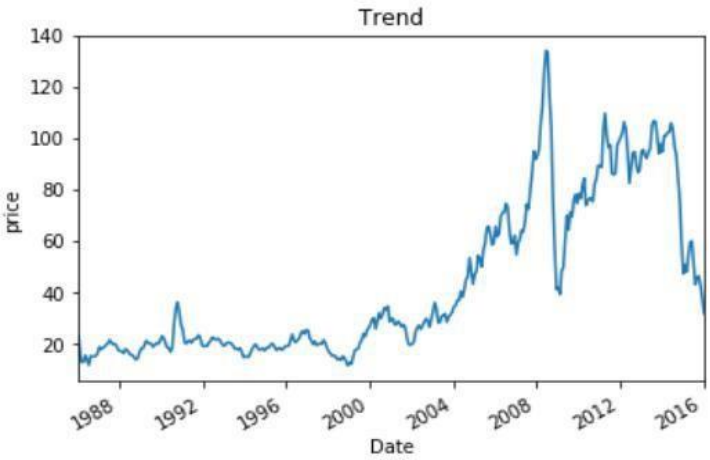


Fig 2: Finding the Trend

Scenario 2 detrend the trend : It shows the graph of detrend, obtained by differencing to achieve stationarity that is mean and variance is constant which facilitates in forecasting easily.

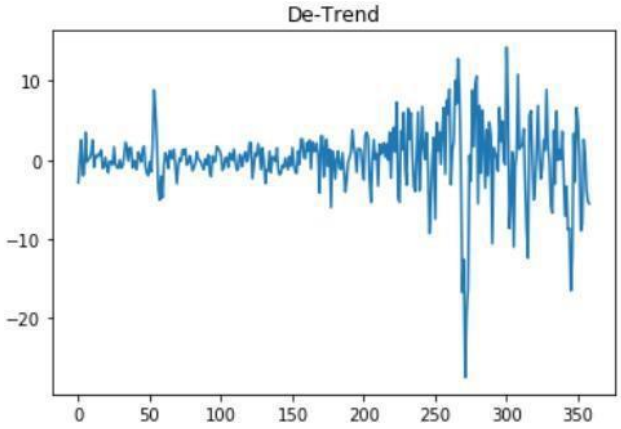


Fig 3: Detrending the Trend

Scenario 3 implementing algorithms: AR model

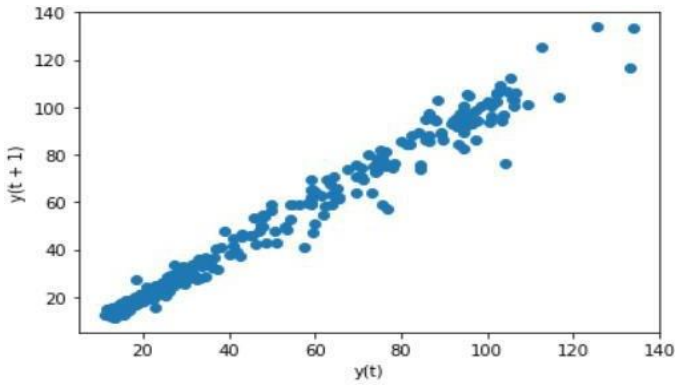


Fig 4: Trend in AR model

Running the illustration plots the raw petroleum information (t) on the x-pivot against the cost on the earlier day (t-1) on the y-hub. We can see an expansive chunk of perceptions along a corner to corner line of the plot. It plainly demonstrates a relationship or some connection.

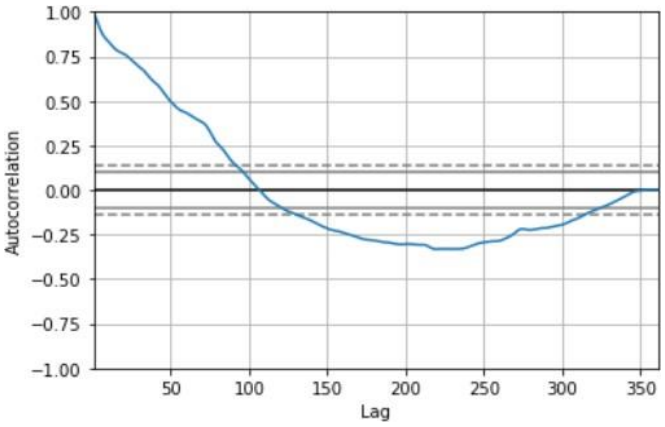


Fig 5: Autocorrelation

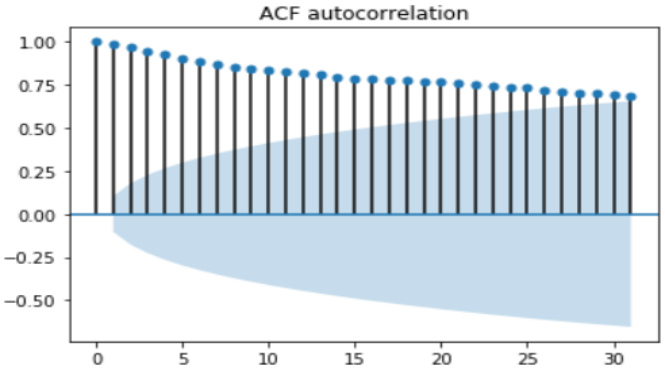


Fig 6: ACF auto correlation

The details models library likewise gives a form of the plot in the plot_acf () work as a line plot.

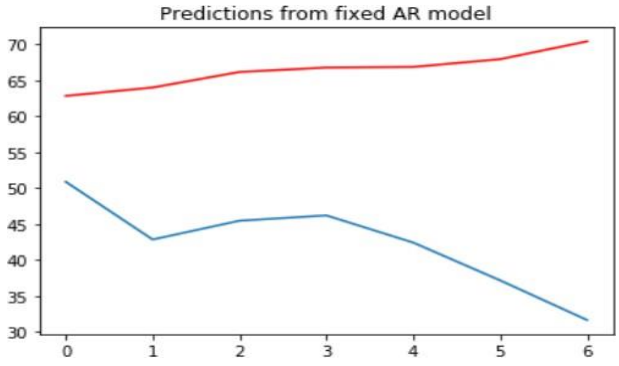


Fig 7: Prediction from fixed AR model

A plot of the normal (blue) versus the anticipated qualities (red) is made. Running the case first prints the picked ideal slack and the rundown of coefficients in the prepared straight relapse show. Test MSE esteem is 640.240

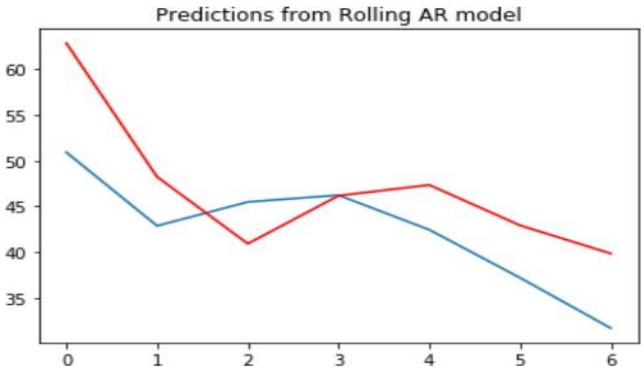
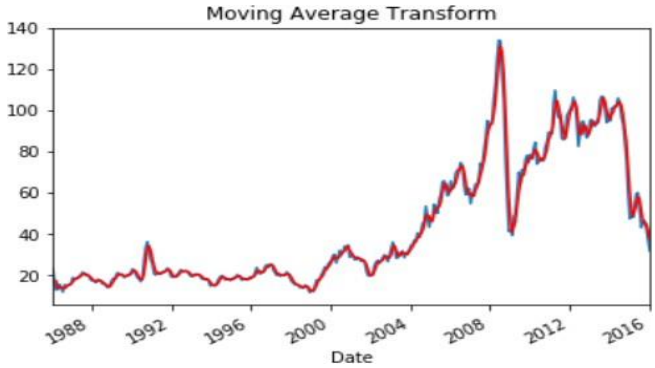


Fig 8: Prediction from Rolling AR model

Running the case prints the estimate and the mean squared mistake. We can see a little change in the conjecture when contrasting the mistake scores. Test MSE esteem is 45.055



Moving average model

Fig 9: Moving Average Transform

The crude perceptions are plotted (blue) with the moving normal change overlaid (red).

ARIMA Model

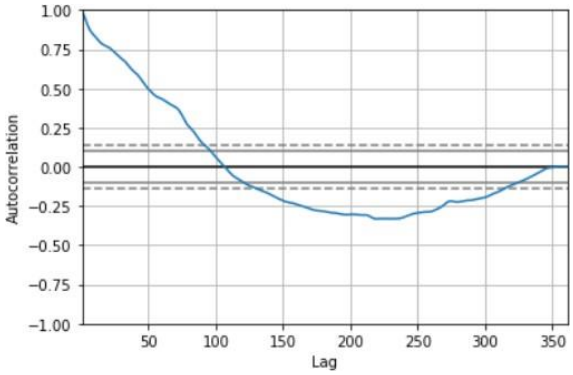


Fig 10: ARIMA Model

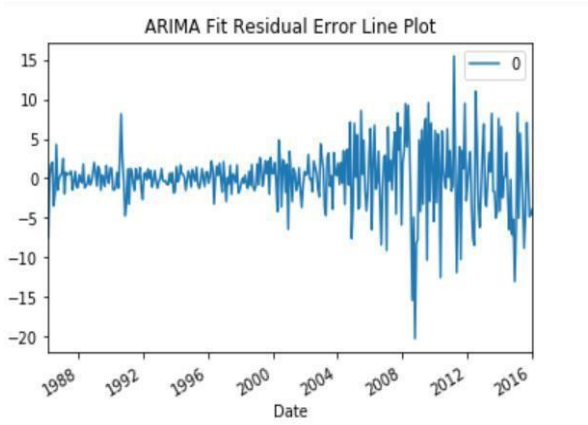


Fig 11: ARIMA fit Residual error line plot

V. CONCLUSION

From the observations we infer that AR model has the highest MSE value. When compared to AR, MA has a massive change in the MSE value. Finally we can see that ARIMA model is having the least MSE value.

Models	Test MSE Value
Auto Regression	**45.454**
Moving Average	**44.552**
ARIMA	**41.246**

REFERENCES

[1] Khashman, Adnan, and Nnamdi .Nwulu."Intelligent prediction of crude oil price using Support Vector Machines." Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on.IEEE, 2011.

[2] Malliaris, A.G., and Mary Malliaris."Time series and neural networks comparison on gold, oil and the euro." Neural Networks, 2009.IJCNN 2009.International Joint Conference on.IEEE, 2009.

[3] Rashmi, T. V., and Keshava Prasanna. "Load Balancing As A Service In Openstack-Liberty." International Journal of Scientific & Technology Research 4.8 (2015): 70-73.

[4] Khanum, Salma, and L. Girish. "Meta Heuristic Approach for Task Scheduling In Cloud Datacenter for Optimum Performance." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4

[5] Jammazi, Rania, and Chaker Aloui."Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling." Energy Economics 34.3 (2012): 828-841.

[6] L. Girish and S. K. N. Rao, "Mathematical tools and methods for analysis of SDN: A comprehensive survey," *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, Noida, 2016, pp. 718-724, doi: 10.1109/IC3I.2016.7918055.

[7] L, G. (2019). "Anomaly Detection in NFV Using Tree-Based unsupervised Learning Method". International Journal of Science, Technology, Engineering and Management - A VTU Publication, 1(2), Retrieved from http://ijesm.vtu.ac.in/index.php/IJESM/article/view/232

[8] Jain, Anshul, and Sajal Ghosh."Dynamics of global oil prices, exchange rate and precious metal prices in India." Resources Policy (2012).

[9] Yi, Yao, and Ni Qin."Short term load forecasting using Time series analysis." Grey Systems and Intelligent Services, 2009.GSIS 2009.IEEE International Conference on.IEEE, 2009.