Indian Institute of Information Technology Allahabad

End-Semester Report On

Predicting Energy Trading Prices Based On Weather And Stock Related Features

(A Multivariate Time-series Analysis)

Under the Guidance of :
Prof. O.P. Vyas
IIIT Allahabad

Submitted by :
Aditya Bahukhandi : IIT2017142
Manav Vallecha : IRM2017007
Neelaksh Trehan : IIM2017002
Nikhil Goyal : IRM2017005

Contents

1 ABSTRACT

Time series datasets have been a subject of interest in recent times. The task is imposing as the datasets are disorganised and have to be manually combined which is a tedious task. The datasets have to be merged based on the timestamp, but before that personalised weights have to be calculated for every dataset. In this report, we have presented how we can combine various time series datasets to have multivariate features. Here we illustrate this with the example of solar energy trading prices dataset of the U.S. Weights are calculated for various states' dataset based on individual state's contribution to the total solar production for the hub. We have implemented LSTM to train our final combined dataset to predict energy trading prices based on various features. Finally we compare the results of ARIMA and LSTM models, which showed that LSTMs are a much better choice in this case, for handling multivariate time series data. Better results are obtained with LSTMs for price prediction.

2 INTRODUCTION

Time series data is of utmost importance in the modern world where every industry wants to enhance its revenue by predicting the future and adjusting their production,consumption or other things according to their business domain. Not only predicting the future, past can also be understood better statistically. The trends learned can help leaps and bounds in increasing the yield of the company. Other than business domain it can be used in monitoring physical systems,software systems,financial trading systems, weather forecasting and various other domains.

The time series dataset is a real backbreaker. Dataset combining is a hectic task in itself especially in-case of a multivariate time series dataset. The datasets usually are not easily available in a single file. One has to manually combine the datasets after studying the individual datasets and getting appropriate weight for each one of them. In this report we are going to present the right approach to combine these datasets by the exemplification of the solar energy trading prices datasets of the US government. The datasets used were Wholesale electricity Data for the PJM West Hub provided by EIA,SEDS dataset,SP index daily price data, Yahoo finance and PSM v3.0 data, National Solar Radiation Database. The final dataset consisted of various features along with the daily energy trading price for the PJM West Hub.[9] We have used LSTM to predict the prices based on the various features. Model is trained in such a way that it learns the trend for the past thirty days and forecast the price of the current day. LSTM was used as we had a large training set and deep learning models have been giving exceptional results in recent research

journals on a comparatively vast dataset.[6] The traditional ARIMA model also requires one to find certain parameters (p,q,d) based on the dataset, which are of no need when LSTM is used as it learns the trend by itself. LSTM also works better on non-stationary time series dataset, where major statistical values like mean and variance are changing over periods of time, which gives LSTM a colossal edge over time-honored time series models.

3 MOTIVATION

The motivation for doing this project was primarily an interest in undertaking a challenging project in an interesting area of research. Nowadays, time series analysis is being extensively used in the area of research . We can see a variety of forecasting problems around us ranging from stock price prediction to weather forecasting, all, in some or the other way, deals with time series data. Time Series forecasting is the use of a model to predict future values based on the trends and variations observed at earlier timestamps. We further researched the topic and came across various papers that implement Machine Learning and Deep Learning strategies in the domain of forecasting problems, this further sparked our interest to take on this project. The best thing about time series analysis is that it can be used to understand the past as well as predict the future. Further, as time series analysis is based on past data plotted against time, it is rather readily available in most areas of study.

This all intrigued us to explore the time-series analysis and hence, we took this project that focuses on determination of energy trading price of a hub (comprising of many states), using a multivariate time series.

4 PROBLEM DEFINITION

We need to combine the necessary data from various datasets of various hub(in each hub there are various states), the data is of electricity prices. The availability and consumption of both of these resources is dependent on the weather, hence involve a high degree of uncertainty and a good degree of variability.

Hence these renewable energy resources are directly linked to the weather parameters. So, in this project we do the “Predicting Energy Trading Prices Based On Weather And Stock Related Features”. We also need to build a suitable and accurate model capable of handling the multivariate time series data and analyse its results.

In this Time Series Data analysis following datasets were used:

- **Wholesale electricity Data for the PJM West Hub provided by EIA**: This dataset contains daily data on electricity prices, trading volume etc for the PJM-West electricity trading hub.
- **State Energy Data System (SEDS), U.S. Energy Information Administration dataset**:
  This dataset is also provided by the US Energy Information Administration(EIA) and was used to extract the solar energy production estimates for the states covered by the PJM West hub from 2000-2013.
- **SP index daily price data, Yahoo finance**:
  This dataset contains the daily prices of the SP IS equity index from 2000-2013.
- **PSM v3.0 data, National Solar Radiation Database**:
  Data of weather parameters like the Direct Normal Irradiance, recorded by the different Class 1 weather stations of 13 states under the PJM West hub's territory.

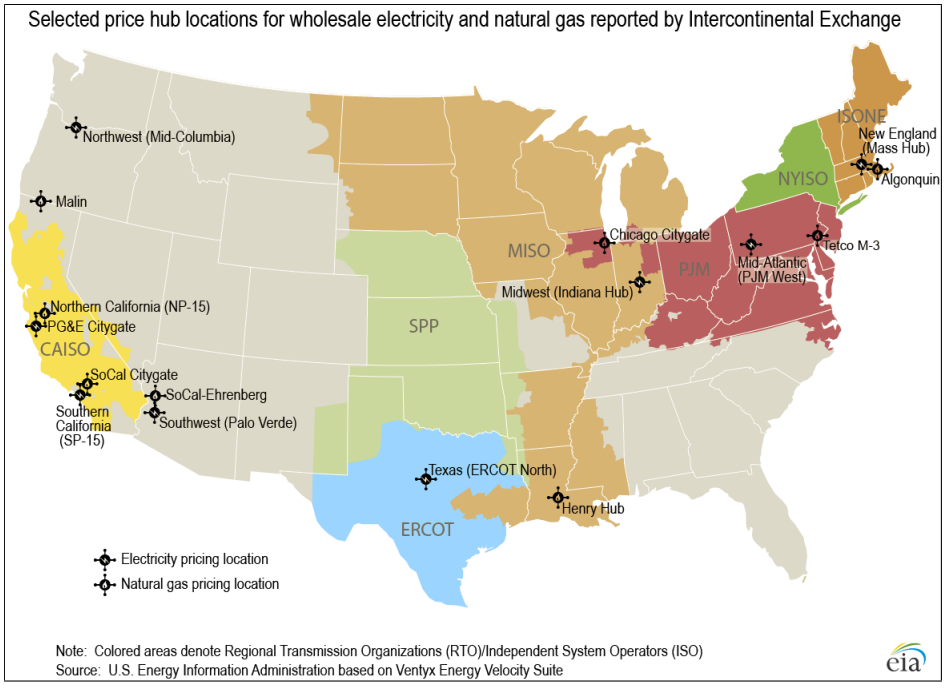

Figure 1: Map showing various energy hubs of U.S. [1]

5 LITERATURE REVIEW

Time series has been explored for centuries, evidence of which can be traced from the field of astronomy. From studying planetary movements to finding solutions for different business issues, time-series analysis finds its uses in every aspect. Time-series analysis has been mostly considered as a statistical problem under the regression analysis. Many regression models exists for time series, and they can be used for representing various stochastic processes.

The main statistical models/techniques used in the analysis of time series data are ARIMA (Auto Regressive Integrated Moving Average) and ARMA (Auto Regressive Moving Average). They are very much efficient for estimating appropriate values, when used on linear/univariate data.[2]
The major difficulty while dealing with time series data is that sometimes measured data goes missing. Acknowledging this, many statistical methods, like Mean substitution, Hotdecking (substitution by other similar, complete data based on correlated characteristic) etc., have been developed for the imputation of the missing values.[3]

While the results predicted by these regression models fit the actual trend well, they fail to do much good on time series that are often influenced by unpredictable external factors, like price series.[4] In such cases, these models are better-off predicting a range rather than precise values. This leads to the classification of time series into two types, Multivariate and Univariate. In the Univariate Time Series, only one variable varies over time while in the Multivariate Time Series, more than one variable affects the result. Multivariate time series models can give more accurate predictions because of more number of variables involved, which makes the model more general and aware of all those variables affecting the trend.[5]

This clearly demands the need of some model, that can effectively capture the multivariate time series.

While some machine learning techniques are well introduced along with the traditional time-series forecasting techniques, deep-learning techniques remains less explored. The paper by Bhaskar P. Murkhoty and Vikas Maurya on Solar Irradiation Forecasting explores how sequence to sequence deep learning models can be applied to time series data of solar irradiation to obtain good accuracies.[6] LSTMs were found to give better accuracies than the traditional ARIMA models for long term modelling and were able to capture the more complex non linear trends.

To prevent the overfitting and to prevent bad accuracy on the test set the cross validation technique can be used, which is basically splitting the dataset into training set, cross validation set and the test set.[7] RMSE(root-mean-square error) can be used to evaluate the performance of the model. To further decrease the error, the

k-fold cross validation can be increased, which will also obtain the final results much faster.

Many other complex deep learning architectures have also been explored, like using a combination of LSTMs and Convolutional Neural Networks (CNN).[8] CNN can be applied with machine learning by viewing the multivariate time series as a sequence of space time images.
The paper on Predictive Modelling of Electricity Trading Prices and the Impact of Increasing Solar Energy Penetration by Soumyo V. Chakraborty and Sandeep K. Shukla presents how predictive modelling may be done using standard Machine Learning models like SVM, Random Forests, and Gradient Boosted Decision Trees.[9]This research analyzes the daily electricity price at two major power trading hubs namely the PJM West and Palo Verde. The data is analysed over a 16 year period . The major highlighting factor was that as solar penetration increased in a region , Solar irradiance became more dominant weather parameter instead of Temperature. The solar irradiance is the output of light energy from the entire disk of the Sun, measured at the Earth. 16 years of daily data was available and 75% of the daily data was used to train the model and 25% as a test set. TMY3 dataset contained about 1020 locations. TMY3 is Typical Meteorological Year (TMY) data. For each trading hub weighted averages of the weather parameters were computed. Weighted averages were proportional to the states contribution to the total solar energy for the hub. TMY3 datasets were obtained from the source National Solar Radiation Database (NSRDB), National Renewable Energy Laboratory. Hourly values of weather variables including solar irradiance and temperature were obtained at desired locations.
The major weather parameters to keep in mind for the prediction purpose are the everyday DNI (Direct Normal Irradiance), which measures the irradiation coming from the sun in a straight line, the Dry Bulb Temperature, the DHI, which accounts to the irradiation due to scattering coming from other directions, and the GHI which is DNI and DHI measured together.

Taking inspiration from these papers we use LSTMs for the predictive modelling of electricity prices with hopes of getting a better accuracy.

6 PROPOSED METHODOLOGY

6.1 Dataset Preparation

Dataset combining, the real backbreaker, was done using the following datasets:

1. Wholesale electricity Data for the PJM West Hub provided by EIA : containing

daily data on electricity prices, trading volume etc for the PJM-West electricity trading hub.[10]

2. SEDS dataset : This dataset is also provided by the US Energy Information Administration(EIA) and was used to extract the solar energy production estimates for the states covered by the PJM West hub from 2000-2013.[11]
3. SP index daily price data, Yahoo finance : This dataset contains the daily prices of the SP IS equity index from 2000-2013.[12]
4. PSM v3.0 data, National Solar Radiation Database : Data of weather parameters like the Direct Normal Irradiance, recorded by the different Class 1 weather stations of 13 states under the PJM West hub's territory.[13]

As part of the dataset preparation first the PSM data of the 13 states under the PJM West hub territory was extracted. This data included information about different weather parameters of the class 1 weather stations in the 13 states. One weather station from each state is chosen as it's representative and the weighted average for each weather parameter is computed based on that state's contribution to the total solar production for the hub. The weights were extracted from the SEDS dataset which contains the solar production estimates of each state from the years 1960-2017. We used the data from 2000-2013 for the estimation of weights.
We also computed the daily mean of the weather parameters for each date since as they were available in 30 min interval gap for each day.
For the next step we merged the data from the SP index data and The Wholesale electricity data and the PJM west data based on the timestamp and merged it with the processed PSM data obtained from the first step. The features that were extracted are depicted in Table 1.

The data thus obtained was cleaned further to remove any missing values and

S.No.	Feature Name
1.	High Price of the day
2.	Low Price of the day
3.	Wtd Avg Price
4.	Daily Volume MWh
5.	Number of Trades
6.	Average DNI fro the day
7.	Mean Temperature of the day
8.	Average DHI
9.	Mean Pressure
10.	Wind Speed
11.	Opening S&P index price
12.	Highest S&P price of the day
13.	Lowest S&P price of the day
14.	Closing S&P index price
15.	Volume traded

Table 1: List of Features

redundancies that might have persisted and was normalised using the *MinMaxScaler* function from the *sklearn* library in python. The data was then split into training (80%) and test sets (20%) after sorting it by the timestamp.

6.2 Model Training

6.2.1 ARIMA

AutoRegressive Integrated Moving Average (ARIMA) has been the conventional time-series data prediction technique for quite some time. Time series data can be broken into three parts trend, seasonality and noise. Before applying any statistical model, we need to ensure that our time series is stationary. As ARIMA is also a statistical model, we did the same for it. Here the dataset consisted only of the Energy Trading Prices where indexing was done datewise.
Following two tests were used to make sure that the time series is stationary:

1. **Rolling Statistics**: Plot the rolling mean and rolling standard deviation against time. If they remain constant then the time series is stationary.
2. **Augmented Dickey-Fuller Test**: Time series is stationary if the p value according to the null hypothesis is low and the critical values for 1%, 5%, 10% confidence intervals are as close as possible to the ADF Statistics.

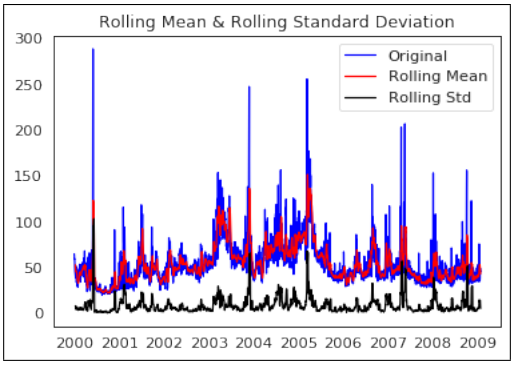

Figure 2: Rolling statistics showing the non-stationary time series

As can be seen in Fig. 2, the rolling mean is showing irregularity so we applied log to the dataset as it decreases the rate of irregularity, which can be seen in Fig. 3.

Now we applied a sequence of steps to make our dataset stationary. We subtracted

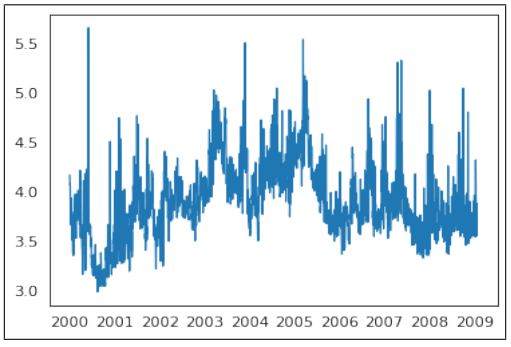

Figure 3: Plot of the log of the dataset

the rolling mean on a window of 12 to get the new dataset. The plot in Fig. 4 shows the rolling statistics for the obtained dataset.

Now upon applying time shifting by subtracting every point by the point that preceded it, the time series became almost stationary as can be seen from the plot in Fig. 5.

6.2.2 LSTM

- With the necessary data cleaned we built an LSTM model with a look back value of 30. This means that the electricity price and weather parameter values of 30 days would be used in predicting the value for the 31st day at a given instant of time.

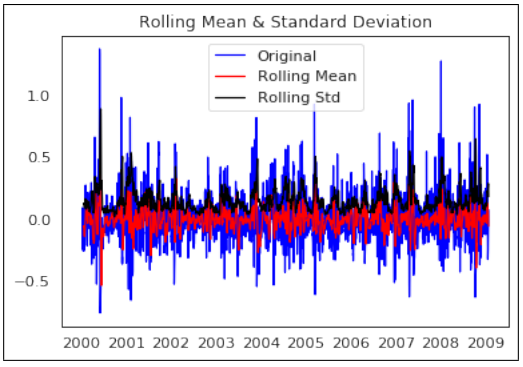

Figure 4: Rolling statistics after subtracting the rolling mean

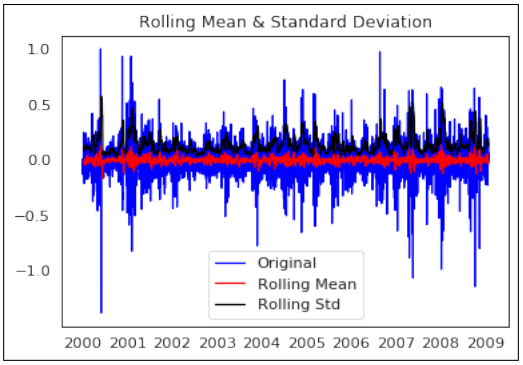

Figure 5: Rolling statistics plot showing the shifted dataset.

- Using the keras API with tensorflow 1.2.1 as the back-end we trained an LSTM model on the data across 100 epochs using a single Dense layer keeping the Dropout as 0.2. This is done so as to introduce some randomness into the network so that it does not over-fit the training set.
- Mini batch gradient descent was used with batch size as 64 and ADAM as the optimizer, ensuring faster convergence to an optimum value. Early stopping is also used to control over-fitting to the training set.

7 REQUIREMENTS

7.1 Hardware Requirements

- Any machine that supports the Google Chrome web browser
- GPU for training and testing the machine learning model

7.2 Software Requirements

- Google Chrome
- A google account that has access to google collaboratory
- Python 3.6
- Tensorflow v1.14 with keras API
- Pandas, Numpy, Sklearn, statsmodels python libraries
- matplotlib and seaborn libraries for plotting purposes
- Spyder IDE or Jupyter Notebook

8 IMPLEMENTATION PLAN

8.1 Pre Mid-Sem

1. We successfully extracted combined the SP index data from yahoo finance and PJM west electricity price data available from EIA.
2. Studied machine learning techniques Artificial Neural Networks and LSTM and how these could be applied for time series forecasting.

8.2 Post Mid-Sem

1. Explored the TMY3 dataset that contained old data of weather parameters from the years 1990-2003. We also observed that the data was not continuous which would have proven to be detrimental to the modelling process. Also different class I weather stations had missing data on different dates so combining them had become a real problem.We therefore used the PSM v3.0 data from NSRDB instead which had continuous data of the weather parameters of the class I weather stations in a continuous fashion and from the required years 2000-2013.
2. Fully combined the necessary data from four datasets, namely Wholesale electricity Data for the PJM West Hub provided by EIA, the SP index daily price data from Yahoo finance PSM v3.0 data from the National Solar Radiation Database and the SEDS dataset from EIA and obtained the dataset that was used for the training of the models.
3. Built an ARIMA model and compared it with the LSTM model.

4. Built an LSTM model and trained it on google collab to predict the next day's electricity price of the hub based on 30 previous days parameter values.

Task	Start Date	End Date
Literature review	11-08-2019	05-10-2019
Finalizing Problem Statement and Goals	25-08-2019	27-08-2019
Gathering and Pre-processing of data	20-08-2019	25-10-2019
Training and testing of models	01-11-2019	12-11-2019

Table 2: Implementation Timeline

9 RESULTS

Three integers (p, d, q) are typically used to parametrize ARIMA models.
p: number of autoregressive terms (AR order)
d: number of nonseasonal differences (differencing order)
q: number of moving-average terms (MA order)
We finally applied the dataset after all the transformations into the ARIMA model, with p, d and q set as 2,1 and 2 and got the result shown in Fig. 6.
Finally we inverted all the previously applied transform on the original dataset, to

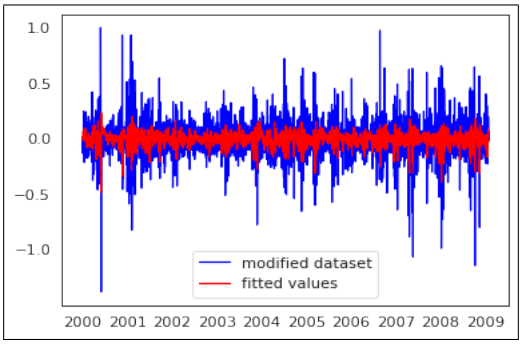

Figure 6: Predicted and the modified dataset Prices, ARIMA

get the real predicted value with respect to the original dataset and got the result shown in Fig. 7.
Root mean square error in ARIMA model came out to be **7918.05562703935**.

The LSTM model that we used for training yielded the results that can be seen through the plots in Fig.8 and Fig. 9.

LSTM, Train Mean Absolute Error: **21.701400477562824**

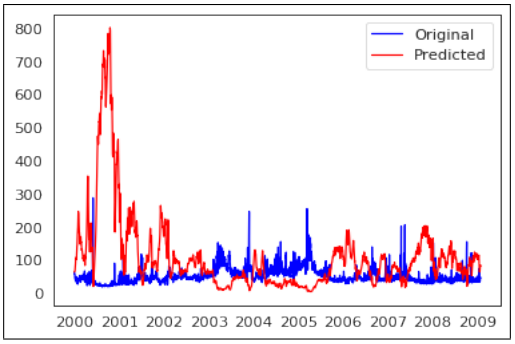

Figure 7: Predicted and Original Prices, ARIMA

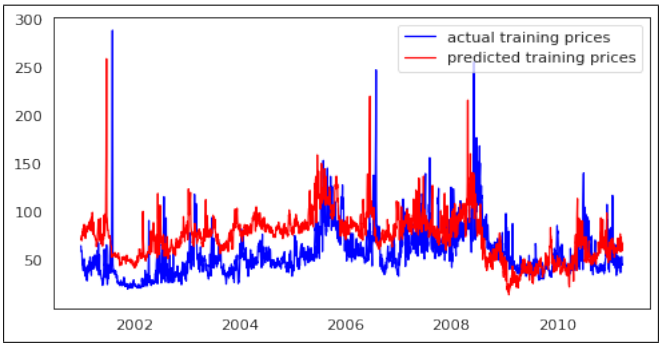

Figure 8: Plot showing the predicted vs the original prices in the training dataset, LSTM

LSTM, Train Root Mean Squared Error: **25.186087818346813**

LSTM, Test Mean Absolute Error: **21.619826146137868**
LSTM, Test Root Mean Squared Error: **24.058920905440996**

10 CONCLUSION

Throughout the entire project phase we came across various researches presented in the field of TIme Series Modelling . There are many time series models used in the recent past like ARIMA but they fail to work well on multivariate time series. We also saw the impact of LSTMs when used with Time Series. We combined the 4 datasets that are defined in the report and created a combined time series data with timestamp of one day that was used to train the LSTM.
In our analysis of energy trading prices forecasting we came to conclude that LSTM's

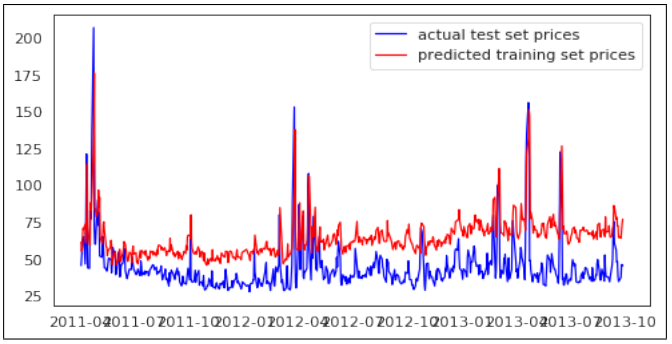

Figure 9: Plot showing the predicted vs the original prices in the test dataset, LSTM

outperformed classical ARIMA model as we can see root mean square error ***7918.05562703935*** of ARIMA is more than the root mean square error ***24.058920905440996*** of LSTM model . The final outcome was shown by prediction of Energy trading prices proving LSTM to be better in case of multivariate time series data.

11 REFERENCES

1. Image taken from `https://www.eia.gov/electricity/wholesale/` [Online Image]

2. Pasapitch Chujai, Nittaya Kerdprasop, and Kittisak Kerdprasop, "*Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models*", International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong

3. Y. S. Afrianti, S. W. Indratno and U. S. Pasaribu, "*Imputation algorithm based on copula for missing value in timeseries data*", 2014 2nd International Conference on Technology, Informatics, Management, Engineering Environment, Bandung, 2014, pp. 252-257. doi: 10.1109/TIME-E.2014.7011627

4. Y. Zhao and L. Shen, "*Application of time series auto regressive model in price forecast*", 2011 International Conference on Business Management and Electronic Information, Guangzhou, 2011, pp. 768-771. doi: 10.1109/ICBMEI.2011.5921078

5. S. Lei, C. Sun, Q. Zhou and X. Zhang, "*The research of local linear model of short term electrical load on multivariate time series*", 2005 IEEE Russia Power Tech, St. Petersburg, 2005, pp. 1-5. doi: 10.1109/PTC.2005.4524543

6. Bhaskar Pratim Mukhoty, Vikas Maurya, Sandeep Kumar Shukla, "*Sequence*

to sequence deep learning models for solar irradiation forecasting", Indian Institute of Technology Kanpur.

7. A. Sarah, K. Lee and H. Kim, "*LSTM Model to Forecast Time Series for EC2 Cloud Price*", 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Athens, 2018, pp. 1085-1088.
8. Vladimir Stojov, Nikola Koteli, Petre Lameski, Eftim Zdravevski, "*Application of machine learning and time-series analysis for air pollution prediction*", CiiT 2018 - 15th International Conference on Informatics and Information Technologies, 2018, Mavrovo, Macedonia.
9. Soumyo V. Chakraborty and Sandeep K. Shukla, "*Predictive Modeling of Electricity Trading Prices and the Impact of Increasing Solar Energy Penetration*", Indian Institute of Technology Kanpur.
10. Wholesale electricity Data for the PJM West Hub provided by EIA, [Online]. Available: `https://www.eia.gov/electricity/wholesale/`
11. State Energy Data System (SEDS), U.S. Energy Information Administration dataset, [Online]. Available: `https://www.eia.gov/state/seds/`
12. SP index daily price data, Yahoo finance, [Online]. Available: `https://finance.yahoo.com/quote/%5EGSPC/`
13. PSM v3.0 data, National Solar Radiation Database, [Online]. `https://maps.nrel.gov/nsrdb-viewer/?aL=UdPEX9%255Bv%255D%3Dt%26f69KzE%255Bv%255D%3Dt%26f69KzE%255Bd%255D%3D1&bL=clight&cE=0&lR=0&mC=4.740675384778373%2C22.8515625&zL=2`