International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Forecasting Household Electricity Usage

Utkarsh Sahaya[1], *Yash Raj Kesarwani*[2]

PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur

ABSTRACT :

With the development of smart electricity metering technologies, huge amounts of consumption data can be retrieved on a daily and hourly basis. Energy consumption forecasting facilitates electricity demand management and utilities load planning. Every year, energy consumption grows worldwide. Therefore, power companies need to investigate models to better forecast and plan the energy use. One approach to address this problem is the estimation of energy consumption at the customer level. Energy consumption forecasting is a time series regression task. Machine learning techniques have shown promising results in a variety of problems including time series and regression problems. In this work, we propose a system to predict daily energy consumption using Machine Learning and Deep Learning. One deep learning model was studied: LSTM Neural Network, and two Machine Learning models: Multiple Regression, Random Forest Regression.

Keywords: Household Energy Forecasting, Multiple Regression, Random Forest Regression, Time Series, Recurrent Neural Networks, Machine Learning, Deep Learning

INTRODUCTION :

The consumption of electricity has increased substantially over the years, fuelled primarily by the growth in population, rapid increase in production and supply. Electricity has reached each and every corner of the world, there is electricity supply even in the most backward villages of India. Electricity consumption per capita in India is 940 kWh (in 2020) which continues to grow rapidly by 4% year-over-year [1]. Currently India's electricity sector is dominated by coal, which during the 2018-19 fiscal year produced about three-quarters of the country's electricity, whereas renewable energy only accounts for 20% of the electricity production in India [2]. Coal being a fossil fuel makes the production of electricity quite expensive as well as it contributes to the worsening of climate because of its carbon emission property. Hence efficient energy production as well as efficient consumption is a general concern from people to governments.

Forecasting is aimed at anticipating the future as precisely as possible so that it helps the government as well as people on knowing about how to manage electricity consumption. Energy management systems (EMS) were created to efficiently monitor, control, and optimise the performance of the generation or transmission system. The demand for electricity consumption by customers is affected by many uncertainties, and an accurate forecasting model has an important role in improving the operation and planning of the power system so that a balance is achieved between power consumption and power production in order to reduce operating costs.

Considering the aforementioned, the objective of this paper is to review a few forecasting methods which may be used to predict energy consumption. The main contributions of this paper are:

- Literature review of the existing studies for the forecasting of energy consumption, exposing their contributions and limitations.
- Analysis of the different types of methods used in the forecasting of energy consumption from multiple perspectives.

LITERATURE SURVEY :

Computational intelligence models are developed by measuring the inputs and outputs of the system and fitting a linear or non-linear mathematical model to approximate the operation of the building. These models are based on the implementation of a function deduced only from samples of training data describing the behaviour of a specific system, being this way well suited when physical relations are not known. For building such advanced computational intelligence models over physical methods is that the former does not require any physical phenomena to deduce an accurate prediction model. However, the lack of proper data can become an issue for the use of computational learning methods because the accuracy is strongly dependent on the quality and amount of available data.

According to the mentioned works, in which many cases were analysed, these techniques are found to be quite effective. Among these methods, Recurrent Neural Networks (RNN) are the primary models employed to forecast household electricity usage [3]. The main input data used to feed this technique is based on house-related parameters.

Considering the house-related parameters, the total energy consumption data is the most used variable (as an endogenous variable), followed by parameters such as voltage, global reactive power, and sub-metering. The prediction horizon of reviewed studies is segmented in daily fractions with varying prediction time steps.

ANN were used which are trained to overcome the restriction of the traditional methods to solve complex problems. Neural networks consist of an interconnection of a number of neurons. There are many varieties of connections in literature. However, this study focuses only on one type of network, which is called the multi-layer perceptron [4].

In one study K-means clustering approach was also used to estimate power and detect anomalies over time [5]. This study was aimed at monitoring energy loads and detecting aberrant loads and behaviour, which is critical to power grid.

PRELIMINARY ITEMS

About Dataset

The Household Power Consumption dataset is a multivariate time series dataset that describes the electricity consumption for a single household over four years. The data was collected between December 2006 and November 2010 and observations of power consumption within the household were collected every minute. It is a multivariate series comprised of seven variables (besides the date and time):

global_active_power: The total active power consumed by the household (kilowatts).
global_reactive_power: The total reactive power consumed by the household (kilowatts).
voltage: Average voltage (volts).
global_intensity: Average current intensity (amps).
sub_metering_1: Active energy for kitchen (watt-hours of active energy).
sub_metering_2: Active energy for laundry (watt-hours of active energy).
sub_metering_3: Active energy for climate control systems (watt-hours of active energy).

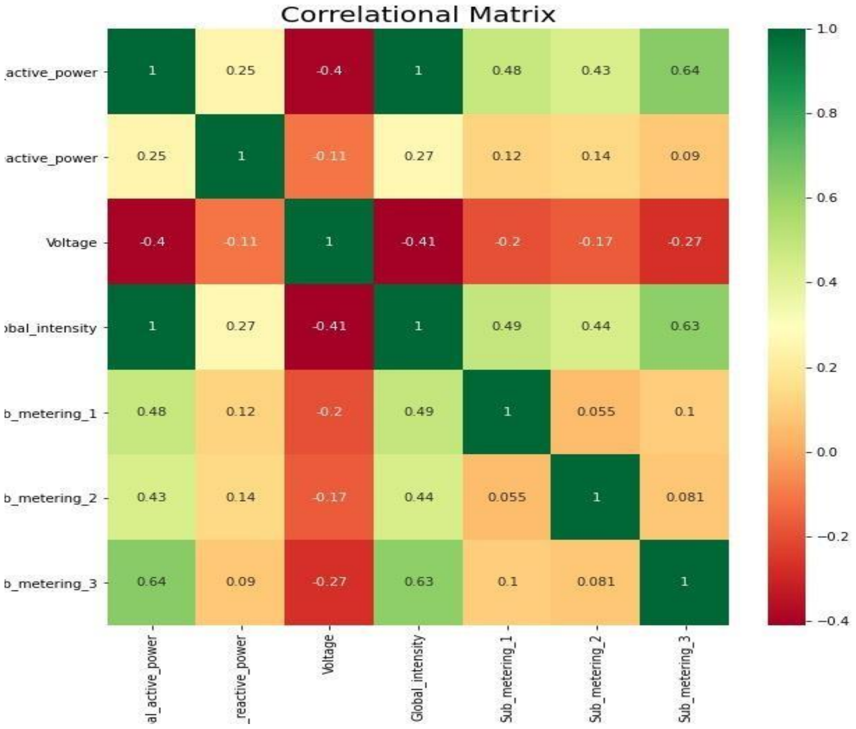
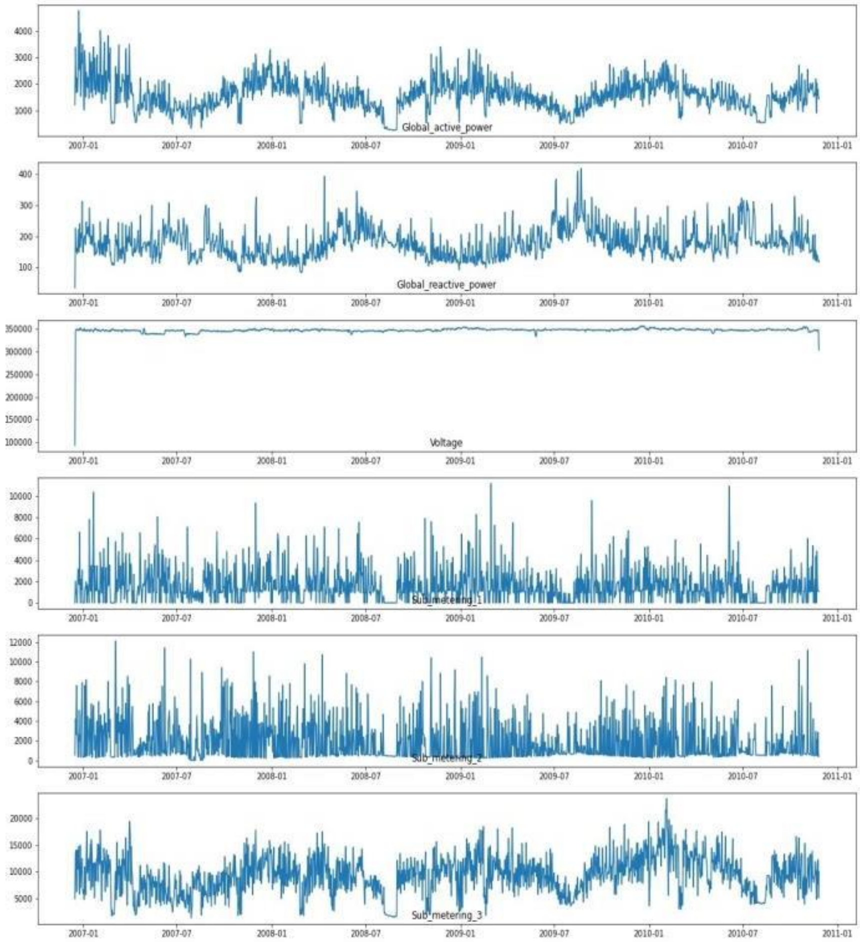


Figure. 2. Correlation Matrix of all variables.

Figure. 1. Line Plots of all variables.



The original data had total 2075259 time series data point of minute-based time. In our analysis the dataset was resampled to 1 day time-period which resulted in 1442 data points with frequency of daily data. The resampling allowed lesser consumption of processing power on the RNN and the predicted result was on a daily basis and not minutely basis. It also allowed for better data visualization. Also, upon pre- processing it was found that the variable global_intensity was having perfect correlation with global_active_power. In order to avoid bias in our supervised machine learning models this variable was entirely dropped from our implementations.

The time series data also had missing values. In order to have a holistic view of the data. The missing values were dealt in three ways.

df_dropped: The dataset obtained by dropping the missing values of time series.
dataset_mean: The dataset obtained by filling the mean value of that variable in place of the missing values of the time series.
dataset_median: The dataset obtained by filling the median value of that variable in place of the missing values of the time series.
dataset_mode: The dataset obtained by filling the mode value of that variable in place of the missing values of the time series.

Models Implemented

Models ranging from the popular Neural Networks models (RNN) to supervised machine learning namely Multiple Regression and Random Forest Regression were implemented on all the four pre-processed dataset namely df_dropped, dataset_mean, dataset_median, dataset_mode.

In Multiple Regression it involves more than one explanatory variable. Hence the other 5 variables served as the explanatory variables for the independent variable (global_active_power). It used least square method of estimation. In time series analysis it is possible to do regression analysis against a set of past values of variables. This is known as Autoregression (AR).

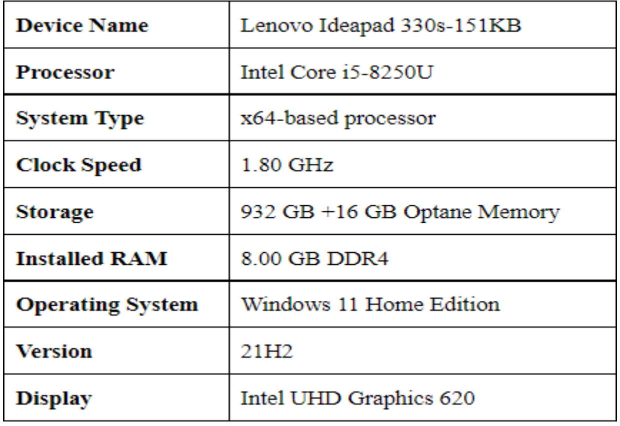
The next supervised method used is Random Forest Regression which uses ensemble learning method for regression. A random forest operates by constructing several decision trees during training and outputting the mean of the classes as the prediction of all the trees. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The third model which uses deep learning LSTM is an artificial recurrent neural network (RNN) architecture. An artificial neural network is a layered structure of connected neurons, inspired by biological neurons. Unlike the standard feedforward neural networks, LSTM has feedback connections. LSTM module has three gates which provide them power to selectively learn, unlearn and retain information.

EXPERIMENT

System Configuration

The experiment was performed on system with following capabilities.



Device Name	Lenovo Ideapad 330s-151KB
Processor	Intel Core i5-8250U
System Type	x64-based processor
Clock Speed	1.80 GHz
Storage	932 GB +16 GB Optane Memory
Installed RAM	8.00 GB DDR4
Operating System	Windows 11 Home Edition
Version	21H2
Display	Intel UHD Graphics 620

Models

We have used three models namely Multiple Regression, Random Forest Regression and LSMT(Long Short-Term Memory) on four preprocessed datasets namely df_dropped (where null tuple values have been dropped), dataset_mean (where null tuple values have been replaced by mean of remaining values), dataset_median (where null tuple values have been replaced by median of remaining values) and dataset_mode (where null tuple values have been replaced by mode of remaining values). Finally we have checked two metrics (**r squared score** and **RMSE score**) in predicted vs. test data.

Multiple Regression

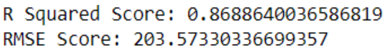
The estimating equation of Multiple Regression is.

$$\widehat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2}$$

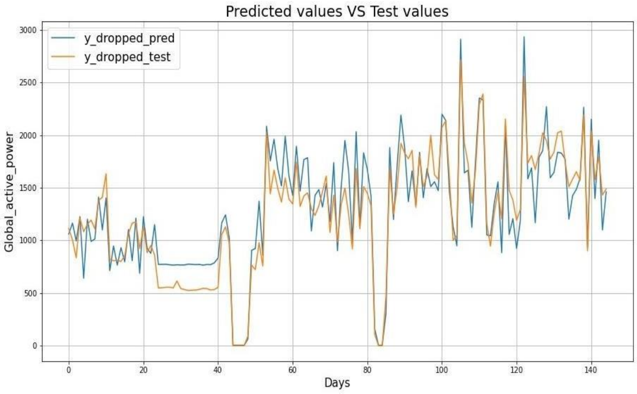
The model is fit by Least square method and the prediction can be done as

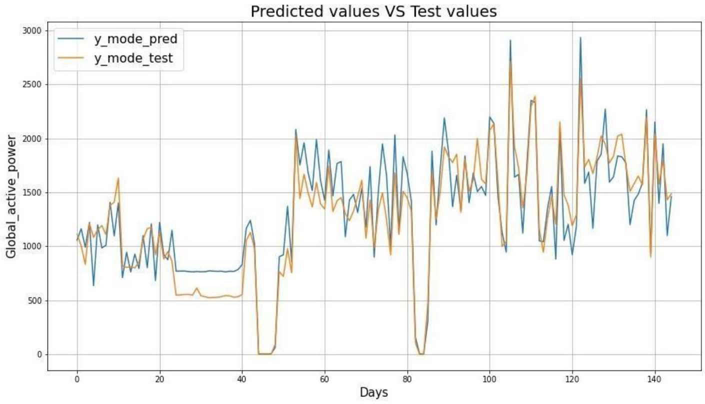
$$\mathrm{y = Xb + e}$$

1.a. **df_dropped:**

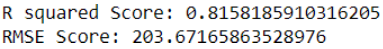


```
R Squared Score: 0.8688640036586819
RMSE Score: 203.57330336699357
```

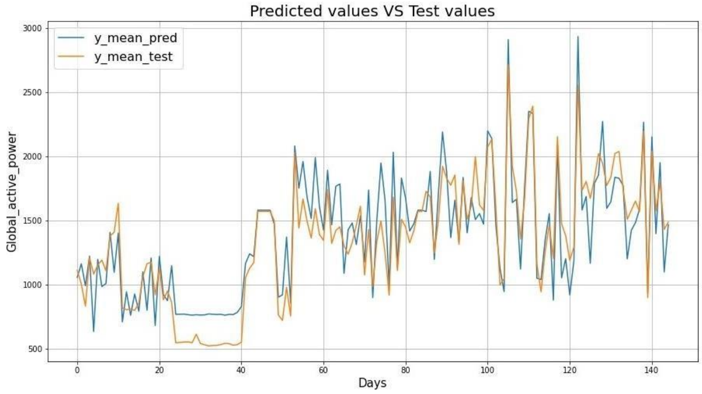




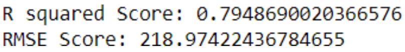
1.b. **dataset_mean**:



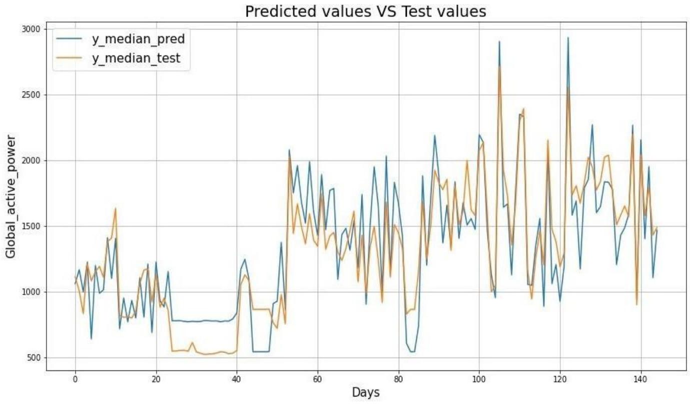
```
R squared Score: 0.8158185910316205
RMSE Score: 203.67165863528976
```



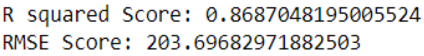
1.c. **dataset_median**:



```
R squared Score: 0.7948690020366576
RMSE Score: 218.97422436784655
```



1.d. **dataset_mode:**



```
R squared Score: 0.8687048195005524
RMSE Score: 203.69682971882503
```

Random Forest Regression

2.b. dataset_mean:

The mathematical equation of the decision tree in Random Forest is.

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

Where :

RFfi sub(i) = the importance of feature i calculated from all trees in the Random Forest model.

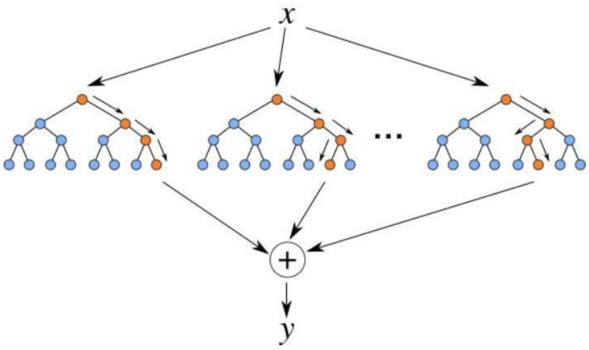
normfi sub(ij) = the normalized feature importance for i in tree j

T = total number of trees

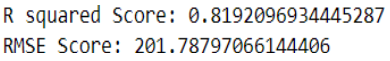
The equation that decides the distance of each node from the predicted value, helping to decide which branch of the tree is the better decision of the forest is.

$$MSE = \frac{1}{N}\ \sum_{i=1}^{N}(fi - yi)^2$$

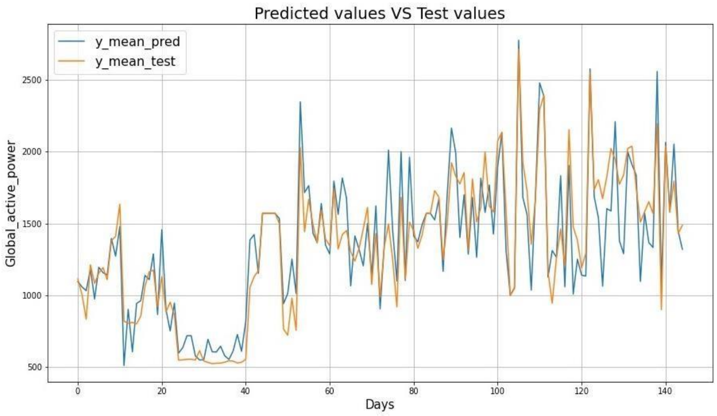
Figure. 3. Figure showing working of Random Forest Regression



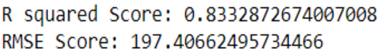
df_dropped:



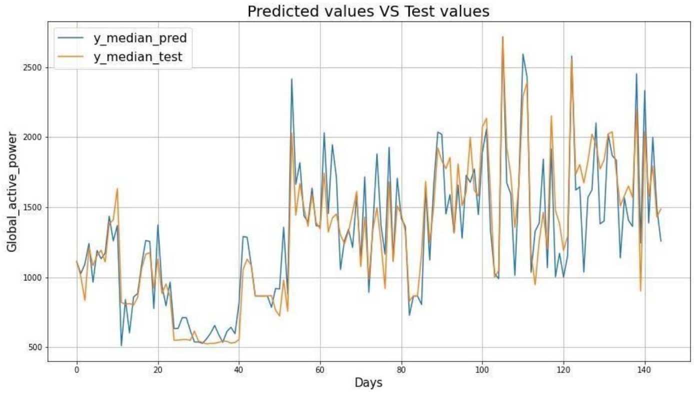
```
R squared Score: 0.8192096934445287
RMSE Score: 201.78797066144406
```



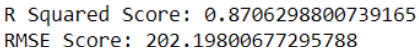
dataset_median:



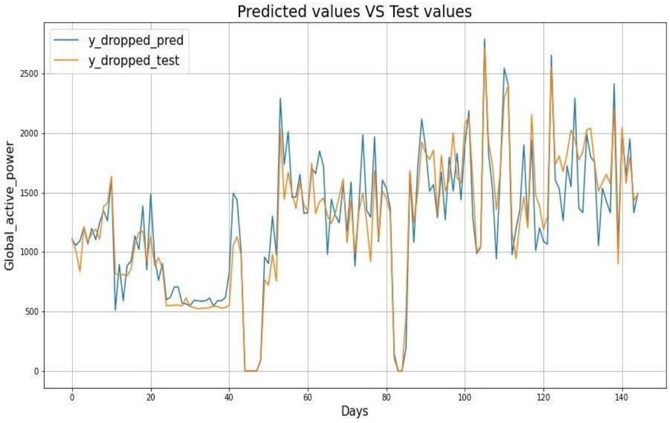
```
R squared Score: 0.8332872674007008
RMSE Score: 197.40662495734466
```



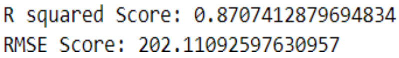
dataset_mode:



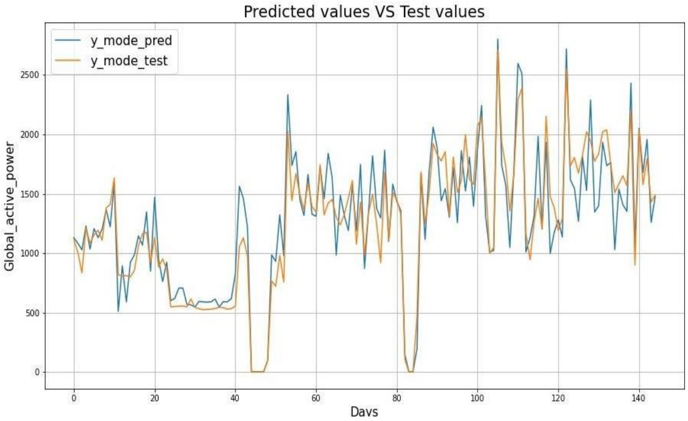
```
R Squared Score: 0.8706298800739165
RMSE Score: 202.19800677295788
```



3.



```
R squared Score: 0.8707412879694834
RMSE Score: 202.11092597630957
```



LSTM 3.b. **dataset_mean**:

The equation for the gates in LSTM are

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

Where

$$i_t \rightarrow represents\ input\ gate.$$

$$f_t \rightarrow represents\ forget\ gate.$$

$$o_t \rightarrow represents\ output\ gate.$$

$$\sigma \rightarrow represents\ sigmoid\ function.$$

$$w_x \rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.$$

$$h_{t-1} \rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).$$

$$x_t \rightarrow input\ at\ current\ timestamp.$$

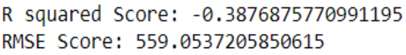
$$b_x \rightarrow biases\ for\ the\ respective\ gates(x).$$

The equation for cell state, candiate state and final state are.

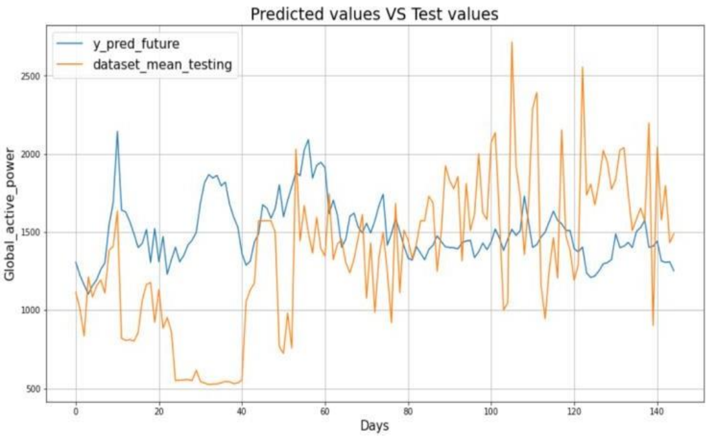
$$\tilde{c}_t = tanh(w_c[h_{t-1}, x_t] + b_c)$$
$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$
$$h_t = o_t * tanh(c^t)$$

$$c_t \rightarrow cell\ state(memory)\ at\ timestamp(t).$$
$$\tilde{c}_t \rightarrow represents\ candidate\ for\ cell\ state$$
$$at\ timestamp(t).$$

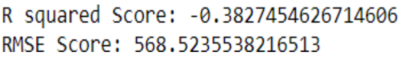
df_dropped:



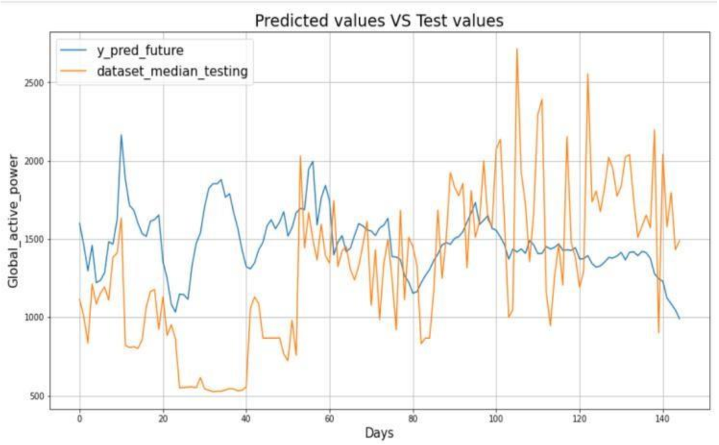
```
R squared Score: -0.3876875770991195
RMSE Score: 559.0537205850615
```



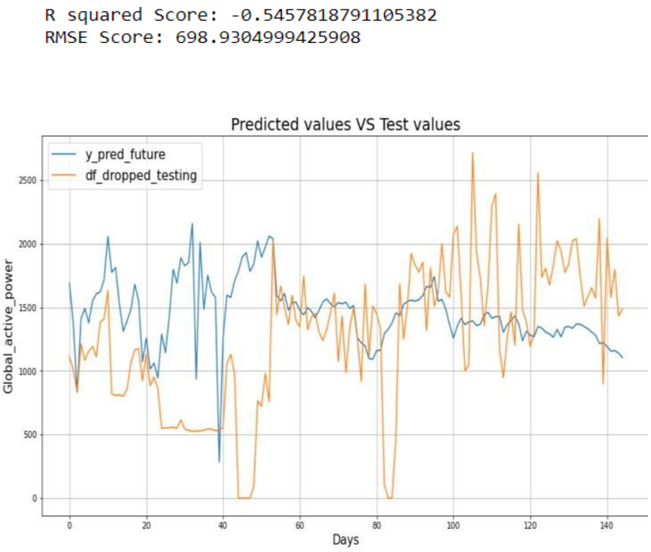
dataset_median:



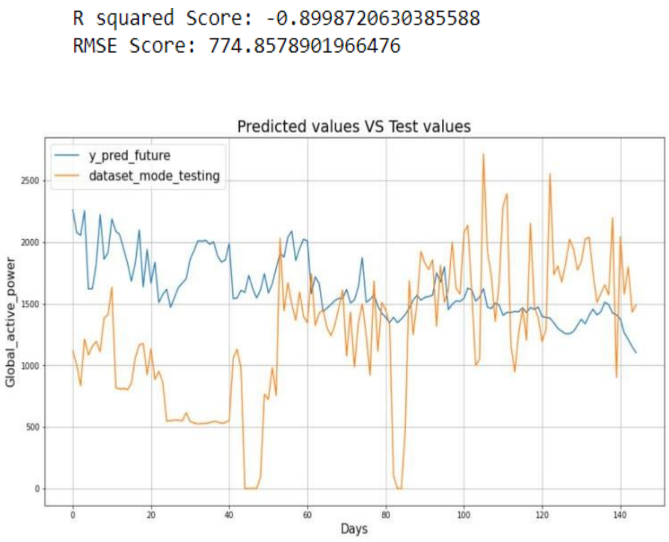
```
R squared Score: -0.3827454626714606
RMSE Score: 568.5235538216513
```



dataset_mode:



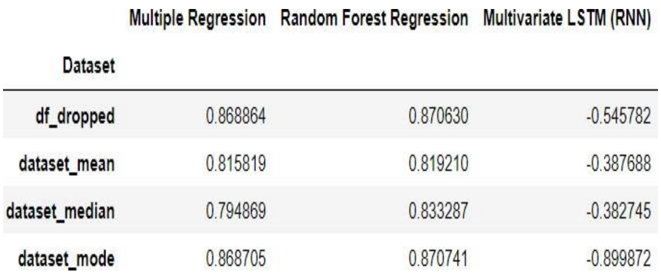
```
R squared Score: -0.5457818791105382
RMSE Score: 698.9304999425908
```



```
R squared Score: -0.8998720630385588
RMSE Score: 774.8578901966476
```

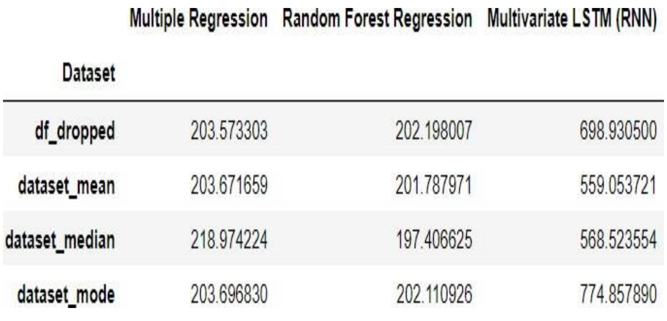
Table Comparing the Models Tested

Table. 1. R-squared score of all Models for each dataset



Dataset	Multiple Regression	Random Forest Regression	Multivariate LSTM (RNN)
df_dropped	0.868864	0.870630	-0.545782
dataset_mean	0.815819	0.819210	-0.387688
dataset_median	0.794869	0.833287	-0.382745
dataset_mode	0.868705	0.870741	-0.899872

Table. 2. RMSE score of all Models for each dataset



Dataset	Multiple Regression	Random Forest Regression	Multivariate LSTM (RNN)
df_dropped	203.573303	202.198007	698.930500
dataset_mean	203.671659	201.787971	559.053721
dataset_median	218.974224	197.406625	568.523554
dataset_mode	203.696830	202.110926	774.857890

Result

Random Forest Regression is found to perform better overall with R-squared score of 0.87 for both df_dropped and dataset_mode and also low RMSE scores of 202 and 202 respectively for the two above mentioned datasets.

In Multiple Regression the 4 datasets produce considerable different accuracy score. But here also df_dropped and dataset_mode gave better results than the other two datasets.

For LSTM model the predicted values fail to match the short-term volatility in our approach.

Hence overall Random Forest Regression proved to be the better among the three models by a considerable margin. And among the datasets df_dropped (obtained by dropping null values) and df_mode (obtained by replacing missing values with mode) performed better than the other two datasets.

PROPOSED METHOD :

Random Forest Regression is the proposed model. On top of giving the best results it has several benefits over other algorithms. Random Forest Regression is popular because of its simplicity and high accuracy. It can solve a variety of problems ranging from needs to predict a continuous time series value as in our case of even classification problem.

The main advantage of using random forest is that it scales well thus new variables or samples can be easily added to the dataset. Random Forest is easy to use and understand. It is not sensitive to missing data and can handle the outliers very well. It has much higher speed than the LSTM model. Whereas Random Forest is fast and with impressive accuracy with large datasets. A random forest is a collection of Decision Trees, Each Tree independently makes a prediction, the values are then averaged (Regression) / Max voted (Classification) to arrive at the final value. The strength of this model lies in creating different trees with different sub-features from the features. The Features selected for each tree is Random, so the trees do not get deep and are focused only on the set of features. Finally, when they are put together, we create an ensemble of Decision Trees that provides a well-learned prediction.

CONCLUSION :

Electricity generation, transmission and distribution facilities require an investment of billions of dollars. Therefore, forecasting electricity consumption is very important for the investors and companies. Adequate capacity planning requires accurate forecasts of the future demand variations and timing of electricity demand. Since short-term electricity consumption data are nonlinear data with strong volatility and instability. The prediction model of Random Forest Regression proposed in this paper can predict data with strong holiday effects and seasonal fluctuations well and can also handle outliers well, it can effectively reduce the prediction error and improve the prediction accuracy. Therefore, the model can effectively forecast short-term electricity consumption, which is conducive.

The analysis done in the paper involved pre-processing the and arriving at 4 different datasets. df_dropped is the dataset obtained when the missing values are dropped. dataset_mean is the dataset obtained by filling mean value in place of missing values, similarly dataset_median and dataset_mode is also there. On doing through analysis, it was found that -

- In general, Random Forest Regression works better among the 3 models. It has fairly high R-squared score for all dataset. It has relatively low RMSE value for all dataset.
- The dataset df_dropped and dataset_mode is comparable in results and better than the other two datasets used.
- We can use either of the dataset (df_dropped or dataset_mode) to perform our forecasting.
- If the data has lot of missing values and the time series data is almost stationary in nature then filling the missing values with mode will be more suitable i.e., dataset_mode should be used.
- If the dataset has few missing values (<2%) and it is non-stationary in nature then dropping the missing values will be more suitable i.e., df_dropped should be used.

REFERENCES :

1. : Center for Policy Research (https://cprindia.org/news/6519)
2. : Government of India Ministry of Power (powermin.gov.in/en/content/power-sector-glance-all-india)
3. : Semantic Scholar (Short-Term Electricity Consumption Forecasting Based on LSTM Method)
4. : Forecasting electricity consumption: A comparison of regression analysis, neural networks. (Elsevier)
5. : Power consumption Prediction and Anomaly Detection based on K mean (Frontiers in Energy Research)

Journals Referred

1. Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines (Fazil Kaytez , M. Cengiz Taplamacioglu, Ertugrul Camb, Firat Hardalac) (ELSEVIER)
2. Short-Term Electricity Consumption Forecasting Based on LSTM Method (Guorong Zhu, Sha Peng , Yongchang Lao, Qichao Su , Qiujie Sun)
3. A Review of Energy Consumption Forecasting in Smart Buildings: Methods, Input Variables, Forecasting Horizon and Metrics (Deyslen Mariano-Hernández , Luis Hernández-Callejo , Felix Santos García, Oscar Duque-Perez, Angel L. Zorita- Lamadrid)
4. Forecasting Electricity Consumption in Residential Buildings for Home Energy Management Systems (Karol, Antonio Ruano, Maria da Graça Ruano)
5. Monthly Energy Consumption Forecast: A Deep Learning Approach (Rodrigo F. Berriel, Andr´e Teixeira Lopes, Alexandre Rodrigues, Fl´avio Miguel Varej˜ao, Thiago Oliveira-Santos)

6. Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means(Junfeng Zhang, Hui Zhang, Song Ding, Xiaoxiong Zhang)
7. A Review of Deep Learning Techniques for Forecasting Energy Use in Buildings (Jason Runge and Radu Zmeureanu)
8. Forecasting Residential Energy Consumption Using Support Vector Regressions (Xiaoou Monica Zhang, Katarina Grolinger, Miriam A. M. Capretz)
9. Short-term power load forecasting based on EMD-grey model (American Journal of Electrical Power and Energy Systems),(J. Dong, P. Wang, and X. Dou,)

Websites Referred

1. analyticsvidhya.com
2. towardsdatascience.com
3. medium.com
4. keboola.com
5. javatpoint.com
6. tutorialspoint.com
7. geeksforgeeks.org