Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

Weather based Electricity Consumption Prediction using Big Data Analytics Model

Mr. Asif Karim
Department of Computer Science And Engineering
SRM Institute of Science and Technology,Kattankulathur

Mr. Rootul Patel
Department of Computer Science and Engineering
SRM Institute of Science and Technology,Kattankulathur

Mrs. B. Ida Seraphim
Department of Computer Science and Engineering
SRM Institute of Science and Technology,Kattankulathur

***Abstract* — Electricity is being consume in every home at present time , we are using electricity for lots of purposes like watching television, charging cell phones, Electric bulb and many more things. Analyzing this electricity data will give us a better understanding of how consumer consumes electricity in their daily life. Apart from user daily electricity consumption there are also External factor which plays a major role in defining the amount of electricity consumed. One of the major external factor is Weather , like in a sunny day a family can use Cooler , AC, Refrigerator which increases the electrical consumption similarly on a cold day user can use Heater , geezer kind off things. By inspecting the convolution, volatility, and variability of the users' electricity consumption behavior and weather data, this paper proposes a machine learning model using data analytics to analyze this behavior. A Big Data Analytics model for tracking and monitoring household electricity consumption based on External Factor like weather is developed. Using the big data analytics techniques, machine learning algorithms and descriptive analysis of data obtained from weather and consumers, the model provides the required information and prediction.**

Keywords : Electricity consumption , Weather forecast , BDA models , Data Analytic techniques, Machine learning

I. INTRODUCTION

In this era of digital world we can see that electricity has become an integral part of our day to day life. From small equipment like bulb, blender, toaster to large appliances like television, washing machine, refrigerator that we use at our home runs on electricity, similarly in each and every sector like Aerospace, transport, computer, telecommunication, construction, education, medical industry there are multiple appliances which can't be run without electricity. So from this perspective monitoring and consumption of electricity becomes more paramount. Nowadays we can observe that everyone is recklessly using electricity which can lead to scarcity of electricity in future as well as wrecking situation for upcoming generations. Apart from electrical appliance there are multiple other factors which affects electrical consumption in home as well as in various sectors. Weather is among one of the most important external factor. So it is very important for us to monitor electrical consumption based on weather so that we can save as much of the energy and contribute for the betterment of future generation. Thus using data analytics technique and machine learning method. We propose a model to analyze the users' electricity consumption behavior based on External Factors like weather.

II. LITERATURE SURVEY

Roimah Dollah, Hazleen Aris from [3] were motivated to predict the household electricity consumption and also monitor it. The electricity consumption data can be easily extracted from the smart meters, billing system and also power station. Big data analytics techniques have been used for data pre-processing, data exploration and also preparing data which is fit for model prediction using machine learning. They used previous electricity consumption dataset and performed descriptive and predictive analytics. Linear regression is the algorithm that they used to predict the electricity consumption. The main limitation of the paper is that there is no usage of external factors like weather dataset in electricity consumption which otherwise plays a major role in electricity consumption.
W. Zhang, X. Dong, H. Li, J. Xu and D. Wang from [1] were motivated on finding the abnormal electricity consumption behaviour with the help of feature engineering and unsupervised learning. The main focus of the paper was on feature selection and not on algorithm selection.Using unsupervised learning they were able to detect if there is any electricity theft by finding anomaly in the data. Deep Learning algorithm was used to find the feature. Dimension were reduced with the help of factor analysis.The dataset which was used was of the industrial park.The paper is not only useful in finding the abnormality in electricity consumption behaviour but also helps in extracting the best feature on which the consumption prediction can be performed. The paper has a disadvantage of lack of imformation on which feature selection strategy to be followed for acurate consumption prediction.
S. Shan, B. Cao and Z. Wu from [2] were motivated in building a model which would predict the electricity consumption more accurately and stability. They proposed a ensemble model and called it gravity gated recurrent unit electricity consumption model. They used two years electricity consumption dataset of a five-star hotel building in shangai and to test the generalization capacity of the model they used dataset of office building. The paper has 4 evaluation metrics for knowing the comparing the accuracy of model and extracting the best model name maximum MAPE, minimum MAPE, average MAPE, variance. The paper has a disadvantage of no external

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

factor being considered for predicting the electricity consumption.

Prachi Kulkarni,D.K Chitre from [4] were motivated in designing techniques to collect electrical consumption data from household. With the help of Internet of things they proposed a energy management system for smart home. Energy management system would contain data acquisition module from Internet of things and has a unique IP address which is connected to wireless network, useful for transferring data from home to cloud servers. The consumption data is collected by system on chip module from each device present in the household. The data collected can further be used for performing analysis and prediction. The paper had no algorithm proposed on how analysis can be performed on the data collected.

III. ADDITIONAL WORK

In our system other than normal electrical consumption we are also including that how external factors like weather and many other factors affect the electrical consumption at home. The Algorithm and approach we will use can vary based upon the accuracy rate. We will use data mining features like cleaning pre-processing clustering, Descriptive and predictive algorithm and also Machine learning concepts.

The various stage of modeling will include:

a) Data Collection and Extraction

b) Data Transformation

c)Data Analytics and Machine learning Algorithms

The python libraries like Matplotlib, Numpy and pandas will be used.

IV. IMPLEMENTATION

In this model we will use data analytics techniques as well as machine learning algorithms which will sum up to 5 layers. The layers are Data source, data exploration, data pre-processing, data visualization and model implementation.

A. Data source

This layer is very first stage of our model where we look for electrical consumption data and weather data from various available resources. This comprises of datasets like how much electricity a particular room uses in 10 minutes duration in addition to the temperature, humidity and various other factors. The consumption data will come various appliances from different rooms like washing machine from bathroom, refrigerator from kitchen, laptop from living room, air-conditioner in bed room.

The data acquired at this layer will be used as an input for other layers. This data will first go through various data analytics techniques and then we will use processed data into machine learning model for training and testing. So that we can get consumption prediction as an output.

```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   date         19735 non-null  object
 1   Appliances   19735 non-null  int64
 2   lights       19735 non-null  int64
 3   T1           19735 non-null  float64
 4   RH_1         19735 non-null  float64
 5   T2           19735 non-null  float64
 6   RH_2         19735 non-null  float64
 7   T3           19735 non-null  float64
 8   RH_3         19735 non-null  float64
 9   T4           19735 non-null  float64
 10  RH_4         19735 non-null  float64
 11  T5           19735 non-null  float64
 12  RH_5         19735 non-null  float64
 13  T6           19735 non-null  float64
 14  RH_6         19735 non-null  float64
 15  T7           19735 non-null  float64
 16  RH_7         19735 non-null  float64
 17  T8           19735 non-null  float64
 18  RH_8         19735 non-null  float64
 19  T9           19735 non-null  float64
 20  RH_9         19735 non-null  float64
 21  T_out        19735 non-null  float64
 22  Press_mm_hg  19735 non-null  float64
 23  RH_out       19735 non-null  float64
 24  Windspeed    19735 non-null  float64
 25  Visibility   19735 non-null  float64
 26  Tdewpoint    19735 non-null  float64
 27  rv1          19735 non-null  float64
 28  rv2          19735 non-null  float64
dtypes: float64(26), int64(2), object(1)
memory usage: 4.4+ MB
```

Figure 1:Data variables

In figure 1 we have 29 columns and 4 rows with each column consisting of 19735 values.

B. Data exploration

In this layer of our model we will explore the data in first stage, we will use python libraries like pandas, numpy for importing various function which can help us exploring data in a much better way. First we will rename the column name of our data set according to our convenience for the model.

	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000
mean	97.694958	3.801875	21.686571	40.259739	20.341219	40.420420	22.267611	39.242500	20.855335	39.026904
std	102.524891	7.935988	1.606066	3.979299	2.192974	4.069813	2.006111	3.254576	2.042884	4.341321
min	10.000000	0.000000	16.790000	27.023333	16.100000	20.463333	17.200000	28.766667	15.100000	27.660000
25%	50.000000	0.000000	20.760000	37.333333	18.790000	37.900000	20.790000	36.900000	19.530000	35.530000
50%	60.000000	0.000000	21.600000	39.656667	20.000000	40.500000	22.100000	38.530000	20.666667	38.400000
75%	100.000000	0.000000	22.600000	43.066667	21.500000	43.260000	23.290000	41.760000	22.100000	42.156667
max	1080.000000	70.000000	26.260000	63.360000	29.856667	56.026667	29.236000	50.163333	26.200000	51.090000

8 rows × 28 columns

Figure 2:Dataset description

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

Figure 2 depicts the count, mean, standard deviation, minimum and maximum value of dataset.

We will use describe method to get more insight details of our dataset.

```
date                    0
Tdewpoint               0
Visibility              0
Windspeed               0
RH_out                  0
Press_mm_hg             0
T_out                   0
ParentRoom_humidity     0
ParentRoom_temp         0
TeenRoom_humidity       0
TeenRoom_temp           0
Ironing_humidity        0
Ironing_temp            0
rv1                     0
outside_humidity        0
Bathroom_humidity       0
Bathroom_temp           0
office_humidity         0
office_temp             0
Laundry_humidity        0
Laundry_temp            0
Livroom_humidity        0
Livroom_temp            0
Kitchen_humidity        0
kitchen_temp            0
lights                  0
Appliances              0
outside_temp            0
rv2                     0
dtype: int64
```

Figure 3:Null value

Figure 3 describes null value in each column.

Since this raw dataset may consist of null values, so we will use isnull method to check how many rows contain null values. Now we will rearrange the data columns like we will merge all temperature into single array, similarly for humidity we will do the same.

Lowest temperature inside the house is 14.89 Deg & highest temperature inside the house is 29.85 Deg, outside the house the lowest temperature is

-6.06 Deg and the highest being 28.29 Deg. The lowest humidity inside the house is 20.60% and the highest humidity inside the house is 63.36%. 3/4th of Appliance consume less than 100Watt.

Now we will separate dependent variables and independent variables for training and testing.

C. Data preprocessing

In this layer of our model we will remove null values and divide the dataset into dependent variables and independent variable. The dependent variables are the variables which rely gets affected by independent variable where as independent variable are not affected by any factors. In our model appliance is a dependent variable whereas others are independent variables.

This process is then followed by dividing the dataset into training and testing dataset which is required for machine learning model. Our model will use 75% of dataset as training dataset and 25% as testing dataset as. As depicted in figure 4.

From, data exploration we were able to notice that light variable consisted of around 14000 zero values, so we decided to drop the light column from our dataset for the betterment of our prediction accuracy.

The data that we have contain large difference like an appliance consumption vary from 1 to 1000 and other variables are varying from 1 to 100. So if we directly feed this data to our machine learning model. The model will become biased and it won't predict correct consumption, so to remove this anomaly we will normalize of each variable into a common range. For this we will use standard scaler from sklearn library.

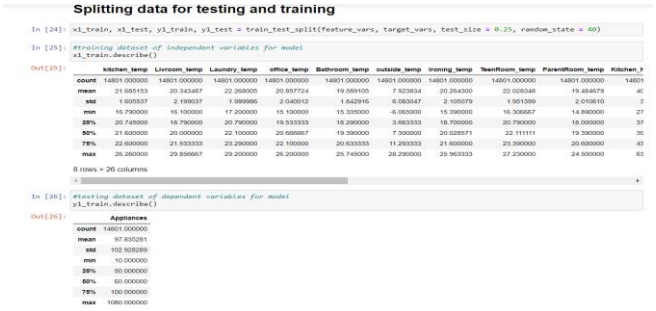

Figure 4:Testing and training dataset

```
array([[ 0.56982678, -0.38358411, -0.38401803, ...,  1.92227254,
         1.25047328,  1.25047328],
       [ 1.25497881,  0.2985558 ,  1.56106102, ..., -1.10404466,
         1.46879046,  1.46879046],
       [ 0.32068059,  0.0545013 ,  0.36601245, ...,  0.14042222,
         0.04873098,  0.04873098],
       ...,
       [-0.0530387 ,  0.84426774,  0.21600636, ..., -0.9484863 ,
        -0.17320537, -0.17320537],
       [ 0.92797443,  1.14289343,  0.51101835, ..., -0.01513614,
         0.50539964,  0.50539964],
       [-0.0530387 , -0.74739207,  0.11600229, ...,  0.81922234,
        -0.04022971, -0.04022971]])
```

Figure 5:Normalization

Figure 5 shows the normalization of dataset.

D. Data visualization

Datasets available are in csv format, which have easily imported and processed using python libraries using Jupyter IDE. Built-in features like Scikit learn, numpy & pandas have been used for the whole process and pyplot, matplotlib is used for visualization. Initial exploratory analysis is done to understand which features are most impactful and critical to the research. The most differentiating feature here is temperature and humidity.

Figure 6 shows the distribution of variables, it is used to detect the anomaly present in the dataset, from figure 6 it is evident that variable light consist of numerous zero values. All the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

temperature variable follow normal distribution except parent room temperature. All the humidity variables follow normal distribution except outside humidity as it has many zero values. The two random variables also contain numerous zero values. Data visualization helps in making the model more accurate in predicting the consumption.

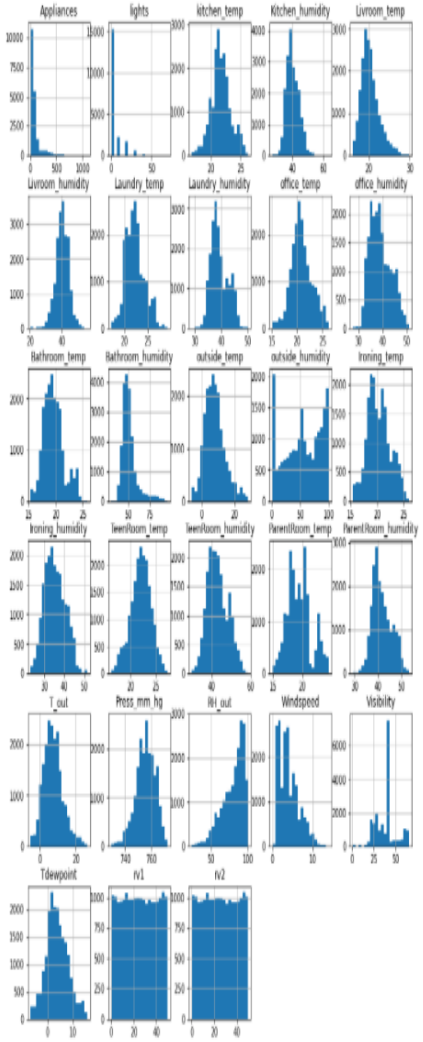

Figure 6:Distribution of variables

E. Model implementation

In this layer we will use processed data to feed our various machine learning model and predict electrical consumption based on that. We will implement following machine learning model.

a)Linear regression
b)Support vector machines
c)k-nearest neighbors
d)Random forest

a)**Linear regression**

It analyze specific relationship between two variables and gain information about a variable through values of other variable.

There are two types of variables dependent and independent variables.

Independent variables are the variables which are given as an input to the system. It can also be called as predictor.

Dependent variables are the one whose value gets affected by change in other values. They can also be called as response variable.

Linear regression widely focuses on relationship between dependent and independent variables.

First order linear model is given by

$$Y = b_0 + b_1 X + \in$$

Y – dependent variable
X – Independent variable
b_0 – Y intercept
b_1 – slope of the line
€ - Error variable

In linear regression we have cost equation which gives us information about how much difference is there between values predicted by model and actual values. Thus we use gradient descent to loop through our data and find the lowest cost error, which gives us most accurate linear regression model.

b)**Support vector machines**

Support vector machine is a powerful supervised machine learning algorithm that works on linearly and non-linearly separable data it finds an optimal hyperplane that best separates our data, so that the distance from a nearest point in a space to itself is maximized.

Hyperplane is a plane of (n-1) dimensions in n dimensional feature space, that separates the two classes, like for a 2-D feature space it will be a line and for 3-D feature space it will be a plane. Hyperplane is basically the decision boundary for our classifier, we will separately classify data based on which region of hyperplane the data falls.

h(x) = wT. x + b
w – weight vector
b – scalar bias

c) **k-nearest neighbors**

It is used for both classification and regression problems. It stores all available cases and classifies new cases based on similarity measures.

A case is classified by a majority vote of its neighbors. The Case is assigned to the class most common amongst its K nearest neighbors measured by a distance function. Classification becomes difficult when the number of dimension increases.

It provides high speed and better performance as it takes less time for model training.

d)**Random forest**

Random Forest is a popular algorithm which is a part of the supervised mastering algorithm, it can be used to solve both regression and classifications problems. Random forest is based on the idea of ensemble learning which happens to be a procedure of pairing several classifiers in order to solve a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

complicated problem and to boost the overall performance of the algorithm.

We are using root mean squared error to calculate the efficiency of our model.

RMSE = $\sqrt{\sum_{i=1}^{n}(P-O)^2/n}$

P-predicted value
o-observed value
n-number of observation

With the help of this formula the accuracy of model is calculated and it shows the error value of the model.

V. RESULTS DISCUSSION

After performing all this machine learning models, we get to know there train time and root mean squared error.

	ML Model	Train_Time	Test_RMSE_Score
0	linear regression	0.005983	0.915530
1	Support vector machine	5.701513	0.879290
2	K nearest neigbour	0.000997	0.790869
3	Random Forest	22.547622	0.642474

Figure 7:Comparision of models

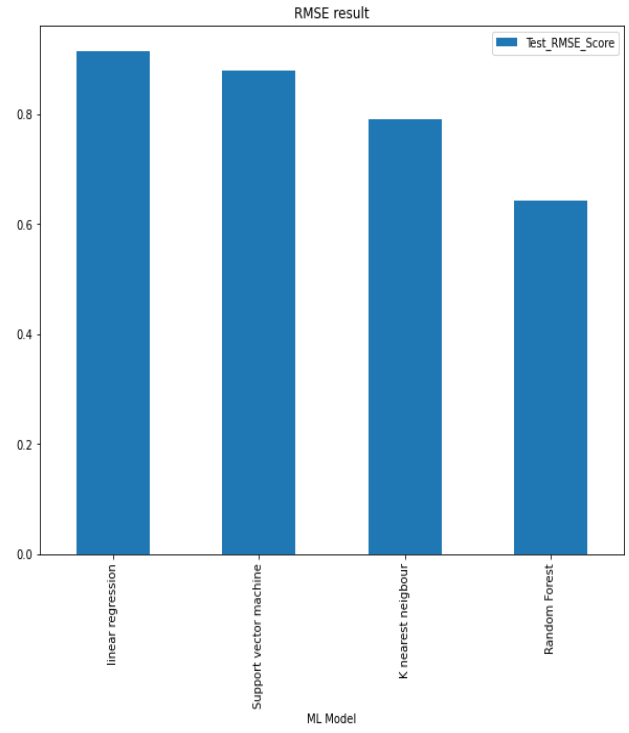

Figure 8:RMSE Comparision graph

Figure 8 shows the comparison of various model with respect to the root mean squared error obtained to predict the accuracy of each model. It is evident from the graph that linear regression is worst performing model among the four models whereas support vector machine is slightly better than linear regression, k nearest neighbor show drastic improvement in accuracy in comparison with above two. But the best performing model out of all four models is random forest model.

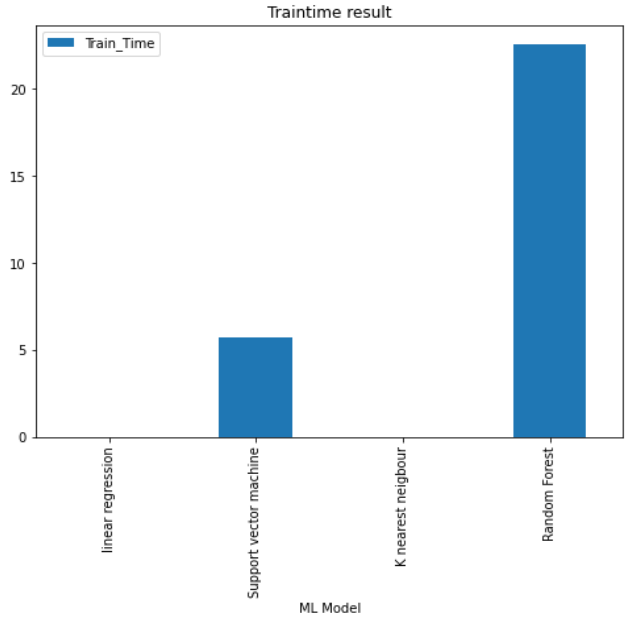

Figure 9:Training time comparison

Figure 9 shows the training time comparison among different models, it is evident from the graph that random forest model takes highest time to train whereas support vector machine takes second highest time to train, linear regression and k nearest neighbor takes very less time in comparison with other two with k nearest the lowest among all.
1)Least RMSE score is by random forest 0.63
2)Linear regression regularization was worst performing model.

VI. CONCLUSION

Existing system monitors electrical consumption without including any external factors. People will recklessly use electrical equipment and it will be difficult to manage consumption and save energy using this model. With the help of externals factor such as weather dataset we will be able to predict daily electrical consumption behavior in a much efficient and a correct way. Apart from linear regression we are using other training model which will predict outcome with less error percentage. Because we are using the pre-trained model in our project, which will reduce the cost of our model and also reduce the cost of daily electrical consumption. With the arrival of electrical cars in market it seems in future most probably there will be crisis of energy so next generation can efficiently use this model to somehow control the energy crisis. This model will help predict the electricity consumption with the help of external factors like weather data. Future works should include research on industrial electricity consumption and how to reduce the electricity consumption in order to reduce the pollution..

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

REFERENCES

[1] W. Zhang, X. Dong, H. Li, J. Xu and D. Wang, "Unsupervised Detection of Abnormal Electricity Consumption Behavior Based on Feature Engineering," in IEEE Access, vol. 8, pp. 55483-55500, 2020.

[2] S. Shan, B. Cao and Z. Wu, "Forecasting the Short-Term Electricity Consumption of Building Using a Novel Ensemble Model," in IEEE Access, vol. 7, pp. 88093-88106, 2019.

[3] Roimah Dollah,Hazleen Aris "A Big Data Analytics Model For Household Electricity Consumption Tracking and Monitoring" IEEE Conference on Big Data and Analytics (ICBDA) 21-22 November 2018

[4] Prachi Kulkarni,D.K Chitre "Enegy Consumption Using IOT and Big Data Analytics Approach in Smart Home" IJIRSET Vol.7,Issue10,October 2018

[5] A. K. Pandey,C.P Agrawal,Meena agrawal "A hadoop based weather prediction model for classification of weather data" 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT) 2017.

[6] P. Louridas,C. Ebert, "Machine Learning," in IEEE Software, vol. 33, no. 5, pp. 110-115, Sept.-Oct. 2016.

[7] H. Hu, Y. Wen, T. Chua and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," in IEEE Access, vol. 2, pp. 652-687, 2014.