Comparision Of Machine Learning Techniques To Predict Electricity Usage.

Analysis and comparision of different algorithms to predict electricity usage for a city, an industry and a house

Yogaraj Kori
x23241365
x23241365@student.ncirl.ie
National College of Ireland, Dublin

***Abstract*—UNESCO's Sustainable Development Goals speak about sustainable energy and reduction of energy consumption. The first step towards this process is to understand the consumption of energy and ultimately predict the energy consumption. For this purpose, we have considered three datasets addressing the electricity consumption of a city, an industry and a house. To predict the energy consumption from the datasets, we have utilized the performance power of two identically named machine learning techniques, Random Forest regression and Gradient Boosting Machine. We have applied these techniques on the three datasets in python and in R to understand how different implementations of each algorithm make an impact on the accuracy measures considered.**

Keywords—sustainable energy, energy prediction, identically named machine learning techniques, Random Forest Regression, Gradient Boosting Machine, accuracy measurement.

I. INTRODUCTION

The Sustainable Development Goals are the UNESCO's official goals to combat worlds problems and to achieve equal rights across the world. Their SDG 11 and SDG 12 talk about sustainable energy and development. To save and provide electricity for everyone we need to understand power consumption and be able to predict future requirements. The main issue regarding this prediction is the understanding of power consumption across different fields. An Industry manufacturing steel will consume more electricity than a house and to manage the power availability to everyone we will also need to understand power consumption between an industry and a house and be able to provide for both without any outages.

To understand this better we have gathered huge data from different sources and is explained below.

1. Electricity consumption for a household: A University of California Irvine dataset containing power usage by appliances and lights along with the weather data near the house to understand when and how these appliances are utilized the most.
2. Stell Industry Power consumption: A University of California Irvine dataset containing power consumption by the industries to make steel plates and a categorization checking if load type is heavy or less.
3. Power consumption of Tetouan city: A University of California Irvine dataset containing power consumption across different zones in the city.

For the purpose of prediction, we have used two different yet similar machine learning techniques which are Random Forest Regression and Gradient Boosting Machine Regression and we have applied these algorithms in both python and R to understand how the implementation of these algorithms vary in both environments. While both of these techniques are ensemble techniques which work with decision trees, they differ in how they approach the data in the training and testing phase.

Random Forest Regression is an ensemble technique which creates multiple decision trees and combines each trees' prediction into a final prediction. Since these trees are created and trained without considering the results of the other trees, it makes this technique less prone to overfitting. It also helps in generalization because of its randomness in feature selection and training.

Gradient Boosting Machine Regression is also an ensemble technique which builds decision trees but varies from RF by creating its trees sequentially and each tree correcting the mistakes made from the previous tree. This algorithm helps to find subtle patterns in the data. This algorithm is more sensitive to overfitting than RF when the tree count is high.

Generally GRBT provides higher accuracy scores when compared to RF, but when in certain scenarios when there is high dimensional data or if there are large numbers of outliers and noise in the data.

II. RESEARCH QUESTIONS

A. Research Question 1(RQ1)

Checking the implementations of Random forest and Gradient Boosting perform exactly the same in both python and R? If not, what are the best performing implementations of these techniques for specific empirical designs, and evaluation metrics? That is, if implementations in python and R for these techniques perform differently, which implementation is better.

B. Research Question 2(RQ2)

How do the datasets for electricity consumption of house, steel industry and city differ from each other? Specifically, how these datasets are different from each other in terms of their characteristics which can impact the effectiveness of RF and GRBT?

III. METHODOLOGY

One of the most important steps towards the development of a prediction model is the Exploratory data analysis. In this project we have applied EDA in python and any subsequent transformations on the data in python as well. We will be

reading the dataset using python pandas, since all of our datasets are in CSV format. Our main goal of the EDA is to explore the dataset and to understand its Characteristics, this will help us answer the questions in RQ2. The characteristics of the datasets are explained in the below tables.

A. Features

Features	Explanation
DateTime	Date and time of the measurement
Temperature	Temperature in the city
Humidity	Humidity in the city
General diffuse flows	Overall movement of electricity
Diffuse flows	Actual movement of electricity
Zone 1 power consumption	Electricity utilized by zone 1 – Quads
Zone 2 Power Consumption	Electricity utilized by zone 1 – Smir
Zone 3 Power Consumption	Electricity utilized by zone 1 – Boussafou

Table 1: Features of city dataset

Now let us look at the features of the dataset 2 which is the steel industry power consumption dataset.

Features	Explanation
Datetime	Date and time of the measurement
Usage_kWh	How much electricity was used.
CO2(tCO2)	Carbon di oxide in the machines which increase power consumption
NSM	Seconds from midnight
Weekday	Day of the week
Week_status	Weekday or non week day

Table 2: Features of steel industry dataset

Now let us look at the important features of our third dataset which is the electricity consumption of house.

Features	Explanation
Appliances	Power consumption of appliances turned on
Lights	Power consumptions of lights turned on
T1,T2…..T9	Temperature in various parts of house
RH_1,RH_2…RH_9	Humidity in various parts of house
T_out	Outside temperature
RH_out	Outside humidity
Windspeed	Speed of wind outside
Rv1	Random variable 1
RV2	Random variable 2

Table 3: Features of House dataset

B. Characteristics of data

Now that we have looked at the features in the datasets, let us look at the functions we have applied to look at the data and check the outputs for them in the below table.

Functions applied	City dataset	Industry dataset	House dataset
Shape	(52416, 9)	(35040, 11)	(19735, 29)
.isnull().sum()	0	0	0
.isna().sum()	0	0	0
Infinity values	0	0	0

Table 4: Basic Characteristics of dataset

Now that we know the shape of the datasets and checked if NULL,NA and INF values are present in the dataset. Another important thing to know about the characteristics of dataset is to check the mean, median and standard deviation of the dataset. For the purposes of the report, we have only written for the dependent variable.

Value name	City Dataset	Industry dataset	House dataset
count	52416.000000	35040.000000	19735.000000
mean	32344.970564	27.386892	97.694958
std	7130.562564	33.444380	102.524891
min	13895.696200	0.000000	10.000000
max	52204.395120	157.180000	1080.000000

Table 5: Basic Characteristics of dataset

C. Boruta Technique

Now that we know the basic characteristics of data, let us try to find which of the features are important. The Boruta algorithm is a feature selection method primarily used in machine learning for identifying relevant variables in a dataset. The output of Boruta technique gives us the importance of features with respect to the dependent variables. Figure below shows the graph of importance from the Boruta technique for the city dataset.

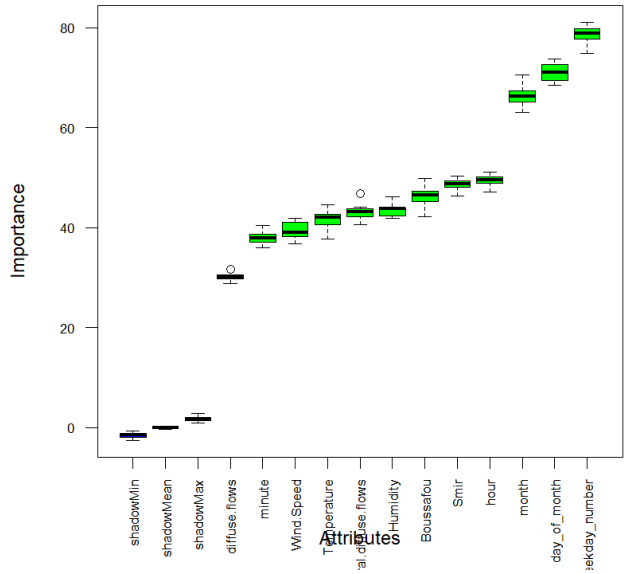


Fig 1: Boruta technique output for the city dataset.

Now let us look at the hierarchy of importance for the Industry dataset.

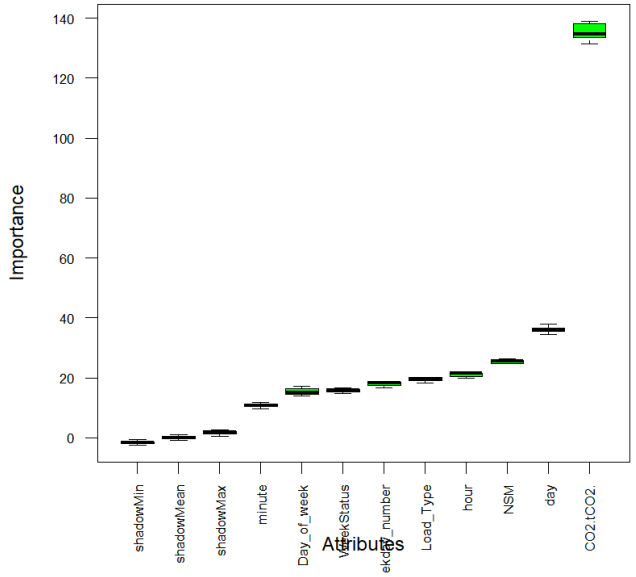


Fig 2: Boruta technique output for the Industry dataset.

Now let us look at the hierarchy of importance for the House dataset.

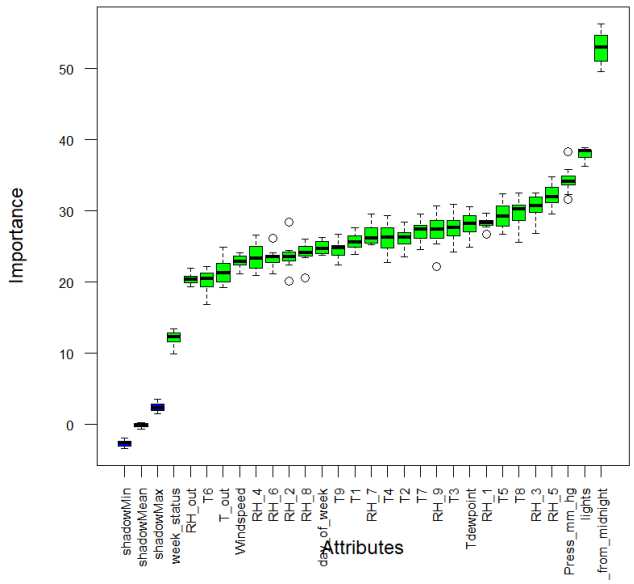


Fig 3: Boruta technique output for the House dataset.

Now that we know the importance of each feature with the dependent variable , we can work on deleting the less important values from the dataset.

D. Corelation

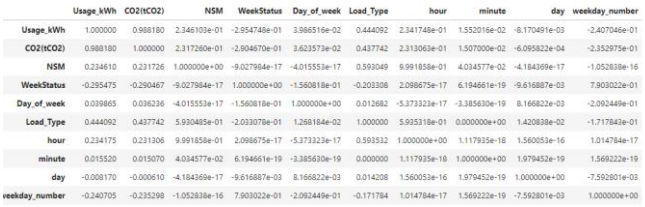
A correlation tells us how each variable is dependent on each other, if one of the highly corelated features then the other one also increases. The below figure shows the correlation for the city dataset.

	Temperature	Humidity	Wind Speed	general diffuse flows	diffuse flows	Quads	Smir	Boussafou	hour	minute	month	day_of_month
Temperature	1.000000	-0.460243	0.477109	0.460294	0.196522	0.440221	0.382428	0.489527	1.971305e-01	-6.724601e-05	2.843350e-01	1.926774e-02
Humidity	-0.460243	1.000000	-0.135853	-0.468138	-0.256886	-0.287421	-0.294961	-0.233022	-2.426916e-01	4.929864e-04	-1.741931e-02	-4.540307e-02
Wind Speed	0.477109	-0.135853	1.000000	0.133733	-0.000972	0.167444	0.146413	0.278641	4.148645e-03	-3.136897e-04	1.683554e-01	1.640667e-01
general diffuse flows	0.460294	-0.468138	0.133733	1.000000	0.564718	0.187965	0.157223	0.063376	1.299766e-01	-1.913001e-04	-2.055450e-02	3.225334e-02
diffuse flows	0.196522	-0.256886	-0.000972	0.564718	1.000000	0.080274	0.044667	-0.038506	1.309090e-01	-1.828407e-03	-1.297793e-01	-2.827775e-02
Quads	0.440221	-0.287421	0.167444	0.187965	0.080274	1.000000	0.834519	0.750733	7.279525e-01	3.426961e-04	-5.346049e-03	2.627309e-02
Smir	0.382428	-0.294961	0.146413	0.157223	0.044667	0.834519	1.000000	0.570932	6.633590e-01	3.307693e-05	3.202665e-01	5.294130e-02
Boussafou	0.489527	-0.233022	0.278641	0.063376	-0.038506	0.750733	0.570932	1.000000	4.542566e-01	3.886920e-05	-2.339260e-01	-4.782920e-03
hour	0.197130	-0.242692	0.004149	0.129977	0.130909	0.727953	0.663359	0.454257	1.000000e+00	1.912651e-17	-5.886319e-16	-1.024999e-16
minute	-0.000067	0.000493	-0.000314	-0.000191	-0.001828	0.000343	0.000033	0.000039	1.912651e-17	1.000000e+00	7.656029e-15	5.190033e-16
month	0.284335	-0.017419	0.168355	-0.020554	-0.129779	-0.005346	0.320266	-0.233926	-5.886319e-16	7.656029e-15	1.000000e+00	4.349864e-03
day_of_month	0.019268	-0.045403	0.164067	0.032253	-0.028278	0.026273	0.052941	-0.004783	-1.024999e-16	5.190033e-16	4.349864e-03	1.000000e+00
weekday_number	-0.013710	-0.017114	0.032116	0.009666	-0.024540	-0.069708	-0.122682	0.005294	-5.420622e-17	2.017312e-17	6.387829e-03	-1.941801e-02

Fig 4: Correlation for City dataset.

We can see from the above figure that while some features like temperature and hour play a crucial role with respect to our dependent variable Quad, we also have variables which are less corelated with our dependent variable.

Now let us look at the correlation of the industry dataset in the below figure.

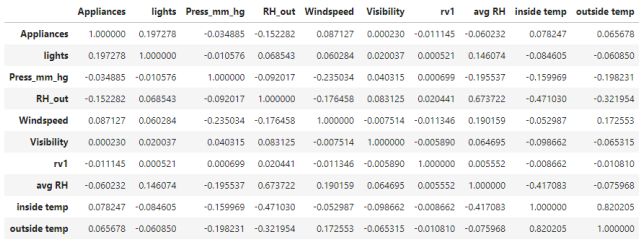


	Usage_kWh	CO2(tCO2)	NSM	WeekStatus	Day_of_week	Load_Type	hour	minute	day	weekday_number
Usage_kWh	1.000000	0.988180	2.346103e-01	-2.954748e-01	3.986516e-02	0.444092	2.341748e-01	1.552016e-02	-8.170491e-03	-2.407046e-01
CO2(tCO2)	0.988180	1.000000	2.317260e-01	-2.904670e-01	3.623573e-02	0.437742	2.313063e-01	1.507000e-02	-6.095822e-04	-2.352975e-01
NSM	0.234610	0.231726	1.000000e+00	-9.027984e-17	-4.015553e-17	0.593049	9.991858e-01	4.034577e-02	-4.184369e-17	-1.052838e-16
WeekStatus	-0.295475	-0.290467	-9.027984e-17	1.000000e+00	-1.560818e-01	-0.203308	2.098675e-17	6.194661e-19	-9.616887e-03	7.903022e-01
Day_of_week	0.039865	0.036236	-4.015553e-17	-1.560818e-01	1.000000e+00	0.012682	-5.373323e-17	-3.385630e-19	8.166822e-03	-2.092449e-01
Load_Type	0.444092	0.437742	5.930485e-01	-2.033078e-01	1.268184e-02	1.000000	5.935318e-01	0.000000e+00	1.420838e-02	-1.717843e-01
hour	0.234175	0.231306	9.991858e-01	2.098675e-17	-5.373223e-17	0.593532	1.000000e+00	1.117935e-18	1.560553e-16	1.014784e-17
minute	0.015520	0.015070	4.034577e-02	6.194661e-19	-3.385630e-19	0.000000	1.117935e-18	1.000000e+00	1.979452e-19	1.569222e-19
day	-0.008170	-0.000610	-4.184369e-17	-9.616887e-03	8.166822e-03	0.014208	1.560553e-16	1.979452e-19	1.000000e+00	-7.592801e-03
weekday_number	-0.240705	-0.235298	-1.052838e-16	7.903022e-01	-2.092449e-01	-0.171784	1.014784e-17	1.569222e-19	-7.592801e-03	1.000000e+00

Fig 5: Correlation for Industry dataset.

Our dependent variable is the Usage_kWh, features like CO2 is highly correlated with the dependent variable and features like day of the week and minutes are very less correlated.

Let us also look at the correlation of the house dataset in the below figure.



	Appliances	lights	Press_mm_hg	RH_out	Windspeed	Visibility	rv1	avg RH	inside temp	outside temp
Appliances	1.000000	0.197278	-0.034885	-0.152282	0.087127	0.000230	-0.011145	-0.060232	0.078247	0.065678
lights	0.197278	1.000000	-0.010576	0.068543	0.060284	0.020037	0.000521	0.146074	-0.084605	-0.060850
Press_mm_hg	-0.034885	-0.010576	1.000000	-0.092017	-0.235034	0.040315	0.000699	-0.195537	-0.159969	-0.198231
RH_out	-0.152282	0.068543	-0.092017	1.000000	-0.176458	0.083125	0.020441	0.673722	-0.471030	-0.321954
Windspeed	0.087127	0.060284	-0.235034	-0.176458	1.000000	-0.007514	-0.011346	0.190159	-0.052987	0.172553
Visibility	0.000230	0.020037	0.040315	0.083125	-0.007514	1.000000	-0.005890	0.064695	-0.098662	-0.065315
rv1	-0.011145	0.000521	0.000699	0.020441	-0.011346	-0.005890	1.000000	0.005552	-0.008662	-0.010810
avg RH	-0.060232	0.146074	-0.195537	0.673722	0.190159	0.064695	0.005552	1.000000	-0.417083	-0.075968
inside temp	0.078247	-0.084605	-0.159969	-0.471030	-0.052987	-0.098662	-0.008662	-0.417083	1.000000	0.820205
outside temp	0.065678	-0.060850	-0.198231	-0.321954	0.172553	-0.065315	-0.010810	-0.075968	0.820205	1.000000

Fig 6: Correlation for House dataset.

Our dependent variable is the Appliances, we can see from the figure that none of the features are very much correlated with the dependent variable. This might play a huge role in the accuracy of our techniques and impact the implementation of the Machine learning techniques.

From all the techniques and information gathered so far, we can clearly answer the Research Question 2. Let us generalize the above steps into a table.

Attribute	City	Industry	House
Shape row count	High	Medium	low
Shape feature	Medium	Medium	High
Max-min	Low	Low	High
Std	Low	Low	High
Correlations with other features	High	High	low
Missing/NULL values	None	None	None

Table 6: Characteristics comparison for datasets

From the above table , we can guess if the output accuracy of our datasets will be good or bad. This also sheds a light on our House dataset that we might get lower accuracy score for the dataset for regression.

Now let us try to apply our machine learning techniques RF and GBRT on the filtered dataset.

IV. IMPLEMENTATION

We will be taking one variable as a dependent variable and the rest of the dataset as independent variables. We will be splitting the data into training and testing data with a ratio of 75/25 where 75% of the data is our training data and will be used to train our model, while the rest 25% is our testing data which will be used to test the predictions from the model. We will be implementing RF algorithm for all three datasets in both python and R for each of the datasets.

There are multiple accuracy scores which can be used to predict the accuracy of models. Since we are using regression model we will be using the following 4 values as our accuracy metrics.

1. R2 score: a metric used to evaluate the performance of regression models, It measures the proportion of the variance in the dependent variable (the target) that is predictable from the independent variables (the features).
2. MAE: Mean Absolute Error, another metric used to evaluate the performance of regression models. It measures the average absolute difference between the predicted values and the actual values.
3. MSE: Mean Squared Error, another commonly used metric for evaluating regression models. It measures the average of the squares of the errors or deviations—specifically, the average squared difference between the predicted values and the actual values.
4. RMSE: Root Mean Squared Error, and it's derived from the Mean Squared Error (MSE). It provides a measure of the average magnitude of the errors in the predictions.

We have implemented 2 algorithms in python and in R, the below table summarizes the results for R2 score for both.

Let us summarize the R2 scores in a table below.

Dataset	Python		R	
	RF	GRBT	RF	GRBT
City	0.9921	0.9745	0.9906	0.9463
Industry	0.9862	0.9862	0.9844	0.9766
House	0.4904	0.2738	0.4398	0.1669

Table 7: R2 value comparison

We have also summarised RMSE scores for both in the below table.

Dataset	Python		R	
	RF	GRBT	RF	GRBT
City	630.0426	1134.6688	687.8236	1647.766
Industry	3.9557	3.9664	4.1785	5.1215
House	71.4105	85.2498	78.1402	95.2856

Table 8: RMSE values comparison

The last summarization is for the MAE scores for the 2 different implementations for RF and GRBT.

Dataset	Python		R	
	RF	GRBT	RF	GRBT
City	411.6938	786.6203	451.1755	1177.646
Industry	2.3542	2.4395	2.3609	2.5747
House	36.377	47.9879	39.7044	53.6125

Table 9: MAE value comparison

V. RESULTS

Let us try to answer the Research questions mentioned above. We will be splitting the research questions into multiple questions for better understanding.

A. *RQ1*

1. CHECKING THE IMPLEMENTATIONS OF RANDOM FOREST AND GRADIENT BOOSTING PERFORM EXACTLY THE SAME IN BOTH PYTHON AND R? :

   Although the implementations were similar in both python and R, we see a slight increase in the accuracy metrics R2 for both RF and GRBT in python, similarly we see a lesser values of MAE and RMSE scores in python implementations.
2. What are the best performing implementations of these techniques for specific empirical designs, and evaluation metrics:

   We see the performance of python is better than R in both of our implementations for both RF and GRBT.

B. *RQ2*

As we can see from Table 6, there are a couple of features that affect the accuracy of our prediction models.

1. Presence of High number of features with low relatability with the independent variable.
2. A low count of records can also impact accuracy of prediction models.
3. A huge difference between the maximum values and the minimum values of a dependent variable.
4. A high standard deviation in the data.
5. Low correlation with all or most of the features.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.