Steel Industry Energy Consumption Dataset

Gundu Lalithendranath

11904466

KM018

Abstract

The fast development of urban advancement in the past decades has seen a lot of difficulties involving heavy usage of energy, infrastructure and many more. To overcome these concerns in the future decades it requires reasonable and realistic solutions. This project presents and explores predictive energy consumption models based on data exploration and Machine Learning techniques for a smart small-scale steel industry in South Korea. Energy consumption data is collected using IoT based systems and used for forecasting. The main goal of this project is to predict how much energy is used when required features are given. So, that a fair idea is created on the patterns of energy consumption in future. In general Smart Industries have various and enormous energy requirements for managing all the components in a smart Industries and cities which present new advances in an organized manner by exploiting all these energy sources, in an ideal way.

Literature Review:

1. In research paper of Sathishkumar V E, I can summarize that he is taking this data to find a patten in energy consumption overcome real world urbanization advancement problem that he has seen in past decades and this is useful for construction of smart Industries and cities where energy is used in efficient ways. Used models prediction:General linear regression,Classification and regression trees, Support vector machine with a radial basis kernel,K nearest neighbours, CUBIST. Root mean squared error, Mean absolute error and Coefficient of variation are used to measure the prediction efficiency of the models.

2. In research paper of Sathishkumar V E, Myeong-Bae Lee, Jong-Hyun Lim, Chang-Sun Shin, Chang-Woo Park and other. The theme is for Predictions of Energy Consumption for Industries gain an important place in energy management and control system,as there are dynamic and seasonal changes in the demand and supply of energy. The models are same as above mentioned.

3. In research paper of Andrea Maria N. C. Ribeiro and other paper, I can draw this matter that he has worked on to minimize the environmental impact to avoid regulatory penalties, and to improve competitiveness, energy-intensive manufacturing firms require accurate forecasts of their energy consumption so that precautionary and mitigation measures can be taken. Models used Autoregressive Integrated Moving Average (ARIMA), simple Recurrent Neural Networks (RNN),Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), Support Vector Regression (SVR) and Random Forest.

My unique methods:

I have used Multiple linear Regression, Polynomial Regression and Decision Tress Regressor. For Ensembled model I have used Stacking Regressor. These models are not evaluated by any other persons.

1 Introduction

This report deals with various machine learning regression models to predict the energy consumption based on features of consumption. South Korea's production industry has begun to evolve at an increasing pace since the 1990s and has become the primary pushing power behind South Korea's continued fast economic development. In the 1990s, primary power usage grew at an yearly pace.

With the growing issue of coal and oil shortages, the energy resource-related forecast and planning issues have drawn considerable problem solving from both the research and industrial practice perspective. To make reasonable use of by-product and non-renewable gasses in the steel industry, scheduling operators must be aware of the quantities of generation, consumption and storage overtime in Realtime. Thus, the precise forecasting of these energy flow units will provide a fair idea to their planning and distribution. The iron and steel industries are always energy-intensive energy consummators. Recently, with the rising energy resource shortage, the energy supply condition in the Industrial sector has become highly challenging. Developing an energy-saving policy has become an increasingly prominent job that can be achieved in respects such as technical advancement & refurbishment.

High energy usage will possibly lead to elevated costs for Iron & Steel products and will result in more pollution and emissions. To this purpose, certain steps, such as optimizing the manufacturing structure and encouragement of techniques to save energy and reduce emissions, are required to guarantee the efficient supply of energy in the manufacturing industry in South Korea.

One of the realistic solutions is to analyze the pattern of energy consumption like at what time there is high usage with help of technologies. So, the industrial expects can formulate solution according like replacing certain parts or machinery in industry for energy efficiency and to reduce cost as well in long run of the factories.

1.1 Objectives

The objectives of the methodology are discussed as follows:

- To predict the usage of energy based on given attributes.
- To execute the system using 3 main regression models and an ensembled model.
- To evaluate the results using Correlation, RSquare, Accuracy, RMSE, MSE and MAE.
- To cross validate all the results using validation approach.

2 Dataset Description

This dataset(from UCI Machine Learning Repository) contains the usage of energy of a factory named DAEWOO Steel Co. Ltd in Gwangyang, South Korea. This dataset is collected in the year 2018 monthly wise consumption of the factory based on load type and week status with a total of 35040 records and 11 columns(broadly).

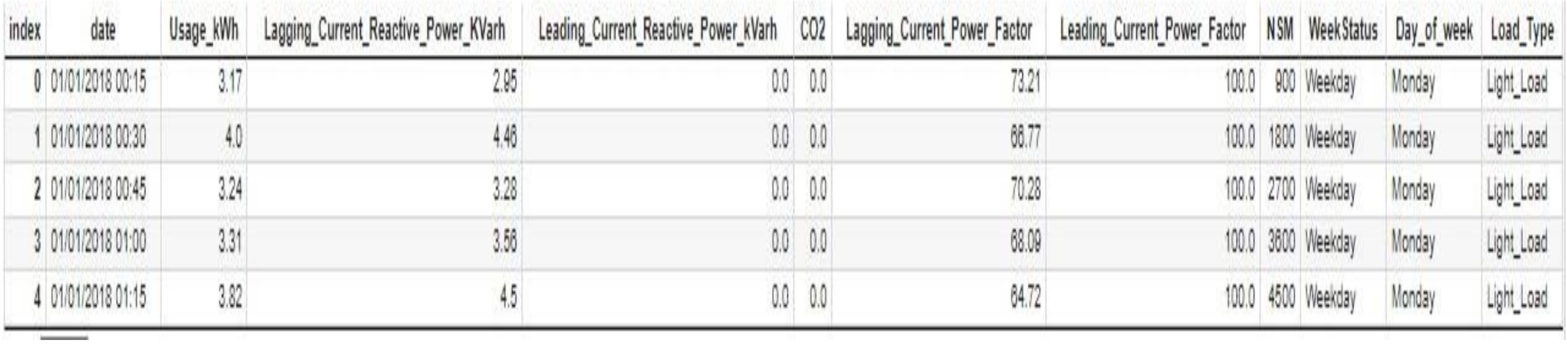
The dataset consists of number of attributes those date, Usage_kWh, CO2, Lagging_Current_Reactive_Power_KVarh, Leading_Current_Reactive_Power_kVarh,

Lagging_Current_Power_Factor, Leading_Current_Power_Factor, NSM (Number of seconds from Midnight), categorical variables: WeekStatus: {weekday, weekend}, Day_of_week: {Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday, Sunday}, Load_Type: {Light, Medium, Maximum}.

The main objective of the dataset is to predict the energy consumption (Usage) of this factory. Features used in this methodology are shown in table 1.

Table 1: Description of the features used

Features	Description
Date	Continuous-time data taken on the first of the month.
Usage_kWh	Industry Energy Consumption Continuous kWh (Units).
Lagging_Current_reactive_power	Load current lags behind the supply voltage Continuous kVarh (Units).
Leading_Current_reactive_power	Load current leads the supply voltage Continuous kVarh (Units).
CO2	Continuous ppm (parts per million).
Lagging_Current_Power_Factor	load current is capacitive in nature
Leading_Current_Power_Factor	load current is inductive in nature
NSM	Number of Seconds from midnight Continuous Seconds (Units).
Week status	Categorical (Weekend (0) or a Weekday (1)).
Day of week	Categorical Monday, Tuesday…. Sunday.
Load Type	Categorical Light Load, Medium Load, Maximum Load.



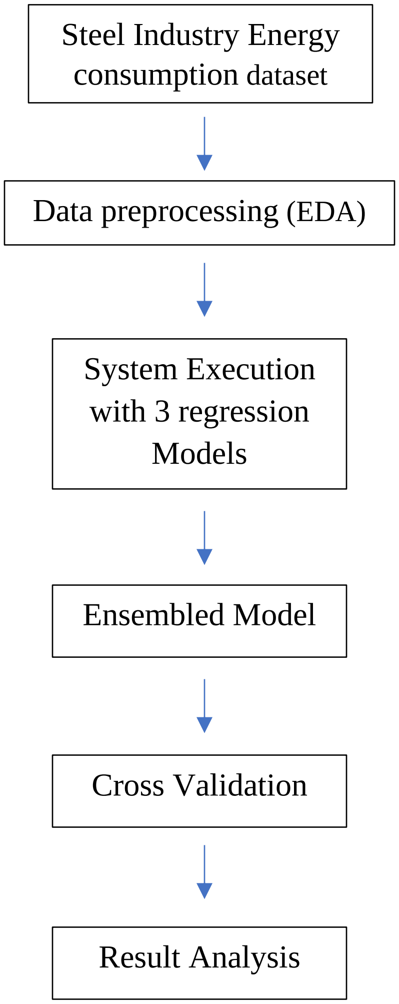
index	date	Usage_kWh	Lagging_Current_Reactive_Power_KVarh	Leading_Current_Reactive_Power_kVarh	CO2	Lagging_Current_Power_Factor	Leading_Current_Power_Factor	NSM	WeekStatus	Day_of_week	Load_Type
0	01/01/2018 00:15	3.17	2.95	0.0	0.0	73.21	100.0	900	Weekday	Monday	Light_Load
1	01/01/2018 00:30	4.0	4.46	0.0	0.0	66.77	100.0	1800	Weekday	Monday	Light_Load
2	01/01/2018 00:45	3.24	3.28	0.0	0.0	70.28	100.0	2700	Weekday	Monday	Light_Load
3	01/01/2018 01:00	3.31	3.56	0.0	0.0	68.09	100.0	3600	Weekday	Monday	Light_Load
4	01/01/2018 01:15	3.82	4.5	0.0	0.0	64.72	100.0	4500	Weekday	Monday	Light_Load

Figure 1: Sample Dataset

3 Approach

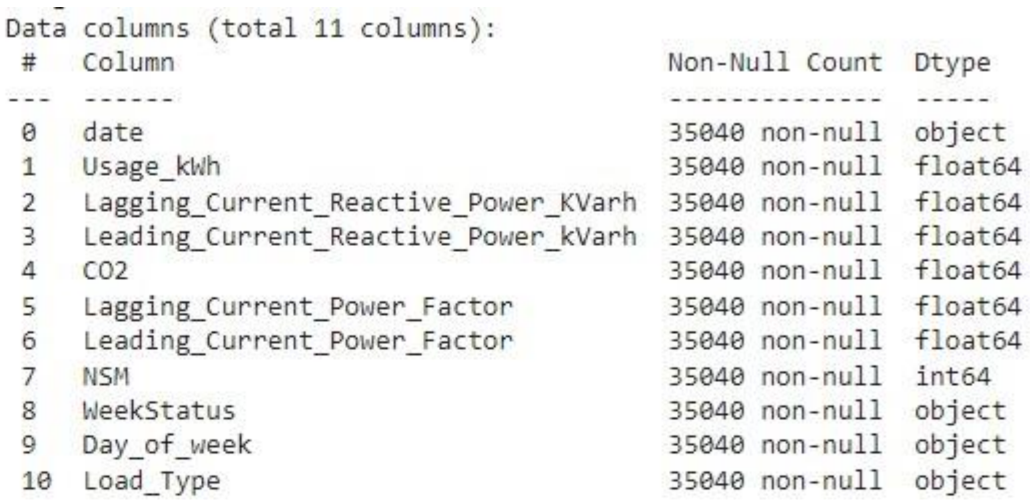
In this approach all three regression models and an ensembled model of machine learning are applied on the training dataset to predict the results. 70% of the data from the dataset is used to train the system and results are predicted by using 20% of the test data and 10% of the validation data. Features are selected from the dataset to improve the results. after executing three models an ensembled model is created for the best performance. Ensembling enhance the results of the resultant prediction. Cross validation is then taken into consideration. K-fold validation is a category of cross validation which measures the robustness of the model. Graphically the approach is shown in figure 2.

Figure 2: Methodology Used


Steel Industry Energy consumption dataset
Data preprocessing (EDA)
System Execution with 3 regression Models
Ensembled Model
Cross Validation
Result Analysis

4 Feature Selection

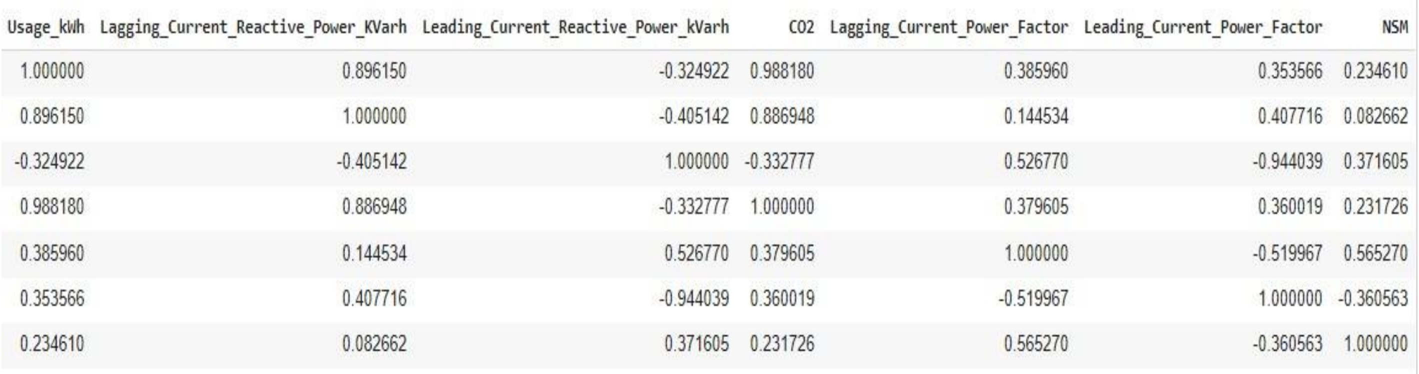
The main objective of feature selection is to find out the most reliable features, as they act as an important factor in the whole prediction process. In my dataset 2 factors are playing a prominent role but we I have trained and test with only 2 features the result was less as compared to when all the features are taken in as all the features here are contributing towards the accuracy of the model so I here conclude that these following features in figure 3 are considered for Models.



```
Data columns (total 11 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   date                                  35040 non-null  object
 1   Usage_kWh                             35040 non-null  float64
 2   Lagging_Current_Reactive_Power_KVarh  35040 non-null  float64
 3   Leading_Current_Reactive_Power_kVarh  35040 non-null  float64
 4   CO2                                   35040 non-null  float64
 5   Lagging_Current_Power_Factor          35040 non-null  float64
 6   Leading_Current_Power_Factor          35040 non-null  float64
 7   NSM                                   35040 non-null  int64
 8   WeekStatus                            35040 non-null  object
 9   Day_of_week                           35040 non-null  object
 10  Load_Type                             35040 non-null  object
```

Figure 3: Selected Features

Correlation between the variables is also one of the factors involved for feature selection. Correlation matrix is shown in Figure 4.



Usage_kWh	Lagging_Current_Reactive_Power_KVarh	Leading_Current_Reactive_Power_kVarh	CO2	Lagging_Current_Power_Factor	Leading_Current_Power_Factor	NSM
1.000000	0.896150	-0.324922	0.988180	0.385960	0.353566	0.234610
0.896150	1.000000	-0.405142	0.886948	0.144534	0.407716	0.082662
-0.324922	-0.405142	1.000000	-0.332777	0.526770	-0.944039	0.371605
0.988180	0.886948	-0.332777	1.000000	0.379605	0.360019	0.231726
0.385960	0.144534	0.526770	0.379605	1.000000	-0.519967	0.565270
0.353566	0.407716	-0.944039	0.360019	-0.519967	1.000000	-0.360563
0.234610	0.082662	0.371605	0.231726	0.565270	-0.360563	1.000000

Figure 4: Correlation between all features

5 Machine Learning models used

There are 3 Machine Learning Models used and they are shown in table 2

Table 2: Machine Learning Models

Model	Function	Package	Tuning parameter
Multiple Linear Regression	LinearRegression()	sklearn.linear_model	none
Polynomial Regression	PolynomialFeatures(), LinearRegression()	sklearn.linear_model, sklearn.preprocessing	degree
DecisionTree Regressor	DecisionTreeRegressor()	sklearn.tree	Max_depth
Stacking Regressor	StackingRegressor()	sklearn.ensemble	estimators
RandomForest Regressor	RandomForestRegressor()	sklearn.ensemble	n_estimators

6 Model evaluation

Description of Model Evaluation Parameters:

Correlation(r): The correlation coefficient is a measure of linear association between the predicted numeric target value and the actual numeric value. Value of the correlation coefficient always lie between -1 and +1. A correlation coefficient of +1 means that two variables are perfectly related in a positive linear manner, a correlation coefficient of -1 means that two variables are perfectly related in a negative linear manner, and a correlation coefficient of 0 means that there is no linear relationship present between the two variables. The correlation between two x and y variables are calculated by using equation.

$$Corr(r) = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma\,(x - \bar{x})^2\,(y - \bar{y})^2}}$$

R Square(R^2): The Square of the Correlation(r). This value can be interpreted as the proportion of the information in the data that is explained by the model.

$$R^2 = Correlation^2$$

RMSE: The Root Mean Square Error (RMSE) metric is defined as a distance measure between the predicted value and the actual value. The smaller the value of the RMSE, the better is the predictive accuracy of the model. RMSE value 0 means a model has perfect and correct predictions. RMSE is calculated by using equation.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

Where $\hat{x}_i$ ·is predicted and x_i is actual values.

MSE: The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. For every data point, you take the distance vertically from the point to the corresponding y value on the curve fit (the error), and square the value.

$$MSE = \frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}$$

Where $\hat{x}_i$ is predicted and x_i is actual values.

MAE: The Mean Absolute Error (MAE) refers to the magnitude of difference between the prediction of an observation and the true value of that observation. It takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

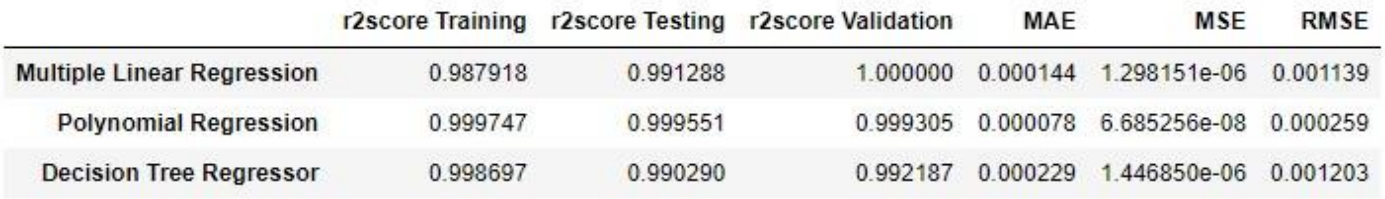
$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\hat{x}_i - x_i|}{n}$$

Where $\hat{x}_i$ is predicted and x_i is actual values.

7 Result Analysis

7.1 Performance comparison

Performance comparison of all the models is shown in figure 5. It has been found that all the three models have performed well with some parameter tuning done on them.

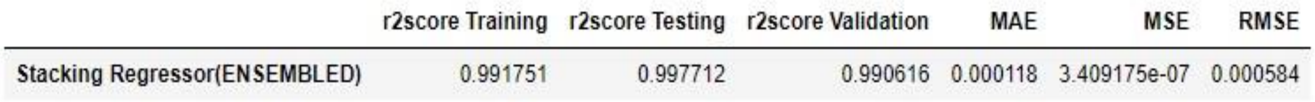


	r2score Training	r2score Testing	r2score Validation	MAE	MSE	RMSE
Multiple Linear Regression	0.987918	0.991288	1.000000	0.000144	1.298151e-06	0.001139
Polynomial Regression	0.999747	0.999551	0.999305	0.000078	6.685256e-08	0.000259
Decision Tree Regressor	0.998697	0.990290	0.992187	0.000229	1.446850e-06	0.001203

Figure 5: Evaluation of 3 models

7.2 Ensemble Model

Ensemble learning involves combining multiple predictions derived by different techniques in order to create a stronger overall prediction. The Stacking Regressor used estimators for all the base modes to be defined in it and also uses a final estimator as I have taken to be random Forest Regressor. Stacking often consider heterogeneous learner, learns then in parallel and combines them by training a meta model to output a prediction based on base learners output as input to meta-model. while final_estimator is trained using cross-validated predictions of the base estimators using cross_val_predict. The function cross_val_predict has a similar interface to cross_val_score but returns, for each element in the input, the prediction that was obtained for that element when it was in inputs of base model. The value of evaluation parameters after ensembling are given in figure 6.



	r2score Training	r2score Testing	r2score Validation	MAE	MSE	RMSE
Stacking Regressor(ENSEMBLED)	0.991751	0.997712	0.990616	0.000118	3.409175e-07	0.000584

Figure 6: Evaluation of ensembled

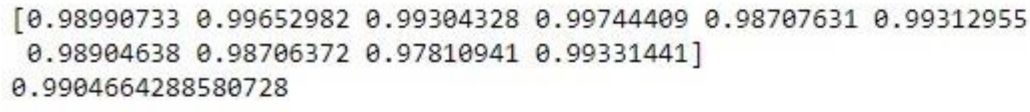
7.3 Cross Validation

Cross validation technique is used to validate the predictive models and analyze statistical results. It estimates how accurately any predictive model will perform. In this technique the original sample is partitioned into a training set to train the model, and a test set which is used for system evaluation. In this methodology validation approach of cross validation is used, in which data get shuffled on random basis.

MLR has no turning parameters, PR has degree 2 is perfect for it as when changed to 3 it is overfitting the data, DTR has max_dept = 7 giving perfect quality split for rest values the accuracy is no high same with RFR as well for n_estimators = 23 giving high accuracy and for rest it is not accurate.

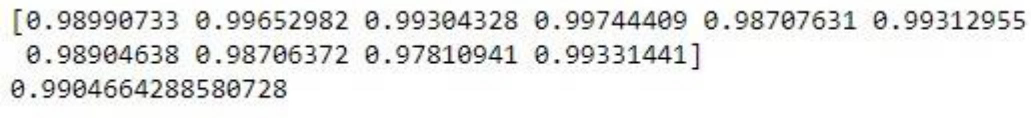
I have done cross validation on test only for all models

Linear Regression CV:



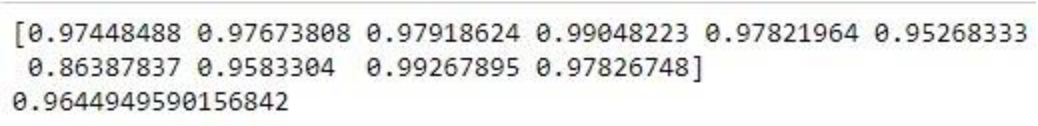
```
[0.98990733 0.99652982 0.99304328 0.99744409 0.98707631 0.99312955
 0.98904638 0.98706372 0.97810941 0.99331441]
0.9904664288580728
```

Polynomial Regression CV:



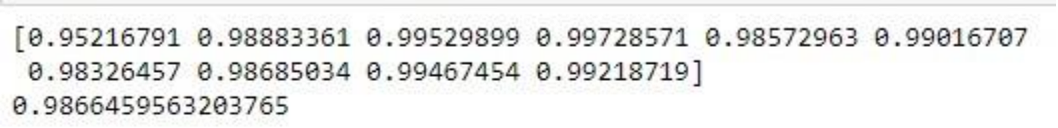
```
[0.98990733 0.99652982 0.99304328 0.99744409 0.98707631 0.99312955
 0.98904638 0.98706372 0.97810941 0.99331441]
0.9904664288580728
```

Decision Tree Regressor CV:



```
[0.97448488 0.97673808 0.97918624 0.99048223 0.97821964 0.95268333
 0.86387837 0.9583304  0.99267895 0.97826748]
0.9644949590156842
```

Stacking Regressor CV:



```
[0.95216791 0.98883361 0.99529899 0.99728571 0.98572963 0.99016707
 0.98326457 0.98685034 0.99467454 0.99218719]
0.9866459563203765
```

7.4 Scatter Plot

Accuracy plot after validation approach is shown in following figures.

Multiple Linear Regression:

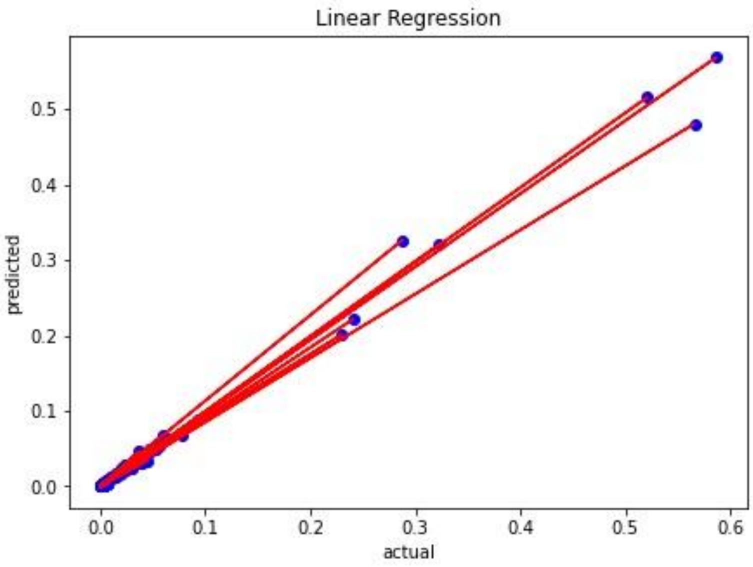


Figure 7

Polynomial Regression:

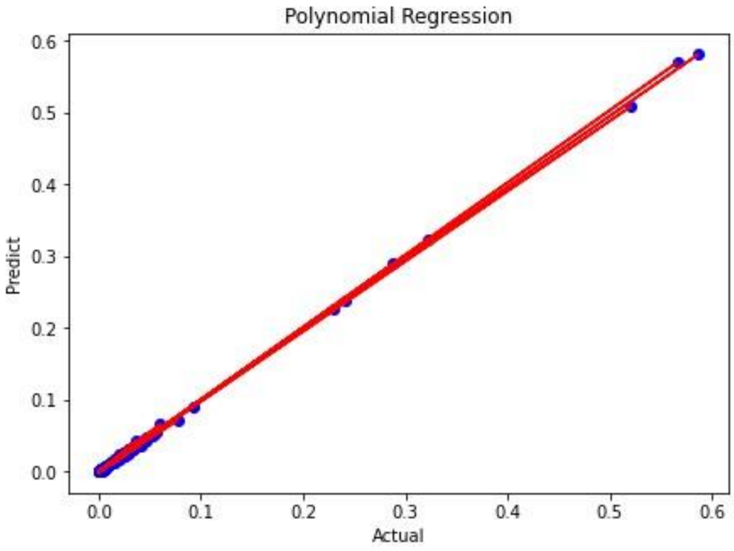


Figure 8

Decision Tree Regressor:

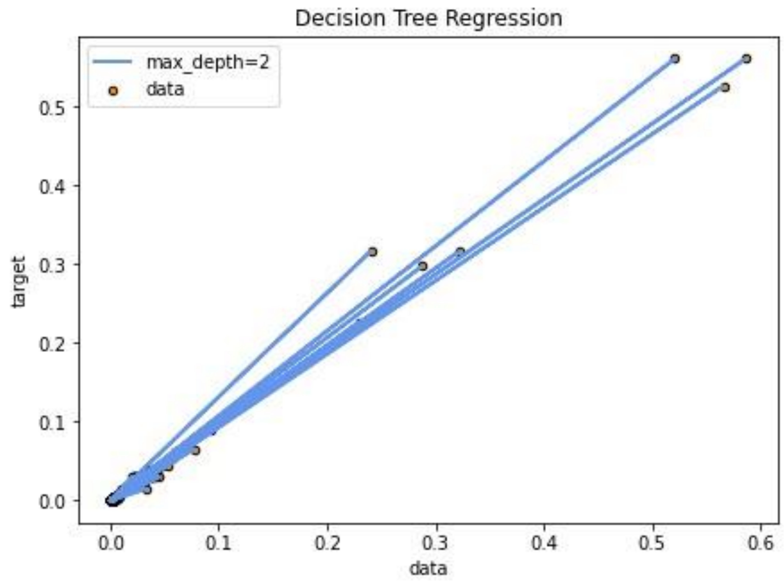


Figure 9

Stacking Regressor:

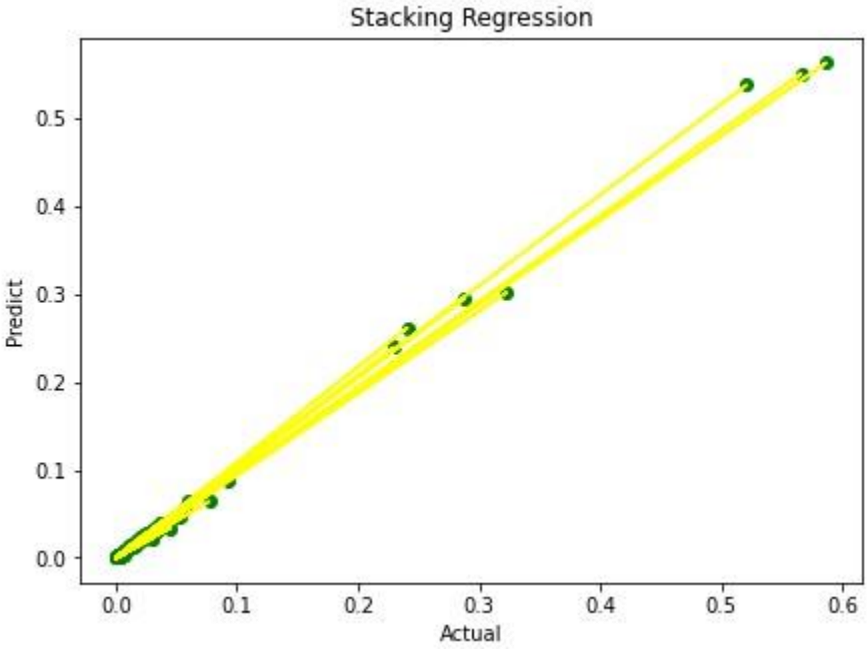


Figure 10

8 Conclusion

Here by we can conclude that we have successfully made 3 models and an ensembled model for Forecasting the energy consumption in steel Industry. The system is trained with 70% of the data, 20% is used for testing purpose and rest 10% of the data is used for validation purpose all the models have achieved excellent accuracy of 99% through ensembled model and as well as based models and we have also seen the average accuracies using cross validation technique and say that this accuracy of 99% is values and we have achieved minimized error which in turn helped us to gain good accuracy.

My project GITHUB link:

https://github.com/Lalith2001/Steel-Industry-Energy-Consumption-Dataset

References:

1. He, Kun, and Li Wang. "A review of energy use and energy-efficient technologies for the iron and steel industry." *Renewable and Sustainable Energy Reviews* 70 (2017): 1022-1039.
2. Holappa, Lauri. "A general vision for reduction of energy consumption and CO2 emissions from the steel industry." *Metals* 10, no. 9 (2020): 1117.
3. Sun, Wenqiang, Qiang Wang, Yue Zhou, and Jianzhong Wu. "Material and energy flows of the iron and steel industry: Status quo, challenges and perspectives." *Applied Energy* 268 (2020): 114946.
4. Conejo, Alberto N., Jean-Pierre Birat, and Abhishek Dutta. "A review of the current environmental challenges of the steel industry and its value chain." *Journal of environmental management* 259 (2020): 109782.
5. Wang, R. Q., Long Jiang, Y. D. Wang, and Anthony Paul Roskilly. "Energy saving technologies and mass-thermal network optimization for decarbonized iron and steel industry: A review." *Journal of Cleaner Production* 274 (2020): 122997.
6. https://scikit-learn.org/
7. http://archive.ics.uci.edu