

Capstone Project

Appliances Energy Prediction

Chetan Chavan

Contents:

Al

- Problem Statement.
- Dataset Variables.
- Energy Crisis and its Solution.
- Dataset Division.
- Training process.
- Visualization.
- Results.
- Conclusion
- Hyperparameter Tuning

Problem Statement:



- As the energy crisis increasing day by day it affects on various factors. Its not only affecting the world energy requirement issues but also affects the economic and social health of any country, mostly in the countries which are going through developing phase. To overcome several related issues, it is necessary to have the outages in order to compensate the load demand. And hence the prediction of energy used by appliances plays a vital role in this concept.
- This project illustrates the energy consumed by house and the data has been collected with the help of sensors. The readings have been taken for each 10 min interval for consecutive 4.5 months. Saving of energy can be done by controlling the energy usage. And thus, prediction of usage comes into picture. This study can save the money of consumer as well as if extra energy is generated then it can also fed back to Grid(Also called as regeneration). We are going to focus on many machine learning regression technique to find out the energy consumption prediction.

Variables:



- date: time → given dat time month and day
- lights: energy used by lights in Wh
- T1: Temperature given in kitchen area, in Celsius
- T2: Temperature given in living room area, in Celsius
- T3: Temperature mentioned in laundry room area
- T4: Temperature of office room, given in CelsiusT5: Temperature recorded in bathroom area, in Celsius
- T6: Temperature given outside the building area particularly (north side), in Celsius
- T7: Temperature provided in ironing room, in Celsius
- T8: Temperature in teenager room 2, in Celsius
- T9: Temperature in parents' room, in Celsius



- T_out: Outside temperature (from Chievres weather station), in °C
- Tdewpoint: (from Chievres weather station),
- RH_1: Kitchen area Humidity %
- RH_2: Living room area Humidity, in %
- RH_3: Laundry room area Humidity, in %
- RH_4: Office room Humidity, in %
- RH_5: Bathroom area Humidity, in %
- RH_6:Outside the building Humidity (north side), in %
- RH_7: Ironing room Humidity, in %
- RH_8: Teenager room 2 Humidity, in %
- RH_9: Parents' room Humidity, in %
- RH_out :Outside Humidity (from Chievres weather station), in %
- Pressure: (from Chievres weather station), in mm Hg



- Wind speed: (from Chievres weather station), in m/s
- Visibility:(from Chievres weather station), in km
- Rv1 :Random variable 1, non-dimensional[1]
- Rv2 :Random variable 2, non-dimensional[1]
- Appliances: Total energy used by appliances, in Wh[1]



Energy Crisis and its Solution:

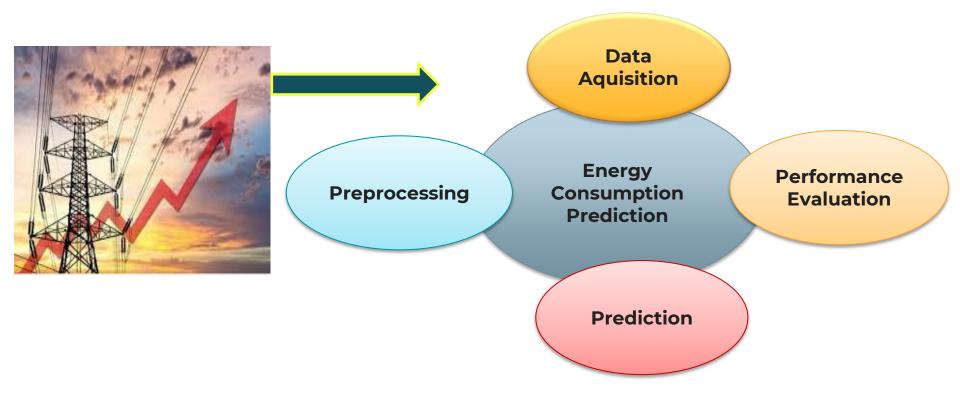


Fig: Prediction methodology for energy consumption using different machine learning algorithms

Data set Division:



There are total 19735 no of rows available. We will go to divide this data into pretraining set, training set, validation set and testing set. The division of data is as follows:

- 75% of data will be put into training set whereas 25% of data will be put into testing set.
- The pre train set was used to find the best models for the given dataset. We have taken best 6 models using pretest set. Their performance will be compared based on their mean absolute errors.
- Once the best 5 models will be obtained, hyper parameters for these models will be tuned and the best parameter will be selected.

Training Process:



We will use following 5 regression techniques to train the data:

It is the regression which uses the shrinkage technique. Which means data values will be shrunk towards the central point. The Lasso regression is very useful when data parameters are few. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator.

Lasso solutions are quadratic programming problems, which are best solved with software (like Matlab). The goal of the algorithm is to minimize: lasso regression

$$\sum_{i=1}^{n} (y_i - \sum_{i} x_{ij} \beta_j)^2 + \lambda \sum_{i=1}^{p} |\beta_j|$$

Which is the same as minimizing the sum of squares with constraint Σ $|B| \le s$ (Σ = summation notation). Some of the βs are shrunk to exactly zero, resulting in a regression model that's easier to interpret.



2] RIDGE Regression: This regression method is mainly used when data having multicollinearity. The method performs L2 regularization. Whenever multi-collinearity problem occurs, least-square are unbiased and variance are large. Because of it predicted value being far away from the actual values.

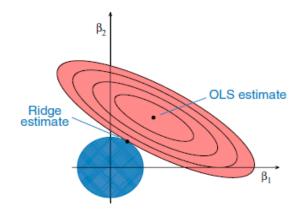


Fig. Ridge Regression



3] Random Forest: Random Forest Regression method is a supervised learning algorithm which uses ensemble learning method for regression technique. Ensemble learning method is nothing but a technique which combines predictions from various machine learning algorithms to prepare a more accurate prediction as compare to the single model.

Test Sample Input

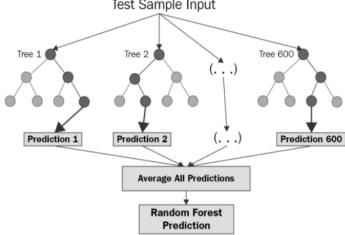


Fig: Random Forest regression



- **4] Gradient Boosting Classifier:** This regression technique calculates the difference between the current predicted value and well known correct target value. This residual is then added to the existing model and this pushes the model towards correct values. To improve the performance of the model we can repeat the process again and again.
- **5] ExtraTree-regressor:** It is a type of ensemble learning technique of regression that adds the results of different de-correlated decision trees which are similar to Random Forest Classifier. Extra Tree can also achieve a good or better prediction than the random forest. The main difference between Random Forest and Extra Tree Classifier is as given below:



- Extra Tree Classifier algorithm never performs bootstrap aggregation as in the random forest. This means, it takes a random subset of data without any replacement. Hence, nodes are always split on random splits but not on best splits.
- In Extra Tree Classifier algorithm randomness doesn't come from bootstrap aggregating but comes from the random splitting of the data.

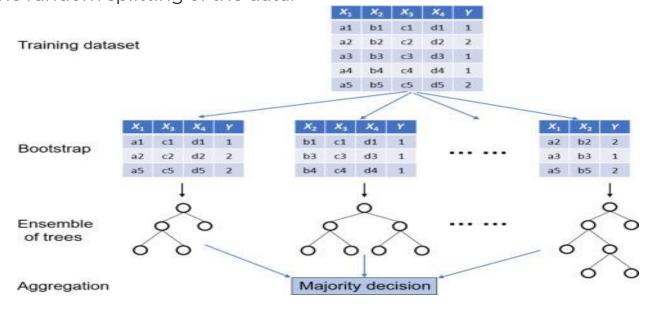


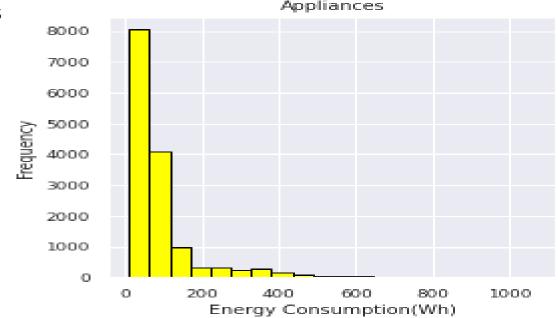
Fig. ExtraRegressor

Visualization:



1] Appliance vs. Energy consumption:

When we will plot the appliance against the energy consumption graph, we can see that percentage of appliances consumption is less than 200Wh. The representation is as follows



next slide.

Appliances	1	0.2	0.055	0.086	0.12	-0.06	0.085	0.036	0.04	0.017	0.02	0.007	0.12	-0.083	0.026	-0.056	0.04	-0 094	0.01	-0.051	0.099	-0 035	-0.15	0.087	0.00023	0.015	-0.011-0	011
lights	0.2	1																									0.00050.0	
Т1	0.055	-0.024	1	0.16	0.84	0.002	0.89	-0.029	0.88	0.098	0.89	-0.015	0.65	-0.62	0.84	0.14	0.83	0.0064	0.84	0.072	0.68	-0.15	-0.35	-0.088	-0.076	0.57	0.00620.	.0062
RH_1	0.086	0.11	0.16	1	0.27	0.8	0.25	0.84	0.11	0.88	0.21	0.3	0.32	0.25	0.021	0.8	-0.03	0.74	0.12	0.76	0.34	-0.29	0.27	0.2	-0.021	0.64	0.00070.	.0007
T2	0.12	-0.0056	0.84	0.27	1	-0.17	0.74	0.12	0.76	0.23	0.72	0.03	0.8	-0.58	0.66	0.23	0.58	0.069	0.68	0.16	0.79	-0.13	-0.51	0.052	-0.07		-0.011-0	.011
RH_2	-0.06	0.051	0.0025	0.8	-0.17	1	0.14	0.68	-0.047	0.72	0.11		0.0097	0.39	-0.051	0.69	-0.041	0.68	0.055	0.68	0.034	-0.26	0.58	0.069	-0.0054		0.00630.	0063
ТЗ	0.085	-0.097	0.89	0.25	0.74	0.14	1	-0.011	0.85	0.12	0.89	-0.066	0.69	-0.65	0.85	0.17	0.8	0.044	0.9	0.13	0.7	-0.19	-0.28	-0.1	-0.1	0.65	0.00520.	0052
RH_3	0.036	0.13	-0.029	0.84	0.12	0.68	-0.011	1	-0.14	0.9	-0.05	0.38	0.077	0.51	-0.25	0.83	-0.28	0.83	-0.2	0.83	0.12	-0.23	0.36		0.017		0.00048.0	0004
T4	0.04	-0.0089	0.88	0.11	0.76	-0.047	0.85	-0.14	1	-0.049	0.87	-0.076	0.65	-0.7	0.88	0.044	0.8	-0.095	0.89	-0.026	0.66	-0.075	-0.39	-0.19	-0.1		0.00180.	0018
RH_4	0.017	0.11	0.098	0.88	0.23	0.72	0.12	0.9	-0.049	1	0.092	0.35	0.26	0.39	-0.13	0.89	-0.17	0.85	-0.045	0.86	0.29	-0.25	0.34	0.3	0.0026	0.62	0.00180.	0018
T5	0.02	-0.079	0.89	0.21	0.72	0.11	0.89	-0.05	0.87	0.092	1	0.033	0.63	-0.63	0.87	0.15	0.82	0.016	0.91	0.072	0.65	-0.17	-0.27	-0.15	-0.084		0.00550.	0055
RH_5	0.007	0.14	-0.015		0.03		-0.066	0.38	-0.076		0.033	1	-0.078	0.26	-0.14	0.33	-0.087	0.36	-0.14	0.27	-0.053	-0.12	0.19	0.082	-0.013	0.078	-0.011-0	.011
Т6	0.12	-0.079	0.65		0.8	0.009	0.69	0.077	0.65		0.63	-0.078	1	-0.67	0.62	0.26	0.48	0.074	0.67	0.18	0.97	-0.14	-0.57	0.17	-0.081	0.76	-0.015 -0	.015
RH_6	-0.083	0.15	-0.62		-0.58		-0.65	0.51	-0.7		-0.63		-0.67	1	-0.75	0.36	-0.67	0.49	-0.74	0.39	-0.64	-0.066	0.72	0.098	0.11	-0.26	0.012 0	.012
77	0.026	-0.14	0.84	0.021	0.66	-0.051	0.85	-0.25	0.88	-0.13	0.87	-0.14	0.62	-0.75	1	-0.034	0.88	-0.21	0.94	-0.078	0.63	-0.098	-0.41	-0.19	-0.11		0.00390.	0039
RH_7	-0.056	0.035	0.14	0.8		0.69	0.17	0.83	0.044	0.89	0.15			0.36	-0.034	1	-0.12	0.88	0.028	0.86		-0.27			-0.0072	0.64	0.00180.	0018
Т8	0.04	-0.071	0.83	-0.03		-0.041	0.8	-0.28	0.8	-0.17	0.82	-0.087	0.48	-0.67	0.88	-0.12	1	-0.21	0.87	-0.16		-0.16	-0.3	-0.22	-0.06		0.00320.	.0032
RH_8	-0.094	0.013	-0.0064	0.74	0.069	0.68	0.044	0.83	-0.095	0.85	0.016	0.36	0.074	0.49	-0.21	0.88	-0.21	1	-0.11	0.86	0.12	-0.23	0.49		0.046		0.00450.	0045
eT	0.01	-0.16	0.84	0.12	0.68	0.055	0.9	-0.2	0.89	-0.045	0.91	-0.14	0.67	-0.74	0.94	0.028	0.87	-0.11	1	0.0087	0.67	-0.16	-0.32	-0.18	-0.1		0.00120.	0012
RH_9	-0.051	-0.0088	30.072	0.76	0.16	0.68	0.13	0.83	-0.026	0.86	0.072		0.18	0.39	-0.078	0.86	-0.16	0.86	0.0087	1	0.22	-0.18	0.36		0.0087		-0.003-0	.003
T_out	0.099	-0.074	0.68	0.34	0.79	0.034	0.7	0.12	0.66		0.65	-0.053	0.97	-0.64	0.63	0.29	0.5	0.12	0.67	0.22	1	-0.14	-0.57	0.19	-0.077	0.79	-0.015 -0	.015
Press_mm_hg	-0.035	-0.011	-0.15	-0.29	-0.13	-0.26	-0.19	-0.23	-0.075	-0.25	-0.17	-0.12	-0.14	-0.066	-0.098	-0.27	-0.16	-0.23	-0.16	-0.18	-0.14	1	-0.092	-0.24	0.04	-0.24	0.00070.	0007
RH_out	-0.15	0.069	-0.35		-0.51	0.58	-0.28		-0.39		-0.27	0.19	-0.57	0.72	-0.41	0.38	-0.3	0.49	-0.32	0.36	-0.57	-0.092	1	-0.18	0.083	0.037	0.02	0.02
Windspeed	0.087	0.06	-0.088		0.052	0.069	-0.1		-0.19		-0.15	0.082	0.17	0.098	-0.19		-0.22	0.2	-0.18		0.19	-0.24	-0.18	1	-0.0075	0.13	-0.011 -0	.011
Visibility	0.0002	3 0.02	-0.076	-0.021	-0.07	-0.0054	4 -0.1	0.017	-0.1	0.0026	-0.084	-0.013	-0.081	0.11	-0.11	-0.0072	2 -0.06	0.046	-0.1	0.0087	-0.077	0.04	0.083	-0.007	1	-0.042	0.00590.	0059
Tdewpoint	0.015	-0.036	0.57	0.64	0.58	0.5	0.65	0.41	0.52	0.62	0.59	0.078	0.76	-0.26	0.47	0.64	0.39	0.5	0.58	0.54	0.79	-0.24	0.037	0.13	-0.042	1	0.00390.	0039
rv1	-0.0110	0.0005	20.0062	0.0007	7-0.011	0.0063	30.0054	0.0004	8 .0018	0.0018	90.0055	-0.011	-0.015	0.012	-0.0039	90.0018	0.003	20.0045	0.0012	2-0.003	-0.015	0.0007	0.02	-0.011	-0.0059	0.0039	1	1
rv2	-0.011	0.0005	20.0062	0.0007	7-0.011	0.0063	30.0051	0.0004	8 .0018	0.0018	0.0055	-0.011	-0.015	0.012	-0.0039	90.0018	0.003	20.0045	0.0012	2-0.003	-0.015	0.0007	0.02	-0.011	-0.0059	0.0039	1	1
	Appliances	lights	Ľ	H_1	77	M_2	Е	판	74	RH_4	Б	. 품	12		Д	RI_7	22	E 8	EL.	R.	Tout	Press_mm_hg	RH_out	Windspeed	Visibility	Tdewpoint	Ν	N2

- 0.8

- 0.6

- 0.4

- 0.2

- 0.0

- -0.2

- -0.4

- -0.6



3] Feature Importance: From feature importance graph we can analyzed that most important features are 'RH_out', 'RH_8', 'RH_1', 'T3', 'RH_3'. Whereas least important features will be - 'T7','Tdewpoint','Windspeed','T1','T5'.

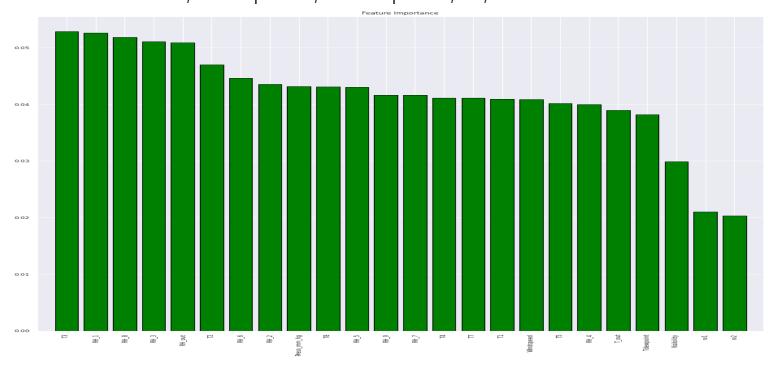
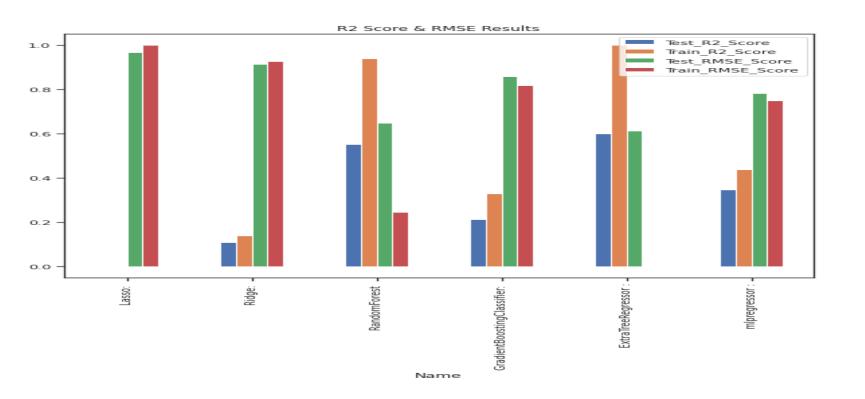


Fig. Feature graph



4] Comparison Graph:







	Name	Train_R2_Sc ore	Test_R2_Sco re	Train_RMSE_Sco re	Test_RMSE_ Score		
0	Lasso	0.000000	-0.000517	1.000000e+00	0.968239		
1	Ridge	0.140903	0.109054	9.268748e-01	0.913684		
2	RandomForest	0.939403	0.552971	2.461646e-01	0.647200		
3	Gradient Boosting Classif ier	0.329343	0.214281	8.189363e-01	0.858033		
4	ExtraTreeRegressor	1.000000	0.599255	1.326785e-15	0.612780		
5	mlpregressor	0.438139	0.346108	7.495736e-01	0.782750		



Conclusion:

1) It is clearly seen that best results for test set is being given by Extra Tree Regressor with R2 score of 0.599255

- 2) Least RMSE score is also by Extra Tree Regressor 0.612780
- 3) Lasso regression model over Linear regression was not giving good result and hence proven to be the worst model.

Parameter Tuning and observation:



Depending on parameter tuning we can conclude that

• Best possible parameter combination are - 'max_depth' is 100, max_features is 'sqrt', 'n_estimators' is 260 and random state is 40.

- Training set R2 score of 1.0 shows the overfitting on training set.
- Using hyperparameter tuning the R2 score can be improved from 0.59 to 0.60 of the Test set.
- Test set RMSE score is 0.60 which is get improved from 0.61 achieved using hyperparameter tuning.



THANK YOU.....