Computational Machine Learning
Assignment - 2

Predicting Energy use by appliances using Regression

By
Ayaz Aziz Mujawar

Abstract

The aim of this project is to find the best regression model to predict the energy use of the appliances in the low energy building based on the house temperature and humidity conditions. Regression models namely Polynomial Regression, Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regressor were adopted and trained using K-Fold cross validation with a split of 5. The report concludes that Random forest regressor and Gradient Boosting regressor are the best performing model to predict the energy use among all other models. It is recommended to use ANN model to further increase the accuracy and to decrease the error rate in the model.

Methodology

Data Retrieval

The dataset "Appliances Energy prediction dataset" used for the regression approach is provided by UCI repository. This dataset consists of 27 descriptive features and "Target_energy" as target features. This dataset consists of numerical and datetime features with 19735 observations and 28 columns including target variable.

Attribute Information

T1...T9 - Temperature in different rooms like Kitchen, laundry room, bathroom, etc.
RH1...RH9 - Humidity in different rooms like kitchen, living room, laundry room, etc.
Wind Speed - wind speed in m/s.
Visibility - visibility in km.
Dew Point - dew point in degree celsius.
rv1, rv2 - Random variables

Data Exploration

Data Exploration is mainly done to understand the characteristic of data in the dataset. The two main goals of data exploration are to understand the distribution of data across the range and to determine whether the data contains outlier or missing values. To check the distribution of the data we calculated the descriptive statistics of the data as shown in Fig 1.1.
This feature shows the central tendency of each and every features column in the dataset.

Missing values: We found no missing values in the dataset.

Outliers: Outliers are the values that fall outside the given range of values. As shown in the figure 1.2(histogram) and 1.3(Boxplot) we can say that most of the data points in the columns are distributed normally and also most of the data points lie close to one another so it is difficult to tell whether the given column contains outliers. So we did not remove any data points from the dataset to avoid data loss.

Data Preparation

We started by separating all the descriptive features from the target features. Then we checked the datatype for all the features. As we found that the datatype of "date" column was "object" so we converted that date column data type to "datetime" datatype.

Feature engineering was done on the datetime column to generate some other features like day, month, year, week, hours and minutes and date feature was dropped. We then converted the 'hours' feature to 'session' feature which contains values like 'morning', 'afternoon', 'evening', etc. As we can see the 'session' feature is of string type we converted that feature into integer type by using label encoder.

Further we can observe that the most of the features in our dataset vary in magnitude, units and ranges. For example, the units and ranges of "wind speed" and "visibility" features are different. So to overcome those issues we applied feature scaling using a robust scalar method that scales the feature values to the same level and also reduces the effect of outliers. Feature selection was done using Random Forest Importance(rfi) which is used to select 10 most important features from the dataset.

Correlation matrix was calculated that shows the relationship between all the independent features and target variable. Correlation shows the strength of association between variables and target variables. If the value of a relationship is 1 then there is a strong correlation and if the value is -1 then there is a negative correlation between the variables.

Data Modelling

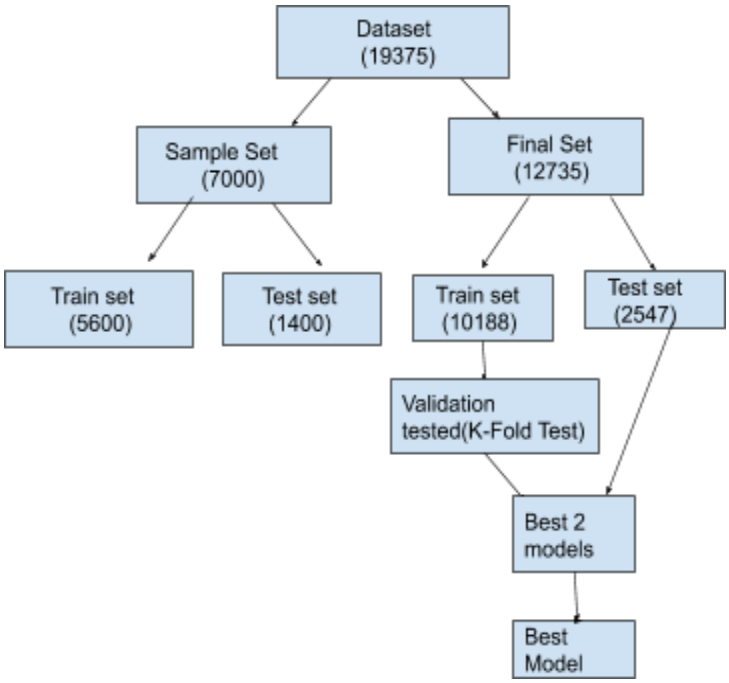

Fig: 1.5

From the above figure 1.5 we can see that the data is divided into two parts 'Sample set' and 'Final set'. Sample set contains 7000 data points from the dataset. The sample set is further divided into 80% of the training set and 20% of the testing set.
The Final set consists of 12735 data points. Out of which 80% are given to the training set and 20% to the test set. Different regression models are applied on the sample set and out of that 4 best models are selected.
These 4 selected models along with hyperparameter tuning and K-Fold cross validations with the split of 5 are then applied to the training data of the Final set. This validation set was used to give the best model out of that 4 selected models and finally the 2 selected models are applied on the unseen test data from the final set.

Evaluation Metrics

Mean Absolute Error (MAE) - MAE defines the average of absolute difference between the predicted and the observed value. The main reason for using MAE over MSE is because we have seen that there are chances of outliers and we did not remove it because of the data loss. So MAE is more robust to outliers as compared to MSE.

Adjusted R2 - Adjusted R2 helps in measuring the effect of independent variables on the dependent variables. The main reason for selecting Adjusted R2 over R2 is that Adjusted R2 value will change only when an independent variable holds a significant relationship with the dependent variable. But in case of R2 the value of R2 changes even if we add features to the model irrespective of the significance.

Model Training

Regression model that we applied on the sample dataset are:
Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Gradient Boosting Regressor, Support Vector Regressor and Random Forest Regressor.
The performance of the model is based on accuracy R-Squared and Adjusted R-Squared and error measures were Mean Absolute Error(MAE), Mean Squared Error(MSE) and Root Mean Squared Error(RMSE)

From the table --- we can say that the best 4 models were selected after observing the accuracy and error terms are: Random Forest Regressor, Gradient Boosting Regressor, Decision Tree Regressor and Polynomial Regression. All these models are applied on K-Fold cross validation with the split of 5.

Random Forest Regressor

Random Forest Regression is an ensemble based learning based on decision trees. In random forest regressor predictions are calculated based on averaging the predictions of each decision tree. Random Forest is good at handling numerical and categorical features and also works well on non-linear datasets. The Random Forest Regressor is imported from the 'sklearn.ensemble' package. For the Random Forest Regressor we got max depth of 20,
Min sample split of 20 and total features on which model was applied was 10.

Gradient Boosting Regressor

Gradient boosting Regression produces a prediction model in the form of a collection(ensemble) of weak prediction models like decision trees.
Tuned Parameter- Performed parameter tuning for the number of max depths. The n_estimators default is 100, criterion is mse and minimum sample split is 2.
Selected Features- The random forest importance is used for feature selection for every split and the regressor is performed on the features with the best scores.
The quality of the models are assessed on the basis of Mean Absolute Error.

Decision Tree Regressor

The DT algorithm regressor repeatedly divides the data into sub-parts based on the information gain. The default values of the parameter control the size of the tree leads to unpruned trees.
Tuned Parameter- Performed parameter tuning for the number of max depths and minimum sample splits.
Selected Features- After executing the random forest importance for feature selection the classifier is
The quality of the models are assessed on the basis of Mean Absolute Error.

Polynomial Regression

Polynomial regression is used to fit the non-linear relationship between the independent variable and dependent variable by considering the nth degree polynomial. For the polynomial regressor we have taken the degree from 1 to 6 and finally calculated the average of mean absolute error

Results

The results are calculated on the Training and the Validation set on the Final set data. For the Random Forest Regressor, the MAE is calculated. The results are found from the K-Fold cross validation. The best parameters that we got are max_depth of 20, min_samples_split of 20, important features are 10. Finally the MAE score that we got is -37.87. For the gradient boosting regressor model we got the best maximum depth of 15 and MAE score is -31.61. For the decision tree regressor the max_depth and min_samples_split are selected as 20 and MAE score is -40.36 and for the polynomial regression the MAE score is -46.65.

For the test set, the best models that have less absolute error are Random Forest Regressor and Gradient boosting regressor. These two models are now tested on a test set and the results were recorded.

Random Forest Regressor - MAE = -43.77
Gradient boosting Regressor - MAE = -42.92

It was observed that there is not much difference between the MAE results of both the models when tested on test data. As it is observed that the random forest gives the best result and also prevents overfitting of data. Also random forest is easy to use and we can apply it anywhere when we want a decent result but GBR doesn't work without cross validation. So out of both of these Random Forest Regressors are the best among all others for this dataset.

Independence Evaluation

1. Predicting Electric Energy Use of a Low Energy House: A Machine Learning Approach

In this paper data was collected and preprocessed and feature selection was done using PCA and F-test. Normalization of data was done using min-max scaling. The dataset was divided into 80% training set and 20% testing set and K-Fold cross validation is also applied to this training set. They have applied the BP-ANN and SVM model and error was calculated using MSE and RMSE. The BP-ANN model is very efficient in predicting electric energy. We have selected the model based on mean absolute error and in this paper authors have selected the best model using MSE and RMSE. The reason to use MAE over MSE is due to the fact that MAE is more robust to outliers since it doesn't make use of squares.

Conclusion

The random forest regressor and Gradient boosting regressor are the two best models among all other regression models that we applied on the dataset. For all the models "Days" is marked as an important feature followed by the 'T3' and 'T_out'. The mean absolute error for random forest regressor and gradient boosting regressor are -43.77 and -42.92 respectively. The regressor worked well by selecting the top 10 features by using random forest importance(rfi) and along with that it also prevents data from overfitting.

Appendix

Models	R-Square	Adjusted R-Square	MAE	MSE	RMSE
Linear Regression	0.05	0.04	69.69	14244.17	119.34
Polynomial Regression	0.163	0.15	63.09	12553.11	112.04
Ridge Regression,	0.05	0.04	69.68	14244.75	119.35
Lasso Regression	0.03	0.02	70.86	14499.75	120.41
Decision Tree Regressor,	0.428	0.425	40.0	8565.35	92.54
Gradient Boosting Regressor,	0.25	0.24	59.05	11224.16	105.94
Support Vector Regressor	-0.081	-0.087	57.79	16215.51	127.34

Random Forest Regressor.	0.54	0.54	38.90	6771.82	82.29

Table 1

	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4	T5	
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	1973
mean	21.686571	40.259739	20.341219	40.420420	22.267611	39.242500	20.855335	39.026904	19.592106	5
std	1.606066	3.979299	2.192974	4.069813	2.006111	3.254576	2.042884	4.341321	1.844623	
min	16.790000	27.023333	16.100000	20.463333	17.200000	28.766667	15.100000	27.660000	15.330000	1
25%	20.760000	37.333333	18.790000	37.900000	20.790000	36.900000	19.530000	35.530000	18.277500	4
50%	21.600000	39.656667	20.000000	40.500000	22.100000	38.530000	20.666667	38.400000	19.390000	4
75%	22.600000	43.066667	21.500000	43.260000	23.290000	41.760000	22.100000	42.156667	20.619643	5
max	26.260000	63.360000	29.856667	56.026667	29.236000	50.163333	26.200000	51.090000	25.795000	9

Fig: 1.1

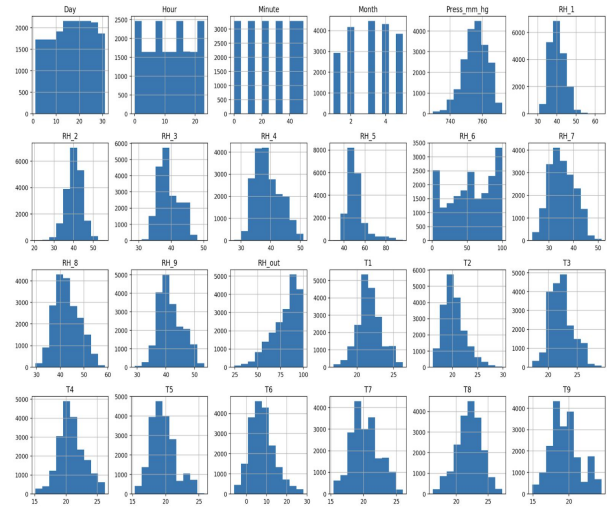

Fig: 1.2

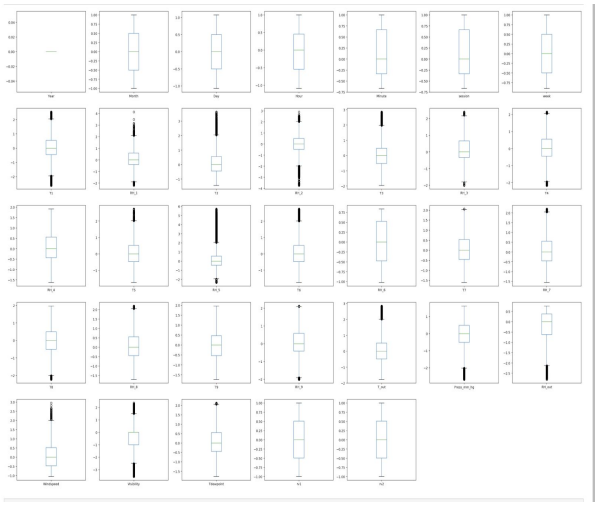

Fig: 1.3

References

1. https://www.researchgate.net/publication/331890052_Predicting_Electric_Energy_Use_of_a_Low_Energy_House_A_Machine_Learning_Approach
2. https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction
3. https://www.featureranking.com/