**Appliances Energy Prediction**

**Abhijeet Dutta**

**Data science trainee**

**Alma Better, Bangalore**

**Abstract:**

Supervised learning, also known as supervised machine learning, is **a subcategory of machine learning and artificial intelligence**. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. Regression is **a supervised machine learning technique which is used to predict continuous values**. The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data.

***Keywords: supervised machine learning, regression, energy***

**1.Problem Statement**

* The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters. Weather from the nearest airport weather station Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru) and merged together with the experimental data sets using the date and time column. Two random variables have been included in the data set for testing the regression models and to filter out non-predictive attributes (parameters).

**Input variables:**

* date time year-month-day hour: minute: second
* Appliances, energy use in Wh (Dependent variable)
* lights, energy use of light fixtures in the house in Wh (Drop this column)
* T1, Temperature in kitchen area, in Celsius
* RH1, Humidity in kitchen area, in % T2, Temperature in living room area, in Celsius RH2,
* Humidity in living room area, in %
* T3, Temperature in laundry room area
* RH3, Humidity in laundry room area, in % T4, Temperature in office room, in Celsius RH4,
* Humidity in office room, in %
* T5, Temperature in bathroom, in Celsius
* RH5, Humidity in bathroom, in % T6, Temperature outside the building (north side), in Celsius
* RH6, Humidity outside the building (north side), in %
* T7, Temperature in ironing room, in Celsius
* RH7, Humidity in ironing room, in % T8, Temperature in teenager room 2, in Celsius RH8,
* Humidity in teenager room 2, in %
* T9, Temperature in parent`s room, in Celsius
* RH9, Humidity in parent’s room, in % To, Temperature outside (from Chievres weather station), in
* Celsius Pressure (from Chievres weather station), in mm Hg RHout, Humidity outside (from
* Chievres weather station), in %
* Wind speed (from Chievres weather station), in m/s
* Visibility (from Chievres weather station), in km
* Tdewpoint (from Chievres weather station), Â°C
* rv1, Random variable 1, nondimensional
* rv2, Random variable 2, nondimensional

**2. Introduction**

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

## **3. Target Variable**

* Appliances

## **4. Reasons for consumption**

* Extensive use of electric appliances to maintain the temperature inside the house suitable for living.

.

**6. Steps involved:**

**Exploratory Data Analysis**

After loading the dataset we performed this method to explore the hidden pattern of our target variable that is Y with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

It also helps to take decisions about feature engineering.

**Null values Treatment**

There is no missing data in our dataset.

**Encoding of categorical columns**

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features as well as manually map different categories of the variables to numerical values. Because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

**Feature Selection**

In these steps we used algorithms like Extra Tree classifier to check the results of each feature i.e which feature is more important compared to our model and which is of less importance.

**VIF:**

Variance inflation factor (VIF) is **a measure of the amount of multicollinearity in a set of multiple regression variables**. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

**Fitting different models**

For modelling we tried various classification algorithms like:

1. **Lasso**
2. **Ridge**
3. **SVR**
4. **Random Forest Regressor**
5. **Extra Trees Regressor**
6. **Gradient Boosting Regressor**
7. **XGB Regressor**
8. **Stacking (KNN, DT Regressor, SVR, Extra Trees, Random Forest Regressor)**

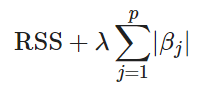
**Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models like Random Forest Regressor and ExtraTrees Regressor.

**7.1. Algorithms:**

**1. Lasso regression:**

### Lasso stands for least absolute shrinkage and selection operator

****

### **Where λ is a tuning parameter that seeks to balance between the fit of the model to the data and the magnitude of the model's coefficients:**

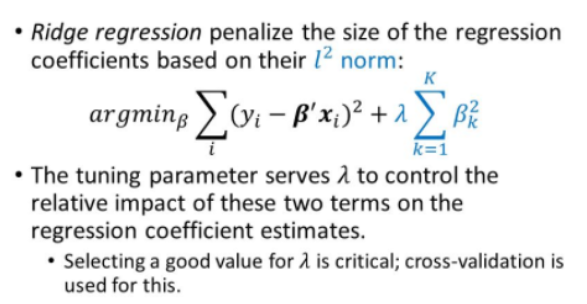
### **A tiny λ imposes no penalty on the coefficient size, and is equivalent to a normal linear regression.**

### **Increasing λ penalizes the coefficients and thus shrinks them towards zero.**



.**2. Ridge regression:**

Ridge regression is **a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated**. It has been used in many fields including econometrics, chemistry, and engineering.

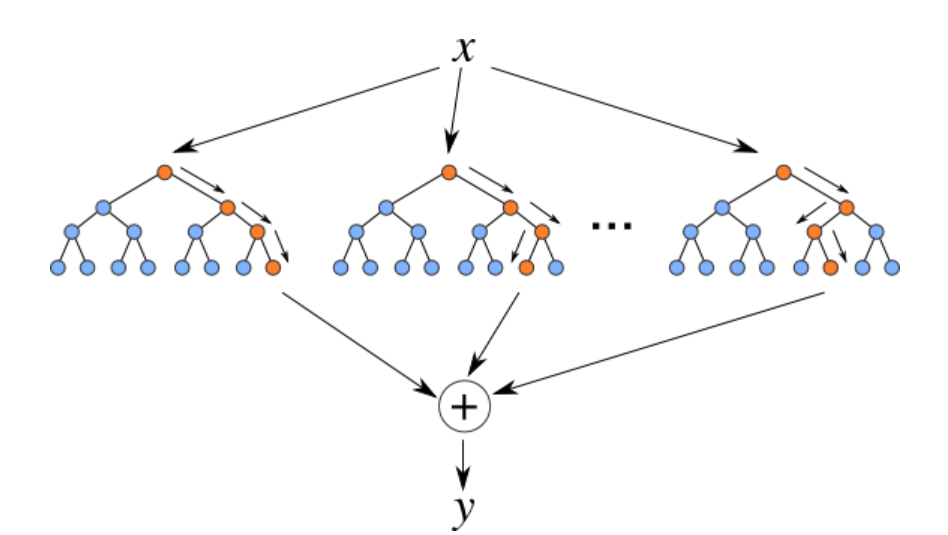


**3. Support Vector Regressor:**

Support Vector Regression is **a supervised learning algorithm that is used to predict discrete values**. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

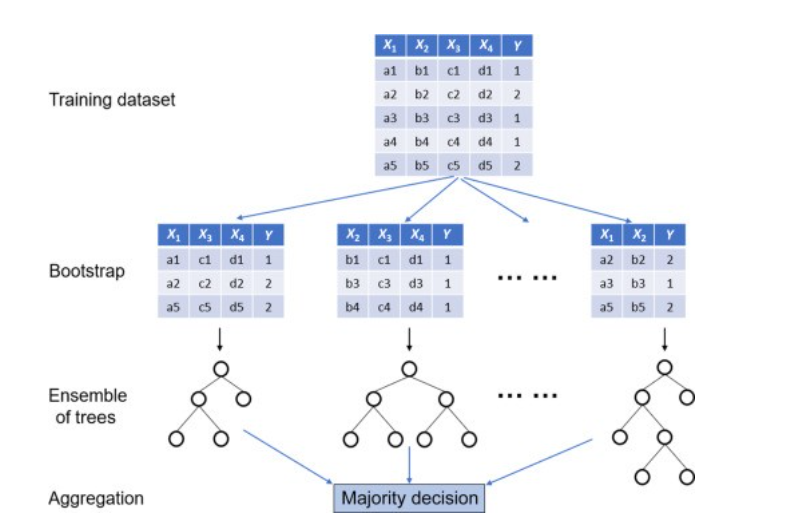
**4. RandomForest Regressor:**

Random Forest Regression is **a supervised learning algorithm that uses ensemble learning method for regression**. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



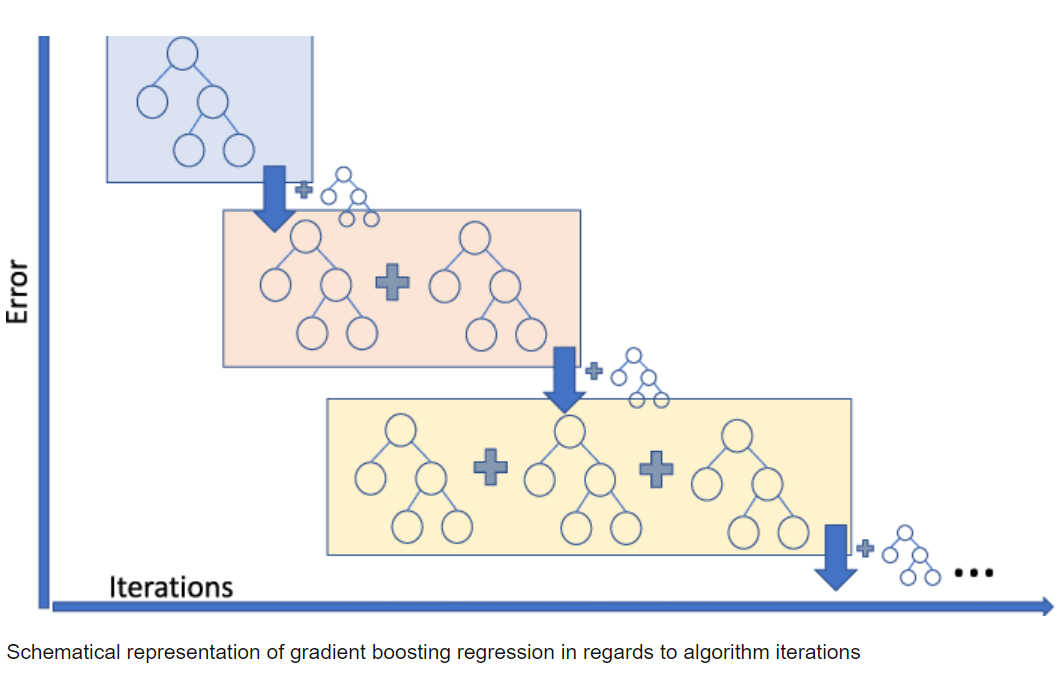
**5. ExtraTrees Regressor:**

An extra-trees regressor. This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

****

**6. GradientBoosting Regressor:**

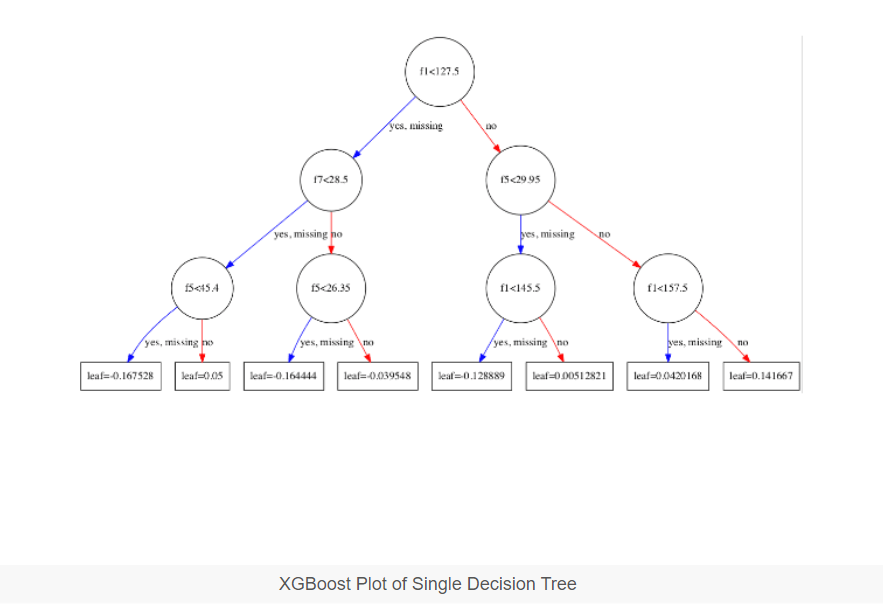
Gradient boosting is **a machine learning technique used in regression and classification tasks**, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

****

**7. XGB Regressor:**

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm.

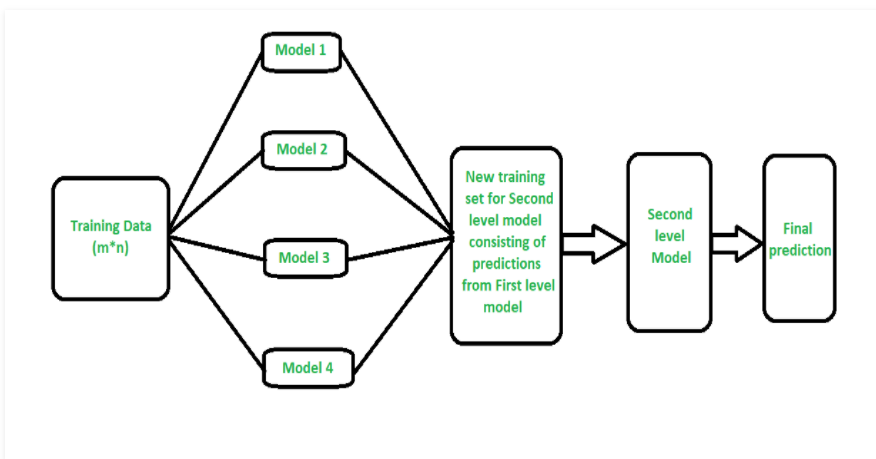
Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. **XG Boost** can be used directly for **regression predictive modeling**.

****

**8. Stacking:**

Stacking is a way to ensemble multiple classifications or regression model. There are many ways to ensemble models, the widely known models are ***Bagging***or ***Boosting***. Bagging allows multiple similar models with high variance are averaged to decrease variance. Boosting builds multiple incremental models to decrease the bias, while keeping variance small.

Stacking (sometimes called *Stacked Generalization*) is a different paradigm. The point of stacking is to explore a space of different models for the same problem. The idea is that you can attack a learning problem with different types of models which are capable to learn some part of the problem, but not the whole space of the problem. So, you can build multiple different learners and you use them to build an intermediate prediction, one prediction for each learned model. Then you add a new model which learns from the intermediate predictions the same target.  
This final model is said to be stacked on the top of the others, hence the name. Thus, you might improve your overall performance, and often you end up with a model which is better than any individual intermediate model. Notice however, that it does not give you any guarantee, as is often the case with any machine learning technique.

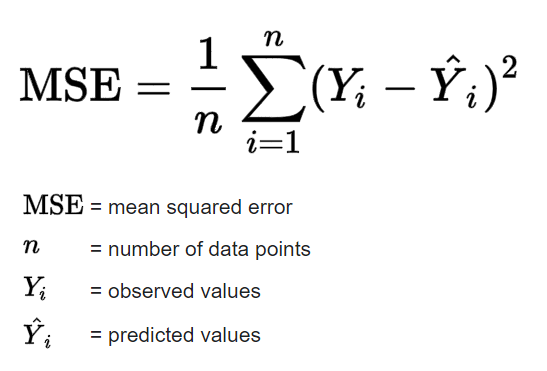
****

**7.2. Model performance:**

Model can be evaluated by various metrics such as:

**1. Mean Squared Error (MSE**):

The mean squared error (MSE) **tells you how close a regression line is to a set of points**. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs.



**2. Root Mean Squared Error (RMSE):**

Root mean squared error (RMSE) is **the square root of the mean of the square of all of the error**. The use of RMSE is very common, and it is considered an excellent general-purpose error metric for numerical predictions.

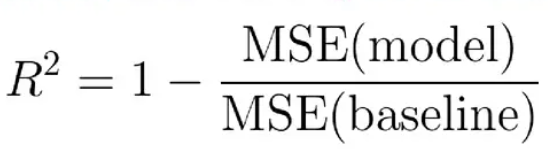
**3. R Squared:**

Coefficient of Determination or R squared is another metric used for evaluating the performance of a regression model.

This metric helps us to compare the current model to a constant baseline model and tells us how much our model is better.

The constant baseline is chosen by taking the mean of the data and drawing a line at the mean.

It is a scale free score that implies it doesn’t matter whether the values are too large or too small, R squared will always be less than or equal to 1.



**7.3. Hyper parameter tuning:**

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Randomized Search CV for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into 5 folds.

**Randomized Search CV-** In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist’s control

**Grid Search CV –**

GridSearchCV **tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method**. Hence after using this function, we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.

**8. Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA (Univariate and Bivariate), encoding of categorical columns,

feature selection (using **VIF**),

Through pipeline we have used SVR, RandomForest, ExtraTrees, Gradient Boosting regressor and XGB regressor, since there is a scenario of overfitting so we have implemented L1 and L2 regularization techniques to penalize the RSS of the train set to overcome the effect of overfitting but the result is not satisfactory.

Only ExtraTrees regressor and RandomForest regressor produce sustainable R squared value for the test set. So, the next step is to tune the parameters of the above mentioned algorithms. ExtraTrees gives 60% and Random Forest produces 59% test model score.

We even tried stacking method in order to get optimal model.

But the test score is 54% only.

**References-**

1. sklearn documentation

2. GeeksforGeeks

3. Analytics Vidhya

4. Towards Data Science