**Appliance Energy Prediction**

Karan Malhotra

**Data science trainee,**

**AlmaBetter**

**Abstract:**

The electricity consumption of a house is mainly related to the type and quantity of household appliances, and the household appliances have an influence on the indoor environment, such as temperature, humidity, and light.Therefore, it is possible to predict the energy consumption of household appliances by establishing relevant models and using different environmental data and electricity consumption data. Energy prediction of appliances requires identifying and predicting individual appliance energy consumption when combined in a closed chain environment. This experiment aims to provide insight into reducing energy consumption by identifying trends and appliances involved. In this paper, support vector machine, k-nearest neighbor, random forest, extremely random forest, extra-tree regressor are used to build models for predicting the energy consumption of household appliances, and the performance between them are compared.

***Keywords: energy prediction,random forest, support vector machine ,closed chain enviroment***

**1.Problem Statement**

### Data-driven prediction of energy use of appliances, 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).

· 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 parents room, in Celsius

· RH9, Humidity in parents room, in %

· Tout, 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**

With the continuous development of China's cities and the growth of residential construction, the energy consumption of residential buildings has increased in recent years. The increase in the economic level of residents has led to an increase in the purchase of household appliances. Researchers have studied between residential electricity consumption and per capita income, electricity prices, gas prices, and climatic conditions. The relationship predicts the development trend of domestic electricity consumption. Forecasting residential energy consumption not only helps to understand the energy consumption of residential buildings in China,  
but also provides analysis data for China's energy conservation and emission reduction work, and can also be used to detect abnormal energy consumption patterns

Hence, the problem of energy consumption is addressed and data factors affecting the consumption are analyzed.

**3. Research method**

For the prediction of household appliance energy consumption, the typical research in traditional machine learning methods is to use multiple regression, neural network, support vector machine and other methods to establish related models to predict the energy consumption of household appliances.In this project, support vector machine, k-nearest neighbor, random forest, extremely random forest, long short-term memory are used to build models for predicting the energy consumption of household  
appliances, and the performance between them are compared.

1. **Model parameters**

* In the experiment, the prediction model based on the traditional machine learning method has many parameters, so the grid search method is used to determine the optimal parameters of the model.
* The experiment uses R2, MSE,MAE and RMSE as the  
  optimization indicators.
* All experiments are done in python.
* According to the evaluation index, the model with lower RMSE and higher R2 will have better performance

**5. How energy consumption prediction works**

In the age of smart homes, ability to predict energy consumption can not only save money for end user but can also help in generating money for user by giving excess energy back to Grid (in case of solar panels usage). In this case regression analysis will be used to predict Appliance energy usage based on data collected from various sensors. we need to develop a supervised learning model using regression algorithms.

Regression algorithms are used as data consist of continuous features and there are no identification of appliances in dataset. The dataset was collected by sensors placed inside the house and outside readings came from the nearby weather station. The main attributes are temperature, humidity and pressure readings. Each observation measures electricity in a 10-minute interval. The temperatures and humidity have been averaged for 10-minute intervals.

**6. Steps involved:**

● **Exploratory Data Analysis**

Date column is only used for understanding the consumption vs date time behavior and given this is not a time series problem. I added one more column temporarily (WEEKDAY)which focuses on if a day was weekday or weekend in order to check the difference in appliance consumption.Then we check the

descriptive stats of temprature and humidity, weather and appliance columns.

● **Null values Treatment**

Our dataset contains no null values except lights coloumn which was droped after analysis.

● **Encoding of categorical columns**

There are no categorical or ordinal features.

● **Feature Selection**

In these steps we selected and analyzed different features given in the dataset to gain some insights for each feature and for better understanding of the data.

● **Data Visualization**

Our main motive through this step was to present our data in the form of different graphs as not much can be understood just with the data so it is better to use some pictorial representation of the data and then do the analysis.

● **Standardization of features**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

● **Fitting different models**

For modelling we tried various classification algorithms like:

1. **Linear Regression**
2. **Lasso,Ridge Regression**
3. **SVM Regression**
4. **Random Forest Regression**
5. **ExtraTree Regressor**
6. **XGBoost**

● **Tuning the hyperparameters for better accuracy**

Extra Trees Regressor performed the best with default parameters. I used grid search cross validation using the GridSearchCV function of the sklearn.model\_selection library. The parameters which were tuned :

1. n\_estimators: The number of trees to be used

2. max\_features: The number of features tob e considered at each split

3. max\_depth : The maximum depth of the tree , If no param is provided then splitting will continue till all leaves are pure or contain less the min\_samples\_split specified

**7.1. Algorithms:**

This is a Regression problem. Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). The regression methods used are

**1.Linear Models** :

**Linear Regression**

In linear regression we wish to fit a function in this

Form Ŷ = β0+β1X1+β2X2+β3X3 where X is the vector of features and β0, β1 ,

β2, β3 are the coefficients we wish to learn. It updates β at every step by

reducing the loss function as much as possible. As modification to Linear regression model, we can apply Regularization techniques to penalize the coefficient values of the features, since higher values generally tend towards overfitting and loss of generalization.

**Ridge Regression**

This loss function includes two elements. Sum of distances between each prediction and its ground truth. The second element sums over squared β values and multiplies it by another parameter λ. The reason for doing that is to “punish” the loss function for high values of the coefficients β.

It enforces the βcoefficients to be lower, but it does not enforce them to be zero. That is, it will not get rid of irrelevant features but rather minimize their impact on the trained model.

**Lasso Regression**

The only difference from Ridge regression is that the regularization term is in absolute value. But this difference has a huge impact on the trade-off. Lasso method overcomes the disadvantage of Ridge regression by not only punishing high values of the coefficients β but actually setting them to zero if they are not relevant. Therefore, we might end up with fewer features included in the model than we started with, which is a huge advantage.

**2.Support Vector Machine**

**Support vector regression**

The Support Vector Regression (SVR) uses the same principles as the SVM for classification . In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem.

**3.Nearest neighbour Regressor**

**KNeighborsRegressor**

KNeighborsRegressor retrieve some k neighbors of query objects, and make predictions based on these neighbors . It computes the mean of the nearest neighbor labels.

**4.Tree based Regression models**

We divide the predictor space — that is, the set of possible values for X1, . . . , Xp — into J distinct and non-overlapping regions, R1, . . . , RJ . For every observation that falls into the region Rj , we make the same prediction, which is simply the mean of the response values for the training observations in Rj .

Our goal is to find boxes R1, . . . , RJ that minimize the RSS given by

RSS = X J j=1 X i∈Rj (yi − yˆRj ) 2 ,

where yˆRj is the mean response for the training observations within the jth box. Tree based models are less affected by outliers as compared to Linear models. Given there isn’t a linear relation between any input and the target variable, so it is likely that Trees will work better than Linear models.

**Ensemble methods**

It combines several decision trees to produce better predictive performance than utilising a single decision tree. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner.

- **Bagging :** Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees. Average of all the predictions from different trees are used which is more robust than a single decision tree.& Boosting

- **Boosting** : Boosting is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.

**Random Forests**

A Random Forest is an ensemble technique capable of performing both regression tasks with the use of multiple decision trees and a technique called bagging. and works well on high dimensional data

**Gradient Boosting Machines**

Gradient Boosting is an extension over boosting method. It uses gradient descent algorithm which can optimize any differentiable loss function. An ensemble of trees are built one by one and individual trees are summed sequentially. Next tree tries to recover the loss . Gradient Boosting= Gradient Descent + Boosting.

**Extremely Randomized trees**

The Extra-Trees algorithm builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. It splits nodes by choosing cut-points fully at random and that it uses the whole learning sample to grow the trees.

**5.Neural Networks**

A multilayer perceptron (MLP) is a deep, artificial neural network. It is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP. MLPs with one hidden layer are capable of approximating any continuous function.

**7.2. Model performance:**

Model can be evaluated by various metrics such as:

1. MAE - Mean absolute error regression loss
2. MSE - Mean squared error regression loss
3. RMSE - is a square root of value gathered from the mean square error function

4. RMSPE - Root mean Square Percentage Error

5. R^2 - Coefficient of determination also called as R2 score is used to evaluate the performance of a linear regression model.

**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 Grid Search CV,This also results in cross validation and in our case we divided the dataset into different folds.

1. **Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

**8. Conclusion:**

That's it! We reached the end of our exercise. Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building.

 ExtraTree Regression provides the best accuracy of 58% without tuned parameter.

And there is no such improvement in accuracy score even after hyperparameter tuning.

If we drop the injested coloumns then the accuracy gets increased to 63% for tuned parameters.

So the accuracy of our best model is 63% which can be said to be good for this dataset. This performance could be due to various reasons like: no proper pattern of data, not enough relevant features.