Appliances Energy Prediction

Ajit Kumar Toppo
Data Science Trainee, Almabetter

Abstract

The purpose of this research is to forecast the electricity consumption of a particular household in Belgium based on the temperature and humidity levels of various rooms in the facility and surrounding weather information over 4.5 months.

Keywords: machine learning, regression, pandas, scikit-learn

Introduction

It is important to study the energy consuming behaviour in the residential sector and predict the energy consumption by home appliances as it consumes the maximum amount of energy in the residence. This project focuses on predicting the energy consumption of home appliances based on humidity and temperature.This project aims to predict the energy consumption of home appliances. With the advent of smart homes and the rising need for energy management, existing smart home systems can benefit from accurate prediction. Energy prediction is important to predict future energy energy needs to achieve demand and supply equilibrium, have control over cost, prevent energy wastage.

About Dataset

The data set is at 10 min for about 4.5 months. The houe 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. 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 year-month-day hour:minute:second.

Appliances energy use in Wh (Dependent variable).

lights energy use of light fixtures in the house in Wh.

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 the laundry room area.

RH3 Humidity in laundry room area, in %.

T4 Temperature in office room, in Celsius.

RH4 Humidity in the 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 the 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 %.

T_out Temperature outside (from Chievres weather station), in Celsius.

Press_mm_hg Pressure (from Chievres weather station), in mmHg.

RHout Humidity outside (from Chievres weather station), in %.

Windspeed (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.

Data Cleaning

In the 'lights' column mostly there are null values because of which this column will not add much value to our prediction so this column was dropped from the dataset. Outliers from the 'appliances' columns were removed. Dropped the column Date because it will be of no use here for prediction.

Data Visualization

In order to quickly visualize the distribution of data we plotted histogram for each variable after that we also plotted a heatmap to show correlation between all variables in our data. Almost all the temperature measures in different rooms are highly linearly correlated with each other. However there is little to no correlation between temperature features and target variables also there is little to no correlation between humidity features and target variables.

Standardize the dataset

Well, the idea is simple. Variables that are measured at different scales do not contribute equally to the model fitting & model learned function and might end up creating a bias. Thus, to deal with this potential problem, feature-wise standardized (μ=0, σ=1) is used prior to model fitting. From Python sklearn library, StandardScaler() function was used to standardize the data values into a standard format.

Splitting the dataset for training and testing purpose

We need to split the dataset into train and test sets to evaluate how well our machine learning model performs. The train set is used to fit the model and the second set is called the test data set, this set is solely used for prediction,testing accuracy.The scikit-learn library was used with the model_selection module in which we have the splitter function train_test_split() to split the data into 75% for the training of model and 25% for the testing of model.

Models used for training

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

Support Vector Regression (SVR) - Supervised Machine Learning Models with associated learning algorithms that analyze data for classification and regression analysis are known as Support Vector Regression. SVR is built based on the concept of Support Vector Machine or SVM.

K-nearest Neighbours regression - non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

Random Forest Regression - 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.

Gradient Boosting 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.

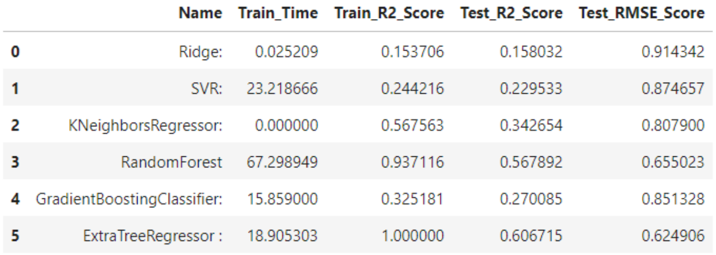
Extra Trees Regressor - This class implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Evaluation-metrics

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

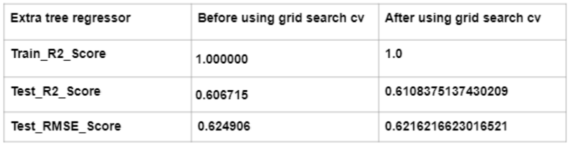
root-mean-square error (RMSE)- a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

Model Performance



	Name	Train_Time	Train_R2_Score	Test_R2_Score	Test_RMSE_Score
0	Ridge:	0.025209	0.153706	0.158032	0.914342
1	SVR:	23.218666	0.244216	0.229533	0.874657
2	KNeighborsRegressor:	0.000000	0.567563	0.342654	0.807900
3	RandomForest	67.298949	0.937116	0.567892	0.655023
4	GradientBoostingClassifier:	15.859000	0.325181	0.270085	0.851328
5	ExtraTreeRegressor :	18.905303	1.000000	0.606715	0.624906

Out of all the models used for regression Extra tree regressor is performing best for this dataset as we can see it has highest r2 score and lowest rmse score. So we performed hyperparameter tuning on the extra tree regression model by using grid search cv and the results obtained are shown in the table below.



Extra tree regressor	Before using grid search cv	After using grid search cv
Train_R2_Score	1.000000	1.0
Test_R2_Score	0.606715	0.6108375137430209
Test_RMSE_Score	0.624906	0.6216216623016521

Little or no improvement can be seen after hyperparameter tuning of the model.

Conclusion

Higher accuracy could not be achieved because of almost no correlation between the independent and the dependent variable. Still Extra tree regressor is the best model among all other models with a r2 score of 0.610 and rmse score of 0.621.

References

1.Kaggle
2.rpubs
3.uci machine learning repository