Machine Learning Engineer Nanodegree

Capstone Project Proposal

Appliance energy prediction

Amanchi Thulasi

February 21st, 2019

Proposal:

Appliance Energy prediction dataset

Domain Background:

History:

The background domain of this project is energy usage prediction inside a home. In the usual setting, different sensors are attached inside the home and the energy usage is also calculated. All readings are taken at regular intervals. The goal is to predict energy consumption.

In this era of smart homes, energy usage prediction can lead to efficient energy management.

Related academic research and earlier work:

In one article related to appliance energy prediction they used neural networks and statistical and artificial intelligence models for energy prediction, but am going to use the machine learning techniques to predict the energy appliances through some regressions techniques which gives more accuracy and simple to implement and less time consumption .

Another article I found that is similar to my work is :

https://www.sciencedirect.com/science/article/pii/S0378778816308970?via%3Dihub

Problem Statement:

Predict energy consumption of the appliances inside a home based on various parameters like temperature, pressure and humidity.

Datasets and Inputs:

The dataset that I am working is downloaded from

Link :- http://archive.ics.uci.edu/ml/machine-learning-databases/00374/

Dataset information :-

The dataset has 19,375 instances and 29 attributes including the predictors and target variable. The 29 attributes are described as follows:-

- date: year-month-day hour:minute:second
- T1: Temperature in kitchen area, in Celsius
- RH_1: Humidity in kitchen area, in %

- T2: Temperature in living room area, in Celsius
- RH_2: Humidity in living room area, in %
- T3: Temperature in laundry room area

- RH_3: Humidity in laundry room area, in %
- T4: Temperature in office room, in Celsius
- RH_4: Humidity in office room, in %
- T5: Temperature in bathroom, in Celsius
- RH_5: Humidity in bathroom, in %
- T6: Temperature outside the building (north side), in Celsius
- RH_6: Humidity outside the building (north side), in %
- T7: Temperature in ironing room, in Celsius
- RH_7: Humidity in ironing room, in %
- T8: Temperature in teenager room 2, in Celsius
- RH_8: Humidity in teenager room 2, in %
- T9: Temperature in parents' room, in Celsius
- RH_9: Humidity in parents' room, in %
- T_out: Temperature outside (from Chievres weather station), in Celsius
- Pressure: (from Chievres weather station), in mm Hg
- RH_out: Humidity outside (from Chievres weather station), in %
- Wind speed: (from Chievres weather station), in m/s
- Visibility: (from Chievres weather station), in km
- T_dewpoint: (from Chievres weather station), Â°C
- rv1: Random variable 1, non-dimensional
- rv2: Random variable 2, non-dimensional
- Lights: energy use of light fixtures in the house in Wh

Solution Statement:

To solve this problem, I will be using one or more regression and regularization methods covered in udacity machine learning.
The most common solution to such problems is the method of
Regression.

Some of the Regression methods are:
a. Linear Regression
b. Polynomial Regression
c. Regularization methods such as Ridge and Lasso Regression
Linear regression can be mathematically expressed as:
$Y = a_1x_1 + a_2x_2 + … + a_nx_n + b$ where,
Y = target variable, x_1, x_2 … x_n are the $\boldsymbol{n}$ attributes of data, a_1, a_2, … a_n are coefficients and b is the intercept.
Similarly, in Polynomial Regression, at least one of the attributes has a degree of more than 1.
In regularization methods, the coefficient values are penalized by adding them (L1

Regularization) or their squares (L2 regularization) to the loss function.
This problem can also be treated as multivariate time series prediction

Benchmark Model:

However, the problem lies in finding a dataset where the results are given in such a fashion which is easily comparable with our classification values. In datasets, it is intrinsically difficult to compare the scores given with our outputs. Therefore, we will use a simple algorithm like multiple linear regression as our benchmark model and try to improve upon its performance by using other algorithms like SVM with Radial Kernel, Random Forest, Gradient Boosting Machines (GBM) . Out of these four models, GBM was able to explain 78% of the variance in the training data
and 57% in test data according the R2 score.

Evaluation Metric:

Some common evaluation metrics for Regression analysis are:
a. Mean Absolute Error
b. Mean Squared Error
c. R2 Score

Project Design:

The general sequence of steps are as follows.

Data Visualization: Visual representation of data to find the degree of correlations between predictors and target variable and find out correlated predictors. Additionally, we can see ranges and visible patterns of the predictors and target variable.

b. **Data Preprocessing:** Scaling and Normalization operations on data and splitting the data in training, validation and testing sets.

c. **Feature Engineering:** Finding relevant features, engineer new features using methods like PCA if feasible.

d. **Model Selection:** Experiment with various algorithms to find out the best algorithm for this use case.

e. **Model Tuning:** Fine-tune the selected algorithm to increase performance without overfitting.

f. **Training and testing the data with comparison of benchmark model:** Test the model on testing dataset.

Here I will use ensemble based Tree Regression models random forest ,gradient boosting ,extra trees by using R2 score , MAE and MSE for my project.

Finally, I will declare the model with highest accuracy score as the best model for appliance energy prediction