MACHINE LEARNING CAPSTONE PROPOSAL

Title: Appliance Energy Prediction

Domain Background: The background domain of this project is energy usage prediction inside a home.The goal is to predict the electricity usage of heating and cooling appliances in a household based on internal and external temperatures and other weather conditions. Each observation measures electricity in a regular interval. The temperatures and humidity have been averaged for 10-minute intervals.

Related research and previous work: **http://dx.doi.org/10.1016/j.enbuild.2017.01.083**

(Research Paper) **https://github.com/LuisM78/Appliances-energy-prediction-data**

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

Datasets and Inputs link: **http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction** (The author has provided here a separate training and testing data files)

Dataset Information: The dataset has 19,375 instances and 29 attributes including the predictors and target variable. The training data provided by author contains 14803 instances and testing data contains 4932 instances. The 29 attributes are described as follows:-

date: year-month-day hour:minute:second

1) T1: Temperature in kitchen area, in Celsius
2) RH_1: Humidity in kitchen area, in %
3) T2: Temperature in living room area, in Celsius
4) RH_2: Humidity in living room area, in %
5) T3: Temperature in laundry room area
6) RH_3: Humidity in laundry room area, in %
7) T4: Temperature in office room, in Celsius
8) RH_4: Humidity in office room, in %
9) T5: Temperature in bathroom, in Celsius
10) RH_5: Humidity in bathroom, in %
11) T6: Temperature outside the building (north side), in Celsius
12) RH_6: Humidity outside the building (north side), in %
13) T7: Temperature in ironing room, in Celsius
14) RH_7: Humidity in ironing room, in %
15) T8: Temperature in teenager room 2, in Celsius

16) RH_8: Humidity in teenager room 2, in %
17) T9: Temperature in parents’ room, in Celsius
18) RH_9: Humidity in parents’ room, in %
19) T_out: Temperature outside (from Chievres weather station), in Celsius
20) Pressure: (from Chievres weather station), in mm Hg
21) RH_out: Humidity outside (from Chievres weather station), in %
22) Wind speed: (from Chievres weather station), in m/s
23) Visibility: (from Chievres weather station), in km
24) T_dewpoint: (from Chievres weather station), Â°C
25) rv1: Random variable 1, non-dimensional
26) rv2: Random variable 2, non-dimensional
27) Lights: energy use of light fixtures in the house in Wh
28) Appliances: energy use in Wh (Target Variable)

Where indicated, hourly data (then interpolated) from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis, rp5.ru. Permission was obtained from Reliable Prognosis for the distribution of the 4.5 months of weather data.

Solution Statement: The 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 = a1x1 + a2x2 + … + anxn + b where,

Y = target variable, x1, x2, … xn are the *n* attributes of data, a1, a2, … an 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 Models:

As well I understand the term “benchmark”, it means a standard against which I can compare my own solution. As I will be trying various algorithms and techniques, the benchmark for me will be around the accuracy achieved by the author in his work.The author of the dataset used 4 models in his research which are listed below: a. Multiple Linear Regression

b. SVM with Radial Kernel

c. Random Forest

d. Gradient Boosting Machines (GBM)

The GBM had a R2 score of 0.97 and RF of 0.92 in the training set. For the testing set the R2 score for GBM was 0.58. As I will be trying various models and techniques, I think achieving an accuracy of above 80% in training and above 50% in testing data will be a good benchmark for me.

Evaluation Metrics : The metrics commonly used to evaluate regression models are:

1) Mean absolute error
2) Mean squared error (MSE)
3) Root Mean Squared error (RMSE)
4) R2 score

Project Design :

I will follow the following sequence of steps.

1. **Data Visualization**: Visually representing data to find correlations between attributes and target variable. This will also help in finding visible patterns in the dataset.

2. **Preprocessing the data**: Scaling and Normalizing the data. As we will use the data provided , I will have to split the data into training, validation and testing datasets.

3. **Feature Engineering**: Finding relevant features, engineering new features, etc.

4 **Model Selection**: Experiment with various models and algorithms to find the best one. Ex. Multi-layer perceptron which has not been used by the author. Others are of course regression techniques, SVM and gradient boosting.

5. **Model Tuning**: Fine tune the selected algorithm to increase performance while making sure it does not overfit.

6**. Testing**: Test the model on the testing dataset.