Machine Learning Nanodegree – Capstone Project

Application Energy Prediction

Domain Background :

In this time of global uncertainty one thing is clear the world needs energy -- and in increasing quantities to support economic and social progress and build a better quality of life, in particular in developing countries. But even in today's time there are many places especially in developing world where there are outages .

These outages are primary because of excess load consumed by appliances at home . Heating and cooling appliances takes most power in house. In this project we will be analyzing the appliance usage in the house gathered via home sensors .All readings are taken at 10 mins intervals for 4.5 months . The goal is to predict energy consumption by appliances .

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).

Related academic research and earlier work:

http://dx.doi.org/10.1016/j.enbuild.2017.01.083
https://github.com/LuisM78/Appliances-energy-prediction-data

Problem Statement :

We should predict Appliance energy consumption for a house based on factors like temperature , humidity & pressure .

Dataset & Inputs :

The dataset has been taken from UCI Machine Learning repository.

Dataset link -
http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

Data Set Information:

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). The dataset has 19375 instances and 29 attributes including predictors and target variable. The training data provided by author contains 14803 instances and testing data contains 4932 instances.

Attribute Information:

1. date time year-month-day hour:minute:second
2. Appliances, energy use in Wh
3. lights, energy use of light fixtures in the house in Wh
4. T1, Temperature in kitchen area, in Celsius
5. RH_1, Humidity in kitchen area, in %
6. T2, Temperature in living room area, in Celsius
7. RH_2, Humidity in living room area, in %
8. T3, Temperature in laundry room area
9. RH_3, Humidity in laundry room area, in %
10. T4, Temperature in office room, in Celsius
11. RH_4, Humidity in office room, in %
12. T5, Temperature in bathroom, in Celsius
13. RH_5, Humidity in bathroom, in %
14. T6, Temperature outside the building (north side), in Celsius
15. RH_6, Humidity outside the building (north side), in %
16. T7, Temperature in ironing room , in Celsius

17. RH_7, Humidity in ironing room, in %
18. T8, Temperature in teenager room 2, in Celsius
19. RH_8, Humidity in teenager room 2, in %
20. T9, Temperature in parents room, in Celsius
21. RH_9, Humidity in parents room, in %
22. To, Temperature outside (from Chievres weather station), in Celsius
23. Pressure (from Chievres weather station), in mm Hg
24. RH_out, Humidity outside (from Chievres weather station), in %
25. Wind speed (from Chievres weather station), in m/s
26. Visibility (from Chievres weather station), in km
27. Tdewpoint (from Chievres weather station), Â°C
28. rv1, Random variable 1, nondimensional
29. rv2, Random variable 2, nondimensional

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 :

Regression is used for problems like this . 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 Regression : In linear regression we wish to fit a function in this form
   $\hat{Y} = \beta0+\beta1X1+\beta2X2+\beta3X3$ 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. Once we reach the minimum point of the loss function we can say that we completed the iterative process and learned the parameters.

2. Logistics Regression (If target variable is categorical) : Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.

3. Polynomial Regression : A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation: $\hat{Y} = \beta0+\beta1X1^2$

4. Ridge regression : Regularized machine learning model, in which model's loss function contains another element that should be minimized as well. $L = \sum(\hat{Y}i– Yi) ^2 + \lambda\sum \beta2$. 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

5. Lasso regression : Lasso is another extension built on regularized linear regression . The loss function of Lasso is in the form: $L = \sum(\hat{Y}i– Yi) ^2 + \lambda\sum |\beta|$. The only difference from Ridge regression is that the regularization term is in absolute value.

Benchmark Models :

The author has used following models in his research

1. Linear Regression (Multivariate)
2. SVM
3. Random forest
4. Gradient boosting machine (GBM)

Out of these models used in research , GBM was able to predict 97% variance in testing set and 57% in training data as per R2 scores published by the author

Evaluation Metrics :

The regression metrics used as standards to measure regression models are

1. Mean Absolute Error
2. Mean Squared Error
3. Root Mean Squared Error
4. R Squared (R2)

Project Design :

The steps to be followed are mentioned below:

1. **Data Visualization** : Visual plots to detect the correlation between different independent variables and between independent and dependent variables . Ranges and other statistical data can also be verified

2. **PrePocessing** : In this process we will be organizing and tidying up the data, removing what is no longer needed, replacing what is missing and standardizing the format across all the data collected.

3. **Feature Engineering :** Find all the features which impacts the models and reduce the number of features if possible using PCA

4. **Choosing a Model :** Check all the applicable models and select the one which provides best metrics .

5. **Hyperparameter Tuning** : Find best possible combination of selected algorithm in order to maximize the performance using Grid Search

6. **Prediction :** Using Test set predict the dependent variable and check accuracy