Capstone Proposal

Kumar Anurag

Appliances Energy Prediction

Domain Background

With the advent of smart homes and rising need for energy management, existing smart home systems can benefit from accurate prediction. If the energy usage can be predicted for every possible state of appliances, then device control can be optimized for energy savings as well. So, this project presents and discusses the data-driven predictive models along with the data filtering to remove non-predictive parameters and feature ranking.

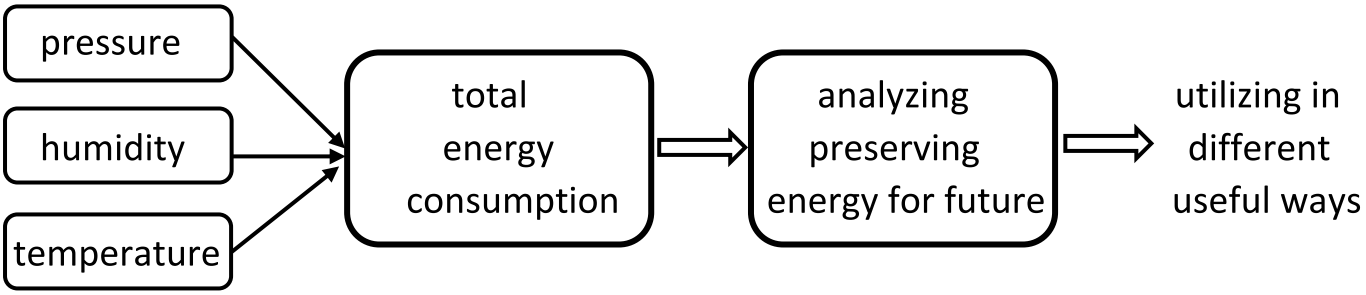
Problem Statement

Based on some parameters like pressure, humidity and temperature,

(i) calculating the total energy usage of electrical appliances used in homes and

(ii) then analyzing and preserving the extra unnecessary used energy for future generations and

(iii) finally suggesting the ways to utilize the preserved energy in different ways


pressure
humidity
temperature
total energy consumption
analyzing preserving energy for future
utilizing in different useful ways

Datasets and Inputs

The dataset is obtained from UCI Machine Learning repository. Luis Candanedo donated this repository. His research paper and GitHub repository demonstrating his work can be viewed from the links as follows:

Paper: https://www.sciencedirect.com/science/article/pii/S0378778816308970?via%3Dihub [1]
GitHub: https://github.com/LuisM78/Appliances-energy-prediction-data [2]
Dataset: http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction [3]

__Dataset Description__: Dataset is having 25 attributes and 19,375 instances including the target variable and predictors. The 25 attributes are described as follows: -

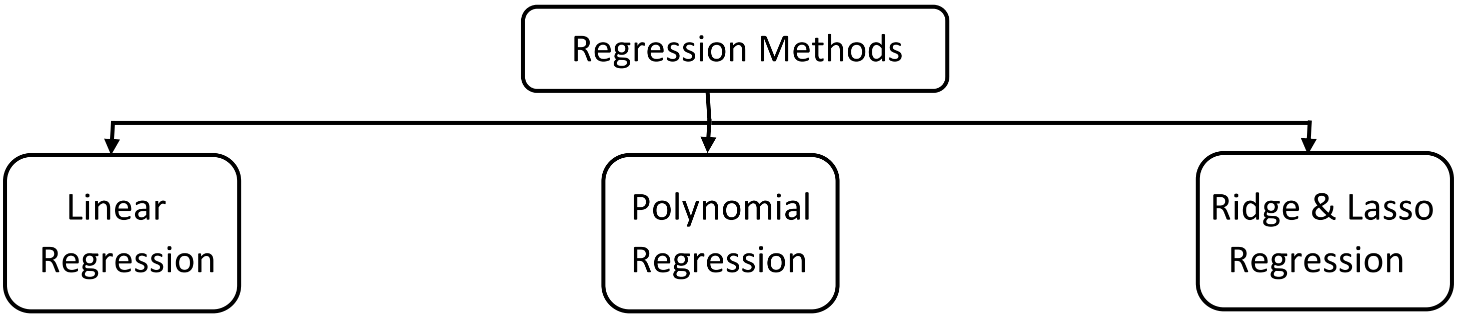
S.No.	Attributes	Description
1	date	year-month-day hour:minute:second
2	T1	Temperature in kitchen area, in Celsius
3	RH_1	Humidity in kitchen area, in %
4	T2	Temperature in living room area, in Celsius
5	RH_2	Humidity in living room area, in %
6	T3	Temperature in laundry room area

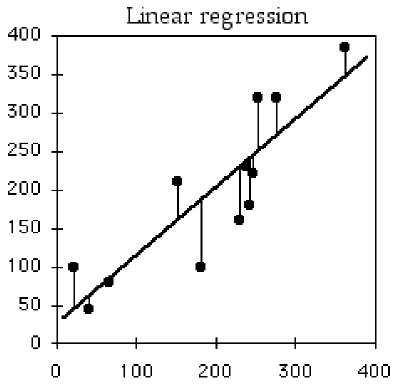
7	RH_3	Humidity in laundry room area, in %
8	T4	Temperature in office room, in Celsius
9	RH_4	Humidity in office room, in %
10	T5	Temperature in bathroom, in Celsius
11	RH_5	Humidity in bathroom, in %
12	T6	Temperature outside the building (north side), in Celsius
13	RH_6	Humidity outside the building (north side), in %
14	T7	Temperature in ironing room, in Celsius
15	RH_7	Humidity in ironing room, in %
16	T_out	Temperature outside (from Chievres weather station), in Celsius
17	RH_out	Humidity outside (from Chievres weather station), in %
18	Pressure	(from Chievres weather station), in mm Hg
19	Wind Speed	(from Chievres weather station), in m/s
20	Visibility	(from Chievres weather station), in km
21	T_dewpoint	(from Chievres weather station), Â°C
22	RV1	Random variable 1, non-dimensional
23	RV2	Random variable 2, non-dimensional
24	Lights	energy use of light fixtures in the house in Wh
25	Appliances	energy use in Wh (Target Variable)

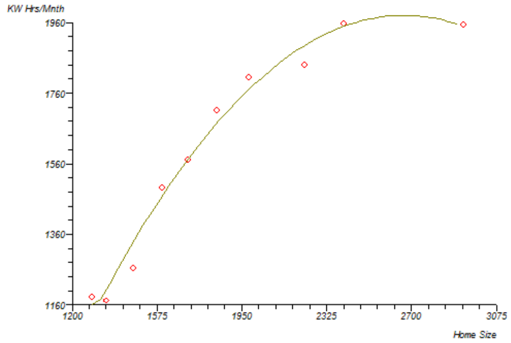
Hourly data, gathered from airport weather station (Belgium, Airport) was downloaded from a public data-set, from Reliable Prognosis, rp5.ru. Permission was obtained from Reliable Prognosis for the distribution of the 4.5 months of weather data. (ref in [3]).

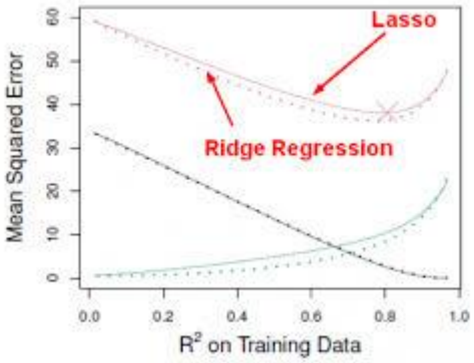
Solution Statement

The most common solution to such problems is the method of Regression. Some of the Regression methods are as follows:









Linear Reg Equation

$$\overset{\text{Dependent Variable}}{Y_i} = \underbrace{\overset{\text{Population Y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1} \overset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

Polynomial Reg Equation

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \text{----} + \beta_n x^n$$

Lasso & Ridge Reg Equation

$$Cost\,(W) = RSS(W) + \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

Benchmark Models

Four statistical models were trained with repeated cross validation and evaluated in a testing set: -

i) Multiple Linear Regression

ii) Gradient Boosting-Machine

iii) Random Forest Classifier

iv) Support Vector Machine with Radial-Kernel

The best model (GBM), among above four models, was able to explain 97% of the variance (R^2) in the training set and with 57% in the testing set when using all the predictors.

Evaluation Metrices

For the regression analysis, here are some of the common evaluation metrices:

i) Mean Absolute Error

ii) Variance R^2 Score:

$$R^2 \equiv 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}$$

where,
SS_{res} = Residual sum of squares
SS_{tot} = Total sum of squares

iii) Mean Squared Error:

$$\sqrt{\frac{1}{N} * \sum_{i=1}^{N} (y_i - y_i')^2}$$

where,
N = number of observations
yi = actual value of target variable
yi' = predicted value of target variable

Project Design

- Data-Visualization: In this, data represented visually and correlation degree is to be find out between predictors and target variable and finally correlated predictors.
- Data-Preprocessing**:** Data is to be operated by scaling and normalization. Also, it is splitted into training, testing and validation sets.
- Feature Engineering: Engineer new features using methods like PCA if feasible.
- Model Selection: Results of mentioned algorithms in terms of accuracy score, is to be consider for selecting the best algorithm.
- Model Tuning: Tuning of the selected algorithm for enhancing the performance without overfitting.
- Testing: Here, testing of model on datasets is to be done.