MACHINE LEARNING CAPSTONE PROPOSAL

Title: Appliance Energy Prediction

1. **Domain Background:** 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:

   http://dx.doi.org/10.1016/j.enbuild.2017.01.083 [1]
   https://github.com/LuisM78/Appliances-energy-prediction-data [2]

2. **Problem statement:** We have a problem of predicting appliances energy consumption of a house with 9 rooms. In order to perform a prediction, we will be utilizing room's temperature and humidity along with weather details such as temperature, visibility, pressure, wind speed etc. Since there is no fixed rule for figuring it out, we will be using a Machine learning approach by training the system with a training data set.

3. **Datasets and Inputs:** The dataset is obtained from UCI Machine Learning repository. It is donated by Luis Candanedo. His research paper and GitHub repository demonstrating his work can be viewed from links [1] and [2] respectively.

   Dataset link:- http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction [3]
   Luis Candanedo LinkedIn Profile:- https://www.linkedin.com/in/luis-miguel-candanedo-7b8b2a11/ [4]
   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
- **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. (cited in [3]).

4. **Solution Statement:** 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:
$Y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$ where,
Y = target variable, $x_1, x_2, \ldots x_n$ are the ***n*** attributes of data, $a_1, a_2, \ldots a_n$ 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.

5. **Benchmark Models:** 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)

    Out of these 4 models, GBM was able to explain 97% of the variance in the training data and 57% in test data according the R2 score. (Training and testing data as used in author's GitHub repository [2])
    Links:
    Training data - https://github.com/LuisM78/Appliances-energy-prediction-data/blob/master/training.csv
    Testing data - https://github.com/LuisM78/Appliances-energy-prediction-data/blob/master/testing.csv

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

7. **Project Design:** The general sequence of steps are as follows-
    a. **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. **Testing:** Test the model on testing dataset.