MACHINE LEARNING CAPSTONE PROPOSAL

Title:- Appliances Energy-Prediction

1. **Domain Background:** This project presents and discusses data-driven predictive models for the energy use of appliances. Data used include measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station and recorded energy use of lighting fixtures. This project also discusses data filtering to remove non-predictive parameters and feature ranking. In this epoch of smart-homes, energy-consumption prediction can lead to efficient energy-management. Our mission is to predict the energy consumption so that it can be useful to manage energy for future generations.

2. **Problem statement:** Prediction of energy usage of the electrical appliances used in homes based on some parameters such as pressure, humidity and temperature.

3. **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:

   http://bit.ly/AppliancesEnergyPredictionResearchModel [1]
   http://bit.ly/LuisCandanedo [2]

   Dataset link:- http://bit.ly/EnergyPredictionDataset [3]

   Dataset description :-
   This dataset is having 29 attributes and 19,375 instances including the target variable and the predictors. The 29 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	T8	Temperature in teenager room 2, in Celsius
17	RH_8	Humidity in teenager room 2, in %
18	T9	Temperature in parents' room, in Celsius
19	RH_9	Humidity in parents' room, in %
20	T_out	Temperature outside (from Chievres weather station), in Celsius
21	Pressure	(from Chievres weather station), in mm Hg
22	RH_out	Humidity outside (from Chievres weather station), in %
23	Wind speed	(from Chievres weather station), in m/s
24	Visibility	(from Chievres weather station), in km
25	T_dewpoint	(from Chievres weather station), Â°C
26	rv1	Random variable 1, non-dimensional
27	rv2	Random variable 2, non-dimensional
28	Lights	energy use of light fixtures in the house in Wh
29	Appliances	energy use in Wh (Target Variable)

Hourly data is gathered from 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. (ref in [3]).

3. **Solution Statement:** The most common solution to such problems is the method of Regression. Some of the Regression methods are:
   i. Polynomial-Regression
   ii. Linear-Regression
   iii. Ridge and Lasso Regression (Regularization methods)

Linear regression can be expressed mathematically as:
$Y = B + A_nX_n + A_{n-1}X_{n-1} + \text{-----} + A_2X_2 + A_1X_1$, where,
Y = target-variable, X_n, X_{n-1}, … X_1 are the ***n*** attributes of data, A_n, A_{n-1}, … A_1 are coefficients and B is the intercept.
Similarly, one of the attributes has degree at least more than 1 in Polynomial-Regression.
For regularization-methods, the coefficient values are modified by adding them (L1-Regularization) or their squares (L2-regularization) to their loss-functions.
Also, this can be solved by multivariate-time-series prediction.

5. **Benchmark Models:** Four statistical models were trained with repeated cross validation and evaluated in a testing set:
    i. Multiple Linear Regression - (MLR)
    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.

6. **Evaluation Metrics:** For the regression analysis, here are some of the common evaluation metrices:
    i. Mean-Squared Error
    ii. Variance R^2 Score
    iii. Mean-Absolute Error

7. **Project Design:** Steps are as follows:
    i. **Data-Visualization:-** In this, data represented visually and correlation degree is to be find out between predictors and target variable and finally correlated predictors. Moreover, we can also analyze visible patterns and ranges of the target variable and predictors.
    ii. **Data-Preprocessing:-** Data is to be operated by scaling and normalization. Also, it is splitted into training, testing and validation sets.
    iii. **Model-Selection:-** Results of mentioned algorithms in terms of accuracy score, is to be consider for selecting the best algorithm.
    iv. **Model-Tuning:-** Tuning of the selected algorithm for enhancing the performance without overfitting.
    v. **Testing:-** Here, testing of model on datasets is to be done.