MACHINE LEARNING PROJECT

CAPSTONE PROPOSOL

Title: Appliance Energy Prediction

Author: venkat

1.Domain Background: The background domain of this project is energy usage prediction inside house. We all are observes that usually in houses we set up different sensors to calculate the energy consumption. Actual in our home all readings are taken at regular intervals. Here by this project the main motto is to predict the energy consumption.

At present in the world, coming to the concept of smart homes energy management with efficient is play a lead role.

Actually the energy consumption from day to day life is increases and in the same scenario many companies provide some electronic equipments also to predict the energy consumption. and here I predict and reduce this problem of the energy consumption by using the supervised learning.

Reference:https://www.sciencedirect.com/science/article/pii/S0378778816300305

2.Problem statement: Here in this the prediction of the energy consumption of the appliances inside a house which is based on the certain parameters like pressure, temperature in a room and humidity. By reducing consumption of the energy, it is better to save more and it is useful to the future generation.

3.Datasets and inputs:

The data is obtained from the UCI machine learning repository and it is donated by the luis candanedo.

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

Dataset information:

The dataset has 19,375 instances and 29 attributes including the predictors and the target variables. And I here take all the features to predict more accurately and make correct predictions. The 29 attributes are explained below

- 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, in Celsius
- RH_6: Humidity outside the building, 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, in Celsius

- Pressure: in mm Hg
- RH_out: Humidity outside, in %
- Wind speed: in m/s
- Visibility: in km
- T_dewpoint: Â°C
- rv1: Random variable 1, non-dimensional
- rv2: Random variable 2, non-dimensional
- Appliances: energy used in wh, it is the target variable
- Lights: energy use of light fixtures in the house in Wh

All above indicated hourly data climate conditions are collected from the chievres weather station it is the airport whether station.

4.Solution statement: we can use many methods to predict this data but regression is the best model among all those methods. And below are the some of the regression methods are:

1. Linear regression
2. Polynomial regression
3. Lasso regression and extra tree regression

And here the linear regression is mathematically represented as a Line equation y = mx +c

Where y is the target variables, m is the coefficient and 'c' is the intercept

Y=m1x1+m2x2+……+m(n-1) x(n-1) +mn xn +c

It is the linear regression equation for the more features
In the same scenario polynomial has at least one of the attributes has a degree of more than the 1. In regularization methods, the coefficient values are penalized by adding them or their squares to the loss function.

In this problem my solution will use the features like as declared in the above in the dataset inputs and by using the supervised model I will predict the output as the accuracy or r2_score of that features help in reducing the power consumption of a house. Here we have to train the data which does not make

any overfitting. And the main metrics will use in this are r2_score,mean squared error and variance and these metrics will decide which model is best.

5.Benchmark models: There are different models to predict the data here in the supervised learning they are

1. Decision trees
2. Multiple linear regression
3. Svm
4. Ensemble models (random forest, adaboost)
5. Gradient boosting machines(GBM)

In all of the above models I will go through the linear regression and calculate the score of that model, after selecting the model we will train/test the data on that model and find the scores,if it gives the best score I will decide it is the best model if not I check with the another model and I compare all the models with their scores respectively .

Suppose my benchmark model is linear regression and after train and testing the data and after that I will do the same procedure to the random forest if it give the more score than my bench mark model then i can decide it is best model than my bench mark model and it will predict the data more accurately than my bench mark model.

References:
Training data - https://github.com/LuisM78/Appliances-energypredictiondata/blob/master/training.csv
Testing data - https://github.com/LuisM78/Appliances-energypredictiondata/blob/master/testing.csv

6.Evaluation Metrics: As we made predictions on the data we have to make some common evaluation metrics for regression are

1. Mean absolute error
2. Mean squared error
3. R2_score

By all of these metrics we can decide the which will be best and suitable for the data and which give accuracy score more.

And I think that by the evalution metrics we can decide our model which may be GBM also as I declared as my opinion.

7.Project design:

Manually some of the steps be follow to predict data and workflow is given below:

1. Data visualization: By the visualizing of the data we can find the degree of the correlation between the predictors and target variables .in this by the data visualization we can also see ranges and the patterns of the target and predictor variables.
2. Data Preprocessing: Preprocessing means attain some operations normalization and the scaling and splitting data which is in the training validation and testing sets.
3. Feature engineering: here is the responsibility to the featuring engineering is to find the correct and relevant features if we want we also drop that feature.
4. Model selections: as I said before we have to made some experiments to find the best algorithm from the various algorithms.
5. Model tuning: After finding the best model we have to tune the model with our data sets to increase the performance without overfitting.
6. Testing: Finally we find the testing the model based on the testing data set.

Workflow:

Initially load the data set and the data set contain different number of columns and rows with a different features. After loading the data we have to check is there any null values present in the data set of each column if there any null values then we will drop that column otherwise we will continue. Here is the main thing to remember when we make a prediction is we have pure knowledge on the dataset and observe the data clearly . here by only observing the data it will solve the some prediction and if there any outliers we have to remove. Now visualise the pure data by using plots like histogram and make some intuition about the data among the correlation of the data, after the visualization we have to take the best features which are better to predict the data and drop the unused features

and then predict the data with score to our bench mark model and to train and find the scores we can perform cross validation and hyper parameter tuning of the model. And after that as I said we will make the scores of some more models and then visualize the performance of the model. Finally give the inputs to the data and it will give the output by the training and testing of the data sets.