Applied AI - Project 5 CS-514

Project Report: ASHRAE - Great Energy Predictor III

Introduction:

In this project, the improvements of energy consumption is focussed with the ASHRAE - Great Energy Predictor III dataset obtained from kaggle platform. The lightGBM model (a Microsoft open source library) is constructed on this dataset to predict the meter_reading. At first to get the important predictor variables, the exploratory data analysis part is executed. After gathering enough information, the model is constructed and the dataset is given to train the model. This model gets trained and the dataset is given for prediction. The resultant prediction values is then exported in the required format.

Handling missing values:

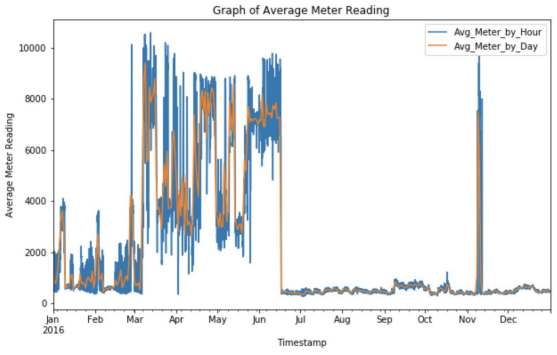
Firstly, the dataset is fed into the notebook using pandas library. The dataset is then checked for its values using '.info()' methods and '.isna()' methods. The site_id and timestamp do not have any missing values. But the variables such as 'precip_depth_1_hr' and 'cloud_coverage' have missing values in them. Since 'floor_count' has a larger number of missing values, this column is dropped from the dataset. The missing values are estimated with the mean values.

Exploratory data analysis:

Using the pandas library again, the dataset is explored using '.describe()' method and the values of each variable is analysed. Here, one data preparation task is done where the variable 'meter' is slightly modified with named values such as 'Electricity', 'Chilled Water', 'Steam', 'Hot Water'. This modified column is used to get the number of unique building types with the distribution of meter id.

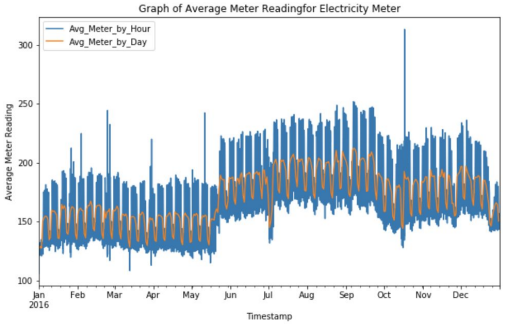
a). Average consumption by day and hour

We can see some surprising trends here, the meter reading is low from Jan to March, however from March it shoots up until mid June, then it almost reaches 0
till Mid november and then briefly shoots up again and then drops to zero.



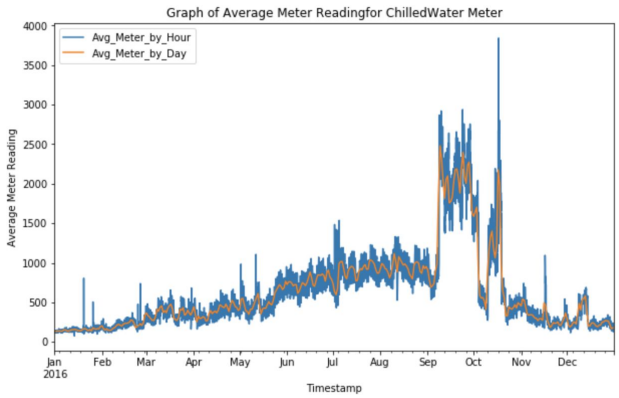
b). Average Meter Reading for Electricity meter

The increase and decreasing trend can be attributed to the usage during the weekdays and during the weekends when it drops.



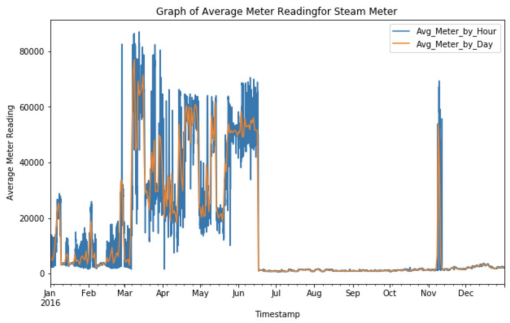
c). Average consumption for ChilledWater

Consumption gradually increases and reaches its peak during septembet to November months.



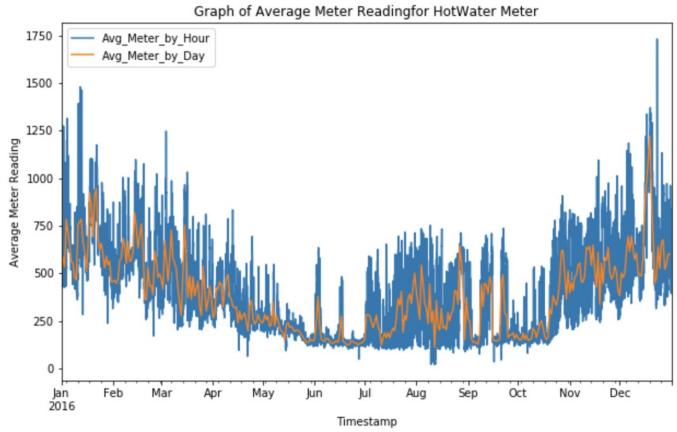
d). Average Meter Reading for Steam Water

This is almost similar to that of the overall trend.



e). Average meter reading for Hot Water meter

Hot water meter reading is high during the winter months and reduces during the summer months.



Finding Important Variables:

The important columns are drawn using the correlation matrix and the threshold value 0.9. If the threshold value is less than the correlation value of the variable and the meter reading then the corresponding variable is dropped. Also, the important variables are proofread using the regression feature importance variables (LightGBM model).

Feeding the dataset to LightGBM model:

The dataset is split using the train_test_split method with the test size as 0.25. Later, the split train and test datasets are given to the 'lgb.train()' method and the dataset is trained. The root mean square error value gets reduced at the 3000th level.

Prediction:

The ‘prediction.extend()’ method is used to predict the meter_reading for the test dataset with the step count as 10000. This process takes a CPU time of 5hours 7mins and 33s, wall time as 44min and 19s. Also, the required format is printed in the required format using the ‘to_csv()’ method which took a CPU time of 2min 5s and it’s the same for wall time as well.

```
%%time
prediction = []
step = 100000
for i in range(0, len(test), step):
    prediction.extend(np.expm1(reg.predict(test.iloc[i: min(i+step, len(test)), :], num_iteration=reg.best_iteration)))
```

```
CPU times: user 5h 7min 33s, sys: 1min 43s, total: 5h 9min 16s
Wall time: 44min 19s
```

1). Prediction Execution

```
%%time
Submission['meter_reading'] = prediction
Submission['meter_reading'].clip(lower=0,upper=None,inplace=True)
Submission.to_csv("Twentysix.csv",index=None)
```

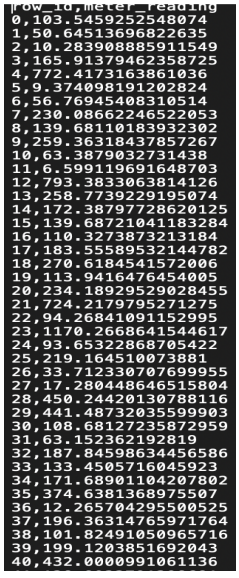
```
CPU times: user 2min 5s, sys: 13 s, total: 2min 18s
Wall time: 2min 23s
```

2). Output print Execution

Instructions to run:

1. Install python version 3 on your Mac/Windows operating system.
2. Install the required packages using ‘pip install’.
3. The required packages are listed below
   ['builtins','builtins', 'os', 'numpy', 'pandas', 'matplotlib.pyplot', 'seaborn', 'datetime', 'gc','lightgbm', 'pip', 'types'].
4. Run the python file from command line/Jupyter Notebook/Pycharm IDE

Output:



```
row_id,meter_reading
0,103.5459252548074
1,50.64513696822635
2,10.283908885911549
3,165.91379462358725
4,772.41731638610 36
5,9.374098191202824
6,56.76945408310514
7,230.08662246522053
8,139.68110183932302
9,259.36318437857267
10,63.3879032731438
11,6.599119691648703
12,793.3833063814126
13,258.7739229195074
14,172.38797728620125
15,139.68721041183284
16,110.3273873213184
17,183.55589532144782
18,270.6184541572006
19,113.9416476454005
20,234.18929529028455
21,724.2179795271275
22,94.26841091152995
23,1170.2668641544617
24,93.65322868705422
25,219.164510073881
26,33.712330707699955
27,17.280448646515804
28,450.24420130788116
29,441.48732035599903
30,108.68127235872959
31,63.152362192819
32,187.845986344565 86
33,133.45057160459 23
34,171.68901104207802
35,374.63813688975507
36,12.265704295500525
37,196.36314765971764
38,101.82491050965716
39,199.1203851692043
40,432.0000991061136
```

3). Sample Output