FUSEMACHINES AI CENTER
Kathmandu, Nepal

A project report
on

ASHRAE – Great Energy Predictor III

A project report submitted for the partial fulfillment for the machine learning module of Micro-degree in Artificial Intelligence

SUBMITTED BY

Jakesh Bohaju
Salit Dangol

UNDER SUPERVISION OF

T.A Mahendra Thapa

16th January 2020

ABSTRACT

American Society of Heating, Refrigerating and Air-Conditioning Engineers, ASHRAE lunch the third great energy predictor challenge via kaggle with evaluation metric root mean square logarithmic error. The competition provide five dataset train,test, weather_train, weather_test and building_metadata. The dataset has huge amount of missing data so some of them are dropped and some are filled with median of respective feature. Converted timestamp data to hour, day and month, target values into logarithmic value using log1p() function. Split data into training and validation of ratio 4:1. Then light gradient boosting is applied to find out the important feature among the features. With the resultant features training carried out with two model linear regression and light gradient boosting. Evaluate both model for validation set of data and finally prediction carried out for the test data, generate submission data and submitted to the kaggle for the metric evaluation as per mentioned.

INTRODUCTION

ASHRAE, founded in 1894, is a global society advancing human well-being through sustainable technology for the built enviromnent.[1] ASHRAE is working on building systems related to the energy efficiency. ASHRAE – Great Energy Predictor III is the third competition organized by the ASHRAE , Dr. Miller present the plans for the third competition in 30th September 2019.[2] Initially they launch two competition before named as the ASHRAE Predictor Shootout I (1993) and Predictor Shootout II (1995) competitions and publish six winning teams.[3] A question “How much does it cost to cool a skyscraper in the summer?” with answer “A lot! And not just in dollars, but in environmental impact.”[4] at kaggle description of the competition which clearly shows that today’s tall building needed high amount of energy for various purpose and mention the fact about environmental impact for cooling those skyscraper.

Dataset Description

Provide five files of dataset with a sample submission csv file.

i. train.csv
- building_id – foreign key for building metadata
- meter – meter id code. Four type of meter {0: electricity, 1: chilledwater, 2: steam, 3: hotwater}
- timestamp – time of measurement taken
- meter_reading – target variable

ii. building_meta.csv
- site_id – foreign key for weather file
- building_id – foreign key for test and train
- primary_use – category of building type
- square_feet – gross floor area of building
- year_built – year building was opened
- floor_count – no of floor of the building

iii. weather_[train/test].csv
- site_id – location id
- air_temperature – degree celsius
- cloud_coverage – portion of sky covered by cloud
- dew_temperature – degree celsius
- precip_depth_1_hr – milimeter
- sea_level_pressure – millibar/hectopascals

- wind_direction – compass direction (0-360)
- wind_speed - meters per second

iv. test.csv

- row_id – row id for submission file
- building_id – foreign key for building metadata
- meter – meter id code
- timestamp – timestamp for test data period

Problem Statement and Motivation

we address most of people focus on neural network as thinking it give better result than using machine learning model; here implement linear regression and light gbm to evaluate the performance in machine learning model.

Light bgm, a fast learner, less memory usage and better accuracy is applied to extract important features among the selective features and train and predict using two different ml model.

LITERATURE REVIEW

More than 3500 participate on the competition where most of them apply categorical gbm and light gbm with neural network including leak data.

Isamu & Matt, 1st place of the competition remove anomalies, timezone correction and imputing temperature during preprocessing with leak data for site 0, 1, 2, 4 & 15 and apply model CatBoost, LGBM & MLP. Then ensembling prediction with generalized weighted mean for log(pred).[5]

Vopani & team, 2nd place of the competition. They apply preprocessing for site 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and leak data for site 0, 1, 2, 4, 15. Apply XGBoost + LGBM + CatBoost + Feed-forward Neural Network for without leak data and leak data use as validation. Then ensembling with weighted average for each site meter.[6]

Eagle4, 3rd place of the competition. He did preprocessing and feature engineering and apply keras CNN, light gbm & catBoost for training, then ensembling.[7]

MPWARE, 9th place on leaderboard, he also use leaked data mainly for hold out validation. He started training without leak data and different model initially. He also ensemble several models; light gbm, CatBoost, neural network and LiteMort as machine learning model.[8]

Georgi Pamukov, 25th place on the leaderboard. He apply 2 layer learning architecture: base layer and ensemble. After training 19 models he get the better score. He use deep neural network and light gbm at base layer and ensemble with leak data.[9]

From these, its clear that most of the all implement boosting algorithm with neural network. And mostly choose boosting framework is categorical boosting framework.

METHODOLOGY

Project Workflow

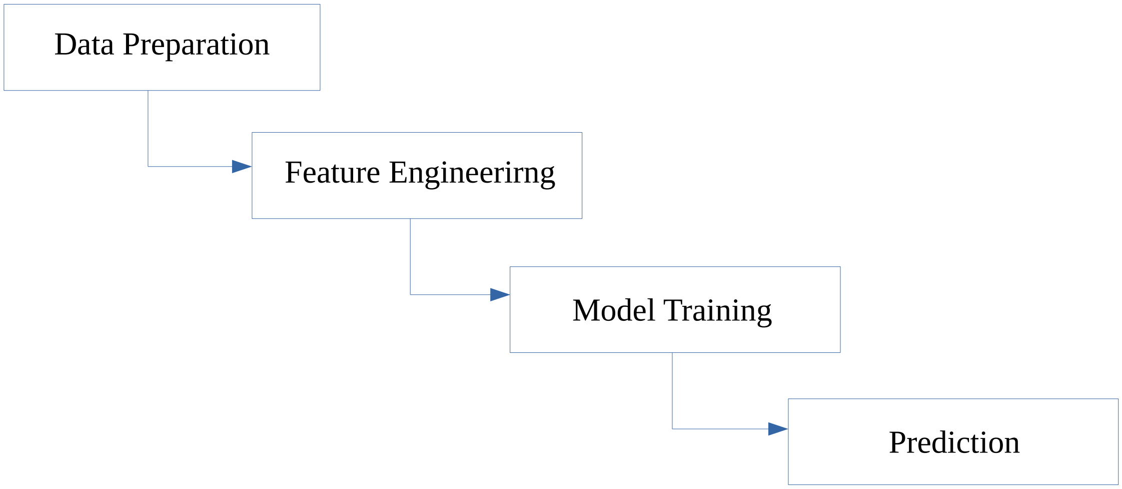


Figure 1 : Project Workflow

i. Data Preparation

- Impute missing data : Figure out the missing data from all dataset Filled some missing data with the median of respective feature column and drop those feature which has huge number of missing data (more than 50% missing data). Feature 'year_built' is with and 'floor_count' from building dataset
- Merge dataframe : Building dataset has site_id as foreign key of weather dataset, merged them. Building dataset has building_id as a foreign key for train and test dataset, merged them and create new dataframe for further process.

ii. Feature Engineering

- Encode categorical data : Encode categorical data using LabelEncoder. The feature primary_use which describe type of building is label encoded.
- Transformation : Transform timestamp data to datetime and create new feature day, hour and month.
- Logarithmic scaling : Separate target dataframe from train dataset and apply logarithmic scaling.

iii. Model Training

Before training the train dataset apply light gradient boosting to figure out top ten important features among the features and we got building_id, square_feet, meter, month, primary_use, air_temperature, site_id, dew_temperature, hour and day as the important features. Using those feature train two model linear regression and light bgm with 5 fold and predict target value for the validation set of data.

iv. Prediction

Initially prediction is applied for validation set of data using both model after getting reliable score apply both model for predicting given test dataset and create submission csv file for submission and generate score for the submitted predicted dataset.

Implemented Model

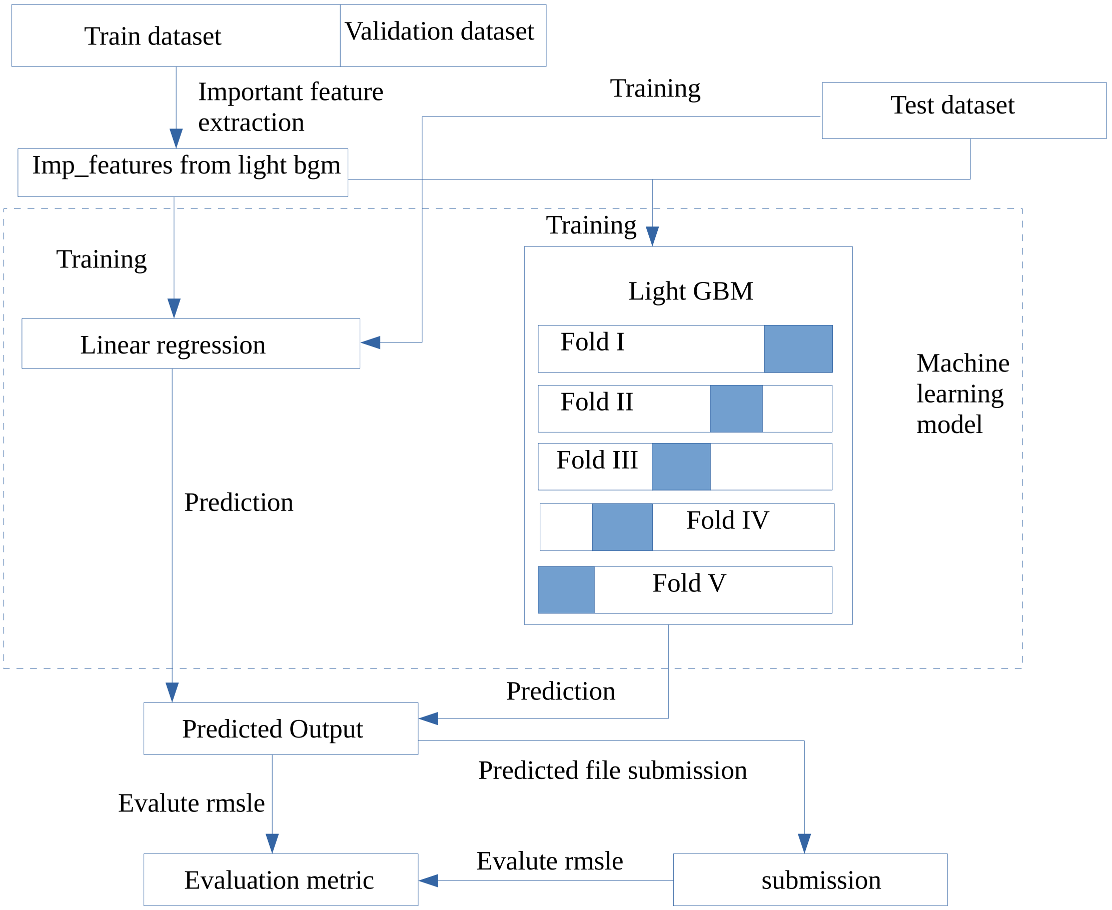


Figure 2 : Implemented model

i. **Train/ Test Split**
Train dataset is split into train and validation dataset in the ratio of 4:1 i.e. 75% train dataset and 25% validation dataset. The split train dataset is further use for extract important feature and validation dataset is used for training the machine learning model for evaluation of the model.

ii. **Extract Important Feature**
75% of splitted dataset fit into light gbm and extract top ten important features. Following parameters are applied for light gbm;

```
parameters = {
                'objective': 'regression',
                'boosting': 'gbdt',
                'metric': {'rmse'},
                'num_leaves': 20,
                'feature_fraction': 0.8,
                'bagging_fraction': 0.8,
                'learning_rate': 0.05,
                'alpha': 0.1,
                 'lambda': 0.1, }
```

Apply 2000 iteration for boost with early stopping criteria 100 and verbose_eval 20 i.e. print evaluation metric on every 20 boosting stage by checking the improvement for 100 time and the process continue for 2000 times. From the process plot the important features, among them select top ten features (building_id, square_feet, meter, month, primary_use, air_temperature, site_id, dew_temperature, hour, day) for training validation and test dataset.

iii. Machine Learning Model

a) Linear Regression

Implement builtin sklearn library linear regression for training the validation and test dataset. Since, have multiple features it implement

$$y_i = m_0 + m_1x_{i1} + m_2x_{i2} + \ldots..+ m_nx_{in} + e \qquad \text{(I)}$$

where,

y_i = dependent variable
xi = explanatory variable
m0 = y-intercept (constant term)
mn = slop coefficient for each explanatory variable
e = model's error (residuals)

$$e = y\_real - y\_pred \qquad \text{(II)}$$

b) Light Gradient Boosting Machine

Light gradient boost machine is the gradient boosting framework which use tree base learning algorithm. Light gbm implement tree vertically while other algorithm uses horizontally i.e. leaf wise and level wise tree. Because its advantages like fast training efficiency, low memory usage, better accuracy, parallel learning supported we also choose light gbm as the second machine learning model for prediction.

We implemented same parameters as in the model use for extracting important features. Apply boosting type as gradient boosting decision tree and evaluation metric as root mean square error.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}} \qquad \text{(III)}$$

iv. Prediction

The validation dataset is fitted into linear regression model first and predict the target value. Similarly fitted into second model light gbm and predict target for the validation dataset. Also apply same for test dataset after finalizing the model.

v. Evaluation Metric

The output predicted dataset apply root mean square log error as evaluation metric.

$$\epsilon=\sqrt{((1/n)\sum(i=1n)(\log(p_i+1)-\log(a_i+1))^2)} \qquad \text{(IV)}$$

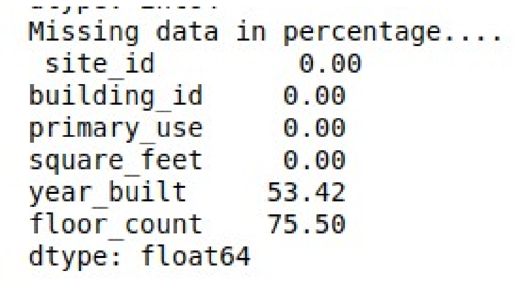
This evaluation metric is apply for both output, predicted data from linear regression and predicted data from light gbm. Among them second model give better than first machine learning model.

vi. **Submission**

The output predicted data is submitted to the kaggle for the evaluation of metric as per competition mentioned.

Output Analysis

From the building dataset huge amount of data is missing, 53.42% year_built and 75.50% floor_count data is missing. On filling the missing data using mean or median give much outlier data so dropped those field.



```
Missing data in percentage....
 site_id         0.00
building_id     0.00
primary_use     0.00
square_feet     0.00
year_built     53.42
floor_count    75.50
dtype: float64
```

Figure 3: Calculating missing data in percentage for building dataset

Among 15 features, select top ten important features among the features. Ten feature from top from figure below are use as feature for the model training and predict target.

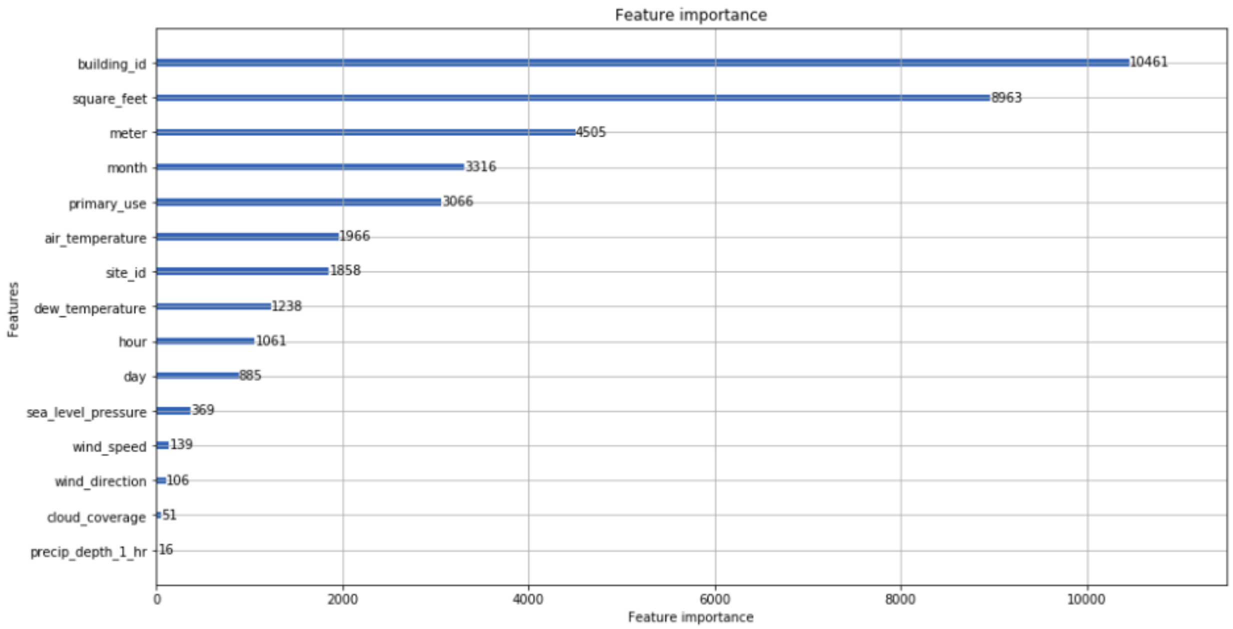


Figure 3: Important feature from light gbm

After selecting important feature training linear regression it took nearly 3 minute and training done using light gbm with 5fold it took about 3 hours. On predicting validation set using selected features evaluation metric rmsle for linear regression is 7.79296 and rmsle for light gbm is 5.51339. This shows that light gbm machine learning model is better than linear regression. Then predicting of target variable on test dataset, the rmsle get better result than on validation set 2.050 for light gbm and 2.191 for linear regression. It shows that both model performing well with good score from light gbm.

LIMITATION AND FUTURE ENHANCEMENT

With just applying machine learning model getting score 2.020 and 2.191 in private score is good but model can not secure best result as per the kaggle score board.

The model can enhance with applying leaked data and more features. As per the score board top competitor secure best score as they implement neural network, so model can enhance using neural network and categorical boosting framework.

CONCLUSION

ASHRAE great energy predictor III competition submitted two model which secure score 2.020 and 2.191 as private score & 1.938 and 1.954 as public score. Data preparation, feature engineering, model training and prediction, four step workflow is applied for the project. Dropping those features which have huge missing data and data pre-processing the remaining data by filling nan field with median value of respective features and merging the csv files is carried out on data preparation step. Train dataset is split into train and validation dataset with ratio 4:1. Lable encoding, transformation of data and logarithmic scaling is apply on feature engineering step with extracting applying top ten features from important features from light gbm. Two machine learning model,linear regression and light gbm a gradient boosting framework is apply for training validation dataset and further applying for training given test dataset. Then predict the target value for validation using both model and apply root mean square error as evaluation metric. Finally the predicted data for testing data is saved as csv file and submitted to kaggle which provide sore. From the score, it conclude that among two implemented machine learning model light gbm secure better score than linear regression model. And the score can be enhance using neural network with applying leaked data.

CITATION

[1] ASHRAE, https://www.ashrae.org/about

[2] Dr. Miller. (2019). Data Science Meets ASHRAE: The Great Energy Predictor Competition III, https://eta.lbl.gov/seminar/data-science-meets-ashrae-great-energy

[3] kaggle, Prior Competition, https://www.kaggle.com/c/ashrae-energy-prediction/overview/prior-competitions

[4] kaggle, Description, https://www.kaggle.com/c/ashrae-energy-prediction/overview/description

[5] Matt Motoki. (2019). 1st Place Solution Team Isamu & Matt, https://www.kaggle.com/c/ashrae-energy-prediction/discussion/124709

[6] Vopani. (2019). 2nd Place Solution, https://www.kaggle.com/c/ashrae-energy-prediction/discussion/123481

[7] eagle4. (2019). 3rd Place Solution, https://www.kaggle.com/c/ashrae-energy-prediction/discussion/124984

[8] MPWARE. (2019). 9th place solution, https://www.kaggle.com/c/ashrae-energy-prediction/discussion/125307

[9] Georgi Pamukov. (2019). 25th Place Solution, https://www.kaggle.com/c/ashrae-energy-prediction/discussion/123525