

MGMT6560 Final Project

ASHRAE - Great Energy Predictor III

**How much energy will a building
consume?**

Overview

Objective:

Accurate estimates of energy-saving investment for different buildings.

Feature Selection

- Methods used
- Features selected list

Model applied

- LightGBM

Improvement

- Dimension reduction
 - Model improvement
-

Data Description - Size



Feature Engineering ? ♦
Model Selection ?

Training Dataset: over 20 million rows

♦ Test Dataset: over 40 million rows

Index	meter_reading
5750030	0
14697993	3.61092
6251104	1.54756
14648512	4.06937
8869829	1.57691
3926699	16.5983
6459463	5.65697
11317820	4.27625
8926706	0.156491
6563550	6.01425
12757986	5.75641
16719012	5.63962
19487268	0

Data Description -Original Features

- **Building_id** - primary key
- **Meter** - {0: electricity, 1: chilledwater, 2: steam, 3: hotwater}
- **Timestamp** - time
- **Meter_reading** - target value
- **Site_id** - location
- **Primary_use** - banking, education, healthcare, entertainment, etc.
- **Square_feet** - gross area
- **Year_built**
- **Floor_count**
- **Air_temperature** - degree celsius
- **Cloud_coverage** - portion
- **Dew_temperature** -degree celsius
- **Precip_depth_1_hr**
- **Sea_level_pressure**
- **Wind_direction**
- **Wind_speed**

Feature Engineering

- ❖ **Hour** - categorical factor
- ❖ **Month** - categorical factor
- ❖ **Weekday** - binary factor

Results

- ❖ Dataset: 1% training data from original training dataset by random shuffling
- ❖ RMSLE (Root Mean Squared Logarithmic Error.)

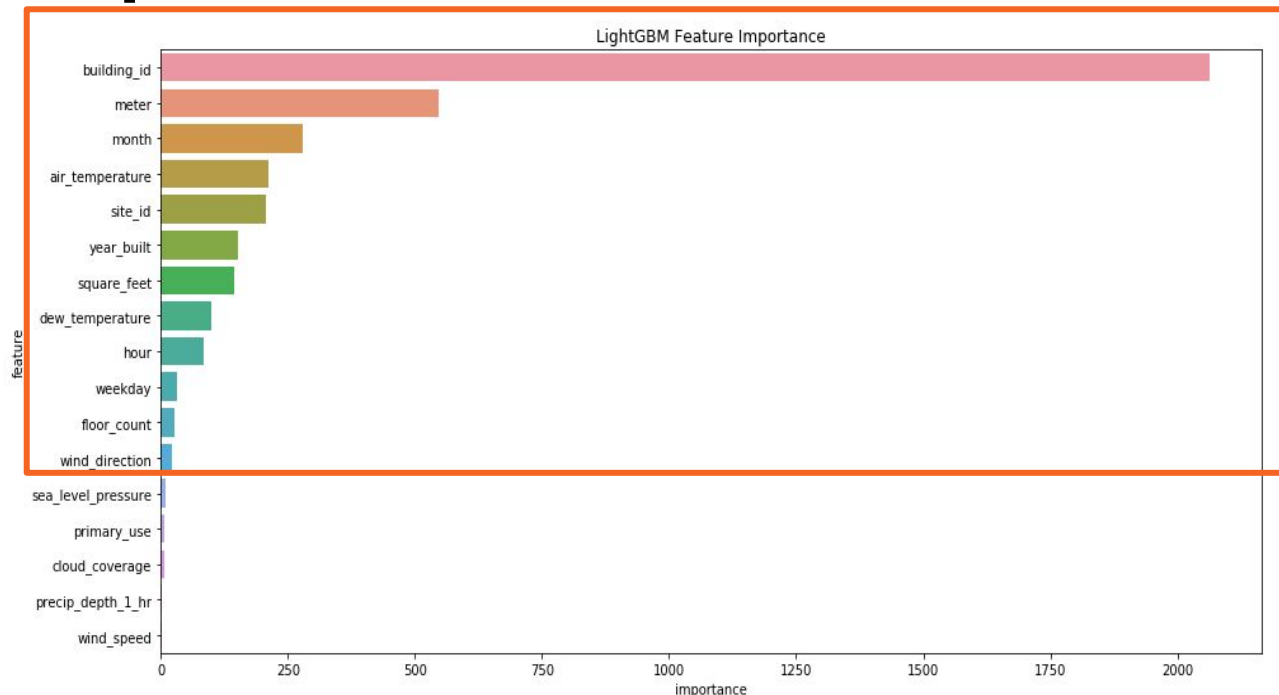
0.39

$$\epsilon = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}$$

Index	pred_y
0	5.00778
1	6.50953
2	5.84678
3	5.50973
4	4.44758
5	4.94788
6	5.46912
7	3.54134
8	4.01248
9	4.77858

Improvement - Feature Selection

- ❖ For the whole dataset
- ❖ Reduce dimensions
- ❖ Half - half LightGBM



Novelty

- ❖ Model based **Feature Selection** method to reduce dimensions
- ❖ **LightGBM model** to avoid more feature engineering work such as one-hot.



```
train_data = lgb.Dataset(train_X, label=train_y,  
                          categorical_feature=["building_id", "site_id", "meter", "month",  
                                              "hour", "weekday", "wind_direction", "year_built"])
```

Q & A
