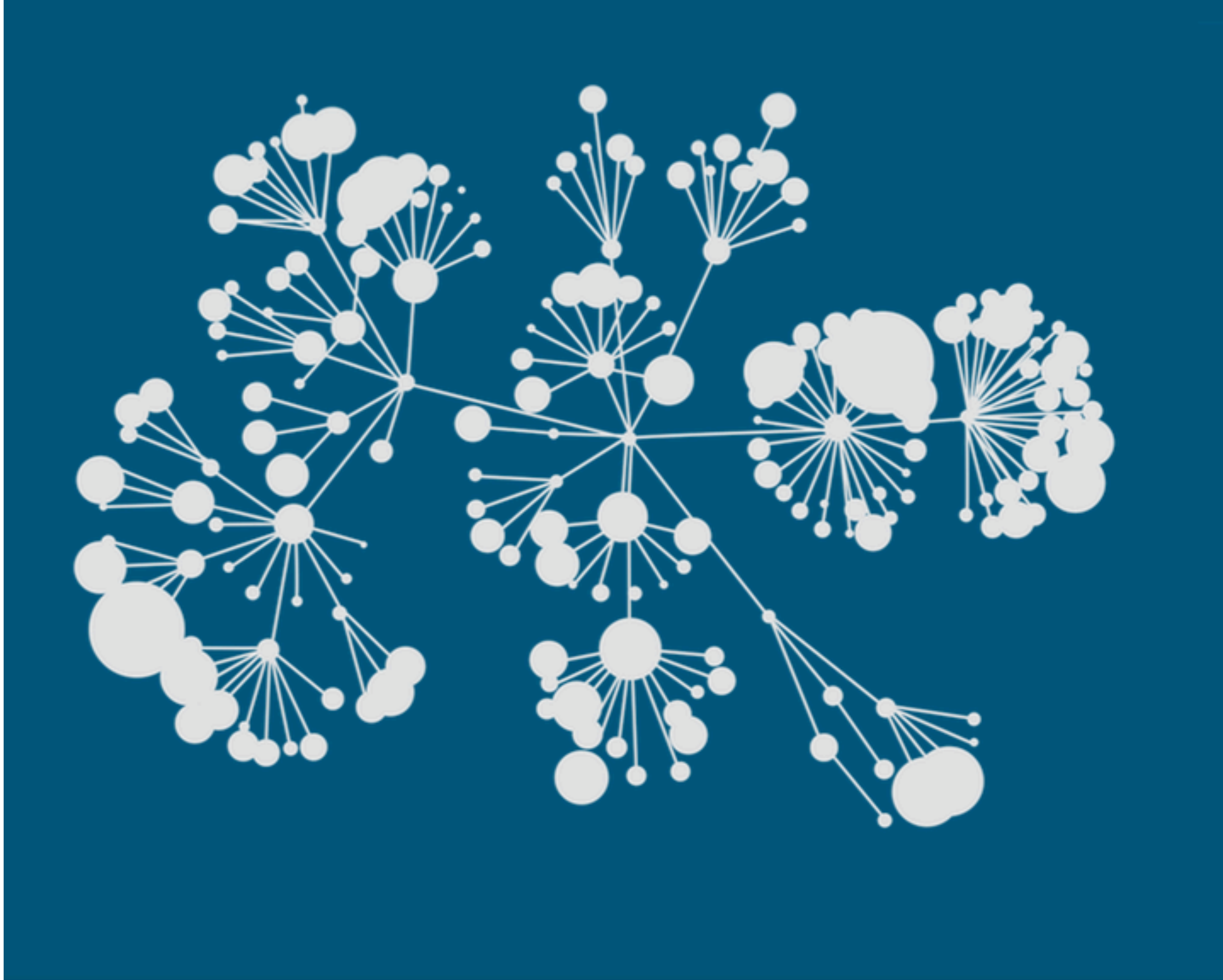
Kaggle

5th place solution



Agenda

1. Background
2. Summary
3. Feature selection & engineering
4. Training methods
5. Important findings
6. Simple model

Agenda

Background

- Tatsuya Sano(Graduate student of University of Tsukuba/ Major: Computer Science, Data Mining)
- Minoru Tomioka(Graduate student of University of Tsukuba/ Major: Computer Science, Numerical Analysis)
- Yuta Kobayashi(Graduate student of University of Tsukuba/ Major: Computer Science, Optimization)

Agenda

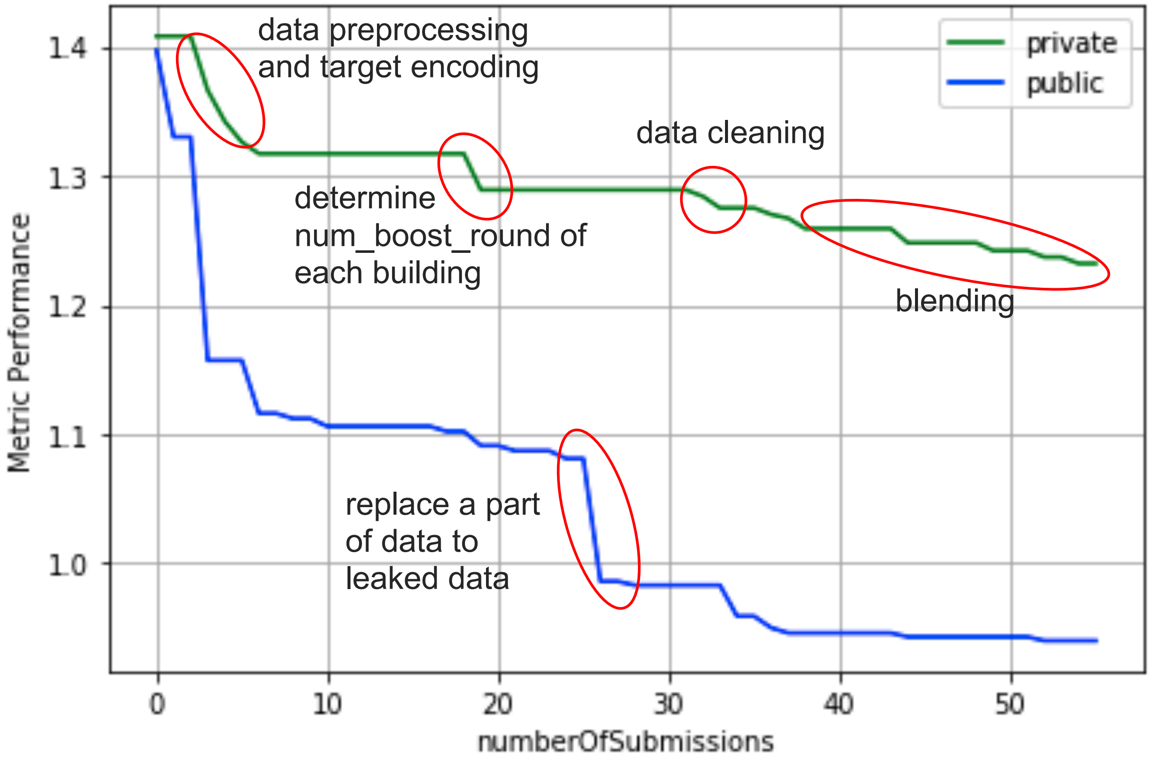
1. Background
2. Summary
3. Feature selection & engineering
4. Training methods
5. Important findings
6. Simple model

Summary

- We only use LightGBM as regressor
- The three most important features were building_id, building_meter_5,building_meter_95(described later)
  •One of our biggest insights was that special target encoding(5% and 95% percentile of target value of each building_id/meter) gave me a big performance improvement
- Used Python (Pandas and LightGBM)
- After ensemble, our score would be 1.236 private / 1.047 public

Summary

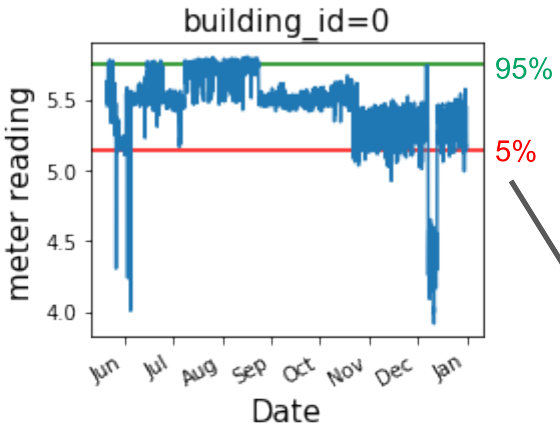
Leaderboard Performance Chart

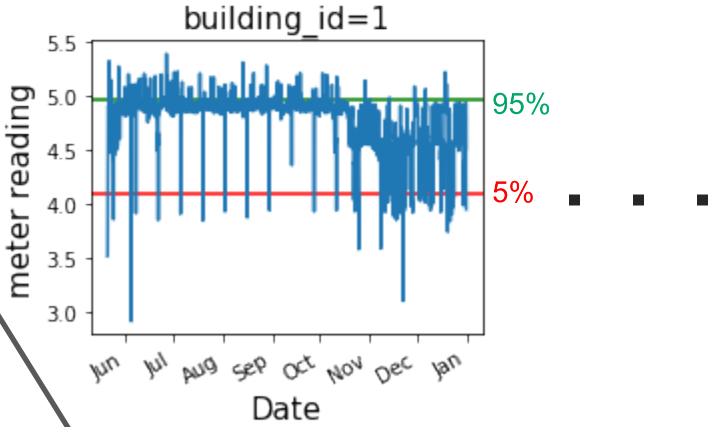


Agenda

1. Background
2. Summary
3. Feature selection & engineering
4. Training methods
5. Important findings
6. Simple model

Target encoding
(5 and 95 percentile)

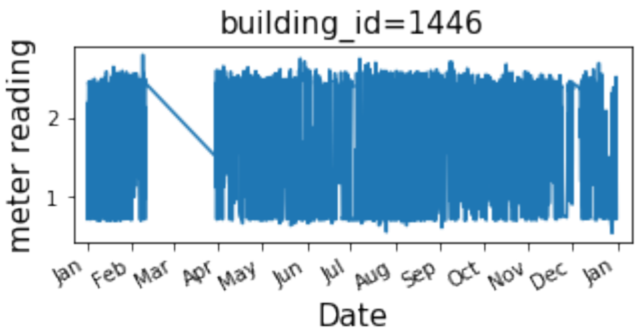




building_id	meter	5%	95%
0	0	5.14	5.75
1	0	4.10	4.96
…	…		
1448	0	1.11	1.77

Target encoding (proportion)

example(building_id =1446)



Target engoding (median) for each day of week

Day of week	Median of Target
0 (Sunday)	0.742
1 (Monday)	2.382
…	
6 (Saturday)	1.194

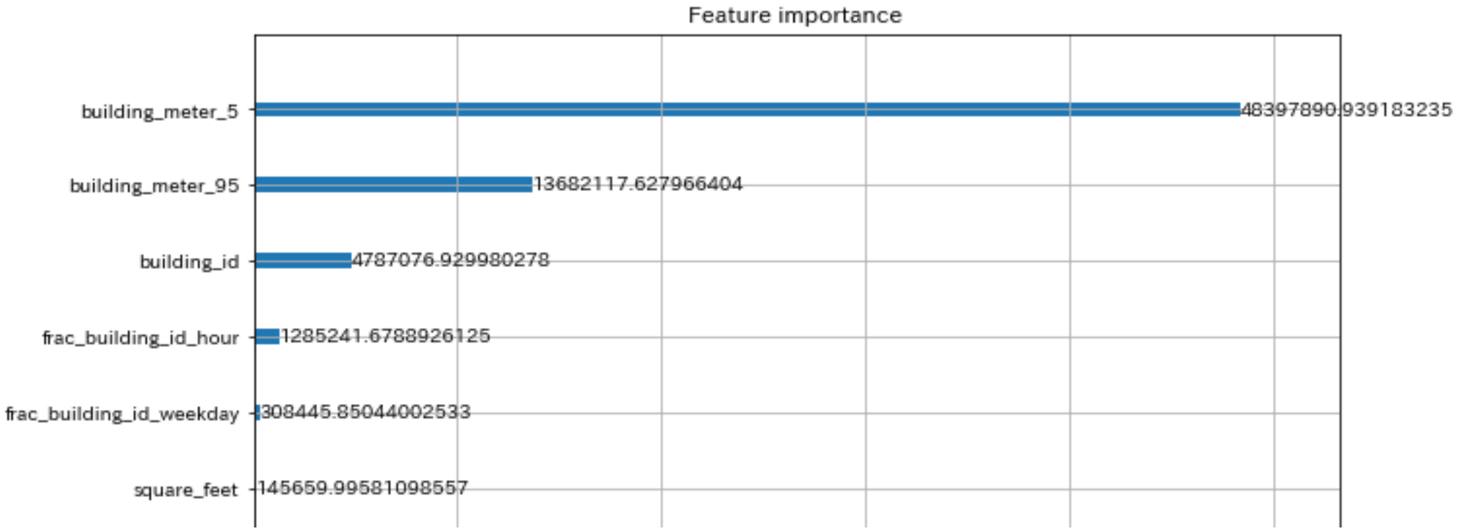
Calculate propotion for each day of week

Day of week	Proportion*
0 (Sunday)	0.054
1 (Monday)	0.173
…	
6 (Saturday)	0.087

$$*\text{Proportion}(i) = \frac{\text{Median of Target}(i)}{\sum_{j=0}^{6} \text{Median of Target}(j)}$$

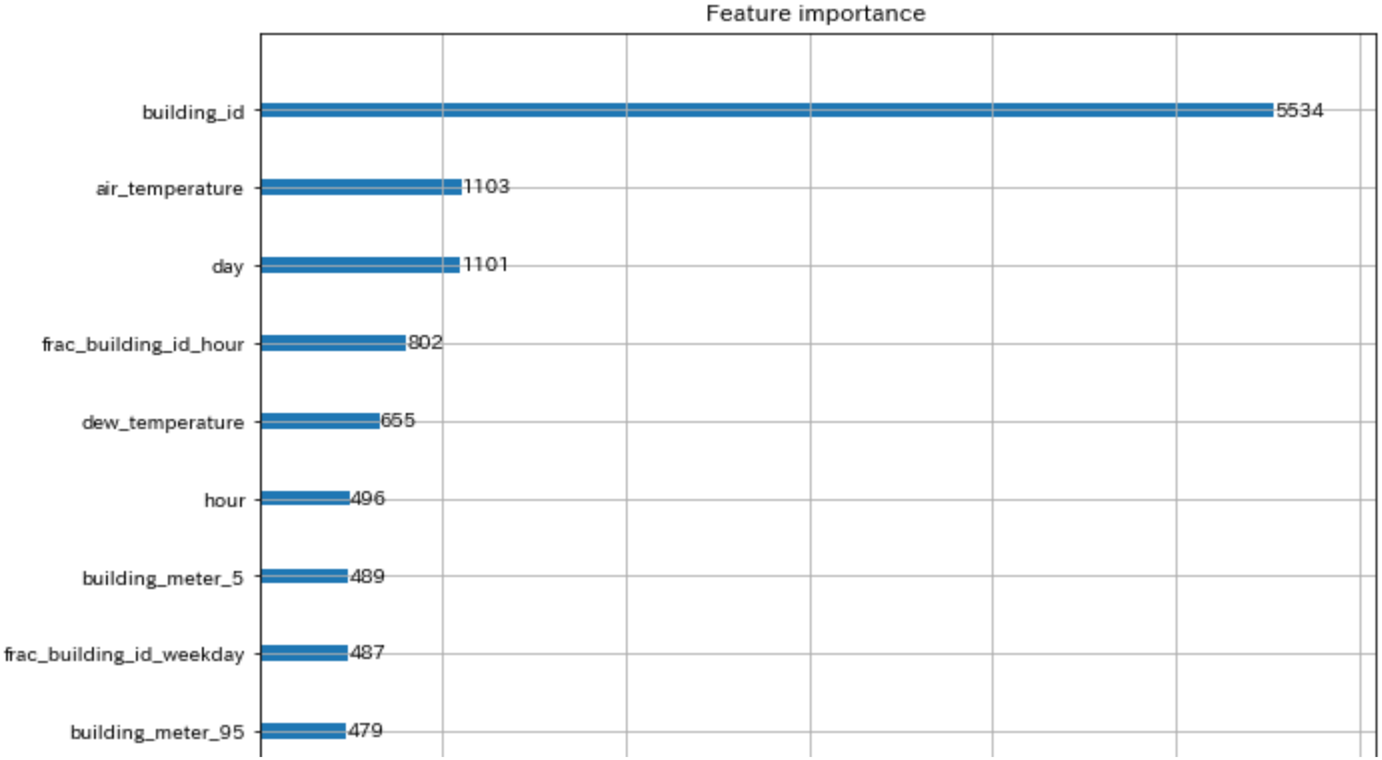
Features Selection / Engineering

Variable Importance Plot(Gain)



Features Selection / Engineering

Variable Importance Plot(Split)

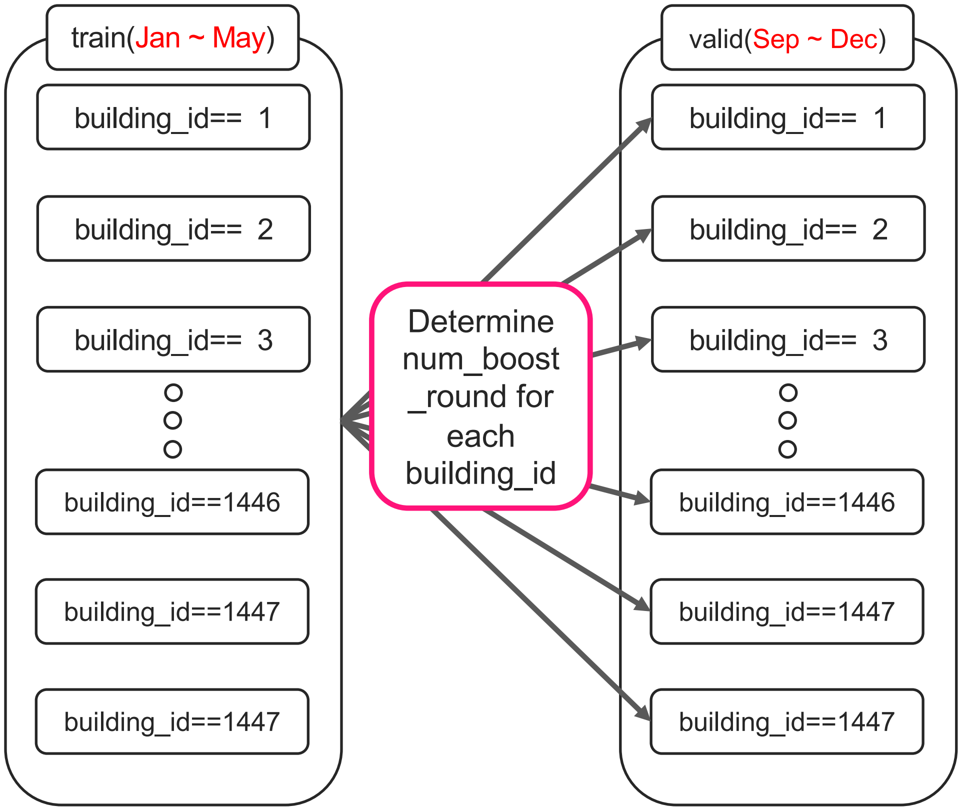


Agenda

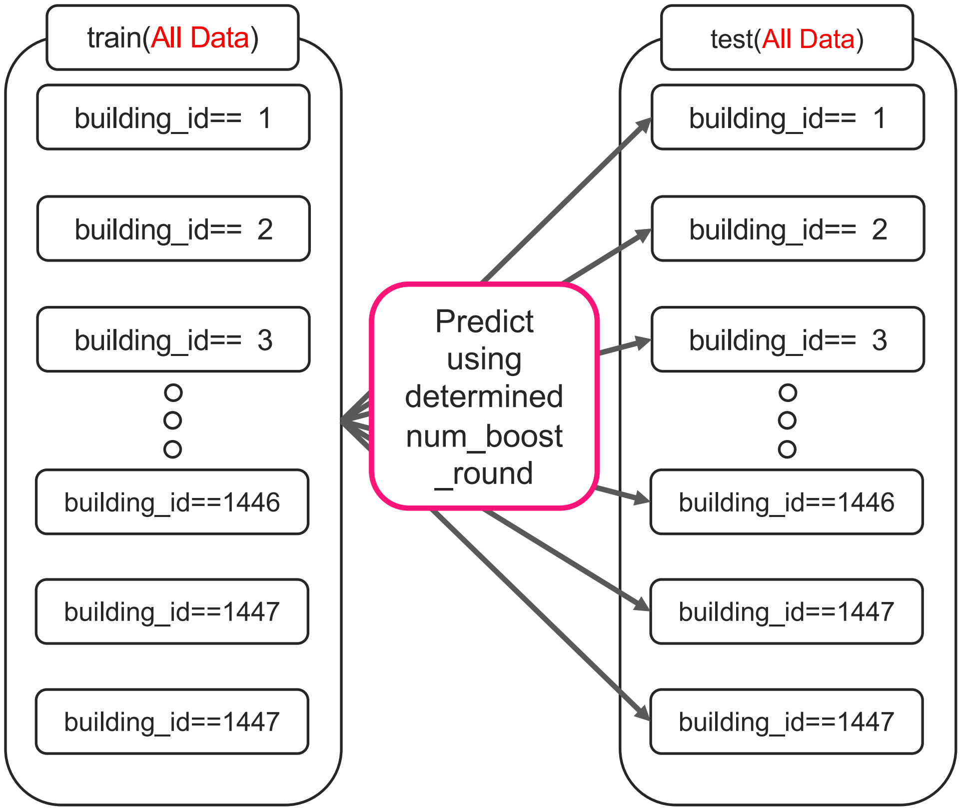
Training Methods

- We used LightGBM
- We determine num_boost_round for each building_id/meater (see next slide)
- We used leaked data for deciding ensemble weight(We also used other competitor’s submission files to ensemble)
  •We chose ensemble weight that minimize RMSLE between submission data and leaked data

Determine num_boost_round -Step1



Determine num_boost_round -Step2



Agenda

1. Background
2. Summary
3. Feature selection & engineering
4. Training methods
5. Important findings
6. Simple model

Important and Interesting Findings

- What set we apart from others in the competition
  - determining num_boost_round of each building/meter
  - data cleaning
  - special target encoding(5% and 95% percentile of target value of each building_id/meter)
  - special target encoding(proportion of target value per week, per hour, per day)
  - ensemble using leaked data
  - ensemble by meter

Agenda

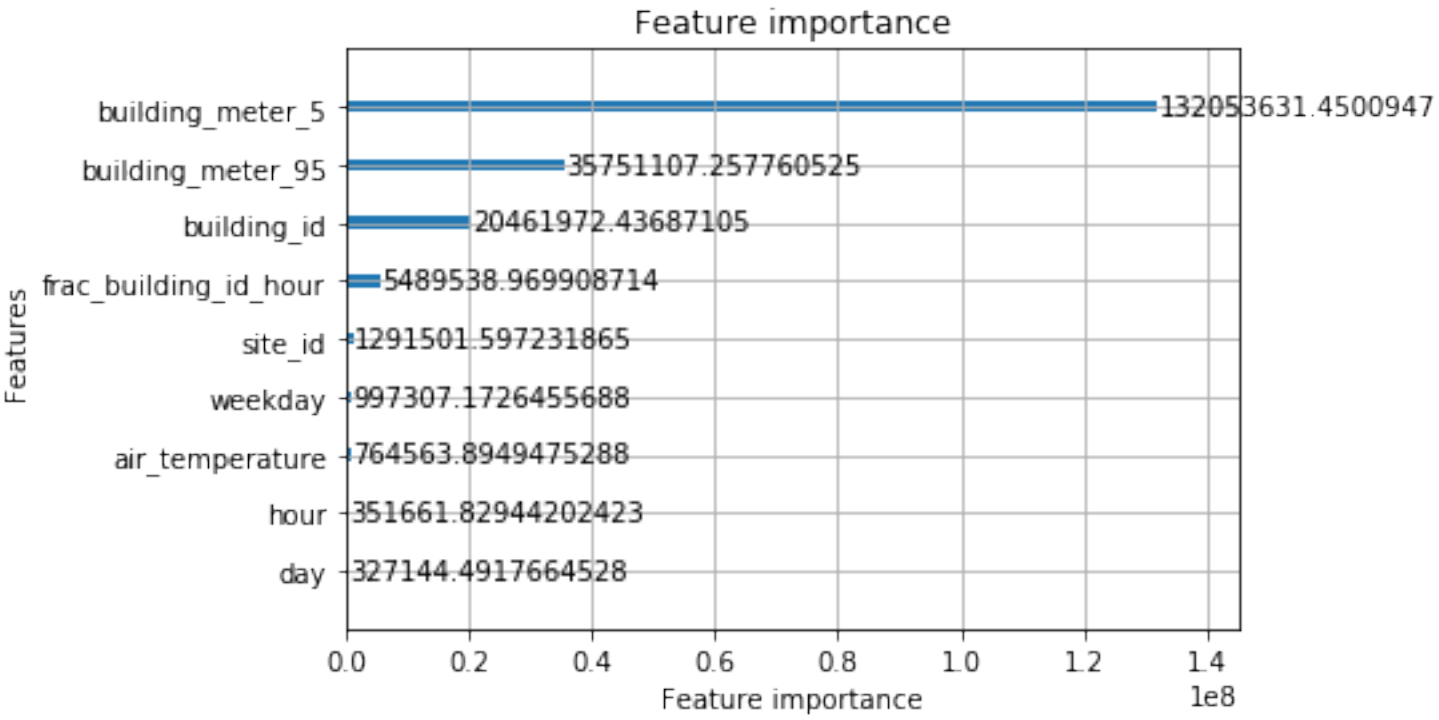
1. Background
2. Summary
3. Feature selection & engineering
4. Training methods
5. Important findings
6. Simple model

Simple Model

- [Outline a subset of features that would get 90-95% of your final performance]
  - We show feature importance in the next slide
- [If you used an ensemble, was there a single classifier that did most of the work? Which one?]
  - We didn’t ensemble for simplified model
- [What would the simplified model score be?]
  - 1.272 private / 1.068 public

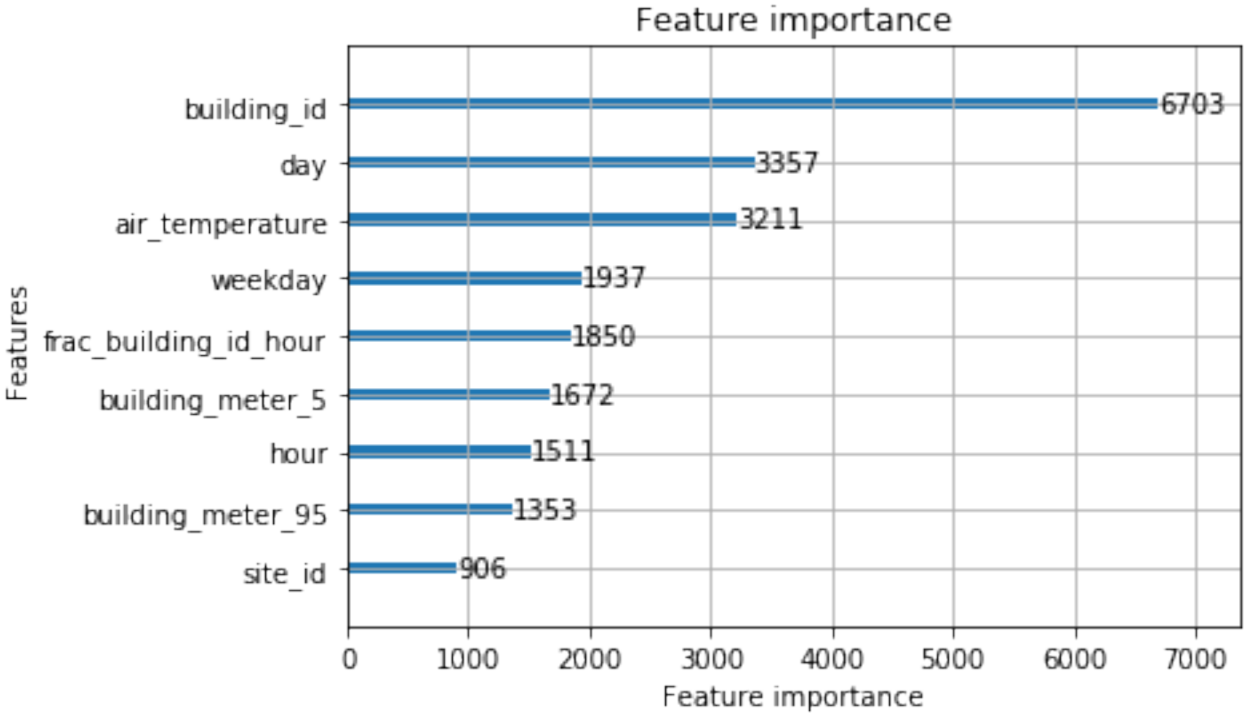
Simple Model

Simplified feature importance(gain)



Simple Model

Simplified feature importance(split)



kaggle™