**A. MODEL SUMMARY**

**A1. Background on you/your team**

* Competition Name: ASHRAE – Great Energy predictor III
* Team Name: mma
* Private Leaderboard Score: 1.237
* Private Leaderboard Place: 5th
* [For each team member]
* Name: Tatsuya Sano
* Location: Japan
* Email: sano.tatsuya.sw@alumni.tsukuba.ac.jp
* Name: Minoru Tomioka
* Location: Japan
* Email: tomioka.minoru.sy@alumni.tsukuba.ac.jp
* Name: Yuta Kobayashi
* Location: Japan
* Email: kobayashi.yuta.xu@alumni.tsukuba.ac.jp

**A2. Background on you/your team**

* What is your academic/professional background?
  + We are graduate student at University of Tsukuba.
* Did you have any prior experience that helped you succeed in this competition?
  + Nothing.
* What made you decide to enter this competition?
  + We are interested in tabular data competition.
* How much time did you spend on the competition?
  + 100 hours total.
* If part of a team, how did you decide to team up?
  + Tomioka’s invitation.
* If you competed as part of a team, who did what?
  + Kobayashi : collect leaked data and analyzed them.
  + Tomioka : read discussion and kernels, share us these knowledge.
  + Sano(Leader) : All of other things(ex. make model, feature engineering).

**A3. Summary**

* The training method we used is LightGBM.
* The most important feature is building\_id.
* The tools we used Python(pandas,LightGBM)
* It takes about 7 minutes to preprocessing, 15 minutes to train, 22 minutes to predict(including ensemble).
* After ensemble, our score would be 1.236 private /1.047 public

**A4. Features Selection / Engineering**

* What were the most important features?
  + Fig.1 and Fig.2 are variable importance plot by gain and by split

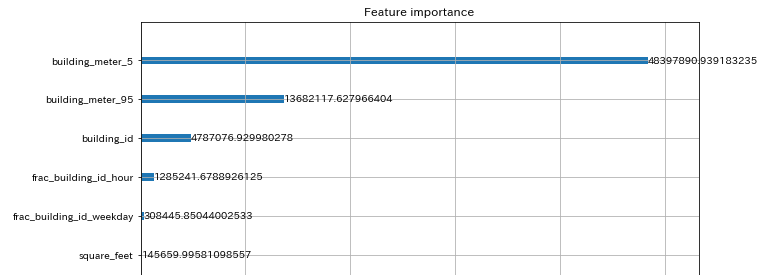


Fig.1 variable importance plot by gain

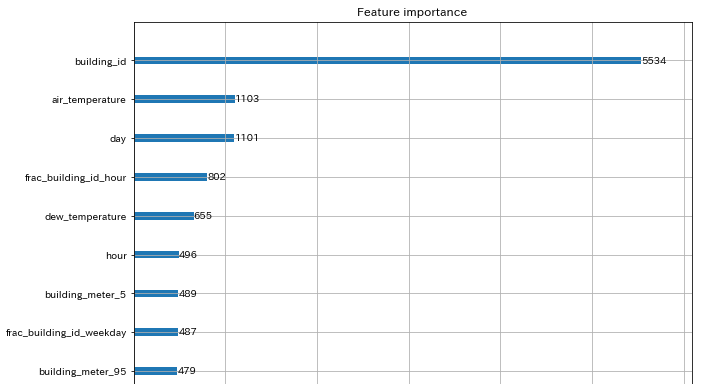


Fig.2 variable importance plot by split

* How did you select features?
  + We make some features and if it works on CV, we adopt the features
* Did you find any interesting interactions between features?
  + Nothing
* Did you use external data? (if permitted)
  + Yes[1]

**A5. Training Method**

* What training methods did you use?
  + LightGBM.
* Did you ensemble the models?
  + Yes, we also use other competitor’s submission files[2][5].
* If you did ensemble, how did you weight the different models?
  + We used leaked data for deciding weight.
  + We chose weight that has smaller RMSLE between submission data and leaked data.

**A6. Interesting findings**

* What was the most important trick you used?
  + 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 day, ... ).
  + determining num\_boost\_round of each building/meter.
  + data cleaning(fill missing value and drop bad rows).
* What do you think set you apart from others in the competition?
  + Tricks written above. Data cleaning(fill missing value and drop bad rows) is most important(referring to [2] [3] [4]).
* Did you find any interesting relationships in the data that don't fit in the sections above?
  + Nothing.

**A7. Simple Features and Methods**

* Our model is simplified enough but we tried restricting fewer than 10 features.
* Is there a subset of features that would get 90-95% of your final performance? Which features?
  + Fig.3 and Fig.4 are simplified model’s variable importance plot by gain and by split.
* What model that was most important?
  + We didn’t ensemble for simplified model.
* What would the simplified model score?
  + 1.272 private/ 1.068 public.

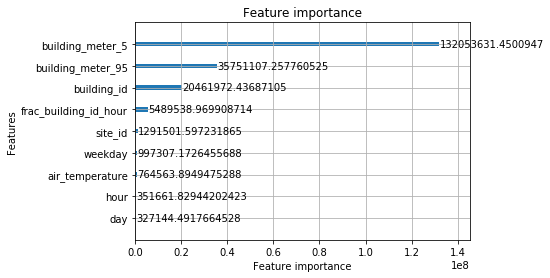
****

Fig.3 simplified model’s variable importance plot by gain

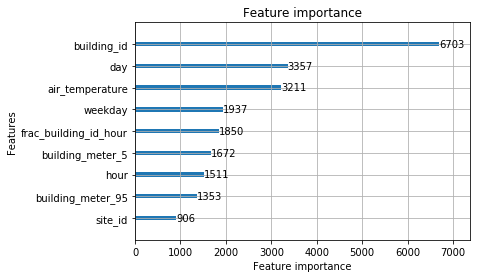


Fig.4 simplified model’s variable importance plot by split

**A8. Model Execution Time**

* How long does it take to train your model?
  + About 10 minutes.
* How long does it take to generate predictions using your model?
  + About 8 minutes.
  + About 15 minutes to ensemble.
* How long does it take to train the simplified model?
  + About 6 minutes.
* How long does it take to generate predictions from the simplified model?
  + About 4 minutes.

**A9. References**

[1] <https://www.kaggle.com/yamsam/ashrae-leak-data-station>

[2] <https://www.kaggle.com/purist1024/ashrae-simple-data-cleanup-lb-1-08-no-leaks>

[3] <https://www.kaggle.com/c/ashrae-energy-prediction/discussion/113774>

[4] <https://www.kaggle.com/aitude/ashrae-missing-weather-data-handling>

[5] <https://www.kaggle.com/rohanrao/ashrae-half-and-half>