

We thank you for your time spent taking this survey.
Your response has been recorded.

Below is a summary of your responses

[Download PDF](#)

(Part 1/4) Your Background Information:

1. What is the name of your representative and email for registration?

Full Name

Zhijian Xu

Email Address

zjxu21@cse.cuhk.edu.hk

2. Which of the following best describe your current expertise?

- ☒ **data scientists, computer science background**
- ☐ data scientists, civil and building engineering background
- ☐ software engineer
- ☐ building service engineer
- ☐ others:

3. Are you familiar with smart buildings? If yes, how long have you been working on smart building-related projects?

- ☒ **No**
- ☐ Yes, less than 2 years
- ☐ Yes, more than 2 years

4. Are you familiar with machine learning and data analytics? If yes, how long have you been working on them?

- ☐ No
- ☒ **Yes, less than 2 years**
- ☐ Yes, more than 2 years

(Part 2/4) Your opinion on data preprocessing

(Please read this first) this competition gives you three types of data, raw data (one day), pre-consolidated data (three months), and model-prepared data (15 months).

5. Have you ever had experiences in data pre-processing where you need to handle complicated datasets similar to the raw dataset of this competition?

- ☐ No
- ☒ **Yes**

6. Do you agree that it would be a tedious procedure for you to extract the required data from the raw dataset of this competition?

Strongly disagree

☐

Somewhat disagree

☒

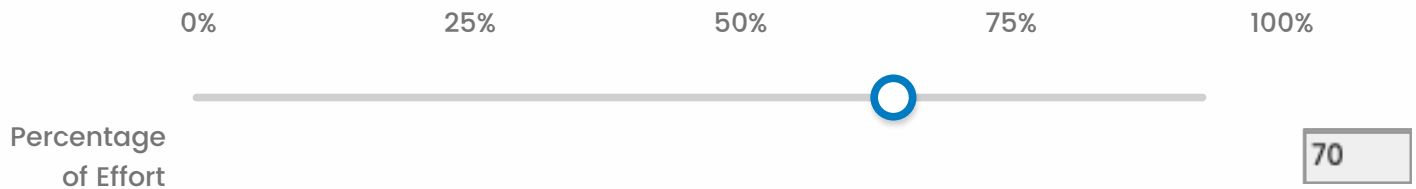
Neither agree nor disagree

☐

Somewhat agree

Strongly agree

7. Please estimate: if given raw data only, how much of effort/time you will spend in data pre-processing (in percentage) to extract data into a form that can be easily used for your model development?



8. Do you agree that it would be a better choice for you to extract data from the raw dataset of this competition by yourself, instead of providing the pre-consolidated data or model-prepared data?

- ☐ No
- ☒ Yes

9. If yes, why do think that this choice is good? If not, please also give your reasons.

By observing the raw data, we can get to know what exactly happened that result in the missing data or outlier.

(Part 3/4) Your opinion on Semantic AI platforms/tools for data preprocessing

10. Please estimate: if given raw data, how much of effort/time a platform, which is similar to Semantic AI Platform and Energon introduced in the workshop, could save for you to handle the raw dataset (in percentage)?

0% 25% 50% 75% 100%

Saved 0%
Effort

25%

50%

75%

100%

25

11. What function(s) of the data preprocessing tool/platform are most useful to you?

(Please read this first) Considering that you are asked to handle a raw dataset from an additional building. It contains the data needed for your model, e.g., flowrate of the chillers, etc. Yet this raw dataset is different in format, sensing structure, etc. to the current EMSD data.

12. Do you agree that handling the raw dataset in this additional building spend significant effort/time again?

☐ No

☒ Yes

13. Please estimate: in this situation, how much effort/time a platform similar to Semantic AI platform and Energon could save for you to handle the dataset in this additional building (in percentage)?

0%

25%

50%

75%

100%

Saved
Effort



50

14. Do you have any comments on Semantic AI platform and Energon?

(Part 4/4) Open Questions

15. What is the most difficult part in this competition (e.g. data

15. What is the most difficult part in this competition (e.g., data preprocessing, model developing or model evaluation)? How did you manage to overcome it?

Data preprocessing. Because there exists a lot of missing data and even missing large chunks of data. This results in that we are not able to obtain enough training data. We used imputation techniques to restore some of the missing points. For the four climate parameters, we use the public data from HK observatory website to restore the missing parts.

16. How did you choose the features for developing your prediction model? What are the features you chose eventually? Also how did you select a proper prediction algorithm/model for this competition?

We conducted a comprehensive data analysis before getting down to the model design. We found that all the climate data have strong impact on the cooling load especially on long term view. But for short term, human activity influence it more such as the working hour and holiday. Thus we add the average waveform of each day of week as its prototype channel besides the given climate data. We selected a model based on the self-attention based on the observation that the cooling load should be similar with similar temperature and humidity.

17. Do you find building-related knowledge helpful in your work? Please briefly describe .

Nope.

18. Have you read academic papers about load forecasting in this competition? To what extent do you find them helpful for your work? Please briefly describe.

Yes. Most of the works are about predicting the load with the data from earlier period as input. But our task is to predict the cooling load out of the given information.

19. Do you find any useful trick or knowledge during the competition that helps you boost the performance of your model?

This dataset have very strong prior knowledge, that is, all the waveform for Monday are very similar. Thus, we added this prior knowledge to our training set and get large performance boost.

