ASHRAE GREAT ENERGY PREDICTOR III

Fanny Ummelen

INTRODUCTION

This Kaggle competition is about predicting the energy usage of buildings based on the data from the previous year and the weather conditions. The goal is to reduce the Root Mean Squared Logarithmic Error. This metric measures the ratio between the predicted and actual value, instead of the absolute difference.

Available data sets:

- train.csv: meter reading for every meter in every building for every hour in 2016
- Weather.csv: weather conditions at every site for every hour in 2016
- Buildling metadata.csv: some data on all investigated buildings, also linking them to a specific site ID.

OUTLINE

- Data cleaning
  - *Outliers and other unusual situations*
  - *Time differences*
  - *Missing weather data*
  - *Missing building years*
- Feature engineering
  - *Time*
  - *Average and standard deviation of meter reading*
  - *Usage time*
  - *Weather influence*
  - *Working day*
  - *Airco*
- Selecting and training a ML model
  - *Descision tree*
  - *Learning curves*
  - *Feature importances*
  - *Compare predictions to actual measurements*
  - *Score on Kaggle competition*
- Conclusions

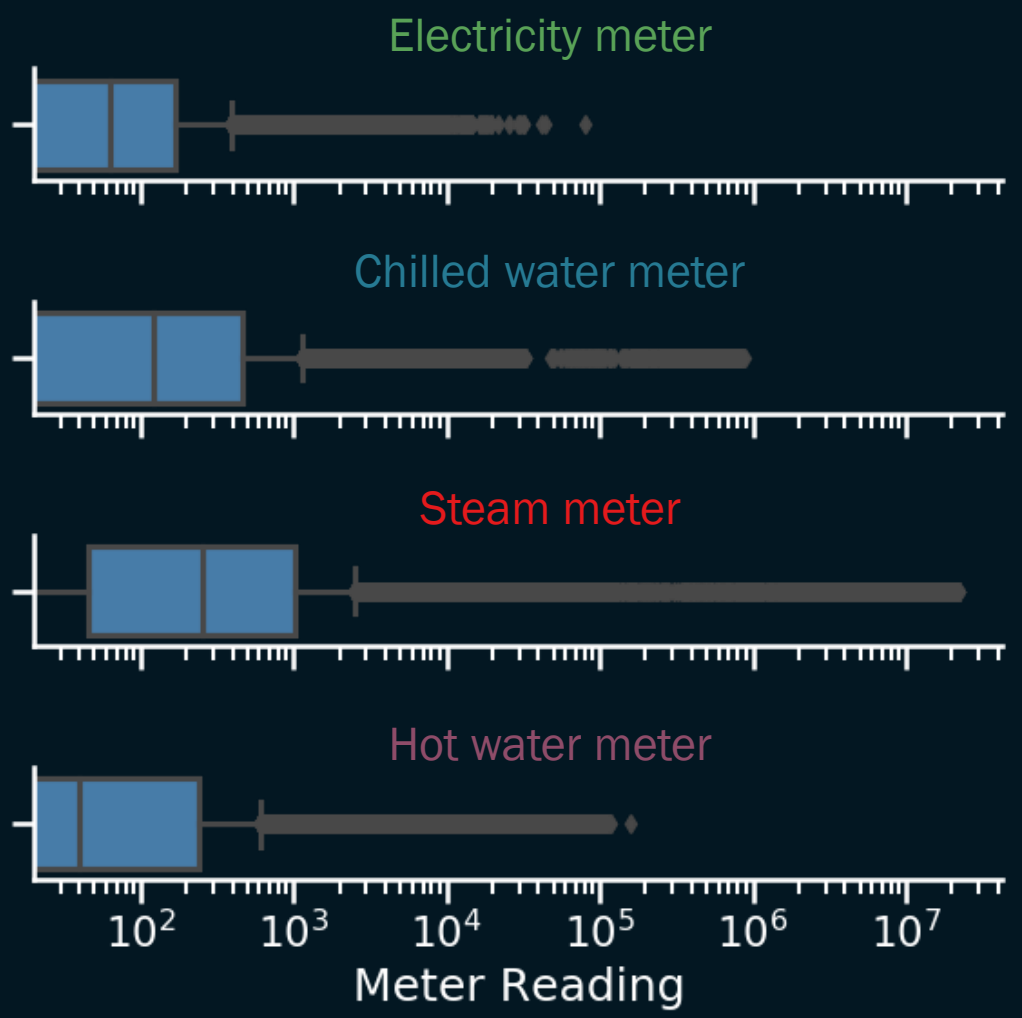
DATA CLEANING

Removing outliers, fixing the time difference between sites, and handling missing values

METER READING DISTRIBUTION

The meter reading is the target value in this project. Here boxplots of the readings of the different types of meters are shown.

- The data contains huge outliers (note the logarithmic axis)
- There are too many outliers to examine each of them individually
- I decided to write a function to detect and remove outliers automatically

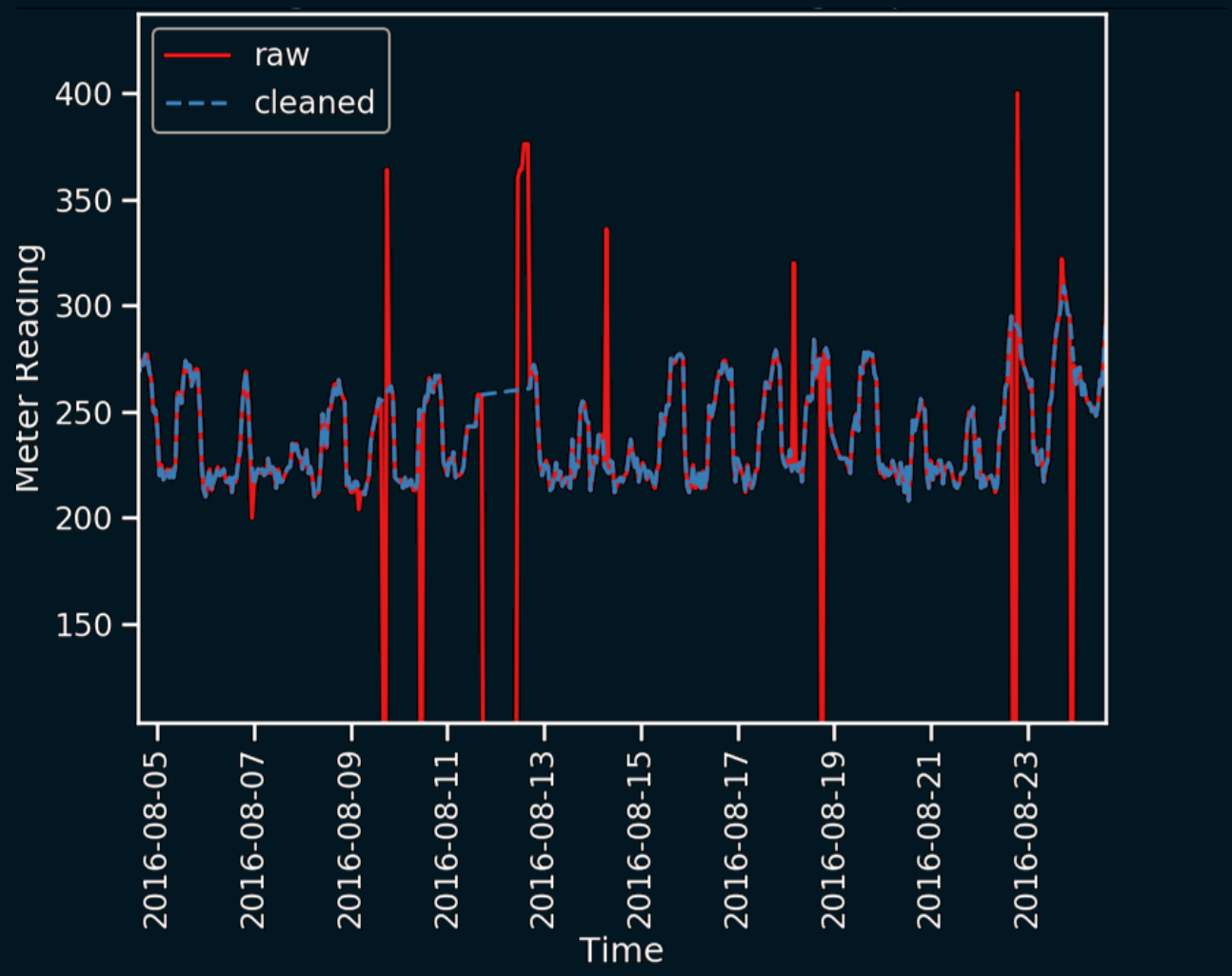


Automated outlier removal

These steps are followed to remove outliers automatically:

- Find the median and inter quartile range of the meter reading for each meter of each building for each month
- For each individual meter reading check if it is acceptably close to its monthly median, and otherwise remove the entry
- Electricity reading before and after outlier removal are plotted for building 1232 as an example
- For meters that have a 0 reading most of the time this method does not work well.

Building 1232 Electricity readings

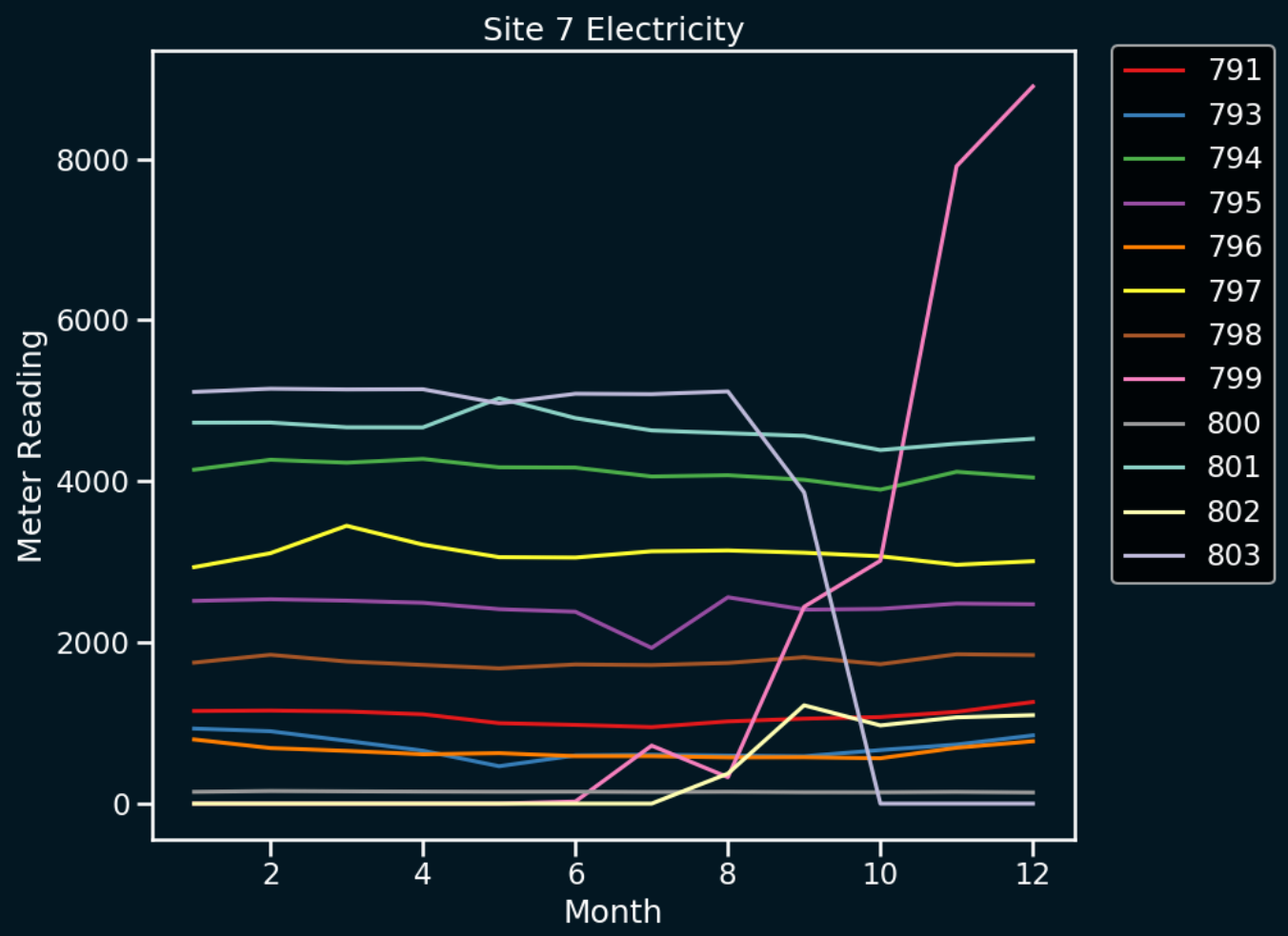


Other unusual situations

Outlier removal does not resolve structural unrepresentative situations. As an example, the electricity readings throughout the year of the buildings on site 7 are plotted.

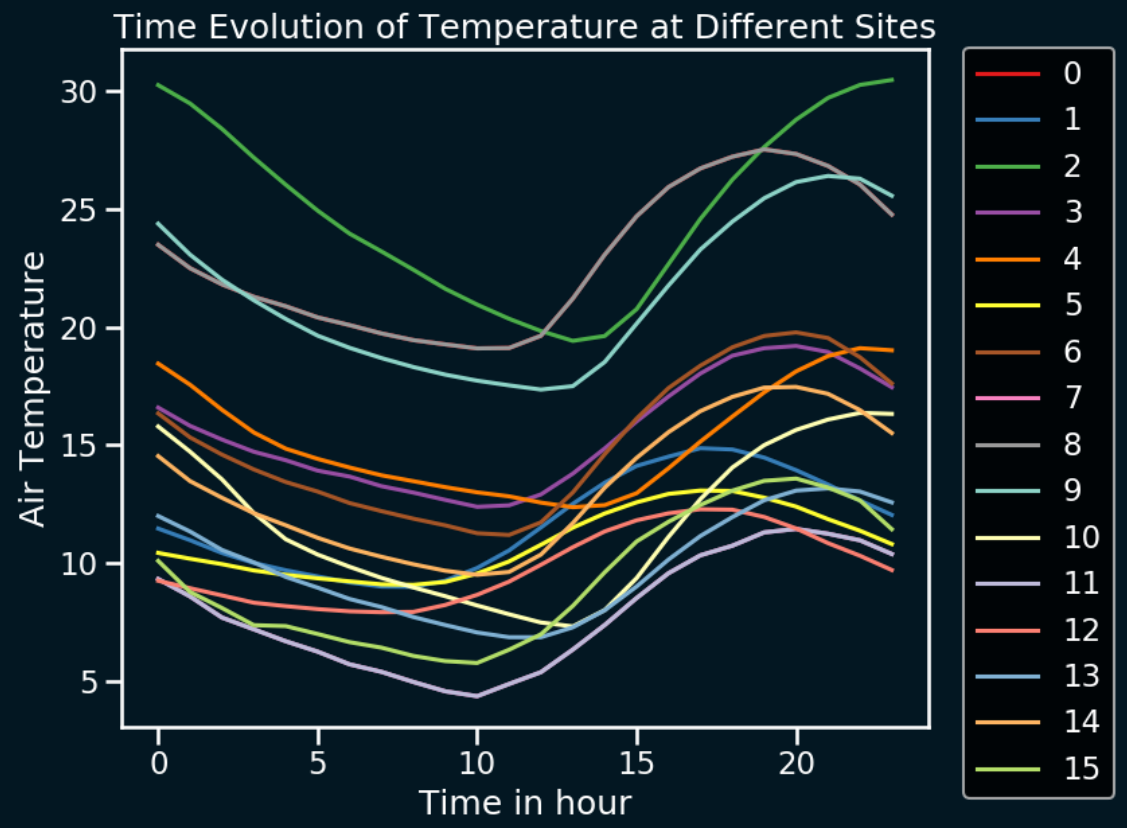
- Building 802 has 0 readings at the start of the year, just like building 799
- Building 803 readings drop to 0 when the readings of 799 and 802 rise up.
- I expect that these kind of situations are related to building renovations, resulting in buildings not being used for a longer time.
- I only include the readings taking in the most recent situations, as they are probably most representative for next years readings.

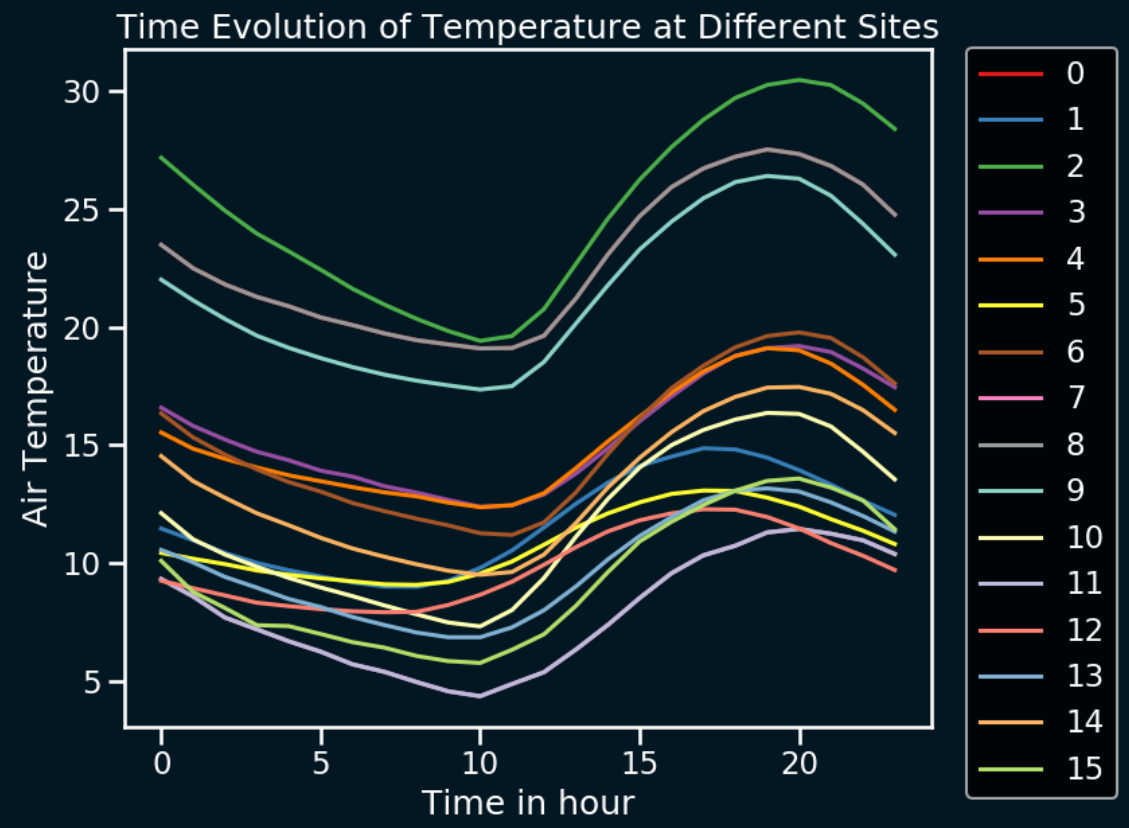
→ I dropped the entries of building 799, 802 and 803 up to October



Time difference correction

When plotting the mean temperature versus the hour of day, the some of resulting curves for the different sites seem shifted horizontally with respect to the rest (left figure). I expect the sites are in different parts of the world, and hence in different time zones. This discrepancy is solved by adding the appropriate amount of hours to the time for the relevant sites (right figure). Site 1, 5 and 12: +3 hours, site 2, 4 and 10: -3 hours, site 9 and 13: -2 hours.





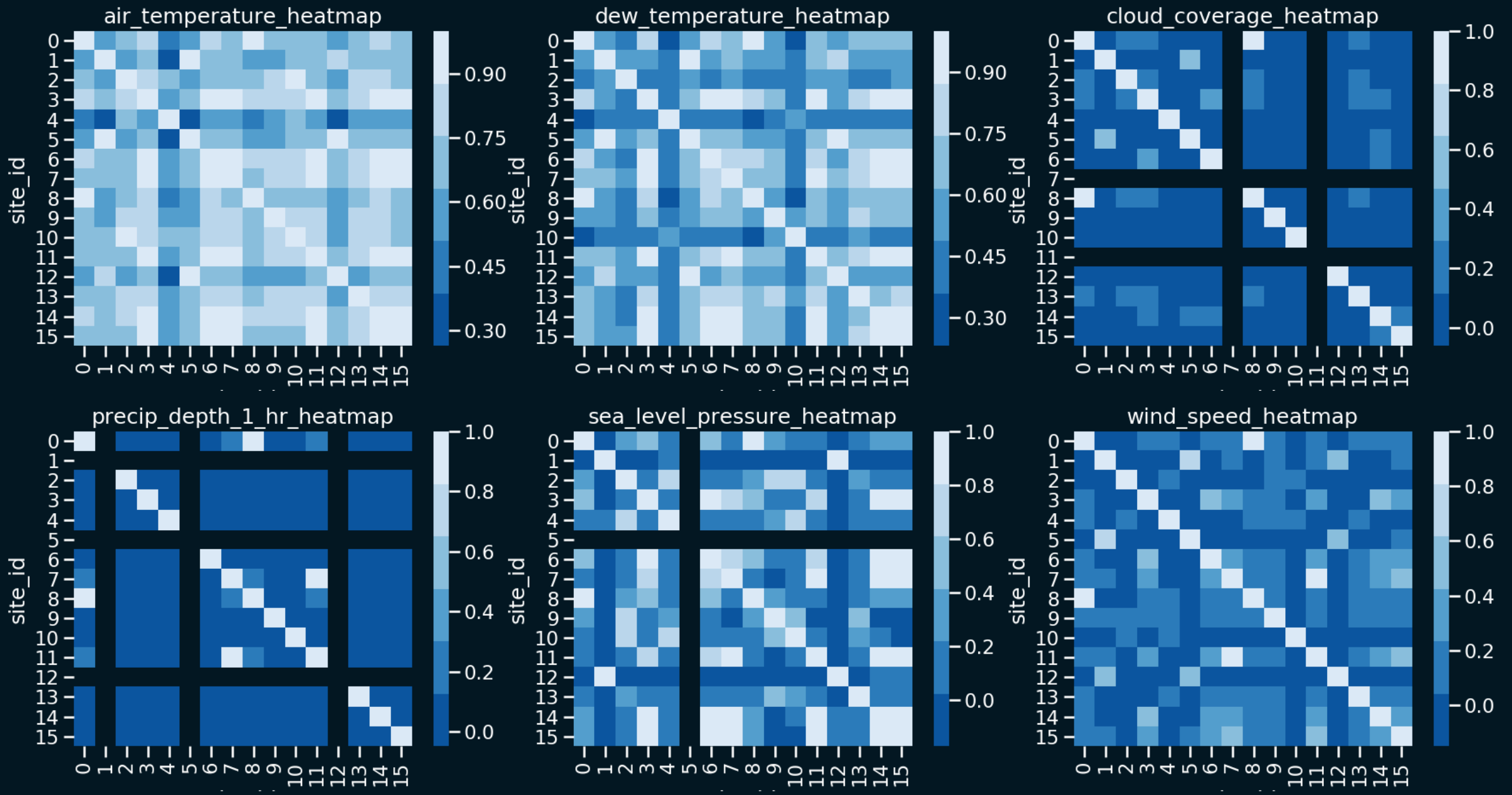
Missing values

The weather data contains a number of missing values. Sometimes a specific weather attribute is missing, sometimes the complete entry of a specific timestamp is missing. I used two strategies to fix this:

- Imputing the value of a nearby site
- Imputing the last recorded value

To know which sites are close to each other, we make heat maps for all weather attributes, so we can see for which sites the weather is highly correlated (see next slide).

Feature	# missing values
Air temperature	55
Cloud coverage	69173
Dew temperature	113
Precip. Depth	50289
Sea level pressure	10618
Wind direction	6268
Wind speed	304


air_temperature_heatmap
dew_temperature_heatmap
cloud_coverage_heatmap
precip_depth_1_hr_heatmap
sea_level_pressure_heatmap
wind_speed_heatmap
site_id
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.90
0.75
0.60
0.45
0.30
1.0
0.8
0.6
0.4
0.2
0.0

Sites grouped

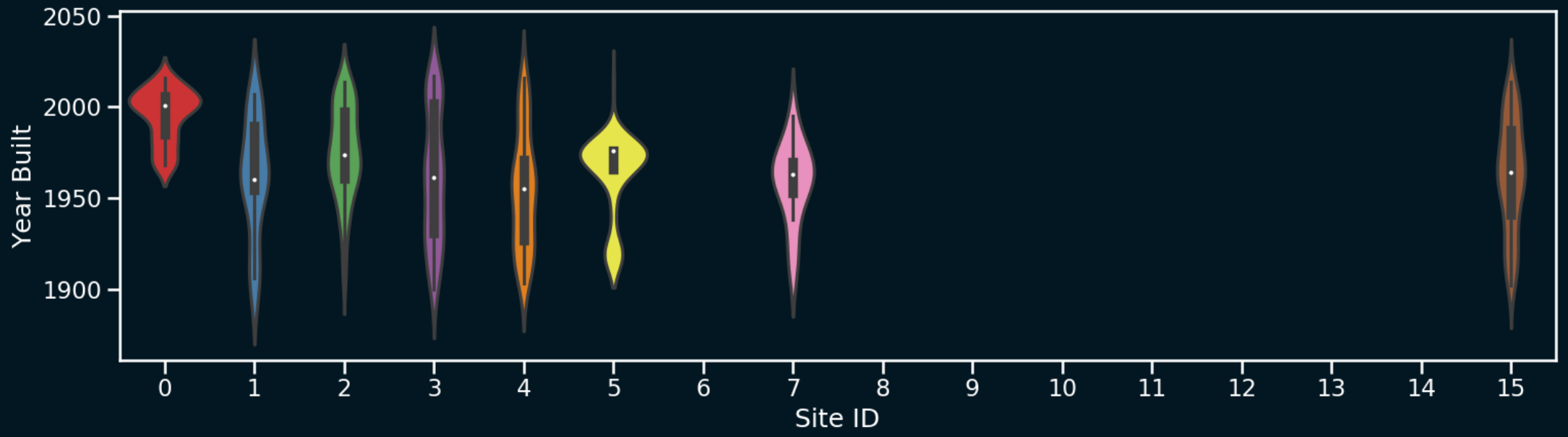
Based on the heat maps on the previous slide, I have identified the following groups of sites:

- Site 1, 5, and 12
- Site 2, 4, and 10
- Site 9 and 13
- All other sites

These groups based on weather readings match the with the time difference analysis I conducted previously!

The map shows where I suspect that the site groups are located, based on weather and time difference.

Missing building years



I expect that the building year is an important factor in how energy efficient a building is. However a very large number of the building years are missing, for some sites none are listed at all! I do not want to dismiss this attribute, so a reasonable value for imputing should be found.

Year built vs meter group

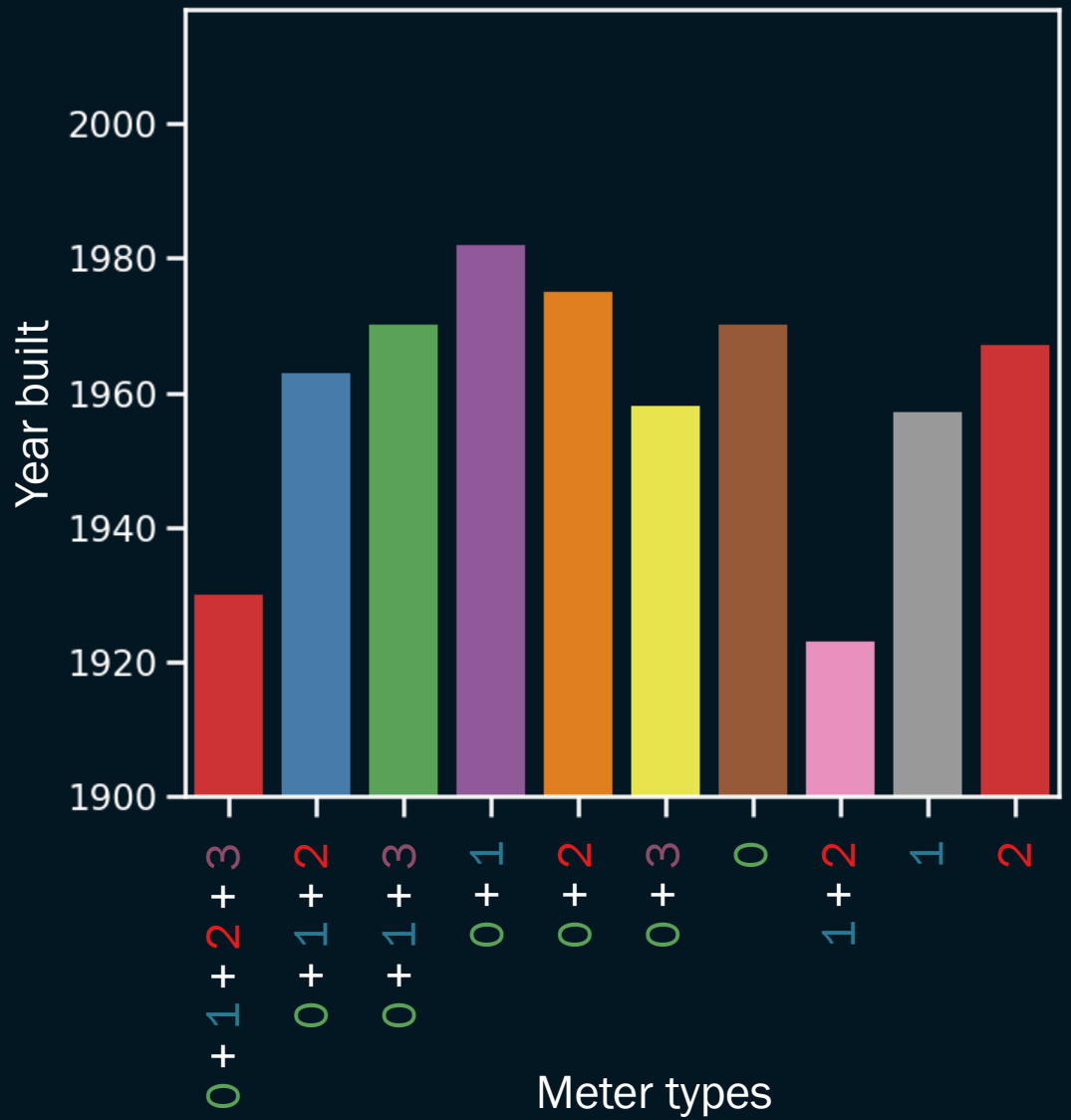
I observed a relation between the types of meters in a building and the building year.

In the figure:

- Electricity = 0
- Chilled water = 1
- Steam = 2
- Hot water = 3

Buildings with only chilled water and steam are the eldest, buildings with only electricity and chilled water are the youngest.

I will use the median of each meter group as values to impute missing building years.



FEATURE ENGINEERING

Time features, features directly based on target, usage time, weather influence, working day, RH, building volume and airco.

Time attributes

A timestamps is included for all entries in the training set. From this I extract:

- The month
- The hour of day

I do not extract the year because all the years in the test set will not occur in the training set

Target as feature

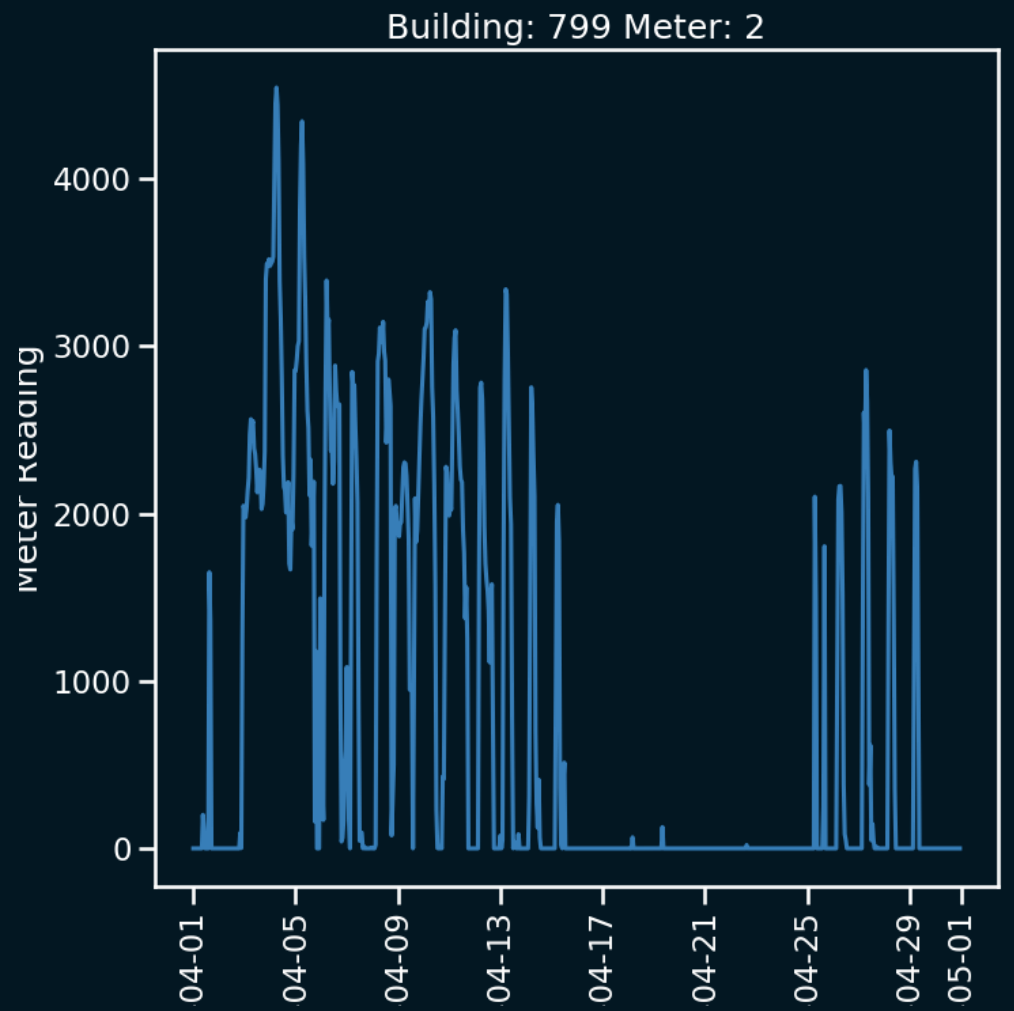
I expect that the average meter reading of a building is a better measure for its energy efficiency than only the building year and building size.

I therefore add the mean meter reading for each meter of each building as an attribute. The standard deviation is saved as an additional feature as well.

Meter usage

Many of the meter reading follow a particular pattern: they are 0 most of the time, but display short, high peaks (see figure).

I added the percentage of time that the reading is not zero as an additional feature (called “meter usage”) for each meter of each building.



Feature importance weather

I expect that the influence of all weather attributes can be summarized quite well by a single feature. Therefore a decision tree is trained with the average meter reading on a day per site as the target value. The feature importances are listed here.

- The most important is the type of meter: steam is required under other conditions than chilled water
- Of the actual weather conditions the air temperature is by far the most important
- Surprisingly, the site ID is also relatively important

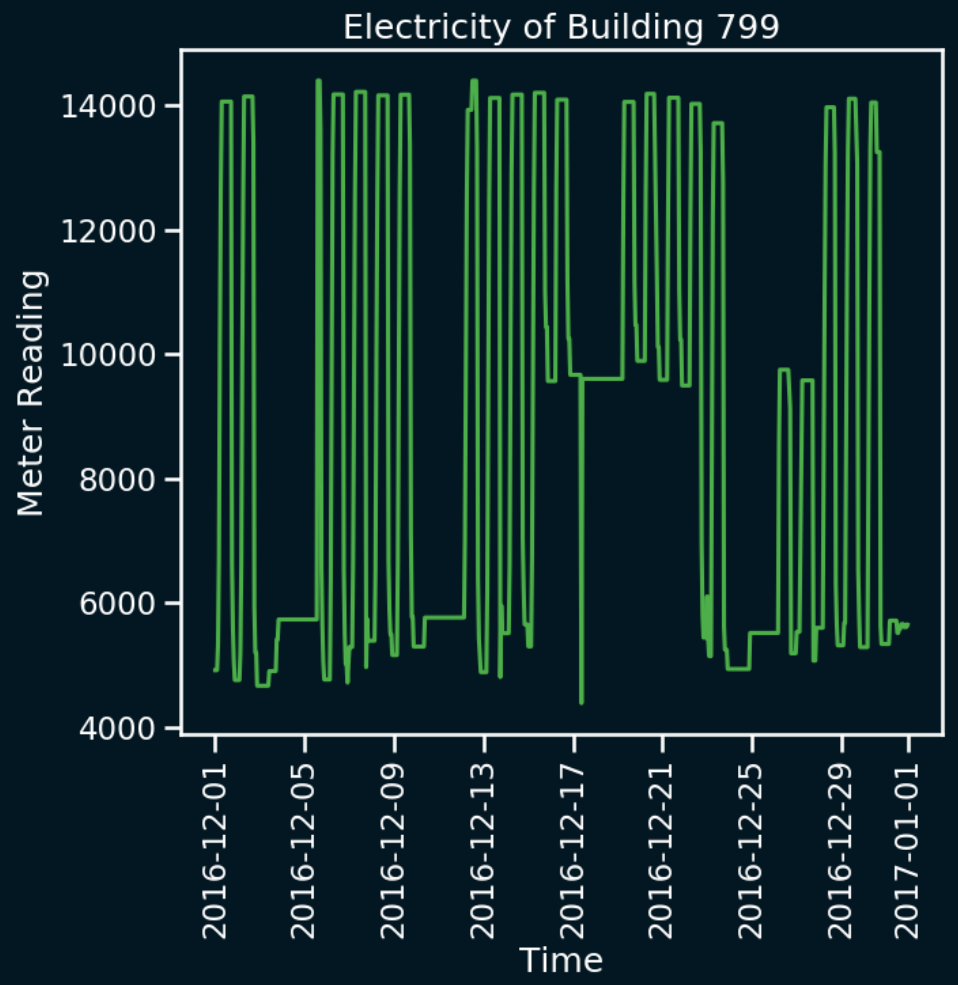
Feature	Importance
Site ID	0.216
Meter	0.383
Air temperature	0.286
Relative humidity	0.041
Wind speed	0.022
Sea level pressure	0.023
Precip depth	0.012
Cloud coverage	0.017

Working days

If we zoom in on the electricity meter readings, we can see a pattern: the readings peak during working hours on week days, but not in weekends.

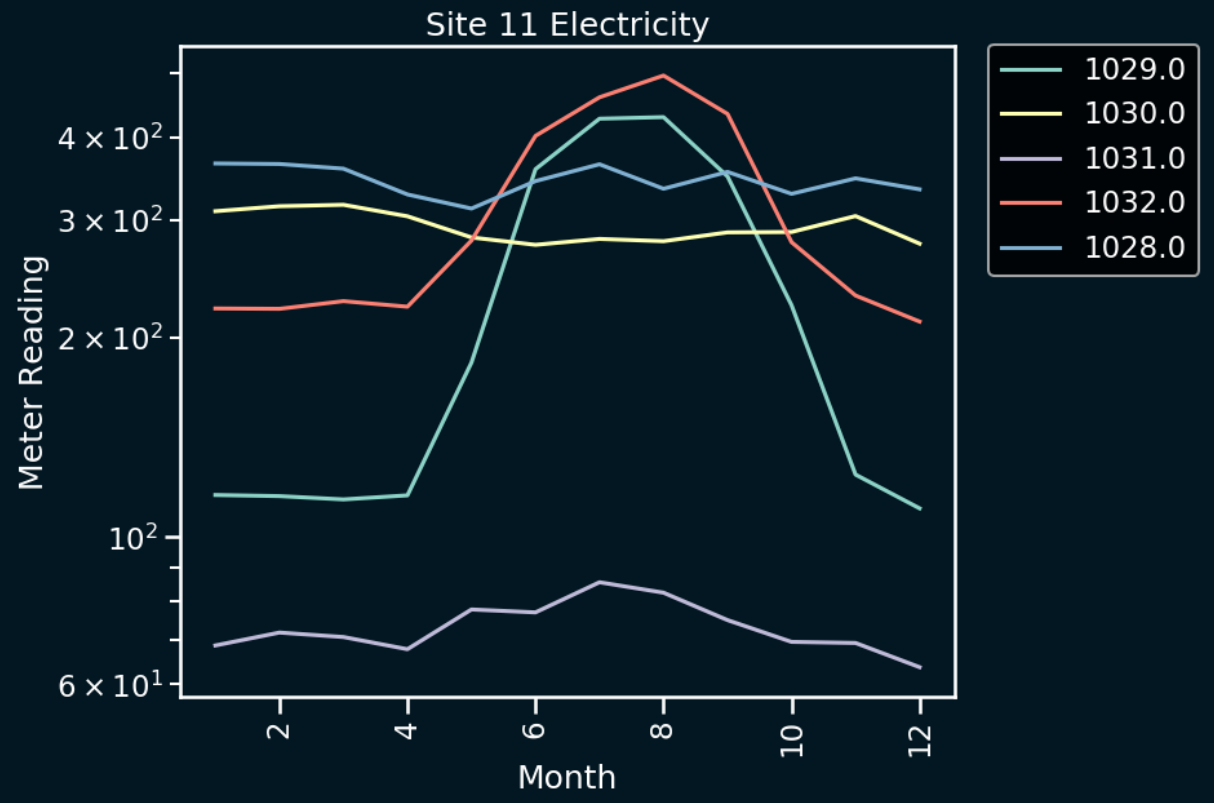
We therefore add a new feature “work” which is zero on weekends and one on week days.

This feature could be further improved by also setting holidays to one.



Airco

For some buildings the electricity reading goes up in summer, while for other buildings it does not. I guess this depends on whether the building has airco or not. I added “airco” as an additional feature, which has value 1 if the median electricity reading in summer is more than 50% higher than the median electricity reading in winter.



MODEL SELECTION & VALIDATION

Model selection

Requirements

The model has to be efficient when the training set is large and has many features. Also, the relation between the features and target are not expected to be linear. The final goal is to optimize the Root Mean Squared Logarithmic Error, so the selected model should be able to do this.

Appropriate models

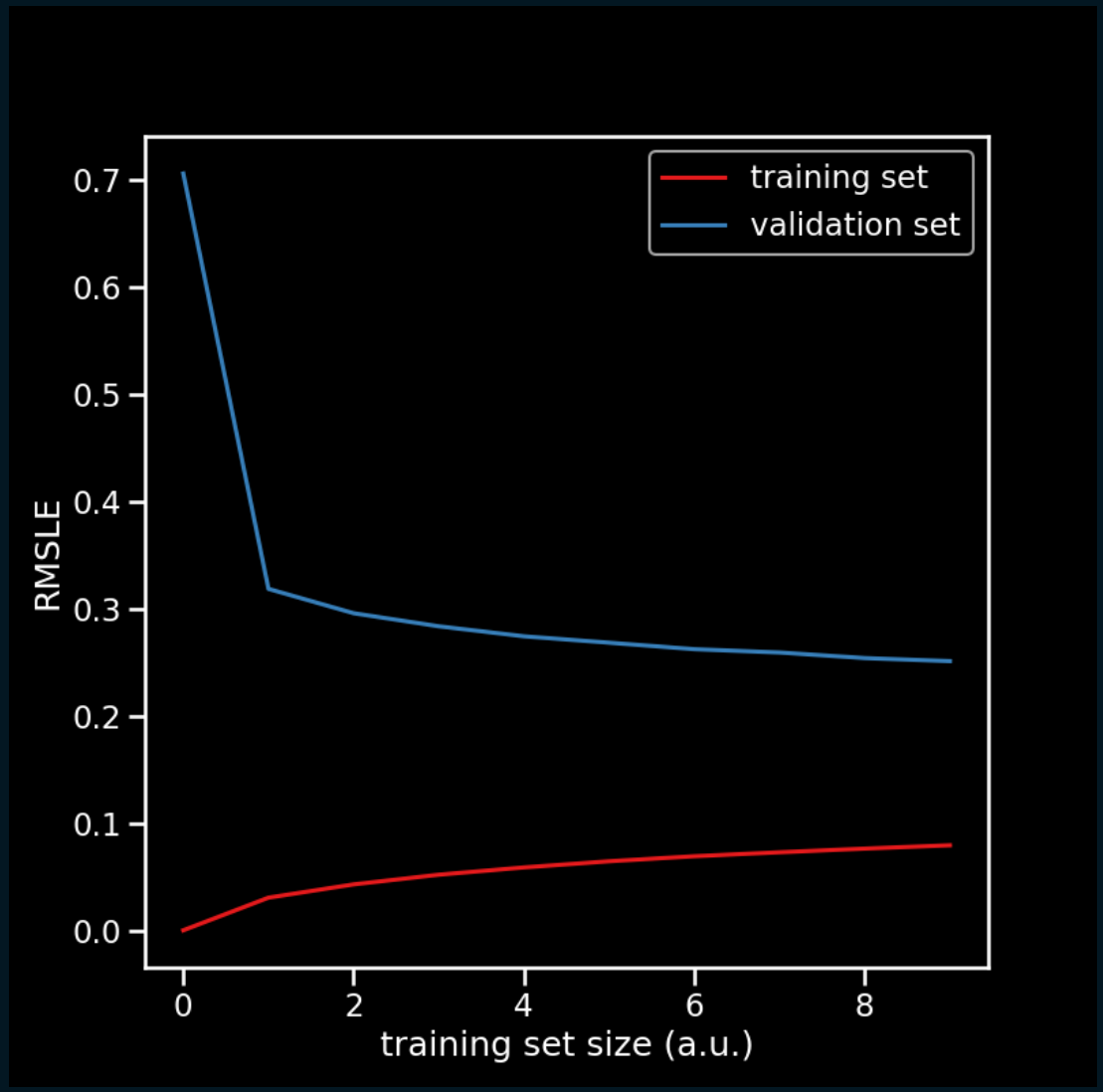
I expect that the data set could be analyzed well with a model based on decision trees, or a neural network. Because I would like to gain insight in importance of the individual features, I choose to work with decision trees.

Details on the used Decision Tree

- Because I expect that the influence of the features could be very different for electricity, chilled water, steam and hot water readings, I trained 4 individual trees.
- In order to optimize the Root Mean Logarithmic Error, the target values fed to the model are the ln(meter reading + 1), instead of the raw meter reading.

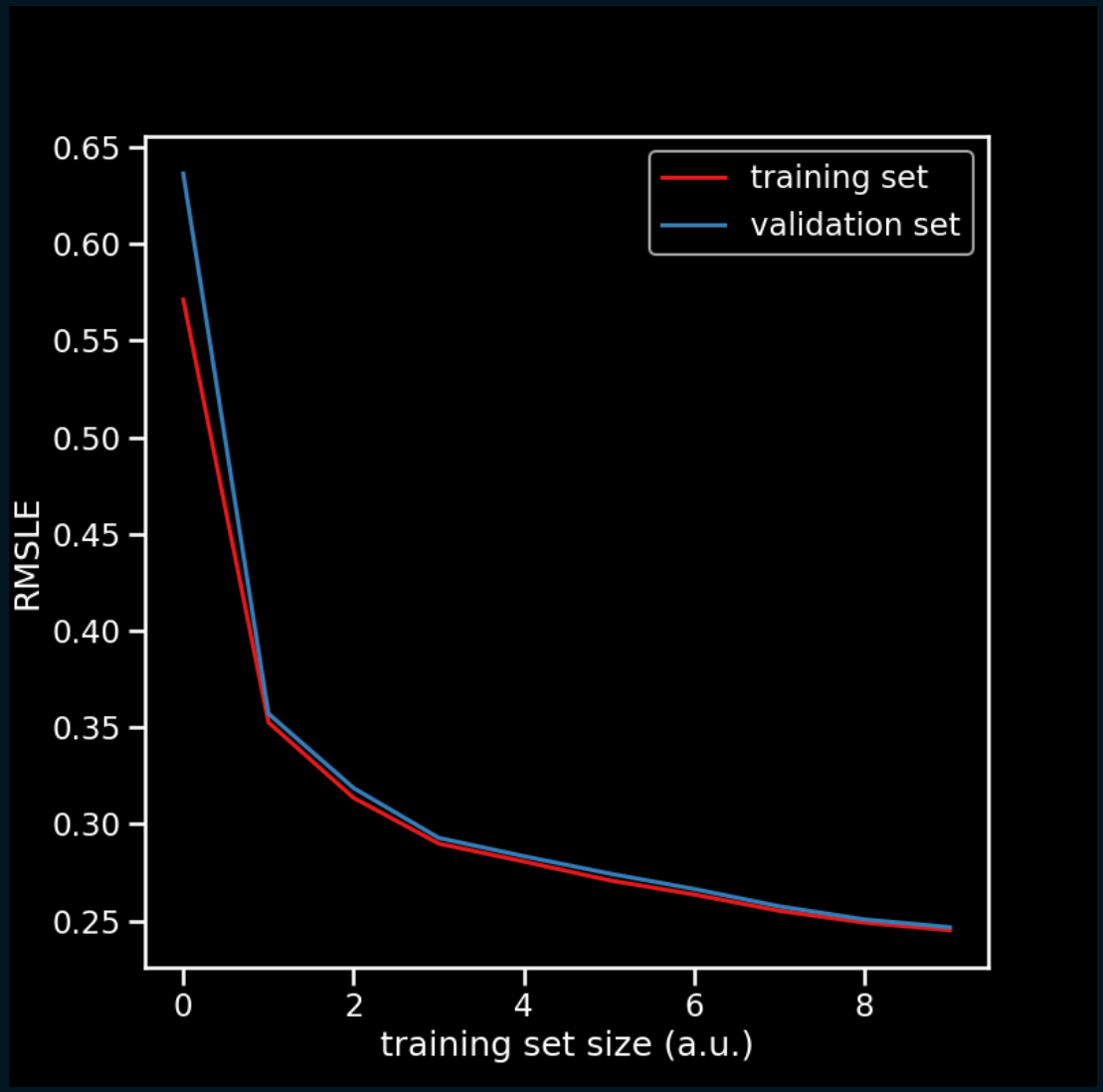
Electricity

- Plotted are the learning curves for the decision tree
- The model is overfitting the data, some form of regularization is needed. I decide to set a minimum to the samples per leaf.



Electricity

- Plotted are the learning curves for the decision tree with 100 samples minimum per leaf.
- Note that the result on the test set will likely be worse than the validation set: the validation and training set are sampled from the same time interval, while the test set is not.



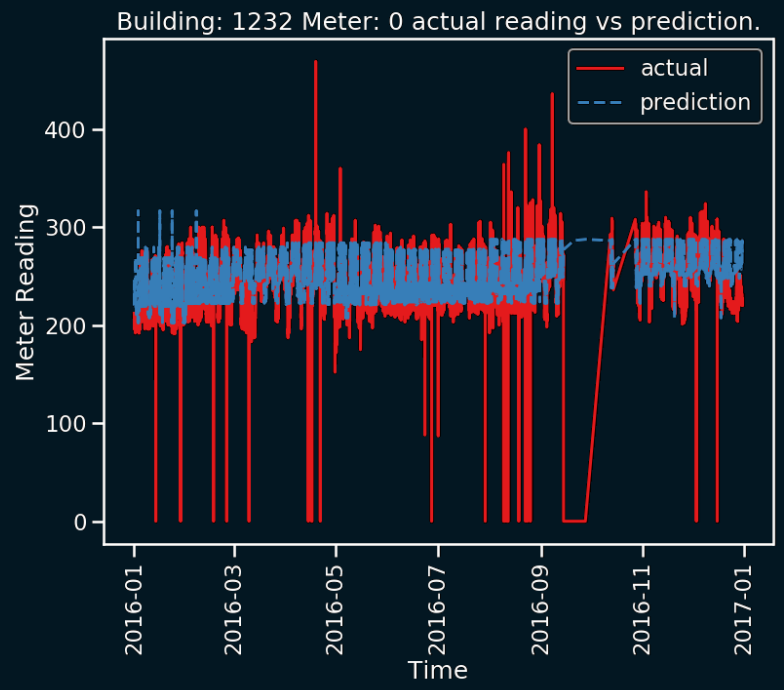
Electricity

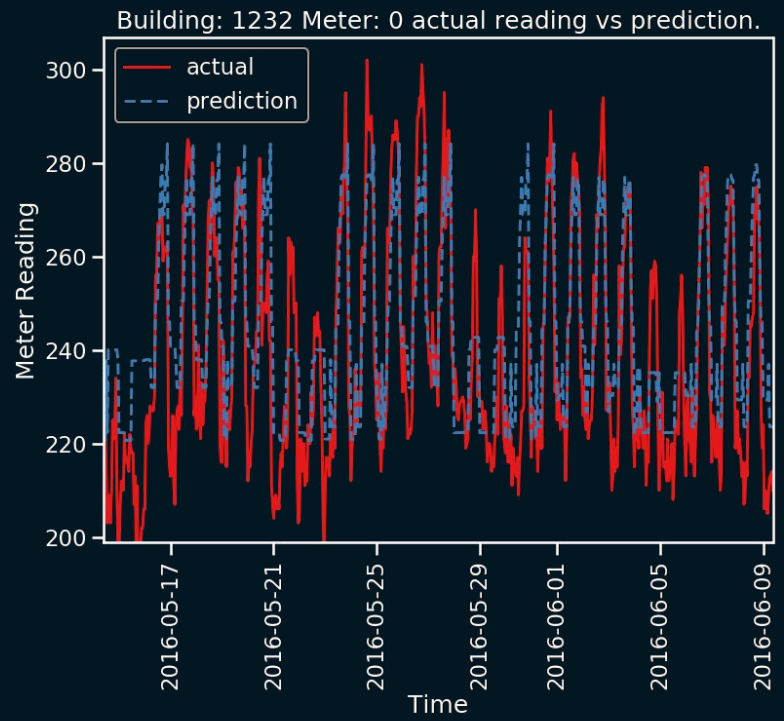
- The mean meter reading for each meter for each building is by far the most important feature
- Other important features are the standard deviation of the meter reading and the month, the rest is all below 1 percent
- Feature engineering turns out to be of utmost importance for electricity predictions, none of the original features has more than 1 percent influence.

Feature	Importance
Site ID	0.00190
Building ID	0.00477
Mean Meter Reading	0.88853
STD Meter Reading	0.02223
Meter Usage	0.00808
Reading Weather Based	0.00264
Hour	0.00930
Month	0.03748
Work	0.00272
Primary Use	0.00209
Year Built	0.00163
Square Feet	0.00237
Floor Count	0.00504
Volume	0.00187
Air Temperature	0.00519
Relative Humidity	0.00118
Cloud Coverage	0.00006
Wind Speed	0.00015
Sea Level Pressure	0.00135
Precip Depth	0.00004

Actual vs prediction electricity

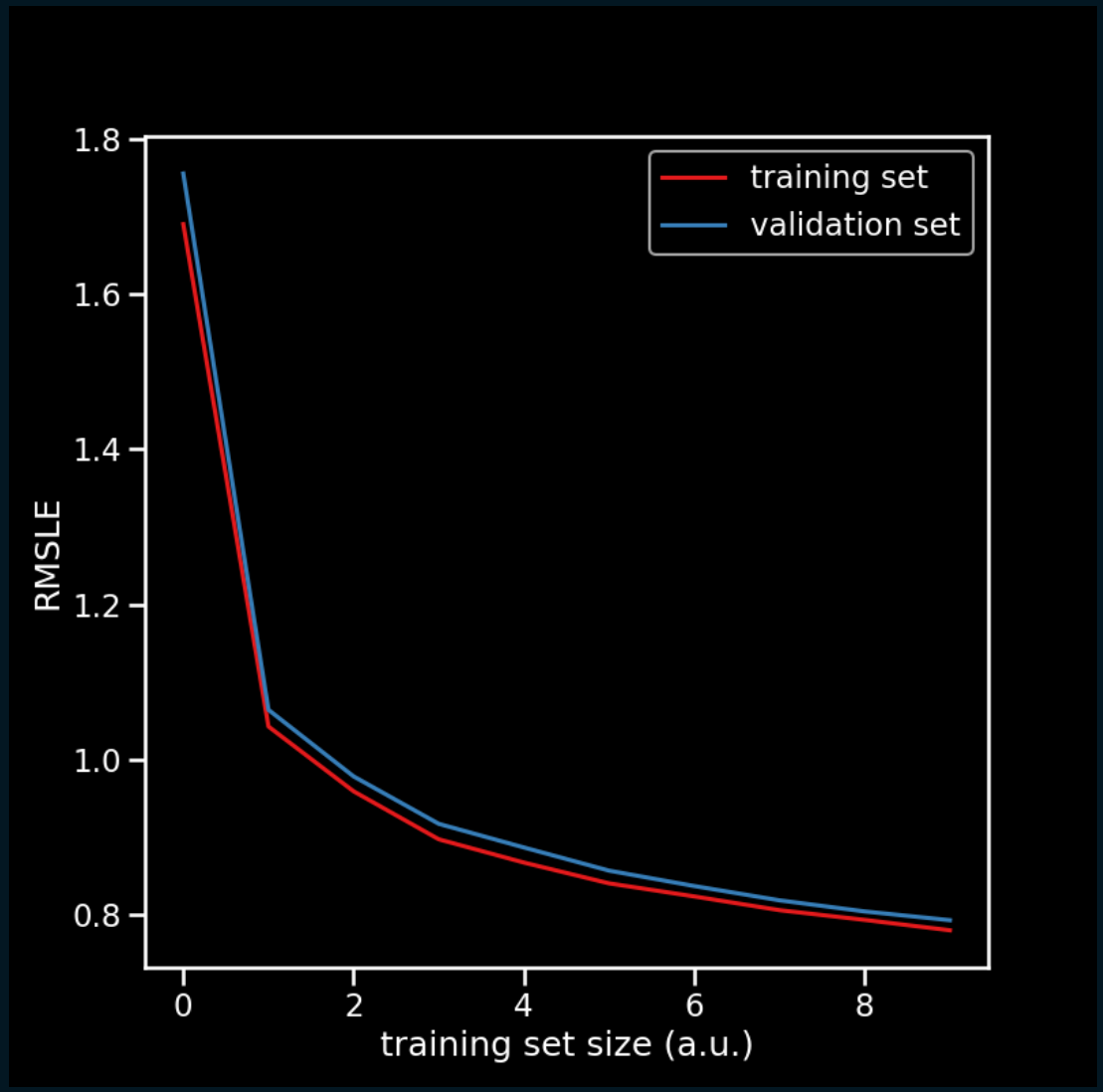
- The prediction contains no outliers (this is good, otherwise the model would not generalize well)
- The predictions follow the general trends in the data, which is especially clear in the bottom graph where a small time period is plotted.
- RMSLE on the validation set was **0.24**





Other meters

- Learning curves for chilled water
- Note that the curves saturate at a significantly higher value than for the electricity readings
- The curves for the training and validation set are close together, no signs of overfitting



Other meters

- The feature importances for chilled water are shown (the importances for steam and hot water are comparable and are not shown).
- The most important features are the mean meter reading per building, the meter usage the reading weather based and the month.
- For chilled water, steam and hot water the RMSLE scores were **0.78**, **0.89** and **1.19.**

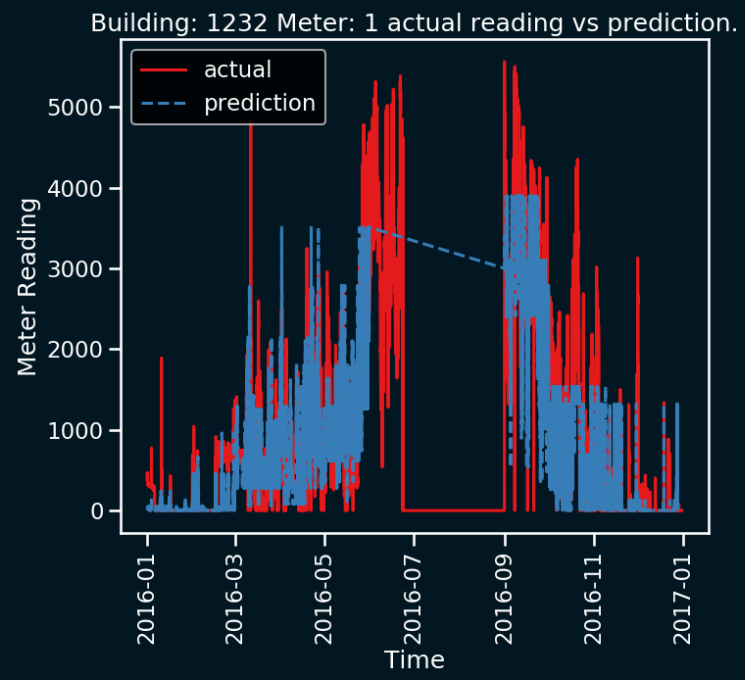
Feature	Importance
Site ID	0.00263
Building ID	0.01856
Mean Meter Reading	0.47785
STD Meter Reading	0.02192
Meter Usage	0.17034
Reading Weather Based	0.09747
Hour	0.02918
Month	0.06114
Work	0.00441
Primary Use	0.00518
Year Built	0.00893
Square Feet	0.00788
Floor Count	0.00022
Volume	0.00842
Air Temperature	0.07725
Relative Humidity	0.00375
Cloud Coverage	0.00024
Wind Speed	0.00039
Sea Level Pressure	0.00345
Precip Depth	0
Airco	0.00069

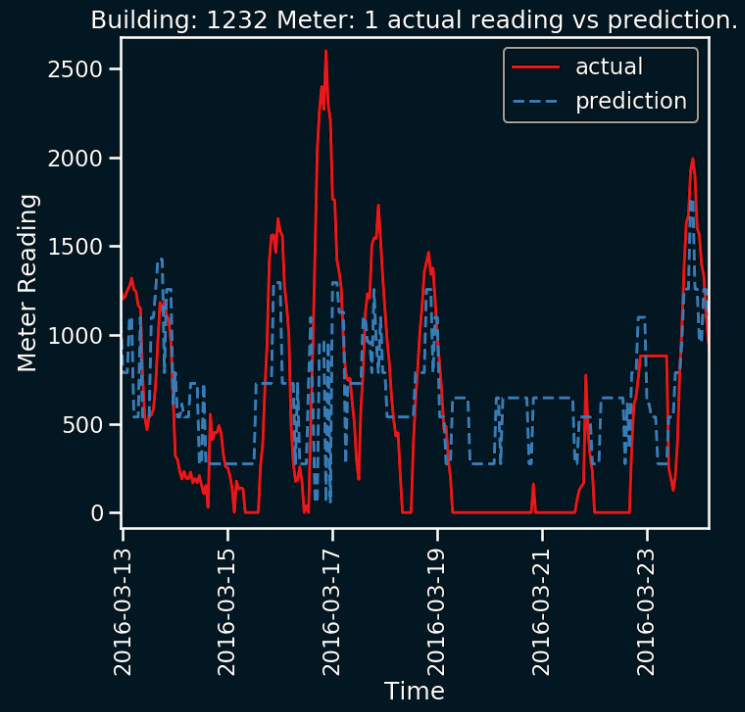
Actual vs prediction chilled water

Some general trends in the data are reproduced by the predictions, for instance the difference between the different months.

However, when we look at a smaller time scale, we see that the daily trends are not captured by the model.

Steam and hot water reading have similar problems.





Score on Kaggle competition

- RMSLE score on test set = **1.23** (number obtained by submitting the results on kaggle.com)
- Judging from the forum discussions, the other participants have focused on identifying the actual sites and finding their energy readings online. This could result in a huge improvement of the score.

CONCLUSIONS

Conclusions

- Feature engineering was of major importance in this project
- Most important features according to decision tree:
  - *Mean energy reading of the building*
  - *Meter usage*
  - *Meter reading weather based (so indirectly air temperature)*
- The prediction of the electricity reading is much better then prediction on chilled water, steam or hot water.
- RMSLE score on test set = **1.23** (number obtained by submitting the results on kaggle.com)