Capstone 2 Great Energy Predictor

# Introduction

## Background

Significant investments have been made to improve building efficiencies to reduce costs and emissions but assessing the value of energy efficiency improvements can be challenging as there's no way to truly know how much energy a building would have used without the improvements. Current methods of estimation are fragmented and do not scale well or are not generalizable to different specific meter types across various building types.

There are two key elements to develop energy saving: first we need to be able to forecast future energy usage without improvement, and we also need to predict energy use after a specific set of improvement have been implemented. One issue hindering the growth of energy markets are the lack of cost-effective, accurate and scalable procedures for forecasting energy use.

The best we can do is to build counterfactual models. Once a building is overhauled the new (lower) energy consumption is compared against modeled values for the original building to calculate the savings from the retrofit. More accurate models could support better market incentives and enable lower cost financing.

## Potential clients

With better estimates of these energy-saving investments, large scale investors and financial institutions will be more inclined to invest in this area to enable progress in building efficiencies.

## Data source

Data across four energy types based on historic usage rates and observed weather were collected.1 This dataset includes three years of hourly meter readings from over one thousand buildings at several different sites around the world.

train.csv (information of 2016)

* building\_id - Foreign key for the building metadata.
* meter - The meter id code. Read as {0: electricity, 1: chilledwater, 2: steam, 3: hotwater}. Not every building has all meter types.
* timestamp - When the measurement was taken
* meter\_reading - The target variable. Energy consumption in kWh (or equivalent). Note that this is real data with measurement error, which we expect will impose a baseline level of modeling error. (UPDATE: the site 0 electric meter readings are in kBTU)

building\_meta.csv

* site\_id - Foreign key for the weather files.
* building\_id - Foreign key for training.csv
* primary\_use - Indicator of the primary category of activities for the building
* square\_feet - Gross floor area of the building
* year\_built - Year building was opened
* floor\_count - Number of floors of the building

weather\_train csv (information of 2016)

Weather data from a meteorological station as close as possible to the site.

* site\_id
* air\_temperature - Degrees Celsius
* cloud\_coverage - Portion of the sky covered in clouds, in oktas
* dew\_temperature - Degrees Celsius
* precip\_depth\_1\_hr - Millimeters
* sea\_level\_pressure - Millibar/hectopascals
* wind\_direction - Compass direction (0-360)
* wind\_speed - Meters per second

test.csv and weather\_test.csv contain similar information of 2017 and 2018 without meter reading.

# Exploratory Data Analysis

## Basic plots

The most commonly seen meter in train.csv is electricity meter, followed by chilledwater meter and steam meter, and the least frequent one is hotwater meter (Figure 1).

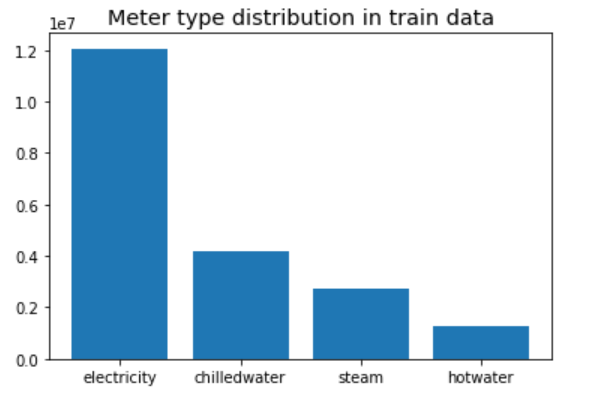


Figure 1 Distribution of meter types in train.csv

As there are a lot of zero values for meter reading in training data and not every building has all four types of meters, we plotted the meter reading of each building along each hour elapsed since the earliest time in training data according to the meter type (Figure 2). In Figure 2, Meter 0 is electricity meter, Meter 1 is chilledwater meter, Meter 2 is steam meter and Meter 3 is hotwater meter; X axis refers to the hours elapsed since 2016-01-01, and Y axis corresponds to building\_id. The yellow streaks mean that meter reading values are not zero, green streaks mean that meter readings are zero, and gray areas indicate that the meter type is not available for the building, which could be caused as certain type of meter is not available in the building, or that the reading is not recorded in dataset for certain time in the building.

We also investigated the building\_meta.csv for original building information. Among the buildings in building\_meta.csv, the most common ones are used for education and office, and the least common ones are for religious worship and utility (Figure 3). Distribution of sizes for buildings with different primary uses is plotted using the violin plot (Figure 4). Buildings for retail, food sales and service, religious worship, utility, technology/science, manufacturing/industrial, warehouse/storage and others have relatively small span of size distribution, while buildings for the rest of uses differ very much in size. Not all buildings have information with regard to building years and floor counts, but for those with known built years and floor counts, a large percentage was built in 1960-1980, meaning they are 36-56 years old when the meter reading data were collected, and most buildings have less than 10 floors.

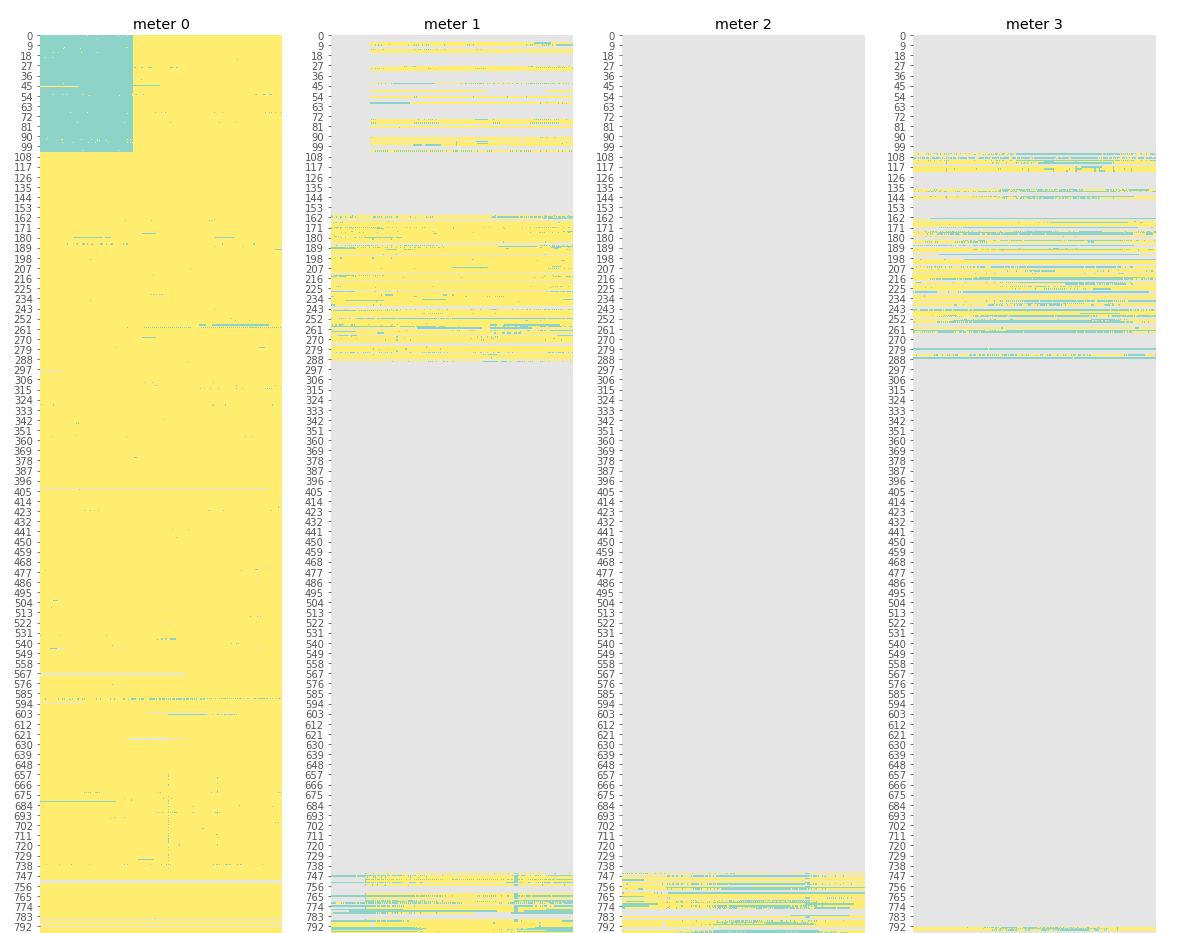
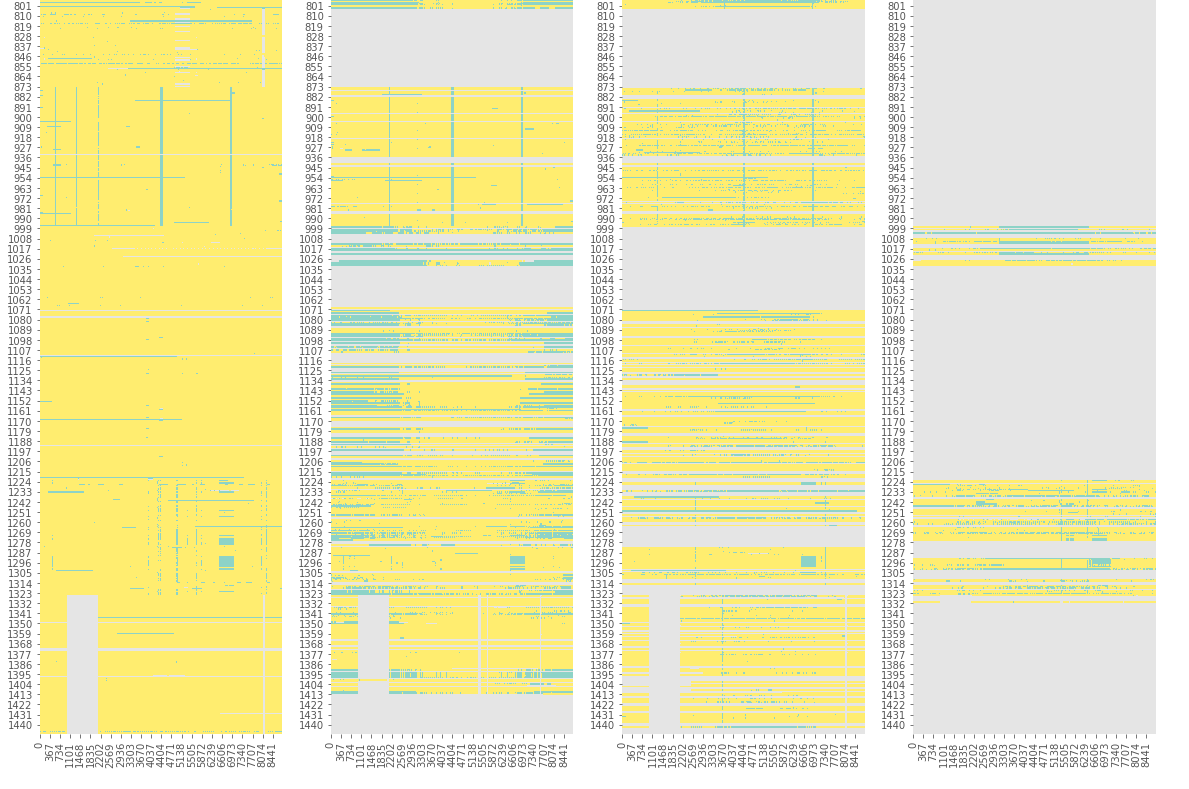


Figure 2 Meter reading of each building by hour since 2016-01-01 in training data per meter type

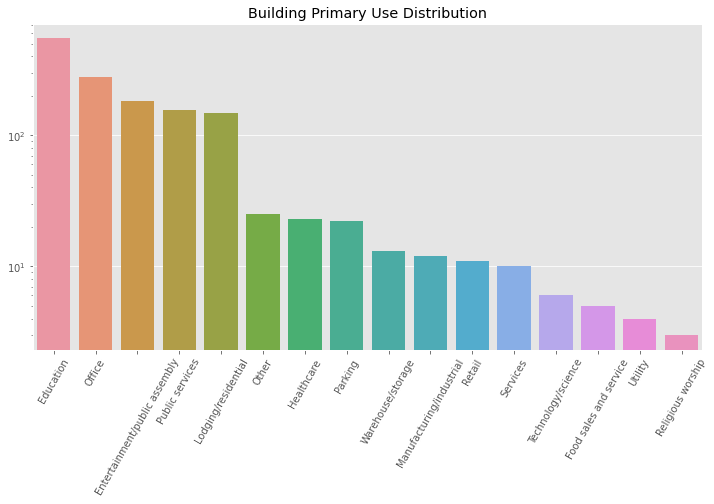


Figure 3 Primary use for buildings in building\_meta.csv

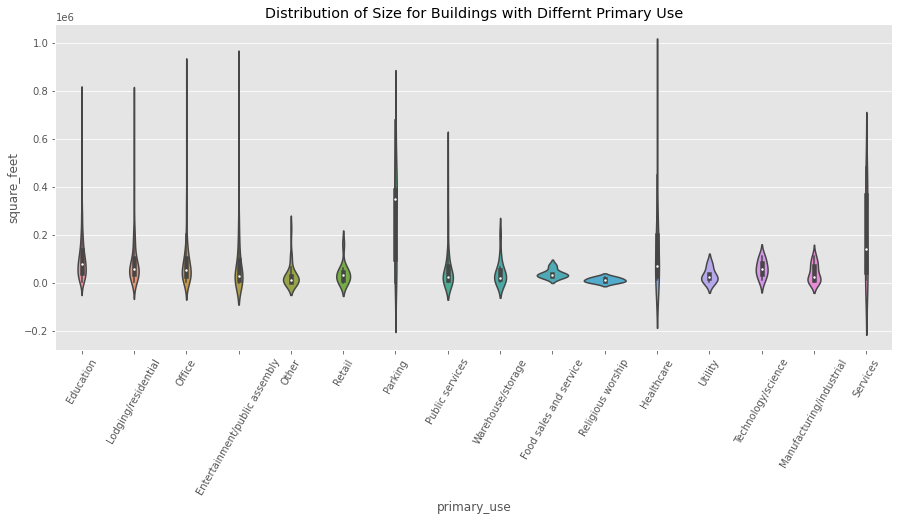


Figure 4 Distribution of sizes for buildings with different primary uses

There are 16 different sites in the weather data, and the average daily air temperature by site is plotted as below (Figure 5). We can roughly see that Site 2 had the highest overall temperature in 2016; Site 15 and 13 had very cold winter in 2016; Site 11 and 7 had identical air temperature and were very cold in February and December. Average daily precipitation is also plotted using line plots (Figure 6). Site 15 had relatively more precipitation among all sites and Site 14 seems dry. Rain was more frequent from June to November in 2016. There are a lot of missing values for Site 15, 11, 7 with regard to precipitation. And precipitation data of Site 1, 5 and 12 are not reported.

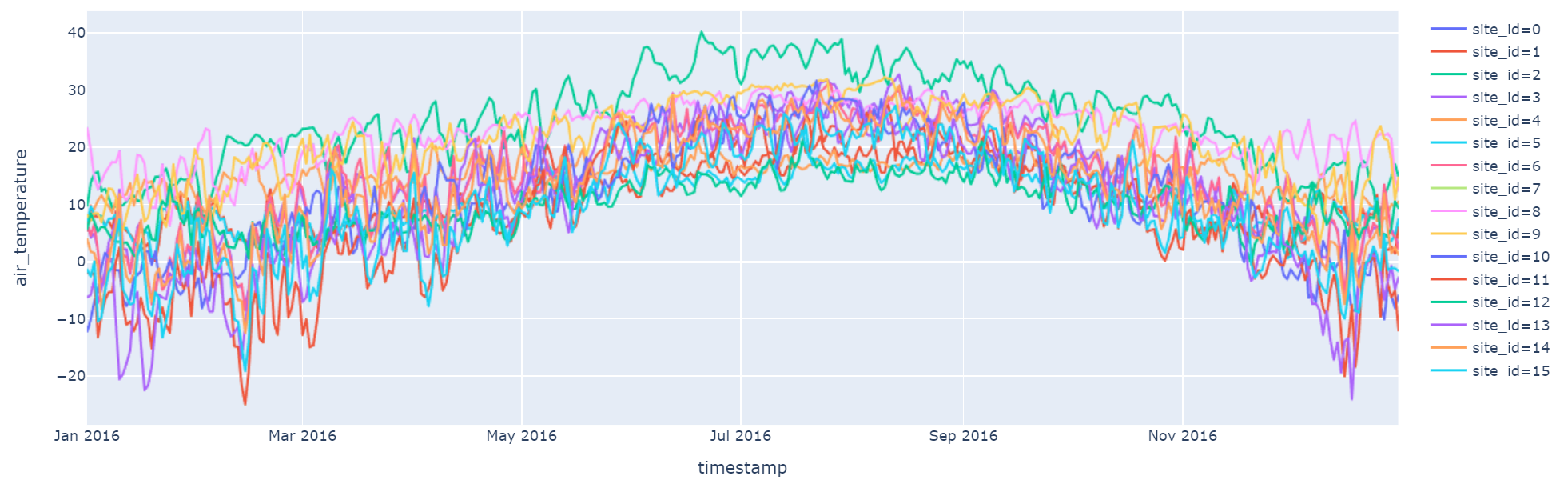


Figure 5 Average daily air temperature by site in weather\_train.csv

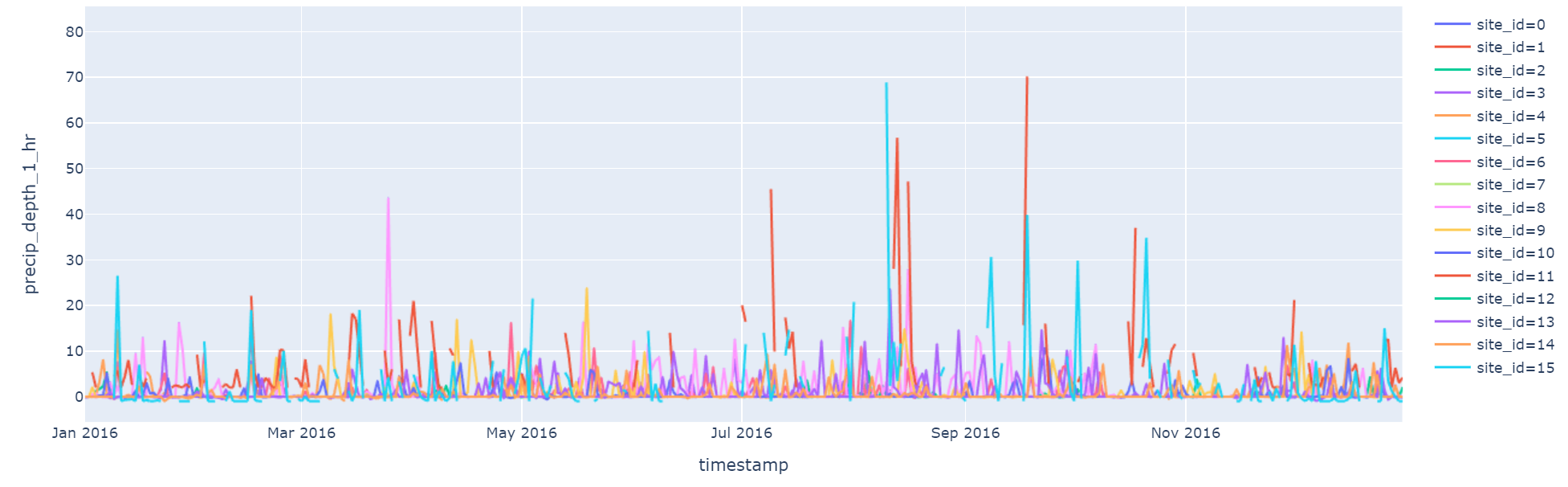


Figure 6 Average daily precipitation by site in weather\_train.csv

## Identifying outliers

According to the Department of Energy (DOE), the average number of kilowatt hours per square foot (kWh/ft2) for a commercial building in a year is approximately 22.5.2 However, there are some abnormal data that are significantly higher than the average yearly energy consumption rate.

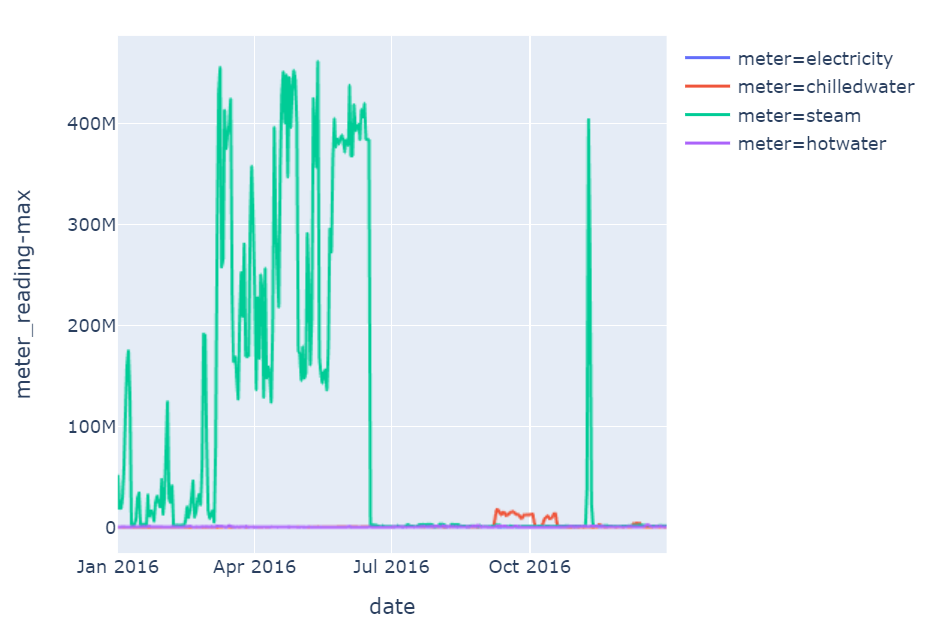
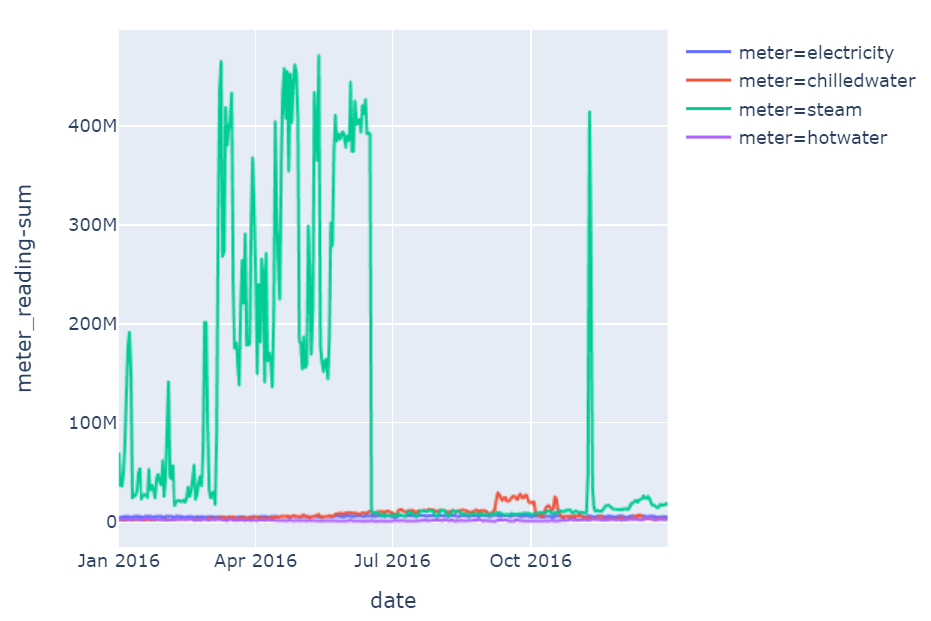


Figure 7 Total daily energy consumed in kwh (left) as per meter type and max daily energy consumed in kwh (right) as per meter type

In Figure 7, the total daily energy consumed in 2016 by all buildings is compared to the maximum of daily energy consumed by a single building in 2016 per energy type. It is clear that there are some extreme values for the total consumed energy and the maximum consumed energy, especially for steam, and it seems that the total and the maximum have similar aberrant patterns, suggesting that only one building for each energy aspect caused the extreme peaks for each day.

For electricity energy, Building 803 often used the most energy daily, followed by Building 801, 799, 1088, 993 and 794. Building 803 consumed about 30,000,000kWh and the building size is approximately 190,000 ft². This is almost 8-fold of the normal average consumption reported by DOE. Building 803 is primarily used for education and should not have such high industrial electricity consumption. This piece of data seems incorrect. Similar abnormal data are seen for Building 801, 799 and 1088. Building 993, as an education center, consumed around 30kWh/ft²/year in 2016, which is reasonable. However, it shows abnormal peaks in July, October and February for electricity meter, and the same abnormal pattern are seen for other types of meter, which could be caused by the start of new semesters. It also has some vacational pauses in July and December. For Building 794, there is a very steady daily consumption with a weekly pattern, and without a seasonal pattern. It consumes 50 kWh/ft²/year, which is about twice the typical consumption. This piece of data seems reasonable.

For steam energy, Building 1099, 1197, 1168 and 1148 often used the most energy daily. Building 1099 have really high steam consumption from January to June and in November. There was no obvious pattern and the data was noisy and this piece of data seems wrong. Other than Building 1099, it seems that Building 1197 has some extreme highs in January and February. Building 1197 has some weekly pattern from June to September, but for the other dates the consumption pattern is very noisy. 8000kWh/ft²/year of steam energy was consumed for this building in 2016, which seems incorrect, and steam should not have been used this much in summer for heating and hot water. Building 1168 has some extreme high consumption from July to September, and in December. Additionally, a 1000000kWh daily average is too big up to a 500000 ft² office. Similar aberrant numbers are seen for Building 1148.

For chilledwater consumption, Building 1284 often has the most daily consumption. But in the maximum chilledwater daily consumption plot (Figure 8), Building 778 and 1088 actually have the abnormal high chilledwater consumption. Building 1088 has very high peaks from September to December, which are about a hundred times more than other times, and this is very abnormal. Building 778 has really high chilled water consumption for a short span of two months and the rest of the time consumed almost no chilled water. There is probably something wrong with the measurement.

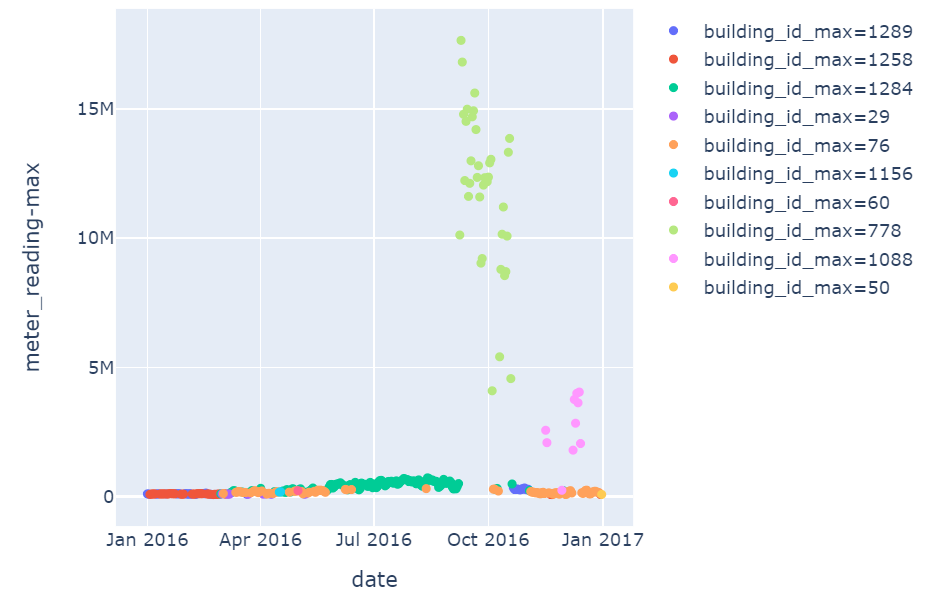


Figure 8 Maximum chilledwater consumption values per day

For hotwater, Building 1021 and Building 1331 have aberrant hot water use. The energy consumption is too large for an entertainment/public assembly building like Building 1021 or for an education center like Building 1331.

# Data Wrangling

## Fixing wrong data

As claimed in <https://www.kaggle.com/c/ashrae-energy-prediction/discussion/119261>, the electric meter reading for site 0 was reproted in kBTU and needs to be converted to units of kWh through multiplying by 0.2931.

## Removing outliers

Taking only the buildings that consume maximum energy per day per type, we can see that there are a lot of measure scale errors.

The error could be caused as:

* The meter is not configured correctly.
* The software has not the units configured correctly.
* The software has not the decimal digits configured correctly.
* Some changes over time irregularly, indicating that some buildings have more than one meter. One error in one meter and the overall measure is unreliable.

This capstone project has only identified outliers that influence the maximum consumption on a daily basis, which is only the tip of the iceberg. A sound analysis should be done to detect and correct these outliers.

Energy consumption outliers per energy aspect that should be removed are:

* electricity- Building 803, 801, 799, 1088
* steam- Building 1099, 1197, 1168, 1148
* chilledwater- Building 1088, 778
* hotwater- Building 1021, 1331

## Handling missing values in weather data

Weather\_train.csv has hourly weather information for 16 sites in 2016, and it should have 140,544 records (16 x 24 x 366, 2016 is a leap year). But this csv only has 139,773 records, meaning 771 hours of data is missing. Weather\_test.csv should have 280320 hours for 16 sites for 2017 and 2018 together, and this csv file is missing 3077 hours. Therefore, we need to align the time by hour and fill in the missing values for weather data.

The missing weather hourly data was first filled by the daily mean, and this method was able to fill in all missing values of air\_temperature, dew\_tempertature, wind\_speed and wind\_direction. Then the missing values were substituted with weekly average and monthly average. But sea\_level\_pressure, cloud\_coverage and precip\_depth\_1\_hr have larger percentage of missing values and can't be filled completely with mean values. Information from sites with similar weather conditions were used to fill in missing values of these features for both training and testing data. Site 1, Site 5 and Site 12 have similar weather conditions, and they resemble Site 14; Site 11 and Site 7 are identical, and they are similar to Site 15. Therefore, the missing sea\_level\_pressure values in Site 5 were filled with monthly average of sea\_level\_pressure in Site 1; cloud\_coverage data from Site 15 are used to fill in the missing values of Site 11 and Site 7; Precipitation data from Site 14 are used to fill in NaNs in Site 1, 5, 12.

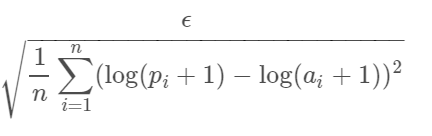
## Handling zeros

After removing max outliers, there are still 9% meter-readings that are zero. The rows that read zero for electricity meters should be dropped because it is not possible for a building to ever have zero electrical usage. Then we identify 48+ hour runs of steam and hotwater meter zero-readings as bad zeros unless they happen during warmer months, because it is common for people to turn off steam or hot water for wall heating during warm days (Here we set warm days as March 1st to Oct 31st). Then bad zero readings for chilledwater meter are identified as 48+ hour continuous zeros in hot months, because chilled water is normally highly consumed in summer for cooling. Here the hot months are set as from June 20th to Sep 20th.

# Data Modeling

The floor\_count and year\_built features are dropped for modeling because there are too many missing values for these features, and another new feature is added to determine if the day is a holiday because holidays might affect energy consumption.

To access evaluation metric RMSLE (where *ϵ* is the RMSLE value (score), *n* is the total number of observations in the (public/private) data set, *pi* is your prediction of target, *ai* is the actual target for *i*, *log(x)* is the natural logarithm of *x*):



target y is transformed to np.log1p(y) which calculates ln(1 + y).

## Linear regression benchmark model

Linear Regression model from sklearn is used as the benchmark model; LinearRegression model has the score of 1.833 for testing data and 1.832 for training data, which is not much overfitting. The correlation coefficients for features are plotted in Figure 9. However, this model does not predict the values very accurately (Figure 10).



Figure 9 Feature importance of linear regression model

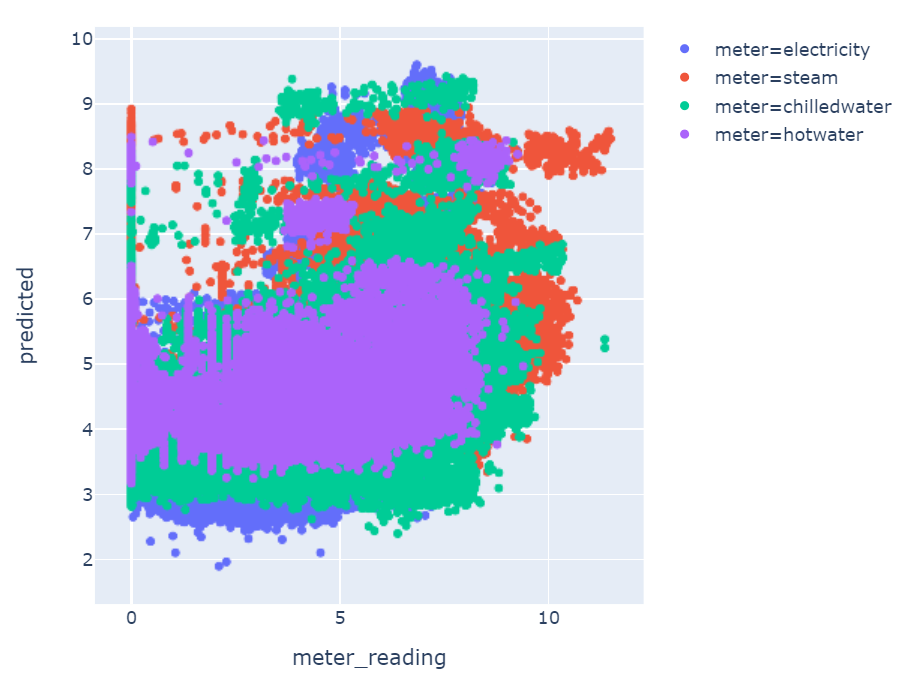


Figure 10 Predicted values vs original values of meter readings using the simple linear regression model

## Light gradient boost machine model

Light gradient boost machine models are then tried because they can deal with categorical data and have higher predicting capacity. Before tuning hyperparameters, the Light GBM model has the score of 0.454 for the testing data and 0.300 for the training data. There is some overfitting, but the prediction is much enhanced (Figure 11). Feature importance is plotted in Figure 12, where dayofyear, building\_id, meter, air\_tempertuare, dew\_temperature, hour and sea\_level\_pressure are important.

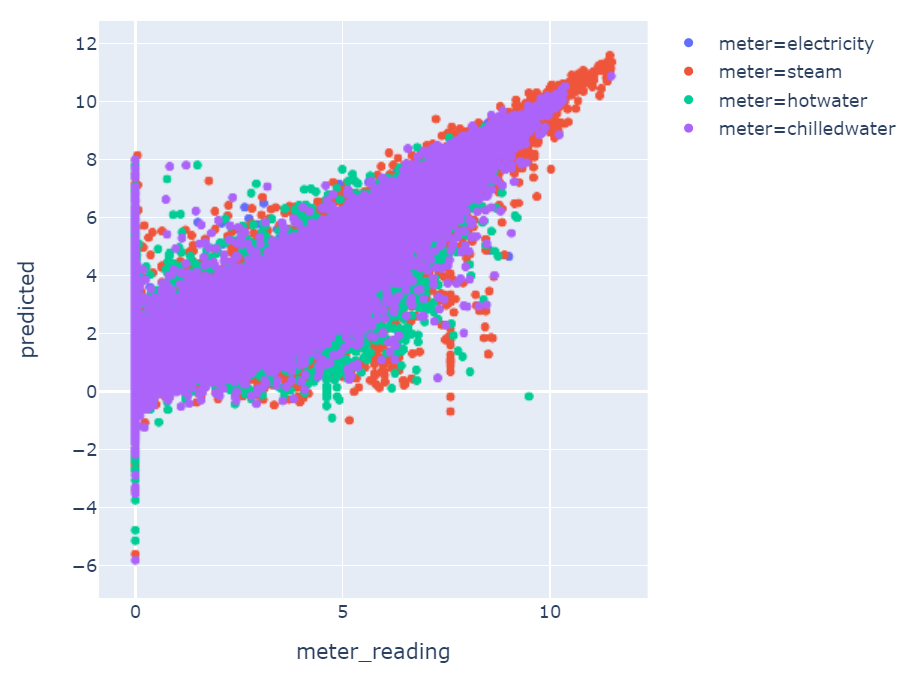


Figure 11 Predicted values vs original values of meter readings using the Light GBM model without hyperparameter tuning

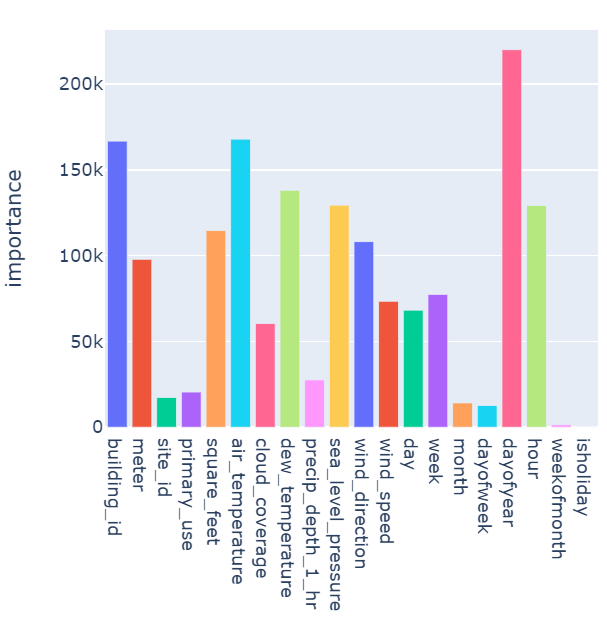


Figure 12 Feature importance of Light GBM model

## Light GBM half-half model

Half-half model is to divide the training data into two parts, and then train one model with the 1st half and validate with 2nd half, and then train another model with 2nd half and validate with 1st half. The prediction test data is done by averaging results from both models. Before tuning hyperparameters, the half-half model has score of 0.469 for testing data and 0.351 for training data. The overfitting is lessened. The predicted values are plotted against the original values (Figure 12).

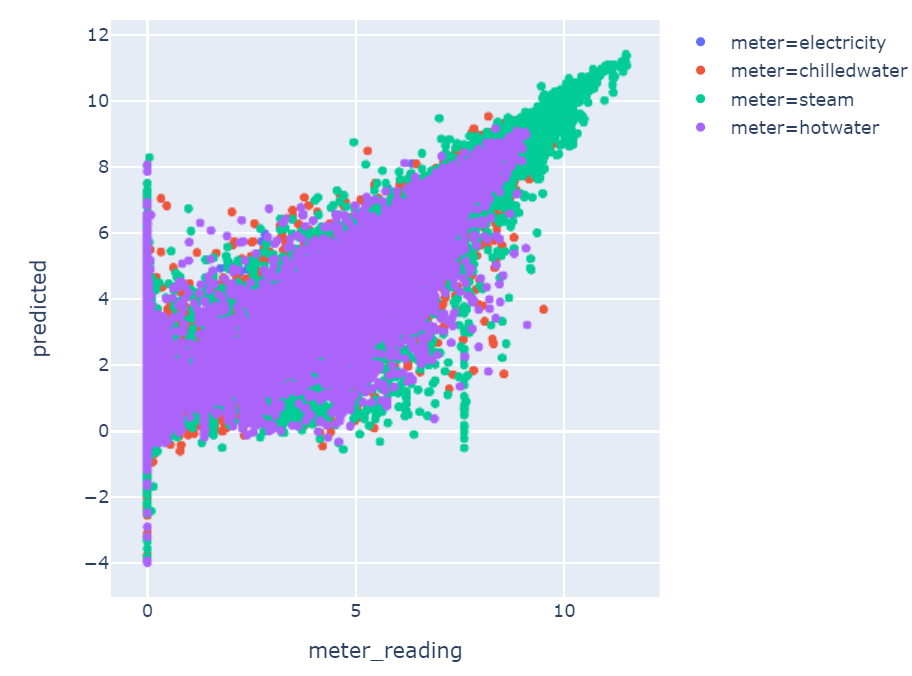


Figure 13 Predicted values against original meter reading values from half-half model without hyperparameter tuning

## Refining dataset

As there are still many zero-readings that can't be predicted by the model correctly, the train dataset would be refined again to eliminate more incorrect zero-readings. Before we set warm days for March 1st to October 31st for steam and hotwater, now we are going to narrow down the warm days (above 10 degrees) for cold regions according to air temperature reported by each station.

* Site 1: May 1st to Oct 31st
* Site 5: May 1st to Oct 31st
* Site 7: May 1st to Oct 20th
* Site 10: May 1st to Oct 31th
* Site 11: May 1st to Oct 20th
* Site 12: May 1st to Oct 20th
* Site 13: April 10th to Oct 31st
* Site 15: March 8th to Oct 31st

And before we defined bad zero-readings for chilledwater as from June 20th to Sep 20th (hot months as summer). Now we modified hot months for hot regions when the air temperature is above 25 degrees.

* Site 0: May 3rd to Oct 15th
* Site 2: April 14th to Nov 3rd
* Site 3: May 27st to Sep 20th
* Site 6: May 27th to Sep 20th
* Site 8: May 3rd to Oct 15th
* Site 9: May 3rd to Oct 15th
* Site 13: May 27th to Sep 20th
* Site 14: May 27th to Sep 20th

After refining the dataset, the benchmark model, Light GBM model and half-half model are re-evaluated and they all perform slightly better.

## Tuning hyperparameter

Hyperparameters tuned for Light GBM and half-half model include min\_child-samples, num\_leaves, lambda\_l1 and lambda\_l2. After tuuning we picked lambda\_l1 as 2.5, lambda\_l2 as 4, min\_child\_samples as 50 and num\_leaves as 50. Then we set smaller learning rate and more estimators. As a result, for Light GBM model, the error for training data is 0.470 and for unseen data is 0.501 This model does not have much of overfitting and would be generalizable to new data. For the half-half model, the score for testing data is 0.519 and 0.52 for training data. The prediction accuracy is slightly lowered but the overfitting is lessened.

References

1. <https://www.kaggle.com/c/ashrae-energy-prediction/data>
2. <https://www.iotacommunications.com/blog/benchmarking-commercial-building-energy-use-per-square-foot/>