MODELLING ENERGY USAGE

MACHINE LEARNING

Dongtao Jiang, Data Science Career Track, Springboard
10/12/2020

DATASET

- Data source: https://www.kaggle.com/c/ashrae-energy-prediction/data
- Features

train.csv

- building_id - Foreign key for the building metadata.
- meter - The meter id code. Read as {0: electricity, 1: chilled water, 2: steam, 3: hot water}. 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.

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 based on EnergyStar property type definitions
- square_feet - Gross floor area of the building
- year_built - Year building was opened
- floor_count - Number of floors of the building

DATASET

weather_[train/test].csv

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

The submission files use row numbers for ID codes in order to save space on the file uploads. test.csv has no feature data; it exists so you can get your predictions into the correct order.

- row_id - Row id for your submission file
- building_id - Building id code
- meter - The meter id code
- timestamp - Timestamps for the test data period

SOFTWARE PACKAGES

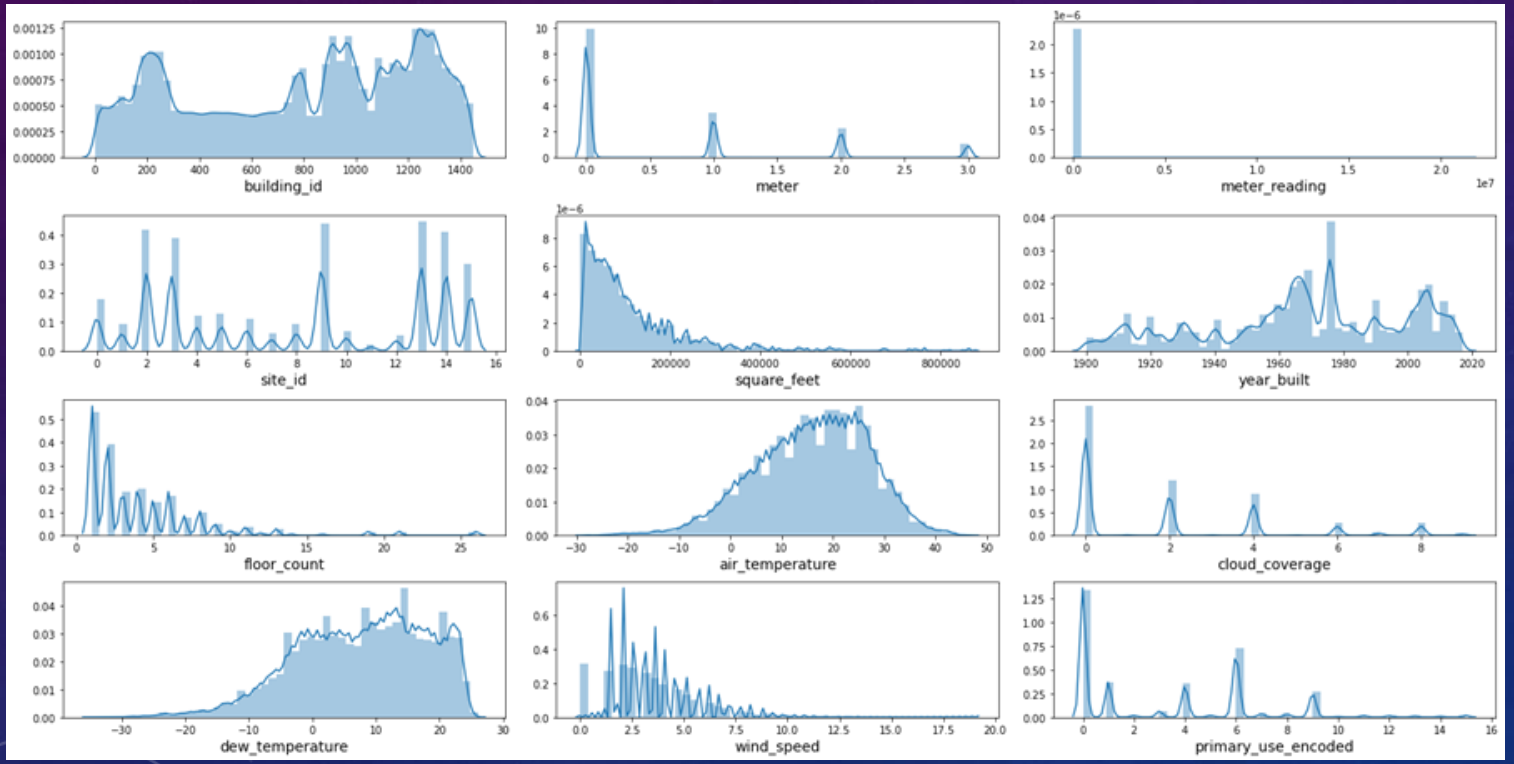
- Python 3.6
- Pandas 1.0.3
- Numpy 1.15.4
- Sklearn 0.23.1
- Matplotlib 3.2.1
- Seaborn 0.10.0
- Scipy 1.10
- Catboost 0.24.1
- LightGBM 2.3.1

DATA WRANGLING

- Conform timestamp to datetime format
- Merge datasets based on building_id, site_id and timestamp
  - train set, test set
  - weather data
  - buidings_meta_df

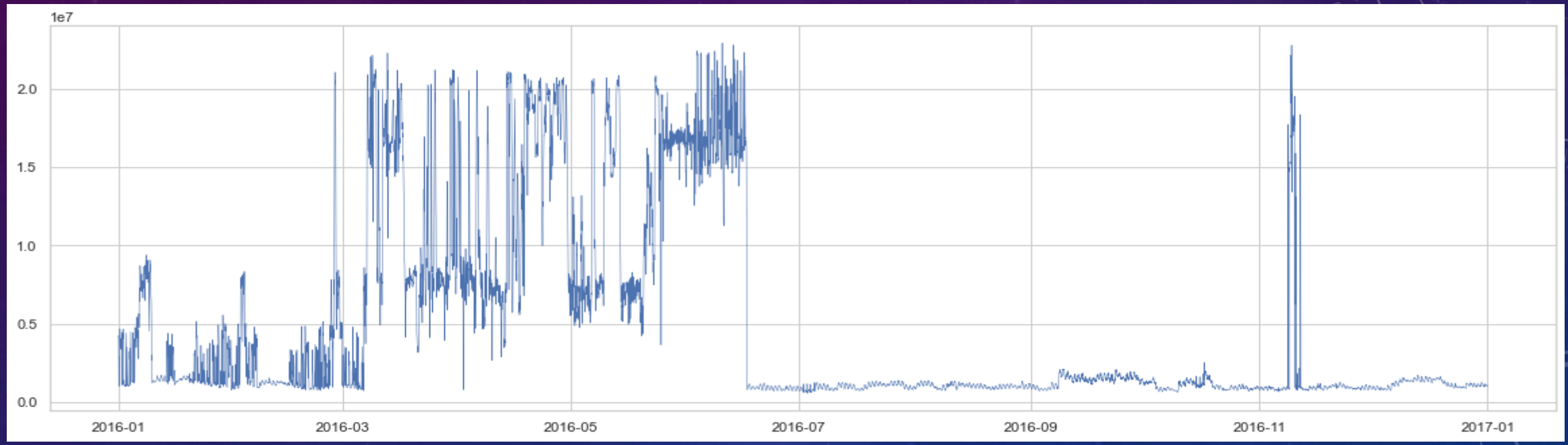
- Change data types to reduce memory usage
- Correct unit
  - Site 0 were not properly converted to units of kWh and are in kBTU
  - Multiply by 0.2931 to get to model inputs into kWh like the other sites, and 3.4118 to get back to kBTU for scoring

EXPLORATORY DATA ANALYSIS



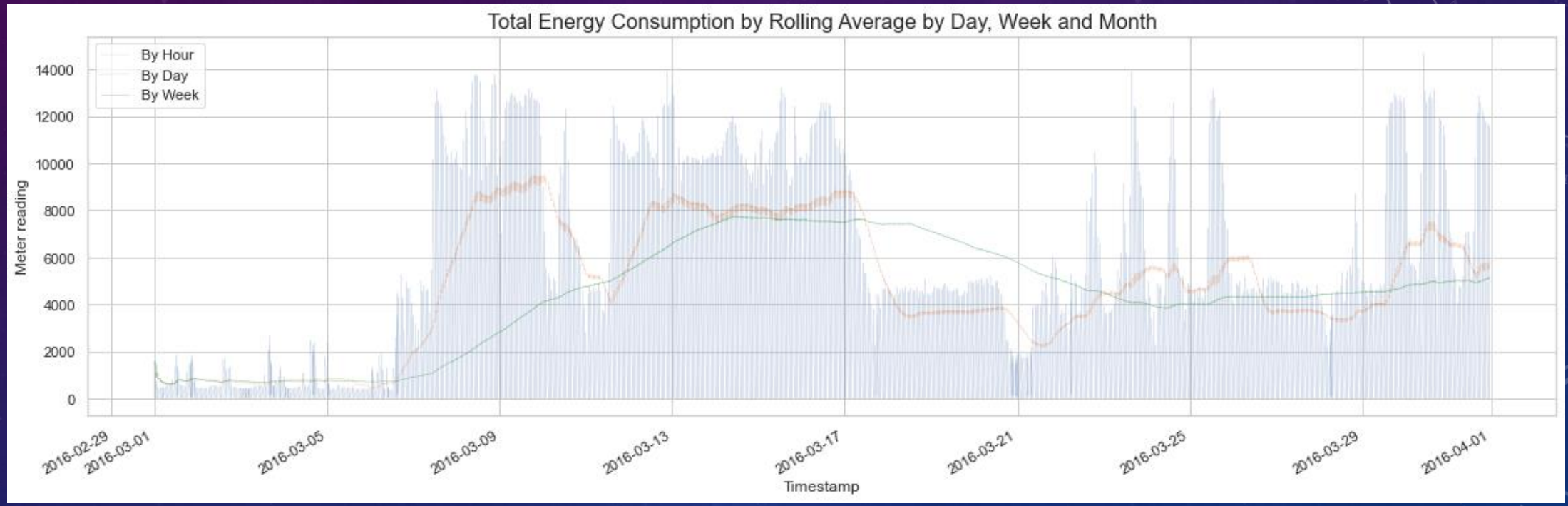
- Meter reading is severely skewed to larger numbers.
- Majority of meter readings are small values.
- Square feet, floor count, wind speed, cloud coverage is also skewed to larger values.
- There are more readings for electricity type of meters. it is followed by {1: 'chilled water', 2: 'steam', 3: 'hot water'}

ENERGY USAGE OVER A YEAR



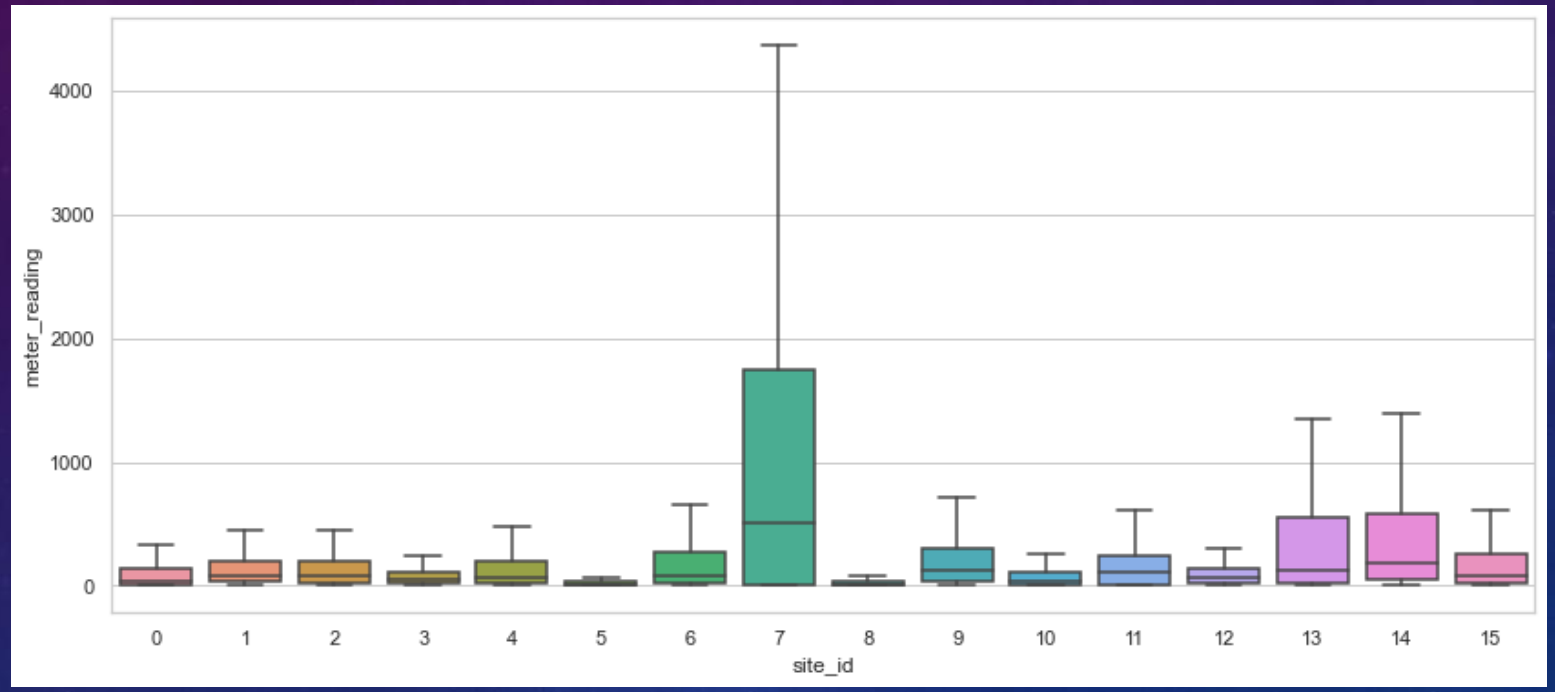
- Unusually high energy usage from March to June and in a few days in November.

ROLLING WINDOWS TO SMOOTH DATA



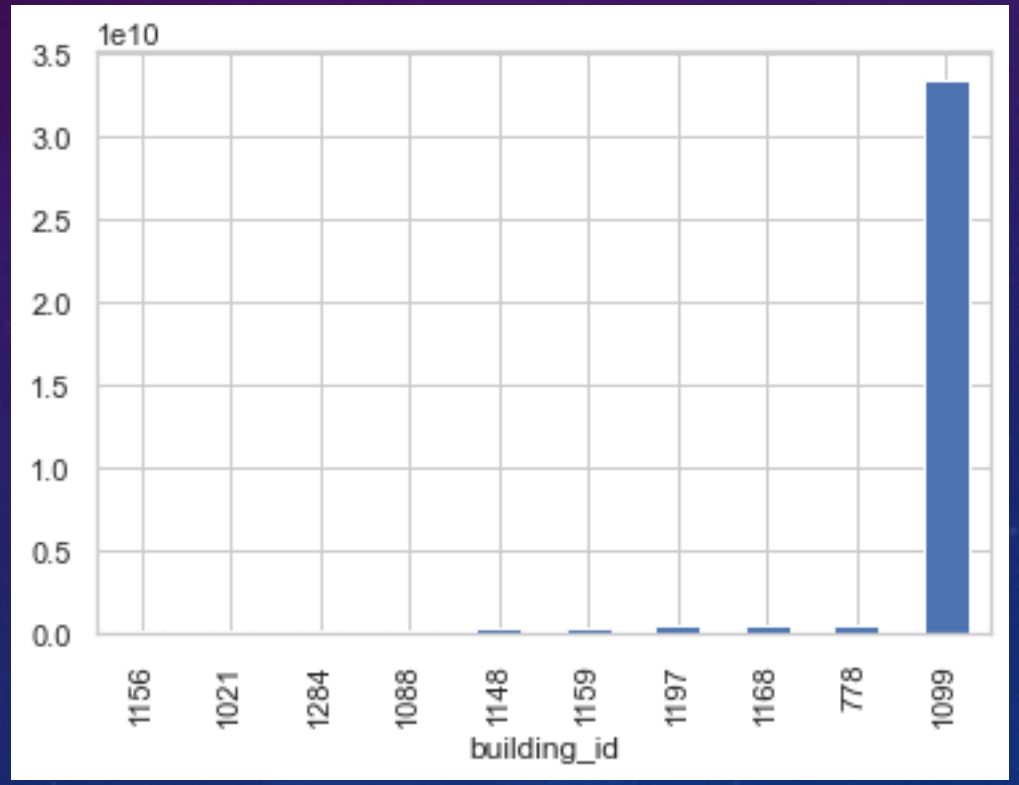
- The larger the window, the smoother the curves, and the more it lags behind.
- The rolling average will be used for modelling in the machine learning.

SITE SPECIFIC ENERGY CONSUMPTION



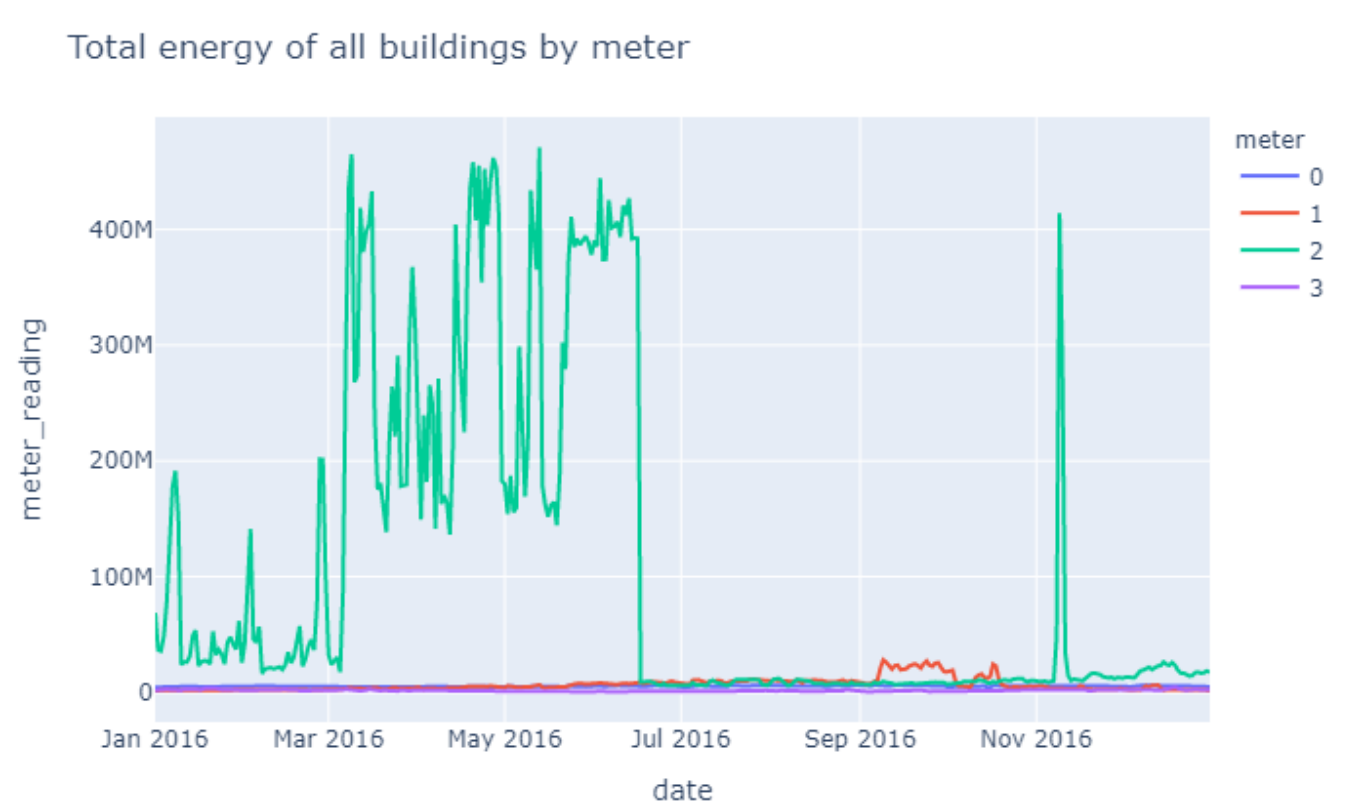
- site 7 consumes many times higher energy than other sites both yearly and monthly.

BUILDING ID SPECIFIC ENERGY USAGE



- building_id 1099 eats up much more energy than all others. It is at least about 100 times higher that of other buildings. The building 1099 belongs to site 13 and is an Educational institution.

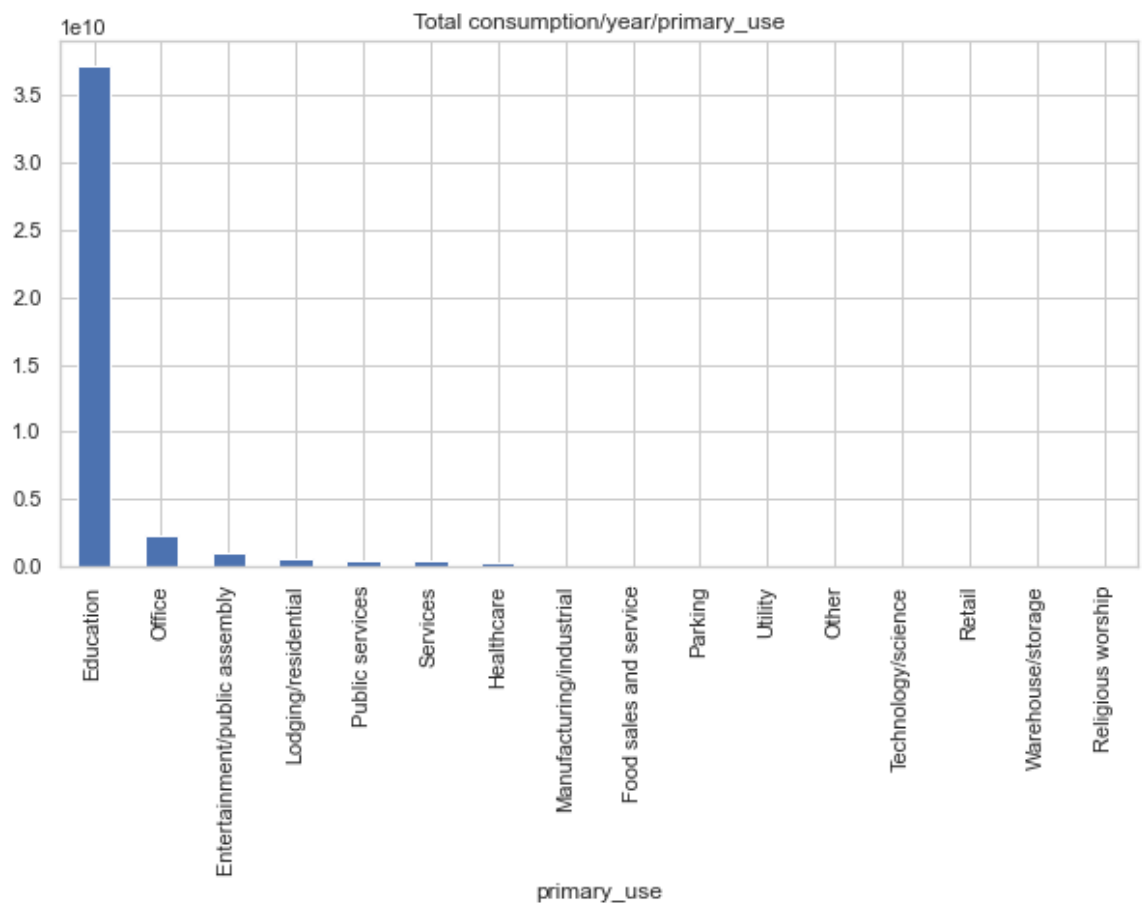
ENERGY USAGE ACCORDING TO ENERGY TYPE



0: electricity
1: chilled water
2: steam
3: hot water

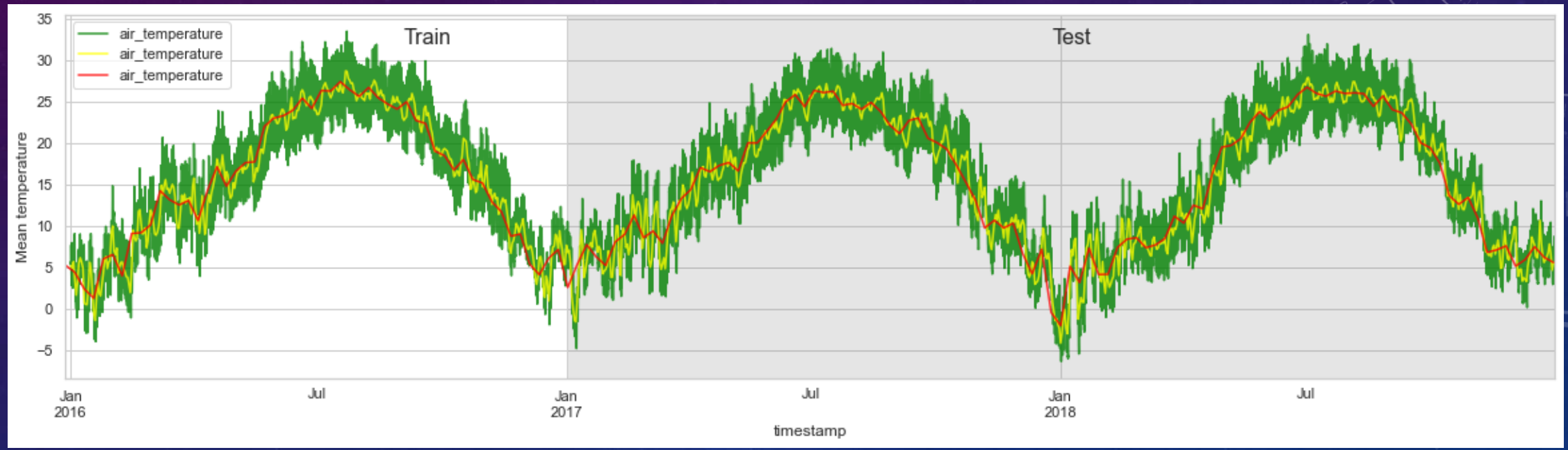
- Steam is the major usage compared to chilled water, electricity and hot water.

PRIMARY USE



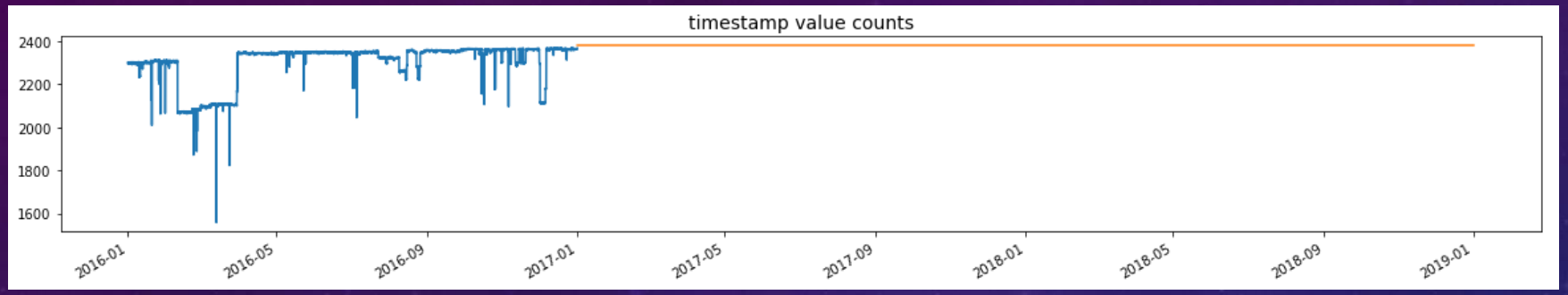
- Education is predominant in energy consumption. It is followed by office, entertainment/public assembly, lodging/residential, etc.

AIR TEMPERATURE

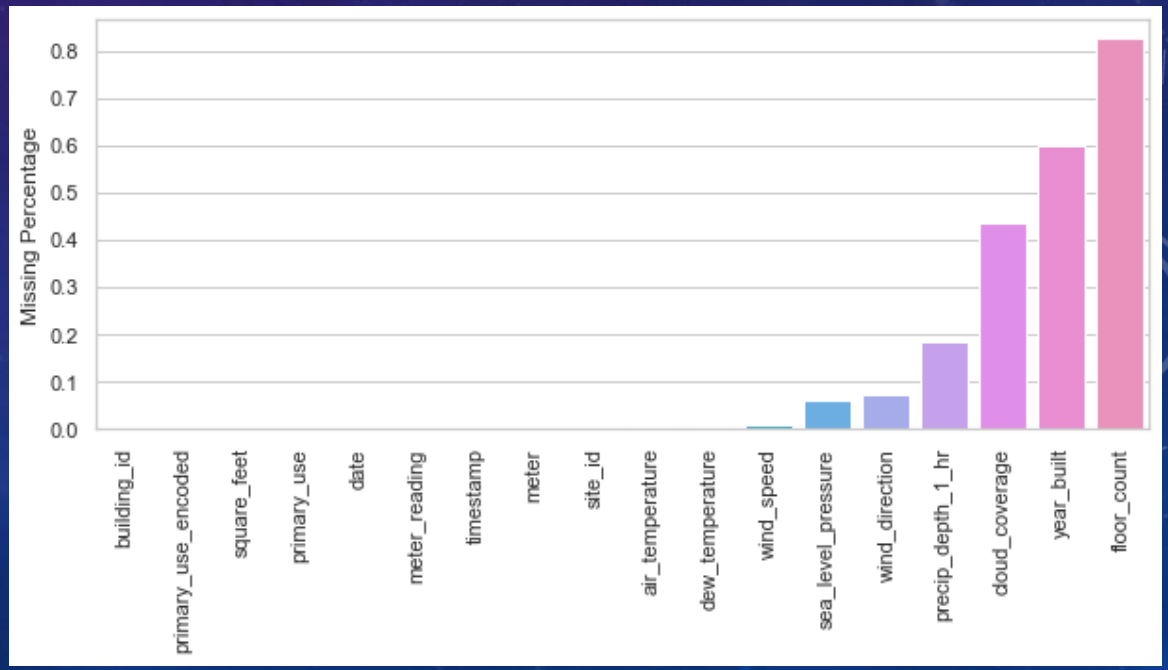


- Air temperature in the test set has very similar pattern as in the train set or following two years.
- The peak consumption of energy during the spring season is not corresponding well with the air temperature pattern.

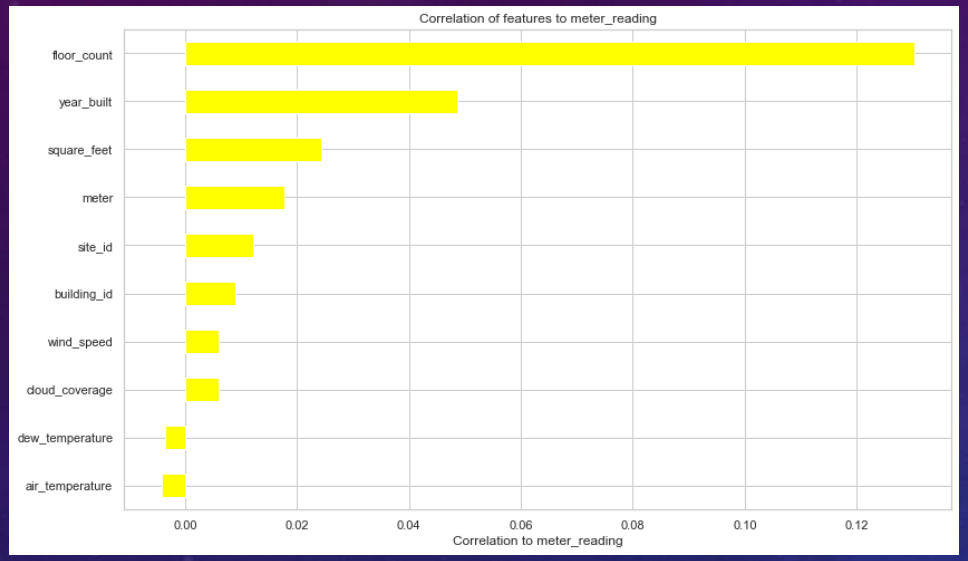
MISSING VALUES

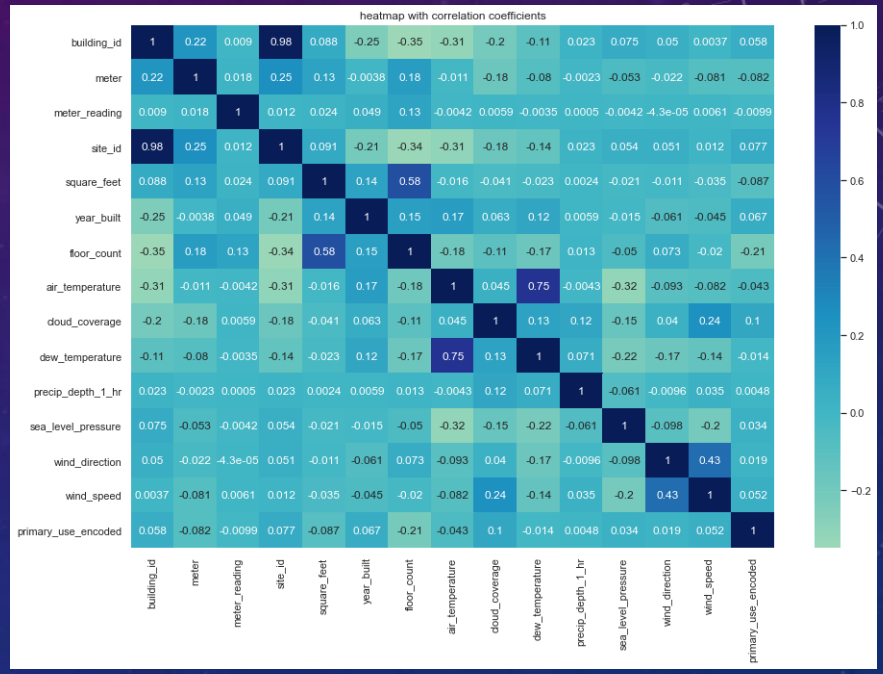


- Compared with test set, the training set has missing records in many hours and days.
- Top ranking in missing values: floor_count, year_built, cloud_coverage with more than 40% missing.



CORRELATION





- Floor count and square footage have strong correlation to meter reading makes sense because the more energy is demanded for larger service area. The lower the air temperature and dew temperature, the higher energy usage is reasonably revealed.
- site id is well correlated to building id. Floor count and square feet are also related to each other.
- The correlation coefficient between dew temperature and air temperature is 0.75, which agrees well with meteorology study.

PRE-PROCESSING

- **Imputation**
  - Missing air temperature is best interpolated linearly instead of using mean value.
  - The missing values in 'year_built', 'floor_count', 'cloud_coverage', 'dew_temperature', 'precip_depth_1_hr', 'wind_speed' were imputed using SimpleImputer with means.
- **Feature engineering**
  - Add hour, day of year, week of year, month as new features.
  - Wind speed is converted to Beaufort scale.
- **One-hot encoding**
  - LightGBM has a built-in mechanism to convert categorical variables to one-hot encoding. It also allows faster computing.

METRICS

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

A measure of how spread out the residuals are.

$$PMAE = \frac{\sum_{j=1}^{n}(|y_i - \hat{y}_i|)}{\sum_{j=1}^{n}|y_i|}$$

This metric gives a better idea to the stakeholders about how far the prediction is away from actual value percentage wise.

MODELS

- Decision tree based models
  - Non-linear nature of problem solving
  - Recommender, ranking, classification, linear regression, etc.
- Gradient boosting
  - Additive strategy starting from a rough decision (mean or median) and get more and more powerful in prediction by adding up all the predictions step by step.
  - Pre-sort to find all possible split points is time-consuming.
- **LightGBM**
  - Emphasizes on larger gradients
- **Catboost**
  - Similar to time series prediction.

CATBOOST

- Computation time is 1 hour 38 min to get the best score 1.140.

LIGHTGBM

LightGBM RMSE test scores, KFold = 5

int: ['hour', 'dayofyear', 'weekofyear', 'month'] cat: ['primary_use'] 'feature_fraction': 0.85	cat: ['hour', 'dayofyear', 'weekofyear', 'month'] cat: ['primary_use'] 'feature_fraction': 0.85	remove ['hour', 'dayofyear', 'weekofyear', 'month'] cat: ['primary_use'] 'feature_fraction': 0.85
0.952	1.424	1.235
0.963	1.209	1.195
0.941	1.277	1.069
1.050	1.225	1.204
1.191	1.278	1.336
average: 1.020	average: 1.283	average: 1.208

- Computation time is about 12 min, which is about 3 times shorter.
- Time features as integer significantly improve the test score from 1.283 to 1.020.

LIGHTGBM-SMOOTHING DATA

Normal KFold Splitting after Smoothing

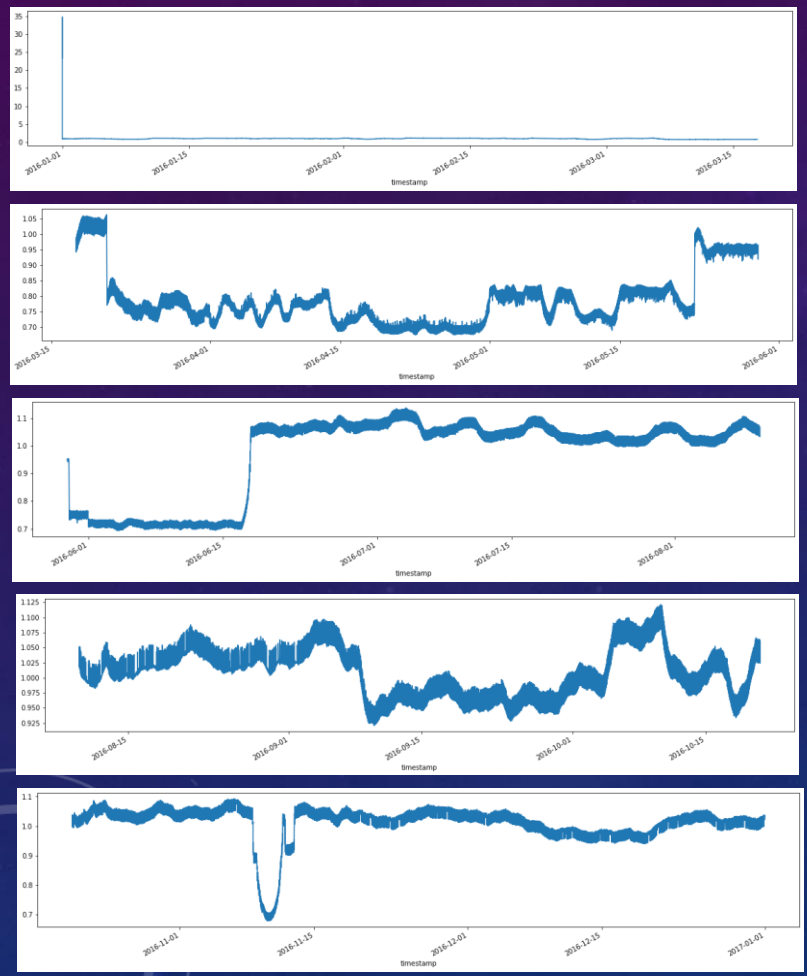
int: ['hour', 'dayofyear', 'weekofyear', 'month'] cat: ['primary_use'] 'feature_fraction': 0.85	
root mean squared error	percent mean absolute error
0.878	0.084
1.943	0.211
1.301	0.129
0.249	0.033
0.491	0.047
average: 0.973	average: 0.102

Time Series Splitting after Smoothing

int: ['hour', 'dayofyear', 'weekofyear', 'month'] cat: ['primary_use'] 'feature_fraction': 0.85	
root mean squared error	percent mean absolute error
1.687	0.191
1.270	0.140
0.114	0.015
0.271	0.039
0.522	0.044
average: 0.773	average: 0.086

- Sample smoothing can potentially suppress the high gradients in the local data, delivering additional benefits to the gradient boosting method.
- Average score decreases from 1.222 to 0.973, a significant improvement in performance.
- Time series splits are also applied to the smoothed dataset. There is a further increase in predication accuracy by the score average.

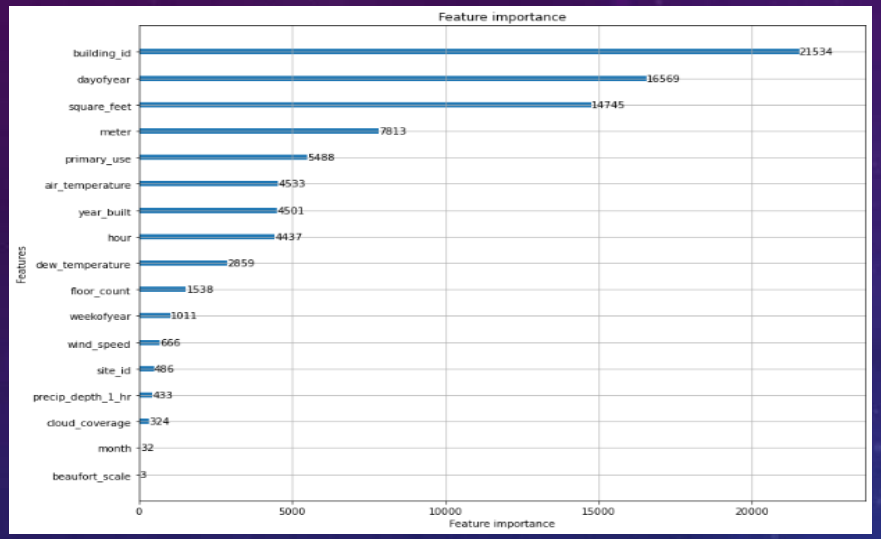
PREDICTED / ACTUAL RATIO OVER THE YEAR



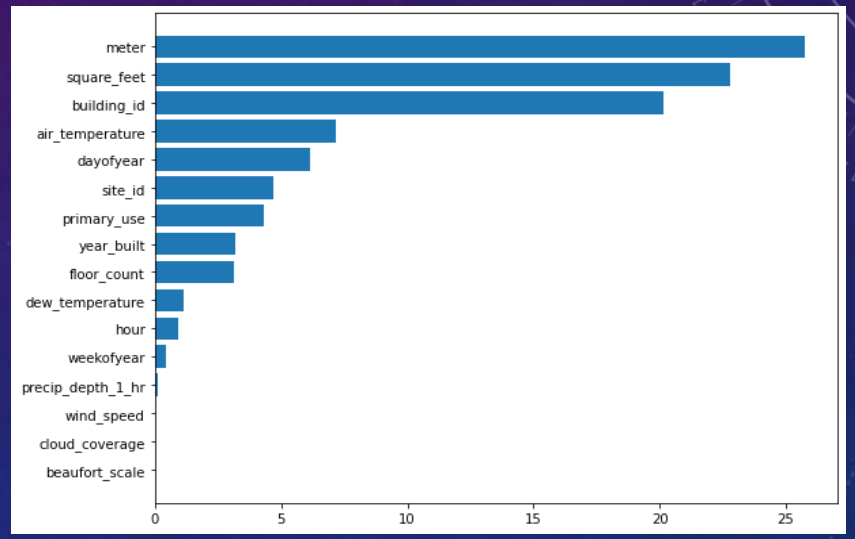
- The exceptionally high ratios close to 35 in the beginning of January in the KFold splitting is caused by data entry error. Removing those data doesn't help the prediction.
- March through July see much less amount of predicted energy usage that the actual, whereas, November through December see the predicted is very close to the actual.
- On March 7th and June 17th, there is a sharp transition in prediction that is very well corresponding with the abrupt transition from low energy usage season to high energy usage season or vice versa.
- The abrupt deep drop in November corresponds with the unusually high usage observed in EDA section earlier. Building 1099 is responsible for it.
- The models fail to catch the transition and result in significant prediction error. It would be a good idea to separate the dataset from the transition and train the models separately.

FEATURE IMPORTANCE

LightGBM



CatBoost



- Top features like building id, day of year, square feet, meter, air temperature, regardless their relative ranking, are captured by both algorithms.
- These are indeed the predominant factors influencing the usage from domain knowledge point of view.
- The model successfully pick out air temperature as an important feature. Therefore, the model can be appropriately used to predict energy usage for the following years based on weather change. This would allow the effect of newer technology investment or change on energy saving to be separated from that of the weather change.
- Weather data like wind speed / beaufort scale, precipitation and cloud coverage remains as the least important features.

CONCLUDING REMARKS

- A rich energy usage dataset coming from over 1,000 buildings around the world combined with local weather data was explored and modeled.
- Exploratory data analysis findings:
  - Distribution of the data is highly skewed with respect to meter readings, energy type, square footage, primary use, etc.
  - Energy usage is exceptionally high from March to June and a few days in November.
  - Significantly higher energy usage on site #7 compared with other sites.
  - Building #1099 on an educational site was found to consume at least 70 times more energy than any other buildings.
  - Steam is the predominant among 4 energy types, over almost all year.

CONCLUDING REMARKS

- Exploratory data analysis findings:
  - Education is predominant in energy consumption. It is followed by office, entertainment / public assembly, lodging / residential, etc.
  - Correlation heat map reveals the lower the air temperature and dew temperature, the higher energy usage.
  - The correlation coefficient between dew temperature and air temperature is 0.75, which agrees well with meteorology study.
- Data Processing
  - Missing air temperature is interpolated linearly.
  - Other missing values in 'year_built', 'floor_count', 'cloud_coverage', 'dew_temperature', 'precip_depth_1_hr', 'wind_speed' were imputed using SimpleImputer with means.
  - New features like hour, day of year, week of year, month are added.

CONCLUDING REMARKS

- Machine Learning
  - Catboost and LightGBM learning algorithms based on gradient boost are applied to the energy dataset.
  - Catboost is about 3 times slower than LightGBM to achieve similar score.
  - Treating time features as integer significantly improve the test score (RMSE) from 1.283 to 1.020. The additional time features seem to have minimal effect on test score.
  - Data smoothing by rolling average (set day as window size) significantly improves the score from 1.222 to 0.973.
  - Time series splitting is found to improve the prediction, as compared with normal KFold splitting.
  - The ratio of predicted value against the actual is used to investigate how well the trained model works at the local level.

CONCLUDING REMARKS

- Machine Learning
  - Possible data entry errors in the first 1000 samples in the dataset, or the very beginning of the year. March through July see much less amount of predicted energy usage that the actual, whereas, November through December see the predicted is very close to the actual.
  - The models fail to catch the transition and result in significant prediction error. It would be a good idea to separate the dataset from the transition and train the models separately.
  - Top important features include building id, day of year, square feet, meter, air temperature, regardless their relative ranking, are captured by both models. These are indeed the predominant factors influencing the usage from domain knowledge point of view. Weather data like wind speed / beaufort scale, precipitation and cloud coverage remains are the least important features.

CONCLUDING REMARKS

- Machine Learning
  - The model successfully pick out air temperature as an important feature. It can be appropriately used to predict energy usage for the following years based on weather data. This would allow the effect of newer technology investment or change on energy saving to be separated from that of the weather change.
  - Lastly, other factors, for example, budget/funding/project situation, decision making by the administrators when to turn on AC, etc., also have great impact on energy usage and its timing. It is necessary to have the human factors in mind when building a realistic model.

FUTURE WORK

- Check with domain experts possible data entry errors in the January meter readings because this the only occurrence with off-the-chart prediction.
- Separate the dataset from the transition where the high energy usage season change to lower or vice versa and train the models separately.
- Separate samples according to meter or energy type for machine learning individually
- Check with domain expert on how budget/funding/project situation, decision making by the administrators when or what time to turn on AC, etc., impact on energy usage and its timing.
- Predict energy usage building wise. This way the local weather data can play a more important role.
- Continue feature engineering by adding new variables such as holidays because it affects energy usage for obvious reasons.