# Feature Engineering and Modeling of Metered Building Energy Usage

Rounok Joardar

## 1. Introduction

Accurate modeling and prediction of energy usage in commercial buildings offers significant cost saving opportunities and have become a major component in optimizing financial performance of many businesses. A key factor in the financing of office buildings is the difference between their actual energy consumption and what they would have used without energy efficiency enhancements. The latter values are typically derived from models. Unfortunately, current modeling techniques do not scale well. Thus, there is a need for developing more accurate and robust models for energy usage in non-residential buildings.

In this project, a model is developed using machine learning techniques for predicting readings of various types of energy meters in a building given the building's characteristics and local weather information. A massive dataset, provided by the American Society of Heating and Air-Conditioning Engineers (ASHRAE), "a global society advancing human well-being through sustainable technology for the built environment", consisting of over 40 million meter readings is used to build the model.

The Python programming language is used. Extensive use is made of pandas as it allows very large dataframes to be manipulated easily. Various regression tools based on ensemble models as well as several numerical packages such as numpy and scipy are also readily available in Python.

## 2. Data Description

The main dataset used in this project includes two years' worth of hourly meter readings from over one thousand buildings at various sites across the US and Europe. Readings from up to eight different meter types such as electricity, gas, etc. are recorded in this dataset. Information is categorized by building identifiers and meter type. It is noted that not every building has all eight types of meters and not every meter has twenty four readings per day. It is noted that all meter readings are in units of kilowatt-hours.

Two supplementary datasets are also provided. The first relates to building characteristics such as square footage, number of floors, primary and secondary usage, etc. Information is categorized by building identifiers and site (i.e. location) identifiers. The second relates to local weather information

from the various sites where the buildings were located. Hourly readings of air temperature, atmospheric pressure, cloud coverage, wind speed etc. are recorded in this dataset. Fig. 1 is a pictorial summary of the three datasets.

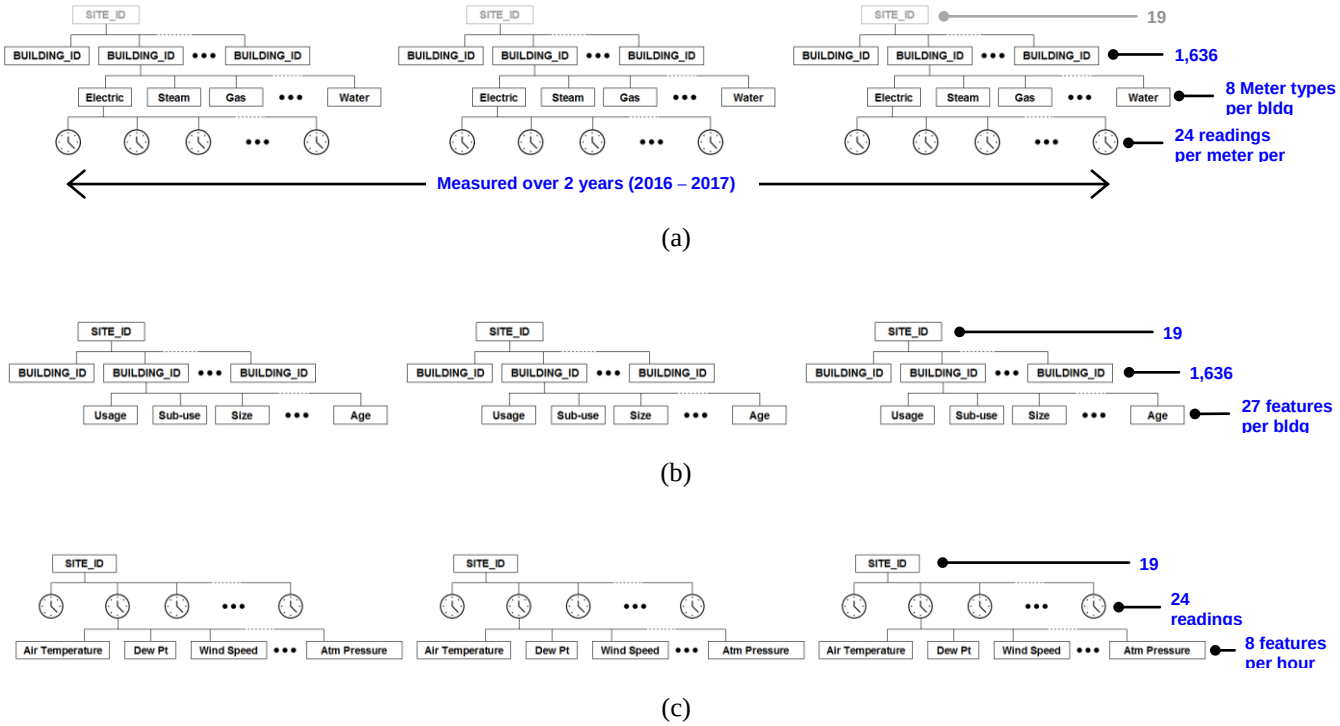
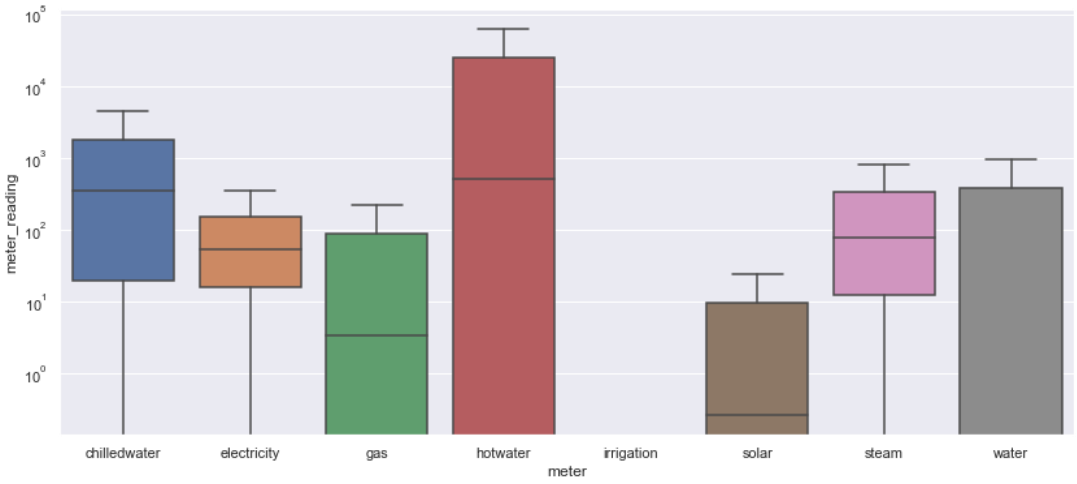


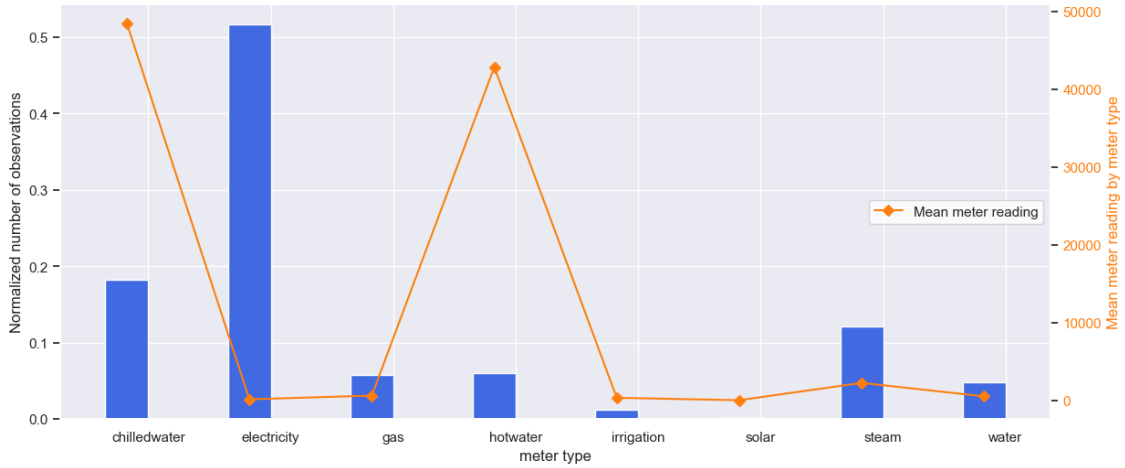

Fig. 1. Schematic view of various datasets available for modeling. (a) The main dataset of energy readings, (b) first supplementary dataset of building metadata, and (c) second supplementary dataset of local weather information.

## 3. Exploratory Data Analysis (EDA)

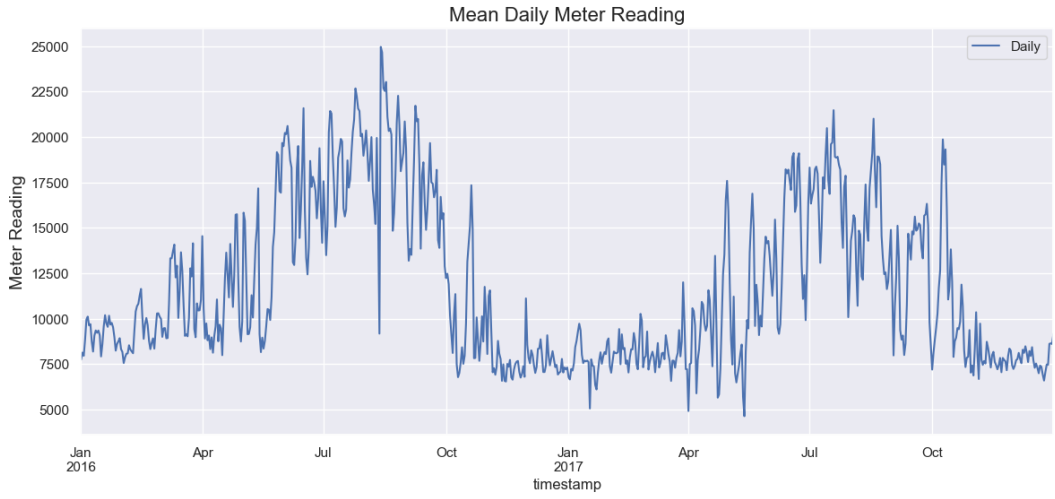
The main energy readings dataset only contains information about building identification, meter type, and energy usage. Some simple statistics such as energy usage by meter type and distribution of meter types can be derived from this data. Fig. 2(a) shows a boxplot of the daily mean values of each type of energy meter. The top two modes of energy usage, by far, are hotwater and chilledwater. Fig. 2(b) shows a distribution of the number of meters of each type. Clearly, the most prevalent meter type is electricity. Fig. 2(c) shows a time domain plot of the daily mean values of all meter readings combined. A repeating seasonal pattern is observed. Based on Fig. 2(c) it may be tempting to consider simply building a time-series model for energy usage. The problem with this idea is that the daily energy usages of different meter types do not all follow the pattern of Fig. 2(c). Per Fig. 2(a), total energy usage is dominated by two meter types, and not all buildings or all sites may have these meter types. Therefore, such a model would not work well for predicting energy usage of buildings without hotwater or chilledwater meters.




(a)




(b)




(c)

Fig. 2. (a) Boxplot of mean daily meter readings on log scale, (b) histogram showing normalized number of meters of each type, and (c) time-domain plot of daily mean of all meter readings.

In order to build an accurate and scalable model, building metadata and site weather information data are merged with the main energy readings dataset. Building metadata contains information about each building shown in Table 1. Weather information is summarized in Table 2.

Once the datasets are merged, correlations between the various numerical features are examined. Fig. 3 shows pair-wise correlation plots between a few selected features. Not all available numerical features are included in this plot for the sake of clarity. Several insights are available from this plot. First, there is a strong correlation between energy usage and air temperature. Air temperature and dew point are very tightly correlated. Therefore it makes sense to include only one of these features in the model. There appears to be little or no correlation between energy usage and square footage of the building. Based on this, building size is not used in modeling. Next, there appears to be a negative correlation between air temperature and atmospheric pressure. This would imply a negative correlation between energy usage and atmospheric pressure, but that is difficult to discern from the correlation plot. For an alternate view, time domain plots of energy usage, air temperature, and atmospheric pressure are shown in Fig. 4. An inverse relation between energy usage and atmospheric pressure is visible. Based on this, atmospheric pressure is retained as a feature for model development.

Table 1. List of building metadata.

| | |
|---|---|
| building_id | building code-name |
| site_id | animal-code-name for the site |
| primaryspaceusage | Primary space usage of all buildings is mapped using the energystar scheme building description types |
| sub_primaryspaceusage | energystar scheme building description types subcategory |
| sqft | Floor area of building in square feet |
| lat | Latitude of building location to city level |
| lng | Longitude of building location to city level |
| timezone | site's timezone |
| industry | Industry type corresponding to building |
| subindustry | More detailed breakdown of Industry type corresponding to building |
| heatingtype | Type of heating in corresponding building |
| yearbuilt | Year corresponding to when building was first constructed |
| date_opened | Date building was opened for use |
| numberoffloors | Number of floors corresponding to building |
| occupants | Usual number of occupants in the building |

Table 2. List of weather data items.

| | |
|---|---|
| timestamp | Date and time |
| site_id | Site identifier |
| airTemperature | The temperature of the air in degrees Celsius |
| cloudCoverage | Portion of the sky covered in clouds, in oktas |
| dewTemperature | The dew point in degrees Celsius (ºC) |
| precipDepth1HR | The depth of liquid precipitation that is measured over a one hour accumulation period (mm) |
| precipDepth6HR | The depth of liquid precipitation that is measured over a six hour accumulation period (mm) |
| seaLvlPressure | The air pressure relative to Mean Sea Level (MSL) (mbar or hPa) |
| windDirection | The angle, measured in a clockwise direction, between true north and the wind direction |
| windSpeed | The rate of horizontal travel of air past a fixed point (m/s) |

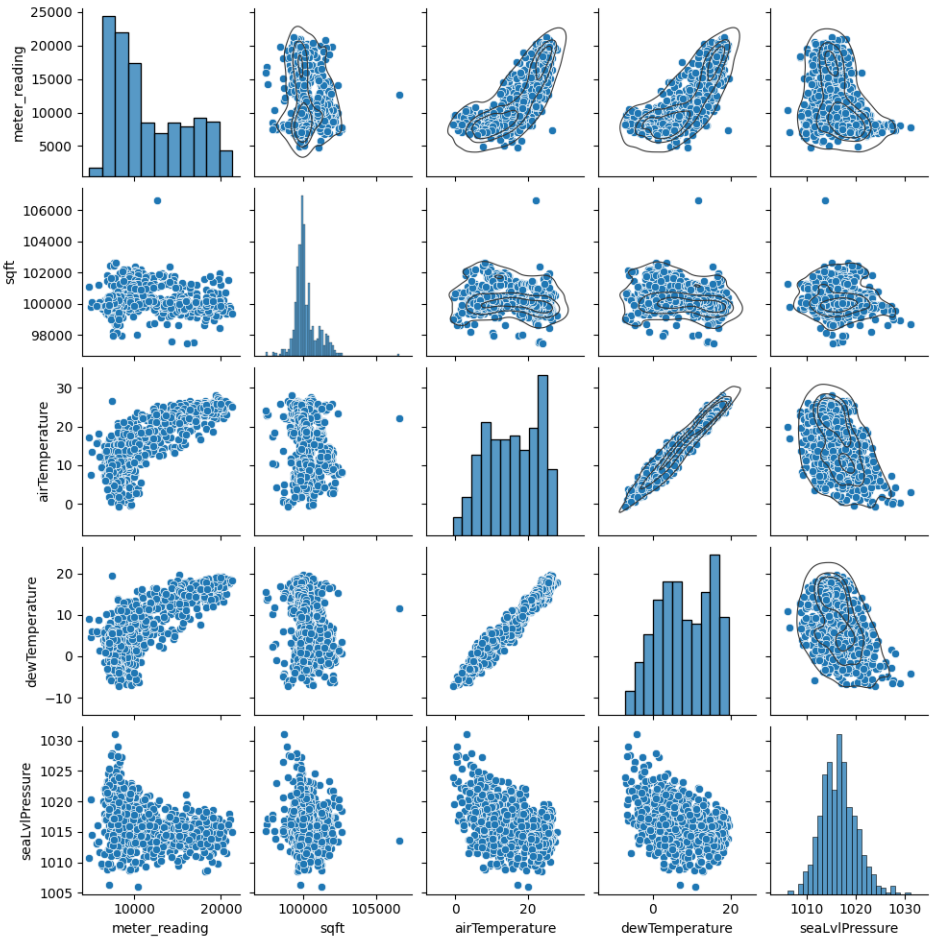



Fig. 3. Pair-wise correlation plots of several available features.

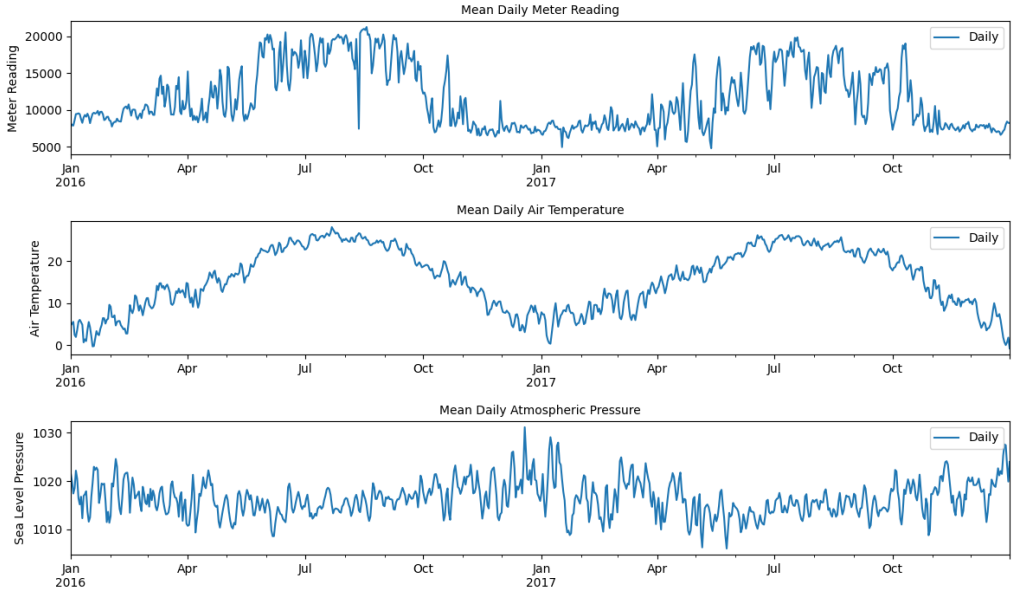



Fig. 4. Time-domain plots of energy usage (top), air temperature (middle) and atmospheric pressure (bottom). An inverse correlation between energy usage and atmospheric pressure is visible.

## 4. Feature Engineering

Although technically some amount of feature engineering has already occurred with the addition of weather and building data to the main dataset, several less obvious features are introduced in this section.

*4.1 Temperature Coefficient*

Fig. 5 shows overlay plots of meter readings and air temperature, split out by each of the eight meter types. It is observed that chilledwater energy usage is in phase with air temperature. On the other hand, hotwater, steam, and gas usage are out of phase with air temperature. These trends make intuitive sense since chilledwater is likely to be more essential in summer when air temperatures are higher, and vice versa for hotwater, gas, and steam. Solar, water, and irrigation, show no discernable correlation. Electricity usage trend versus air temperature is difficult to determine visually.

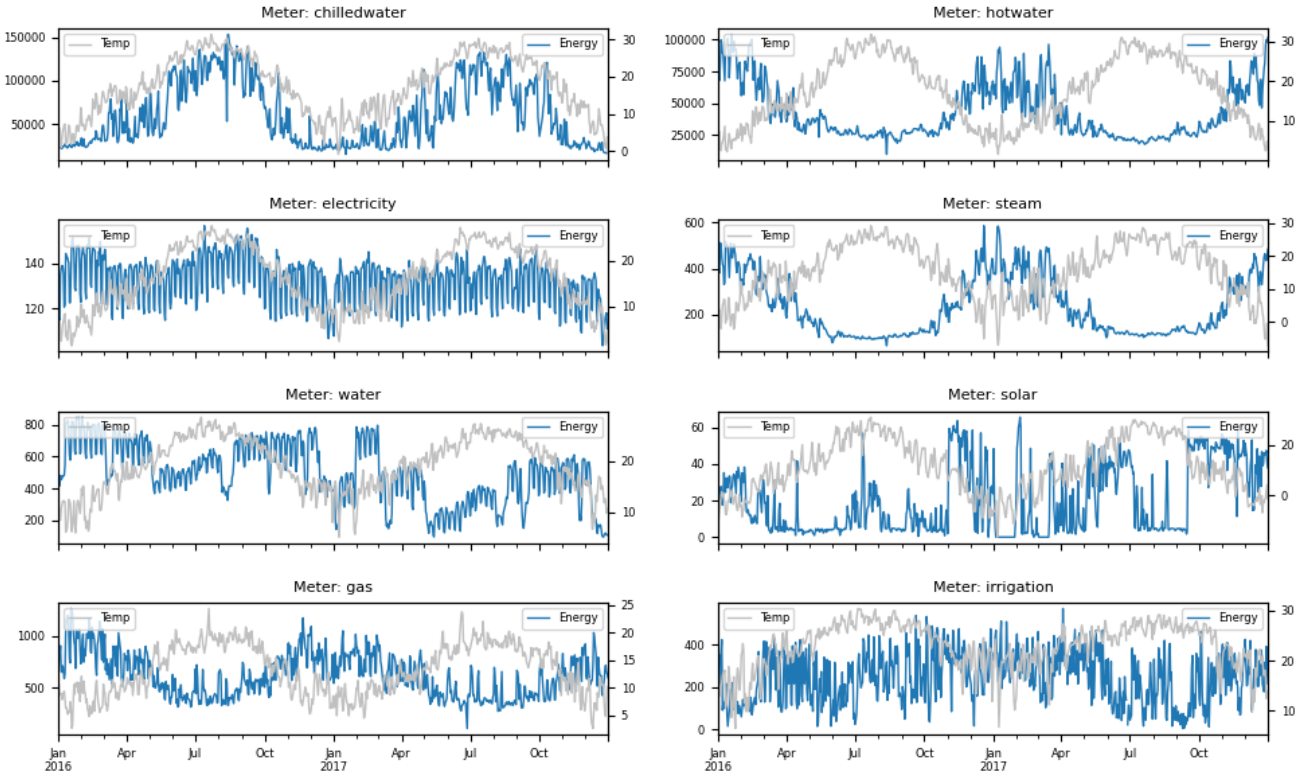



Fig. 5. Time-domain overlay plots of energy usage and air temperature, separated out by meter type.

Fig. 6 shows the same data presented as correlation plots, i.e. energy usage versus air temperature, split out by meter type. The trends and correlations described earlier continue to hold up. Mathematically, electricity usage has a positive correlation with air temperature.

Based on these observations, a new feature, “temperature coefficient”, is introduced. Rather than attempting to estimate a numerical partial derivative of energy use of each meter type with respect to air temperature, the feature is classified into three levels: +1 for chilledwater and electricity (positive coefficient), -1 for hotwater. steam, and gas (negative coefficient), and 0 for the others (indeterminate).

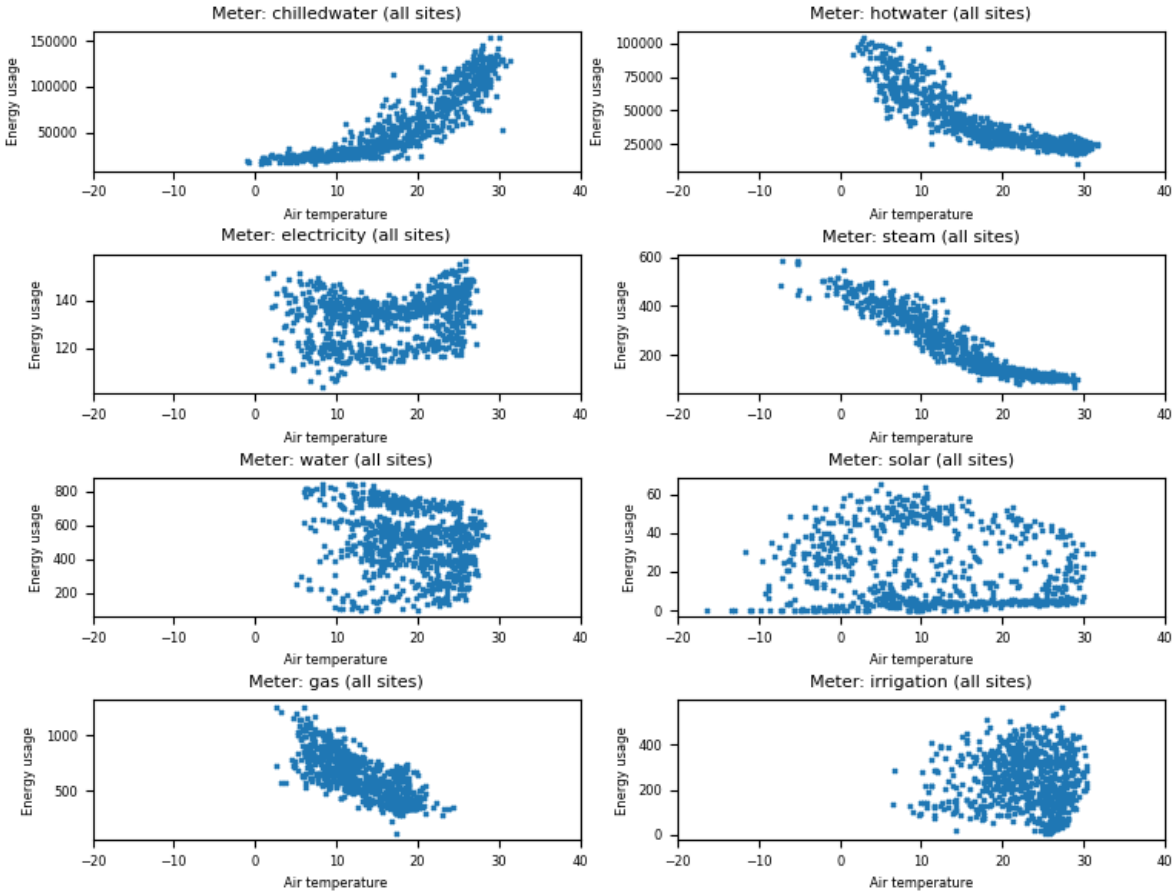



Fig. 6. Correlation plots of energy usage versus air temperature separated out by meter type.

*4.2 Pressure Coefficient*

Fig. 7 shows overlay plots of meter readings and atmospheric pressure, split out by each of the eight meter types. In this case, it is observed that chilledwater energy usage is out of phase with air temperature, while hotwater, steam, and gas are in phase. This is not totally surprising, considering the previously discussed inverse relation between air temperature and atmospheric pressure.

As before, a new feature, “pressure coefficient”, is introduced and classified into three levels: -1 for chilledwater (negative correlation), +1 for hotwater, steam, and gas (positive correlation), and 0 for the others (indeterminate).

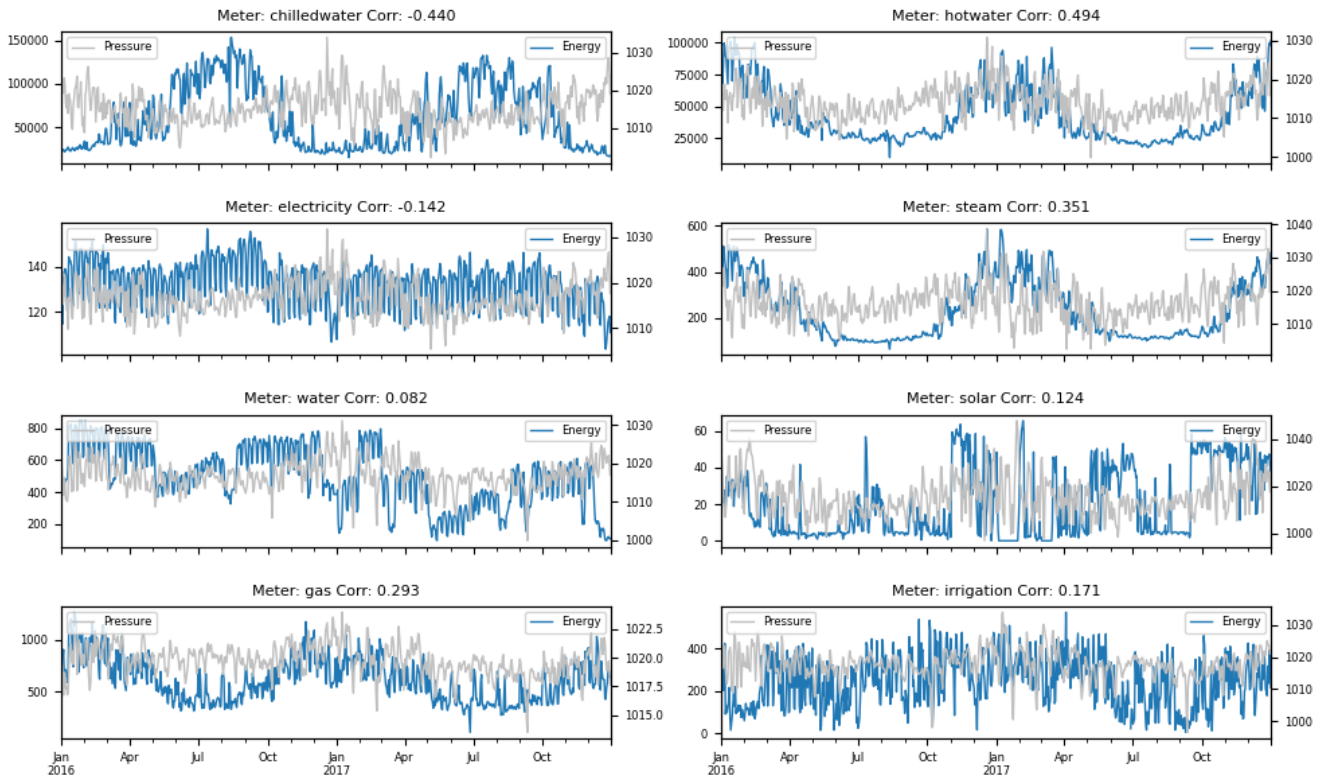



Fig. 7. Time-domain overlay plots of energy usage and atmospheric pressure, separated out by meter type. The exact correlation coefficients are also shown above each plot.

### *4.3 Weekly Periodicity*

Fig. 8 shows time-domain plots of mean daily electricity and water meter readings. There is a clear weekly periodicity, with energy usage dropping on weekends. In view of this, a new weekday/weekend binary feature is introduced, 1 for weekdays, 0 for weekends.

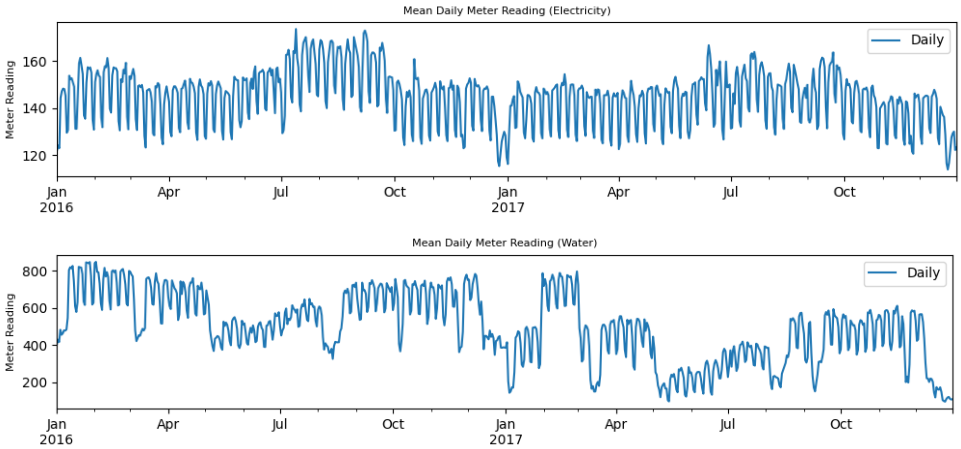



Fig. 8. Time-domain plots of energy usage readings from electricity and water meters showing weekly periodicity.

*4.4 Sub-Primary Usage*

In-depth tracking of the available data reveals that a much clearer classification of building energy consumption by use mode is possible from the sub-primary use data than from the primary use data alone. However, sub-primary usage has 92 unique values and not all of these are relevant for modeling. Fig. 9 shows histograms of energy usage by the top building use modes, for four meter types. It is observed that there is significant commonality, i.e. most of the energy usage is associated with a few types of building use modes. Based on this analysis, the top 20 common use modes are selected and the rest are clubbed into a “other” type.

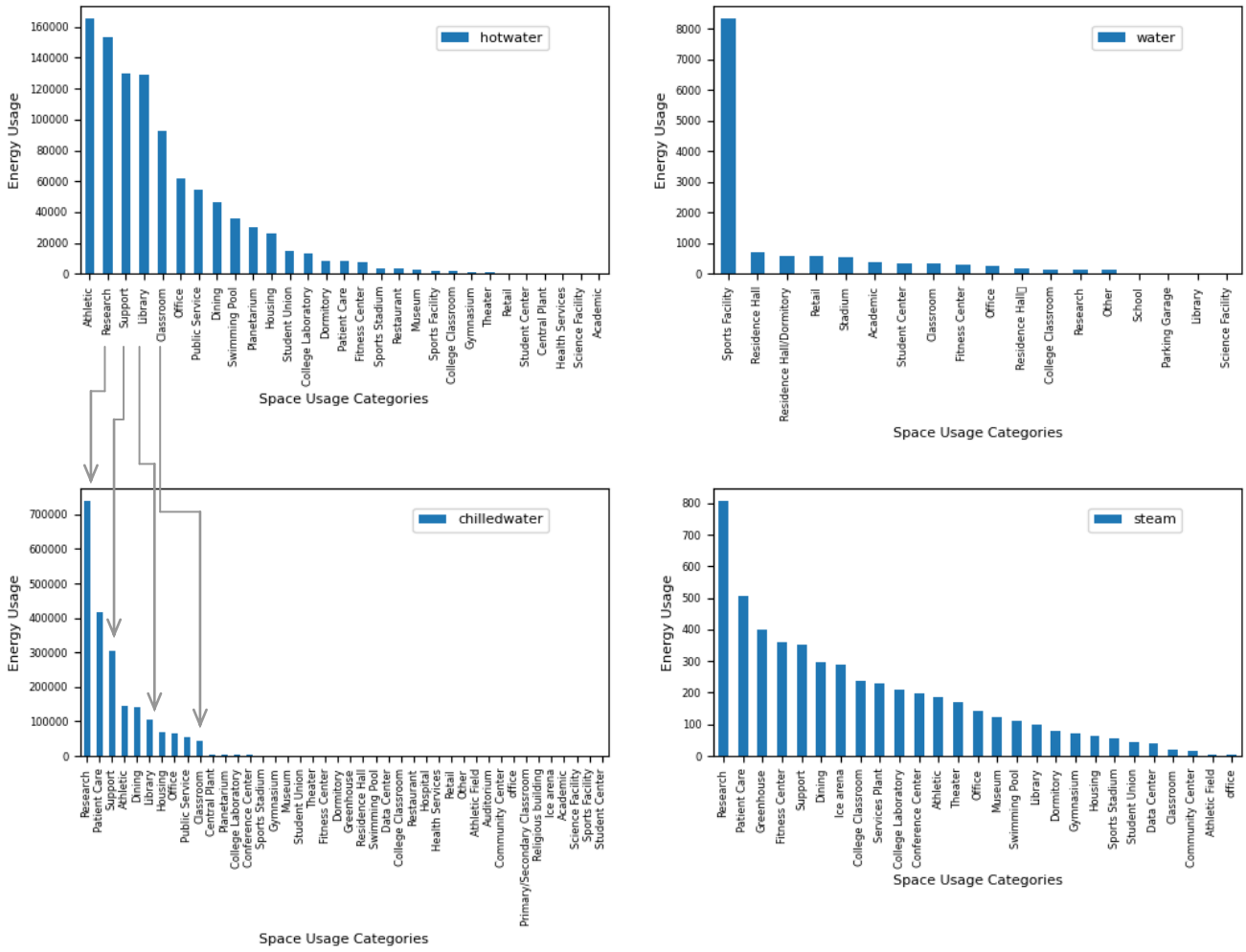



Fig. 9. Histogram plot of energy consumption by building use mode.

*4.5 Aggregation Features*

In addition to the new features described above, four additional aggregation features were added. These are the minimum, maximum, mean, and standard deviation values of air temperature readings, grouped by site. It is noted that these aggregated values need to be merged with the original weather dataset. Code used to accomplish this is shown below. A pictorial summary of feature engineering done in this project is shown in Fig. 10.

```
# Add new features: mean, min. max, and var of air temperature
weather_df['dates'] = weather_df['timestamp'].dt.strftime('%d %b %Y')
aggregations = {
    'airTemperature': ['mean', 'min', 'max', 'var']
}
weather_agg = weather_df.groupby(['dates', 'site_id']).agg(aggregations)
weather_agg.columns = pd.Index([e[0] + "_" + e[1].upper() for e in
weather_agg.columns.tolist()])
weather_df = pd.merge(weather_df, weather_agg, how='left', on=['dates', 'site_id'])
weather_df['airTemperature_VAR'].fillna(0, inplace=True)
weather_df.drop(['dates'], axis=1, inplace=True)
del(weather_agg)
```

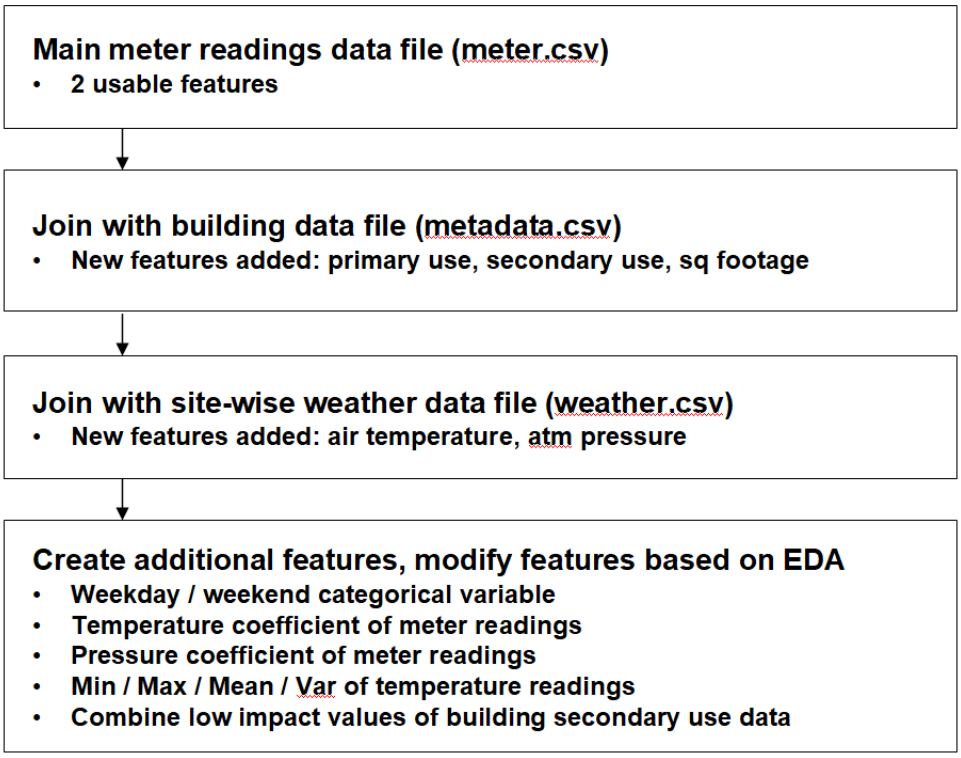



Fig. 10. Summary flow showing all features engineered for this model.

## 5. Modeling

Of the various popular machine learning models available in Python, the Light Gradient Boosting Method[1] (LGBM) package from Microsoft is chosen for this project. LGBM is a gradient boosting framework based on decision tree algorithms. It uses a histogram approach that buckets continuous features into discrete bins. This method offers a very good tradeoff between speed and accuracy, a particularly important consideration given the very large dataset in use.

As discussed earlier, there are four non-numerical categorical features in this energy use prediction model: meter type, building primary usage, building sub-primary usage, and weekday/weekend. Categorical features converted to numerical values by label encoding using sklearn's "labelencoder" function. This method of encoding assigns an integer value to each unique member of a given category. Compared to one-hot encoding, where each unique member gets a binary code, leading to a proliferation in the number of columns in the feature dataset, label encoding is more compact. One concern with label encoding is implied ordinality. However, in LGBM these label encoded categorical features are identified as such, and the algorithm does not assign any ordinality to the codes. In fact, the LGBM user guide recommends label encoding for categorical features over one-hot encoding.

For completeness it is noted that there are six continuous numerical features: square footage, air temperature and its four aggregated measures (minimum, maximum, mean, and standard deviation), and atmospheric pressure. Temperature and pressure coefficients are treated as categorical features although they are assigned numerical values.

For purposes of model training and testing two non-overlapping datasets are needed. In this case, this is accomplished by using the 2016 portion of the dataset for training and the 2017 portion for testing. The split is roughly half-half.

Before a model can be trained, its internal parameters, also known as hyper-parameters, have to be set to their optimum values. This is a time intensive task, particularly if an exhaustive grid search approach is utilized. In this work, in the interests of time, a readymade Python package named hyperopt[2] is used. A small (15%) randomly selected portion of the 2016 training dataset is used for hyper-parameter optimization (takes ~ 2 hours on a 3.6 GHz CPU). The final values are shown in Table 3.

[1] https://lightgbm.readthedocs.io/en/latest/

[2] Bergstra, J., Yamins, D., Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proc. of the 30th International Conference on Machine Learning (ICML 2013)

Table 3. Summary of LGBM hyper-parameters used to build this model.

| | | | | | |
|---|---|---|---|---|---|
| bagging_fraction | 0.83 | learning_rate | 0.175106981 | metric | RMSE |
| boosting | gbdt | max_depth | 23 | min_gain_to_split | 4.06 |
| feature_fraction | 0.93 | max_bin | 255 | num_leaves | 1895 |
| lambda_l1 | 4.100479495861697 | min_data_in_leaf | 200 | objective | tweedie |
| lambda_l2 | 4.727106993281982 | min_data_in_bin | 3 | subsample | 0.662448122 |

## 6. Results and Model Validation

Once the train and test data are separated and optimum LGMB hyper-parameters determined, the required model is built. A limit of 100,000 boosting rounds is used and iterations are stopped if there is no improvement in fitting criterion after 1000 rounds. Model training on the 2016 data, comprising around 20 million rows, takes about 6 minutes on a 3.6 GHz CPU. After training is completed the model is used to predict 2017 energy usage for each building. In order to illustrate advantages obtained from the engineered features, two models are built: one with a minimal set of features and another with the full set of features,

Fig. 11 shows actual 2017 energy usage readings by meter type overlaid on predicted values. The predicted results are from a model built with minimal features, specifically air temperature, air pressure, square footage, and primary use. The RMSE error (in percentage terms) for each meter type is also shown. It is clear that the model is able to capture hotwater and chilledwater readings with reasonable accuracy, but fails to do so for other meters.

In Fig. 12 the actual 2017 energy usage readings are overlaid on predictions obtained from a model built with the full feature set. The improvements are obvious. Not only are the RMS errors reduced for every meter type, but characteristics such as weekly periodicity are very well captured. Particularly pleasing is the accuracy with which hotwater and steam meter readings are captured. On the other hand, solar readings are still not well predicted. This is likely to be a result of not including cloud coverage as a feature in modeling. This will be addressed in a future revision. Fig. 13 shows x-y correlation between actual and predicted energy usage including $r^2$ value from a linear regression. Table 4 is a summary comparing RMS errors obtained using a minimal feature model and a full feature model.

Once trained, the LGBM model implementation provides importance scores for each of the features used. Fig. 14 is a bar graph showing feature importances for this model. It is noted that the non-obvious engineered features are among the top most important ones.

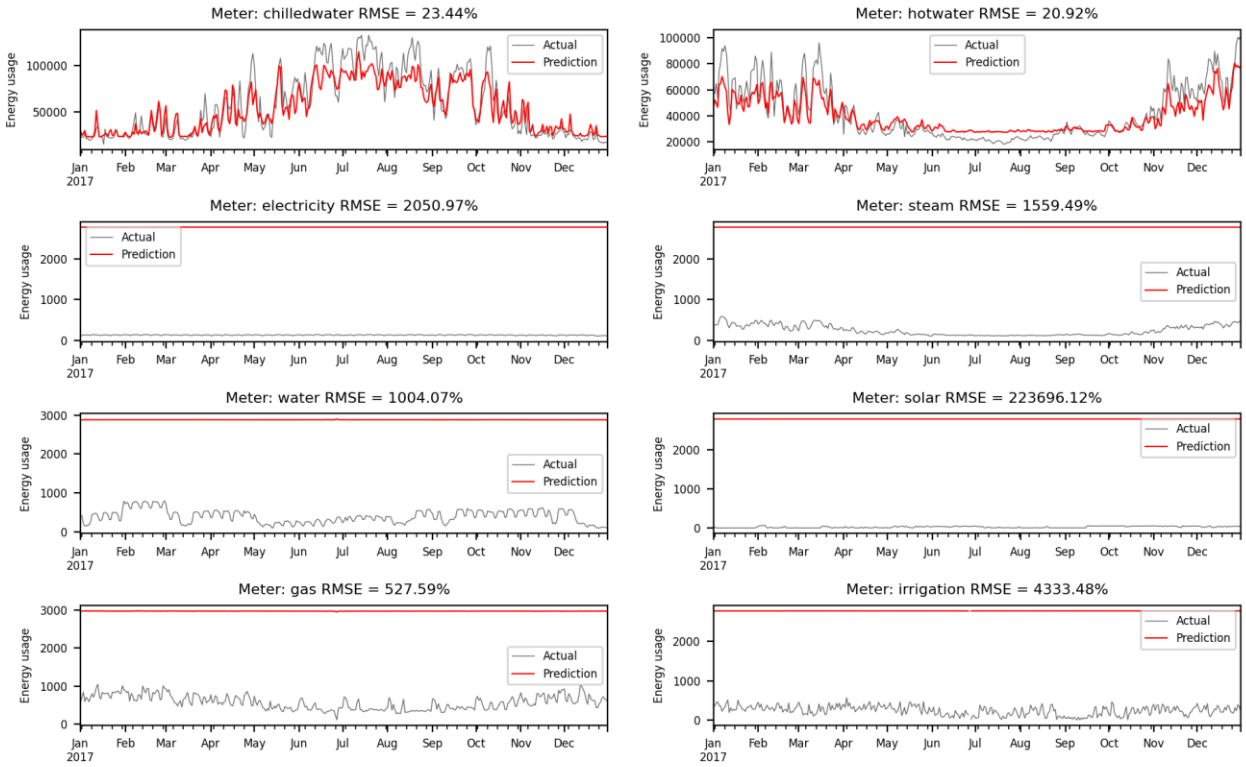



Fig. 11. Time-domain plot of actual energy usage by meter, overlaid with predictions from minimal feature model.

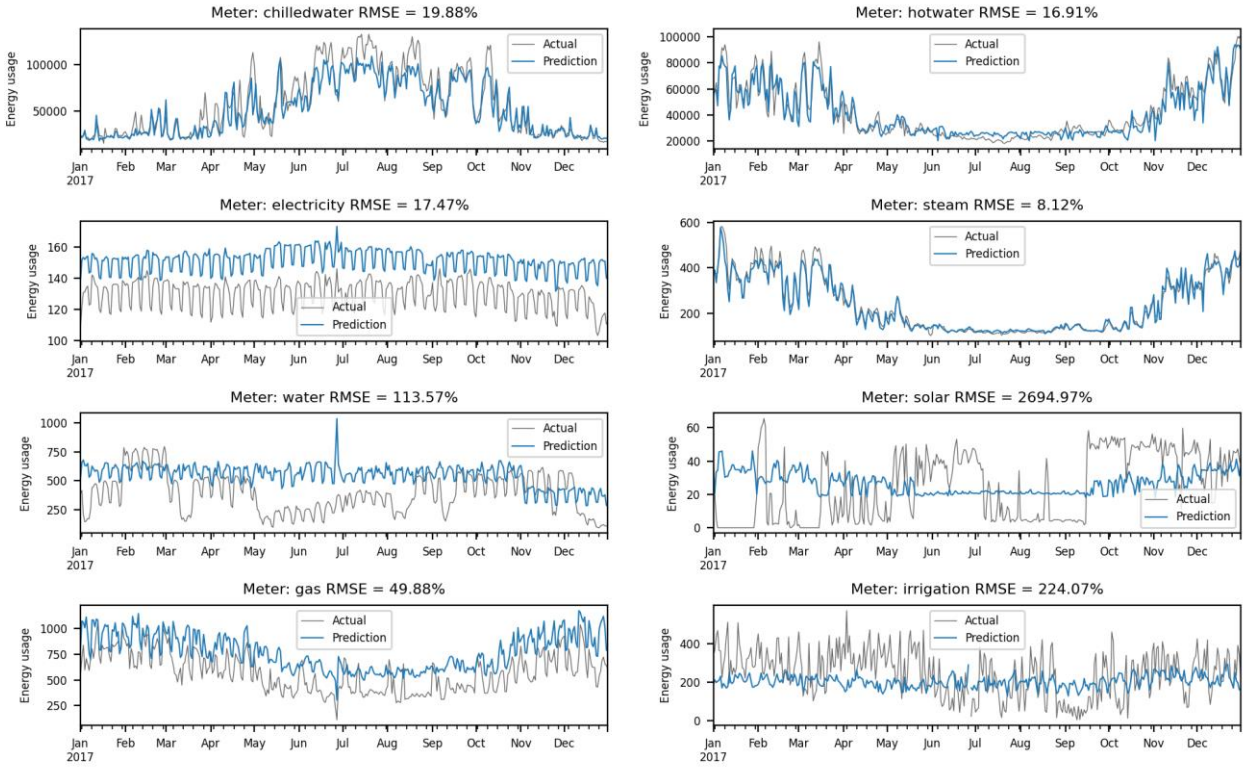



Fig. 12. Time-domain plot of actual energy usage by meter, overlaid with predictions from full feature model.

Table 4. Summary of % RMS errors by meter type with minimal and full feature engineering.

| Meter Type | RMSE (%) Before Feature Engineering | RMSE (%) After Feature Engineering | Comment |
|---|---|---|---|
| Electricity | 2050.97 | 17.47 | Significant improvement |
| Steam | 1559.49 | 8.12 | Significant improvement |
| Hot Water | 20.92 | 16.91 | Some improvement |
| Chilled Water | 23.44 | 19.88 | Some improvement |
| Gas | 527.59 | 49.88 | Significant improvement |
| Water | 1004.07 | 113.57 | Significant improvement |
| Irrigation | 4333.48 | 224.07 | Significant improvement |
| Solar | 223696.12 | 2694.97 | Not reliable |

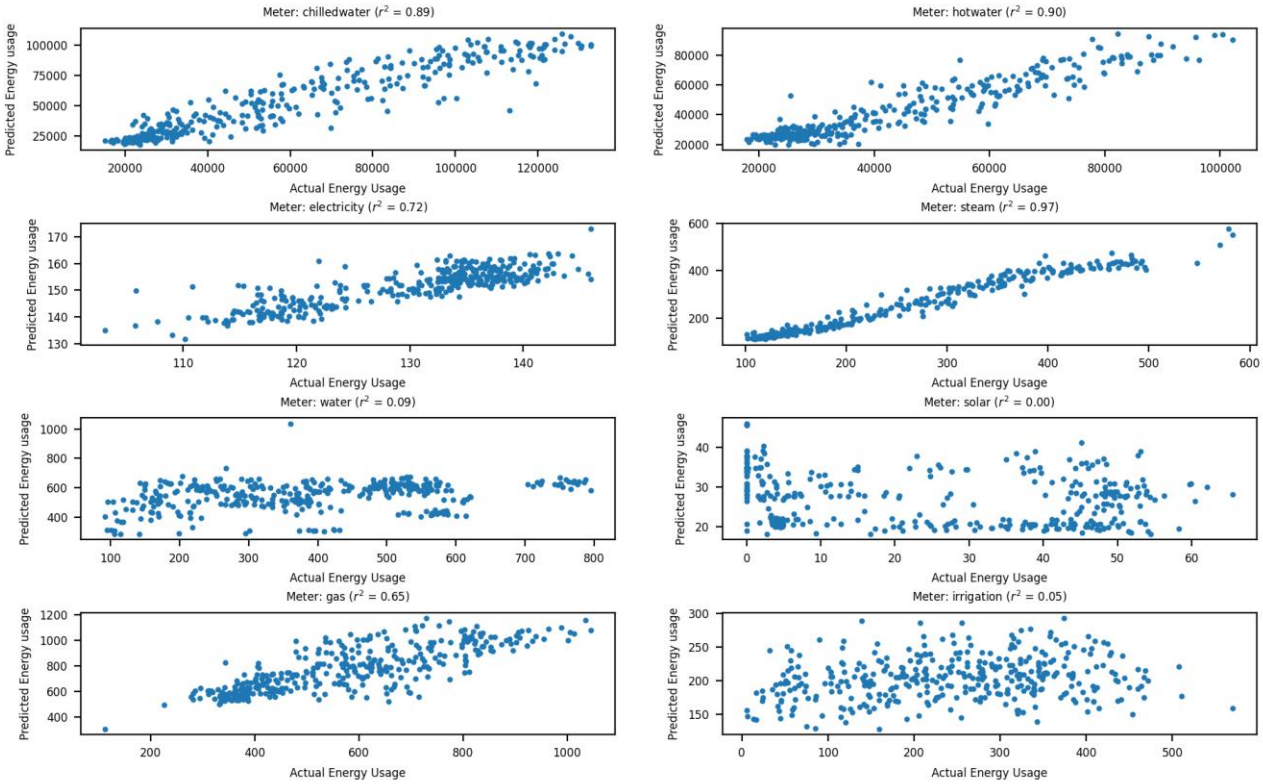



Fig. 13. Correlations between actual and predicted energy usage for each meter type.

As a final point, it is noted from Fig. 14 that square footage is identified as the most important feature. Intuitively this makes sense; the larger the building, the more energy it is likely to use. However, it is interesting that this relationship is not evident in Fig. 3. To investigate this further, correlation between energy usage and square footage is examined by meter type. Results are shown in Fig. 15. It is observed that while hotwater usage has a high positive correlation with square footage, most of the other meter readings are either weakly correlated or not correlated. Considering that hotwater constitutes one of the largest fractions of total energy consumption, it makes sense that square footage comes out as the most important feature. Lastly, since Fig. 3 is essentially a combination of all

of the meters shown in Fig. 14, it also makes sense that no particular correlation is evident from that view.

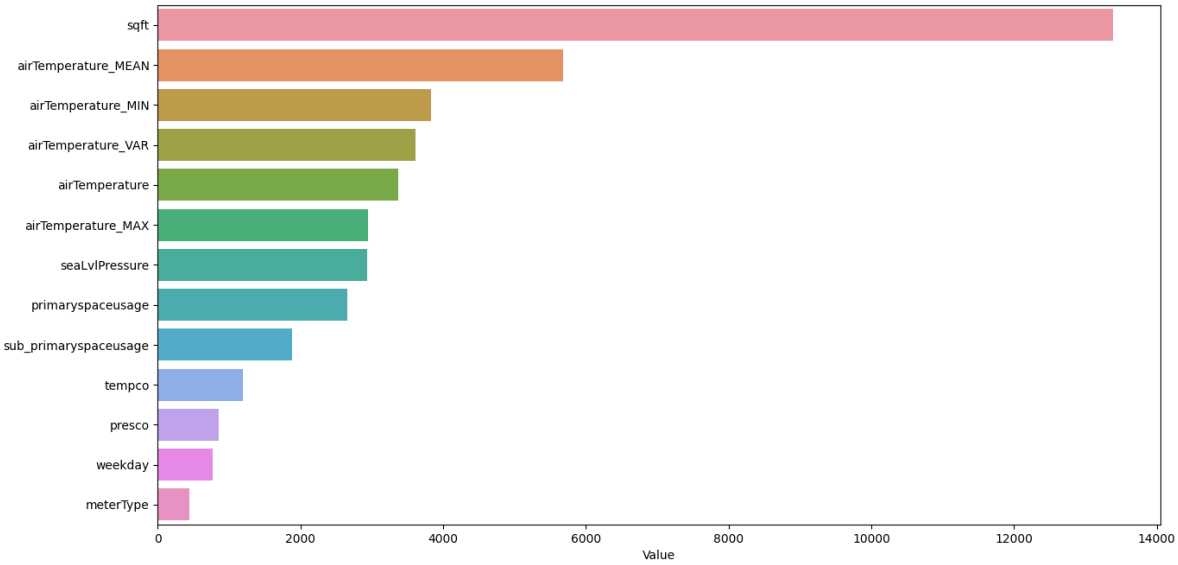



Fig. 14. Feature importance graph as reported by LGBM using 2016 training dataset.

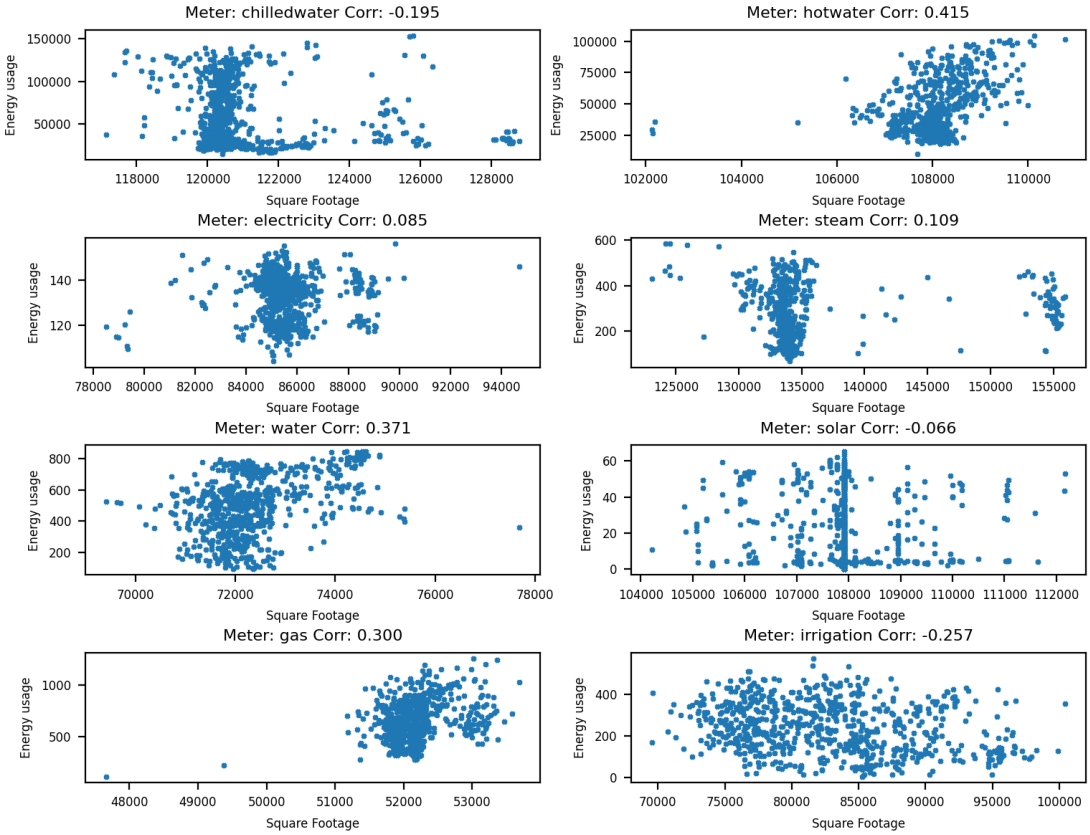



Fig. 14. Correlation plot of energy usage versus building square footage, separated by meter type.

## 7. Summary

A machine learning based regression model is developed that accurately predicts energy usage in buildings. The model utilizes the Light Gradient Boosting Model (LGBM) algorithm from Microsoft. High predictive accuracy is achieved by use of several new features derived from insights obtained through extensive data analysis. An alternate approach that may allow further improvements in accuracy would be to model each meter type individually and sum up the results.