Predicting Energy Usage For Commercial and Residential Buildings
Final Project Report
INFSCI 2725 Data Analytics
David Ball, Jaime Fawcett, Debdas Ghosh, Shruti Gupta, Piu Mallick

Introduction

Our final project attempts to predict energy usage for commercial and residential buildings through predictive modeling of building energy meters. The dataset and initial inspiration for this project came from Kaggle.com.[1] In the process of addressing this challenge, the specific problem we sought to solve shifted, from a more narrow prompt to a problem with broader scope and more valuable results. This paper will describe that evolution, along with the other aspects of our analysis: data understanding and preparation, modeling, evaluation of our model, and planning for deployment.

Business Understanding

For any building owner or user, energy usage represents a significant, ongoing cost. This cost can often be impacted positively by installing energy efficiency improvements to the building. However, evaluating the impact of these improvements can be challenging. What follows represents our initial approach to addressing this problem.

Initial Problem Statement

Do energy efficient modifications actually lead to lower energy usage?

This question was posed in the context of validating the cost-effectiveness of investments in building energy efficiencies after-the-fact, or in other words, justifying a sunk cost. The value in this approach stemmed from the use of pay-for-performance financing in building efficiency retrofitting projects, which in many cases require a direct comparison of "before and after" costs. Our model attempted to answer this problem by modeling energy usage of buildings that had not yet undergone modifications. The model would then, theoretically, be used to predict pre-modification energy use over a post-modification period for the building, producing a predicted "before" cost that could be compared to the real-world "after" cost for that same period, for the buildings that underwent modification. Therefore, comparing "pre-modification" cost and "post-modification" cost would allow owners to see if modifications resulted in energy and cost savings.

Data Understanding

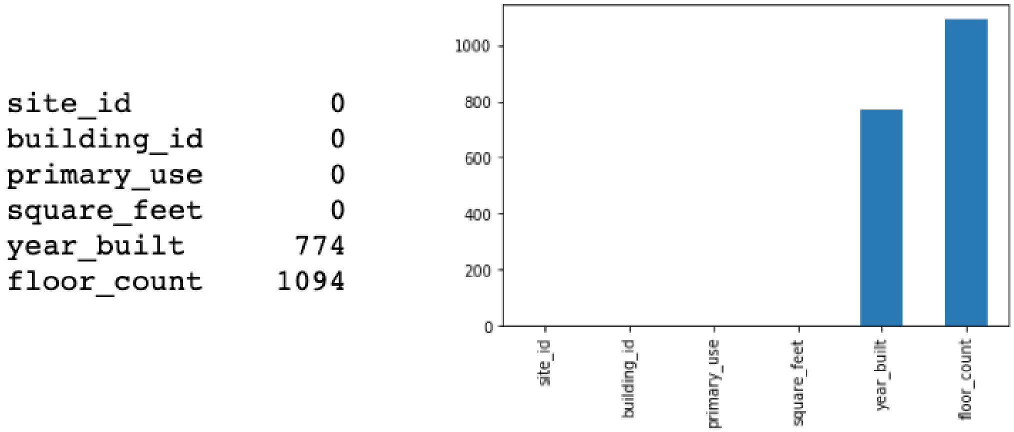
The model was built using three datasets: building data, weather data, and meter data. These were collected *hourly* over the course of 2016, across 16 different cities and 1449 buildings spread among those cities. The building dataset allowed our model to account for different types of buildings based on their use, such as educational, residential, office, healthcare, and others. The weather dataset provided the key predictive features that shifted day-to-day and over the course of the year, as energy usage, particularly for heating and cooling, depends in large part on weather conditions. The meter dataset contained our target variable, the meter readings, as well as the type of meter(s) used in each building.

Data Preparation

In order to ready the data for use in our modeling process, we had to carry out several transformations. We dropped the floor count and year built features in the building dataset due to large proportions of missing data in those features (Figure 1). We also had large portions of missing data in several features of the weather dataset, including cloud coverage, wind direction, and precipitation depth (Figure 2). Others (e.g. air temperature, dew temperature, and wind speed) had smaller percentages of missing data. In order to maintain as much as possible regarding the features' precision and predictive power, we filled in missing values in these features with the median of the available values for that day.

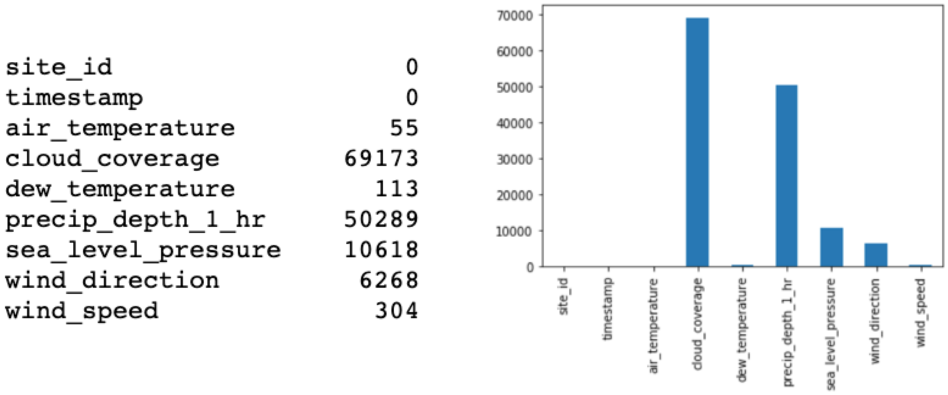
[1] ASHRAE - Great Energy Predictor III, Kaggle. https://www.kaggle.com/c/ashrae-energy-prediction

Our preparation of the meter dataset was determined by our strategic decision to limit the scope of our analysis to those buildings which used only electric meters. We did this in order to avoid complexities that would be difficult to account for in our model, such as different measurement methods for different types of meters and the unpredictable distribution of energy use for a single building across multiple meters. While this does limit the context in which we will be able to deploy our model, we judged this a necessary cost of being able to feasibly model our target variable. Figure 3 shows the frequency of buildings with different meter/energy types (with replacement for buildings with multiple energy types). This decision ultimately affected both our meter dataset and our building dataset. After removing all buildings with non-electric meter readings, were left with 866 buildings, or approximately 60% of the original datasets 1449 buildings.



```
site_id          0
building_id      0
primary_use      0
square_feet      0
year_built     774
floor_count   1094
```

Figure 1. Missing data values in building dataset.



```
site_id                   0
timestamp                 0
air_temperature          55
cloud_coverage        69173
dew_temperature         113
precip_depth_1_hr     50289
sea_level_pressure    10618
wind_direction         6268
wind_speed              304
```

Figure 2. Missing data values in weather dataset.

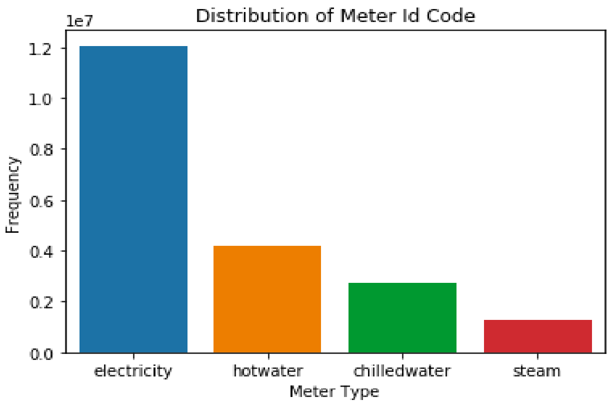


Figure 3. Frequency of buildings with different meter types

We also learned something interesting about our meter and weather datasets. Both datasets included hourly collection (each instance represented one hour of weather readings per weather site, or one hour of meter reading per building). We learned that meter data was collected in local time while weather data was collected in universal time, so we converted weather data to local time, based on the locations of the weather sites (See presentation, slide 25).

After combining the three datasets, we also converted our target variable, meter_reading which was recorded in kWh, to meter_reading_sq_ft (kWh/square feet), by dividing the meter_reading for each instance by the square_feet value for that instance. This provided a target variable more comparable to the industry standard. We also split the timestamp feature into date and time (separate features). Our final dataset had 28 features and 7,526,433 instances.

Modeling

Our model needed to predict a numerical variable, so we made the natural choice to use a multivariate regression. Our problem could not be viably altered to make our target categorical, which ruled out classification models. We also determined there was no need for ranking or ordering, and as a result ruled out a logarithmic model. Given our limited modeling options, our ability to improve the accuracy of our model came down to feature selection and further data cleaning. Initially, we ran four models:

- Model 1: All features included, including date and time. We included all features because we found that dropping any increased our error. This was supported also by examining feature correlations, which showed no strong relationships.
- Model 2: All features included, but instances with 0-value meter_reading_sq_ft removed as we thought these must be erroneous meter readings.
- Model 3: All features included, 0-value meter_reading_sq_ft placed back in as Model 2 actually had higher error rates and lower R-squared. Range of meter_reading_sq_ft reduced based on industry averages.
- Model 4: Same as Model 3, but time included as a feature (24 additional categorical features added).

The most impactful of these decisions was to filter the meter readings by reducing the range (Model 3). We researched the average energy usage for a commercial building (scaled for size) and looked at the distribution of our meter readings, deciding ultimately on a cutoff point roughly one and a half times the average (around 0.009 kWh/square foot; see Figure 4). By eliminating these outliers from our training data, we were able to greatly reduce the error of our model.

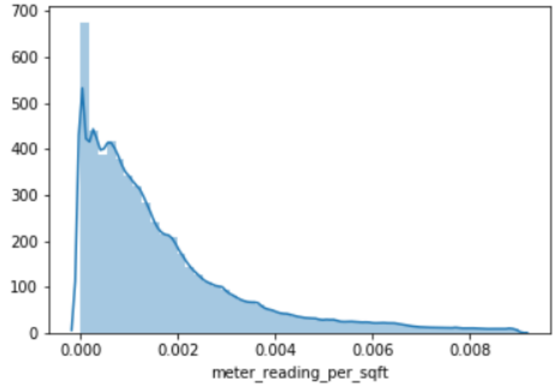


Figure 4. Distribution of meter_reading_square_ft feature

Evaluation

At this stage in our process, though we were making progress with our model, we were also harboring concerns about its limitations. While our error scores were low, our R-squared values for our models were also low, showing a poor fit for our models (Table 1). Our ability to experiment with alternative modeling approaches was limited, due to the nature of our problem. It was not clear to us how to handle the 0 meter readings. While we felt positive that these must be erroneous and the number of 0-value meter readings were numerous. Additionally, after seeking feedback, shortcomings in our problem statement were brought into sharper focus.[2] We decided to revisit this aspect of our project.

Table 1. Error scores and R-squared values for Models 1 - 4

Metric	Model 1	Model 2	Model 3	Model 4
Mean Absolute Error (MAE)	0.0020	0.0021	0.0012	0.0012
Mean Squared Error (MSE)	1.9135e^{-05}	1.9307e^{-05}	2.5962e^{-06}	2.5822e^{-06}
Root Mean Squared Error (RMSE)	0.00437	0.00439	0.00161	0.00160
R-Squared	0.074	0.101	0.124	0.129

Business Understanding Revisited

As previously mentioned, our initial problem was addressing the need for counterfactual pre-modification energy data to compare against real post-modification energy data. This represents a narrow set of circumstances within the larger business context of evaluating building energy efficiency. The model we were building could be used to solve that problem, but another model very much like it could be used to do much more. In order to move towards unlocking this potential, we first had to reframe our problem.

Revised Problem Statement

How much can you expect to save on your energy bill if you make an energy saving upgrade to a building?

This problem encompasses our initial task, while expanding the scope to evaluation of a variety of other possible situations. These include evaluating a specific efficiency upgrade for a certain building or comparing different potential upgrades to see which would be most beneficial. Given this greater range of applications, from a business perspective this is ultimately a much more valuable problem to solve.

Modeling and Data Revisited

After revising our problem statement, we returned to our model and attempted to further reduce the error in our results. Our approach at this time was to break down meter readings by month, rather than by hour, in order to obtain a more reliable target value (Model 5). Energy readings are frequently averaged over a month, which accounts for fluctuations hour to hour and day to day. We also averaged weather data by month, so that our meter data would continue matching correctly with our weather data. We thought this would help improve the fit of our model. Finally, we shrunk our building type feature by merging similar categories, reducing the total count of categories from 15 to 5 (Model 6). While the latter change did not improve our model, adjusting the meter readings to a monthly breakdown did improve the error rate in our results marginally (Table 2).

Notably, the way our model attempted to solve the problem did not shift with our problem statement. This is because our general approach of predicting a meter reading given information on the building and the weather applied just as well to our revised problem statement as it did to our initial problem statement. Where changes in the approach would be needed is in the building information used by the model, which we will discuss as part of the potential deployment of our model .

[2] Thank you for your guidance Dr. Walker!

Table 2. Error scores and R-squared values for Models 1 - 6

Metric	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Mean Absolute Error (MAE)	0.0020	0.0021	0.0012	0.0012	0.7451	0.7516
Mean Squared Error (MSE)	$1.9135e^{-05}$	$1.9307e^{-05}$	$2.5962e^{-06}$	$2.5822e^{-06}$	1.0696	1.0849
Root Mean Squared Error (RMSE)	0.00437	0.00439	0.00161	0.00160	1.0342	1.0416
R-Squared	0.074	0.101	0.124	0.129	0.135	0.122

Deployment

Overall, our model does not answer the revised problem statement as optimally as desired. While our errors were low, our model featured poor fitting to the data as shown by the low r-squared values, even for our best model, Model 5.

The ability of our model to successfully answer our revised problem statement depends on having access to data on the composition of the buildings in the set. Specifically, what is required is data on the type of energy efficient modification the buildings underwent. Access to such data would allow us to retrain our model to predict energy use with this specific modification. Then, the model could be deployed on buildings without the specified modification and provide an estimation of the energy savings the building could experience if it did receive the modification. This would provide an answer to our revised question, namely what cost savings could be experienced given a specific energy saving modification.

Additionally, our model could be further empowered with even more detailed building efficiency data. Ideally, we would expand our dataset to answer questions such as the following: what efficiency-enhancing design or construction elements were a part of its construction? What specific efficiency upgrades has it undergone since construction? To what extent does it meet certain efficiency standards? Acquisition and organization of this data would require substantial domain knowledge and represent a sizeable expansion of the building dataset. However, granular efficiency data such as this would empower our model to not only estimate the savings of individual improvements, but to compare and evaluate different improvements to select the most beneficial.

Conclusion

As a project team, we appreciated the complexity and challenges this project posed and it provided a good opportunity to critically think and analyze the many factors that influence data mining and modeling. Our project presented what seemed to be a simple problem and question that was discovered to be rather complex to solve given the data available. While we were able to overcome many of these challenges through our data preparation, more sophisticated methods could be employed to handle elements of the data we were not able to address, such as better ways to handle missing values, skewed data, or 0-value instances of the data. Overall, this project was beneficial in illuminating the complexities of big data problems and the advantages of data analytic methods and processes, such as the CRISP-DM, to solve such problems.