INFSCI 2725 Data Analytics – Homework 3
David Ball, Jaime Fawcett, Debdas Ghosh, Shruti Gupta, Piu Mallick

Business Understanding

The following describes the problem context and problem statement we are going to attempt to answer and address for our final project.

Problem Context

Homeowners and building owners can spend thousands of dollars a year on electric energy to heat, cool, light, and operate their homes and buildings. Building modifications such as energy efficient windows and insulation can help to reduce costs. However, it can be difficult to know exactly how much energy, and therefore money, the owner is saving after modifications have been made. This is because energy meter readings after modifications have been made have no other readings to which to compare. It is not accurate to compare energy readings post-modification to energy readings from the prior year (pre-modification), as there are many variables that could affect the amount of energy being used year to year. For example, the year prior could have had a warmer summer leading to more air conditioning use. Instead, it is more accurate to be to compare 'pre-modification' and 'post-modification' energy use readings over the same time period. But how does one obtain a 'pre-modification' energy reading once modifications have already been made?

The proposed model will predict a building's energy usage (i.e., meter reading) given specific weather conditions and facts about the building (e.g., size of building, use of the building, etc.) By predicting a building's pre-modification energy usage for a certain time period post-modification, we can compare the expected pre-modification energy usage to the building's *actual* energy usage over the same time period. From this information, stakeholders can learn whether energy-efficient modifications led to lower energy usage, and therefore cost savings. This can help building owners, as well as engineers and building designers, determine whether future investment in the selected energy-efficient modifications is cost effective and should be used to modify other buildings.

Problem Statement

The proposed model will predict electrical energy usage for different buildings given specific weather conditions and facts about the building (e.g., square foot area of the building, use of the building, etc.) so that owners can more accurately assess whether energy efficient modifications actually led to lower energy usage and therefore, cost savings.

Data Understanding

The following describes the datasets we obtained and will use to create our model to answer the problem statement above.

Available Data and Collection

We obtained the dataset for this model and problem from Kaggle.com.[1] We will be using three available datasets for constructing our model: (1) Building Dataset; (2) Weather Dataset; and (3) Meter Dataset. Table 1 under *Data Preparation* presents the features provided in each dataset.

Limited information is provided on how the data was collected. The data was collected over the course of 2016 for a total of 1448 buildings. The data across the three datasets can be combined using the site_id and building_id features to create a complete dataset that will provide weather, building, and meter data for each hour of the entire 2016 calendar year (January 1 – December 31).

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

The *building dataset* will allow our model to account for differences between types of buildings and their use, especially when thinking about the meter data and timestamp features. For example, we will want our model to account for the fact that public buildings (like offices) are more likely to be used during the day compared to residential buildings (that might be used more so in the evening as tenants are typically at work or out of the home during the day). So, we could expect higher meter readings, for example from the hours of 9:00 AM to 6:00 PM for public buildings. Buildings with larger square-foot area will likely use more energy to heat or cool, so we will want to understand and account for these differences.

The *weather dataset* will help the model to account for changes in the time period and weather over the course of the year. We will want our model to account for the fact that energy usage is likely to be highest during the warmest and coolest parts of the year when inhabitants might use a lot of air-conditioning or heating.

The *meter dataset* will ultimately be what our model is trained on (when combined with weather and building data). It contains the meter_reading feature which is our target variable. The meter feature is an interesting feature as it describes the type of meter (and therefore type of energy source) being used. How we will use this feature will be described in the next section.

Data Preparation

The following presents what we have accomplished so far in terms of preparing and cleaning our data. Further data preparation and data cleaning processes may need to take place as we begin preliminary analyses on our data and run our first models.

Cleaning the Individual Datasets

In order to prepare the data for our model, several transformations have been conducted thus far. The primary change involved merging the three datasets (meter, weather, and building datasets) into an overall training dataset. However, we first cleaned the individual datasets. The following describes how we have cleaned and transformed the data into a usable format thus far.

Building Data

For the building dataset, we dropped two features that had a high number of missing values: floor_count and year_built (Figure 1). There is a large proportion of data for these features that are missing. Therefore, this data would not be helpful for contributing to our model. None of the other data features had missing values, so we preserved these features. Therefore, our final building dataset had 4 features and 1,449 instances.

Weather Data

For the weather dataset, we split the timestamp feature into two date and time features and then removed the timestamp feature. This will allow for easier manipulation when analyzing the data by date across buildings or by time across buildings.

We dropped three features that had a high number of missing values: cloud_coverage, wind_direction, and precip_depth_1_hr (Figure 2). Other features (e.g., air temperature, dew temperature, and wind speed, etc.) also had missing values. Removing individual instances of these features that had missing values would remove the entire date and time information that we need to preserve in order to have complete data for each possible date and time of the year when merging the datasets. It is important that we preserve as many of the instances of the data as possible, especially when considering date and time, so that our model

will be more precise and not rely on aggregate values (e.g., means across the day) for these weather data features.

Therefore, in order to preserve these instances of the data, we identified missing values for a feature that occurred in a day. We then calculated the median of that feature for that day using all available values for the day. We used median values because the features had skewed distributions, so median provides a better measure of central tendency compared to mean or mode. So, for example, the data had missing values for air_temperature for January 16 at times 8:00 AM, 9:00 AM, and 10:00 AM. Therefore, we calculated the median air_temperature value using all available air_temperature values for January 16. Then, we replaced the missing values with the median value calculated for the day. We did this for all missing values within each day, across each of the remaining features. Therefore, our final weather dataset had 7 features and 139,773 instances.

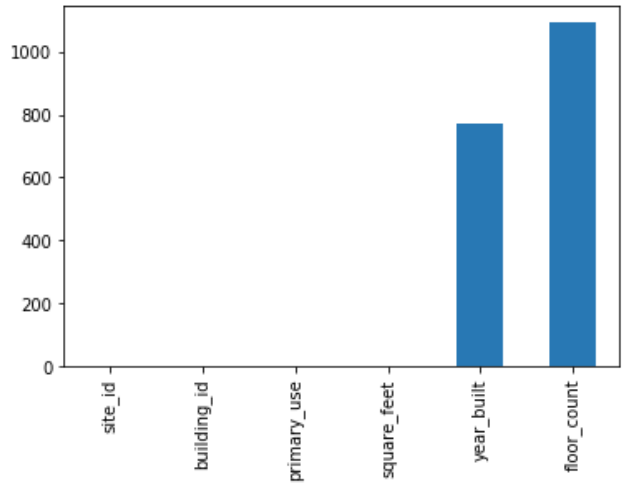

Figure 1. Missing values – building dataset

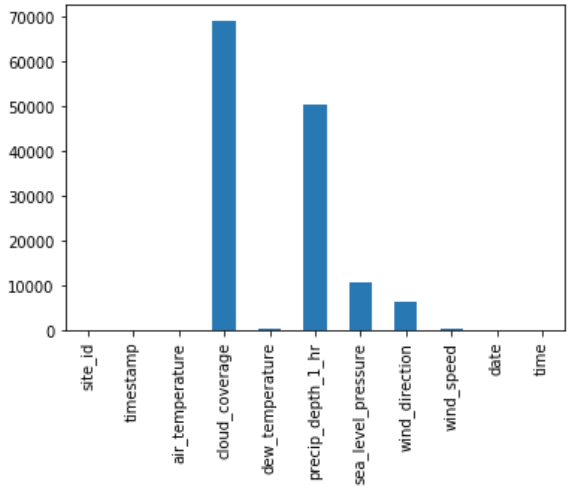

Figure 2. Missing values - weather dataset

Meter Data

For the meter dataset we first dropped any data instances (rows) that were not associated with electricity-powered energy use (value 0 for the meter feature). We did this in order to simplify some of the assumptions in our model. For example, one complexity that is difficult to account for is the fact that different energy types (electricity, chilled water, etc.) are not all measured in the same way. While they are all measured using kWh, chilled water for example only takes a meter reading when the energy level rises above a certain threshold. Accounting for this variability requires a complex understanding of energy use methods, particularly when considering that some buildings combined electrical energy with some other energy type. Therefore, we limited our problem context and problem statement to only predicting energy usage for buildings that solely used electrical energy. One step we still need to complete is that we need to drop all buildings that used multiple energy types. We can then drop the *meter* feature because this feature will now be pure (all 0s) and therefore not useful in predicting energy usage for our model.

Second, similar to the weather dataset, we split the timestamp feature into two features: date and time. This allows for easier manipulation when analyzing the data by date across buildings or by time across buildings. Therefore, our final meter dataset will have 4 features. We cannot determine the final number of instances until we drop multi-energy buildings.

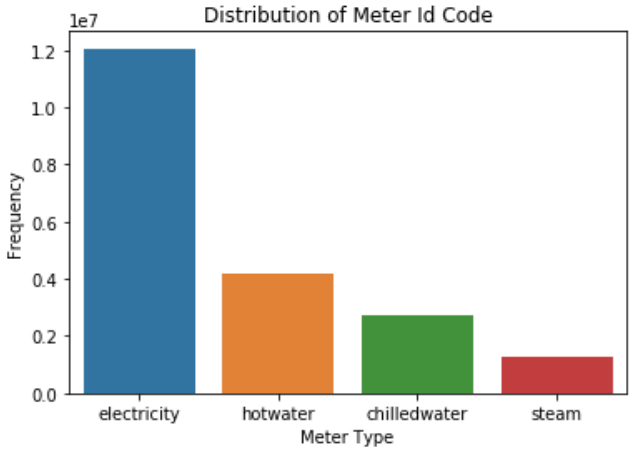

Figure 3. Frequency distribution of different energy use (meter) types within Meter Dataset

Merging the Datasets and Final Transformations

Now that the individual datasets were cleaned, missing values addressed, and unneeded features removed, the final step was to combine the datasets. The datasets were combined using python. After the datasets were merged, we need to complete one more transformation which is converting the meter_reading to meter_reading_square_foot – in other words, the energy usage per square foot of the building.

Table 1. Datasets, Features, and Descriptions

Feature	Description
Building Data	
site_id	An ID code that represents the weather station associated with the building. Numerical.
building_id	An ID code that represents the unique building. Numerical.
primary_use	Represents the buildings primary use: education, lodging, residential, retail, office, entertainment, other, parking, public services, warehouse, food sales, religious worship, healthcare, and technology.
square_feet	Represents the total square foot area of the building. Numerical.
year_built	Represents the year the building was built. Numerical.
floor_count	Represents the total number of floors in the building. Numerical.
Weather Data	
site_id	An ID code that represents the unique weather station. Numerical.
timestamp	Represents when the time period the weather data was collected. Provides year, month, day, and hour. Numerical.
air_temperature	Represents the air temperature reading at the timestamp in Celsius. Numerical.
cloud_coverage	Represents the amount of cloud coverage at the timestamp in oktas. Numerical.
dew_temperature	Represents the dew temperature reading at the timestamp in Celsius. Numerical.
precip_depth_1_hr	Represents the amount of precipitation per hour at the timestamp in millimeters. Numerical.
sea_level_pressure	Represents the sea level pressure at the timestamp in millibar per hectopascals. Numerical.
wind_direction	Represents the wind direction at the timestamp in compass direction (degrees). Numerical.
wind_speed	Represents the wind speed at the timestamp in meters per second. Numerical.
date	Represents the day that the measurements were taken. Numerical.
time	Represents the hour that the measurements were taken. Numerical
Meter Data	

Feature	Description
building_id	An ID code that represents the unique building. Numerical.
meter	The type of meter in the building: 0: electricity, 1: chilled water, 2: steam, hot water: 3.
timestamp	Represents when the time period the meter_reading data was collected. Provides year, month, day, and hour. Numerical.
date	Represents the day that the measurements were taken. Numerical.
time	Represents the hour that the measurements were taken. Numerical.
meter_reading	Energy consumption in kilowatt hours (kWh). Numerical.
meter_reading_sq_ft	The Target Variable. Energy consumption in kilowatt hours (kWh), per square foot.
Orange – removed features; blue – added features; green – target variable.	

Modeling

We propose using a regression model to predict energy usage, given that the target variable is numerical. No ranking or ordering is necessary so a logarithmic model is not suitable. We may propose developing a regression tree model as well if preliminary data analysis supports such a model.

In determining our model, we are first going to want to understand the information gain and entropy of the different features that remain in our final dataset. We also want to run some additional correlations to understand the relationship between or metrics.

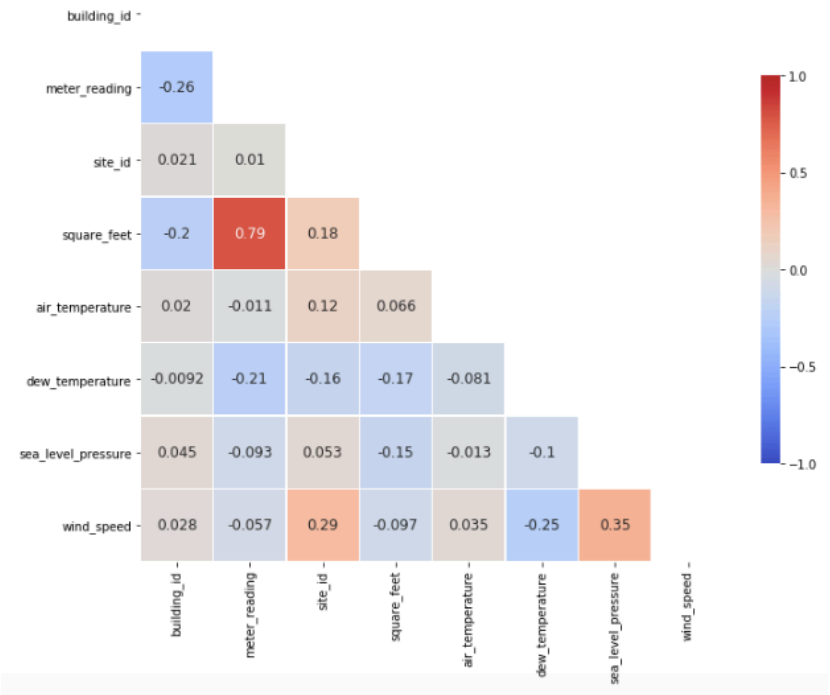

Figure 4. Correlations between features in final merged dataset

Evaluation

The model will be evaluated using a holdout method. One thing to keep in mind is that the data is closely connected to date and time features, and date and time of year greatly influences other features such as weather and meter readings. Therefore, to be sure we have a representative holdout dataset, we need to be sure we are including data from across the year, assuming that weather isn't changing drastically day-

to-day. Therefore, we propose that our holdout data include all Sundays and all Wednesdays for the year. We will use a similar method to also complete a ten-fold cross validation method.

Deployment

One key limitation in the deployment of our model is that this model will only be useful for buildings that solely run on electrical energy. However, many buildings can run on other energy types or a combination of electrical and other energy types (e.g., wind energy, chilled water, steam, etc.) Therefore, our model will be limited during deployment to just estimating energy usage for buildings that only use electrical energy. Additionally, our data has only been collected for certain geographic areas. Therefore, it may not be generalizable to buildings located in other geographic regions.