Predicting Building Energy Consumption

By: Melissa Barona

1. Problem Statement and Objectives

The building sector is worth 10% of the global Growth Domestic Product (GDP), and is expected to continue growing. Buildings use about 40% of global energy and emit 33% of greenhouse gas emissions. of their large contribution to climate change, the implementation of energy efficiency strategies in buildings offers the greatest potential for reducing greenhouse gas emissions. Investing in energy efficiency programs can have significant environmental and financial benefits for our current and future generations.

Measurement and Verification (M&V) techniques are used by facility owners and operations contractors to evaluate and manage energy usage, improve energy conservation measures, and increase energy savings. M&V protocols are also necessary for pay-for-performance (P4P) programs in which energy savings are compensated in the form of payments, and these payments are directly related to the performance of the retrofitted facility. To gain the trust of investors and enhance their chances of being financed, facility owners and contractors must implement M&V plans based on transparent, credible, and accurate projected energy savings.

Governmental entities, energy consumers, financial institutions, utilities, contractors, energy appliance manufacturers, and non-governmental organizations are key stakeholders with vested interest in energy consumption baseline modeling. Baseline models can help develop regulatory policies and standards for buildings and technologies to reduce greenhouse gas emissions. This enables utilities to implement energy efficiency programs that benefit consumers in the form of energy savings. These energy efficiency programs help reduce peak energy demand and benefit utilities that have insufficient generation capacity. This improves grid reliability and avoids costly capital expenditures on new power plants. In their shared interest to promote energy efficiency measures and reduce greenhouse gas emissions, energy service companies and financial institutions can fund these programs and provide financial guidance. Regulation and energy efficiency programs also influence manufacturers in the products they innovate, sell, and market, as well as contractors or technical experts that must be able to provide guidance in the use of the energy-efficient equipment and effectively communicate the benefits of energy efficiency to consumers.

2. Objective

Energy savings are determined by comparing measured energy usage before and after the implementation of an energy efficiency program. An energy consumption baseline is estimated using consumption data before an efficiency measure is implemented (a baseline period of 1-3 years) and projected into a reporting period (1 year following retrofitting) to estimate what the energy use would have been if the measure had not been implemented. This project will use linear, ridge, and random forest regression models to estimate the energy consumption baseline for electric, chilled water, steam, and hot water use.

3. Data

The data for this project is provided by ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) for its Great Energy Predictor III Kaggle competition. The dataset contains 1 year of hourly meter readings from 1,449 buildings at several different sites around the world. The target variable will be the meter reading with the Energy consumption in kWh. The features include building characteristics such as size in square feet, floor count, the year built, and the primary use of the building, as well as weather data including air temperature, dew temperature, cloud coverage, precipitation depth, sea level pressure, and wind direction and speed.

The dataset containing the meter readings contains the following columns:

- **building_id:** ID of the building.
- **meter:** The meter id code corresponding to electricity, chilled water, steam, and hot water meter readings. Units are in kWh.
- **timestamp:** Time at which the measurement was taken.

The dataset containing information about each building contains the following columns:

- **site_id:** ID of the site location of the building.
- **building_id:** ID of the building.
- **primary_use:** Primary use of the building.
- **square_feet:** Gross floor area of the building.
- **year_built:** Year building was opened.
- **floor_count:** Number of floors of the building.

The dataset containing weather information contains the following columns:

- **site_id:** ID of the site location of the building.
- **air_temperature:** Air temperature in degrees Celsius.
- **dew_temperature:** Dew temperature in degrees Celsius.
- **cloud_coverage:** Portion of the sky covered in clouds, in oktas units.
- **precip_depth_1_hr:** Depth of accumulated precipitation in Millimeters
- **sea_level_pressure:** Sea level pressure in Millibar.
- **wind_direction:** Wind direction in compass direction (0-360).
- **wind_speed:** Wind speed in meters per second.

4. Data Cleaning

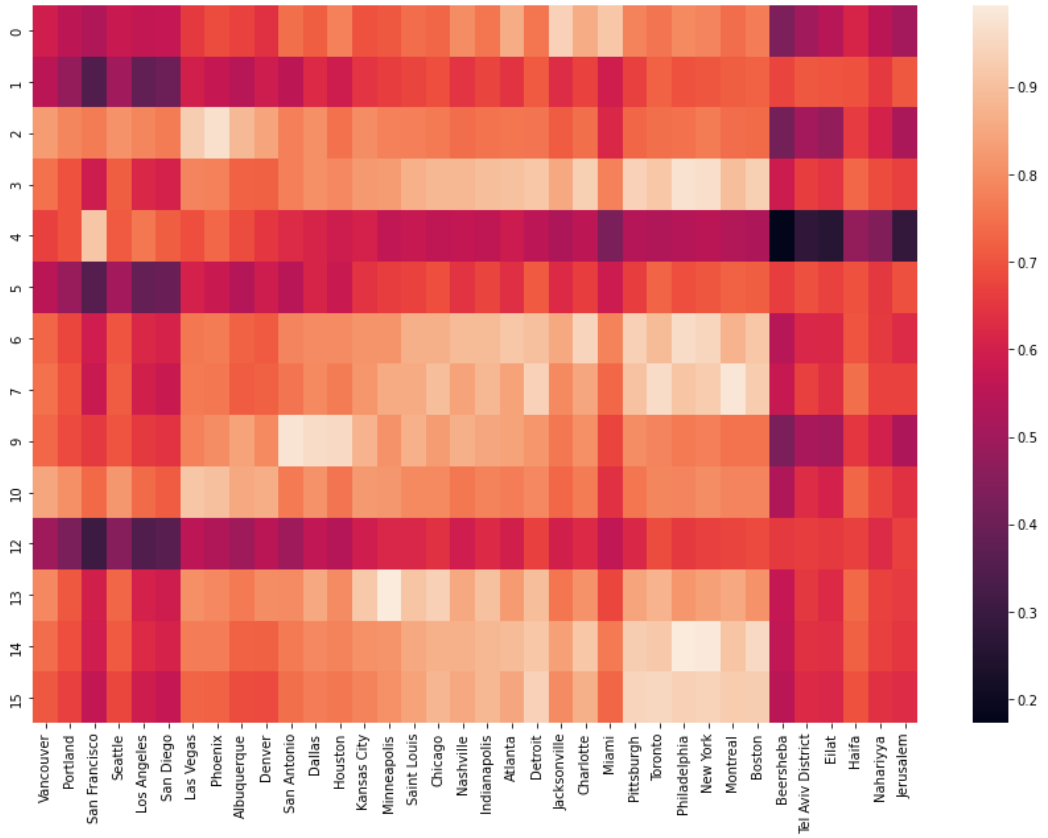
- **Find duplicates:** During data cleaning, I found sites 8 and 11 to be duplicates of sites 0 and 7. I removed sites 8 and 11 from the weather dataset and assigned all buildings in site 0 to site 8, and all buildings in site 7 to site 11.

- **Find number of missing values:** The dataset with the meter readings contained a meter reading value for every hour of the year 2016. No values were null in this dataset. The weather data set is missing an additional 0.49% of all the data it should have for every hour of every day in 2016. Once merged with the dataset containing the meter readings, all meter readings taken at an hour with no corresponding weather information in the weather dataset were left out. This means that in the final dataset for modeling, 0.49% of meter readings were discarded. Additionally, the weather feature missing the most data is cloud_coverage. The total percentage of missing values for precip_depth_1_hr is 40%, which includes 34.6% of null values and 4.4% values labeled as "-1". The dataset containing information about the buildings is missing 75% and 53% of the floor_count and year_built values.

	Count	%
cloud_coverage	56729	46.356691
precip_depth_1_hr	42405	34.651685
sea_level_pressure	10525	8.600613
wind_direction	6018	4.917671
wind_speed	304	0.248417
dew_temperature	87	0.071093
air_temperature	52	0.042492

	Count	%
floor_count	1094	75.500345
year_built	774	53.416149
site_id	0	0.000000
building_id	0	0.000000
primary_use	0	0.000000
square_feet	0	0.000000

- **Align time zones:** During exploratory data analysis, I noticed some odd patterns. For example, electricity usage was higher during the night than during the day, which did not make sense. When I looked this up on the kaggle discussion, a competitor brought to everyone's attention that the weather data was in UTC (universal time coordinated), while the meter readings were in local time. This meant that the timestamps in the weather data had to be shifted to the local times that would correspond to the meter readings. To correct for the timestamps in the weather data, I used external air temperature data for a variety of cities, and calculated the correlation of the external air temperature data with that in the weather dataset for the different sites_ids. This helped me match the site_id to one of the cities in the external air temperature dataset and shift the timestamps in the weather dataset from UTC to local time.



The site_ids were matched to the following locations:

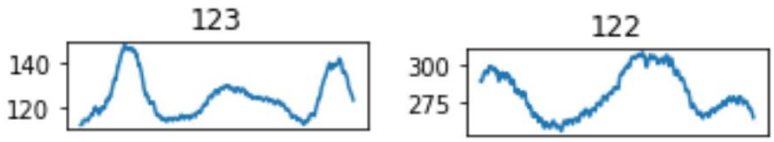
- Site 0: Jacksonville
- Site 2: Pheonix
- Site 3: Philadelphia

- Site 4: San Francisco
- Site 6: Philadelphia
- Site 7: Montreal
- Site 9: San Antonio
- Site 13: Minneapolis
- Site 14: Philadelphia
- Site 15: Toronto

Sites 1, 5, 10, and 12 don't match very well. For these, I may have to look at other possible matches in future iterations of this project.

4. EDA

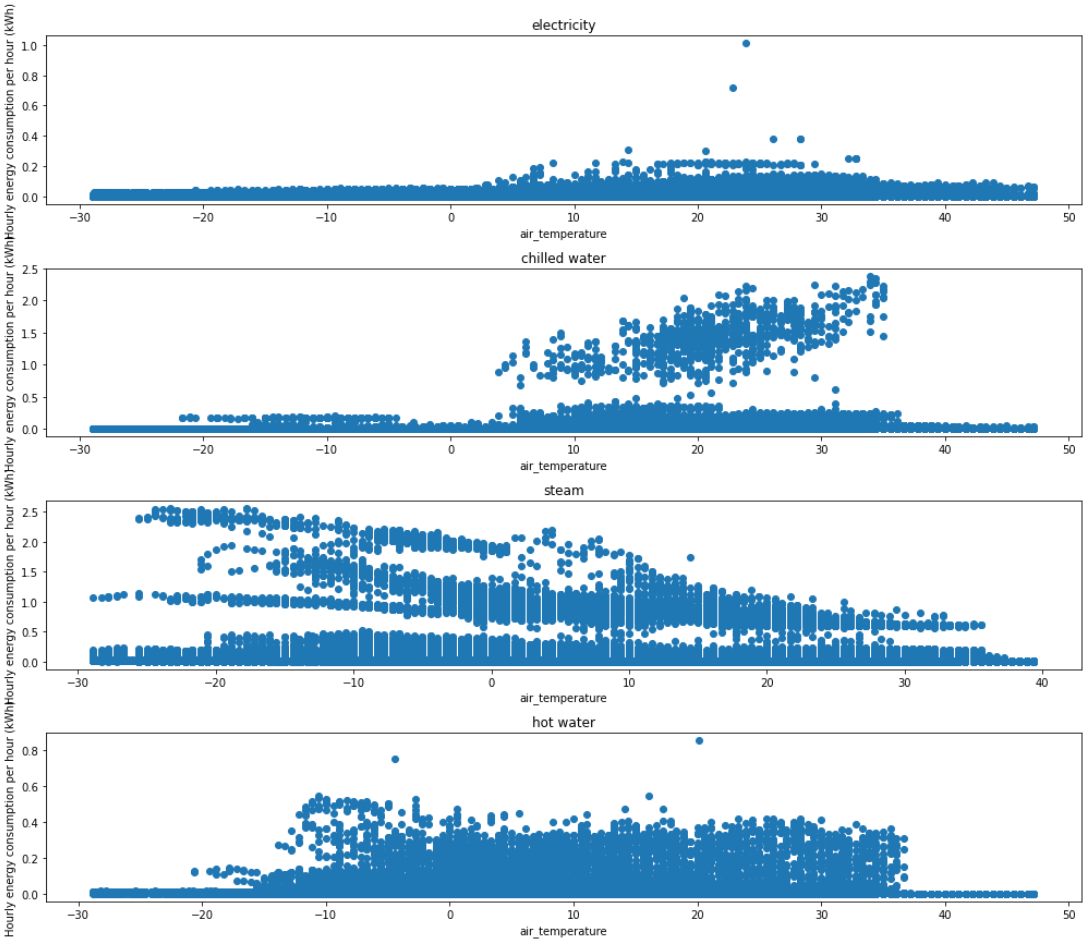
- **Building type and site distributions:** Most buildings are education and office buildings, and most buildings are in site 3 which has weather conditions similar to Philadelphia.
- **Meter readings for each type of building:** All building types have mostly electricity meter readings. Warehouse/storage and services don't have any chilled water meter readings. 'Other', parking, warehouse/storage, manufacturing/industrial, retail, services, utility, and religious worship don't have any hot water meter readings Retail and Religious worship don't have any steam meter readings.
- **Meter readings over time:** Over time, we should observe peaks during winter and during summer for the electricity. This is observed for some buildings as shown below:



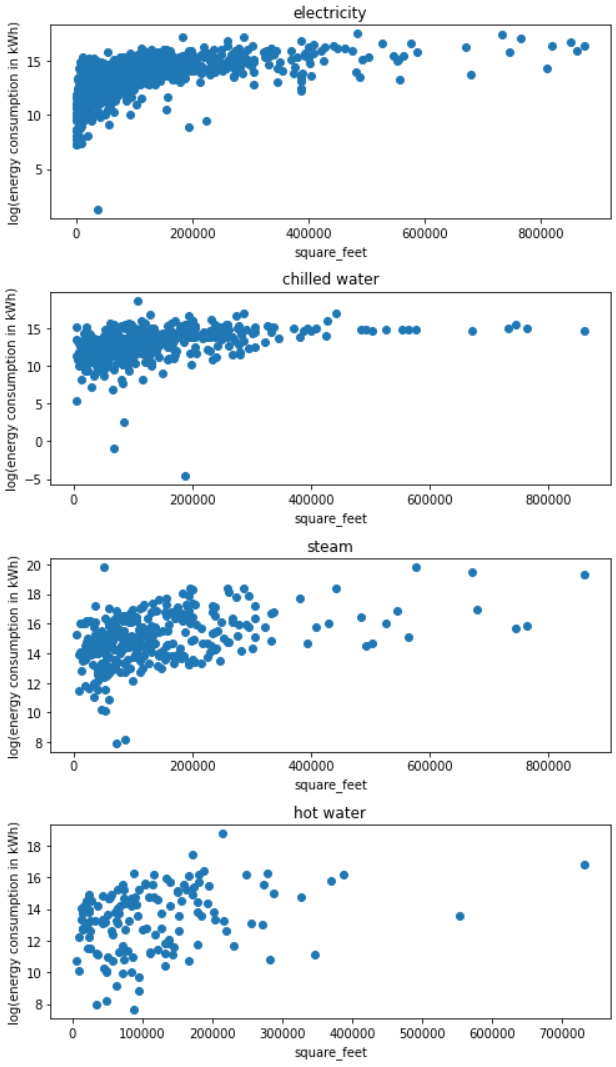
- **How each weather feature can affect energy consumption:**
  - **Size of the building:** Larger buildings consume more energy
  - **Air and dew temperature:** At warmer temperatures, AC will be used, driving up energy consumption. At colder temperatures, the heater will be used, driving up energy consumption
  - **Sea level pressure:** We can try to relate sea level pressure to air pressure. Air pressure is higher during winter because colder air is denser. If air pressure is high against the building, air will enter the building, make the building colder, and the building will work more to get warmer. So high

pressure during winter should lead to higher energy consumption in steam and heat. If air pressure is low against the building, it is warmer outside, warmer air will enter the building, and the building will work more to get cooler. So low pressure during summer should lead to higher energy consumption in chilled water.
    - **Wind speed and wind direction:** Higher wind speed will cool the building. This is good in the summer and bad in the winter. In the summer it helps lower energy consumption, but in the winter, it may lead to higher energy consumption. If wind speed is low, wind direction won't matter if wind speed is high, wind direction will matter more and its effect on energy consumption will depend on the position of the building relative to the direction of the wind.
    - **Precipitation and cloud coverage:** Precipitation and cloud coverage will cool a building in the summer which can result in a decrease in energy consumption. During the winter, precipitation will make a building colder, which may lead to an increase in energy consumption.
- **Energy consumption vs. air temperature and size of the building:** While we have several features for modeling, the main two that more logically could determine energy consumption are air temperature and size of the building. We should expect to observe the following:
    - Higher hot water and steam use during colder temperatures
    - Higher chilled water use during warmer temperatures
    - Higher electricity usage in colder and warmer temperatures
    - Higher energy consumption for bigger buildings

  In the graph below, I am plotting the energy consumption in kWh per square foot vs. the air temperature. We see that the plots corresponding to electricity and hot water usage are not very informative. However, for chilled water use, we do see higher usage for higher temperatures, and for steam, we see lower usage for higher temperatures.



The following graph is a plot of the logarithm of the total energy consumption for each building vs. size of the building in square feet. we see that the bigger the building is, the higher its energy consumption.



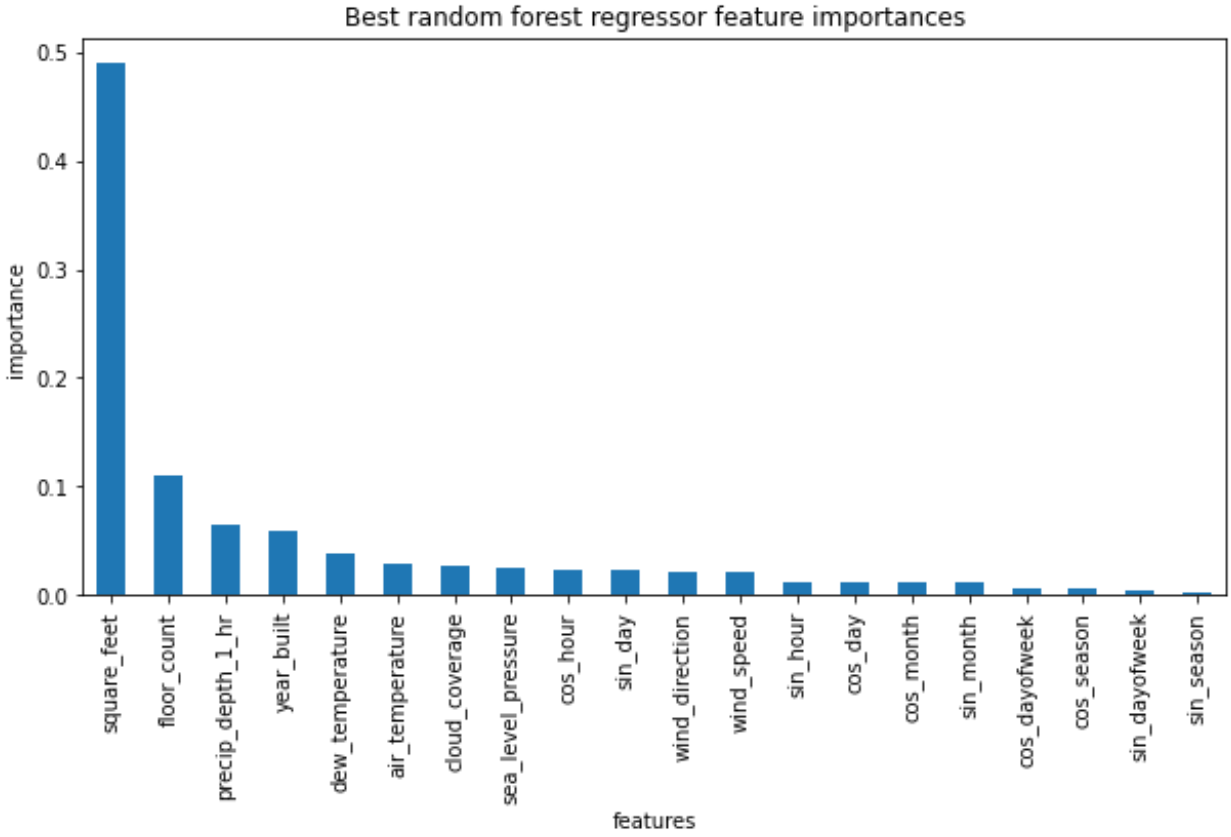
5. Preprocessing, Feature Engineering, and Modeling

Missing data in the weather features were filled in by averaging out the data over time using the moving average. Season, month, day of week, and hour were added as features and transformed into cyclic features using sine and cosine functions.

This dataset contains ~ 20 M rows. It is very difficult to train on this very large dataset using only my computer. Therefore, I took a subsample of a 100,000 rows for the linear regression models, and 10,000 rows for the random forest regression model.

6. Algorithms & Machine Learning

Linear regression, ridge regression, and random forest models were fit to the data. I used the RSME and MAE as metrics to assess the performance of the models. The average score for the linear models was ~0.32, while that of the random forest model was ~0.62. Therefore, our best performance is the random forest regression model as it captures better the non-linearity of the data. In addition to this, Both the linear regression model and the random forest model found square_feet to be the most important feature for predicting energy consumption.



Model	RMSE		MAE	
	Training Set	Testing Set	Training Set	Testing Set
Dummy Regressor	367	374	176	179
Linear Regression	304	307	128	129
Ridge Regression	304	307	128	129
Random Forest Regression	82	205	32	85

7. Predictions

Because the models suggest that the size of the building is the most important feature, the random forest model seems to overestimate the electricity consumption for building 1298. (I should try looking at a different building).

8. Future Improvements

- In the future, I would like to spend more time detecting outliers in the meter readings and addressing them appropriately. For example, the hampel identifier uses moving estimates such as the rolling average to detect outliers in a time series and replace them with the average.

- Both linear regression and random forest regression found that size of the building was the best predictor. This is due to the larger variation in the square_feet feature compared to the rest of the features. If the size of the building were to be used as the only feature to determine the energy consumption baseline, we would not be able to reproduce the seasonality in the energy consumption. Better feature engineering may be needed. What I would do to address this is try to fit a model for a single building to get an idea of the most important weather features to predict energy consumption.