Final Report:

Residential Electricity Prediction

Problem Statement

Residential energy consumption is an important part of the economy and a common focus for improving energy efficiency. Energy consumption can be predicted with an engineering model of a building's thermodynamics, but such models are time-intensive and are subject to error if built on incorrect assumptions about occupant behavior or quality of construction.

Some efforts have been made at streamlining energy estimates through U.S government programs like Energy Star benchmarking or the HUD Utility Schedule Model for utility allowances. However, these make use of limited inputs when more information may be available and are limited in precision. A machine learning model could potentially incorporate more building features to predict consumption for a variety of building types in different geographic regions.

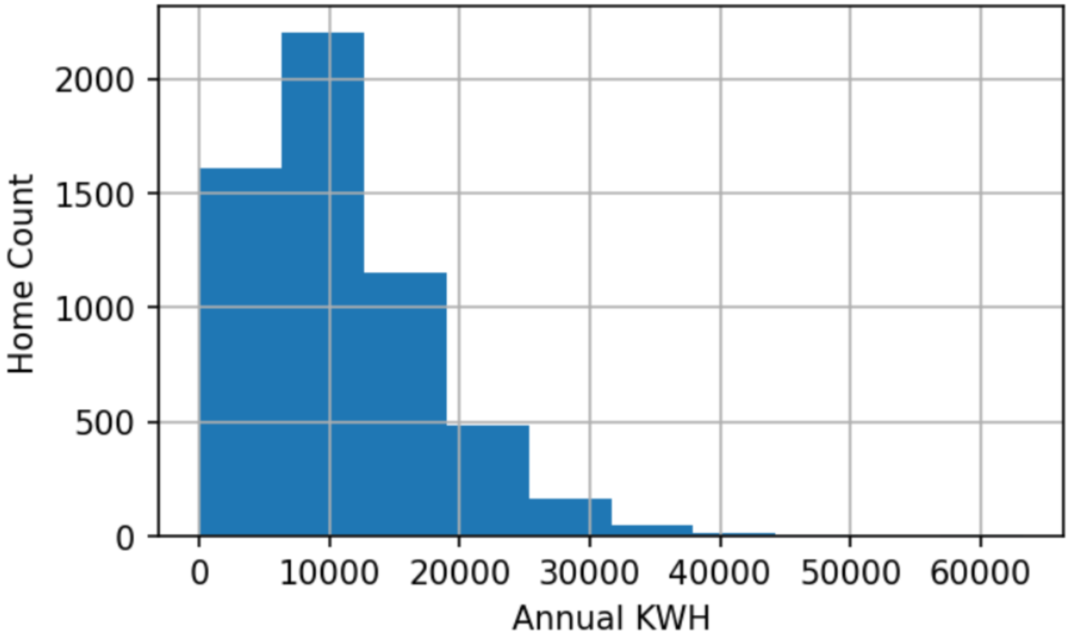
The scope of this problem is limited to electricity prediction and not other utilities such as natural gas.

Exploratory Data Analysis

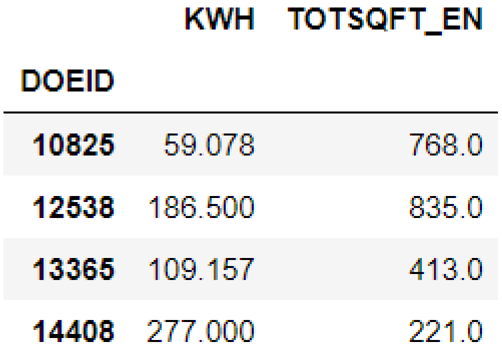
The original form of the dataset is an Excel file downloaded from the Energy Information Administration. It contains 5,686 rows and 758 columns, in which each row represents a unique building.

I performed initial checks on the dataframe by compiling missing values and looking at the distribution of the response variable (KWH). There was only one feature with missing values (NGXBTU), which is irrelevant to this problem because it provides natural gas consumption, which is outside the scope of this problem.

The distribution of KWH consumption is right-skewed. I checked some of the extreme values on either end and the low-values are generally for smaller homes while the large values are for very-large homes. So there do not appear to be any data quality issues with the response variable.

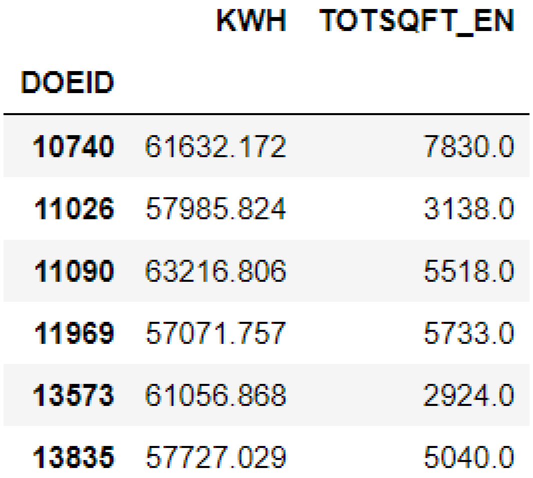


Unusually small kWh values:



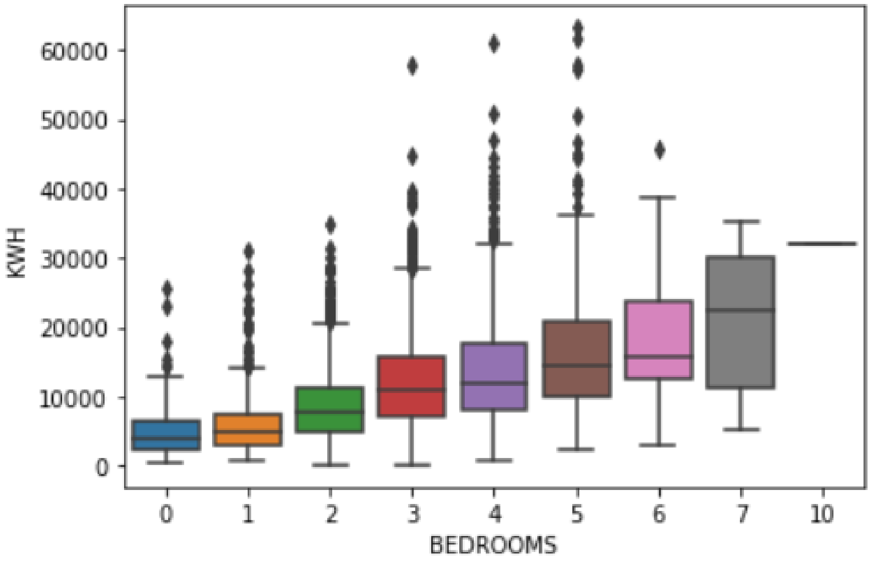
DOEID	KWH	TOTSQFT_EN
10825	59.078	768.0
12538	186.500	835.0
13365	109.157	413.0
14408	277.000	221.0

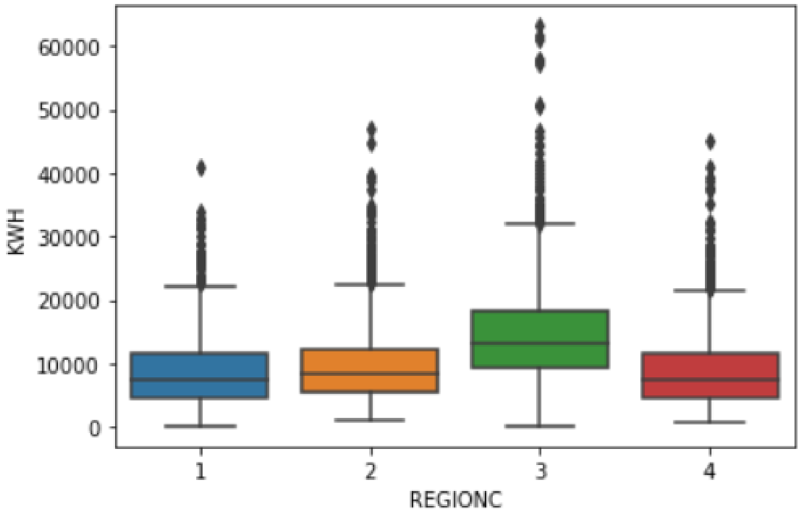
Unusually large kWh values:



DOEID	KWH	TOTSQFT_EN
10740	61632.172	7830.0
11026	57985.824	3138.0
11090	63216.806	5518.0
11969	57071.757	5733.0
13573	61056.868	2924.0
13835	57727.029	5040.0

The dataset is too large to perform some conventional EDA techniques such as correlation heat maps, but I inspected a couple features of expected importance with boxplots to see if any issues or unusual behavior stand out. For the most part, the data looks normal; homes with more bedrooms generally use more electricity, and homes in the south (REGIONC=3) tend to use more electricity, which is probably explained by larger air-conditioning loads.





Data Wrangling

Manual Review

The feature names and variables are not easily interpretable and must be cross-referenced with the Variable and Response Codebook found on EIA's website in Excel format. Inspection of the codebook reveals several obstacles to overcome before the file can be used by an ML algorithm:

- Some variables that appear numeric or ordinal in nature are in fact categorical. An example is that the variables for the feature “REGIONC” are integers 1 through 4. However, the codebook reveals these integers are codes for Northeast, Midwest, South, and West regions.
- It is obvious that some features have no predictive power, such as columns that provide conversion factors between energy units such as Kilowatt-hours to British Thermal Units.
- Some variables are redundant, such as HDD50 and HDD65. These both represent how cold the building’s climate is and differ only by an arbitrary base temperature used to count “heating degree days” over the course of one year.

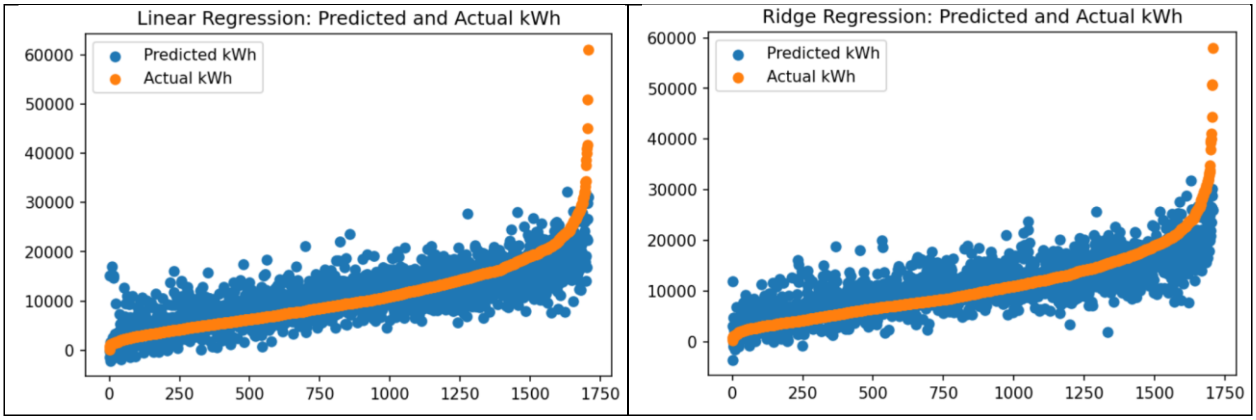
Feature Selection

Redundant and non-predictive variables described above could possibly be eliminated through automated feature selection methods such as Lasso regression or the Boruta algorithm. However, I manually eliminated them in order to apply the algorithm to features with more uncertain significance. I did this by compiling the feature lists in Excel and then reading them into Python and dropping from the dataframe. I then created dummy variables for the list of categorical variables I created. The resulting dataframe contained 309 columns.

I applied Lasso Regression to further reduce features. I experimented with different alpha parameters to test different trade-offs between model simplicity and precision. Alpha=2 reduced the number of features by about 80% while maintaining most of the model accuracy (tested across three different models) and I selected this for the final model.

Model Selection

I tested three models on the selected features: linear regression, lasso regression, and ridge regression. I assessed results using five-fold cross-validation. Most models behaved similarly with only a slight improvement for ridge and lasso regression over linear regression. All had cross-validation scores of 0.6 to 0.62 and correlation coefficients of 0.77 to 0.78 between predicted kWh and test-set kWh. Visually, there is also little difference in how the models perform against the data.



Takeaways

The largest challenge and most significant step in this problem is feature reduction. The number of features was adequately reduced to about 7% of the total in the original dataset. The best model is probably lasso regression because it has built-in regularization, achieves the highest cross-validation score, and is interpretable.

However, there is still a large amount of variation in the outcomes that cannot be explained by the model. The model may be more appropriate for estimating average consumption in a population of similar buildings than for developing a high confidence around the expected consumption of an individual building.

The features with largest coefficients show some potential issues with relying solely on a machine learning method. The most influential feature determined by each of the three models is FUELTUB_21, which is the dummy variable for a hot tub heated by “some other fuel” per the survey questionnaire. Even though electric-heated hot tubs (FUELTUB_5) were also included as a feature using the Lasso selection process, its coefficient is far smaller. This seems illogical considering that a hot tub heated by some other fuel would use zero electricity; only a hot tub heated with electricity would increase electricity consumption.

Suggestions for Future Work

One suggestion is to modify the features of the dataset. Although the dataset was feature rich, it is missing variables that would allow a machine-learning model to approach the same precision as traditional engineering analysis. These include surface area of the buildings, and degree-days and electricity consumption on a monthly basis, rather than just as annual totals. These variables would implicitly define the overall heat transfer coefficient of each building, which would be an improvement because the survey responses in the dataset provide only an ambiguous description of the building construction (for example, “well insulated” versus “adequately insulated” for feature ADQINSUL). The monthly totals are available from the billing data that EIA collects but are not included in the public dataset.

With the building envelope better defined by an overall heat transfer coefficient, an ML algorithm could be used more specifically to estimate consumption created by tenant behavior and their survey responses, with less noise created by uncertainty in the quality of building construction.

Further possible work could include reviewing the ML-selected features to see if they are easily obtainable through surveys, developing confidence intervals on predictions, and developing models for additional utilities besides electricity.