Capstone Project 1: Final Report

Springboard Data Science Career Track
Ishmail Grady
May 2019

Problem Statement

The sociopolitical effects of climate change are shifting the focus of energy sector. Energy providers are facing pressure to be more efficient and environmentally sustainable. How energy providers communicate with their customers about their energy consumption and sustainability will play a crucial role in how their businesses are able to adapt to this changing market. This project will serve as a solution to help energy providers enhance their customer communication by providing a predictive model that will allow its customers to predict their energy consumption based on the characteristics of their household. This predictive model can then be incorporated into an online tool for customers that leverages the existing data of energy providers. This tool will give customers the ability to do the following:

- Predict their energy consumption in a more robust, and precise way using granular information specific to their household.
- Compare their actual consumption to predicted consumption in order to gauge how energy efficient their household is.
- Analyze how various household decisions can affect their energy use, thus providing a means to minimize their consumption.

Ultimately, providing this tool to customers will be beneficial for energy providers for the following reasons:

- It will allow providers to compete with other energy management solutions that can threaten the relevancy of their customer communications.
- Provides a means of collecting data on customers to build better datasets that for applications such as peak load management and targeted marketing for goods and services.

Data Source

The main dataset used to develop this model will come from the microdata of the 2015 Residential Energy Consumption Survey (RECS) Survey conducted by the U.S. Energy Information Administration. This survey is a national sample of housing units that are considered primary residences, as defined by the U.S. Census Bureau. The survey results contain data on 5,686 randomly selected households across the nation. This sample was statistically designed to represent 118.2 million households throughout the country. The RECS Survey contains documentation of their sampling methodology and sampling error. More information on how the sample was designed and how sampling error was computed, see the 2015 RECS Household Characteristics Technical Documentation Summary.

This dataset contains two main types of information: household characteristics and consumption & expenditures. Household characteristics data covers many areas such as appliances, electronics, space heating, household demographics, and more. Consumption & expenditures data contains information on the fuel type(s) used, the end uses of the fuel associated with the various household characteristics, and the dollar values of the energy used. The dataset contains over 700 variables in total; complete list of variables and descriptions can be found can be found in the Variable and Response Codebook. This data is relatively clean and publicly available. Special consideration will be given to the sampling error associated with the data, any imputed values, and the relative weights of each data point.

Data Cleaning

The raw dataset is relatively clean and tidy. Each row in the main dataset represents a respondent in the survey and each column corresponds to distinct survey questions or parameters of the survey construction. All entries column variables contain the expected data types. There are columns in the dataset that correspond to an indicator variable for the usage of various fuel types, household characteristics and demographics of the household of the respondent. If the household does not use a fuel type or has a household type that could not possibly have a certain characteristic these values are blank. Therefore, these values are not truly missing, but will be considered as not applicable.

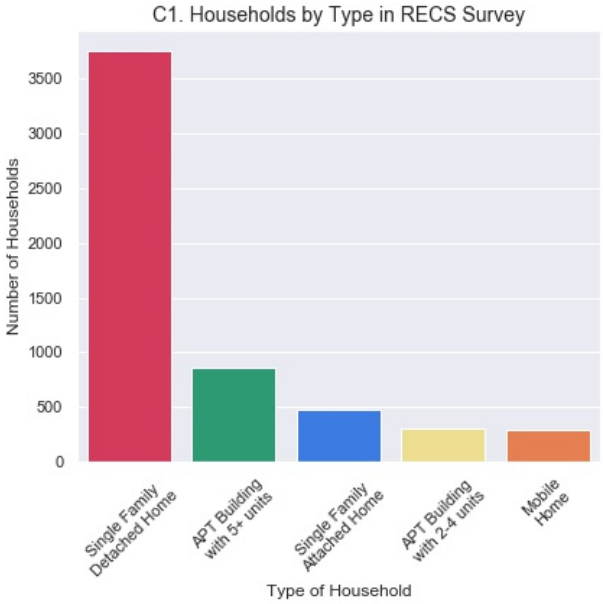
The survey was designed to be statistically representative of all US households. Each observation has an associated weight that corresponds to the number of households the observation represents. Thus, using these weights will ensure that any outliers in the data will be handled appropriately. There is a subset of column variables that may contain imputed values where data was not available. The dataset contains 217 columns that represent imputation flags for 222 columns variables (some imputation flag columns correspond to multiple variables). These imputation flags indicate if the corresponding variable was imputed or not. A separate dataset was constructed that contains observations without imputed values. To do this, for each observation if an imputation flag column indicated an imputed value, the field for the corresponding variable was replaced with NaN.

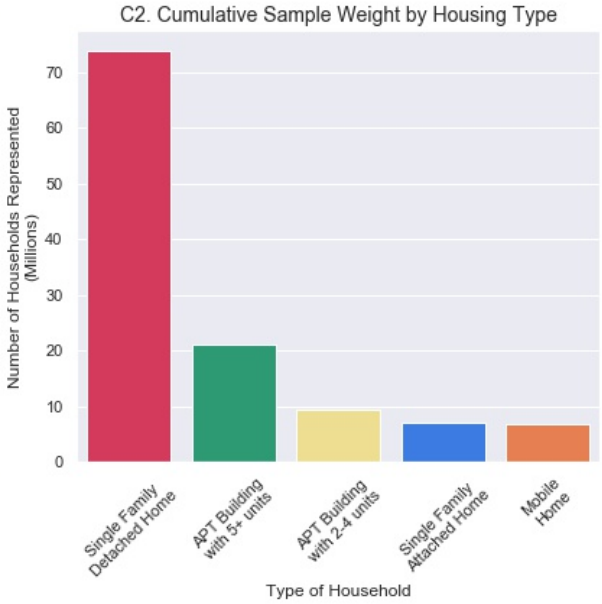
There were 20 questions in the survey that have “Don’t Know” as a possible answer (encoded as -9 in the dataset) the values in the corresponding columns of the dataset were replaced with NaN for all questions that were answered with “Don’t Know”. The rows with the new missing values will remain in the dataset in order to leverage all of the available data. An appropriate predictive model that accounts for missing values will be used in this project. The codebook that contains a description and variable type for all variables, and response labels and encodings for all categorical variables is given in excel file. The information contained in this file was extracted and saved for future use as it will be a valuable tool for understanding the large number of variables. Pickle files containing dictionaries for variable descriptions, variable labels, and mappings for variable responses were made.

Exploratory Data Analysis

To get a better understanding of the dataset before developing a predictive model, exploratory data analysis was conducted with a focus on identifying variables that have a correlation with total

energy consumption. Given that this dataset contains observations for a representative sample of US households, the first step in becoming familiar with the data is understanding what household types the dataset is comprised of.

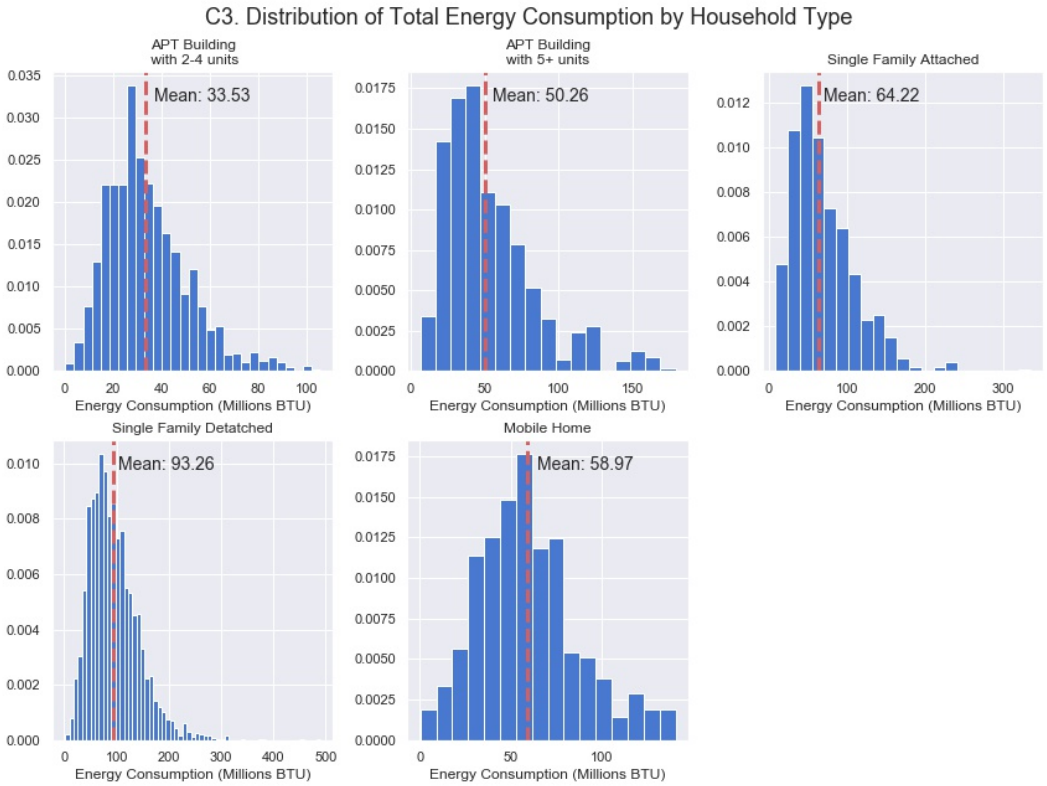




As shown in Figure C1, the majority of households in the dataset are Single Family Detached homes and these households make up 66% of the 5686 observations in the dataset. Figure C2, shows the cumulative sample weights of the observations by housing type. The amount of observations for each household type are roughly related to the cumulative weights of their observations as shown when comparing Figure C1 and C2. The 3,752 observations of Single Family Detached households in the dataset are statistically representative of 73.8 million American households.

According to Figure C3 below, Single Family Detached households consume significantly more energy than other household types. The energy consumption of these households also has a significantly larger variance. As shown in Figure C3, the distribution of energy consumption for each household type is positively skewed, with the exception of mobiles homes which have a more symmetric shape. On average, Single family detached home consumes over twice as much as apartments in buildings with 2-4 units and apartments in buildings with 5+ units. It is surprising to see that mobile homes have a higher mean total energy consumption than apartments, some additional investigation may be necessary to determine the exact cause of this.

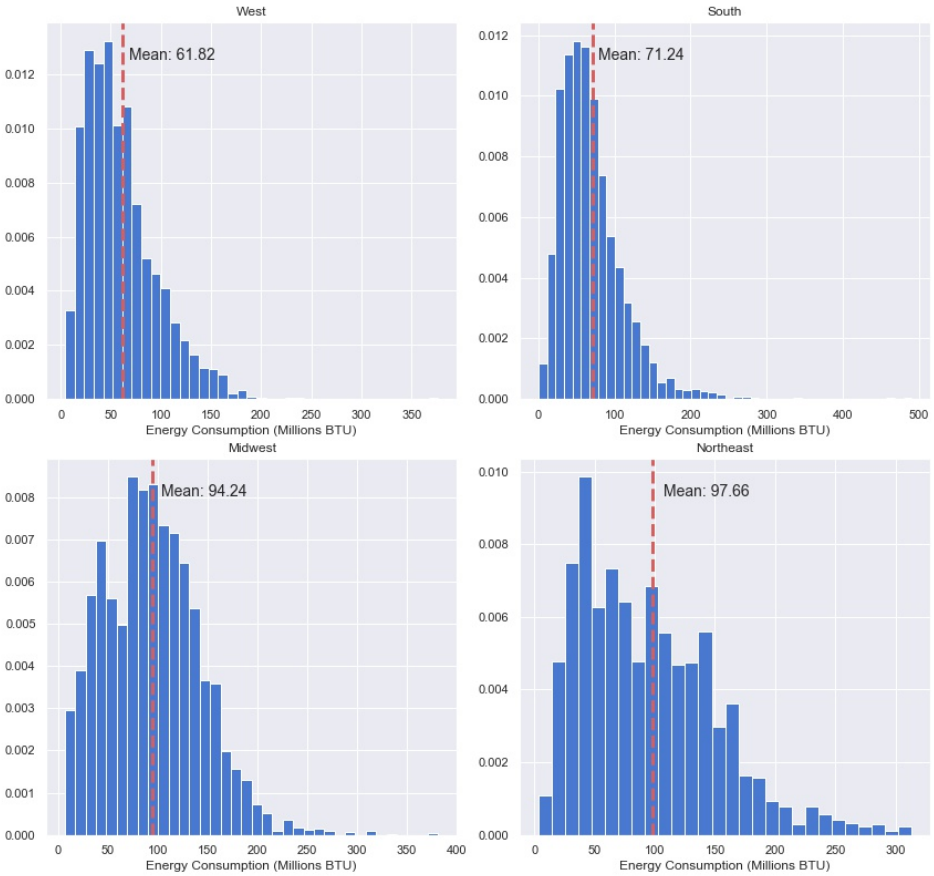
Other important factors to consider in this model are geography and climate. According to Figure C4 shown below, there appears to be a significant difference in the means of the distributions of energy consumption across regions of the country. All four regions have positively skewed distribution. The distributions of energy consumption in the Northeast and Midwest appear to be similar however the distributions of the South and West have significantly lower means. This may suggest that total energy consumption is closely related to the region that a household is in and that households in colder climates consume more energy than those in warmer climates. Further investigation should be done to quantify the significance of the difference in distribution and how other variable related to region or climate affect energy consumption.



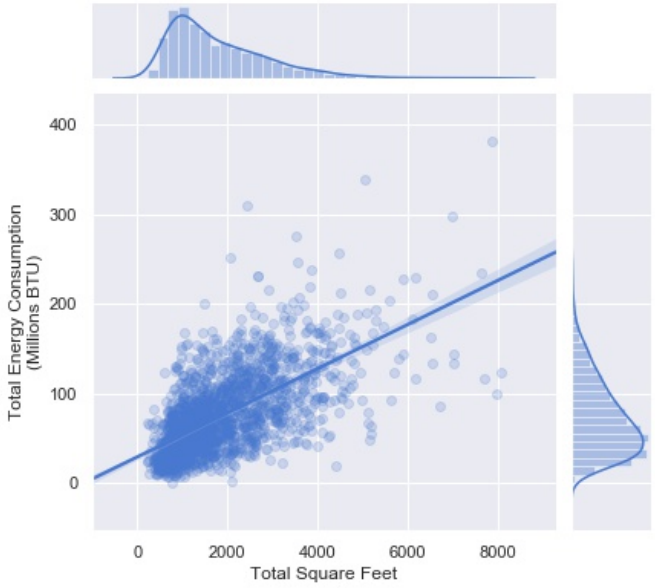
To develop a useful model, it will be important to understand how the variables are related to the primary dependent variable, total energy consumption. The joint plot below in Figure C5 shows the relationship between total square feet in the household and total energy consumption. There appears to a moderate positive correlation between total heated square feet and energy consumption. In fact, these two variables have a pairwise correlation coefficient of .64. Total heated square feet in a household and total energy consumption have a correlation coefficient of .58. The correlation is not very strong; however, total square feet and total heated square feet have the highest correlation to total energy consumption of any numeric household characteristic. A positive correlation between these variables make sense for a variety of reasons. For example, we can expect that households with more square footage to possibly use more energy for heating in cooling. Larger households might also have more inhabitants and more inhabitants could increase the use of electronics household appliances.

This variable will likely be an important feature to include in the final predictive model however more investigation is necessary to fully understand how this relationship changes when conditioned on other variables.

C4. Distribution of Total Energy Consumption by Census Region



C5. Total Square Feet v. Total Energy Consumption



Though the correlation between household size and energy consumption is not surprising it is still important to verify the statistical significance of this correlation within the dataset. After verifying the significance, it will be useful to quantify a range of for the true population correlation between these variables

To verify the significance of the of the sample correlation between total square feet and total energy consumption, a z-test was performed using a Fisher transformation with the following experiment design:

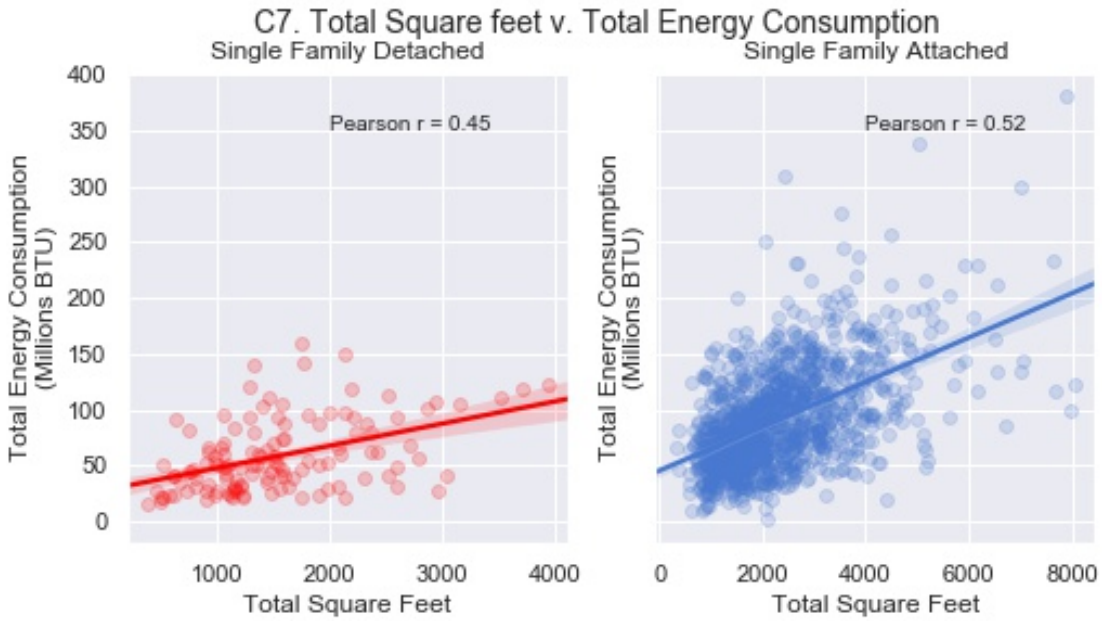
$$H_o: r = 0$$

$$H_a: r \neq 0$$

$$\alpha = 0.05$$

The result of the experiment yielded a p-value of near 0, thus we reject the null hypothesis that there is no correlation between total square feet of a household and total energy consumption in favor of the alternative hypothesis. This suggests that there is positive correlation between total square feet of a household and total energy consumption among all households. A 95% confidence interval for the true correlation is yields a range of [0.62 - 0.65].

Now that the correlation between total square feet and total energy consumption is understood among all households, it will be interesting to observe how the correlation varies across how households. Figure C7 below shows regressions plots of total square feet and total energy consumption for Single Family Detached households and Single Family Attached household to illustrate the difference the difference in correlation.



It is important to note that there are significantly more Single Family Attached households (3752) than Single Family Attached households (479) in the dataset. The correlation for both groups is weaker than the correlation for the entire sample. The correlation for Single Family Attached

households is stronger than Single Family Attached homes. To test the significance of this difference in correlation among the two groups, a z-test was performed using a Fisher transformation with the following experiment design:

$$H_o: r_a = r_d$$
$$H_a: r_a > r_d$$
$$\alpha = 0.05$$

The p-value of the above z-test was 0.027 which is less than the significance level, thus we reject the null hypothesis that the correlation between total square feet of a household and total energy consumption for Single Family Detached Homes and Single Family Attached Homes is equal in favor of the alternative hypothesis. This suggests that the correlation for Single Family Detached homes is greater than that of Single Family Attached Homes.

In conclusion, the analysis detailed above will assist in selecting features that will be used in the predictive model. Having confidence in the relationship between the size of a household and the energy it will consume will drive intuition for selecting variables that are related to household size. Incorporating Additional household characteristics should help better predict energy consumption and account for the difference between household types.

Results & In-Depth Analysis

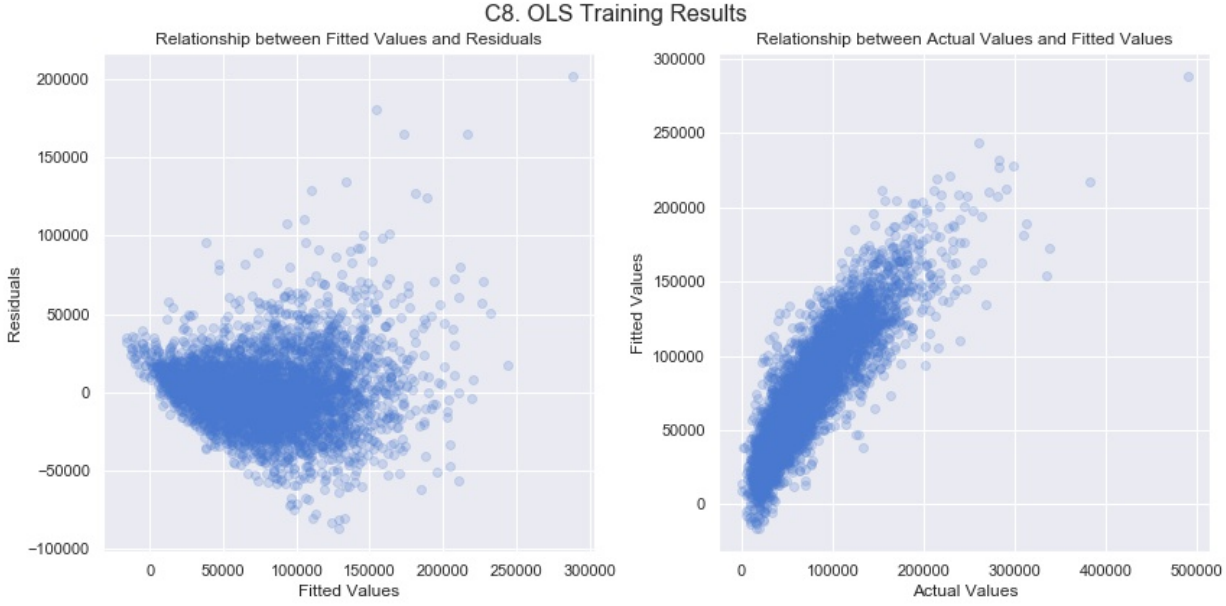
Data Processing

Before experimenting with predictive models, the dataset was processed by scaling features, encoding categorical and ordinal features, and removing low variance Boolean features. Continuous features were scaled to have a mean of 0 and standard deviation of 1. Categorical features were obtained using one hot encoding and removing a one column for each category to avoid introducing collinearity. Boolean features that contained the same value in more that 80% of samples were removed from the data set. The resulting dataset contains 5,686 samples and 660 features.

Linear Models

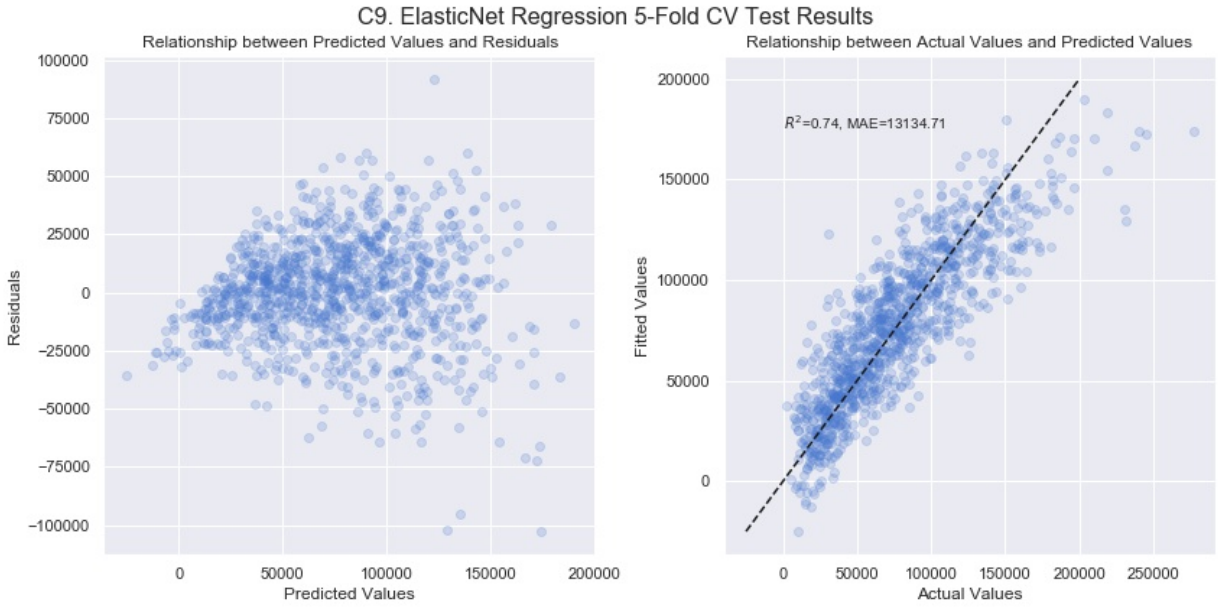
The first class of predictive models that were chosen to experiment with were linear regression models, starting with Ordinary Least Squares Regression. This initial model (and all following models) was trained using 80% of the available data, reserving 20% of the remaining data as a hold out test set. All linear models were fit through the origin as this makes the most practical sense given the features. This initial model yielded a $R^2 =$ 75.9% and a median absolute error (MAE) of 12,664 kBTU on the training set, and a $R^2 =$ 68.9% and a median absolute error (MAE) of 15,008 kBTU. Using this model, outliers were identified to improve prediction. As shown in figure C8, there are several points with high residuals that the model was not able to predict well. These outliers were removed as they represented households that have abnormally high energy consumption.

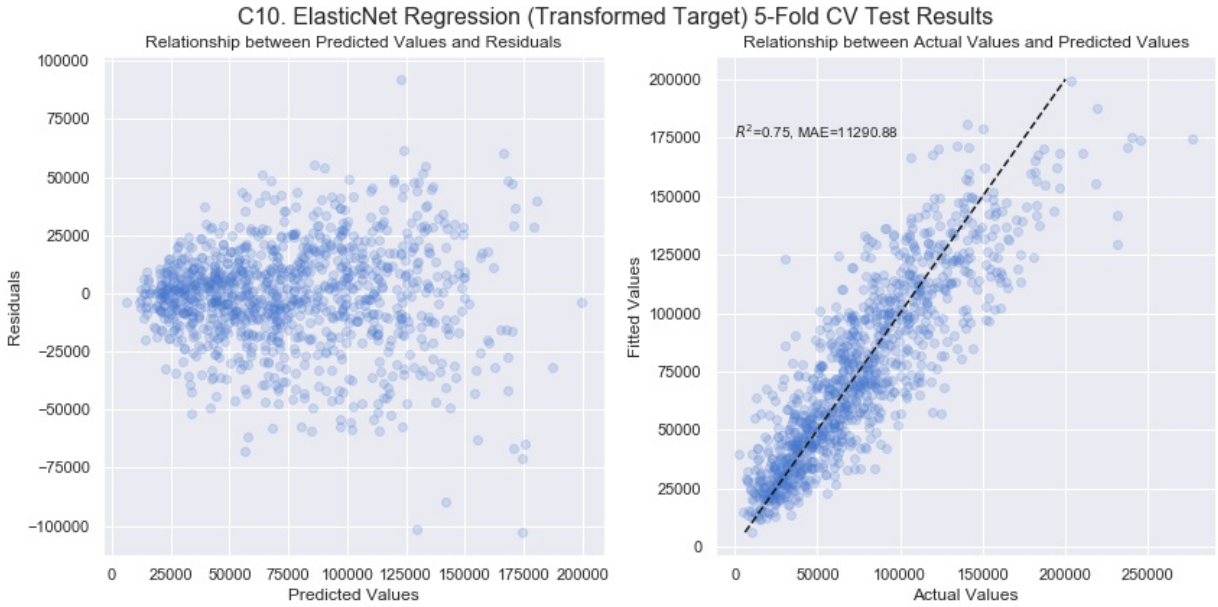
C8. OLS Training Results



After removing these outliers, an OLS model was trained as several other sci-kit learn linear models: Ridge Regression (L2 regularization), Lasso Regression (L1 regularization), and Elastic Net Regression (L1 and L2 regularization). Each model was trained using 5-Fold cross validation on the training set to determine the optimal hyperparameters. The table below shows the results of these models, each of which was able to improve on the performance of the original OLS model. The Elastic Net model performed the best on the test data and the results are shown in C9. As shown in the plot of the actual values vs fitted value, there still remain several marginal outliers and the model does not perform as well for the more extreme small target values and large target values. To improve the prediction of this model. The target space was transformed to a normal distribution to spread out the most frequent values and reduce the impact of outliers. This resulted in the best performing linear model with a R^2 = 75.4% and MAE of 11,291 on the testing data.

	Training Set		Test Set	
Model	**R^2**	**MAE (kBTU)**	**R^2**	**MAE (kBTU)**
Ridge	80.3%	11,706	73.4%	13,360
Lasso	80.3%	11,733	73.4%	13,334
Elastic Net	79.4%	11,899	73.7%	13,134
Elastic Net (w/ Transformed Target)	80.3%	10,608	75.4%	11,291





ElasticNet Regression (Transformed Target) Results by Housing Type		
Housing Type	**Median Consumption (kBTU)**	**MAE (kBTU)**
Mobile Home	57,398	11,291
Single Family Detached	85,531	13,011
Single Family Attached	55,208	10,315
Apartment Building w/ 2-4 units	42,284	7,891
Apartment Building w/ 5+ Units	30,291	5,973

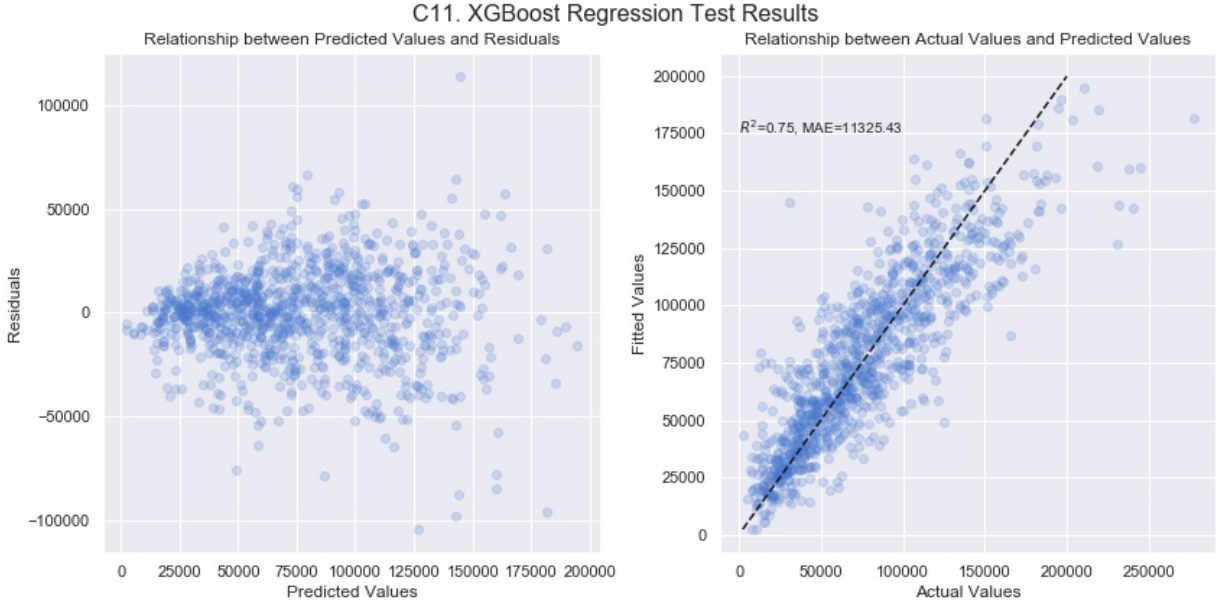
As shown in the table above, the MAE of the predictions of the elastic net (transformed target) model vary substantially across housing types and the error has different practical interpretations. For each household type the model is able to give a reasonable estimate for the total energy consumption.

The model performs best on Single Family Detached households which makes sense given the amount of data points of this type present in the dataset relative to the others. Given the high variability that is to be expected in a problem of this nature, this model can be useful and performs better than a prediction based on the median or mean of energy consumption of all households.

Regression Trees Boosting Model

Ensemble decision tree models such as random forest models, gradient boosting models and Xtreme gradient boosting models were considered and experimented with in order to obtain a model worth more predictive power. Of these models, the Xtreme gradient boosting had the best performance and the results of the model are shown in C11. This model was trained using a combination of randomized parameter selection and grid searching to tune the model using 5-fold cross validation.

C11. XGBoost Regression Test Results



Conclusion

The elastic net model with the transformed target will be used chosen as the best model for the goal of predicting household energy consumption. This model performed relatively well on the test data which gave confidence in its ability to generalize and predict energy consumption given data on future households. MAE of 11,291 kBTU is a substantial

This model was also chosen because of its interpretability which is very important in the use case of this model. This model will not only be used to predict how much energy a household will use annually, but also to give intuition behind what factors influence these predictions to give households a better understanding of how they can affect adjust their consumption. As shown in the table below shows the top 15 factors that have the highest influence on the model in order of absolute weight determined by the loss function.

Rank	Feature Description
1	Use of wood for space heating
2	Use of other fuel type* for space heating
3	Indicator for Apartment Building w/ 5+ Units
4	Use of propane for space heating
5	Categorical Indicator for studio apartment (Yes)
6	Indicator for household built between 2010 -2015
7	Indicator for electricity used for water heating
8	Indicator for water heater in apartment
9	Indicator for receiving assistance for energy bills
10	Use of electricity for space heating
11	Indicator for Energy Star refrigerator
12	Indicator for swimming pool at household
13	Categorical indicator for studio apartment (N)
14	Use of natural gas for water heating
15	Use of built-in electricity units for main space heating

This model is not a finished product and can be further improved further to improve predictive power. Particular areas that may be explored in the future to improve upon the models discussed above is more robust feature selection to eliminate noisy features, tuning more parameters for boosting models, further investigation into the correlation between variables and their effects on the results, and incorporating more model classes that may perform better.