Capstone Project 1

Predicting Household Energy Consumption

Project Proposal

HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

- The sociopolitical effects of climate change are shifting the focus of the energy sector. Utility companies are facing pressure to be more efficient and environmentally sustainable
- How utility companies communicate with their customers about their energy consumption 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

Utility Company and Customer Benefits

HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

Utility Company Benefits

- 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.

Customer Benefits

- Analyze how various household decisions can affect their energy use, thus providing a means to minimize their consumption and save money
- Compare their actual consumption to predicted consumption in order to gauge how energy efficient their household is.

Data Wrangling

Data Source

HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

- 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 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

Data Cleaning

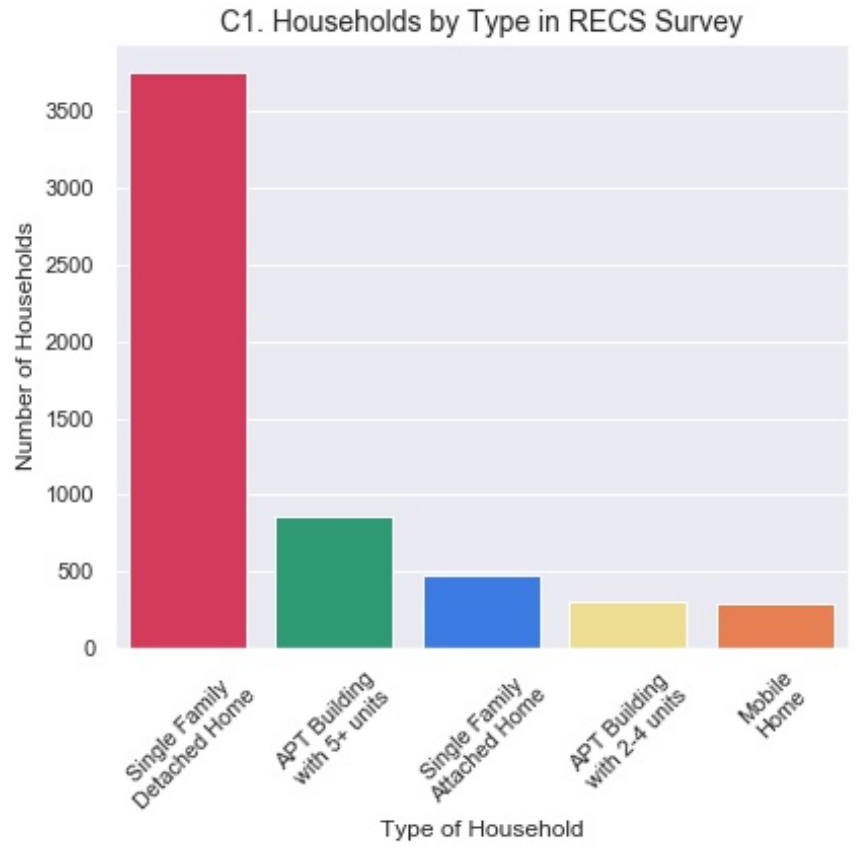
HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

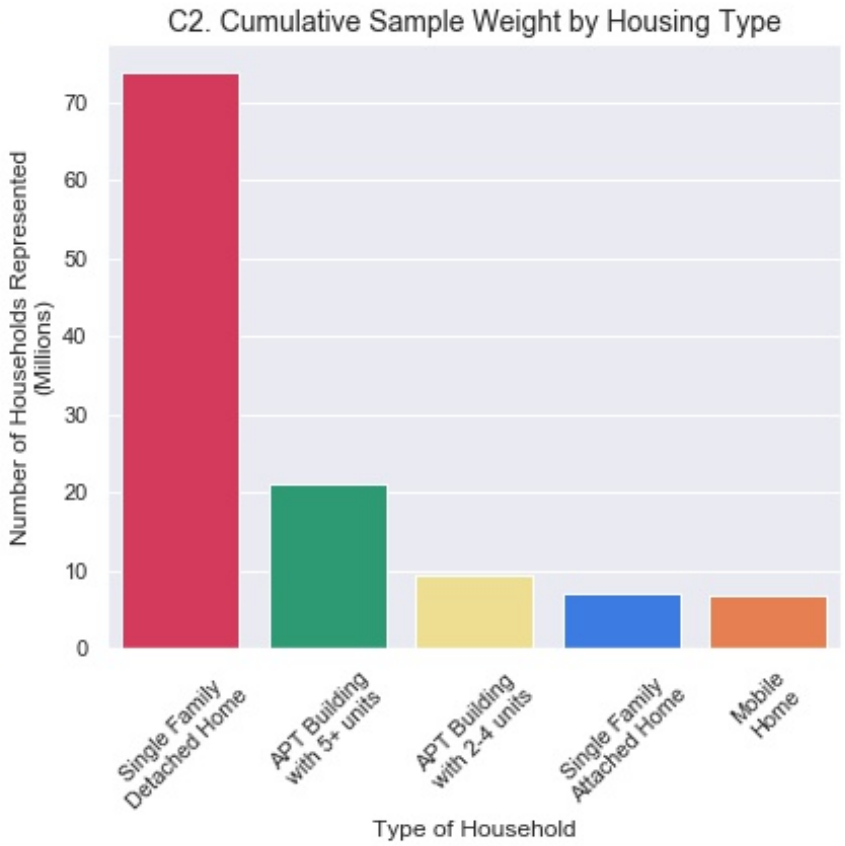
- The raw dataset is relatively clean and tidy. Each row in the main dataset represents a respondent in the survey and each column corresponds to distinct survey questions or parameters of the survey construction
- The survey was designed to be statistically representative of all US households. Each observation has an associated weight that corresponds to the number of households the observation represents
- The dataset contains 217 columns that represent imputation flags for 222 columns variables (some imputation flag columns correspond to multiple variables).
- There were 20 questions in the survey that have “Don’t Know” as a possible answer. These values were also considered to be missing
- 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

Exploratory
Data Analysis

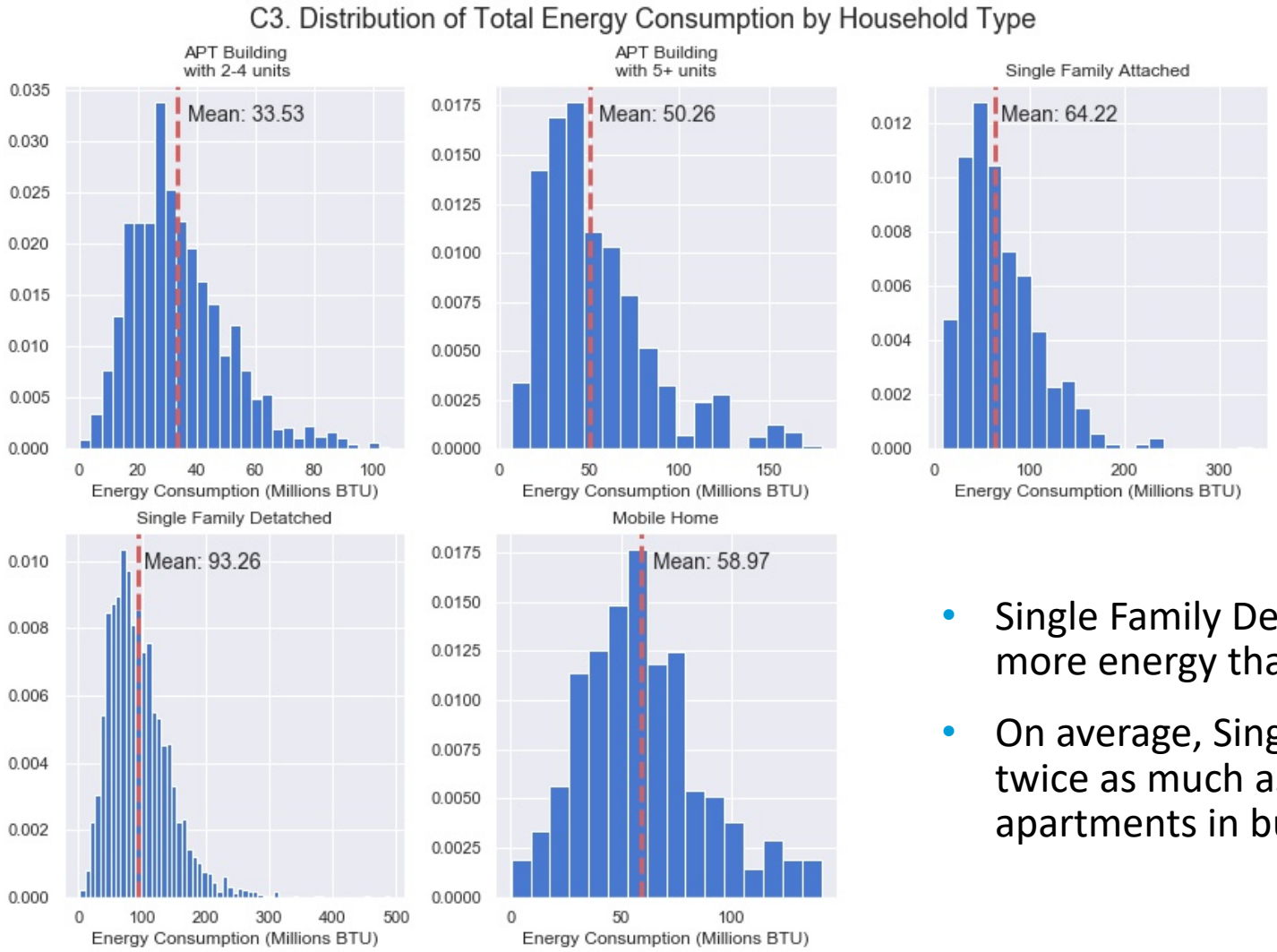
Exploratory Data Analysis (1 of 4)

- 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.
- The 3,752 observations of Single Family Detached households in the dataset are statistically representative of 73.8 million American households.



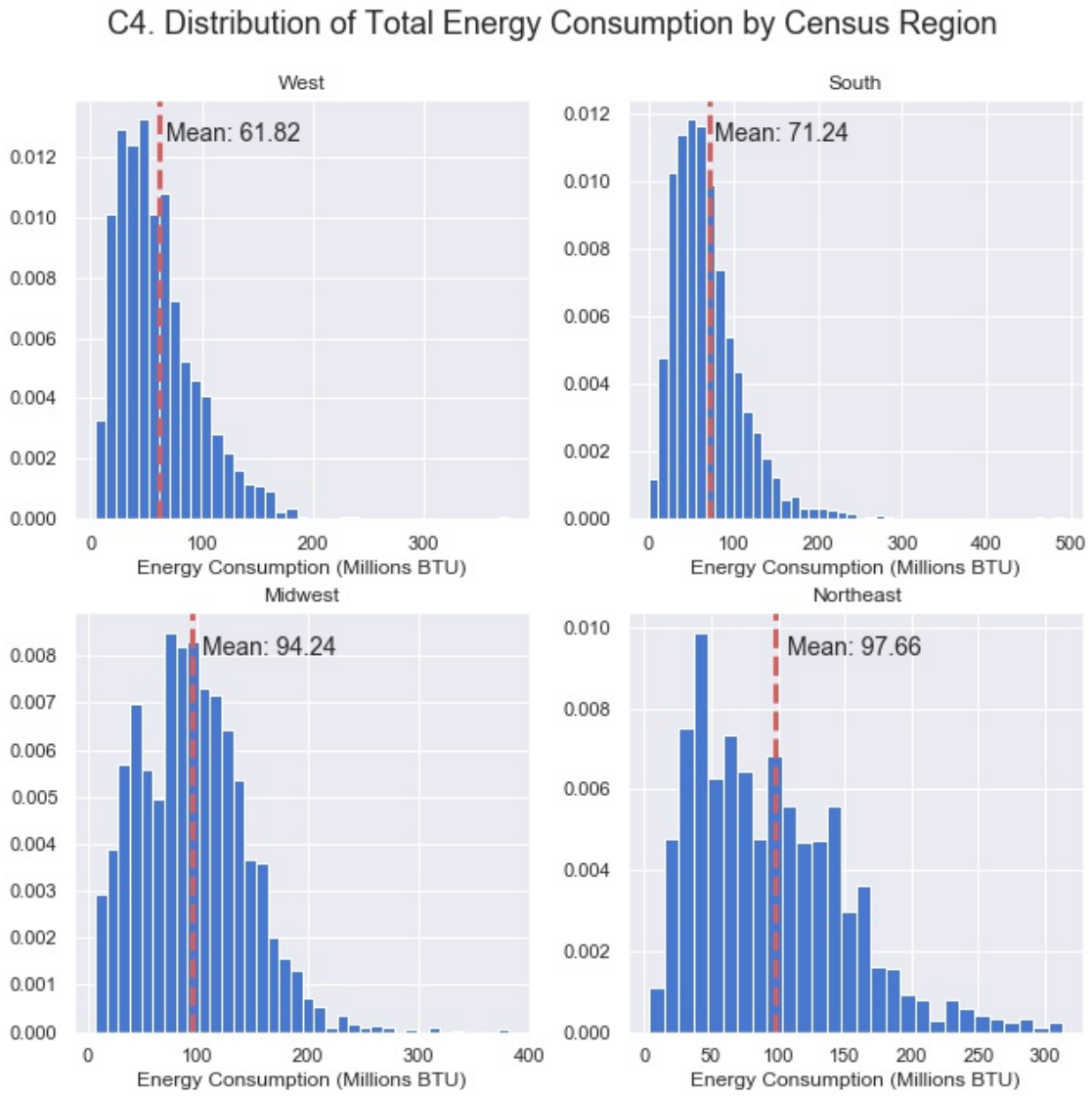


Exploratory Data Analysis (2 of 4)



- Single Family Detached households consume significantly more energy than other household types.
- 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.

Exploratory Data Analysis (3 of 4)



- 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

Results

Results (1 of 4)

Several classes of models were experimented with including linear regression models, random forest models, and ensemble models. Linear models gave yielded the best combination of speed, performance, and interpretability

	Training Set		Test Set	
Model	**R^2**	**MAE (kBTU)**	**R^2**	**MAE (kBTU)**
Baseline	-3.9%	29,626	-4.4%	29,050
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

The ElasticNet Regression (w/ Transformed Target) model was selected has the best model for this project due as it performance on testing data which increases confidence in predictions made on future data. This model also allows for great interpretability by identifying the most influential variables

*Target variable was normalized to decrease influence of outliers

Results (2 of 4)

For each household type the model is able to give a reasonable estimate for the total energy consumption. This model can be useful and performs better a prediction based on the median or mean of energy consumption of all households, while accounting for many different factors

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

The model performs best on Single Family Detached homes which makes up the 66% of the households in the US

Results (3 of 4)

One of benefits of the type of model chosen is the ability to analyze the variables that are influence predictions to gain insight into what the most important factors are in addition to the ability to use these results to verify/improve the model in the future. The model found the following features to be most important:

Top 15 Most Influential Features	
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

*Fuel types other than natural gas, propane, fuel oil/kerosene, electricity, or wood

Conclusion

Conclusions

HOUSEHOLD ENERGY CONSUMPTION PREDICTION TOOL

- The model is useful for predicting total energy consumption for various types of households and can serve as a tool for customers and integrated into existing websites/software applications of utility companies
- This model has much greater accuracy than less sophisticated methods such as using the mean or median to predict consumption, thus making it practically and statistically significant
- This model is not a finished product and can be further improved further to improve predictive power.