Predicting Energy Consumption

Name: Sudhandar Balakrishnan
email: 21sb23@queensu.ca or sudhandar@gmail.com

About the data

- Residential Energy Consumption Survey (RECS) is a national sample survey that collects energy-related data for housing units occupied as primary residences and the households that live in them.
- The RECS 2009 data was collected from 12,083 households with 938 features and our goal is to predict the energy consumption in Kilowatt-hour.

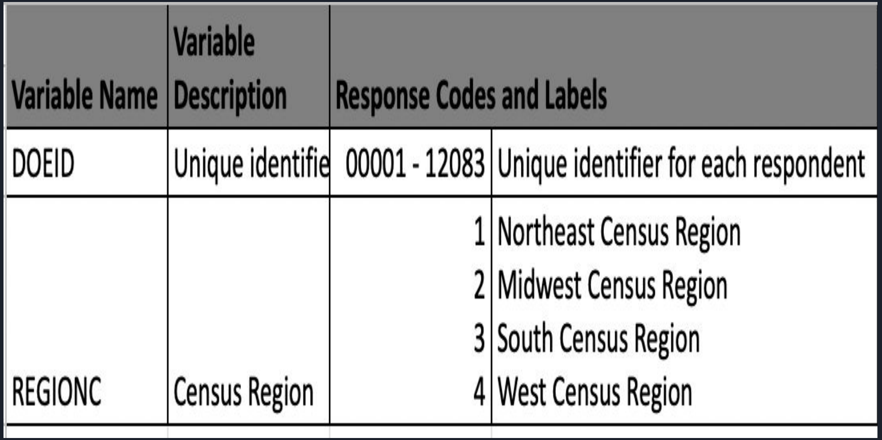
Business Problem:

Create a model to forecast energy usage based on numerous household details and attributes.

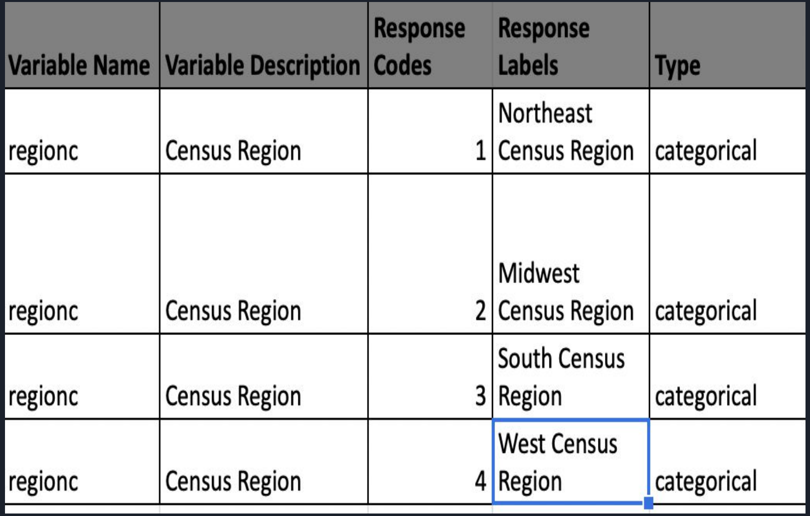
Approach:

1. Preprocess the csv files and store them in an SQLite database.
2. Fetch the preprocessed data and analyze it to understand the data and make relevant assumptions.
3. Clean the data based on the assumptions.
4. Analyze the relationship between the features and understand its characteristics to select the correct model type.
5. Model selection and training.
6. Hyperparameter tuning to improve the performance of the model.

Preprocessing:



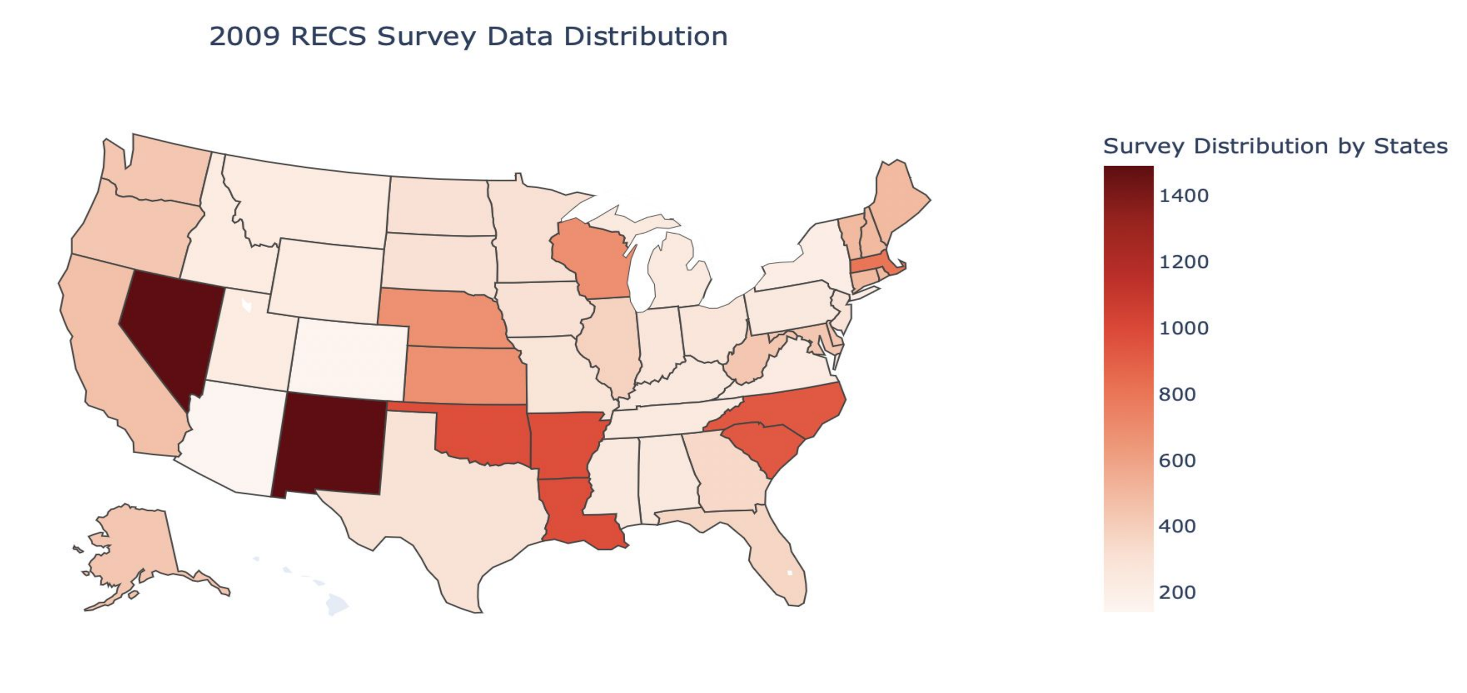
Variable Name	Variable Description	Response Codes and Labels	
DOEID	Unique identifie	00001 - 12083	Unique identifier for each respondent
REGIONC	Census Region	1	Northeast Census Region
		2	Midwest Census Region
		3	South Census Region
		4	West Census Region



Variable Name	Variable Description	Response Codes	Response Labels	Type
regionc	Census Region	1	Northeast Census Region	categorical
regionc	Census Region	2	Midwest Census Region	categorical
regionc	Census Region	3	South Census Region	categorical
regionc	Census Region	4	West Census Region	categorical

Data Exploration:

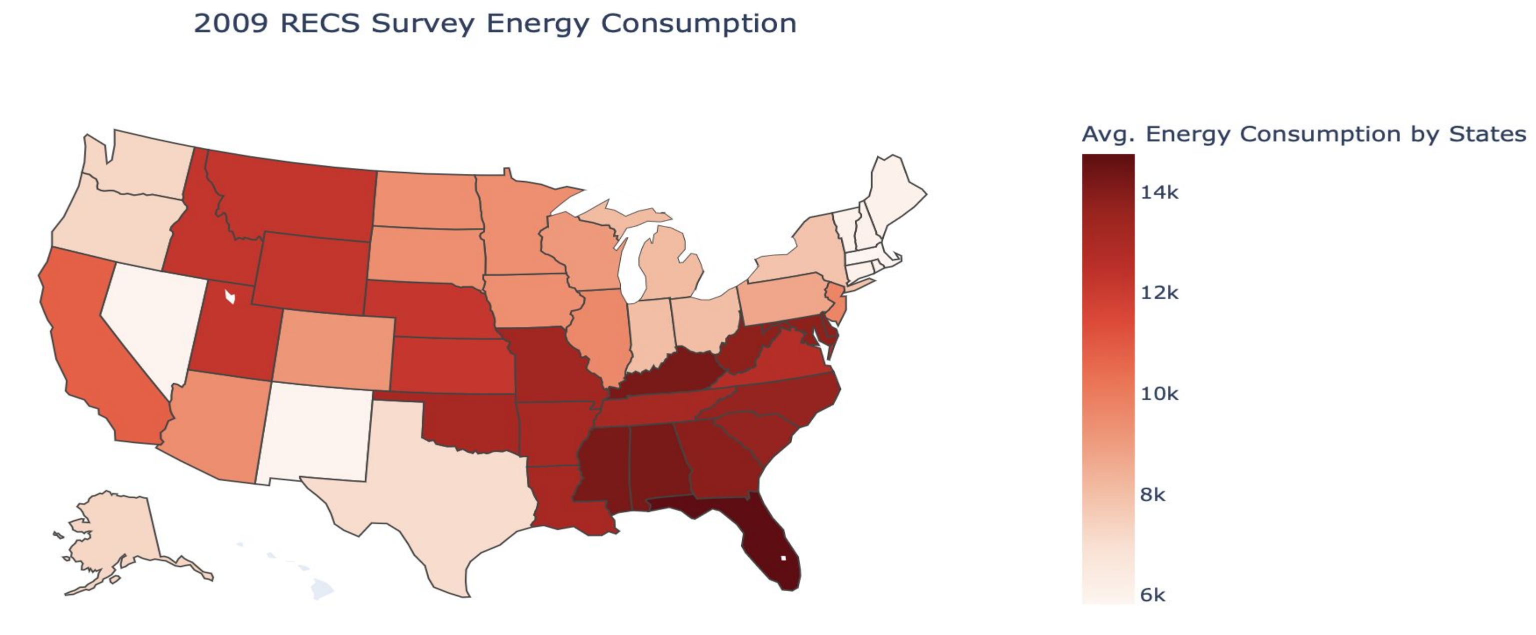
Nevada and New Mexico (combined) have the most number of survey records. They are in the southern part of the country and tend to have warmer weather.



2009 RECS Survey Data Distribution
Survey Distribution by States
1400
1200
1000
800
600
400
200

Data Exploration:

The states with the most records consume lesser energy on average. Florida with just 364 records consumes higher energy than other states. Keep in mind, that Florida is a state with warm weather.


2009 RECS Survey Energy Consumption
Avg. Energy Consumption by States
14k
12k
10k
8k
6k

Assumptions about the data:

- In general, states and nations with colder climates use more electricity than those with hotter climates. Different factors affect the amount of power consumed in hotter and colder states. For instance, in areas with higher temperatures, especially in the summer, air conditioners use a lot of electricity. In colder states, especially in the winter, house heaters use a lot of electricity.
- There is a possible chance that the survey might be collected during the summer time.
- There might be a lack of uniformity in the size of the houses from which the data was collected.

Data Cleaning:

- Numerical columns with more than 50% null values were removed.
- Impute the null values in numerical columns with the median..
- Categorical columns with more than 50% null values were removed.
- Impute the null values in categorical columns with the mode.
- Remove the imputation flags.
- Remove the columns which contain kwh data in British Thermal Units (BTU).
- Remove the individual energy consumption features (kwhoth, kwhcol, etc.)
- Remove Outliers using IQR.

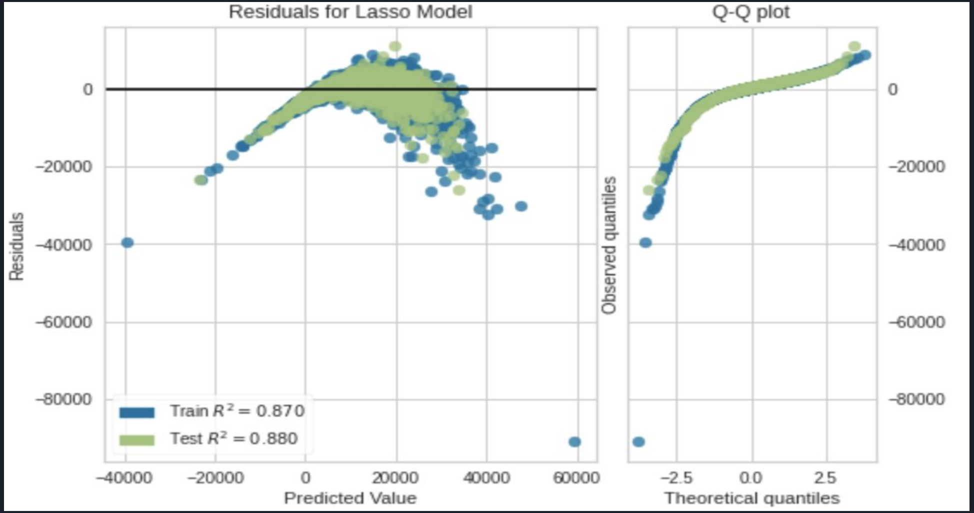
Before Cleaning: 12,082 records, 940 features

After Cleaning: 10769 records, 233 features

Reasons for rejecting a linear model:

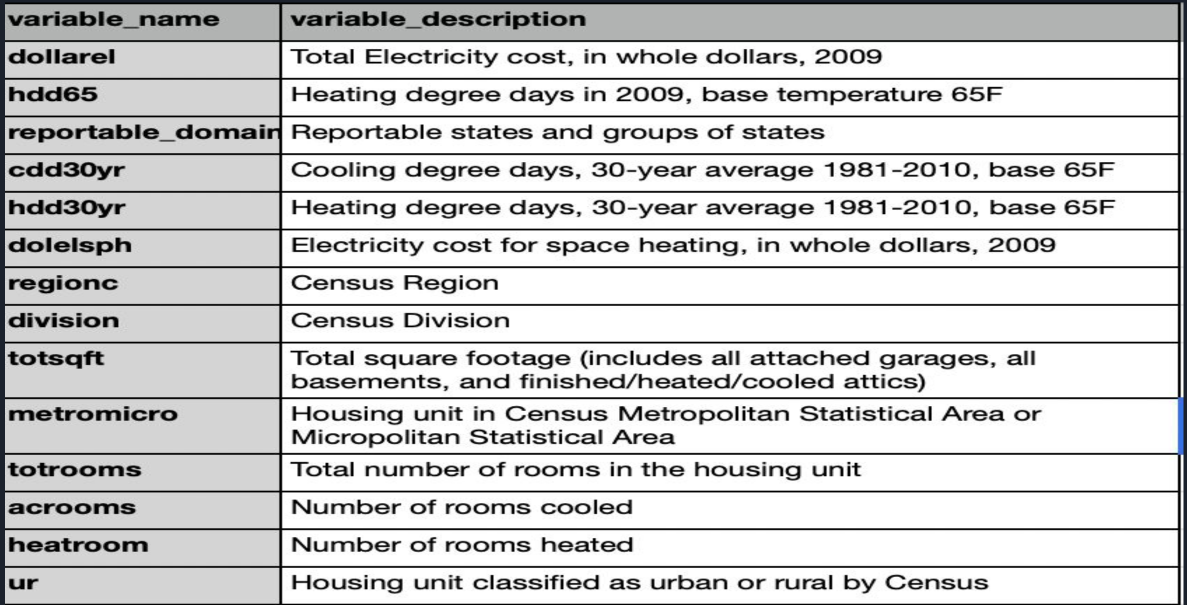
The data fails the following assumptions of a linear regression model,

1. Independence - All the features are dependent on each (multicollinearity). This was verified using the Variance Inflation Factor (VIF)
2. The data is not linear - This was verified by plotting the residuals of a lasso regression model.



Feature Selection:

- The Random Forest Regressor has been used to select the most important features.
- Furthermore, some features were manually added to the feature list in order to give the model access to more details regarding the various elements and characteristics of the home.
- The following features were selected for the model,



variable_name	variable_description
dollarel	Total Electricity cost, in whole dollars, 2009
hdd65	Heating degree days in 2009, base temperature 65F
reportable_domain	Reportable states and groups of states
cdd30yr	Cooling degree days, 30-year average 1981-2010, base 65F
hdd30yr	Heating degree days, 30-year average 1981-2010, base 65F
dolelsph	Electricity cost for space heating, in whole dollars, 2009
regionc	Census Region
division	Census Division
totsqft	Total square footage (includes all attached garages, all basements, and finished/heated/cooled attics)
metromicro	Housing unit in Census Metropolitan Statistical Area or Micropolitan Statistical Area
totrooms	Total number of rooms in the housing unit
acrooms	Number of rooms cooled
heatroom	Number of rooms heated
ur	Housing unit classified as urban or rural by Census

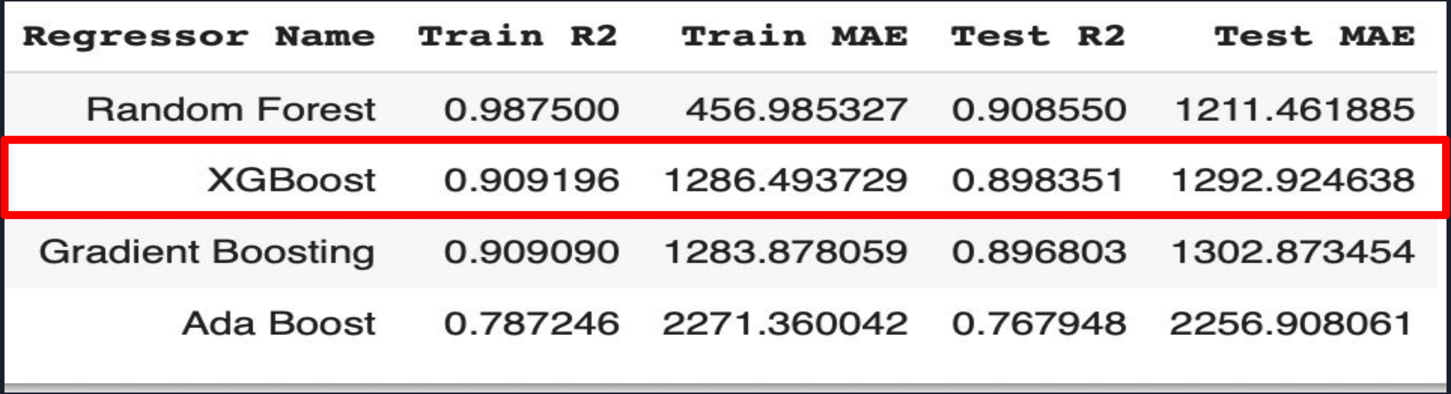
Model Selection

Number of training records: 8615
Number of test records: 2154

Error Metrics:

1. R^2 error - Helps in measuring the goodness of fit and comparing the performance of models.
2. Mean Absolute Error - Helps in finding out the value by which the predictions are wrong.

Performance Comparison:



Regressor Name	Train R2	Train MAE	Test R2	Test MAE
Random Forest	0.987500	456.985327	0.908550	1211.461885
XGBoost	0.909196	1286.493729	0.898351	1292.924638
Gradient Boosting	0.909090	1283.878059	0.896803	1302.873454
Ada Boost	0.787246	2271.360042	0.767948	2256.908061

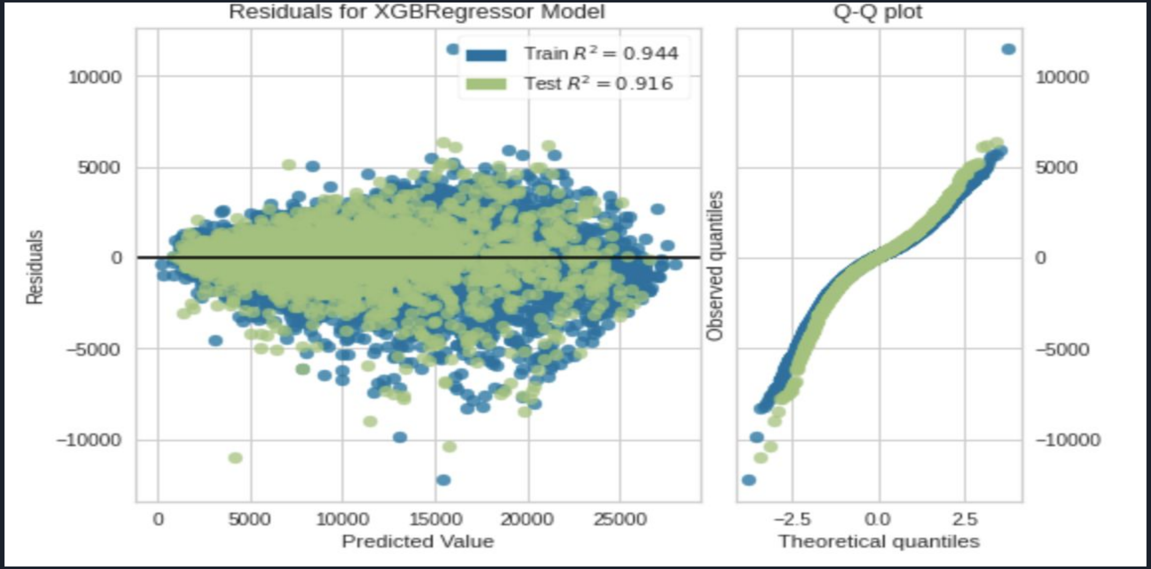
Final Results:

Best Model : XGBoost Regressor

Results:

Train R2: 0.9438   ;   Train MAE: 1012.6

Test R2: 0.9160   ;   Test MAE: 1169.13



Performance Evaluation:

1. The model's performance has been enhanced as a result of fine-tuning the hyperparameters.
2. The test set's mean absolute error is 1169.13 kwh. The predictions made by our model could be off by 1169.13 kwh from the real value.
3. The data shown in the residuals plot contains a few outliers. This has occurred even after removing the outliers using IQR. To address this problem, the dataset has to be further investigated.
4. This model is not perfect, and by experimenting with other feature selectors, fine-tuning the hyperparameters, and employing better models, the performance can still be enhanced.
5. The household's location and the heating degree days are crucial factors in determining energy use.

Future Directions:

1. For the next iteration, instead of banking on the feature selection methods, we can take into account the individual components of the house like the appliances used, size of the house and number of members in the household along with the location. We can build a model using these features and see if these features can help in building a better model.

2. We can incorporate one or two of the individual electricity consumption components, such as kwhcol (energy consumption of air conditioners) and kwhoth (energy consumption of other components) and see if it's adding value to the model.

3. Ignore the features which have a non-linear relationship with the target variable and consider features which have a linear relationship with the target variable. This can be verified by using scatter plots with the features on the x-axis and the target variable on the y-axis.