Predicting Energy Consumption Cost in Residential Units

1. Problem statement

What is the estimated cost of energy in a residential unit, given the attributes available on any listing services website?

US total residential energy consumption in 2019 was 3,487.55 Terawatts-hour[1] representing a $460B ticket for homeowners and rental property investors. This economic burden is difficult to manage since it is usually tied to construction features or building systems that are costly to replace or renew, and house energy consumption expenses are usually known only after the first year of making the purchase decision.

Currently, buyers often find difficulties to estimate the potential real cost of energy bill when scouting the market for a new house or condo. Most listing services and realtors do not have the information or tools to predict bill costs, and it can weigh heavily on the final buying decision.

Also, understanding the factors that determine the energy bill before-hand could lead to higher competitiveness in the real estate market, encouraging energy efficient innovation and practices, and eventually benefiting the final buyer or investor.

Furthermore, knowing the building factors that weighs on the energy consumption costs can be used to calculate accurately the return on investment of systems updates or rehabilitation projects.

2. Objective

The present project aims to provide homebuyers, homeowners, and retail rental property investors a powerful tool to estimate the yearly energy bill, just by providing a few parameters shown in a typical listing services website. The target unit could be a single-family house (attached or detached), a townhome, a single unit in a condominium, and a mobile home.

The intended consumption cost prediction will be created out of variables that can be extracted from any common web listing service, such as RedFin, Realtor.com, or Zillow. Therefore, we aim to generate an accurate prediction out of widely available information of the unit even if the user is not the owner.

This tool is intended to be used by non-professional homebuyers, homeowners, and rental property investors to allow them to make informed and data-driven decisions.

[1] (The U.S. Energy Information Administration (EIA) , 2020)

A secondary objective of this projects is to investigate which building features have the largest impact on energy consumption, so any homeowner could design the home remodel project to yield the best results in terms of energy savings.

3. Methodology

The dataset

We will use the 2015 Residential Energy Consumption Survey data (RECS), from The U.S. Energy Information Administration (EIA). This dataset represents a national sample survey that collects energy-related data for housing units occupied as a primary residence and the households that live in them. Data have been collected from more than 5,600 households selected at random, representing 118.2 million U.S. households[2].

The RECS dataset contains 5,686 observations and 759 variables grouped in seven broader categories:

1. Building attributes
2. Building system features
3. Household uses and habits
4. Social and demographic characteristics
5. Energy consumption features
6. Economic features
7. Climate data

We further selected four target variables from the complete dataset to measure the units' energy consumption:

- Cooling cost: Annual dollar cost of the energy used for space cooling regardless of the fuel employed
- Heating cost: Annual dollar cost of the energy used for space heating regardless of the fuel employed
- Water Heating cost: Annual dollar cost of the energy used for heating water for domestic use regardless of the fuel employed.
- Other electricity cost: Annual dollar cost of general electricity usage excluding all three above categories.
- Total costs: Additionally, we included the prediction of the sum of all above target features.

We also selected some explanatory variables from the 2015 RECS dataset, based on two criteria:

1. Relevance and explanatory potential for predicting the target variables.
2. Availability of information on common real estate listing websites.

[2] Residential Energy Consumption Survey (Recs). The U.S. Energy Information Administration, (https://www.eia.gov/consumption/residential/data/2015/index.php?view=microdata). Retrieved on 11/18/2020.

For instance, climate quantitative data is extremely relevant since their values could clearly explain the variability on consumption cost, however those records for a certain unit are not easily available so the variable was discarded. Alternatively, we chose to include the general climate classification area because it can be easily inferred from the zip code of the unit.

The short-list of the explanatory variables is presented in the table below:

Explanatory Variable Name	Variable Description
ATTIC	Attic above the housing unit
BEDROOMS	Number of bedrooms
CELLAR	Housing unit over a basement
COOLTYPE	Type of air conditioning equipment used
EQUIPAGE	Age of main space heating equipment
EQUIPM	Main space heating equipment type
HIGHCEIL	High ceilings
CLIMATE_REGION_PUB	"Building America" Climate Zone
KOWNRENT	Own or rent
NCOMBATH	Number of full bathrooms
NHAFBATH	Number of half bathrooms
PROTHERM	Programmable main thermostat
SIZEOFGARAGE	Size of attached garage
STORIES	Number of stories in a single-family home
TOTSQFT_EN	Total square footage (used for publication)
TYPEHUQ	Type of housing unit
YEARMADERANGE	Range when housing unit was built

Most of the explanatory variables contain categorical values, except for *Total Square Footage which is numerical;* whereas all target variables contain numerical continuous values.

Preprocessing and modelling

Following an exploratory analysis of the variables, some pre-processing steps were taken to prepare the data for the modeling:

- Feature transformation:
  - Regrouping categorical variables into larger groups and create binary dummy variables for each category.
  - Log transformation of numerical variables to rescale and center the distribution.

- Feature selection:
  - Removal of colinear variables.
  - Dimensionality reduction with a Recursive Feature Elimination algorithm.

Finally, some out-the-box regression algorithms were tested to select the best performer and further adjust its hyperparameters using Grid Search and Cross Validation techniques. The algorithms tested included:

- Linear Regression (OLS)
- Linear Regression with L2 regularization - Ridge
- Linear Regression with L1 regularization - Lasso
- Bagging ensemble method - Random Forest Regressor
- Boosting ensemble method - Gradient Boosting Regressor

The metrics used to assess the model's performance were explained variance and mean absolute error since the final objective is to provide an accurate consumption prediction with a reasonable margin of error.

Gradient Boosting Regressor algorithm was finally selected for the final model, which performed better across all target variables for a thin margin, as shown on the Figure 1 below.

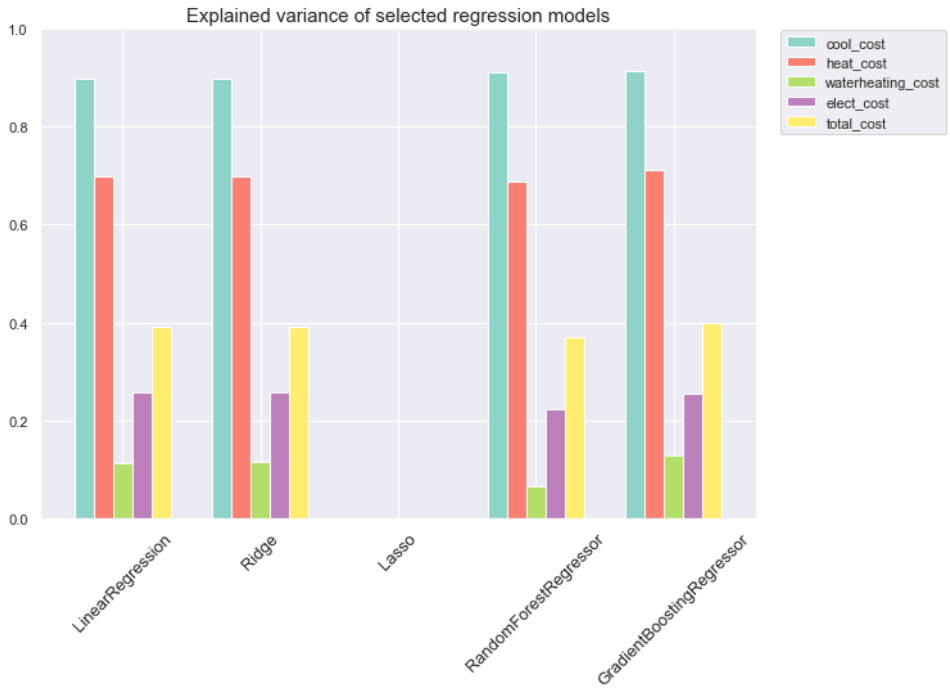

Figure 1 - Explained variance of regression models

Besides the predicted consumption, we created an additional model based on same Gradient Boosting Regressor algorithm to predict the upper prediction interval, so we could provide the maximum consumption bill for the customer with 80% likelihood.

4. Analysis

Continuous variables

An initial data analysis showed that all continuous numerical variables have a left skewed distribution with a long right tail. Figures 2 and 3 represents a selection of these variables and show that the values are mostly distributed around a peak and there are several one-sided outliers.

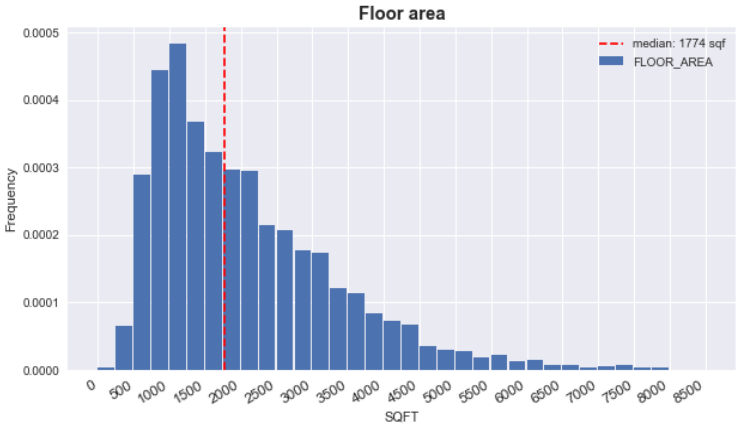

Figure 2 - Total Square Footage

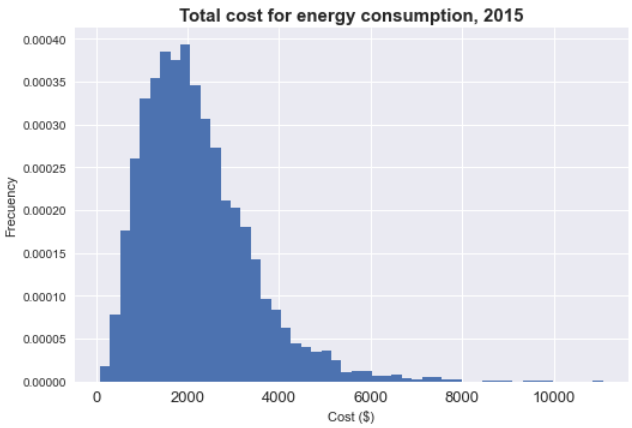

Figure 3 - Total energy costs

The correlation between *Total Square Footage* size and target variables is positive and moderate, except for *Water Heating Costs*. Figures 4 and 5 suggest that *Total Square Footage* is somewhat related to *Total Consumption Costs* but there might be other potential confounding factors to consider. It is worth noting that *Water Heating Costs* is not related to the unit's Floor Area.

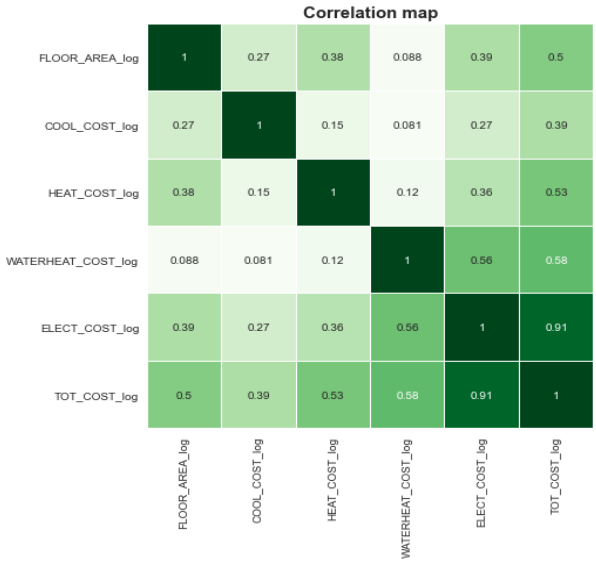

Figure 4 - Correlation map of numerical variables

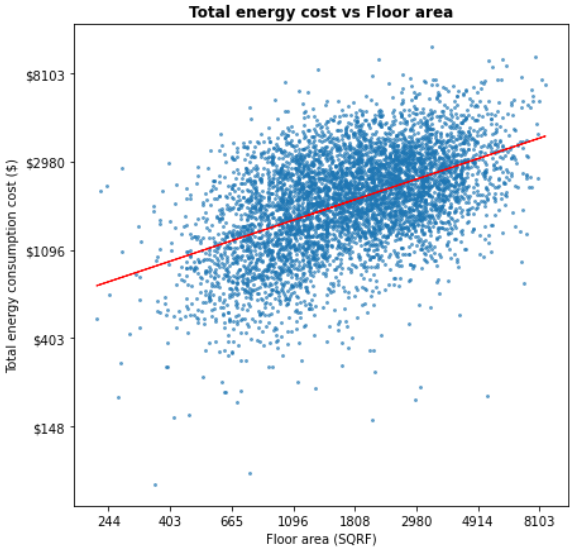

Figure 5 - Scatterplot of Total Energy costs vs Floor Area

Categorical variables

Among the most relevant categorical features in the dataset we can observe some correlation with the target variables:

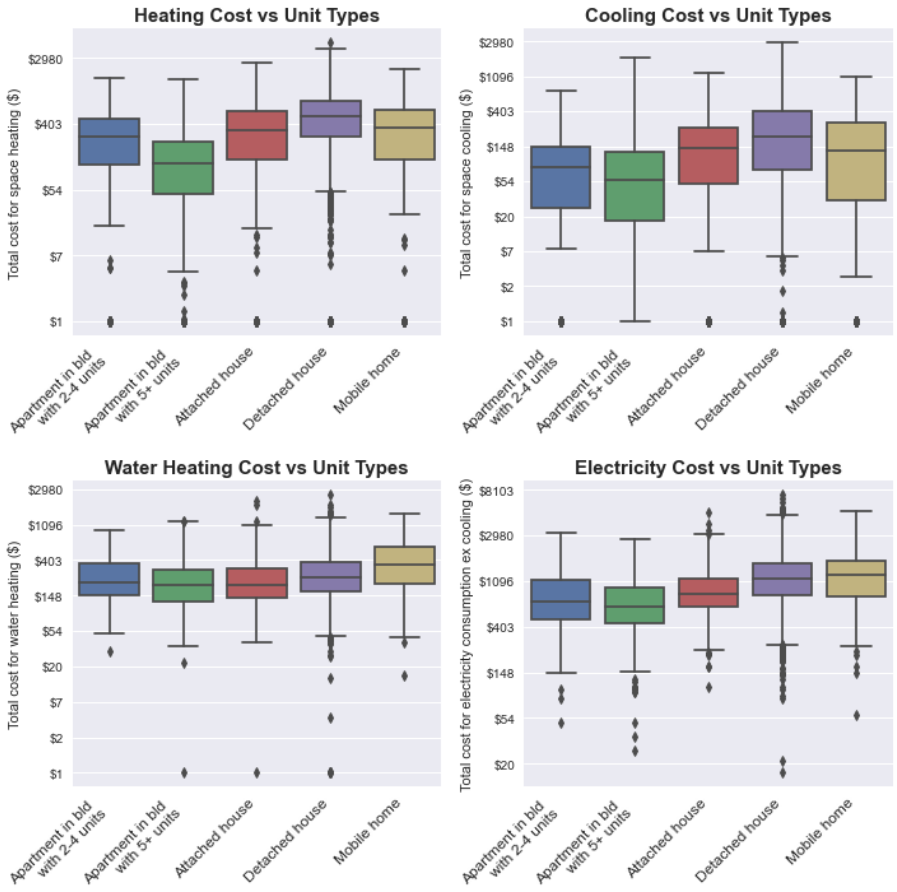

Figure 6 - Energy costs vs unit types

Figure 6 shows that there are clear differences in consumption costs among different unit types. Single-family detached houses have the highest average energy consumption costs across all energy end-uses, whereas apartments in high-rise condominiums account for the lowest consumption cost. However, the overlapping whiskers mean that the correlation is not strong, and the variance cannot be explained only by this feature.

Surely, there are other confounding factors, such as the unit size or the fact that in rural/suburban areas where houses are more common than apartments the energy price is usually more expensive.

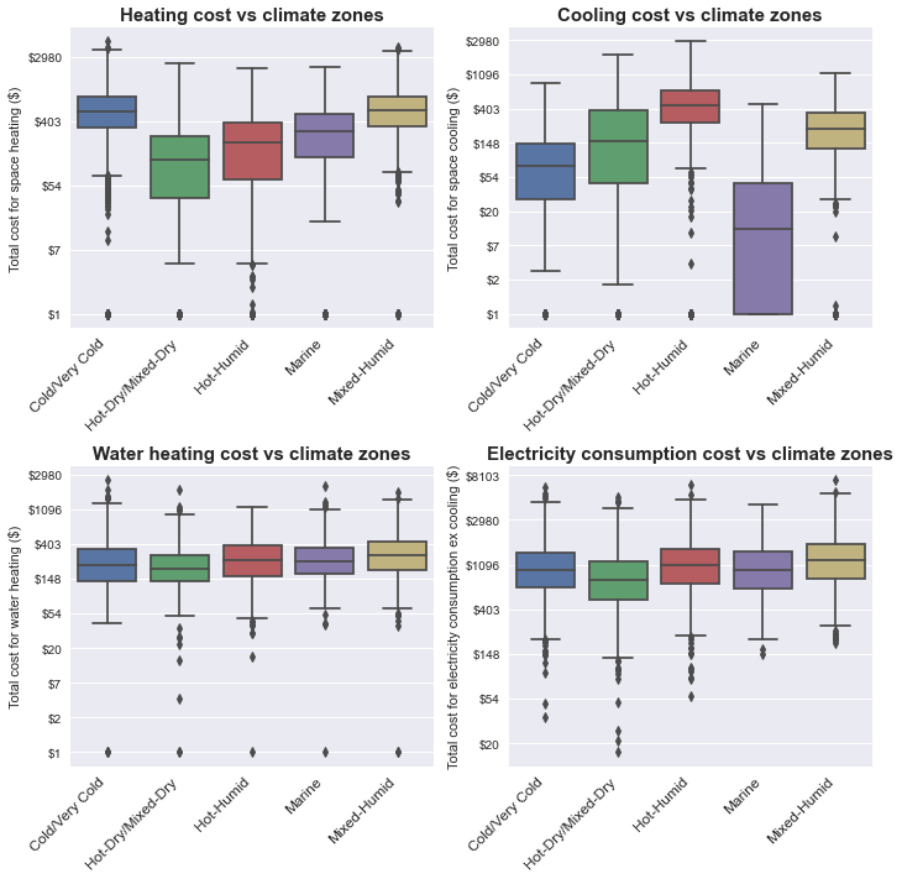

Figure 7 - Energy consumption costs vs Climate Region

As shown in Figure 7 above we can see that residential units located in mix-humid climate zone tend to have higher water heating and electricity consumption. It is also worth noting that heating consumption costs for the units in this climate is also as high as the units located in Cold or Very Cold climates.

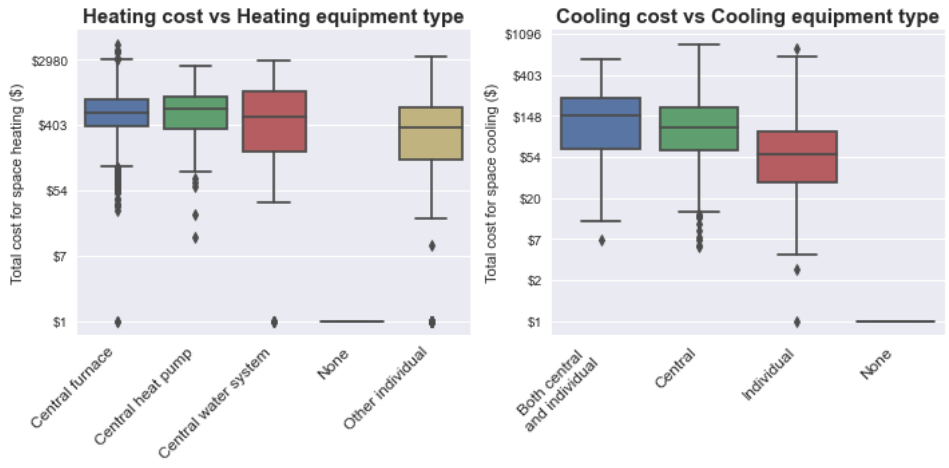

Figure 8 - Heating and cooling costs vs equipment type

Figure 8 shows heating and cooling consumption costs by equipment type. The data has been adjusted by climate zones to remove the climate confounding factor.

We can observe that units with central heating systems tend to have generally higher consumption expenses than others with individual systems, yet the spread of the consumption expenses in the latter is larger and their values overlap among different categories, showing a weak correlation.

Cooling costs tend to be higher in units with central systems in contrast to units with individual systems. This could be explained because individual cooling systems have less energy demand and are installed only in the spaces needed, also smaller units (with lower consumption) tend to have fewer central systems installations. In any case, the correlation between these variables is also weak

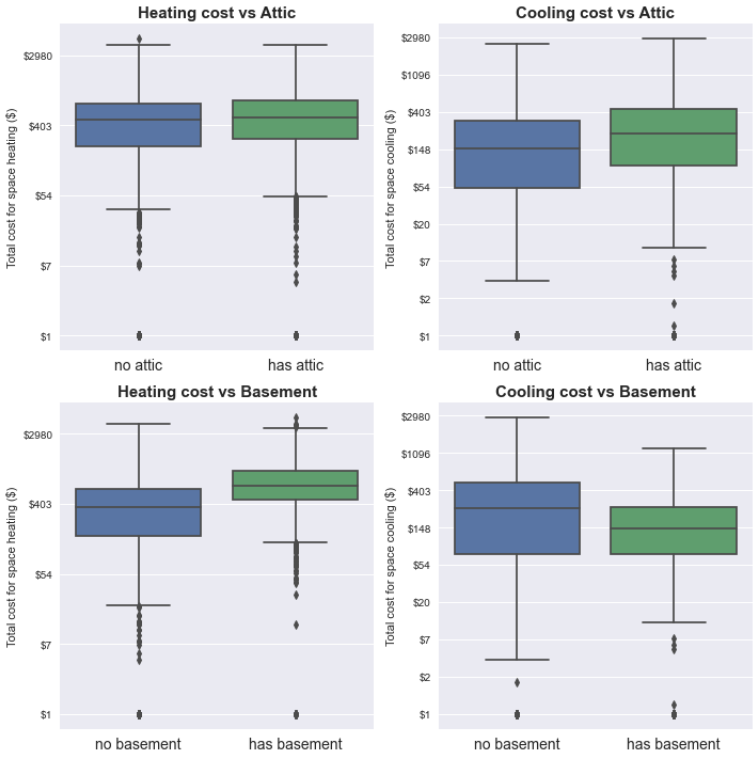

Figure 9 - Heating and cooling costs vs existence of attic and basement

Among construction and building features, the existence of attic and basement in the unit is one of the most relevant when it comes to predict heating and cooling costs.

Figure 9 shows that building an attic over the unit accounts for higher average cooling costs and no significant difference in heating cost. Existence of unit's basement accounts for higher average heating costs and lower cooling cost.

Only the data for detached and attached houses have been selected for this analysis to adjust for the confounding factor of the unit types.

5. Findings

The table below shows the performance metrics for our final five models based on the Gradient Boosting Algorithm, and some descriptive statistics of our target variables; thus, it is possible to compare the mean absolute error with the total value range and the median.

	Features	Explained variance	MAE	Min	Median	Max
0	cooling cost	0.915	99.68	0.00	146.60	2,860.01
1	heating cost	0.708	247.07	0.00	400.84	4,766.66
2	water heating cost	0.144	131.84	0.00	236.44	2,566.61
3	other electricity cost	0.253	434.26	15.43	1,021.74	7,310.58
4	total utility cost	0.396	719.65	76.40	1,997.59	11,078.48

Explained variance

It can be seen a large difference in model performance between target variables prediction. Although they are all related to energy consumption, there are significant differences in how the source of energy is used.

Space heating and cooling in residential units is achieved by an energy flow balance between external climate conditions (temperature, humidity, etc.) and internal comfort conditions. Broadly speaking, in cold climates there is an energy inlet from a building system and an energy outlet through the building envelope (facade, windows, etc.), and the opposite for hot climates. Therefore, Heating and Cooling Costs are largely driven by climate region, and construction and building system features, which is the bulk of our dataset.

Some features have particularly high weight on cooling and heating value prediction:

- Climate zone
- Unit type
- Building age (which is a proxy for insulation level)
- Square footage

Regarding *Water Heating Cost*, its value is mostly driven by number of household members, their behavior, and heater and fuel type. Since social and behavioral information is not available in our dataset, little target variance can be explained with our data and model. Nevertheless, number of bedrooms and

bathrooms do provide some information about the variance of water heating cost, they might be representing a proxy for number of household members. Please note that this response variable accounts for the energy spent on water heating and not the water consumption itself.

When it comes to *Other Electricity Costs*, this target variable is hardly explained by any building intrinsic features, instead it is more likely driven by the energy intensity of the unit, which means the appliance stock, number of lighting fixtures and their type, and their usage pattern. However, just like happens with Water Heating Cost, number of bedrooms, bathrooms, and square footage add a minimal value to the prediction, which makes sense since the larger the unit's area the more appliances and lights it contains.

Error

Prediction error measured by mean-absolute-error score is significantly high in comparison with median values in the test set even in the *Cooling Cost* variable. This might be explained by the spread of the distribution of the target variables, the large number of outliers could be outweighing in the mean error.

Additionally, some other error in the target variable is also due to different energy costs nationwide. All the selected features in our dataset are intended to provide insights on the level of energy consumption, but our final target variables are dollar bill costs.

6. Conclusion

In general, none of the explanatory variables have a strong correlation with the target variables and are not able to explain their variance to a large extent. This means that either consumption cost have a high random component, which does not seem reasonable, or there are many confounding factors that weighs on the final value; including electricity costs in different locations, household socioeconomic characteristics, other building features not included in the dataset, such as building orientation, window area, etc.; to name a few.

At the same time, many features selected do not seem to be individually relevant to predict heating and cooling costs, but still a combination of them are able to explain over 90% variance in case of cooling costs and 70% in case of heating costs. On the contrary, no combination of the available features in the dataset were sufficient to explain the variance of water heating costs and electricity cost.

The inclusion of *Total Consumption Cost* did not bring any significant improvement to the model prediction power. However, predicting all different energy use combined in a single variable is still interesting since the errors from the prediction of individual variables could balance among them.

7. Recommendations

- This model could be used to get an estimate of the cooling and heating cost of a certain unit, but its prediction will not be very useful for units with low consumption costs, in which the prediction error may be too large.

- Regarding Electricity and Total consumption costs, the model has a weak prediction power but still can provide an upper consumption limit with 80% probability.

- Finally, this model definitively will not predict Water Heating Costs with accuracy, since the variables included in the dataset cannot explain its variance. Therefore, the output of the model for this target variable will be close to random.

8. Further research

Potential further research includes develop methods to capture additional information of the target units and translate it into new explanatory variables from the RECS dataset that could be included in the model and enhance its prediction capacity.

Particularly, behavioral, or social features could cast more light on the drivers of the water heating consumption; and appliance stock information will probably generate a more robust prediction of electricity consumption costs.

Other line of further research could be filtering by unit type, climate zones or other broad classification, and perform the same prediction modelling. Perhaps the outcome of the model could be more accurate and with lower mean absolute error.