Predicting Building Annual Energy Use Intensity Using Lasso Regression Method

Yuqi PAN (yuqip) Siliang Lu (siliang1)
yuqip@andrew.cmu.edu siliang1@andrew.cmu.edu

ABSTRACT

In this paper, the lasso (Least absolute shrinkage and selection operator) regression method is used to predict the building annual energy use intensity. The CBECS dataset of 2003 is used for the analysis. The goal of the project is to fit a Lasso model to predict the annual EUI and evaluating the model by testing and comparing. We also try to perform feature selection using Lasso. The literature regarding to energy consumption prediction and analysis is explored and the approach adopted by this study is explained. After that, an explanation of the dataset and how the dataset is preprocessed is given. In the analysis part, we explored the relationship between tuning parameter and corresponding mean square error (indicating the performance of the model) and number of features selected (indicating the level of shrinkage of the model) . The optimum model is selected using cross-validation method. Some properties including: tuning parameter, selected features, MSE of the selected optimum model is investigated is discussed. The investigation of using Lasso to perform feature selection is also conducted and the select features are discussed. The analysis showing that the Lasso is able to predict the EUI for majority of the buildings but do not have the ability to capture outliers. The regression model cannot capture features represented in binary or categorical forms. Using subset to analysis End use is a good way to improve perdition which can be investigated in the future.

1. Introduction

With the development of cities in the 21st century, commercial buildings have played a key role in shaping the modern urban environment. By the end of 2014, the number of commercial buildings has increased to 5.6 million, and the amount of commercial floor space has increased from 51 billion to 87 billion square feet throughout the US [1]. High performance commercial buildings, which integrates and optimizes all major high performance attributes, including energy efficiency, durability, life-cycle performance and occupant productivity have potential to make a better city. Hence, many studies have been conducted to improve the built environment of the commercial buildings, particularly the energy performances. The energy consumption of commercial building accounts for around 18 percent of the total energy consumption in the US.

Traditionally, building scale demand estimates have been performed using engineering software packages (e.g., EnergyPlus) that rely on an in-depth compilation of building properties[2]. However, another forecast method based on analysis of large-scale data with effective machine learning approaches. Many interrelated attributes regarding buildings more or less influence the energy performances and are furtherly applied to forecast and analyze building energy consumption and describe building energy use patterns. Data analysis based on machine learning methods not only predicts the energy consumption level but also has the ability to select the most relevant attributes. Zhun Yu, Fariborz Haghighat et al(2010) eveloped a decision tree method for building energy modeling. As a result, the indoor temperature and whether to use electrical heat are the two most important factors to the building energy performances. In addition, Howard et al. (2012) implemented a robust multivariate linear regression to analyze building energy use by end use in New York. Tian, Song, and Li (2014)explored the spatial patterns of domestic energy in London using spatial error models.

In this report, Lasso (Least absolute shrinkage and selection operator) Regression model was applied to find out the relevant attributes to the energy performances of commercial buildings. The formula of Lasso Regression is shown as follows:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Where λ is the tuning parameter. When λ is zero, informing no penalization is taken and the equation becomes linear regression. When λ getting bigger, more penalizations are taken and more coefficients will turned to zero indicating less feature selected. By adding this penalization term the Lasso can be used to perform feature selection. The Lasso model was then validated with cross-validation test method and the alpha corresponding to the minimum of MSE was selected for feature selection of the datasets.

In this report, the module of sklearn.linear_model.Lasso [6] as well as sklearn.linear_model.LassoCV were used to study the following relations:

- relation between the amount of attributes selected and λ (in the package, lambda is called alpha).
- relation between the MSE and λ
- the validation of linear regression model for the prediction of the EUI with the selected attributes

2. Data Process

The dataset used is the 2009 (year published) Commercial Building Energy Consumption Survey (CBECS) from EIA. The dataset has more than 5000 instances of commercial buildings all around US. The survey included questions related to several aspects of building characteristics and roughly divided that into several major files. The first seven major files include all kinds of building characteristics and the File 15-20 contains the energy consumption and end uses. The files used for this analysis is File 1-7, File 15 and File 17. The imputation files (File 8-14) and end uses files other than electricity and natural gas are not included in the study. The layout of the files is as shown in Table. The dataset consist of three data types including nominal (categorical), numeric and binary forms. The dataset is preprocessed before feeding into models and performing analyses. The preprocessing procedures are below:

1. The first 7 Files are combined to form a dataset containing all attributes
2. The repetition part of the files is eliminated
3. The attributes with a lot of missing data is eliminated from the dataset. This is because imputation of attributes with a lot of missing data can lead to noise and inaccuracies. The attributes which data can only be obtained from a small number of buildings

is assumed to be less relevant and do not influence the big picture of building consumptions

4. The heating and cooling degree day information is also extracted from file 15 since these two values is a better representation of the location climate and is closely related to building total consumptions.

5. The building annual EUI is calculated based on the total square footage and major fuel consumption and added to the dataset

6. The building sample weight is added to the dataset

7. The final dataset has 5215 rows and 162 columns. The first 159 columns contain 159 different building or climate features marked as Predictors. The rest three columns are the total and electricity EUI of building and sample weighting factors, marked as the Targets (Note: weighting factors are also marked as targets only for the convenience of processing)

8 The same Lasso Regression method is applied to study the important attributes to the energy consumption of HVAC subset and results are compared with the overall dataset.

Table 1 Basic Building Information

Basic info.	Basic Building information
FILE 1	Basic Building information and End Uses
FILE 2	Building Activity and Special Measure of the Size
FILE 3	Heating and Cooling Features and Conservation Features
FILE 4	Water Heating, Refrigeration, and Office Equipment; and Special Space Uses
FILE 5	End Uses of Major Energy Sources; Electricity Generation; and Purchasing of Electricity and Natural Gas
FILE 6	Minor Energy Sources and End Uses for Minor Energy Sources
FILE 7	Lighting Percent, Equipment, and Conservation Features

The imputation process is done using python and Scikit Learn package. Three types of data are imputed differently. The imputation method of binary and categorical is to use the value which is most frequently appeared in the column, while for numeric values, the median of the column is used for missing data. We use median here instead of mean to avoid the influences of extreme values.

To better fit the model and assess the performance of the model, the dataset is split into two subsets with equal samples. The first subset is referred as training set and the second subset is referred

3. Exploratory data analysis

The exploratory analysis of the target of prediction is conducted to give general information of the distribution of the target values. A histogram is plotted to show the distribution of EUI values in the whole dataset.

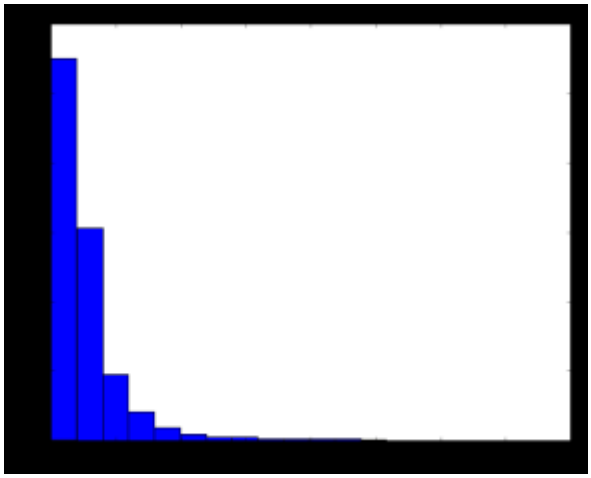

Figure 1 Total EUI Distribution

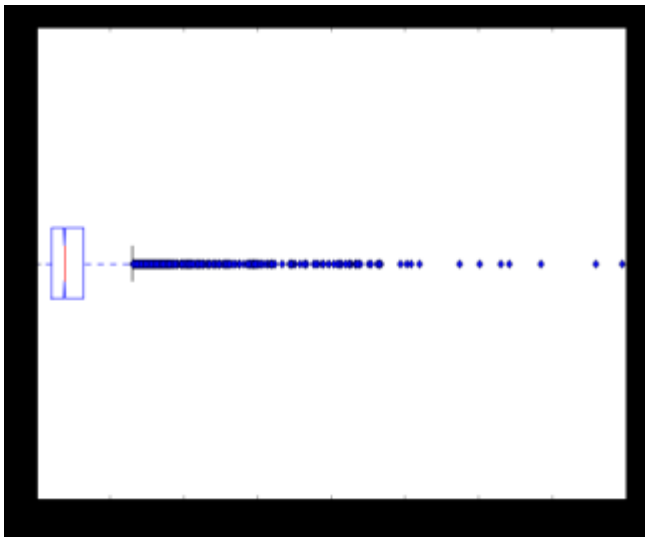

Figure 2 Box plot of building EUI

As shown in the histogram above, most of the buildings have an EUI less than 200 MBtu/sqft, and there are several buildings with higher EUI values reach to 600MBtu/sqft. There are some buildings which have really high EUI values. Since the number of buildings which has extremely high EUI is very small and is not likely to affect the whole dataset, these buildings can be regarded as outliers which can increase the error of the model fitted. Some of the staticstical information regarding total EUI is shown in Table 2

Table 2 statistical information regarding total EUI

Population size	5215
Mean EUI	108.6 MBTU/sqft
Median	76.57 MBTU/sqft
Standard deviation	122.8 MBTU/sqft

As suggested by the table, comparing the mean and median value, the standard deviation is very large indicating the dataset has some outliers which may influence the model fitting in the following analysis.

4. Model Investigation

4.1 Selecting Optimum Model with Varying Tuning Parameter

The model is established using the Scikit - learn package for python which can perform a variety of statistical analysis and machine learning functions. Our first task is to select an appropriate tuning parameter (alpha) which can gives us the best fitted model. This is done by varying the tuning parameter in a pre-defined range and produces a series of models using the training set. These models are tested using the testing set and the model performance is evaluated based on the Mean Square Error. (Method 1) The model with the minimum MSE is selected to be the optimum model. A plot of Mean Square Error (MSE) with different alpha is given. The penalization also determines the shrinkage of the model represented as the number of features selected. If the coefficient of a certain feature is not zero, the feature is considered to be selected by the Lasso model. The plot of number of features selected against tuning parameter is also given.

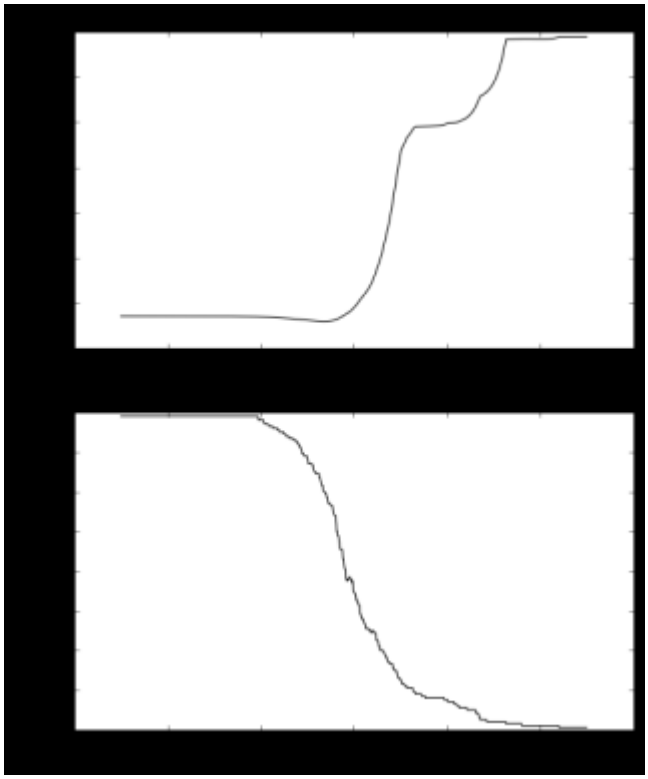

Figure 3 change of MSE and number of selected featureswith Tuning Parameter

The MSE decreases with increasing alpha when alpha is less than 1 and increase rapidly afterwards. Compare the change of MSE and change of number of features selected diagram; it is clear that the number of features selected has a great influence on the MSE. The MSE varies only a little when all features are selected but as the number of features decreases, the MSE drops significantly. The result showing that the optimum number of features lies towards the high end when majority of the parameters are selected (but not all features are selected).

The important parameters discovered are shown in Table 3

Table 3 summary of statistical information of the model

Minimum MSE	9602.5
Corresponding alpha	0.22
R square value	0.39

The selected model is only tested once using the testing set. A better way to select the best model from training dataset is to do cross-validation (multiple tests) and select the model which gives the minimum cross-validation error. (Method 2) The cross-validation is performed to divide the training set into several parts and use one part as the training set and rest to the testing set. This process is performed iteratively until all the parts are fitted and tested. The process can be done by a predefined function in the package: LassoCV(). The LassoCV is capable of fitting the model with a range of alpha values and select one optimum model with least cross-validation error. The cross validation model fitting and selection is performed using the training dataset to produce the selected "optimum model" and the performance of the selected model is tested on the testing set. Since out dataset is not large, the cross validation is defined to be 5 fold (with approx. 400 samples in each fold). The result of MSE of each fold (shown as colored dash line) and mean MSE across all the folds (shown as black colored line) are shown in the plot.

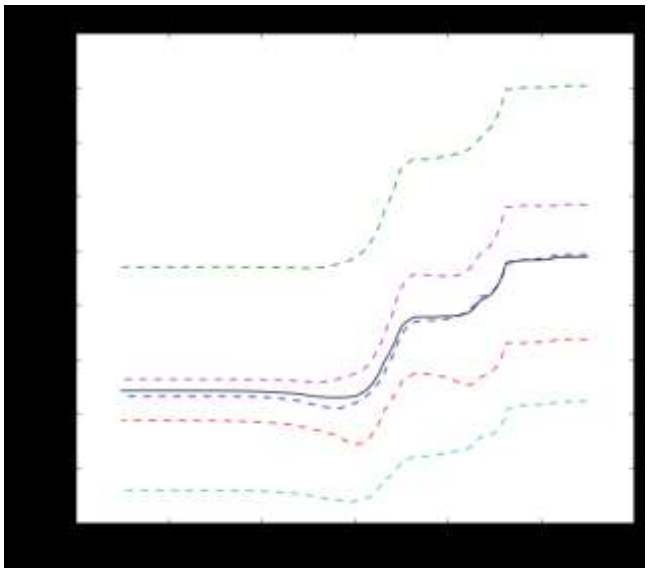

Figure 4 change of MSE with tuning parameter of 5 folds

The important parameters from the selected model is

Table 4 summary of statistical information of the selected model

Minimum MSE	10595.67
Corresponding alpha	0.31
R square value	0.41

Note: The minimum MSE refers to the minimum mean cross-validation MSE.

Comparing the model selected by only doing 1 test and cross-validation of 5-folds, the trend and results obtained are very similar in terms of alpha and minimum MSE. The minimum MSE is a bit larger in cross-validation model as there are more variations during cross validation. The coefficient of determination of only 0.41 shows that the regression model is not well fitted. One possible reason for that is that most attributes are given in Binary or Categorical, only 21 out of total 159 attributes are given in numerical values and also, the numeric values are not being

normalized. Generally, binary and categorical data is not good for regression model. In addition, though attributes with a lot of missing data are eliminated in advance, the remaining data is still not complete. The imputation method can create a lot of inaccurate values and noise which can result in inaccurate predictions.

4.2 Evaluating the Performance of the Selected Model

A better way to illustrate the quality of produced model is to compare the predicted consumption values and real consumption values using the testing data set. The "optimum model" produced by cross-validation is selected as the final model and the corresponding coefficients are used to predict the consumption levels of testing dataset. A plot of predicted EUI and Actual EUI is given.

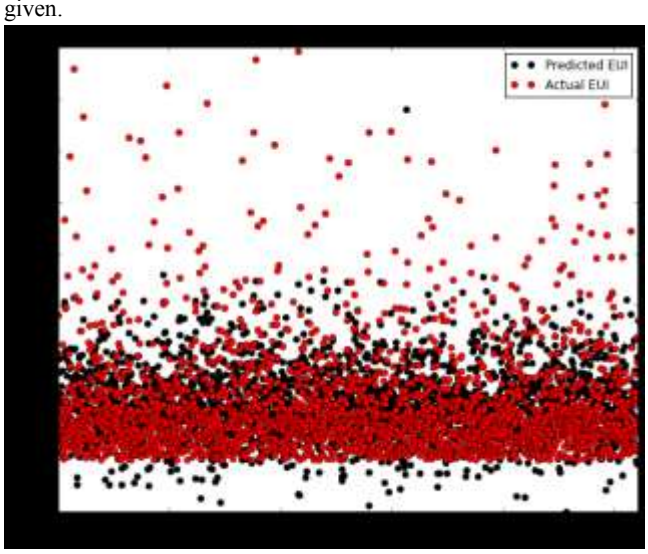

Figure 5 comparison of predicted EUI and actual EUI

From the plot, the result of predicted values and real consumptions are very close for most of the buildings in the general range of EUI (i.e. EUI < 200 MBtU/sqft). However, the Lasso regression model is not good in fitting buildings with very high consumptions (i.e. outliers) the value of which is beyond the range of majority buildings. It is probably because the buildings with high consumption have certain influential features given in binary or categorical form which cannot be easily captured by regression model.

Confidence interval at 95% for each of the coefficient β of the selected model is shown in Figure 6It can be observed that most of the coefficients are less than 50. In addition, the confidence interval for most of β is [-50, 50]. However, one attribute has an interval larger than 200, which is square footage". It makes sense since different buildings have different floor areas and vary a lot with each other.

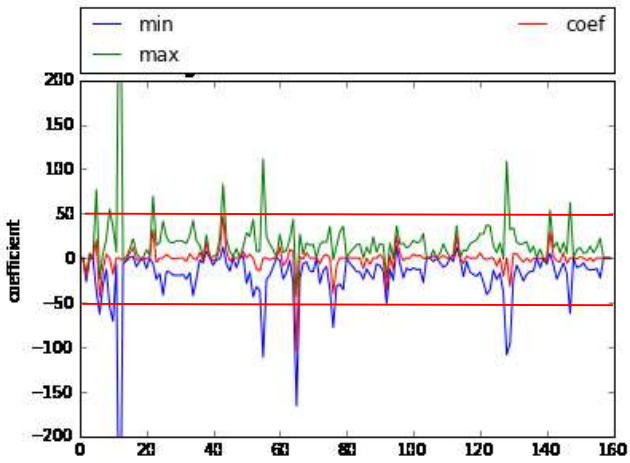

Figure 6 Confidence interval at 95% for each of the coefficient β of the selected model

Besides confidence interval, statistical significance of the final model was also conducted to determine whether the null hypothesis H0: β=0 for each of the coefficient shall be rejected. The hypothesis were tested at the 95% confidence interval.Figure 7 shows the T-value for each of the coefficient. Besides, T of the confidence interval at 95% is between -2.578 and 2.578. As shown in the figure, most of Ts are within the range, which means that those coefficients cannot reject the null hypothesis. In fact, 27 out of 159 attributes can reject the null hypothesis, which consists of all three types—binary, categorical and numeric.

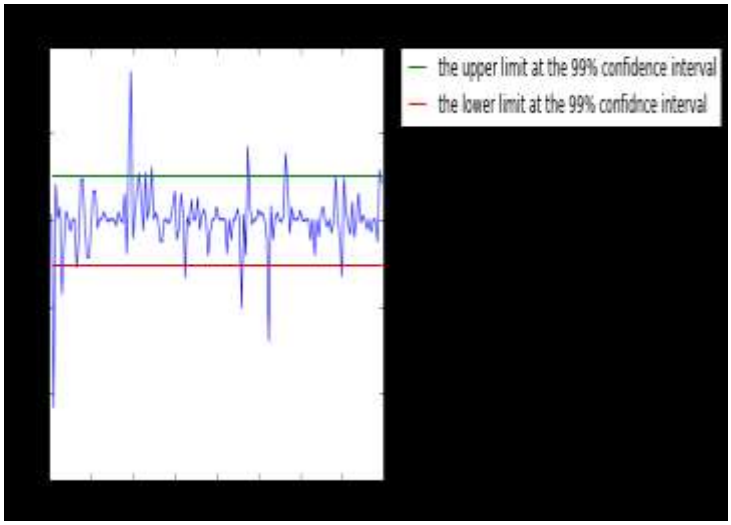

Figure 7 T-value for each of the coefficient

In the section feature selection, since it has concluded that numeric attributes play an important role in the linear regression model, the numeric attributes which has the ability to reject the null hypothesis are considered to be more important than other attributes. Hence, "percent exterior glass", total weekly operating hours", "number of photocopiers", "heating degree days" and "cooling degree days" were selected for future predictions of the energy consumption level of the commercial buildings.

4.3 Selected features

As discussed before, the Lasso regression has the ability to "shrink" and perform attribute selection. Therefore, the selected features are analyzed. Using the "optimum model" produced by cross-validation function, a set of coefficients is generated. The features with corresponding coefficients that are not zero is marked as "selected". By analyzing the selected model, total 114 out of 159 features are selected (not shown). A distribution of features selected into different files is shown below:

Table 5 distribution of features

File	Number selected	Number total	% selected
Basic info.	9	10	90%
FILE 1	33	46	72%
FILE 2	1	1	100%
FILE 3	20	31	65%
FILE 4	22	28	79%
FILE 5	11	20	55%
FILE 6	0	0	--
FILE 7	16	21	76%

It is hard to tell which category of attributes has the most significant influence on the building energy consumption since the number of features in each category differs since a lot of features are eliminated in the pre-processing. A general idea from the table is that the building basic information plays a more important role in total energy consumption. To better understand the level of influence of features, a feature selection analysis is performed.

4.4 Using lasso to perform feature selection

The "optimum model" selection is based on cross-validation error; therefore, a large number of features are selected to better fit the model. With more than 100 features selected, it is hard to determine which feature has the most significant influence on the total consumption. In this section, a method is proposed to determine the relevant importance of the parameters. The investigation is performed as to keep increasing alpha and shrink the number of features selected. To achieve this, we control the number of features to be selected (shrinkage target) and keep increasing the value of alpha until the target is reached. Different number of shrinkage target is defined and the selected features under each target is shown in the table below

Table 6 summary of number of target and selected alpha

Target	Alpha	Feature selected
1	25397.65	Square footage
3	2373.76	Square footage Number of employees during main shift Heating degree days (base 65)
5	662.47	Square footage Number of floors, Total weekly operating hours , Number of employees during main shift, Heating degree days (base 65)
10	266.23	Square footage, Number of floors, Total weekly operating hours , Number of employees during main shift, Computer area percent, Heating degree days (base 65), Number of computers, Percent lit when open, Cooling degree days (base 65)

The general results given in feature selection are consistent with what reported in literatures. The building operational schedule, occupancy density, climate severity (represented as heating and cooling degree-days) seems to be the most influential features towards energy consumption. One interesting finding is that even the total EUI (which normalize the area) is used as prediction target, the square footage appear as the most influential parameter using Lasso method. Also, it is worth noting that all the selected features are of numeric values, which means that some possible important features represented in the binary or categorical form cannot be captured using this method and hence has little use in regression model.

5. Evaluating HVAC Subset

Using the same method as discussed before, a similar Lasso model is built using the HVAC sub dataset (File 3) to fit a model to predict total heating and cooling consumption (End Use) (File 17). As discussed before, considering the complexity of the dataset and features contributing to total consumption, the coefficient of determination is low, indicating a less strong correlation and less accurate model. The assumption of investigating into a subset is that by reducing the total number of features and focusing on subset and related end use, we would be able to get a more accurate model.

As a result, the minimum MSE of the model cross-validated with 5 folds is calculated as 2432.85, which is smaller than that of the overall dataset. Table shows the statistical information of the selected model for HVAC subset.

Minimum MSE	2432.85
Corresponding alpha	0.094
R square value	0.31

Hence, the information shown above confirms that the investigation of subsets could bring out more accurate results.

6. Conclusion

From the studies, conclusions are drawn as follows:

- Generally, Lasso regression can give a relatively accurate prediction to the majority of the buildings in the building stock. However, it lack the ability to predict outliers. This is because that some of the information represented in the binary or categorical form can not be effectively extracted through regression model.
- Numeric attributes are more suitable for regression analysis and can produce better results.

- The lack of accuracy of the model is possibly due to large amount to features are represented in the binary or categorical form rather than numeric form. Other reasons may be that the dataset is not complete and has a lot of missing data. The imputation of missing data is not accurate
- Lasso regression is useful for feature selection and can be integrated with other methods to develop more accurate results.
- Prediction of end use using subset can be more accurate and therefore, further studies can look in to establishing individual models to predict end use.

7. REFERENCES

[1] EIA, "A Look at the U.S. Commercial Building Stock: Results from EIA's 2012 Commercial Buildings Energy Consumption Survey (CBECS)," 2012. [Online]. Available: https://www.eia.gov/consumption/commercial/reports/2012/buildstock/.

[2] S. Petersen and S. Svendsen, "Method and simulation program informed decisions in the early stages of building design," *Energy Build.*, vol. 42, no. 7, pp. 1113–1119, Jul. 2010.

[3] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy Build.*, vol. 42, no. 10, pp. 1637–1646, Oct. 2010.

[4] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, and V. Modi, "Spatial distribution of urban building energy consumption by end use," *Energy Build.*, vol. 45, pp. 141–151, Feb. 2012.

[5] W. Tian, J. Song, and Z. Li, "Spatial regression analysis of domestic energy in urban areas," *Energy*, vol. 76, pp. 629–640, Nov. 2014.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in {P}ython," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.