Automated Feature Selection With Ensemble Methods on the 2012 CBECS Data

Kal Parvanov, Brian Morales, and Wyett Considine

Abstract

In this paper we explore the 2012 CBECS to better understand what specific building features contribute the most to energy consumption. Our approach in achieving this objective includes the use of automated feature selection techniques from the wrapper and embedded method families, and the subsequent training of four different ensemble models which are individually evaluated with the SHAP method for greater insight into the effects of the selected features on each model's power of prediction.

1 Introduction

In 2018 alone, U.S. commercial buildings consumed 6.8 quadrillion British Thermal Units (BTU), summing $142 billion spent on energy(eia,). These vast consumption and expenditures are highly problematic for multiple reasons. Research shows that this level of energy consumption is detrimental to the environment, and contributes greatly to climate change (Akpan and Akpan, 2011). In addition, when buildings pull too much energy from the electrical infrastructure, it can be harmful to the surrounding grid. When there is too great a pull on the energy grid, it can cause blackouts and brownouts, which can cause damage to urban environments. One can attempt to predict locations that will need more energy based on building features. If one can predict the locations that will experience spikes, then higher amounts of robustness can be built into the network to prevent outages. This paper aims to analyze the building features that demand higher electricity expenditures in order to create optimal building planning for future development.

Some commercial buildings consume much more energy than others. We aim to determine the features of these buildings that require more energy, and therefore lead to the usage of more energy. To do this, we will be using machine learning techniques to analyze the Commercial Buildings Energy Consumption Survey (CBECS) to formulate a list of the most impactful features to contribute to the Energy Usage Intensity of commercial buildings in the United States.

1.1 Problem Space

The U.S. Energy Information Administration administers the Commercial Building Energy Usage (CBECS) survey on a representative subset of the commercial buildings across the U.S. The survey data also includes information about numerous features that play a role in determining a building's energy usage. Examples are the total square footage of a building, number of floors, if a building is solar powered, etc. In this paper we attempt to explore the predictability of Total Electricity Usage Intensity (Total EUI) defined as the ratio of the total annual electricity consumption over a building's square footage.

In Fig. 1 we see the distribution of the Total EUI has a positive skew, with most of the observations lying around $[0, 50]$. In particular, the Total EUI mean is $\mu = 18.43$, the standard deviation is $\sigma = 20.83$, and 75% of our observations are exclusively around 22.75. The data cleaning procedure eliminated most of the missing values and outliers. In section 2.2, we discuss the data cleaning and pre-processing procedure more extensively.

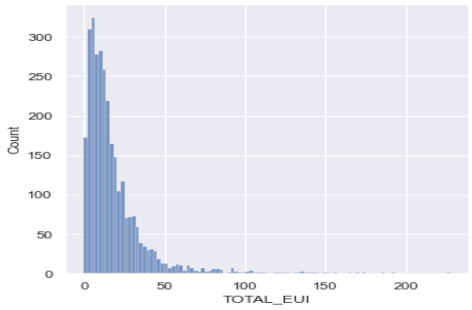

Figure 1: Distribution of TOTAL EUI

2 Approach

2.1 Method Outline

This paper has two objectives: to apply the automated feature selections methods Lasso and RFE; and find the best interpretable model to use in the exploration of feature significance with respect to Total EUI. Much of this analysis is based on feature selection and pre-processing of the CBECS data. We use a combination of feature selection techniques, including Wrapper method and Embedded methods to reduce the total pool of possible features. Then we construct and train models on the CBECS data set to predict new buildings' energy efficiency. Once we have valid models, we use the SHAP method to analyze the best-performing models to express the building features that contribute the most to predicting a building's energy consumption. We recognize that some buildings are much bigger than others. This can make direct comparison of energy consumption between misleading buildings. Instead of direct energy consumption, we are using Total Energy Use Intensity (Total EUI). EUI is an indication of the building's energy efficiency based on the building's design and operation. We find this measure to be a better quantity to observe, as it normalizes the buildings' sizes, and expresses the energy efficiency of the building's features more accurately. The developed EUI models were evaluated and compared in terms of their R^2, adjusted R^2 scores, and RMSE values. Lastly, we utilized the SHAP method to illustrate the most important features in the models.

2.2 Lasso Feature Selection

Feature selection is an important process in the development of building energy models because of the large dimensions in the data set. Reducing the amount of features can largely improve the training time, complexity, accuracy, and interpret-ability of a model.

In the regression setting, the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

is commonly used to describe the relationship between a label, Y, and features, $X_1, X_2, ..., X_n$. However, when dealing with high dimensional data sets, the inference of a model can become complicated. To keep the distinct inference advantage one method we decided to conduct was the Lasso method.

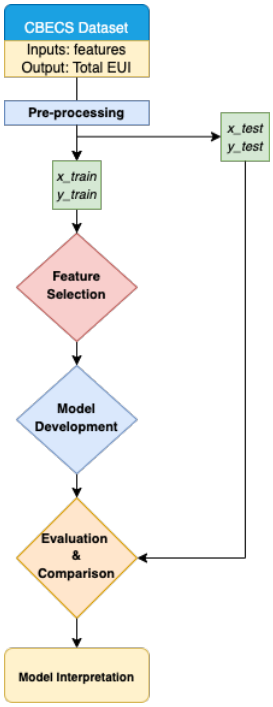

Figure 2: Research Outline

Its often the case that some or many of the features used in multi-regression are in fact not associated with the response. Including irrelevant variables to the model can lead to unnecessary complexity. To address this issue we apply the shrinking method, Lasso with cross-validation, to shrink the irrelevant estimated coefficients to 0. As a result we were able to reduce the amount of features from 1000 to 36.

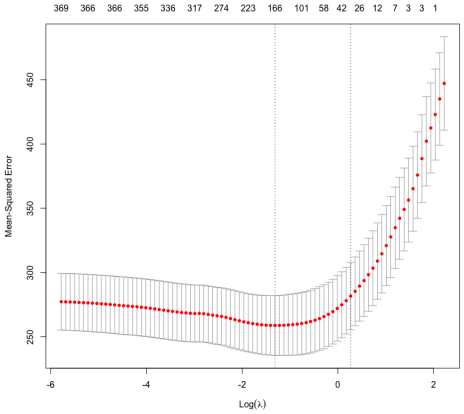

Figure 3: Cross-validation with Lasso

In Fig. 3, we directly estimate the penalty term, λ, to find the best term that produces the most accurate regularization. To find the best λ, we find the term that best minimizes

the MSE by following the equation,

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

and return all its estimated coefficients associated with the model.

The results we found best was $\lambda = 0.26$; the penalty term shrank more than half of the estimated coefficients to 0, resulting in only 36 important features.

2.3 RFE Feature Selection

Recursive Feature Selection is a wrapper-type automated feature selection technique that uses a pre-selected machine learning algorithm and pre-defined n number of features to loop recursively through the data and eliminate all the irrelevant features until the m best possible features are left. This particular feature selection was found to perform particularly well in Liu'swork with healthcare building energy usage data (Liu et al., 2022).

In our paper, we decided to implement a 5-fold cross-validated Recursive Feature Selection (RFECV) with a Decision Tree Regressor as our estimator (using adjusted R^2 as measure of accuracy). The goal of this strategy was to find the subset of best performing features and minimize subsequent model confusion due to a high number of features. After running (RFECV) we learned that there are 314 meaningful features in our dataset. Below is a plot of number of features versus mean test accuracy.

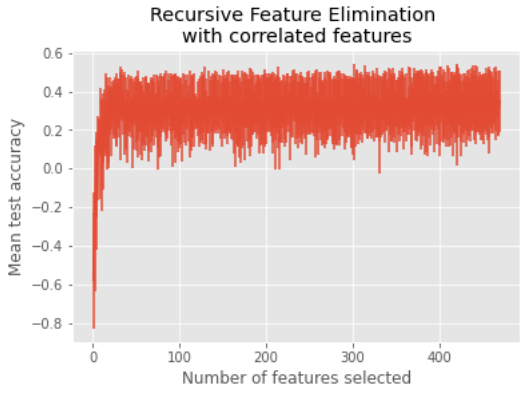

Figure 4: Num. of Feat. v.s. Mean Test Accuracy

2.4 Machine Learning Algorithms

Ensemble learning has been suggested as a powerful machine learning method for developing building energy models that describe the complex relationship between the model output and inputs (Liu et al., 2022). The ensemble methods we decided to proceed with were Catboost, XGBoost, LightGBM, and Random Forest (RF).

XGBoost is a boosting algorithm developed based on the gradient boosting decision tree (GBDT) by Friedman and further improved by Chen and Guestrin (Liu et al., 2022). The improvements in XGBoost mainly includes significant increase in efficiency and performance. XGBoost applies a second-order Taylor expansion to estimate the value of loss and find the best base learner. One advantage of XGBoost is that it prevents model over fitting, which is a drawback of the GBDT (Liu et al., 2022).

Random forest is a bagging algorithm that builds on a number of decisions trees on bootstrapped training samples. However, when building these decision trees, each time a split in a tree is considered, a random sample of f features is chosen as split candidates from the full set of f features. The model error is calculated by the MSE, but better yet, we can average the out-of-bag (OOB) MSE of all trees and provide a more accurate representation.

LightGBM is a gradient boosting ensemble method that creates decision trees that grow leaf-wise. LightGBM can be used for both regression and classification and does well when implemented on sparse datasets. In many ways LightGBM is a competitive alternative to XGBoost which is much slower when it comes to training time. Previous research by (Liu et al., 2022) suggests its relevance to datasets of the type we are working with.

CatBoost is another gradient boosting ensemble method that (unlike XGBM and LightGBM) uses symmetric trees and ordered boosting (better than regular boosting) to cut down on computing time and produce resultsmuch more quickly than XGBM. Out of all three boosting algorithms CatBoost is best at handling categorical features and one-hot-encodes which makes it a goo choice for our proble. Furthermore, we noticed that CatBoost had not been used by other researchers on the 2012 CBECS dataset, which motivated us to attempt to implement it.

2.5 Model Interpretation

In order to be able to interpret the best performing model we decided to use the SHAP method. SHAP is a game theoretic approach to explain-

ing the output of a ML model. In the model the features are treated as players in a game where the objective is the model's prediction. The Shapely value, the average expected contribution of one feature to the output of the model is calculated by considering all possible combinations of features (2^N for N features). In reality, this is a NP-Hard problem so to get around this the SHAP method uses random sampling and approximations to find the SHAP value. Once we have found all the SHAP values we can produce summary plots which serve to inform us of each feature's importance to the predictive power of the machine learning model. This tells us a lot about the hidden relationships between different features and the predicted variable.

3 Data

3.1 Data Cleaning

The 2012 CBECS dataset consists of 1119 features over a sample of 6720 buildings. The dataset is plagued by multiple missing values and various potentially unreliable feature imputations. For our project we decided it best to eliminate all features with majority (details explained below) missing values. The selected values would shrink the number of observations drastically.

The particulars of the data cleaning process can be outlined in two major steps. First, we selected the all the non-imputed or weight variable features (from 2 to 446 and 1094 to 1104 in column index), which resulted in a subset of 455 features with all original samples. Second, we looped through the features sequentially keeping all the features that if added to our list of selected features would not leave us with fewer that 3000 observations when all the row NaN values are dropped. We finally, dropped all the NaN values of the resulting subset and we ended up with 3009 observations across 125 variables. The explanatory variable Total EUI was created as the ratio of Total Yearly Electricity Consumption (ELCNS) in kWh over Total Bulding Square Footage (SQFT). (Note this variable has been suggsted in previous work by (Deng et al., 2018))

3.2 Pre-processing

Based on previous research done by (Liu et al., 2022) on set of Chinese healthcare buildings, we decided to perform one-hot encoding on all of our categorical variables in the hope that we could isolate the most meaningful features. Also, note that the negative consequences of the data sparsity of the one-hot-encoding was mediated by our choice of ensemble models, which have been shown to handle sparse data well (Liu et al., 2022). After implementing the one-hot-encoding we ended up with 455 features and 3009 samples. We then proceeded to split our data set and explanatory variable into train and test data using a 3 : 1 split with random sampling (sklearn train-test-split).

4 Results

Model Results		
Algorithm	Ft. Selection	Adjusted R^2
CatBoost	Lasso	0.57
	RFE	0.68
LightGBM	Lasso	0.56
	RFE	0.60
XGBoost	Lasso	0.54
	RFE	0.61
RF	Lasso	0.57
	RFE	0.58

Even though feature selection methods such as, Lasso and RFE used in this study, can quantify the effect of input features on EUI, the outcome may not be reliable since analysis does not rely on the adjusted R^2 of a model. Therefore, the influence of the input features by the SHAP method is still needed for model interpretation (Liu et al., 2022).

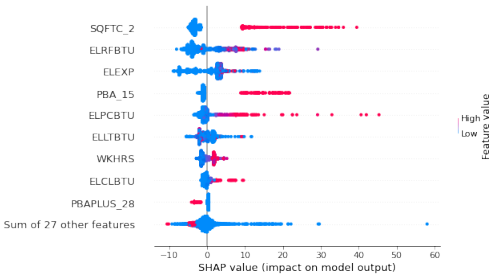

Figure 5: CatBoost SHAP Values Lasso

In Fig. 5, 6 we can see the most substantial features are electricity refrigeration use (ELRFBTU) and the square footage of a building from 1000 to 5000 sqft.

In Fig. 7, 8, we have similar important features in the Catboost SHAP summary plots.

5 Discussion

In order to be consistent with our models we decided to work with adjusted R^2 because of the misinterpretation R^2 can have. Recall that

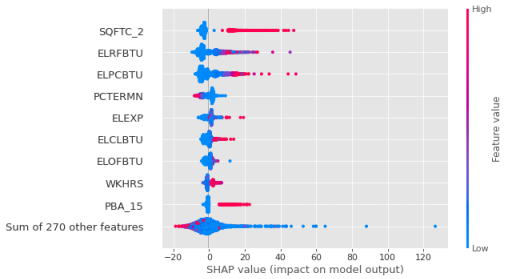

Figure 6: CatBoost SHAP Values RFE

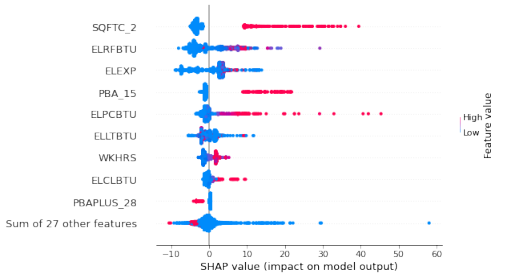

Figure 7: LightGBM SHAP Values Lasso

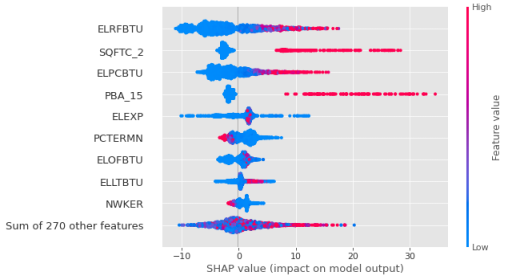

Figure 8: LightGBM SHAP Values RFE

R^2 is defined as $1 - RSS/TSS$, where $TSS = \sum(y_i - \hat{y})^2$ is the total sum of squares for the label. Since RSS always decreases as more variables are added to the model, the R^2 will always increase as more features are added. Hence, we decided to move forward with adjusted R^2 because once all of the correct features have been included in the model, adding additional irrelevant features will lead to only a very small decrease in RSS.

According to the adjusted R^2, CatBoost performed the best for all lasso feature selection methods. This was not surprising since Catboost tunes its hyper parameters by executing grid searches that find the best performance and is also much better equipped at working with one-hot-encoded categorical variables (sparse data).

For Catboost with Lasso model, each features contribution was calculated using the SHAP values as shown in Fig 5. The x-axis indicates the SHAP values for each feature, where positive and negative values indicate the positive and negative correlations of the input features to the response, EUI. The y-axis denotes the feature names ordered by importance and each point for a feature represents an individual sample. The color bar change from red to blue indicates the SHAP value of the feature from high to low (Liu et al., 2022). This is similarly done with the rest of the beeswarm/summary plot figures.

Notice that for CatBoost + Lasso and LightGBM + Lasso, the adjusted R^2 is 0.1 decimal different and the SHAP values for both models chose the same important features. The R^2 results for the CatBoost and LightGBM models with RFE were significantly better than the results for the two models when coupled with the Lasso feature selection method. In particular the CatBoost + RFE model gained an additional 18 percent increase (.10) in R^2 over CatBoost + Lasso suggesting that RFE was better at handling one-hot-encoded data. A interesting observation is that the top features contributing to each model's performance were the mostly the same but often in different order of importance. For example, PBA_15, which indicates food service as the principal building activity appeared in all of the summary hinting at a strong positive correlation with Total EUI.

References

Usenobong Friday Akpan and Godwin Effiong Akpan. 2011. The contribution of energy consumption to climate change: A feasible policy direction. *International Journal of Energy Economics and Policy*, 2(1):21–33, Dec.

Hengfang Deng, David Fannon, and Matthew J. Eckelman. 2018. Predictive modeling for us commercial building energy use: A comparison of existing statistical and machine learning algorithms using cbecs microdata. *Energy and Buildings*, 163:34–43.

Energy information administration (eia)- commercial buildings energy consumption survey (cbecs).

Xue Liu, Hao Tang, Yong Ding, and Da Yan. 2022. Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings. *Energy and Buildings*, 273:112408.