Prediction and Feature Analysis of Electricity Consumption

Zhongyuan Li
zhongyul@andrew.cmu.edu

Shawn Li
shuol2@andrew.cmu.edu

ABSTRACT

In this project, we want to estimate total electricity consumption using the data sets from RECS (Residential Energy Consumption Survey). Maximal Information Coefficient (MIC) is used to select some closely correlated features, Regression Trees are used to train the predictive model and predict the data sets. In addition, a K-fold Cross Validation is introduced to validate the accuracy of the predictive model. To test the accuracy of the features selected by MIC, we select features based on different MIC values and our own experience. Finally, we focus on several single features and analyze their relationships with total electricity consumption. The results indicate that the residential electricity consumption is closely related with some features and MIC method can effectively help us to select those features.

Keywords

Building Electricity Consumption, Regression Trees, Maximum Information Coefficient, Cross Validation

1. INTRODUCTION

As the energy consumption increases rapidly in recent years, many concerns about it have been raised. Institutions such as the International Energy Agency (IEA), the U.S. Energy Information Administration (EIA), and the European Environment Agency record and publish energy data periodically. Improved data and understanding of World Energy Consumption may reveal systemic trends and patterns, which could help frame current energy issues and encourage movement towards collectively useful solutions [1]. RECS data sets from EIA are chosen in this project. One of the main problems in RECS is that there are too many features (more than 900 features). Thus, it is hard to use all of the them to predict the energy consumption. Another problem is that the energy consumption is various and it contains electricity, fuel, gas, etc. In order to simplify the prediction, we only focus on the total electricity usage for residents and MIC method is introduced to filter the features.

The paper is organized as follows. In Sec.2, several main methods used in the project is described. Sec.3 presents the detailed working process including how to code the program, how to implement the methods, etc. The results and analysis are in Sec.4 and Sec.5 provides detailed study on relationship between several single features selected from MIC method and total electricity usage. A summary of the results and a brief discussion of future work are shown in Sec.6.

2. METHOD

2.1 Regression Trees

Regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition [2]. It maps observations about an item (branches) to conclusions about the item's target value (leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees [3]. The detailed descriptions are in [2]. In this project, a useful package named scikit-learn in python is used to train the data sets and make predictions.

2.2 Maximal Information Coefficient

Electricity consumption in the RECS data sets consists of so many various features that we cannot easily find the relationship between these features and electricity consumption. These features can be influenced by the sizes of houses, frequent usage for heating and cooling, number of living residents, etc. In fact, these features work as a whole for the final electricity consumption. Therefore, in the first place, a powerful tool is introduced to filter these features and find the some main related features that can influence the electricity consumption.

According to [4], Maximal Information Coefficient (MIC) can search for pairs of variables that are closely associated, calculate some measure of dependence and rank the pairs by their scores. Hence, MIC method is introduced in this project to filter the factors in RECS data sets.

The definition of MIC is presented in [4] and [5]. First, let D be a set of selected pairs. And for a grid G, let $D|_G$ denote the probability distribution induced by the data D on the cells of G, and $I(-)$ is mutual information. Then let $I^*(D, x, y) = \max_G I(D|_G)$, where the maximum is taken over all x-by-y grids G. MIC is defined as follows:

$$MIC(D) = \max_{xy<B(|D|)} \frac{I^*(D, x, y)}{\log_2 \min\{x, y\}}$$

where B(n) is usually set as n0.6.

3. EXPERIMENT

MIC method is able to determine functional or not functional relationship between two variables and so it is very suitable for our project. In order to implement this method, a library in Python called minepy is used to calculate MIC values for each pair of data in RECS. Features with MIC values above 0.2 are selected for further analysis, shown in the Table.1.

Table 1. Features of MIC values above 0.2

Features	MIC
Type of housing unit	0.2090
Final sample weight	0.2062
Number of floors in a 5+ unit apartment building	0.2056
Number of apartment units in a 5+ unit apartment building	0.2056

Studio apartment	0.2142
Number of floors in an apartment	0.2142
Number of bedrooms	0.2340
Total number of rooms in the housing unit	0.2441
Basement in housing unit	0.2422
Finished basement	0.2422
Heating used in basement	0.2422
Cooling used in basement	0.2422
Portion of basement exclusively used by housing unit in apartment building with 2-4 units	0.2422
Well water pump used	0.2142
Number of rooms heated	0.2470
Swimming pool	0.2210
Number of outdoor lights left on all night	0.2210
Number of windows in heated areas	0.2040
Total square footage.	0.2685
Total square footage. (Used for EIA data tables.)	0.2668
Total heated square footage	0.2460

With respect to the 21 features selected by comparing the MIC value and the total electricity consumption, we can formulate the predictive model by conducting regression trees. The coefficient of determination R^2 of the prediction would be calculated. Plus, the predicted electricity consumption will be compared with the original data. In addition, to improve the accuracy of prediction, we will obtain 10 features by filtering the existing 21 features with two methods. Firstly, we will increase the MIC value from 0.2 to 0.24 and redo selection by the computer. Secondly, we will objectively select the features by our experience. In addition, the K-fold cross validation is introduced to validate the predictive model. In result, 10 features are picked up from the 21 features. The value of R^2 of the two predictions will be compared with each other.

4. RESULTS AND ANALYSIS

4.1 Coefficient of Determination (R^2)

In statistics, the coefficient of determination, denoted as "R squared", provides a measure of how well observed outcomes are predicted by the model, based on the proportion of total variation of the outcomes from the model. The formula of R^2 is shown as following.

$$r^2 = \left[\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}\right]^2$$

where, the r is the correlation coefficient; x is the value in first set of data; y is the value in second set of data and n is the total number of values.

4.2 Results using Regression Trees

According to the 21 features selected by MIC, the predictive model is formulated by regression tree analysis. The predictive value is compared with the original data which is used to train the regression model, shown in Fig.1. Also, we use the model to predict the original data, the coefficient of determination reaches 0.99.

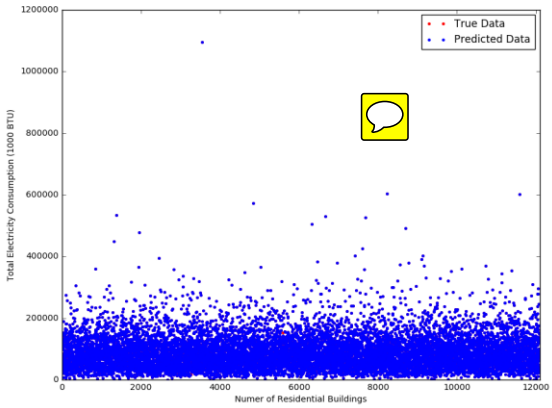


Figure 1: Comparison of true data and predicted data

4.3 Cross Validation

To test the accuracy of the prediction of the regression trees model, the K-fold cross validation is used, which gives an evaluation on the performance of regression trees. We split the data into ten subsets randomly, where nine of them are used to train the predictive model and the other one is used to validate the model. The same process is repeated for 10 times. We would calculate the R squared for each validation. The accuracy of the model is based on combination of all the R squared, shown as following.

$$RA = \frac{1}{n}\sum_{i=1}^{n} R_i$$

Since splitting the data is random, the R squared is various among different operations. The average value is approximately 0.345. Two of the validation with the R^2 of 0.37 is shown in Fig. 2.

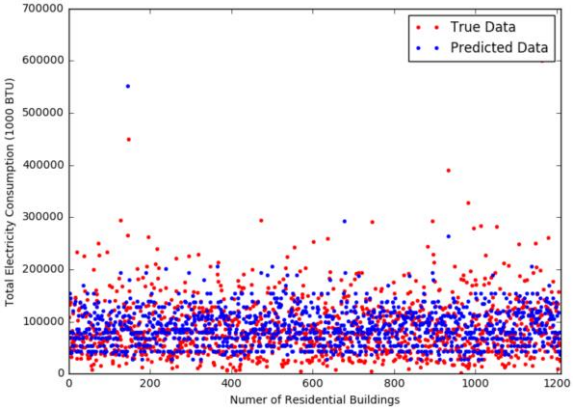
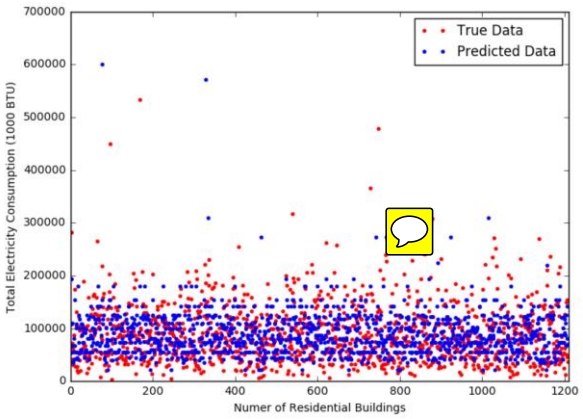


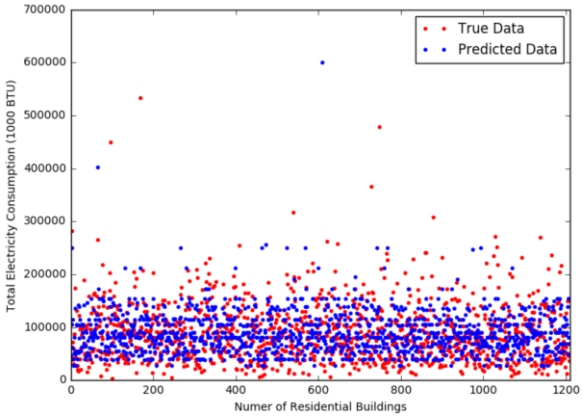
Figure 2: Comparison of predicted value and true value using K-fold cross validation

4.4 Calibration of Prediction

Ten features are selected by increasing the MIC value to 0.24. Also, we objectively select ten features with respect to our experience. To compare the two methods, splitting of the data sets follows the same subsets instead of the random splitting so that we can compare the R^2 with respect to the same subset for validation. The mean value of R^2 of them are 0.325 and 0.361 respectively. One of the validations of them is shown in Fig.3 with respect to the same validation subset.



a. Features selected by computer



b. Features selected by man

Figure 3: Comparison of two methods: predicted value and true value

5. SINGLE FEATURE

The selections obtained from MIC method are very interesting, so we want to study on these features. In this section, the analysis on each pair is presented. The "each pair" means we select one of features and the total electricity consumption as one pair. The following results demonstrates these features indeed have a strong relationship with the total electricity consumption, which also proves the effectiveness of MIC in selecting features.

5.1 Number of Rooms Heated

Firstly, we are interested in the relationship between the number of rooms heated and the total electricity consumption. The number of rooms ranges from 1 to 25 and -2 means not applicable. The number of rooms above 17 is ignored. The number of rooms above 16 is ignored as well, since there few data above 16 and the results will be more clear if we only focus on the range from 1 to 15, shown in Figure 4. It is obvious to see from the figure that the average electricity consumption increases as the number of rooms heated increases.

5.2 Number of Windows in Heated Areas

The feature "number of windows in heated areas" also shows a positive relationship with the total electricity consumption. In the RECS data sets, the number 0, 10, 20, 30, 41, 42, 50 and 60 denotes the number of widows 0, 1 or 2, 3 to 5, 6 to 9, 10 to 15, 16 to 19, 20 to 29 and 30 or more, respectively. In the Fig.5, we can easily notice that when the number of windows in heated areas increase, the electricity consumption increase as well. The reason could be that the windows are easily transferring data than the walls or floor.

5.3 Total Heated Square Footage

It is obvious to know there will be a close relationship between the total heated square footage and the total electricity consumption and it is very likely to be positive proportional. The results from Fig. 6 indicate that there exists an approximately positive proportional relationship between these two variables. The larger an area is, the more energy needs to be used for heating or cooling.

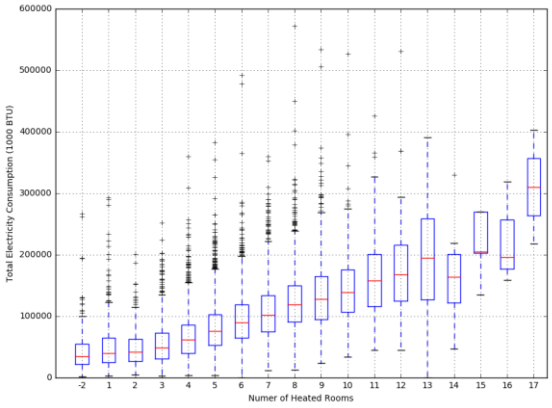


Figure 4: Box plot for number of rooms heated and total electricity consumption

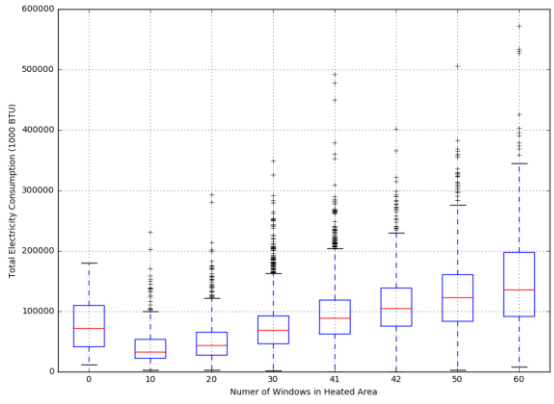


Figure 5: Box plot for number of windows in heated areas and total electricity consumption

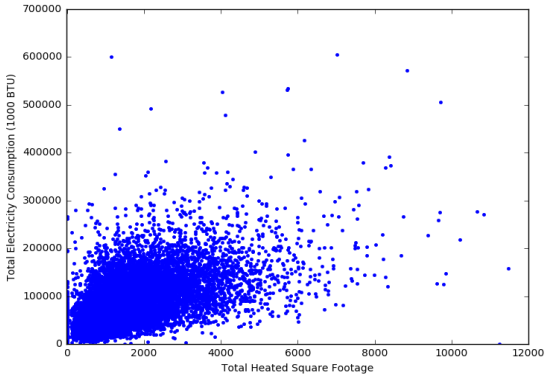


Figure 6: Total heated square footage and total electricity consumption

6. CONCLUSION

Based on the selections from MIC methods, the features such as type of housing unit, number of windows in heated areas, number of rooms heated, which are closely correlated with the total residential electricity consumption. In order to predict the total residential electricity consumption, the regression model is formulated and the K-fold cross validation is used to validate the model. By setting the MIC, 21 features are selected among more than 900 features and then are used for regression analysis. By implementing more strict rules for selecting features (increasing the threshold of MIC value to 0.24), only 10 features are selected. The R^2 of it is a bit lower than that for the 21 features. In addition, we objectively select 10 features from the existing 21 features based on our experience, the R^2 of it is higher than that for the 21 features and the 10 features selected by computer. Thus, the man-made selection results are a little bit more accurate. The reason could be that we objectively remove some similar features and the codes simply make choices based on MIC values. For example, the feature "number of bedrooms", for it is a subset of the feature "total number of rooms in house units", despite their MIC are higher than 0.24. It also indicates that in data analysis, the man-made selection of features is also important so that we can make accurate predictions with as less few features as possible.

In the future, we want to introduce a new method to filter different features and factors and compare its performance with MIC method.

7. ACKNOWLEDGMENTS

Our thanks to Professor Mario Berges and teaching assistant Henning Lange.

8. REFERENCES

[1] Title: World_energy_consumption. https://en.wikipedia.org/wiki/World_energy_consumption. Accessed: 2016- 12- 8.

[2] W.Y.Loh. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14-23, 2011

[3] Title: Decision tree learning. https://en.wikipedia.org/wiki/Decision_tree_learning. Accessed: 2016- 12- 7.

[4] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. Science, 334(6062):1518-1524, 2011.

[5] D.N. Reshef, Y.A. Reshef, M.M. Mitzenmacher, P.C. Sabeti. Equitability Analysis of the Maximal Information Coefficient with Comparisons. arXiv:1301.6314. 2013.