Building Energy Performance Analysis: A supervised learning approach

Eric Zhuang
New York University
Center for Urban Science and Progress

May 9th, 2020

Final Project Paper for Machine Learning for Cities

1. ABSTRACT

During the past decades, traditional energy benchmarking methods have been developed and then widely used by building energy managers. However, they all have some limitations since most metrics only consider a single aspect of the energy benchmarking, which may cause misleading results. (Gao, 2014) This paper aims to address the issue from a supervised learning perspective, predicting whether a building exceeds the national median energy benchmark point by inputting various building features. In summary, this research will talk about why building energy benchmarking is essential, what has been done in the past about machine learning implementation, methodologies of studying this topic, and the achieved results. Instead of one-hot encoding, this research uses the mean encoding to transform categorical variables. The benefit of doing this is that the model is performed on a lower-dimensional dataset. We observed that the model trained with the mean encoding method achieves 5% higher accuracy in predicting the target variables than one-hot encoding. Overall, the best-performed classification model uses random search cross validation with 3 folds and achieves a 67% of the accuracy score. As a preliminary research using classification methods for energy benchmarking, future work is needed to create a much more comprehensive analysis, including the development of accurate performance indices.

2. INTRODUCTION

According to the U.S. Energy Information Administration in 2017, the country's total energy consumption is about 98 quadrillion British thermal units (Btu), accounting for close to 20% of the world's energy consumption. (EIA) Among this vast amount of energy consumption, residential and commercial sectors together contribute nearly 40% of total U.S. energy use. (US

DOE) This happens for a number of reasons. The population increase in the United States drives up the number of residential, schools, commercial, and all other kinds of buildings. The increase in employment is a significant driver of new floor areas in the office and retail buildings along with extended building size in order to accommodate more people. (US DOE) Such changes would bring potential adverse impacts on the global environment as buildings would emit more carbon dioxide. Due to the increasing awareness of environmental issues, the idea of building energy efficiency has become a popular topic among architects, designers, policymakers, and government sectors. Under this topic, energy benchmarking is widely discussed and utilized because building owners and operators can determine whether they need to improve their energy performance compared to similar buildings. (Ruff, 2016) Currently, plenty of energy benchmarking programs have taken remarkable effects on building energy performance. (Curbed) According to the EPA, energy benchmarking could help buildings use roughly 2.4% less energy compared to those that don't use the benchmarking method. (EPA) In this research, we will discuss a supervised learning approach for energy benchmarking. The goal is to understand whether it is possible to predict the energy performance[1] given a building's primary function and characteristics.

3. LITERATURE REVIEW

Most of the energy benchmarking analysis is done using clustering algorithms. One research done by the University of Pennsylvania addresses building energy performance clustering in a new way. Other than clustering by the building types, they applied a more comprehensive approach that incorporated various building features such as occupancy, lighting

[1] Energy performance will be measured by whether the Source EUI of a building exceeds median benchmark metric.

condition, heating and cooling systems, etc. Based on this approach, they developed four clusters and calculated a pseudo-centroid for each cluster as a benchmark point. (Gao, 2014) Another research done by the National University of Singapore discovered that applying the K-shape cluster method first to the dataset could boost the performance of time series forecasting on energy demand. (Yang, 2017) Those are excellent resources from previous literature, which bring much inspiration to our research. Goldstein talks about the two most popular energy performance indices to evaluate the residential and commercial building's performance: Asset Ratings and Operational Ratings. (Goldstein, 2013) However, there is no machine learning method implemented; instead, the paper only addresses how the building energy performance is accessed using those indices. Similarly, we use Source Energy Usage Intensity (Source EUI), namely, how much energy a building consumes to operate as an index to benchmark a building's energy performance. Noneless, we approach the problem with a classification method.

4. METHODS

4.1 Data collection

The commercial building dataset is collected from CBECS commercial building survey data in 2016. This dataset has information of 6720 commercial buildings across the state and has over 1000 features. Due to the extensive amount of features, a comprehensive feature selection procedure is required.

4.2 Data preprocessing

The original dataset contains a large amount of N/A values. The first step is to remove the columns with over 1000 N/A values for the validity of the analysis. After that, we use the microdata codebook to get a sense of what each remaining column represents. Also, since most of the categorical and binary columns are represented by numbers, we map the corresponding category name to the columns for later use.

According to Energy Star, national median source EUI (energy usage intensity) is recommended as a benchmark to measure the energy performance of a building. (energy star) Since the data does not contain source EUI or the median, it is necessary to create those columns to perform further analysis. Source EUI is calculated by the following equation:

$$Total\ Energy\ Consumption\ (kBtu)\ /\ Total\ Floor\ Area\ (SQFT)$$

To think of this as a supervised classification problem, we need to generate two classes as target variables to train the model. Fortunately, with the calculated source EUI column for each surveyed building, we can compare them to the national median SourceEUI (energy star). For this part, we created a column representing "higher than median EUI" as 1 and "lower than median EUI" as 0. For example, EUI value for a college building is compared only with the median EUI of its category. The same approach applied to all other building categories.

4.3 Feature Selection

It is necessary to understand what affects the energy consumption of a building before feature selection. According to research from the University of Moratuwa, building energy consumption can be influenced by many factors. The five major factors are climate, building-related characteristics, building systems and services related characteristics, occupant related characteristics, socio-economic, and legal-related characteristics. (Silva, 2012)

We use their finds to conduct our initial feature selection process. The selected features are building activity, hours open per wee, SQFT, wall_material, main heating methods, main cooling methods, building shape, whether the building has an equal amount of glasses all sides, the climate of the building location, number of workers, window type, whether the building has automation system, whether the building has packaged heating units. With all the above features, we then perform the mean encoding for categorical features.

4.4 Mean encoding

One hot encoding is an efficient method when working with categorical variables; however, it increases the dimensionality of data. For this research, implementing one hot encoding would increase the dimensionality to over 50. Therefore, mean encoding comes into place to mitigate this concern. The essence of the mean encoding is that it is the label encoding directly related to the target. One the other hand, the label for one particular feature is calculated by:

$$Number\ of\ True\ target\ (1\ in\ this\ case)\ under\ that\ label\ /\ Total\ target\ under\ that\ label$$

After completing the mean encoding for each categorical feature, we map them to the original dataset as numerical columns. This encoding method not only reduces dimensionality and accelerates the learning process, but also makes it more convenient to understand the feature importance.

4.5 Baseline Model

We also generated a baseline model to compare with the model accuracy since it is crucial to understand whether the performance of a model is satisfying. (Jason Brownlee) We used the dummy classifier package from sklearn to develop such a baseline with the training and test datasets. We selected the "Uniform" strategy to generate prediction uniformly at random. (Sklearn)

4.6 Model Selection

As a classification problem, we chose two powerful ensemble algorithms: AdaBoost classifier and Random Forest classifier. Their results are compared to see which algorithm performs better in predicting energy benchmarking results. Ensemble methods are selected because we have a mixed type dataset with both numerical and categorical datasets. Tree-based models usually handle this type of dataset pretty well. Another reason is that ensemble learning uses multiple algorithms on a single dataset to achieve better predictive performance than any other constituent learning algorithm. (Wikipedia) Therefore, since the goal is to accurately predict whether the energy performance of a building exceeds its median threshold, we use a high-level machine learning method to start.

5. Results

5.1 Exploratory analysis

During the exploratory analysis, we did some data visualization to show a general picture of the commercial building characteristics. Among all the surveyed buildings, office, education, Non-refrigerated warehouse, inpatient health care and religious worship facilities are the top 5 commercial building types that report their building information. It is worth noting that due to the higher number of observations for those commercial building types, prediction results might be more accurate for them since they have more training samples.

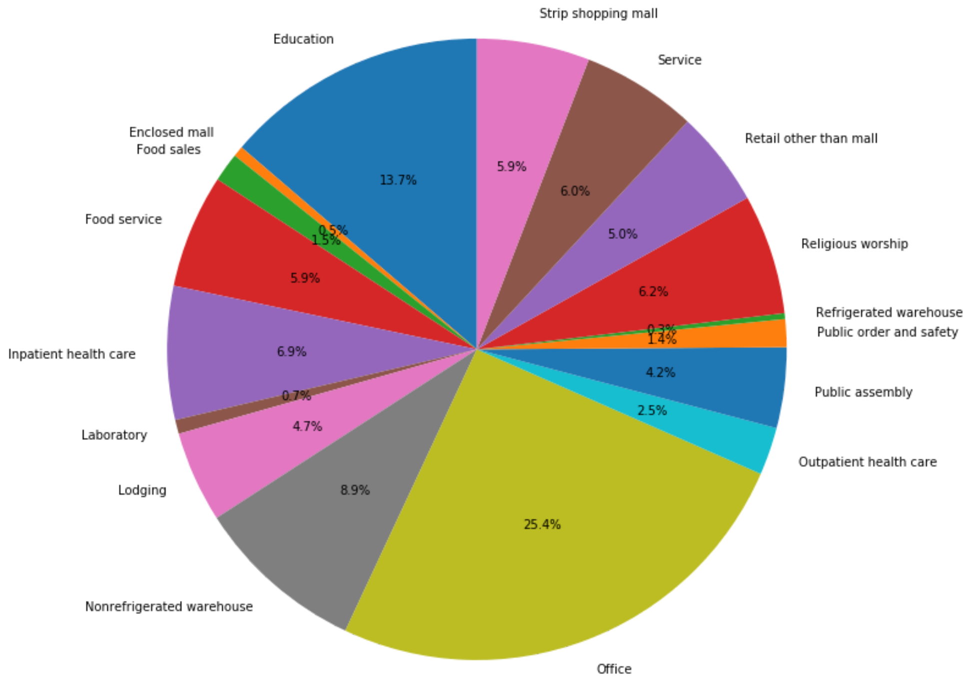

Figure 1: proportion of commercial building types

Figure 2 below indicates how many servers in average does different building types have. Commercial Office buildings undoubtedly have the most servers in use, followed by inpatient health care and education facilities. Those top commercial building types are very labor-intensive and generally require a considerable amount of space, thus require more servers than other types of commercial buildings.

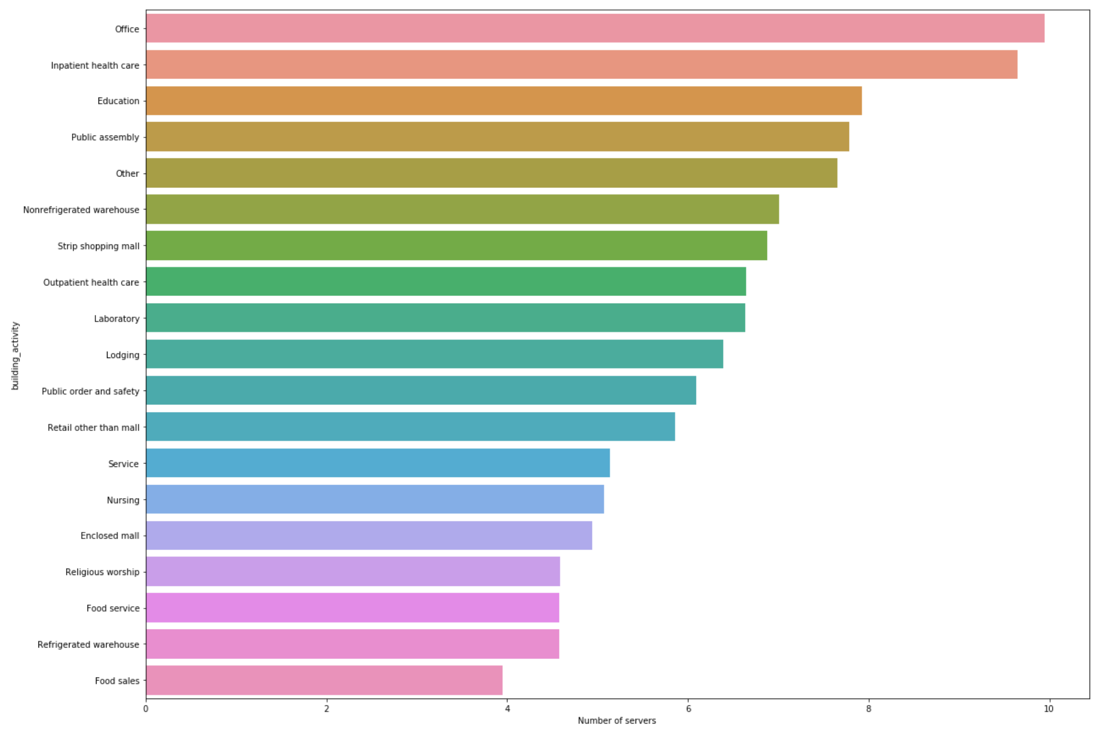

Figure 2: Number of servers by building type

We are also interested in looking into whether the building wall material impacts energy performance. The following figure shows buildings that use glass as wall materials have higher energy usage intensity than others. This is somewhat counter-intuitive since it is reasonable to guess glass or window walls have strong abilities to balance the indoor temperature and lighting conditions. But later we found that those buildings generally have more floor spaces and longer operating hours. Commercial office buildings can be a particular example in this case.

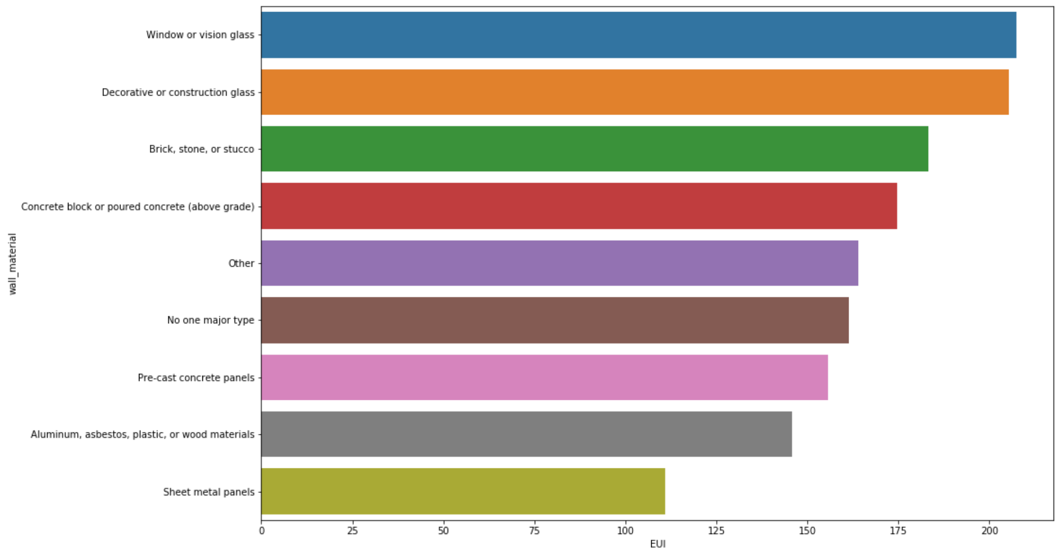

Figure 3: EUI by wall material

Figure 4 reveals which commercial building type has higher violation proportion, namely the number of buildings for each type exceeding the median sourceEUI. The laboratory category seems to be more likely to consume more energy than its median benchmark.

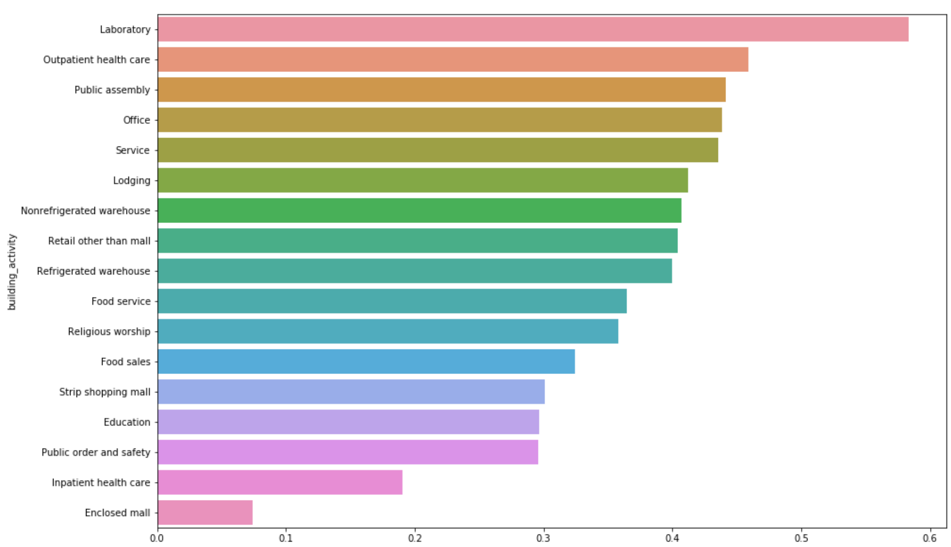

Figure 4: overrun median benchmark

5.2 Baseline Establishment

The baseline accuracy is computed as 0.51 using the original training and test dataset, this number is used to compare with the two ensemble models later on.

5.3 Supervised Learning

AdaBoost classifier

Adaptive boost model obtains the best overall out of sample accuracy of 66% when its n estimator number equals 500. The algorithm also generates feature importance, which is shown below in figure 5.

	variables	importance
1	SQFT	0.338
3	num_of_workers	0.178
14	building_activity_	0.15
5	Number of servers	0.128
0	hours_open_per_week	0.082

Figure 5: adaboost feature importance

Random Forest classifier

We used randomized search cross validation with 3 fold to estimate the best parameters for random forest classifier. The results are shown below: {'n_estimators': 1000, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}. After applying the best parameter to the model, we achieve an overall accuracy of 67%. Feature importance for each variable is also calculated as shown in figure 6.

	variables	importance
1	SQFT	0.161167
3	num_of_workers	0.140848
14	building_activity_	0.139243
0	hours_open_per_week	0.131545
10	MAINCL_	0.0647694

Figure 6: random forest feature importance

Both models have higher learning accuracy than the baseline model, but there are still rooms to improve. It is interesting to see that those two algorithms achieve almost the same accuracy results but generate different feature importance. Adaboost puts extensive weight on the building floor area, whereas random forest gives more averaged out importance to the variables. This could happen because of the nature of these two models where the Adaptive boost uses weak learners (namely, use one feature at a time), and the random forest grows a full tree. (The professional point)

Discussion and limitation

This research uses a novel approach to understand building energy benchmarks with classification methods. The goal is to predict whether a commercial building outperforms or underperforms the median energy usage intensity within its building category. Other than one-hot encoding, state of the art mean encoding is implemented in this research to reduce the dimensionality of the dataset. The random forest model achieves overall 67% of out of sample accuracy and generates the feature importance matrix. Is 67% a good enough result? Not necessarily because there is still much space to improve the algorithms by consulting domain experts about what features of a building truly impact the building energy performance. But as preliminary research using this method, we don't reject the importance of the results since at least it can provide a prescience for the portfolio manager to understand the building energy performance by inputting the building information. In future work, a more careful cross validation should also be conducted to ensure the model does not have an overfitting issue.

Citation

1. M. N. K. De Silva and Y. G. Sandanayake, BUILDING ENERGY CONSUMPTION FACTORS: A LITERATURE REVIEW AND FUTURE RESEARCH AGENDA. 28 – 30 June 2012
2. David B. Goldstein & Charles Eley, A classification of building energy performance indices. 6 December 2013
3. Kumar, Naresh. “Difference between Random Forest and AdaBoost in Machine Learning.” *Difference between Random Forest and AdaBoost in Machine Learning*, theprofessionalspoint.blogspot.com/2019/03/difference-between-random-forest-and.html.
4. Energy star portfolio manager, Technical Manual, 2018. https://portfoliomanager.energystar.gov/pdf/reference/US%20National%20Median%20Table.pdf
5. Energy Efficiency Trends in Residential and Commercial Buildings, United States Department of Energy. October 2008 https://www1.eere.energy.gov/buildings/publications/pdfs/corporate/bt_stateindustry.pdf
6. Roy, Baijayanta. “All about Categorical Variable Encoding.” *Medium*, Towards Data Science, 25 Apr. 2020, towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02.
7. “Sklearn.dummy.DummyClassifier¶.” *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html.
8. “Ensemble Learning.” *Wikipedia*, Wikimedia Foundation, 9 May 2020, en.wikipedia.org/wiki/Ensemble_learning.
9. Brownlee, Jason. “How To Know If Your Machine Learning Model Has Good Performance.” *Machine Learning Mastery*, 15 Mar. 2018, machinelearningmastery.com/how-to-know-if-your-machine-learning-model-has-good-performance/.
10. Rebecca Ruff, Commercial Building Energy Benchmarking and Disclosure Policies in the United States. August 2016
11. “Building Energy Use Benchmarking.” *Energy.gov*, www.energy.gov/eere/slsc/building-energy-use-benchmarking.
12. T. Nikolaou, Review on methodologies for energy benchmarking, rating and classification of buildings, 29 Jun 2011.
13. Junjing Yang, Chao Ning, Fan Zhang, Chirag Deb, K-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, March 2017.