Predicting Energy Usage of School Buildings

Daniel Sambor, Rohith Desikan, Vikhyat Chaudhry,

CS 229: Machine Learning, Stanford University, Fall 2016

Introduction

Providing clean, abundant, and reliable energy is arguably one of the most important issues facing society in the 21st Century. Of particular significance is the building sector, which consumes over three quarters of total electricity in the US and 41% of primary energy in general (Marasco et al).

Specifically in California as more renewable energy projects are installed, the state is faced with a critical problem on the grid known as the "duck curve." The grid often cannot handle the increasing influx of solar energy during the middle of the day and utilities must curtail it or risk damaging grid infrastructure.

To reduce energy consumption and minimize the environmental impact of buildings, measuring and understanding patterns in this energy use is paramount. In terms of energy efficient design, projects can be separated between new construction and existing buildings. In order to outfit a new building with renewable energy, it is important to predict the future energy use based on design parameters in order to match renewable generation with the load. For existing buildings, the goal is to retrofit with the optimal mix of technologies to gain the best efficiency improvements at the lowest cost.

Within the building sector, school campuses are especially interesting to solving these energy issues as they use most of their energy during the middle of the day, or peak times on the grid. Thus, an entire school system could provide benefits to the grid by instituting bulk energy efficiency projects or designing a new campus capable of generating and consuming their own power on site.

As population grows, schools are also faced with the decision whether to build a new campus or retrofit existing buildings. By predicting future energy use for new buildings or the results of retrofitting their current campus, the goal is to make this decision simpler and easier for school officials to make.

Data and Methods

This analysis will be performed using a recently published data set from the California Energy Commission consisting of historical energy use for over 14,000 schools in California. The data is provided open source, mandated by Proposition 39, a bill to fund clean energy projects for schools. Given the recent availability of this data and the results of a thorough literature review, to the knowledge of the authors there has not been an analysis of this scale on a real data set for energy predictions in schools.

The data set is comprised of several features per school including building area (ft^2), annual energy use (kWh), peak demand (kW), and natural gas use (therms). Additionally, for each school the data provides retrofit options that the school implemented, often more than one per school and the resulting energy savings to investment ratio (SIR) per retrofit.

To enrich the data set further, a geographic information systems (GIS) analysis was performed to obtain climate and demographic information for each school or school area. A dataset of mean annual temperatures for California was obtained from the PRISM Climate Group, and overlaid with each school as shown below in Figure 1. Mean temperatures were then assigned to each school and used to calculate total degree-days for each location. A degree-day quantifies the amount of heating and cooling necessary for a building in a specific area; thus, this feature will aid in modeling the climatic impact on energy use.

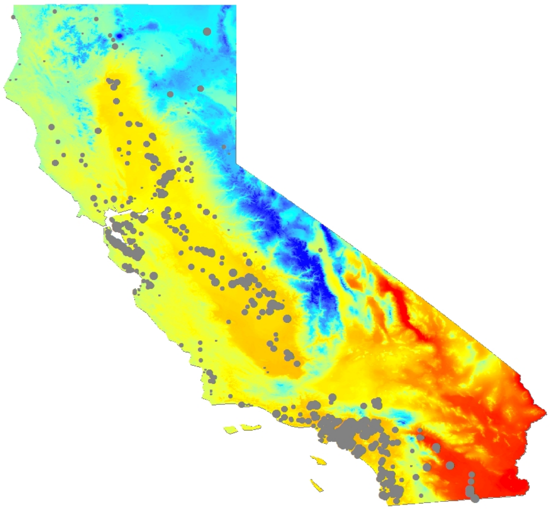

Figure 1: Schools in California with Mean Temperatures of overlaying region

Finally, demographic census data was obtained for each school zip code area, to account for population and spatial factors (US Census, TIGER Data). The features included median age, household income, and total housing units, among others. These are displayed in Table 1. Income was used to model whether a school in a wealthier area was more likely to have funding to undergo a retrofit. Median age serves to model potential emphasis placed on schooling, meaning that an area with a higher median age may not invest as much in schools.

A literature review was performed, which showed that predicting time series of energy use is quite common in the literature, with one of the most widely used and accurate methods being Artificial Neural Networks (ANN) (Mocanu et al 2014). Other methods for energy prediction, based on building energy reviews, include linear regression and support vector machines (SVM) (Foucquier et al, 2013; Tso et al, 2007).

Table 1: Features List Used

Feature	Unit
Site Conditioned Area	ft^2
Average Peak Demand	kW
Electricity Density	kWh/ ft^2
Natural Gas Density	$therms/ft^2$
Energy Use Intensity (EUI)	$kBTU/ft^2/yr$
Energy Efficiency Measure Category (EEM)	-
EEM Estimated Annual Electric Savings	kWh
Total Combined Energy Costs	$/yr
EEM Estimated Measure Cost	$
EEM Savings to Investment Ratio	-
Total Degree Days	-
Total Housing Units	-
Average Age of Housing Structure	-
Median Gross Rent	$
Median Value of Housing Units	$
Median Household Income from last year	$
Median Age of Population	-

In terms of retrofit analysis, which is a less-studied aspect of buildings, some form of logistic regression is the most common method of determining classes of data. Kontokoska used logistic regression to predict whether a building would decide to choose to retrofit in general (Kontokosta 2016). The model resulted in 81% test accuracy for determining whether a retrofit would be chosen. In another study, Kontokosta applied decision lists to determine which retrofit should be installed based on access to questionnaire data (Kontokosta et al 2016). Although a few studies have been done applying machine learning to predict the best retrofit for schools, there is a scarcity of studies that predict which retrofits would have positive/most positive SIR values.

Thus given a set of input features provided from a school superintendent who desires an analysis of whether to build a new campus or retrofit existing buildings, the goal is to output the energy costs of a new building or the energy savings per investment for the modeled optimal retrofit options.

Analysis: New Buildings

Given that several statistical learning methods have been reviewed in the literature for building energy prediction, different models were tested to determine their applicability for school buildings. Based on a thorough literature review, linear regression, SVM and artificial neural networks were chosen for predicting annual energy use for new buildings.

Polynomial Regression

The first approach was to implement a simplified linear regression model after performing holdout cross validation. Thus, to train the model, 80% of the data was chosen at random, leaving 20% for testing. A polynomial regression was fitted to the data and cross-validation used to obtain the optimal degree polynomial. To visualize the data, regression for two features, the site conditioned area (in square feet) and annual electricity use (kWh/yr) were plotted with the resulting fit. This feature subset was determined because they showed the greatest statistical significance and correlation amongst the features for the analysis. In essence, they were the two most important features for energy consumption, which makes sense.

The rest of the new building analysis was performed using all features shown in Table 1. Thus, the quadratic fit was chosen as the best model given the lowest root-mean square error (RMSE). It was also determined to be optimal based upon an analysis of the residuals. The results of the polynomial regression are shown in Figure 2.

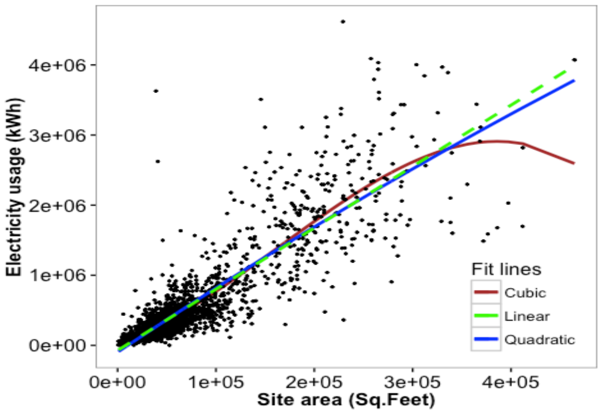

Figure 2: Polynomial Regression (multiple features, linear (green), quadratic (blue), and cubic (red))

Support Vector Machines (SVM)

Using SVM's as a method of building energy consumption regression has recently become popular in energy modeling literature (Ahmad et al 2014). Ahmad reviews several studies that have used SVM's for energy prediction for many applications from short-term to long-term forecasting. The output of the SVM is shown in Figure 3.

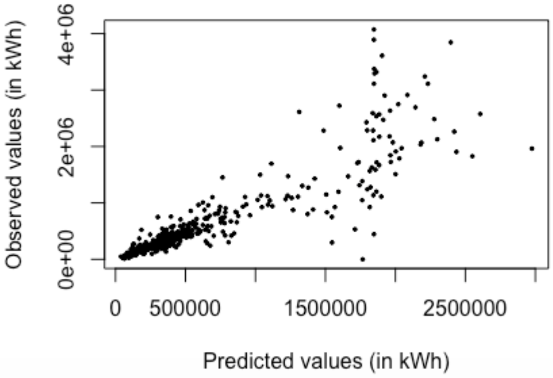

Figure 3: SVM Regression

Artificial Neural Networks (ANN)

ANN is based on the fact that there are hidden layers, called neurons, which take in a linear or non-linear combination of input features and applies weights to them through interconnections (or synapses). Neurons have activation functions, which process these weighted linear or non-linear combinations and give out a predicted value. In this study, the activation functions are sigmoidal and we use back propagation with Batch Gradient descent algorithm for minimizing the cost function. This study uses a 3-layer network and assumes the cost function to be convex in nature. As explained in the literature review, ANN is one of the most powerful and widely used methods for predicting building energy consumption; thus, any building analysis should include it within the overall strategy (Mocanu et al, 2014). The output of the ANN is shown in Figure 4.

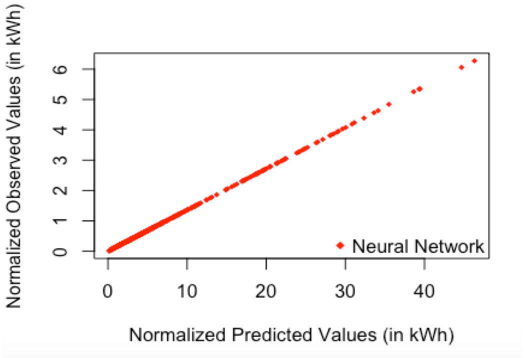

Figure 4: ANN Prediction vs. Observed

To compare the results of the different methods, root mean square loss was used. The final results displaying the RMSE of the three strategies for predicting annual electricity are shown below in Table 2.

Table 2: Energy Prediction Results on Test Set

Model	RMSE
Neural Network	1.40%
Polynomial	39.0%
SVM	55.8%

As expected from the literature, the neural network resulted in the best fit to the data, with polynomial regression and SVM's displaying far less accuracy. While ANN is known to over-fit, increasing bias by reducing the number of features still resulted in superior results for neural nets.

Analysis: Existing Buildings

Modeling retrofits for existing buildings is the natural complement to predicting energy usage for new construction. For this analysis, the goal is to classify a school, based on its input features such as budget, size and electricity use, into a retrofit class that would provide the maximum value for money. As previously mentioned, the data set included which retrofits schools have already installed (sometimes more than one) and their resulting savings; thus, the models were trained on 80% of this data and tested on the rest. The different classes of retrofits possible are shown in Table 3. The two methods chosen for analysis included Softmax and Artificial Neural Networks.

Softmax Model

Given that there are numerous classes of data, classification using a Softmax model based on a multinomial distribution is logical method for the analysis. As the literature used logistic regression to classify whether a retrofit was effective or not, this would classify the exact retrofit that a school should implement. There are nine large classes evident in the data set, which were condensed from 43 and the model was run for all classes. However, as shown in the second column of Table 3, most are rarely implemented and thus not accounted for at all in the prediction model. The model was run using a limited set of classes, though this did not vary the final results significantly.

Table 3: Retrofit Classes in Dataset and Predicted Results

Retrofit Class	Training Data Composition (%)	Model Predictions (%)
Lighting	70.0	93.0
HVAC	27.0	6.1
Envelope	0.21	0.9
Plug Loads	0.96	0
Electrical	0.12	0
Hot Water	0.18	0
Pumps/Motors	0.51	0
Energy Storage	0.3	0
Kitchen	1.0	0

Thus, while lighting accounts for a majority of previously installed retrofits in the data, the model recommends an even larger percentage of lighting retrofits. Intuitively this makes sense as LED lighting technology has become so efficient that it makes economic sense for most areas, especially those regions like California without significant HVAC loads.

The Softmax model was run with several different features. Ultimately, the last feature below in Figure 5, the cost of the retrofit proved to be most important as expected. A Principal Components Analysis was also performed which also showed that retrofit cost along with household income of the neighboring was also important. Thus, wealthier areas may be more likely to invest in the school system.

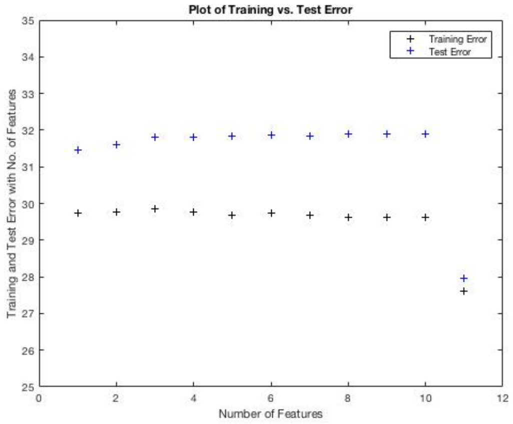

Figure 5: Test vs Train Error for Softmax

Artificial Neural Networks (ANN)

ANN was used again for classification of the various retrofits with the predictors being Savings to investment ratio (SIR) and Annual Energy Consumption (kWh). The back propagation and gradient descent method were used for this classification. The network was a 3-layer network and the errors at each layer were used for adjusting the activity (z) and weights (w) in the back propagation method. These adjusted weights were then used to minimize the cost function. The classification of various retrofits is shown in Figure 5. Most of the savings are dominated by the lighting retrofit (blue cluster) followed by HVAC (green cluster).

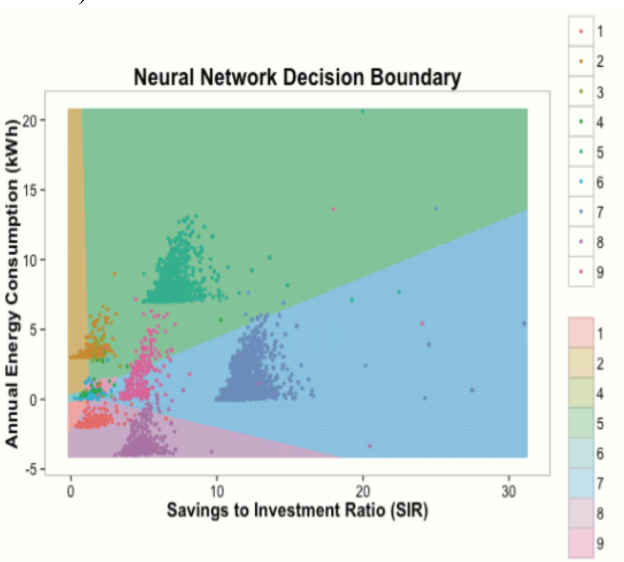

Figure 5: ANN Retrofit Classification

Overall, the classification error on the test set is shown in Table 4.

Table 4: Retrofit Classification Error on Test Set

Model	Test Error
Neural Network	**3.36%**
Softmax	**27.9%**

Again, neural networks proved to be the better method. Whereas the Softmax model was only trained using one retrofit option per school (the one with the best SIR) in the training set, ANN was able to model all retrofits that a school had implemented. Thus, ANN is a more powerful approach that is able to model the relationships between different retrofit classes for the same school, as well as different schools. Additionally, given that the Softmax regression accounts for just one retrofit per school, it could be used to recommend the first installation and ANN could be used to suggest the second and third most important ones.

Conclusion

Thus, given a specific set of design features available, the goal of this analysis is to allow a school official to then receive an estimate of energy consumption if a new campus were built or the best retrofits that lead to the largest savings per investment if they were to remodel the existing campus.

Neural networks were used and in both cases outperformed the other models in predicting the values of the test set. This coincides with the clear emphasis in the literature that neural nets are the best method for energy prediction. Overall these models are valuable for quick analysis and site screening given that many engineering models in the industry are too complex to allow for a single school official to use.

Future Work

These models can also be extrapolated beyond California to predict for other areas of the country. Given California's mild climate, there may be a more significant link between weather and energy use for the rest of the US.

Other algorithms can also be tested including falling rule lists and decision trees, which have been demonstrated in the literature for step-wise installation of retrofits (Marasco et al, 2016). While these models have good predictive power, more work is possible on the retrofit analysis.

References

[1] E. Mocanu, P. Nguyen, M. Gibescu, and W. Kling, “Comparison of Machine Learning Methods for Estimating Energy Consumption in Buildings,” Probabilistic Methods Applied to Power Systems, Conference, 2014.

[2] A. Foucquier, S. Robert, F. Suard, L. Stphan, and A. Jay, “State of the art in building modelling and energy performances prediction: A review,” Renewable and Sustainable Energy Reviews, vol. 23, no. 0, pp. 272 – 288, 2013

[3] D. E. Marasco and C. E. Kontokosta, “Applications of machine learning methods to identifying and predicting building retrofit opportunities,” vol. 128, pp. 431–441, 2016.

[4] C. E. Kontokosta, “Modeling the energy retrofit decision in commercial office buildings,” vol. 131, pp. 1–20, 2016.

[5] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah, H. A. Rahman, F. Hussin, H. Abdullah, and R. Saidur, “A review on applications of ANN and SVM for building electrical energy consumption forecasting,” *Renew. Sustain. Energy Rev.*, vol. 33, pp. 102–109, 2014.

Optional References:

[6] H. Kim, A. Stumpf, and W. Kim, “Automation in Construction Analysis of an energy ef fi cient building design through data mining approach,” *Autom. Constr.*, vol. 20, no. 1, pp. 37–43, 2011.

[7] Y. Heo, R. Choudhary, and G. A. Augenbroe, “Energy and Buildings,” vol. 47, pp. 550–560, 2012.

[8] S. Robert, L. Ste, and A. Jay, “State of the art in building modelling and energy performances prediction : A review,” vol. 23, pp. 272–288, 2013.

[8] J. Zico, J. Ferreira, A. Jr, A. Press, J. Z. Kolter, and J. F. Jr, “A large-scale study on predicting and contextualizing building energy usage," 2016.

[9] H. Zhao and F. Magoulès, “A review on the prediction of building energy consumption,” vol. 16, pp. 3586–3592, 2012.

[10] G. K. F. T. Ã and K. K. W. Yau, “Predicting electricity energy consumption : A comparison of regression analysis , decision tree and neural networks,” vol. 32, pp. 1761–1768, 2007.