Energy Efficiency and Machine Learning: Understanding Technology Adoption Decisions

Aramayis Dallakyan, Texas A&M University, 979-220-7274, dallakyan1988@tamu.edu
Reid Stevens, Texas A&M University, 979-847-5805, stevens@tamu.edu

Overview

The energy efficiency gap, the difference between the cost-minimizing level of energy efficiency investment and the observed level of energy efficiency investment, has attracted significant attention from academic researchers (Allcott and Greenstone, 2012). We contribute to this literature by applying machine learning models to data on energy efficiency technology adoption by small- to mid-size manufacturing firms in the United States that participated in energy efficiency audits sponsored by the government. First, we determine whether machine learning models would better explain technology adoption decisions than the linear models typically used in the literature. We then evaluate whether variables have been overlooked by researchers in the past. Rather than motivating variable selection with economic theory, we select variables for our model of technology adoption using out-of-sample fit. Our results demonstrate that nonlinear machine learning models significantly improve the fit technology adoption models, especially when we use algorithms to select the independent variables. This research suggests that, to some extent, energy efficiency adoption decisions can be better understood using machine learning models that exploit all data available to researachers, rather than the narrow set of variables typically considered by researchers.

Methods

We begin by replicating the payback, cost-savings, and price-quantity technology adoption models from the seminal Anderson and Newell (2004) paper and extending those results using energy audit data from the Department of Energy's Industrial Assesment Centers (IAC) program through 2017. We then determine whether machine learning models would provide a better fit of data than Anderson and Newell's logit models using the same independent variables. We use the least absolute shrinkage and selection operator (LASSO), which is similar to linear regression, but selects only the most useful variables for forecasting. We also use some of the well-known linear classifier models, including Support Vector Machine (SVM), as well nonlinear models, including Gaussian processes for classification, tree based and ensemble methods.

The SVM is appropriate when the classes are not linearly separable. It produces nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature (independent variable) space. The idea behind Gaussian process classification is to place a Gaussian Process (GP) prior over the unknown function f(x) and then squeeze this function through the logistic function to get prior on pr(y=+1|x). Tree based methods partition the feature space into box-shapes regions. Then in each box the model makes the dependant variable average as different as possible. The boxes are defined by the splitting rules, which are related to each other through a binary tree. We also use ensemble methods, in which we combine a group of models to build a prediction model. The usual two families of ensemble methods are averaging methods suchas Bagging and Random forest , where the idea is to build several estimators and then average the predictions. We also use boosting methods in which the combined estimator is constructed sequentially to reduce the bias of the estimator.

Next, we determine whether other variables collected by the energy efficiency auditors are useful in explaining technology adoption. While some research has identified additional variables that can be included in models of technology adoption (Blass, et al., 2014), this topic is ripe for exploration with these statistical learning techniques. Additional indpendent variables include past energy use, auditor experience, and auditor eduction level. We estimate a variety of variable selection models to determine which variables are most useful in predicting technology adoption. We use logistic regression with L_1 and elastic-net penalization for subset selection, as well tree based methods to find the optimal tree size by controlling cross-entropy or Gini index. L_1 penalization shrinks the parameter space and by selecting "important" variables, where importance is determined by out-of-sample fit. The elastic-net selection is the compromise between L_1 and L_2 penalization. It selects the variables, similar to the LASSO model, and simultaneously, like ridge regression, shrinks the coefficients of correlated predictors. In decision tree models, the optimal tree size is the tuning parameter which is chosen from the dta by minimizing cost complexity criterion.

For each of these models, we use cross-validation to ensure the models are not overfit. This involves training the tuning parameters in the choosen models so that the models will use only those variables and values of tuning parameters which increase accuracy of out-of-sample predictions.

Results

First, we compare the out-of-sample forecast accuracy of the Anderson and Newell (2004) model to our machine learning model (random forest) fit using data ending in 2001. The Anderson and Newell model accurately predicts the adoption decision (a binary variable) in about 50 percent of cases, while the machine learning model accurately predicts the adoption decision in about 60 percent of cases. Much of the improvement in out-of-sample forecast accuracy is a result of the non-linearity of the machine learning model. Evidently, the energy efficiency adoption decision is more accurately modelled using a non-linear machine learning model.

We explore the accuracy of different machine learning models over a larger data set and present these results as a horse race between models. Each model is fit on data ending on December 31, 2015. Each of the models then predict technology adoption decisions for every audit in 2016 and 2017. The accuracy of these predictions is assessed in several ways, including most correct predictions, fewest false positives, fewest false negatives, and minimum error. Our initial research has shown that the nonlinear machine learning models offer a significant improvement over the logistic regressions. Morevever, we highlight several overlooked variables, including auditor experience, that have predictive power for energy efficiency adoption decisions. These results demonstrate the importance of nonlinearities in energy efficiency decision making, as well as other variables not typically included in energy efficiency models.

Conclusions

New statistical learning methods have the potential to improve our understanding of the energy economics field. Using energy efficiency as a case study, we apply machine learning methods to better understand the energy efficiency gap, that firms and individuals underinvest in energy efficiency. These models allow for nonlinearities and also allow for the inclusion of independent variables in the model that are often overlooked by researchers. These models improve our ability to forecast energy efficiency technology adoption decisions out-of-sample relative to the logistic models typically used in the field. This paper demonstrates the value of machine learning techniques to the energy efficiency gap, as well as other topics in energy economics.

References

Anderson, Soren T., and Richard G. Newell. (2004). "Information programs for technology adoption: the case of energy-efficiency audits." Resource and Energy Economics 26.1: 27-50.

Allcott, Hunt, and Michael Greenstone. "Is there an energy efficiency gap?." The Journal of Economic Perspectives 26.1 (2012): 3-28.

Blass, Vered, Charles J. Corbett, Magali A. Delmas, and Suresh Muthulingam. (2014). “Top management and the adoption of energy efficiency practices: Evidence from small and medium-sized manufacturing firms in the US.” Energy, 65, 560-571.