Predicting cost savings of households adopting distributed energy technology

Trisha Jani

Abstract

The purpose of this project is to predict how much money households can save if they adopt a distributed energy technology, specifically a home battery, solar panel, or both. Using real smart meter data from around 80,000 households in northern California, we extracted a set of about 200 features describing each household. From here, we implemented a combination of hand selection and forward feature selection to narrow our set of features down to 5. We then performed regression analysis to create a model that predicts how much money a household can save under the three different adoption schemes (solar panel, battery, or both). We also ranked households based on how much money they would save and separated them into four categories to build a four-class classifier. For all three adoption schemes, the linear SVM classifier had over 80% accuracy. Our results suggest that it is indeed possible for households to use features derived from smart meter data to inform decisions regarding adoption of distributed energy technologies.

I. Introduction

The energy industry is witnessing a rise in the use of data analytics to better understand consumers' relationship with energy. Whereas data analytics and machine learning methods have been used by utilities to improve their grid management, these techniques are now being applied behind the meter to characterize the way that end users consume energy. Building models from an individual consumer's time-series data will improve understanding of how much flexibility and variability the grid can offer, which may lead to improved future grid development.

Policy makers, energy utility providers, and the private sector all benefit from understanding the capabilities and merits of a data-driven energy efficiency approach. In particular, insights drawn from consumption data can be effectively harnessed to drive ambitious policy changes that lead to economic growth. Being able to characterize consumers on a fine scale is an essential to improving current day energy management and efficiency programs.

In this project, we used the raw smart meter data of 80,000 households in northern California provided by PG&E to perform data analysis and predictive modeling to characterize households that can save money by using a distributed energy (DE) technology. Specifically, given this raw data, we wanted to figure out how much money households could save if they were to adopt a home battery (such as the Tesla Powerwall), a solar panel, or both.

II. Dataset & Features

i. Time series data

The household energy consumption data comes from about 80,000 smart meters that capture consumption on hourly intervals from California residences over a period spanning from August 2010 to July 2011. The dataset was provided by the Pacific Gas and Electric Company (PG&E) to the Stanford Sustainable Systems Lab (SSSL). Households with very low annual consumption and a high amount of zero reading were excluded from the dataset.

ii. Savings data

Since this project intends to predicts savings households would see if they were to adopt DE technology, we did not have actual cost savings data. The SSSL ran a simulation to predict the net present value (NPV) of the savings of a household that adopts the technology. They assumed that households installed a 7 kWh bat-

tery and 5 kW residential solar panel, and used the technology in an ideal manner. They estimated savings for three adoption options: exclusively adopting a battery, exclusively adopting a panel, and adopting both a battery and panel. We used the simulated savings data as true savings in this project.

iii. Feature extraction

The Stanford Sustainable Systems Lab has developed a platform called VISDOM (Visualization and Insight System for Demand Operations and Management) that extracts features given the raw smart meter data, weather data, geographical information and census information as input. VISDOM then generates a set of roughly 200 features describing each customer. These features are categorical or numerical data that can be calculated or estimated from the raw data inputs.

Several features are determined through basic statistical techniques; these include the mean, median, variance, maximum, and minimum of consumer demand. Additional features include these statistics specified for certain time periods or ranges, such as variance of demand during the winter, mean consumption for the 12th hour of the day, and the ratio of night to day consumption in the month of August. Furthermore, VISDOM implements statistical techniques to estimates features that relate consumption to external factors such as weather.

iv. Exploratory data analysis

We first performed basic exploratory data analysis to get a better idea of how our set of nearly 200 features map into determining a household's annual savings. Since many features generated by VISDOM are related, and it is not practical to build a interpretable model with 200 features, we wanted to get a better understanding of which features could play an important role for different adoption strategies.

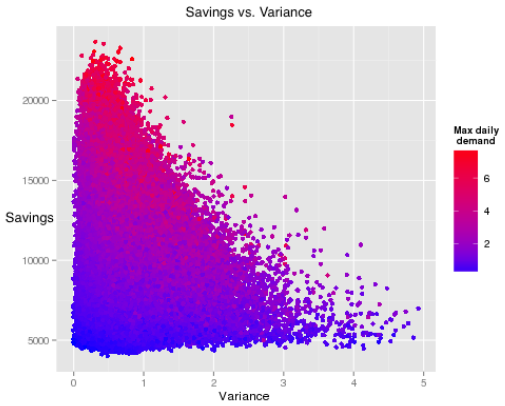

Figure 1: *Savings vs. variance under battery and solar panel adoption*

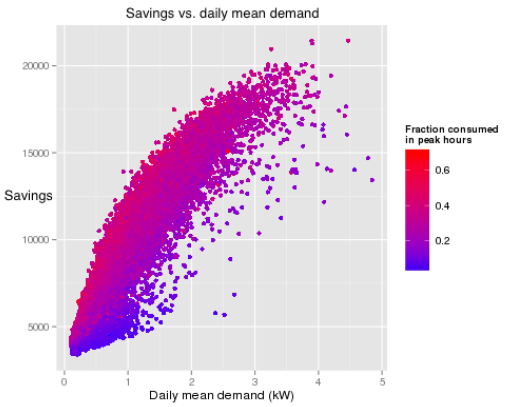

Figure 2: *Savings vs. daily mean consumption under battery adoption*

Figure 1 shows the relationship between cost savings and annual energy consumption variance when households adopt a solar panel and battery. The graph suggests that households with high variance will likely have lower cost savings, while savings of households with lower variance will have a wider range. Households that consistently have high daily demands will likely save the most. This graph suggests that features that capture variance and max demand could be important in developing a predictive cost model.

Figure 2 shows the relationship between cost savings and daily mean demand when households adopt a solar panel. The graph suggests that households with higher demands will save more money than households with lower demands. The relationship between savings and mean demand appears to be linear.

III. Methods

i. Feature selection

Since it is not practical to build an interpretable model with around 200 features, we first needed to trim our set of features to a more manageable size. Using a large number of features may lead to a more complex and less interpretable model. We reduced our feature set in stages. First, we looked at our set of features and manually got rid of features we though would be repetitive or irrelevant in developing our predictive model. For example, we thought it was unlikely that the mean energy consumption at midnight throughout the year would be a helpful feature in determining savings. In this manner, we narrowed down our feature set to about 70 features.

Next, we removed features we thought would be repetitive or correlated. For example a feature that marks how much energy households consume during the daytime is probably correlated to a feature that marks how much energy households consume during peak pricing hours, since peak pricing hours are during the day from 1 p.m. to 7 p.m. This further reduced our feature set to roughly 15 features.

At this point, we decided to implement forward feature selection to narrow down to a final set of features. Specifically, we performed a forward search for the best subset of features that predict cost savings under a linear regression model, using the adjusted-r^2 criteria defined as follows: $r^2_{adj} = 1 - \frac{(1-r^2)(n-1)}{n-k-1}$. We

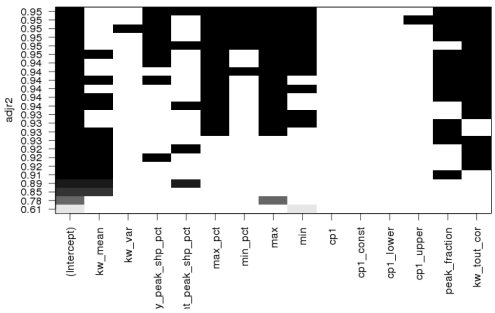

Figure 3: *Best subset of features under adjusted-r^2 criteria for predicting savings under dual adoption*

decided to use this criteria since it would penalize the addition of additional features that do not improve the regression model and could lead to overfit.

Figure 3 provides a graphical representation of the features that are used in each model to predict savings under battery adoption. We performed a similar analysis to predict savings under solar adoption and dual adoption. Given these results, we chose five features that were common amongst all three adoption schemes and interpretable in an energy consumption context. We though it was particularly important that the final features we use to build our models be very interpretable, since everyday people should be able to decide whether or not purchasing some combination of a solar panel and battery is right for them based on these features. The five features chosen were: mean of daily max consumption, mean of daily mean consumption, mean of consumption used during the peak hours of 1 p.m. to 7 p.m., consumption correlation with temperature, and maximum hourly consumption as a percentage of total daily consumption. Throughout the figures, these features are abbreviated as max, mean, peak_fraction, cp1_upper, and max_pct, respectively.

Table 1: *RMSE as a percent of SD*

Model type	Solar	Battery	Both
Linear regression	26	42	25
Boosted LR	27	43	25
Regression Tree	52	64	55

ii. Regression Analysis

We first attempted to build a predictive model that would estimate the exact savings a household could see under the three different technology adoption schemes. We built all of our models using the Classification And REgression Training (CARET) package in R. Each model was trained using 5-fold cross validation, repeated 5 times. Our training set had 63104 examples and our test set had 15777 examples. As a preprocessing step, we centered and scaled all of our data before implementing any models.

Models we implemented include a linear model, generalized linear model with boosting, and CART regression tree, each using the 5 features that were selected using forward selection. Our model's results on test data are summarized in Table 1. We found that basic linear regression was the strongest performer and the regression tree was the weakest performer.

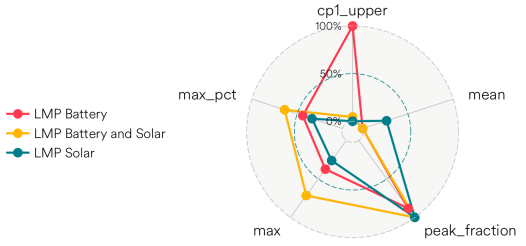

Figure 4: *Best subset of features under adjusted-r^2 criteria for predicting savings under dual adoption*

A linear model with the five chosen features is highly interpretable. We found that different features play different roles in determining savings based on the technology adoption scheme. First, we saw that all the weights for the features in our linear model were positive; this makes sense since in all cases, as the value of a features increases, the potential savings should also increase. We also found that in the case of solar adoption and dual adoption, the feature indicating the fraction of energy consumption during the peak hours of 1 p.m. to 7 p.m. is the most important in determining savings. This makes sense, since a solar panel would provide energy during those hours. Lastly, we found that the correlation between energy consumption and temperature is the most important feature in determining savings under battery adoption. This suggests that people who have high energy consumption during the summer, presumably by running AC during the daytime when energy prices are high, could save money by using a battery to store energy when electricity is cheaper.

iii. Classification Analysis

Our regression analysis does an okay job predicting savings using five features. In a practical context, however, households are not interested in the exact amount of money the would save if they were to adopt some combination of technology. Rather, they are more interested in knowing whether they would be on the higher end of the savings spectrum. This information will allow them to determine if they should adopt certain technology.

Table 2: *Group labels: Savings range for each adoption scheme*

Group	Battery	Solar Panel	Both
1	17-690	3,385-6,370	4,025-7,810
2	690-1,027	6,370-8,072	7,810-10,012
3	1,027-1,374	8,072-10,139	10,012-12,584
4	1,374-4,908	10,139-23,350	12,584-25,273

Using this rationale, we decided that classifying households as high or low potential savers would be a good next step. To do so, for each adoption scheme, we

Table 3: *Model training accuracy comparison*

Model type	Solar	Battery	Both
Linear SVM	89	80	92
Polynomial SVM	87	73	89
Naive Bayes	79	67	81

Table 4: *Model test accuracy comparison*

Model type	Solar	Battery	Both
Linear SVM	89	80	89
Polynomial SVM	86	73	88
Naive Bayes	78	67	81

ranked all households based on their savings and split them into four equally sized groups. Group labels are summarized in Table 2. We then build a five-feature classifier to predict which group a household belongs to.

Models we implemented include a linear SVM, polynomial SVM, and Naive Bayes. Model accuracy on our training data is summarized in Table 3. Model accuracy on our test data is summarized in Table 4. Prediction accuracy for dual adoption is generally the highest, followed by solar adoption and then battery adoption. Overall, the linear SVM model is the strongest performer across all adoption schemes for the test data.

IV. Conclusions

The results presented above show that it is possible to make meaningful predictions about whether or not a households should adopt distributed energy technology based on their smart meter data. A five-feature linear regression model gives decent estimates of how much money households would save under the three technology adoption schemes. In particular, households with high energy consumption, especially during the daytime hours, can end up saving the most money. A five-feature linear SVM classifier can predict with accuracy between 80% and 90% which quarter of the cost savings spectrum a household lays in. This information would be very helpful for households to decide whether purchasing new technology will help them save money long term.

Potential future work would include building a classifier with a larger number of classes. It would also be interesting to examine the effect of using more features in regression analysis, since there was still room for improvement in the regression models. It would also be worth exploring different types of regression models.

Over the course of the upcoming years as solar and battery technology becomes cheaper and better, adoption of the technology will become more widespread. Utility companies will have real residential savings data, and at that point, it would be interesting to validate the models presented in this paper against true savings data rather than the simulated data used in this project.

We would like to thank PG&E for the smart meter times series data and the Stanford Sustainable Systems Lab for access to the features and simulated cost savings data.

References

[1] Max Kuhn (2017). Classification and Regression Training. R package version 6.0-78. https://cran.r-project.org/web/packages/caret/caret.pdf
[2] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. https://CRAN.R-project.org/package=e1071
[3] J. Kwac, J. Flora, and R. Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data," IEEE Transactions on Smart Grid, vol. 5, no. 1, pp. 420-430, Jan 2014.