2014 IEEE International Conference on Systems, Man, and Cybernetics
October 5-8, 2014, San Diego, CA, USA

# Smart meter data analytics: prediction of enrollment in residential energy efficiency programs

Michael Zeifman
Building Energy Technologies
Fraunhofer Center for Sustainable Energy Systems
Boston, MA
mzeifman@fraunhofer.org

***Abstract*—Massive rollout of residential smart meters has spurred interest in processing the highly granular data available from these devices. Whereas the majority of smart meter data analytics is devoted to characterization of household electric appliances and their operational schedules, little work has been done to leverage these data to predict household propensity to enroll in energy efficiency and demand response programs. The state-of-the-art methodology for household enrollment prediction involves measurable household characteristics (e.g., age, household income, education, presence of children, average energy bill) and a multivariate logistic regression that connects these predictor variables with the probability to enroll. Unfortunately, the prediction accuracy of this method is just slightly better than 50%, and the required household data are not freely available to utilities/ program contractors. We developed a new method for prediction of household propensity to enroll using only hourly electricity consumption data from households' smart meters, collected over twelve months. The method implements advanced machine learning algorithms to reach an unprecedented prediction accuracy of about 90%. This level of accuracy was obtained in our study of a US West Coast behavior-based residential program.**

***Keywords—electricity consumption, disaggregation, utilities, classification***

## I. INTRODUCTION

A massive rollout of residential smart meters is currently underway with some 65 million units to be deployed by 2015 in the US [1], 30 million units planned in the UK [2] and similar numbers of installations in many other technologically advanced countries. A smart meter records household electricity consumption at least every hour, and it is this three orders of magnitude resolution increase over conventional analog power meters that has spurred significant interest in smart meter data analytics. These analytics mainly focus on disaggregating smart meter data into the energy consumption data for individual household appliances [3], [4], [5]. This reveals information about the energy consumption and operating patterns of major household electric appliances, which can be valuable for assessing energy savings and/or demand response potentials. However, it is not clear that the disaggregation results provide insights into households' likelihood to actually save energy.

The likelihood to save energy is closely related to household propensity to enroll in energy efficiency/ demand response programs [6]. Conventionally, households are recruited to participate in such programs without regard to their propensity to enroll. Since the fraction of successfully recruited households is usually low, significant resources are wasted on recruitment efforts [6]. Moreover, recruitment within a sub-population that is likely to enroll could be made more effective by targeted marketing.

The state-of-the-art methodology for household enrollment prediction involves measurable household characteristics (e.g., age, household income, education, presence of children, average energy bill) and a multivariate logistic regression that connects these predictor variables with the probability to enroll [7]. Unfortunately, the classification accuracy of this method is just slightly better than 50%, and the required household data are not freely available to utility companies.

In this paper, we present a new method for prediction of household propensity to enroll in energy efficiency programs that requires only twelve months of hourly household smart meter data. The method implements machine learning algorithms to reach an unprecedented prediction accuracy of about 90%. This level of accuracy is obtained in our study of a US West Coast behavior-based residential program.

## II. KEY MODEL ASSUMPTION

The electricity consumption of a household is defined by a particular household's set of appliances and how the appliances are used, i.e., its residents' behavioral patterns and habits. Using electricity consumption data obtained at hourly resolution, researchers can deduce the presence of major household electric appliances and their operational profiles by applying relatively simple disaggregation algorithms [3], [4], [5]. Therefore, hourly data embed significant information about household behaviors. This embedded information, in turn, may correlate with households' propensity to participate in energy saving programs. Consequently, we assume that the hourly electricity data of residential utility customers can be a predictor of the propensity to enroll. This is a strong assumption, but, fortunately, it can be easily tested.

This work was supported by Kauffman Foundation

978-1-4799-3840-7/14/$31.00 ©2014 IEEE

## III. METHODOLOGY

### A. Data description

In 2012, a major West Coast Utility Company partnered with a private contractor to launch a new behavior-based residential energy saving Program[1]. In this opt-in Program, participants were recruited by several channels that included local educational institutes, social media and news advertisement. The participants could control their electricity usage by monitoring their hourly electricity consumption data. Significant awards were offered for energy savings to the participating households.

To be eligible for this Program, residential customers must reside within a specific geographic area (section of a major city). Out of approximately 470,000 eligible customers, about 5,600 customers had enrolled as of September 2012.

We received the following data:

- Information about 5,600 households that enrolled in the Program and 32,000 households located right outside the Program area[2] , including:
  - Hourly household electricity consumption over the twelve month before the Program started (i.e., a set of 8,760 data points per household), and
  - Household zip code.

Due to strict utility regulations, no personal information of any kind accompanied any household data.

### B. Proposed Method

For the binary treatment, the usual choice of a model for propensity estimation is either a linear or a logit probability model [8]. Such models are feasible for traditional low-dimensional socio-econometric covariates, e.g., household income or house floorspace. The sheer size of the energy consumption data array of a household (8,760 data points for one year), however, makes direct application of the linear/logit models unfeasible.

Our patent-pending methodology is explained in detail elsewhere [9]. Briefly, we use a matrix of weights and an indicator response matrix. The indicator response matrix has $G$ columns and $N$ rows, where $G$ is the number of classes and $N$ is the number of observations, each observation being a "curve," i.e., a 1D array of data ($M$ by 1 in size, where $M$ is the number of data points in the "curve"). If observation $n$, $n$ = 1,2, .., $N$, belongs to class $g$, $g$ = 1, 2, .., $G$, the $g^{th}$ column of $n^{th}$ row takes value 1, and 0 otherwise. Once the indicator response matrix is built with training data, this response matrix and the $N$ curves are used for algorithm training, i.e., for estimating the matrix of weights β ($M$ + 1 by $G$ in size). For the classification, each curve is assigned to the class having the largest estimated response value calculated using the matrix β and the curve.

In our case, $G$ = 2, $M$ = 8,760 and $N$ is the number of households for training. We use 2,000 of randomly selected households of class 1 (enrolled) and another 2,000 randomly selected households of class 2 (not enrolled) for training. To minimize possible effects related to different levels of household electricity consumption, we normalize each household data set by subtracting its average value and dividing by its standard deviation. After training, we can calculate the response matrices for validation and testing. Figure 1 shows a block scheme of this process.

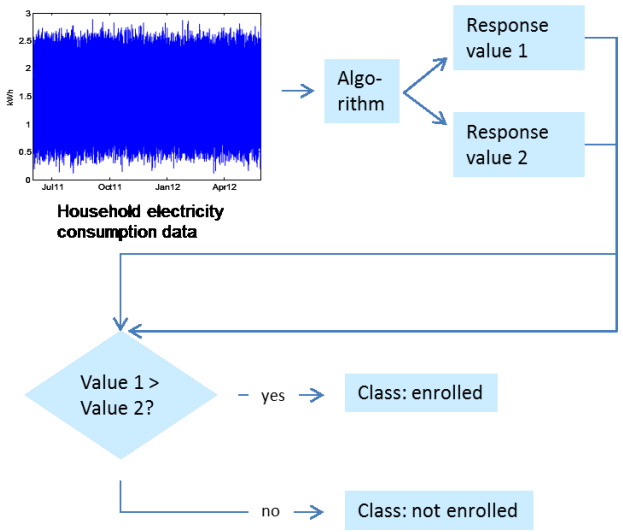


Fig. 1. Block scheme of classification based on smart meter data.

## IV. RESULTS

Initially, we trained our algorithm using a random sample of normalized data from 2,000 enrolled and a random sample of normalized data from 2,000 not enrolled households. For testing, we randomly selected a sample of 2,000 enrolled households from the remaining pool of 3,600 enrolled households and 2,000 of not enrolled households from the remaining pool of 30,000 not-enrolled households. The obtained results were as follows.

For validation (using the same sample for estimation and prediction):

- Number of correctly classified households, class enrolled: 1,848 (92.4%). Number of correctly classified households, class not enrolled: 1,834 (91.7%).

For testing (using different not overlapping samples for estimation and classification):

- Number of correctly classified households, class enrolled: 1,825 (91.2%). Number of correctly classified households, class not enrolled: 1,809 (90.4%).

Encouraged by these results, we performed multiple cross validation using random sub-sampling. In this setting, we repeated the above-described process 1,000 times. At each run,

[1] Our research agreement prevents us from publishing the contractor name. The contractor requested us to keep the Utility name confidential.

[2] These 32,000 households were in the same micro climate zone as those enrolled – see Section V.B for more information.

we draw at random two samples of 2,000 households each, one from 5,600 enrolled households and one from 32,000 not enrolled households. Using these two samples, we estimate matrix β. Then we draw at random a sample from the 3,600 remaining households that enrolled and perform classification of this sample using the estimated matrix β. Finally, we draw at random a sample of 2,000 households from the 30,000 remaining households that did not enroll and classify this sample. Table I below shows the classification results in terms of the 95% confidence intervals.

TABLE I. CLASSIFICATION ACCURACY IN CROSS-VALIDATION, 95% CONFIDENCE INTERVAL

| **Samples used** | **Enrolled households** | **Not enrolled households** |
|---|---|---|
| Training samples | 92.4±1.1 % | 91.7±1.3 % |
| Testing samples | 91.2±1.1 % | 90.5±1.4 % |

Figure 2 shows histograms of the propensity scores for a sample of 2,000 enrolled households. The scores were calculated using matrix β estimated from this sample and a sample of 2,000 not-enrolled households. It can be observed that the scores are not exactly centered around 1 and 0 as they would under a perfect separation.

The particular choice of the sample size (2,000) is based on

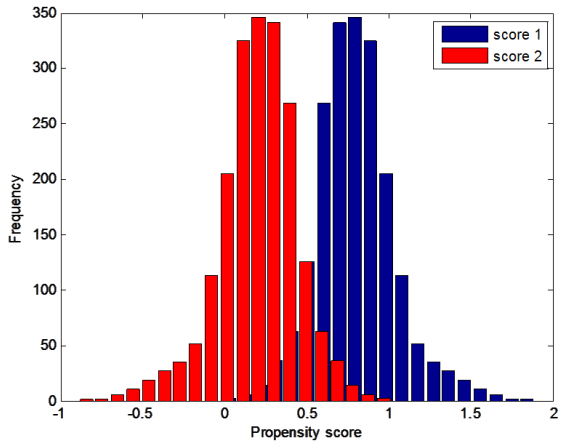


Fig. 2. Distributions of propensity scores (1 - for enrollment class, 2 for non-enrollment class) for a sample of 2,000 enrolled households.

a separate analysis of the classification accuracy vs. the sample size. This analysis is beyond the scope of this paper. Briefly, the classification accuracy decreases to about 80% as the sample size decreases to 1,000 and to 65% as the sample size decrease to 500. The choice of the data array size (12 months) is motivated by the necessity to include seasonal variations in residential electricity consumption.

## V. DISCUSSION

The results of the previous section suggest that household smart meter data can be a surprisingly strong predictor for the household propensity to enroll in an energy efficiency program. We can only speculate about the origins of the found relationship between appliance usage patterns (latent in the smart meter data and apparently isolated by our algorithm) and the behavioral tendency to save energy. Consequently, we explored various aspects related to the data we used in this work to determine if there might be additional factors that inadvertently contributed to the high classification accuracy.

### *A. Not enrolled vs. did not get chance to enroll*

In principle, the pool of households that we marked "not enrolled" actually consists of the households that did not get a chance to enroll, because they are located outside of the eligible geographic area (see section III.A). However, given the fact that only about 1% of eligible households has enrolled, the difference between the not enrolled households and the households that did not get a chance to enroll is small as compared to our classification error.

The reason we did not use the potentially eligible but not enrolled households is related to the unavailability of such data to us: the primary goal of this research was to evaluate the energy savings achieved by the Program and correspondingly by constructing a valid quasi-experimental control group [the evaluation part of this work is confidential]. The use of households that were exposed to recruitment efforts but did not enroll for control group construction would create a selection bias [6]; [also see section VI]. Accordingly, the utility company did not supply us the data for such households.

### *B. Different climate zones*

Since the enrolled and not-enrolled households are located in different parts of a large city, the latent differences in energy consumption uncovered could potentially be attributed to different climate zones within the city. To prevent this effect, we used a climate zone classification system for the US State where the Program was administered. Accordingly, we defined the climate zones of the enrolled households; in fact, all enrolled households belong to the same climate zone. The pool of 32,000 not-enrolled households that we used in this work was selected from a larger pool of about 200,000 candidate households by matching the household zip codes with the zip codes of this single climate zone. Therefore, both "enrolled" and "not enrolled" households that we used are located in the same climate zone.

### *C. Different socio-economic characteristics*

If the enrolled households are located in a wealthier neighborhood and the not-enrolled households are located in a poorer neighborhood, the latent differences in energy consumption might be attributed to differences in the socio-economic population parameters. We used the US Census data [10] to assess potential differences within and between the enrolled and not-enrolled populations. We studied such parameters as median age, fraction of family households, average family size, and median household income for the zip codes related to the two populations. We found no statistically significant difference between the populations.

### D. *Different energy consumptions*

The average hourly energy consumption of the 5,600 enrolled households during the pre-Program period is 0.5957 kWh and the standard deviation of this average is 0.012 kWh. The average hourly energy consumption of the 32,000 not-enrolled households during the pre-Program period is 0.6126 kWh and the standard deviation of the average is 0.007 kWh. Therefore, the difference between the average hourly energy consumptions of these two large samples is within two standard deviations, i.e., statistically insignificant.

In addition, it is of interest to check if average hourly electricity consumption correlates with the estimated response variable or propensity score (see section III). Figure 3 shows a scatter plot of the average hourly electricity consumption of the enrolled households versus the calculated propensity score. No noticeable dependence between these two quantities can be observed. The estimated value of a correlation coefficient is -0.05, which is insignificant for the sample size. In other words, the household propensity to enroll does not correlate with the household electricity consumption.

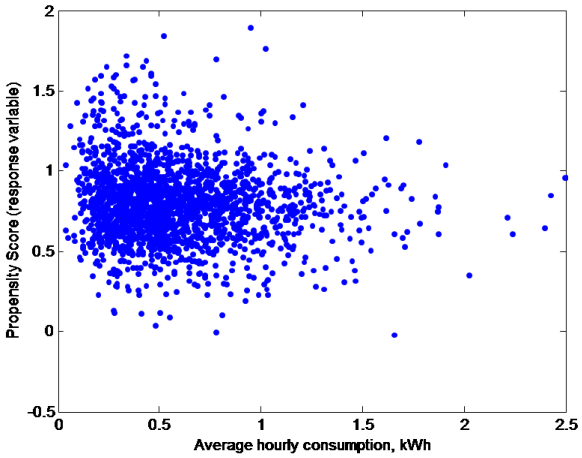


Fig. 3. Average hourly electricity consumption vs. propensity score for a sample of 2,000 enrolled households.

## VI. CONCLUSIONS

The proposed method provides a significant improvement in accuracy of enrollment prediction over the state-of-the-art: about 90% of classification accuracy of our method vs. about 50% classification accuracy implied in Ref. [7]. Whereas both ours and the state-of-the-art method require data for several hundred households for training, our method requires only hourly smart meter data that are freely available to utilities. In contrast, the state-of-the-art method requires numerous socio-economic characteristics of each household that are not freely available.

We foresee two immediate applications of our method. First, it can be used to pinpoint households that are likely to participate in energy efficiency programs, thus saving resources for customer acquisition in such programs and potentially increasing participation rates. Second, it can be used to construct a valid quasi-experimental control group for evaluation of energy savings from these programs [6]. Many behavior-based energy efficiency programs cannot implement an experimental control group using, e.g., the randomized control trials, because of the program design or budget limitations [6]. The current practice of constructing quasi-experimental control groups usually involves matching by energy consumption, i.e., for each enrolled household a baseline or "control" household is selected from a pool of candidate households by minimizing the difference in energy consumption. Our Figure 3 indicates that the propensity to enroll is not correlated with energy consumption; instead, a higher propensity to enroll in a program reflects behavioral profiles that also make them more likely to modify their energy-related behaviors. Thus, using the households that are unlikely to enroll as the baseline for evaluation of energy saving introduces a well-known selection bias [6].

Our work poses many open questions. For example, our method worked well (enrollment prediction with 90% accuracy) for a given region/program. Will its performance be as good for a different region or program? Can the enrollment data from one program be used as a proxy for another program in the same region? Can our algorithm, trained on data from one region, be applied to data from another region? What are the requirements to the region (e.g., size, homogeneity)? Additional research work is needed to answer these questions.

## ACKNOWLEDGMENT

The author is grateful to Peter Klint and Dr. Kurt Roth for insightful discussions and support.

## REFERENCES

[1] Utility-Scale Smart Meter Deployments, Plans, & Proposals, Report by Institute for Electric Efficiency, the Edison foundation (2012).

[2] Smart Metering Summary Plan, Policy paper, Department of energy and climate change, UK (2013).

[3] A. Molina-Markham, et al. "Private memoirs of a smart meter," Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building. ACM, 2010.

[4] C. Armel, K., et al. "Is disaggregation the holy grail of energy efficiency? The case of electricity." Energy Policy 52 (2013): 213-234..

[5] B.J. Birt, et al, "Disaggregating categories of electrical energy end-use from whole-house hourly data," Energy and Buildings 50, pp. 93-102 (2012).

[6] "Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations." US Department of Energy, 2012.

[7] M. Harding, and A. Hsiaw, "Goal Setting and Energy Efficiency," Working Paper (2012).

[8] M. Caliendo, S. Kopeinig, "Some Practical Guidance for the Implementation of Propensity Score Matching," Journal of Economic Surveys 22, pp. 31–72 (2008).

[9] Zeifman M., "System and Method of Prediction of Household Enrollment in Energy Saving Program," patent application (2013).

[10] US Census Bureau community facts, available at http://factfinder2.census.gov/faces/nav/jsf/pages/community_facts.xhtm.