Analytics for understanding customer behavior in the energy and utility industry

Y. Kim
A. Aravkin
H. Fei
A. Zondervan
M. Wolf

Understanding customer behavior is becoming critical for electric grid operations. Grid operators need to model customer behavior from multiple perspectives, in part due to recent changes of the customer role from passive loads to prosumers. Customers are active agents rather than "passive loads," and this change in behavior comes with a variety of challenges. Solving these challenges requires a 360-degree view of the customer, which calls for machine learning techniques for classification, time-series analysis, and uncertainty quantification. This allows utilities to actively work in the challenging landscape of active customer behavior. In this paper, we describe several important techniques for practical modeling of active customer behavior, together with corresponding areas of behavior modeling: energy savings potential, adoption of sustainable products and services, prediction of photovoltaic adoption, and fraud detection. For each of these applications, we briefly survey machine learning tools that have allowed demonstrable practical impact. Where possible, we illustrate results using datasets from Alliander N.V. (a Dutch energy company), as well as other companies. We present quantitative results, describe their qualitative impact on the companies concerned, and provide some practical insight into model building and validation.

Introduction

The design principle of modern electric grids is based on several assumptions about grid components, including generation plants, transmission systems, distribution systems, and energy consumers. Traditionally, utility customers have been considered only as energy consumers, i.e., passive loads, divided into three customer classes: residential, commercial/midsize, and industrial. Analysis was limited to understanding the electrical load characteristics of each category, in order to plan supply and provide reliable services. The key research agenda underlying these aims was identifying load patterns and developing load prediction algorithms at various aggregation levels (e.g., distribution transformers, substation transformers, and transmission components) [1, 2].

In recent years, the energy landscape has changed rapidly. Disruptive technologies such as distributed energy resources, rooftop solar panels, and electric vehicles drive the uncertainty of customer behavior to highest levels to date. The rise of prosumers, compounded by environmental concerns, pushes the limit of customer engagement capabilities of the utilities industry. Without a deep understanding and a 360-degree view of each customer, grid operation faces significant risks. For example, continued operations that do not take customer behavior into account have led to failure of electric grids, including unexpected blackouts for many customers due to high supply surge induced by photovoltaic generation [3].

Unfortunately, utility companies have been relying mainly on survey data to understand their customers. This reliance is vulnerable to bias (e.g., self-selection for surveys), and leads to an ineffective customer engagement. For example, typical response rate for surveys is 10% to

Digital Object Identifier: 10.1147/JRD.2015.2503988

©**Copyright** 2016 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the Journal reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied by any means or distributed royalty free without further permission by computer-based and other information-service systems. Permission to republish any other portion of this paper must be obtained from the Editor.

0018-8646/16 © 2016 IBM

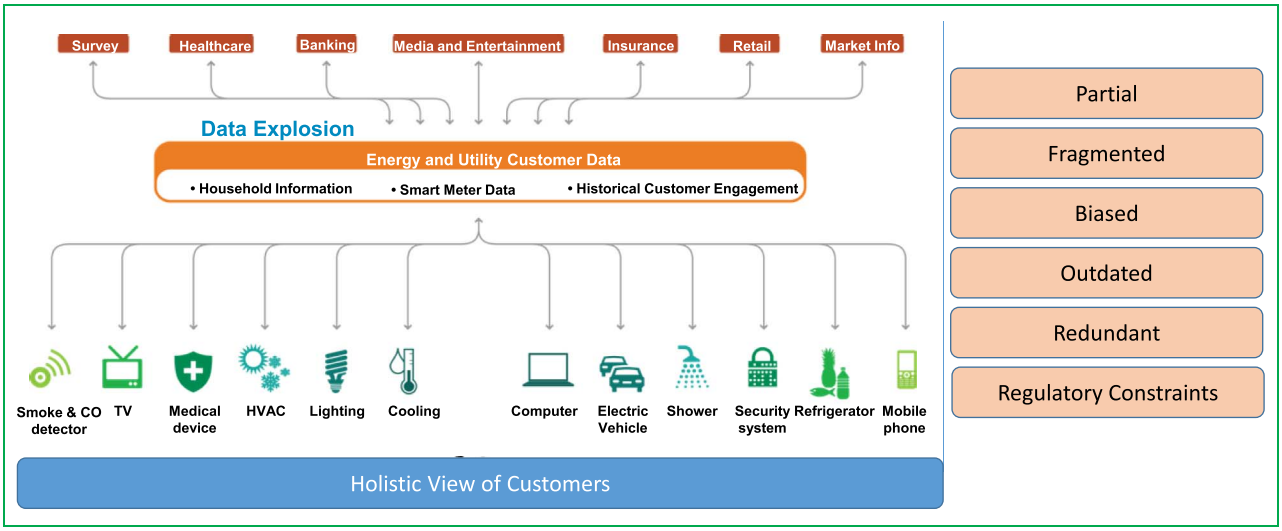

Figure 1

Customer data landscape and challenges. (HVAC: heating ventilation and air conditioning; CO: carbon monoxide.)

15% [4], and survey data are often specific to each use case. A typical survey is designed to address specific concerns such as sentiment and customer satisfaction regarding service quality and product delivery [4].

Fortunately, utility companies already own significant amount of customer data. Building information, demographic attributes, lifestyle information, and growing number of smart meter data offer enormous opportunities to build data-driven analytics on customers. Despite data fragmentation and other limitations depicted in **Figure 1**, we believe that these data sources can provide a holistic understanding of customers when combined with careful curation and consideration of data sources.

In this paper, we present a high-level summary from a research program developing customer analytics, carried out in collaboration with Alliander N.V. (a Dutch utility company). We describe the data landscape of an electric company, develop analytics use cases and tangible business cases, and discuss outcomes. Highlighted applications include sustainability behavior modeling, energy savings potential, load forecast, generation forecast, and fraud detection. In each of the examples, we use analytics to help achieve yearly key performance metrics, including meeting a 3.4% energy savings target and reducing revenue loss by €3 million by 2018 with improved fraud detection.

We begin this paper with an overview of data issues in energy and utilities. Next, we provide a brief description of the modeling and analytics process. We then present three use cases, discussing the problems, methods, and results. The use cases come from collaboration between the Smarter Energy Research Institute (SERI) at IBM Research with Alliander, a large energy provider in the Netherlands. Specific problems discussed are 1) energy-savings potential modeling, 2) photovoltaic adoption modeling, and 3) fraud detection.

Customer data landscape in the energy and utility industry

Data associated with residential customers comprises a set of attributes, time-series data (e.g., smart meter and billing data), and associated location/address data. Typical attributes include house size, house type, household income, education, and number of occupants. Marketing data vendors offer more detailed household characteristics such as shopping patterns and loyalty card transactions. Aside from the fact that the use of the private data needs to follow strict privacy regulations, the number of these attributes can be greater than 1,000, covering a comprehensive snapshot of each household. Electricity usage data, historical billing information, and payment history capture usage trends as well as the state of each household's financial status.

Historical marketing data provides insights on energy efficiency programs, electric heat pump promotion outcomes, photovoltaic adoption, and other information. Since the data is being collected from partial markets, the data capture only positive labels (e.g., the data captures only a subset of the true ownership).

Figure 1 illustrates a rich set of data allowing deep understanding of customers from different perspectives. Challenges in the data warehouse include (a) a significant

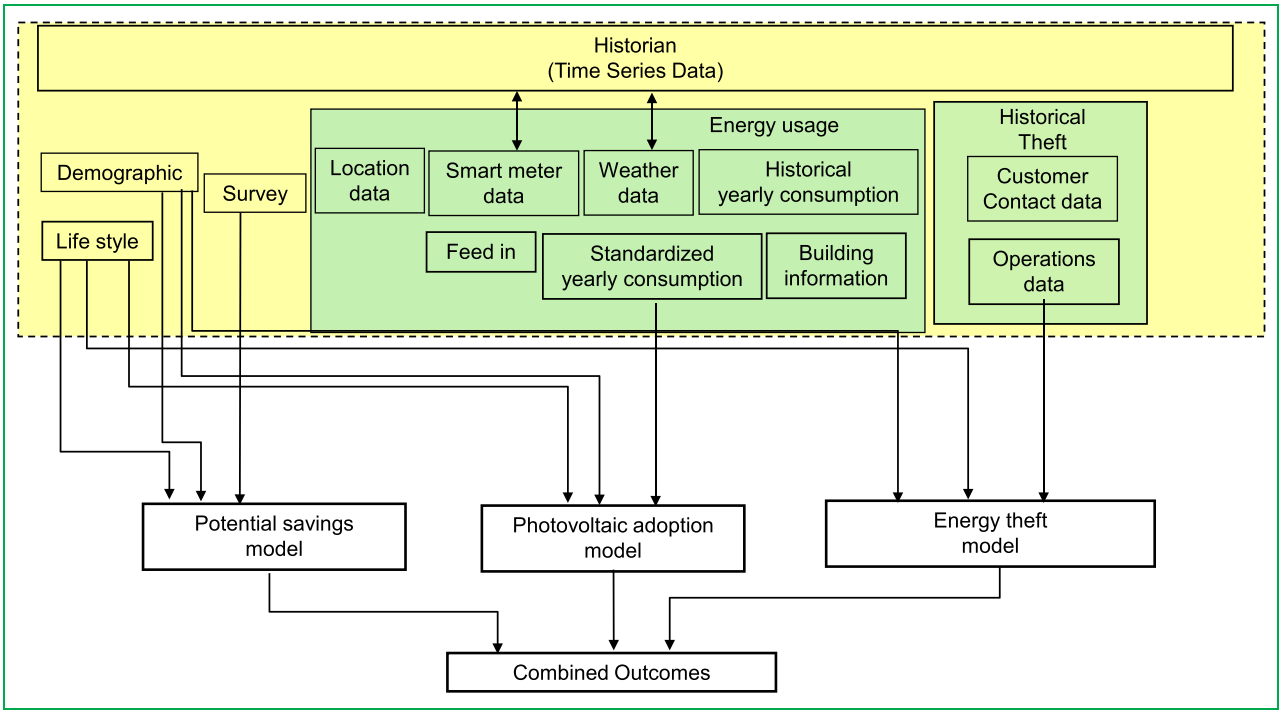

Figure 2

Curated data inputs for various models.

amount of outdated data, (b) incomplete data coverage, and (c) clerical and other errors in the data.

Data curation and correlation are fundamental precursors to analytics development for customer data sets. The process involves several steps. Customer data typically reside in heterogeneous data silos. Aside from regulatory constraints, mapping these data to each customer involves careful consideration. Typically, the fundamental attributes used for mapping are spatiotemporal; in particular, to perform analysis, we need to identify a time period during which a household is associated with a particular address. This data is available in a customer registry database where a combination of name, account information, and address information is stored. When an entry is identified, a set of merging and searching mechanisms seeks data that map to the address information for the particular time period. For analysis, the data curation needs to fill in incomplete data where possible. A common practice to fill incomplete data is to extrapolate the missing attributes from similar attributes such as income level and house value.

Analytics and data process overview

Once data are consolidated and curated, a series of steps are required to build a model. The first step is to choose relevant attributes from a list. To choose the set of attributes, several methods can be applied. Attribute selection methods include the following: (a) manual attribute selection using domain expert knowledge, (b) merging and clustering of attributes by use-case context, (c) deriving composite attributes, and (d) deletion of noisy attributes. Once the attributes are processed, a set of machine learning tools are used to evaluate the feasibility of model applications (**Figure 2**). Clustering analysis and other unsupervised machine learning techniques are used to identify similar customers in a particular context, which is key to identifying desired market characteristics.

Ideally, a customer behavior model can be built using supervised learning techniques, but this requires training data. Unfortunately, often only partial observations of markets are available, requiring novel approaches to build a classification method. As an example, consider the problem of ownership detection of a particular appliance. Suppose one wants to identify owners of electric heat pumps in premises using a combination of building attributes and coarse grained electric meter data. The training data contains only a short list of customers who purchased electric heat pumps (e.g., less than 0.5% of the entire population).

This simple practical problem, which is common across most of the customer engagement questions, has unique

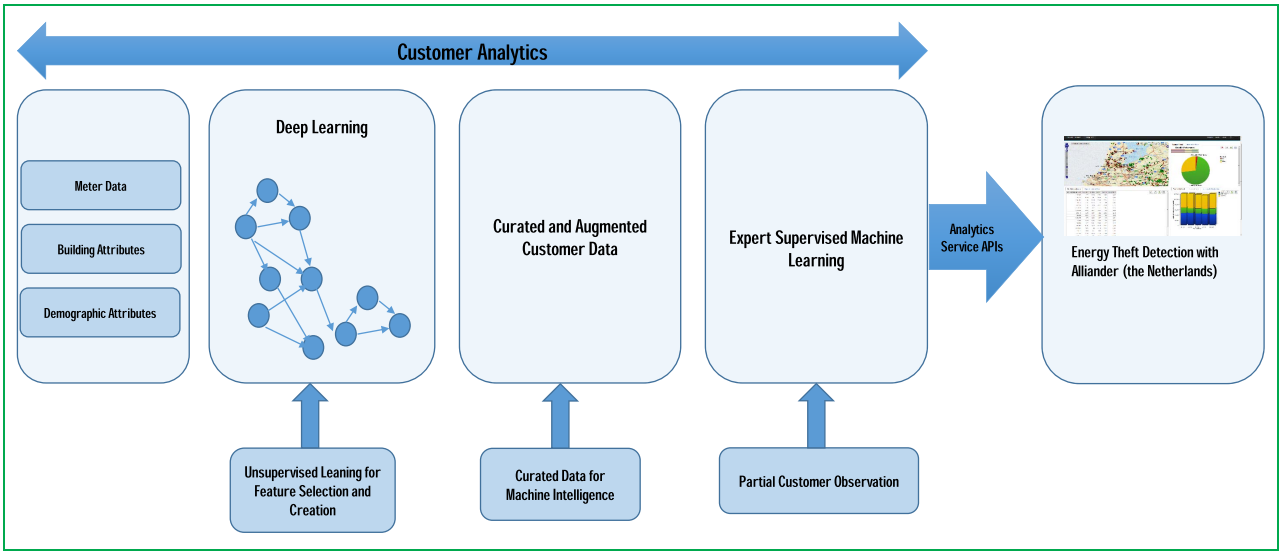

Figure 3

Example customer analytics flow.

challenges. First, the detection algorithm needs to be trained using partial (only positive) labels (since we only know ownership information). Second, due to the current deployment setting of the smart meter reading infrastructure, only coarse-grained energy consumption data are available for the detection algorithm.

Traditional methods requiring complete label information cannot be directly used in this setting because these methods require: 1) temporally fine-grained smart meter data that captures appliance-level activities, 2) a rich set of training data for a specific household including feature libraries such as appliance usage patterns, and/or 3) ground-truth data of target appliance usage. None of these requirements is satisfied in the current problem setting.

Therefore, the recipe for the modeling building needs to allow data-driven analytics to extend traditional supervised machine learning with the ability to tune feature choices, smart feature extraction from both time-series and survey data, methods to draw inference from partially labeled data, and the ability to support quality measurements from various perspectives. **Figure 3** shows an analytics flow diagram describing the flow from the curated data, feature extraction, feature derivation, to model training and applications where any model can be built following the steps. We leverage the state-of-the-art machine learning techniques covering unsupervised learning such as *K*-means, two step clustering, and Kohonen self-forming map, supervised learning such as decision trees (e.g., CHAID [Chi-squared Automatic Interaction Detection]) [5], support vector machines (SVMs) [6], Classification and Regression Trees, Generalized Linear Models (GLMs), neural network (NN) [7] and semi-supervised learning techniques such as regularized support vector machines (R-SVMs) for each category of problems. A particular recipe for each use case is discussed in the following section.

Use cases

In this section, we provide details on three use cases between the Smarter Energy Research Institute (SERI) at IBM Research, and Alliander N.V., a utility partner based in the Netherlands. SERI is a research program that partners energy and utility companies with IBM Research to develop analytics solutions for key businesses challenges. The SERI program also provides an environment for the platform team to design, build, and validate these capabilities. SERI focuses on joint research efforts in application areas such as outage planning optimization, asset management optimization, integration of renewable and distributed energy resources, wide-area situational awareness, and customer analytics for a participatory network.

SERI engaged with Alliander to gain insight into customer behavior in order to help manage grid-operations. Three primary focus areas included quantifying potential energy savings, predicting solar panel (photovoltaic) adoption to support the upgrade/management planning for low-voltage transformers, and identifying possible energy theft cases to increase public safety and protect revenue-loss (Alliander has the

goal of reducing loss by €3 million by 2018 due to energy theft).

Energy savings potential modeling

Motivation

Modeling of energy savings potential has three motivations. First, Alliander has 3.4% energy savings as a key performance indicator (KPI), which is thus far an unverified mandate. Second, energy-savings pilot programs have been designed and run in an *ad hoc* manner. Third, there is no current capability to perform budget planning and optimization for these activities. For example, it is difficult to answer questions such as, "How much of savings per euro invested can be achieved (kWh/euro)?"

Currently, the energy savings potential per customer is determined with respect to typical household categories such as a single family home or an apartment unit. A further refinement of customer categories is achieved when combined with regional boundaries. The savings potential then is calculated as the difference (if positive) between the customer's consumption and the median energy consumption within the customer's category. This median reference value is rather arbitrary, and other measures may be better for achieving maximum savings. In fact, the difference between the customer's consumption and the group's mode (peak of the distribution) can be a better measure than the median, because the reference energy consumption should be the most common consumption within a category.

This approach has been used to provide the customers with insight into their energy consumption in comparison to the customers' category, which in turn allows the customer to assess whether it is worthwhile to engage in energy savings measures.

Unfortunately, this limited segmentation of customers does not capture a realistic estimate of the energy-savings potential. The minimum energy requirements per customer show stochastic variation due to the inherent variety within a group of customers. This makes per-customer estimates of the energy savings potential unreliable, since few parameters used in the model contribute to large between-customer variation.

Approach

The model for quantifying energy savings potential relies on the following modeling hypotheses. First, similar households should have similar energy consumption patterns. Second, building characteristics, demographic attributes, and historical energy consumption can be effectively used to predict potential energy savings. Third, normal energy consumption can be described by regression variables such as house size, house age, and income level. Finally, deviated consumption from regression output is a proxy for quantifying potential energy savings.

To obtain a reliable estimate of energy savings potential, the model considers two key aspects: fine grained segmentation of customers and robust reference consumption calculation.

To obtain clusters of similar energy consumption groups, we use various datasets including building surveys, demographic information, lifestyle data, and historical energy consumption. We choose attributes that are related to energy consumption trends. The key attributes include gross floor area, house volume, income level, education level, and energy consumption.

To further control variability due to the inherent uncertainty in the historical energy consumption, we perform an aggregate analysis. For example, before running clustering analysis, we normalize energy consumption and calculate descriptive statistics such as weekly average energy consumption, monthly average energy consumption, sample variance of consumption, and minimum and maximum of consumption. We also filter out outliers in the energy consumption range due to the ownership of photovoltaic systems, geothermal heat pumps, and other extreme cases.

Outcome

To identify a best practice and a model for building the energy savings potential model, we trained various clustering models. We chose the best model (according to a held-out or excluded dataset, i.e., data not used for fitting) from a set of commonly used state-of-the-art clustering techniques: *K*-means, two step clustering, and Kohonen self-forming map. **Table 1** summarizes the outcome of the clustering algorithm.

Table 1 shows that a *K*-means approach gives a best clustering quality with respect to the cluster silhouette score. However, the number of clusters is only four, giving less detailed potential savings quantification. On the other hand, the Kohonen clustering algorithm [8, 9] provides 24 clusters which are more finer-grained than the *K*-means. Although the clustering performance is weaker than the outcome from *K*-means (10% worse), this outcome is favorable since it offers more detailed potential savings quantification. Our qualitative assessment of the 24 identified clusters can have more detailed and compelling description of customers, which capture the energy consumption behavior of customers better. Note that the finer grained clusters give more conservative estimate of the energy savings potential, which is due to that fact that the reference energy consumption in a smaller cluster is closer to the other customers in each cluster.

Table 1 Energy savings potential clustering models: converged clusters only.

Clustering algorithm	*Silhouette score*	*Number of clusters*	*Smallest cluster* (%)	*Largest cluster* (%)	*Savings potential* (%)
K-means	0.268	4	21	29	11.2
TwoStep	0.252	2	0	52	8.2
Kohonen	0.242	24	2	8	11.6

The identified best practice to quantify energy savings potential of customers involves two steps:

Step 1 Apply the Kohonen clustering algorithm on three category of variables: (a) building characteristics such as gross floor area, building type, insulation type, house age, and occupancy level; (b) demographic attributes such as income level, education level, age, job category; and (c) life style attributes such as support of sustainability promoting products and services, support of organic products, and support of nature conservation societies.

Step 2 Define a normal energy consumption in each of clusters and calculate the achievable energy consumption. The normal energy consumption is the average energy consumption of each cluster after removing outliers.

The analysis revealed that as of year 2014, the service territory has energy savings potential of 8.2%.

Photovoltaic (PV) adoption forecast modeling

Motivation

The Alliander service territory has experienced rapid growth of PV at a rate of more than 100% in the past two years. While fortunately there were no catastrophic events, neighboring countries experienced PV driven blackouts [3]. To mitigate risks associated with the wide adoption of PV systems, Alliander needs to analyze the trend of PV adoption, and be able to predict it for the near future, at several geographic aggregation levels (postal code, city, and province).

Various statistical models can be used to predict PV adoption as a function of demographic factors (based on historic data). These models can typically provide relative probability of PV adoption for different customers, which in turn can be used to accomplish Alliander's aims. One of the simplest models that can be used is logistic regression, focusing on outcome (PV adoption) as a function of demographic predictors.

The PV adoption modeling uses a combined modeling approach. First, we build a logistic regression model to calculate an adoption probability of PV adoption. Second we extend the static adoption probability to future forecast using the survival analysis.

In this section, we describe logistic regression model development and interpretation. We also describe probabilistic validation of the logistic regression model using techniques from diagnostic tests. Finally, we describe how model-based predicted probabilities can be used to develop predictions at different geographic levels.

A weakness of the logistic regression model is that it does not account for temporal change; and in the last portion of this section, we propose survival analysis as an example of a model that can address this short-coming. After providing a brief overview of survival analysis, we discuss difficulties in interpreting model output in a real-world modeling context.

Approach

A logistic regression model relates a binary outcome to a set of predictors. PV adoption is the main outcome of interest, and logistic regression works with a particular transform of the PV adoption probability. In our context, a vector of fitted coefficients provides important information about the effect of demographic variables on the probability of adopting PV. Each coefficient represents the "status" of customers with respect to a variable, for example a type of housing, income, and so on. The logistic regression model works with overall estimated *baseline probability* of adopting PV. Most fitting programs choose a default baseline customer, but this reference can be changed if of interest to the user. Each coefficient, when exponentiated, gives the odds ratio of adopting PV (probability of adoption at each particular level of each demographic variable relative to probability of adoption at the baseline). The fitted model allows us to compute a predicted probability of adopting PV for each customer. These predicted probabilities are especially useful for comparing customers in different groups in terms of relative likelihood of adopting PV, and obtaining province-level and zip-code-level predictions.

The starting logistic regression model begins by including all available levels of all available demographic variables (e.g., house volume and house area) and descriptions of the customers (e.g., stage of life, social

class, income, and education level). The fitted model provides odds ratio estimates and their 95% confidence intervals, which allows us to evaluate both the significance of the predictive variable, and the magnitude of its effect.

Fitted coefficients allow us to compute fitted values (for data points in the development sample), as well as predicted values (for data points outside the development sample). Both fitted and predicted values can be interpreted as probabilities. For a given threshold p in the interval (0,1), data points with values being greater than p can be considered as "predicted yes." To obtain a measure of model quality, for a given threshold probability, one can look at different measures of agreement between predicted and actual values. All measures of agreement are obtained from the so-called "confusion matrix," a 2×2 matrix that compares predicted to actual PV adoption. Two important measures are true positive fraction (TPF), the fraction of PV adoptions correctly identified, and false positive fraction (FPF), the fraction of non-PV adoptions incorrectly predicted to adopt.

These values of course depend on the chosen threshold probability. It is therefore particularly useful to look across the entire range of p in (0,1). As p runs from 0 to 1, it affects both the FPF and TPF (the reader can easily conclude that for $p = 0$, $(\text{FPF}, \text{TPF}) = (1, 1)$, and as p increases, both statistics decrease monotonically to 0). Pairs (FPF, TPF) corresponding to each threshold can then be drawn on a graph. This graph is called a Receiver-Operator Characteristic curve (ROC), which is frequently used to validate quality of diagnostic tests. It contains all TPF/FPF information for all thresholds, and describes the natural tradeoff between these quantities. The area under this curve (AUC) gives a one number summary of the quality of the test. The higher the AUC is, the better; AUC is necessarily between 0 and 1, with $AUC = 0.5$ meaning the model is not useful, and $AUC = 1.0$ meaning that the model can perfectly separate adopters from non-adopters.

Validation of logistic regression model

We exploit ROC techniques to build a validation strategy for logistic regression. We split the data into training groups (70%) and validation groups (30%), and fit a logistic regression model on 70% of the data, obtaining estimates. Using these estimates, we build ROC curves for both the development sample and held-out sample, and compute the AUC in each case. The AUC can evaluate the quality of the model, especially when computed for the held-out set. It is interesting to note that our model performed equally well $(AUC = 0.79)$ for both training and testing as shown in **Figure 4**. This means that the logistic model has some predictive power, and, furthermore, the AUC based on training was a good estimate of the out-of-model AUC.

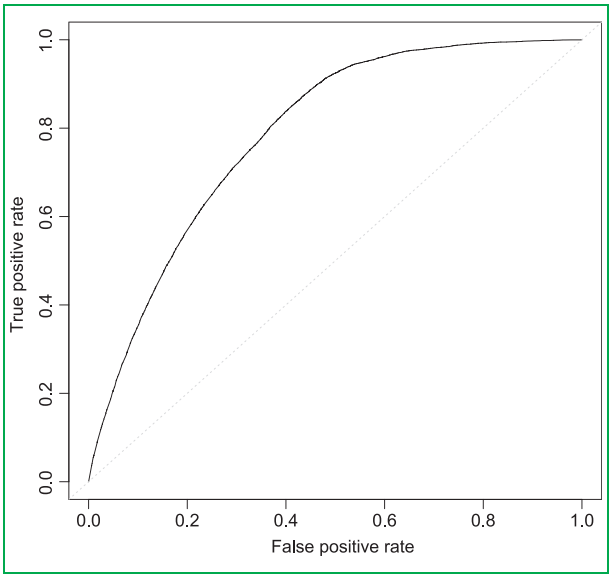

Figure 4

Receiver-Operator Characteristic Curve for final PV model. Here, AUC = 0.788.

While AUC above 0.8 is considered "good," this is an understandably vague criterion. Interpreting the AUC is challenging from a practical perspective, but there are other means to discuss performance, in terms of sensitivity and specificity thresholds that are relevant to specific applications. For example, based on our analysis, we know that if we pick a threshold that correctly identifies 90% of customers who would adopt PV (from the current dataset), we can expect this model to have roughly a 47% false positive rate. We can also use it to obtain relevant thresholds. For example, if the entire dataset for non-PV customers is used, the predicted value threshold for probability of adoption that corresponds to the 90% sensitivity and 47% specificity is 0.01. Thus, if we label a customer as "will adopt PV" if her probability of adoption predicted by the model is 0.01 or higher, this labeling will correctly identify 90% of customers who adopted PV, but also include 47% of the customer base who didn't adopt PV. The low threshold probability is not surprising, since the proportion of customers who adopted PV in the database is quite low, and this proportion influences the decision threshold. ROC-based analysis allows us to state and evaluate outcomes with respect to FPF and TPF, without reference to particular model-based thresholds.

Use of prediction outcomes

In this section, we describe a method to propagate large-scale adoption rate scenarios (e.g., one scenario may be that 200,000 customers will adopt PV by 2020) to a finer geographic resolution using a predictive probability

model. Utilities are faced with the tough problem of not just maintaining assets, but evolving grids to a more distributed model in which energy is injected at multiple points into the grid like wind farms, solar farms and PV "behind the meter." It is impossible to maintain and plan assets without an accurate prediction of PV adoption in the service territory, and thus PV adoption is an essential building block of the capital planning process.

When we use the term predictive probability model, we mean a model that, given a household, can compute a probability of adoption of the product by that household. Logistic regression is an example of such a model, as pointed out in the previous section. Inputs to the model include demographic variables, such as income level, education level, and household information. The output (at the household level) is a scalar probability estimate in (0, 1). Once these probabilities are computed at the household level, they can be used for various tasks. One such task is ranking households in order of likelihood of adoption, from most likely to least likely. Another task, particularly useful for prediction and uncertainty quantification, is to provide a relative weighting scheme by which customers can be selected (for promotional campaigns, for example).

A large-scale scenario (such as the 200,000 customer selection in the above example) can then be translated into further geographic boundaries as follows. We sample the predicted number of customers, using the predicted weights as sample probabilities. Once the required number is sampled, we can group the sample by the required geographic level (zip-code, city, and province). A single sample or grouping is treated as a *random realization* of model-based predictions. By repeating the process multiple times, "reasonable intervals" for plausible customer distributions according to the required geographical level are obtained. The custom algorithm we developed for the Alliander project is detailed below:

(a) *Initialization*: Input the times series of projection scenarios, along with low, medium, and high scenario estimates for the total number of customers who will adopt PV.
(b) For each random realization (1 to 100,000):

  (i) For each year, for each estimate (low/medium/high), sample the required number of customers using model-based probability estimates.
  (ii) Count up and record customers within each province/PC4/PC6 area, where PC refers to a Postal Code.

(c) *Output*: For each year, for each estimate (low/medium/high), and for each locale (province/PC4/PC6), report the required quantile interval (e.g., 0.025–0.975 for 95% reasonable range) of counts.

Limitations of stationary models

The main limitation of using predicted probabilities based on stationary models is the underlying assumption that current trends (i.e., effects of demographic variables observed up to now) will continue into the future. This clearly will not be the case in general. Models that allow changes over time are therefore preferable; however, they are also necessarily more complicated, and harder to interpret, validate, and use for prediction. Survival analysis is a particular extension we considered; although it was not fully implemented for the project, we include a brief description of the method and some inherent modeling difficulties.

Time-series and survival analysis

Survival analysis models the time it takes for events to occur. While used in medical studies (hence the name), it is often used to model phenomena close in nature to technological adoption (time to default, etc.). Survival analysis is based on the *hazard function* $h(t)$, the instantaneous rate of PV adoption at time t, given no adoption prior to time t. The Cox Proportional-Hazards model is very popular, because it is semi-parametric (it requires no functional assumptions on the hazard function).

Cox survival model

We fit a Cox regression model, using the same covariates as for Logistic Regression. The fundamental assumption of Cox Proportional Hazards (i.e., where effects of each covariate is multiplicative relative to the baseline time-dependent hazard function)) makes it easy to consider this a non-stationary (or dynamic) extension of logistic regression. In particular, the coefficients corresponding to demographic variables are interpretable in much the same way as for the logistic regression model. All time dependence is accounted for by the baseline hazards model, which can potentially capture temporal trends *over the time period* that the model is fitted.

Survival model diagnostics

There are several issues one can test for to determine whether or not a model is valid. The most common issues include the following: (a) violation of the assumption of proportional hazards, (b) influential data points, and (c) nonlinearity in the relationship between the log hazard and the covariates. We ran a basic diagnostic test for violation of proportional hazards. Unfortunately, assessing violations are based on interpreting plots; and plots for large-scale real data were difficult to interpret. We therefore focused on testing assumption (a), violations of the proportional hazards assumption, which can be quantified using p-values.

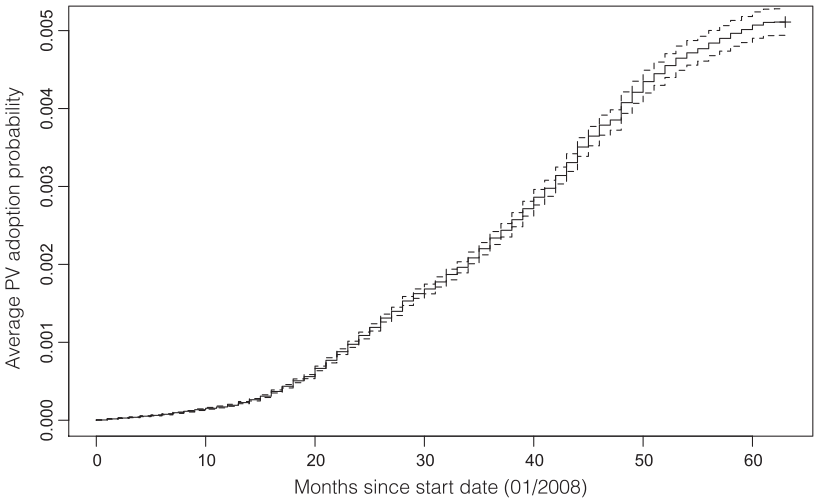

Figure 5

Cumulative hazard rates (for adoption of PV) by month over a 5-year period. The middle curve is the estimated cumulative hazard rate (probability of adopting PV by date given on x-axis), with the top and bottom curve giving the 95% estimated confidence interval.

We observed strong evidence of non-proportional hazards for some demographic levels, in particular housing type. This evidence raised questions of how to proceed with model building—in particular, one could include additional terms, such as interaction terms; otherwise, one can build multiple survival models for different subsets of data. Both of these directions are potentially problematic, as they lead to an increased variable space for an already sparse dataset (when time is taken into account, PV adoptions in each time period can be very small).

Difficulties in prediction using survival analysis

One of our main goals is to use time-series modeling to improve prediction of PV adoption. While logistic regression provides expected probabilities of adoption, these are fixed in time; however, more difficulties occur when using time-dependent outputs.

Survival analysis holds promise for time-dependent analysis, since it gives us a way to estimate expected hazards. The key issue is how to extrapolate the interval of fitting, meaning the time period for which we have data. In other words, the fitted hazard function $h(t)$ is valid over the time period for which the model was constructed; it is less clear how to project this estimate to the future. The estimated hazard function for PV adoption estimated over a 5-year period is shown in **Figure 5**. Note that the hazard function is increasing over time, so as we expect, there is an acceleration in PV adoption over time, and the hazard function is useful in capturing/quantifying this trend. Prediction using the hazard function is more difficult.

One potential strategy would be use the estimate of the hazard at the final time point for each group/individual, and then use this information to predict future adoption, by assuming that the hazard remains fixed over time or applying different analyses, including Bass curves [10].

All of these questions, together with concerns about data sparsity and validity of the proportional hazard assumption, relegated the survival analysis extension of PV adoption modeling to future research.

Best practice recommendation

The best practice to date in PV penetration modeling involves the following steps.

Step 1 Build an initial logistic regression model with all levels of all demographic variables.
Step 2 Based on the initial model, merge levels into groups according to effect on outcome variable (PV adoption).
Step 3 Using a training set (random subset of initial data, e.g., 70%), build a model on the grouped variables.
Step 4 Test the predictive power of the model on the held-out testing set (remaining 30%).

The resulting model was tested using the ROC curve, as described in the model validation section above. The final ROC curve (with AUC measure of 0.78 on testing data) is presented in Figure 4.

Fraud detection modeling

Motivation

Alliander estimates total losses from electricity theft across its service territory at approximately €150 million to €160 million annually [11]. Ninety-five percent of all energy theft in the Netherlands is believed to be from marijuana growing operations, and Alliander estimates that approximately 0.3% to 1% of homes have some kind of illegal growing operation. Alliander aims to boost its annual recoveries to beyond €15 million, from €10 million today. It is estimated that there are between 30,000 to 100,000 marijuana plant growers in the Netherlands. Alliander loses 350 GWh annually through theft (equivalent to the consumption of 100,000 households). 7,500 police-controlled forced entries include: charged theft (€35 million), collected theft (€10 million), and department cost (€10 million).

Alliander has investigated many direct power analysis techniques for identifying potential cases, but significant privacy and legal constraints severely restrict such efforts. Alliander hopes to build advanced data analytics techniques that can help the fraud detection department in its efforts.

Alliander faces considerable constraints in its fraud detection efforts. First, Dutch law imposes significant data privacy controls that restrict the collection and sharing of certain customer data. These privacy concerns, coupled with limited AMI (advanced metering infrastructure) across the Alliander service territory, result in severely limited power consumption insight at the customer level. Alliander's yearly self-reporting policy worsens the situation. Alliander must rely on customer self-reports once per year and an Alliander reading only every third year.

The existing model has much room for improvement in terms of detecting potential fraud cases [12]. For example, one of the existing models has 0.7% true positive fraction (TPF). The goal of this work is to improve the accuracy of existing fraud-detection models without requiring additional data, i.e., using only the available data (demographic, consumption, and known fraud cases). The proposed classification models are built using the following observations. First, significant consumption-change behaviors will be identifiable (e.g., the customer may revert to reporting the known average consumption in the year they begin growing operations, while in prior years they were well over or well under). Clearly, the availability of yearly self-reported consumption makes it much harder to identify such anomalies. Also, participants in growing operations will have consumption patterns different from others with similar characteristics. Again, the low resolution of consumption data (only yearly self-reported with Alliander readings every third year) makes this effect more difficult to measure. Our approach is to use automatic data mining approaches to identify indicative variables for possible fraud cases by correlating past fraud cases with all the available attributes.

Approach

Feature selection and normalization

The first step is to select the key features for the model training. The features are chosen based on statistical criteria and the quality of the data such as the availability, diversity, and bias of each attribute. By performing the initial feature filtering, obvious errors and noise from the raw data were removed.

To further improve the quality of the features for model building, feature normalization and iterative feature elimination methods were applied. Feature normalization involves the following steps: (a) normalize the standard yearly energy consumption of each year by maximum consumption, (b) normalize each attribute to the scale of 1, and (c) calculate year-to-year consumption difference based on the normalized energy consumption.

The iterative feature elimination method was then applied to check the importance of the features in relation with overall accuracy. The method starts by using all the features and gradually removes one feature at a time until it finds the classification accuracy to be unacceptable. The purpose is to determine the importance of the features in a multi-variant analysis framework and to identify the changes in important features across different years.

Supervised machine learning model training

Once model features were selected, the team investigated several model options as well as choices of time-horizons. Classification models considered were C5 tree, Logistic Regression, Decision List, Bayesian Network, Discriminant, Support Vector Machine, C&R (Classification and Regression) Tree, Quest Tree, CHAID Tree, and Neural Networks.

The classification accuracy curve for this problem was monotonically decreasing, which means that each attribute contributes to the classification accuracy positively in most cases. Another interesting observation was that the rank of the feature importance differs from year to year, with only a few attributes appearing at the top of the list across different years.

Outcome

We used the historical fraud investigation data from 2003 to 2013. The positive cases include 9,688 cases which is about 0.3% of the total investigation cases. The data set is used to train and validate various supervised machine learning models. The models have been validated using

Table 2 Best models for fraud theft detection with different validation methods.

	Cross-fold validation	*Sliding window within year*	*Sliding window year to year*	*Sliding window validation over multiple years*
Best models with high accuracy	SVM and Bayesian network	SVM	C5 tree	Harmonic model

Table 3 Outcome summary.

Use cases	*Energy savings potential modeling*	*Photovoltaic adoption modeling*	*Energy theft detection modeling*
Key attributes for model training	Building characteristics, demographic attributes, lifestyle attributes, energy consumption, and normalized energy consumption	Building characteristics, income level, location	Building characteristics, income and education level, energy consumption, normalized energy consumption, and location
Outcomes	24 customer segments with 5.2% average savings potential	0.788 AUC and 0.005 cumulative adoption hazard value in 5 years	61%–72.3% detection accuracy
Best model combination	Kohonen clustering with descriptive statistics on energy consumption	Cox hazard analysis and Logistic Regression	C5.0 decision tree with historical theft investigation data
Model building recipe	Step 1. Apply the Kohonen clustering algorithm Step 2. Define a normal energy consumption in each of clusters, and calculate the achievable energy consumption Step 3. The reference energy consumption is the average energy consumption of each cluster after removing outliers	Step 1. Build an initial logistic regression model with all demographic variables Step 2. Based on the initial model, merge levels into groups according to effect on outcome variable (PV adoption) Step 3. Using a training set, build a model on the grouped variables Step 4. Test the predictive power of the model on the held-out testing set	Step 1. Apply a feature selection method using the historical energy theft investigation data Step 2. Train a C5.0 decision tree model using the chosen attributes with the training data Step 3. Validate the performance of the trained model Step 4. Apply the model to the rest of the population

several validation methods, including cross-fold validation, sliding window model validation, and ROC analysis. The SVM and Bayesian Network models performed the best in the cross-fold validation, and the SVM performed the best in the sliding window, within a year time frame. The C5 tree and harmonic models performed the best for the year to year sliding window validation. It is interesting to note that the best performing models are different for each validation scenario. However, each validation scenario shows that an application of the machine learning models is feasible with high accuracy (**Table 2**).

Among all the possible models, the most practical model is the C5 decision tree model for year-to-year detection. The recipe for building the model is as follows:

Step 1 Apply a feature selection method using the historical energy theft investigation data.
Step 2 Train a C5.0 decision tree model using the chosen attributes with the training data.
Step 3 Validate the performance of the trained model.
Step 4 Apply the model to the rest of the population.

Conclusion

A deeper understanding of customer behavior is key to successful operation of electric grids. While a variety of data sources exist, their use is made more difficult by their fragmented nature, partial observations, noise in the data, duration of time since the last update, and difficulty in data curation. We investigated the feasibility of building a framework that integrates multiple data sources to build applications for several use cases.

In the course of the SERI engagement with Alliander, several models were built and trained to accommodate a

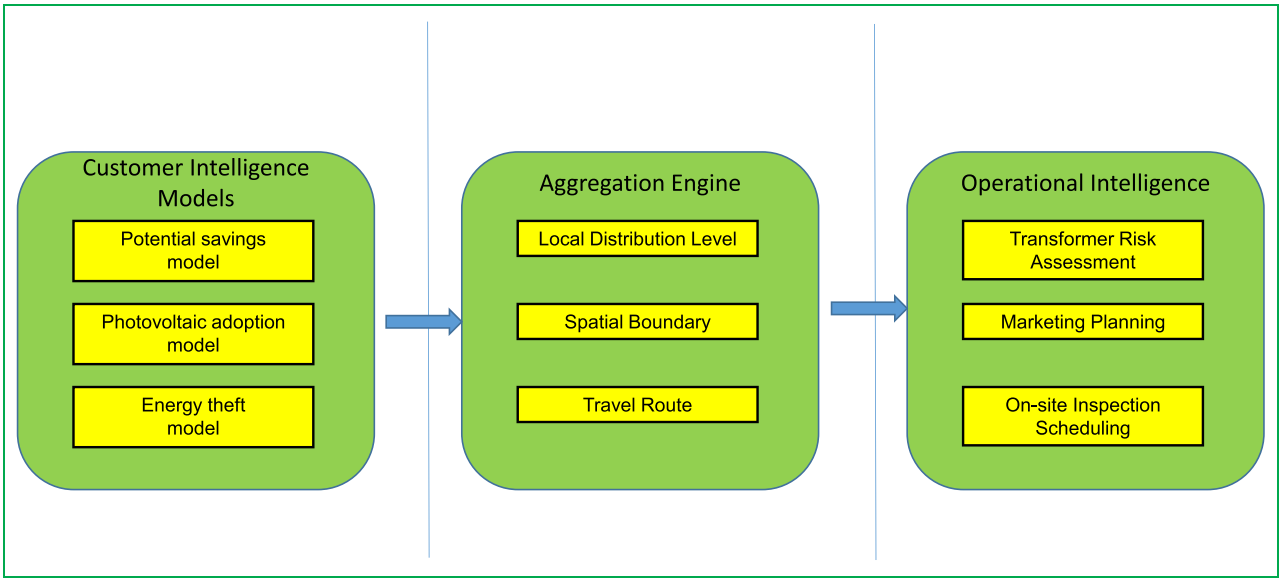

Figure 6

Customer models to operational intelligence.

range of business needs, including energy saving potential modeling, photovoltaic adoption forecast modeling, and fraud detection modeling. In all three cases, we were able to build machine learning models to address business goals, using temporally coarse-grained and low-quality data. Cross-validation of the models showed that it is feasible to add predictive and business value using only existing partial data sources. **Table 3** summarizes the best practices and outcomes for each use.

In addition to the use cases presented, we believe further applications can be built using the same framework. One such application of interest is the detection of appliance ownership information, including electric heat pumps, electric water heaters, refrigerators, and heating ventilation and air conditioning systems (HVACS).

A viable modeling solution must be translatable to operational outcomes. In particular, aggregation of customer behavior variables by meaningful distribution grid levels and geographical levels requires a sufficiently detailed level of electric connectivity information. **Figure 6** illustrates a logical flow diagram of analytics outcomes from a customer understanding framework applied to grid operations. The process begins with building customer models (including selection of customer attributes for each area of interest). Predictions from these models are then accumulated at functionally useful geographic levels, and culminate in actionable items that address particular issues or problems. For example, a photovoltaic (PV) adoption model can be used to aggregate PV predictions at local levels to predict PV adoption within transformer locations; once this information is available, decisions can be made as to which transformers to upgrade to prepare for PV adoption effects.

Modeling of customer behavior has become more of a problem in the age of big data, and many avenues exist for future research and development. Customer modeling is crucial for various contexts. For the energy and utilities industry, however, the use of the customer behavior, in the near future, is centered around understanding of the impact of customer behavior on the stability of electric grids. Increasingly, utility industries are using smart meter data for many activities. Combined with smart meter data analytics, customer behavior analytics in the energy and utilities industry poses a unique set of challenges. One of the key requirements of such analytical service is the usability of the analytics. For this, a good user interface, an acceptable computation latency, and system scalability are necessary. In particular, the data size for a simple task is generally very large. For example, for a simple task, analytics needs to process 9.70 billion data points for 2.97 million customers of Alliander. Therefore, scalability and latency of modeling algorithms for customer behavior become key considerations going forward.

References

1. D. P. Buse, P. Sun, Q. H. Wu, and J. Fitch, "Agent-based substation automation," *IEEE Power Energy Mag.*, vol. 1, no. 2, pp. 50–55, Mar./Apr. 2003.
2. W. R. Cassel, "Distribution management systems: Functions and payback," *IEEE Trans. Power Syst.*, vol. 8, no. 3, pp. 796–801, Aug. 1993.

3. J. C. Boemer, K. Burges, P. Zolotarev, J. Lehner, P. Wajant, M. Fürst, R. Brohm, and T. Kumm, "Overview of German grid issues and retrofit of photovoltaic power plants in Germany for the prevention of frequency stability problems in abnormal system conditions of the ENTSO-E region continental Europe," in *Proc. 1st Int. Workshop Integr. Solar Power Power Syst.*, Aarhus, Denmark, 2011, pp. 1–6.
4. B. E. Hayes, *Measuring Customer Satisfaction: Survey Design, Use, and Statistical Analysis Method*. Milwaukee, WI, USA: ASQ Quality Press, 1998.
5. M. Ture, F. Tokatli, and I. Kurt, "Using Kaplan-Meier analysis together with decision tree methods (CRT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients," *Expert Syst. Appl.*, vol. 36, no. 2, part 1, pp. 2017–2026, 2009.
6. J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
7. K. Gurney, *An Introduction to Neural Networks*. London, U.K.: UCL Press, 1997.
8. T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, Jan. 1982.
9. R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.
10. F. M. Bass, T. V. Krishnan, and D. C. Jain, "Why the Bass model fits without decision variables," *Market. Sci.*, vol. 13, no. 3, pp. 203–223, 1994.
11. S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," *Critical Inf. Infrastructures Security*, vol. 6027, pp. 176–187, 2010.
12. P. McCullagh, "Generalized linear models," *Eur. J. Oper. Res.*, vol. 16, no. 3, pp. 285–292, 1984.

Received April 1, 2015; accepted for publication May 1, 2015

Younghun Kim *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (kimy@us.ibm.com)*. Dr. Kim is a technical lead and IBM Research Staff Member of the Smarter Energy Research Group at the IBM T. J. Watson Research Center. He received his Ph.D. degree in electrical engineering from the University of California, Los Angeles, in 2010. He completed his M.S. and B.S. degrees in the Electrical Engineering Department from Seoul National University in 2006 and 2004, respectively. Since October 2010, Dr. Kim has been with IBM. He leads the predictive analytics research stream for smarter energy research, including utility customer analytics, asset health and management analytics, and outage management analytics. His research interests are in the field of smart grids, wireless sensor networks, computer-human interaction, and sustainability research areas. His work won the best paper awards from the ACM SenSys (Sensor Systems) 2008 Conference and from ACM CHI (Human-Computer Interaction) 2013 Conference, and he won the Computational Sustainability Award from the CCC (Computing Community Consortium). His work on smarter energy conservation received the IBM Research Division Award.

Aleksandr Aravkin *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (saravkin@us.ibm.com)*. Dr. Aravkin is a Research Staff Member at the IBM T. J. Watson Research Center and an adjunct professor at Columbia University. He received his Ph.D. degree in optimization and M.S degree in statistics from the University of Washington in 2010. His current research interests include convex and variational analysis, algorithm design with applications to data science, robust and sparse inference and learning, dynamic systems, PDE (partial differential equation) constrained optimization, log-linear models, and matrix decomposition and recovery problems.

Hongliang Fei *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (hfei@us.ibm.com)*. Dr. Fei received a B.S. degree in computer science from the Harbin Institute of Technology, China, in 2007 and a Ph.D. degree in computer science from the University of Kansas in 2012 with honors. He is now working as a Research Staff Member at the IBM T. J. Watson Research Center. His research focuses on leveraging structured input/output information from data for enhancing supervised and unsupervised learning algorithms, including sparse learning, feature selection, anomaly detection, and multitask learning. He received the ICDM (International Conference on Data Mining) 2011 Best Student Paper Award, and CIKM (Conference on Information and Knowledge Management) 2009 best paper runner-up award, and he serves as a reviewer for leading journals including *IEEE Transactions on Knowledge and Data Engineering*, *ACM Transactions on Knowledge Discovery from Data*, *Pattern Recognition,* and *Bioinformatics*.

Arjen Zondervan *Alliander N.V., Arnhem, The Netherlands (arjen.zondervan@alliander.com)*. Dr. Zondervan works as a Data Scientist and Business Data Analyst in the Customer & Market domain of Alliander, a power grid operator in the Netherlands. He provides insights from large datasets of both internal and external customer data, using statistical techniques, data mining, and predictive modeling. As part of the SERI-project, a worldwide analytics collaboration of utilities facilitated by IBM Research, he is the product owner for the development of predictive models for renewable energy adoption and fraud detection based on customer data. His other activities include analyses of energy savings potential, forecasting of customer demand, and providing marketing insights from customer data.

Maarten Wolf *Alliander, N.V., Arnhem, The Netherlands (maarten.wolf@alliander.com)*. Dr. Wolf received his M.Sc. degree in physical chemistry from the University of Amsterdam, The Netherlands, in 2003 and his Ph.D. degree in biotechnology from the University of technology Delft, The Netherlands, in 2008 *cum laude*. After staying at the Max-Planck institute for biophysical chemistry he joined Alliander as a consultant in 2013. He now uses a range of modeling techniques to address important issues to Alliander and stimulates the transition to an efficient data driven distribution service operator. As part of the SERI project, a worldwide analytics collaboration of utilities facilitated by IBM Research, he is the product owner for the development of predictive models for renewable energy adoption and fraud detection based on customer data.