ENERGY DATA ANALYTICS FOR IMPROVED RESIDENTIAL SERVICE QUALITY AND ENERGY EFFICIENCY

Research in Progress

Hopf, Konstantin, University of Bamberg, Germany, konstantin.hopf@uni-bamberg.de
Sodenkamp, Mariya, University of Bamberg, Germany, mariya.sodenkamp@uni-bamberg.de
Kozlovskiy, Ilya, University of Bamberg, Germany, ilya.kozlovskiy@uni-bamberg.de

Abstract

Utility companies generally have an extensive customer base, yet their knowledge about individual households is small. This adversely affects both the development of innovative, household specific services and the utilities' key performance indicators such as customer loyalty and profitability. With the goal to overcome this knowledge deficit, persuasive systems in the form of customer self-service applications and efficiency coaching portals are becoming the getaway of data exchange between utility and user. While improved customer interaction and the collection of customer data within respective information systems is an important step towards a service-oriented company, the immediate value generated from the collected data is still limited, mostly due to the small fraction of customers actually using such systems. We show how to utilize the knowledge gained from the sparse number of active web users in order to provide low-cost and large-scale insights to potentially all residential utility customers. We do so using machine-learning-based Green IT artifacts that allow for improving decision-making, effectiveness of energy audits, and conservation campaigns, thus ultimately increasing the customer value and adoption of related services. Moreover, we show that data from the publically available geographic information systems can considerably improve the decision quality.

Keywords: Energy Data Analytics, Household Characteristics, Green Information Systems, Volunteered Geographic Information (VGI)

1 Introduction

In both the academic literature and the practitioner sphere, much attention is currently being paid to the value that organizations could create through the use of big data and business analytics (Constantiou and Kallinikos, 2015; Gillon et al., 2012; Mithas et al., 2013; Sharma et al., 2014). This especially holds for energy utilities for which a plethora of new efficiency regulations, new market models, and increasing customer expectations regarding a clean energy supply result in severe pressure on their revenues (Gebauer et al., 2014). Consequently, utility companies regard an intensive customer engagement as an important means to protect their existing customer base, to tap new sources of growth, and to establish new business models.

Despite their large and valuable customer base, the knowledge of utility companies about individual customers is small, especially when compared to the retail or online service sector. Data analytics and machine learning may help utilities to overcome this information deficit and to improve their key performance indicators. In this context, energy efficiency campaigns and automated home audits are seen as a door opener that helps to collect initial customer insights. Recently, Loock et al. (2013) have shown that specific interventions are extremely valuable for energy consultancies, e.g., to identify households that show a mismatch between energy demand and household characteristics. Customer insights help to formulate suitable saving advice that reflect disposable income, appliance structure,

etc., and to design targeted motivational cues that engage customers into energy efficiency campaigns. The initial success of such systems led to ever-more persuasive systems, i.e., websites with the purpose of motivating people to change their attitudes and behaviors (Fogg, 1998; Graml et al., 2011), while establishing a better customer interaction and collecting customer data. In the same vein, multiple studies have shown that home energy audits can bring benefits to the households in terms of energy efficiency and monetary savings (Kastner and Stern, 2015; Palmer et al., 2012).

Yet, for a variety of reasons and despite the early adoption of lead users, the great majority of utility customers do not to take advantage of these service offerings (Sodenkamp et al., 2015). Consequently, the existing portals alone may be successful in collecting in-depth data about a fraction of users but do not suffice to collect data about the large share of passive customers.

In this work, we show how utility companies can use the existing data from the small number of active portals users to infer customer specific insights also about the users who do not subscribe to such online offerings. We build upon previous work (Sodenkamp et al., 2015) and provide new insights from a field study in cooperation with a university spinoff-company that develops and runs efficiency and customer insights software for about 30 European utility companies. In greater detail, we show how IT artifacts that utilize machine learning tools can help utility companies to transfer the customer knowledge gained from the limited number of portal users to their entire customer base. Energy consumption data that is available for billing purposes underlies the classification. This leads to improvement of the effectiveness and timeliness of energy consultancies, conservation campaigns, empowers decisions about personalization of products and services, and increases cost-efficiency of marketing initiatives at the large scale at low cost.

Thus, the overall goal of our research is to ***investigate how machine learning artifacts can be used in combination with state-of-the-art customer engagement portals and publicly available volunteered geographic information (VGI) to infer customer specific information on the entire customer base of a utility.***

We aim to answer the following research questions:

Research question 1 (RQ1): *Is it possible to use machine learning to create an IT artifact able to find energy-efficiency related household characteristics from the yearly electricity consumption data?*

Research question 2 (RQ2): *Can volunteered geographic information (VGI) improve the predictive power of the IT artifact for household classification?*

From the technical perspective, plenty of research about utility customer segmentation based on electricity consumption exists. Researchers apply clustering methods (such as Self-Organizing Maps or k-means) to identify groups of similar consumption pattern (Figueiredo et al., 2005; Kwac et al., 2013; Räsänen et al., 2008; Sánchez et al., 2009; Verdu et al., 2006). The resulting clusters need then manual interpretation of an expert wherefore such methods cannot be applied in IT at scale.

Research on non-intrusive load monitoring (Hart, 1992; Zeifman and Roth, 2011) showed, that the identification of specific appliances is possible using high-frequency meter readings (multiple measurements per second). Due to the high data volume and privacy aspects (Quinn, 2009), such fine-grained information is not appropriate for use in utility companies for practical purposes.

Based on 15- or 30-minute smart meter and customer survey data, methods have been proposed to predict household characteristics such as age of house, number of appliances, family and social status, etc. (Beckel et al., 2013; Hopf et al., 2014; Sodenkamp et al., 2014; Beckel et al., 2014). These methods are promising, but due to the current status of European smart grid infrastructure (Einhellig et al., 2014), the majority of households is still equipped with conventional electricity meters. Therefore, we transfer household classification methods and apply it to broadly available yearly electricity consumption data.

In a previous study, our team showed recently (Sodenkamp et al., 2014), that the prediction of customer participation on an energy efficiency web portal is feasible based on annual electricity consumption and address information. In this work, we extend a similar prediction approach to infer three household characteristics (household type, living area and number of residents).

Besides address information and electricity consumption data, we rely on publically available data

sources. The first reason for that is a growing availability of open public sector information (Arzberger et al., 2004; Janssen et al., 2012). The direct and indirect economic value of such data sources in the European Union is considered in the order of EUR 140 billion annually (Vickery, 2011). The second reason is, that currently a large amount of volunteered geographic information (VGI) data have been collected and published from by non-professional individuals (Goodchild, 2007; Sester et al., 2014). These individuals have varying expertise and contribute their work to a community. Therefore, the data is free of charge and very attractive for companies. Examples for VGI projects are world mapping projects (incl. OpenStreetMap, Wikimapia), social media (incl. Flickr, Twitter, Facebook, Youtube), environmental, ecological, disaster, crime and outdoor activity mapping, such as business directories.

The rest of the paper is organized as follows: In section 2 we describe our research methodology including our developed IT artifact with the core components. In section 3 we show how the supervised machine learning artifact can be operationalized to answer both research questions. Finally, in section 4 we give a conclusion and name future research topics.

2 Energy data analytics methodology

Our research is designed using the design science approach (Hevner et al., 2004) with the principles underlined by Peffers et al. (2007). Our focus lies thereby on building a supervised machine learning IT artifact where we develop a novel artifact in order to solve an existing organizational problem.

Our IT artifact is schematically illustrated in Figure 1. As an input, we consider household annual electricity consumption data and the address (both available to utilities for billing purpose). As an output, additional customer information in the form of household characteristics is obtained (household type, living area, etc.). This information is known for a part of the customers who use the engagement portal and represents a ground truth underlying our supervised machine learning model. The data is normalized, prepared and enriched with VGI data in a feature extraction component.

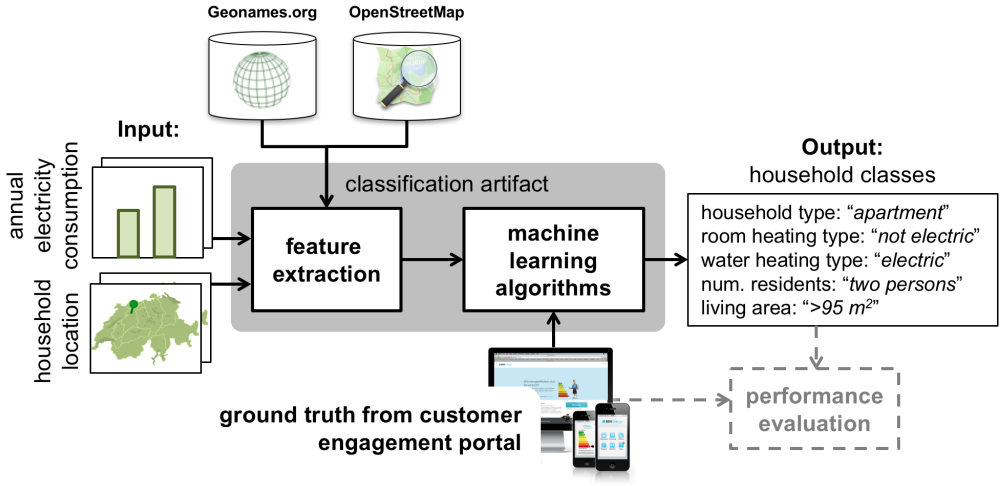

Figure 1: Household class prediction methodology

2.1 Data and variables

For the algorithm training and evaluation, we rely on a customer dataset from a Swiss utility company provided by our praxis partner BEN Energy AG in an anonymized way. The complete dataset (we name it A) encompasses 10'482 customers. For each customer, the household address (street, postal code and city) and the annual electricity consumption from the years 2009 – 2012 (four cumulated values) are available. The subsample $B \subset A$ ($|B| = 3'986$) consists of customers who use a customer engagement portal and completed an online-survey about household characteristics there. The relevant variables of B for our research are listed in Table 1. We assume that the sample B is representative of the whole set of customers A. We find evidence for this assumption in the fact that the electricity consumption of A and B follow the same distribution (Kolmogorov-Smirnov test statistic with $D = 0.0284$ and $p \leq 0.0001$).

Variable	Type	Description
ID	categorical	Unique customer identifier
PLZ	categorical	Billing address, postal code
City	categorical	Billing address, city
Street	categorical	Billing address, street and street number
Cons_2009, Cons_2010, Cons_2011, Cons_2012	numerical	Electricity consumption in one year (min: 0, median: 4483, max: 25'680)
Days_2009, Days_2010 Days_2011, Days_2012	numerical	Number of days in which the electricity consumption was created (min: 0, median, 356, max: 366)
HouseholdType	categorical	Type of the household (alternatives: apartment, house)
LivingAreaM2	numerical	Living area in m^2 (min: 10, median: 125, max: 5'443)
NumResidents	numerical	Number of persons living in the household (min: 1, median: 2, max: 10)

Table 1: Variables in the customer dataset

2.2 Definition of properties (dependent variables)

Household *properties* are derived from the survey data. In this work, we consider three properties that are defined as follows:

pHouseholdType - The variable household type was raised with the categories 'apartment' and 'house'. We use these classes directly in our IT artifact.

pLivingArea - The variable living area takes integer values in the range of 10 to 5'443. Therefore, any definition of this property is ambiguous. We defined the class borders at 95 m^2 and 145 m^2 based on the following motivation: First, the class borders are empirically defined and based on quantiles. The 33% quantile is 100m^2, the 66% quantile is 150m^2, and the 99% quantile is 400m^2. Since we assume that people estimate their living area in a survey to the next upper bound, we define the categories 5m^2 below this round number. Second, we find further evidence in our class definition in European statistics (Statistical Office of the European Communities, 2014, p. 54): the average dwelling size in the EU-28 countries is 95.9 m^2, in Switzerland it is according to the statistics 117.1 m^2.

pNumResidents - The number of residents in a household takes fewer values than the living area, but the variable has nevertheless a range of 1 to 10 household and the class borders can be defined ambiguously. We tested a set of definitions in the classification: a) 1 / 2 / >2, b) 1 / 2 / 3-5 / >5, c) 1 / 2 / 3 / 4 / >4, d) 1 / >1. Our results show that the definition (b) has the best trade-off between gained information, number of classes and classification performance. Therefore, we include only this definition in this paper.

2.3 Definition of features (independent variables)

The definition of features is a crucial step during the development of a machine learning model and the quality of features is a key success factors for the classification performance. In the considered classification problem, we use two sources for the feature definition: 1) consumption data and 2) VGI. Both are described in detail below.

2.3.1 Electricity consumption features

We use features that represent the electricity consumption of one household, its development over time and the consumption compared to the neighborhood of the household. Thereby, we rely on electricity consumption features that showed a high importance (Sodenkamp et al., 2015).

1. The annual consumption normalized by the consumption days:
   Mean daily consumption = Total annual consumption / Number of consumption days

   We apply the log transformation to achieve a symmetric distribution of the variable. Since the consumption of different years correlates highly (correlation coefficient $\rho \geq 0.87$), we use the mean value of all years instead of including features for every year.
2. The consumption trend as the relative change between the consumption of different years is used, obtained with a linear regression model of the four years of consumption.

3. We use neighborhood comparison as a feature. To obtain this, we calculate the mean logarithm consumption and standard deviation in the postal code region and calculate the Z-score of the household's consumption deviation from it's neighborhood.

2.3.2 Features from volunteered geographic information (VGI) systems

Using the address information of one household, the IT artifacts retrieves geographic information from two popular VGI projects: OpenStreetMap and GeoNames.org.

GeoNames.org is a large collection of geographic places. The database contains currently over 10 million geographical names and their locations worldwide. We use the data source to obtain the distances to nearby city centers with a population of 1'000, 5'000 and 15'000.

OpenStreetMap.org (OSM) is the largest map maintained by volunteers (Ballatore et al., 2013). OSM contains vector-based geographical information that consists of points and lines between points, annotated with semantic information.

The calculation of features is schematically illustrated in Figure 2. In total, we implemented 66 geographic features from the available data in the VGI projects. We adopted spatial landscape metrics from geographic information science (Baskent and Jordan, 1995; Gustafson, 1998) to define the features.

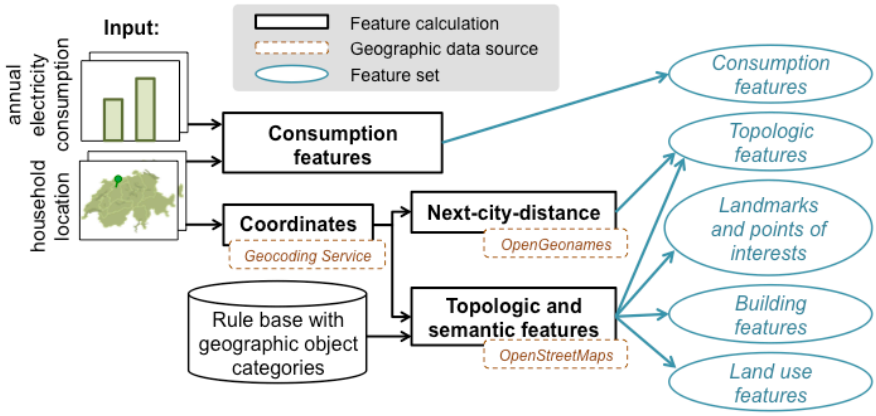

Figure 2: Overview to the feature calculation process, information sources and feature sets

All geographic features can be subsumed under four categories:

1) *Topologic features*: describing the structure of and relations between one household and spatial neighbors (e.g. lon./lat., frequency objects in the surroundings, distance to city center)
2) *Landmarks and points of interests*: Meaning of an object within the spatial context it appears (frequency, distance, and other measures to sights, public institutions, shops, cafes, etc.)
3) *Features about buildings* (e.g. mean/variance of the surface area, the distance to buildings, and the type of buildings in the surrounding)
4) *Features about land use* (land use type embracing the household, area distribution in different land use types, etc.)

2.3.3 Correlation based feature selection

Since the number of features (3 consumption features and 66 geographic features) would hamper the classification performance due to the course of dimensionality (Guyon et al., 2003), we applied correlation-based feature selection (Hall, 1999) with the implementation of Romanski and Kotthoff (2014) before the classifier training.

In our evaluation, the logarithmized mean daily consumption and the the lat./lon. coordinates of the household have been mostly selected by the CFS method for all properties. Beside these features, the area of the next building was selected for pHouseholdType and pNumResidents. The number of public institutions and the existence of public transportation showed a high importance for pLivingArea and pNumResidents.

2.4 Classification algorithms

We tested four classifier for their suitability in our classification problem and implemented our IT artifact in the statistical programming environment GNU-R. The algorithms are explained shortly:

k Nearest Neighbors (kNN) infers the class by considering those *k* training instances with the lowest Euclidean distance to the example that is to be classified. We use normalize all variable inputs to a codomain of [0;1], because the kNN classifier is sensitive to the ranges of the input variables (Han et al., 2012). We rely on the implementation of Wing et al. (2015) and tested various values for *k* and found out that *k=15* work best for our classification problem.

Support Vector Machine (SVM) was proposed by Vapnik and Vapnik (1998). The algorithm searches for a hyper plane in the vector space that separates all training examples with a maximal margin. In the case of not separable training data, a kernel-function is used that transforms the training vector into a higher dimension. We tested four different kernels, varied the parameters of SVM in 317 configurations, and found that the radial basis kernel having a coefficient of 50 and a cost of misclassification parameter of 50 leads to the best results with the SVM-implementation of (Meyer et al., 2014).

Random Forest (RF) – This algorithm generates multiple low correlated decision trees that are learned and evaluated with ensemble methods (Breiman, 2001). In our artifact, we use the implementation of (Hothorn et al., 2006; Strobl et al., 2008).

Naïve Bayes (NB) – Bayesian classifier predict the class membership based on a probability that a given data point belongs to the class. The probabilities needed for this prediction are calculated by means of the Bayes' theorem. In our analysis we use the implementation of (Meyer et al., 2014).

We choose kNN and SVM because of good performance in previous works (Beckel et al., 2014; Sodenkamp et al., 2014). Since both algorithms are unsuitable to handle categorical features (e.g. building type, land-use type), we additionally considered RF and NB.

3 Demonstration and evaluation of the methodology with real-world data

In this section we operationalize our IT artifact to address both research questions. To quantify its performance, we count the number of correct and misclassified examples in comparing the predicted household classes with ground truth data (as described in Section 2.1). To obtain a robust calculation of the performance measures, we use 5-fold cross-validation (Stone, 1974) and calculate two performance measures: Precision and Accuracy. However, a number of 10 folds is considered as ideal for general classification problems (Kohavi, 1995; Kuzey et al., 2014), but in our case we choose 5-fold cross-validation, because of the class distribution of the household properties (we took into account for at least 20 examples to belong to each class for evaluating the performance).

Precision (Pr) is a measure for *one single class* and quantifies the amount of correct classified examples (true positives) among the positively predicted examples (true positive and false positive):

$$Pr = true\ pos. / (true\ pos. + false\ pos.)$$

The codomain of precision is [0;1], where 0 indicates that no household was correctly predicted, and 1 means that all positive predicted households belong truly to the predicted class.

Accuracy (Ac) is a measure for *one household property* that quantifies the correct classified examples (true positives and true negatives) among all examples:

$$Ac = true\ pos. + true\ neg. / (true\ pos. + true\ neg. + false\ pos.\ false\ neg.)$$

We compare the classification accuracy with the size of the biggest class as a proxy for randomly guessing the right class. Both performance measures are commonly used in classifier evaluation (Han et al., 2012; Sokolova and Lapalme, 2009).

RQ 1: Is it possible to use machine learning to create an IT artifact able to find energy-efficiency related household characteristics from the yearly electricity consumption data?

To answer this question, we look at the classification accuracy for all properties as depicted in Figure 3. The precision results for all classes are included in Figure 4. Both plots show the average classification accuracy and precision calculated with the cross-validation for each classifier (4 different symbols) and with three different feature sets: consumption features (red / leftmost symbols) consumption and VGI features (green / symbols in the middle), and only VGI features (blue / rightmost symbols).

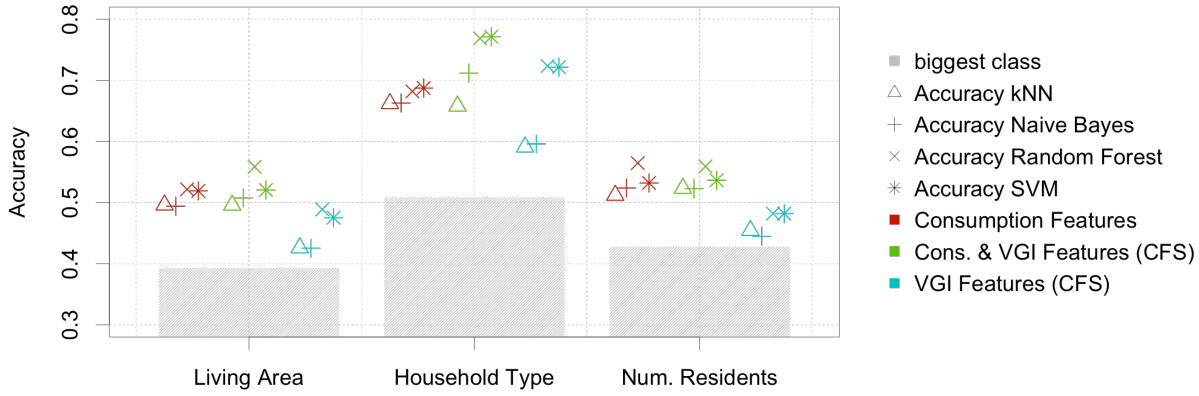

Figure 3: Average classification accuracy with consumption and geographic features, compared with the class size as minimal benchmark for the classification performance

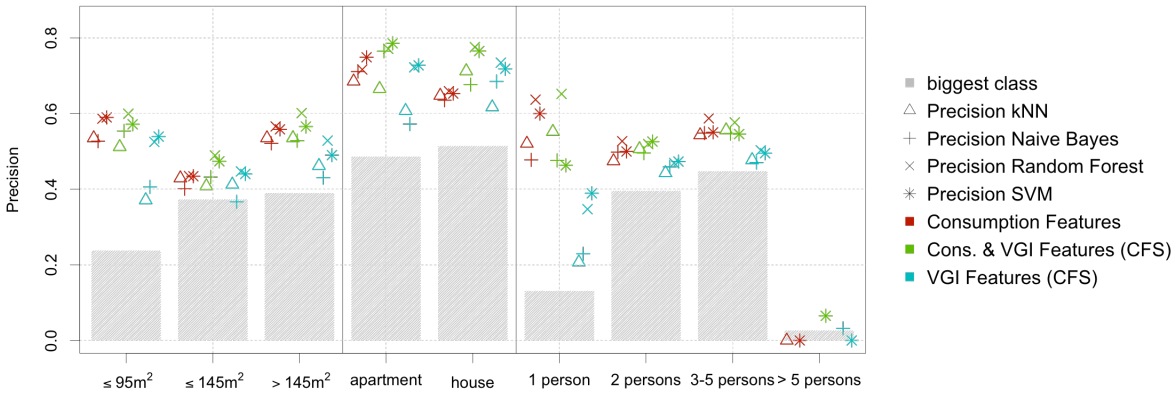

Figure 4: Classification precision for each class with consumption and geographic features, compared with the class size as minimal benchmark for the classification performance

The classification results show that supervised machine learning based on electricity consumption data and the household address (consumption features) can predict household classes with an accuracy between 49.4% and 68.7%. This is, averaged over all properties and classifier 28.7%, higher than a random guessing of the biggest class. On single-class level we can see a significant improvement in precision, too: the average improvement of the classification with consumption features is 80%. The largest achievement of classification is thereby the recognition capability of small classes (small dwellings with lower than 95m^2 floor area and single households).

Our first research question can therefore be answered positively: Information on customers that are active on an energy-efficiency web portal can be used to enrich customer data of non-portal users.

RQ 2: Can VGI improve the predictive power of the IT artifact for household classification?

To answer this question, we compare the classification results not only with the class sizes, but also with the classification results of our classification with solely consumption features.

For property *living area* we see an improvement of 12.55% due to the addition of VGI features with the RF classifier (from Pr 43.47% to Pr 48.92% in class '95m^2 – 145m$^{2'}$). Especially households with a size > 95m^2 can be recognized with higher performance by using geographic features and RF / SVM classifier. The largest improvements in classification with VGI data can be found in property *household type* ('apartment' by 7.63% and 'house' by 17.68% with the RF classifier). There is no actual improvement of VGI data for the recognition of the *number of residents*, because the geographic features have not been selected by the CFS method for this classification settings.

It is particularly interesting that small dwellings and the household type can be recognized without any electricity consumption on a high level (decrease in accuracy 6.8% compared to the classification with geographic and consumption data, see blue rightmost symbols in Figure 4). Compared to the relative class size, the property household type can be recognized 42% better than a random guessing (Ac 72.35% compared with 50.87% size of the biggest class). This opens the possibility to use the household classification methodology not only for electric utilities, but also to all companies that deal with large end-use customer data.

We can answer our second research question positively: VGI data can improve the classification of household characteristics. We could achieve a mean improvement of 7.0% for pLivingArea and of 12.7% for pHouseholdType resulting from the use of VGI data.

4 Summary and outlook

In this paper, we have shown the potential of data analytics to becoming a valuable and scalable decision support mechanism for energy utility companies. In particular, we have demonstrated how machine learning artifacts can be used in combination with state-of-the-art customer engagement portals and publicly available data sources to infer household specific information on the entire customer base of a utility. As a result, even a limited number of data points per household (yearly electricity consumption) is sufficient to extract properties of the residencies that are crucial for targeted efficiency campaigns and personalized customer communication. In a nutshell, the developed artifact enables utility companies to gain insights about millions of individual customers within short time and at scale. Furthermore, publically available data (e.g., from volunteered geographic information (VGI) systems) can significantly improve the recognition quality.

Our future work will be dedicated to further investigate the results of the best classification algorithms with respect to interesting patterns between single VGI features. Besides that, we are going to include further open data sources (e.g., public statistics, cadastral data, satellite pictures) in our artifact and apply further feature selection methods to improve classification results. Furthermore, we will expand the number of potentially valuable household characteristics and test applicability of different analytical methods. We also plan to estimate the robustness of the presented methodology using other datasets (customers from other geographical regions). Finally, field studies on the enabled interventions toward selected household classes and their economic and ecological effects will be provided to complete the validation.

Acknowledgments

The research presented in this paper was financially supported by Swiss Federal Office of Energy (Grant number SI/501202-01), and Eureka member countries and European Union (EUROSTARS Grant number E!9859 - BENgine II).

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., Wouters, P., 2004. An International Framework to Promote Access to Data. Science 303, 1777–1778.

Ballatore, A., Wilson, D.C., Bertolotto, M., 2013. A survey of volunteered open geo-knowledge bases in the semantic web, in: Quality Issues in the Management of Web Information. Springer, pp. 93–120.

Baskent, E.Z., Jordan, G.A., 1995. Characterizing spatial structure of forest landscapes. Canadian Journal of Forest Research 25, 1830–1849.

Beckel, C., Sadamori, L., Santini, S., 2013. Automatic socio-economic classification of households using electricity consumption data, in: Culler, D., Rosenberg, C. (Eds.), Proceedings of the Fourth International Conference on Future Energy Systems. ACM, Berkeley and California and USA, pp. 75–86.

Beckel, C., Sadamori, L., Staake, T., Santini, S., 2014. Revealing household characteristics from smart meter data. Energy 78, 397–410.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Constantiou, I.D., Kallinikos, J., 2015. New games, new rules: big data and the changing context of strategy. J Inf technol 30, 44–57. doi:10.1057/jit.2014.17

Einhellig, L., Behrens, K., v. Preysing, L., 2014. Einführung von Smart Meter in Deutschland. Deutsche Energie - Agentur GmbH, Berlin.

Figueiredo, V., Rodrigues, F., Vale, Z., Gouveia, J.B., 2005. An electric energy consumer characterization framework based on data mining techniques. IEEE Transactions on Power Systems 20, 596–602.

Fogg, B.J., 1998. Persuasive computers: perspectives and research directions. Presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press/Addison-Wesley Publishing Co., pp. 225–232.

Gebauer, H., Worch, H., Truffer, B., 2014. 5 Value Innovations in Electricity Utilities. Framing Innovation in Public Service Sectors 30, 85.

Gillon, K., Brynjolfsson, E., Mithas, S., Griffin, J., Gupta, M., 2012. Business Analytics: Radical Shift or Incremental Change?

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.

Graml, T., Loock, C.-M., Baeriswyl, M., Staake, T., 2011. Improving Residential Energy Consumption at Large Using Persuasive Systems. 19th European Conference on Information Systems (ECIS), 19th European Conference on Information Systems (ECIS), Helsinki, Finland, June 2011.

Gustafson, E.J., 1998. Quantifying Landscape Spatial Pattern: What Is the State of the Art? Ecosystems 1, 143–156. doi:10.1007/s100219900011

Guyon, I., André, Elisseeff, 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hall, M.A., 1999. Correlation-based feature selection for machine learning. The University of Waikato, Hamilton, New Zealand.

Han, J., Kamber, M., Pei, J., 2012. Data mining: Concepts and techniques, 3. ed, The Morgan Kaufmann series in data management systems. Elsevier, Amsterdam.

Hart, G.W., 1992. Nonintrusive appliance load monitoring. Proceedings of the IEEE 80, 1870–1891. doi:10.1109/5.192069

Hevner, A.R., March, S.T., Park, T., Ram, S., 2004. Design Science in Information Systems Research. MIS Quarterly 28, 75–105.

Hopf, K., Sodenkamp, M., Kozlovkiy, I., Staake, T., 2014. Feature extraction and filtering for household classification based on smart electricity meter data. Computer Science-Research and Development 1–8.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., Laan, M.J.V.D., 2006. Survival ensembles. Biostat 7, 355–373. doi:10.1093/biostatistics/kxj011

Janssen, M., Charalabidis, Y., Zuiderwijk, A., 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management 29, 258–268.

Kastner, I., Stern, P.C., 2015. Examining the decision-making processes behind household energy investments: A review. Energy Research & Social Science 10, 72–89. doi:10.1016/j.erss.2015.07.008

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, in: The Proceedings of the 14th International Conference in AI. Presented at the ICAI, Morgan Kaufmann, San Mateo, CA, pp. 1137–1145.

Kuzey, C., Uyar, A., Delen, D., 2014. The impact of multinationality on firm value: A comparative analysis of machine learning techniques. Decision Support Systems 59, 127–142. doi:10.1016/j.dss.2013.11.001

Kwac, J., Tan, C.-W., Sintov, N., Flora, J., Rajagopal, R., 2013. Utility customer segmentation based on smart meter data: Empirical study, in: Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on. IEEE, pp. 720–725.

Loock, C.-M., Staake, T., Thiesse, F., 2013. Motivating energy-efficient behavior with green IS: an investigation of goal setting and the role of defaults. MIS Quarterly 37, 1313–1332.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2014. e1071: Misc Functions of the Department of Statistics (e1071). TU Wien.

Mithas, S., Lee, M.R., Earley, S., Murugesan, S., Djavanshir, R., 2013. Leveraging Big Data and Business Analytics [Guest editors' introduction]. IT Professional 15, 18–20.

Palmer, K., Walls, M., Gordon, H., Gerarden, T., 2012. Assessing the energy-efficiency information gap: results from a survey of home energy auditors. Energy Efficiency 6, 271–292. doi:10.1007/s12053-012-9178-2

Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems 24, 45–77. doi:10.2753/MIS0742-1222240302

Quinn, E.L., 2009. Privacy and the new energy infrastructure. SSRN Electronic Journal 2009. doi:10.2139/ssrn.1370731

Räsänen, T., Ruuskanen, J., Kolehmainen, M., 2008. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Applied Energy 85, 830–840.

Romanski, P., Kotthoff, L., 2014. FSelector: Selecting attributes.

Sánchez, I.B., Espinos, I.D., Moreno Sarrion, L., Quijano López, A., Burgos, I.N., 2009. Clients segmentation according to their domestic energy consumption by the use of self-organizing maps, in: Energy Market, 2009. EEM 2009. 6th International Conference on the European. pp. 1–6.

Sester, M., Arsanjani, J.J., Klammer, R., Burghardt, D., Haunert, J.-H., 2014. Integrating and Generalising Volunteered Geographic Information, in: Burghardt, D., Duchêne, C., Mackaness, W. (Eds.), Abstracting Geographic Information in a Data Rich World, Lecture Notes in Geoinformation and Cartography. Springer International Publishing, pp. 119–155.

Sharma, R., Mithas, S., Kankanhalli, A., 2014. Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. European Journal of Information Systems 23, 433–441.

Sodenkamp, M., Hopf, K., Staake, T., 2014. Using Supervised Machine Learning to Explore Energy Consumption Data in Private Sector Housing, in: Tavana, M., Puranam, K. (Eds.), Handbook of Research on Organizational Transformations through Big Data Analytics. p. 320.

Sodenkamp, M., Kozlovskiy, I., Staake, T., 2015. Gaining IS Business Value through Big Data Analytics: A Case Study of the Energy Sector. Proceedings of the Thirty Sixth International Conference on Information Systems (ICIS), Fort Worth, USA, 13-16 December.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45, 427–437.

Statistical Office of the European Communities, 2014. Living conditions in Europe: 2014 edition. Publications Office of the European Union, Luxembourg.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological) 111–147.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional Variable Importance for Random Forests. BMC Bioinformatics 9, 307. doi:10.1186/1471-2105-9-307

Vapnik, V.N., Vapnik, V., 1998. Statistical learning theory. Wiley New York.

Verdu, S.V., Garcia, M.O., Senabre, C., Marin, A.G., Franco, F.J.G., 2006. Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps. IEEE Transactions on Power Systems 21, 1672–1682.

Vickery, G., 2011. Review of recent studies on PSI re-use and related market developments. Information Economics, Paris.

Wing, M.K.C. from J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R.C. team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., 2015. caret: Classification and Regression Training.

Zeifman, M., Roth, K., 2011. Nonintrusive appliance load monitoring: Review and outlook. IEEE Transactions on Consumer Electronics 76–84.