Weisser, Christoph; Lenel, Friederike; Lu, Yao; Kis-Katos, Krisztina; Kneib, Thomas

Article

Using solar panels for business purposes: Evidence based on high-frequency power usage data

Development Engineering

Provided in Cooperation with:
Elsevier

Suggested Citation: Weisser, Christoph; Lenel, Friederike; Lu, Yao; Kis-Katos, Krisztina; Kneib, Thomas (2021) : Using solar panels for business purposes: Evidence based on high-frequency power usage data, Development Engineering, ISSN 2352-7285, Elsevier, Amsterdam, Vol. 6, pp. 1-19, https://doi.org/10.1016/j.deveng.2021.100074

Development Engineering 6 (2021) 100074

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

Using solar panels for business purposes: Evidence based on high-frequency power usage data

Christoph Weisser [a,b], Friederike Lenel [a,*], Yao Lu [c], Krisztina Kis-Katos [a,b], Thomas Kneib [a,b]

[a] Georg-August-Universität Göttingen, Göttingen, Germany
[b] Campus-Institut Data, Science (CIDAS), Göttingen, Germany
[c] RWTH Aachen, Aachen, Germany

ARTICLE INFO

Keywords:
Rural electrification
Off-grid energy
High-frequency electricity usage data
Solar panels
Tanzania
Risk management
Credit default
Big data
Supervised machine learning
Time-dependent proportional hazards model
XGBoost

ABSTRACT

Access to electricity is typically the main benefit associated with solar panels, but in economically less developed countries, where access to electricity is still very limited, solar panel systems can also serve as means to generate additional income and to diversify income sources. We analyze high-frequency electricity usage and repayment data of around 70,000 households in Tanzania that purchased a solar panel system on credit, in order to (1) determine the extent to which solar panel systems are used for income generation, and (2) explore the link between the usage of the solar system for business purposes and the repayment of the customer credit that finances its purchase. Based on individual patterns of energy consumption within each day, we use XGBoost as a supervised machine learning model combined with labels from a customer survey on business usage to generate out-of-sample predictions of the daily likelihood that customers operate a business. We find a low average predicted business probability; yet there is considerable variation across households and over time. While the majority of households are predicted to use their system primarily for private consumption, our findings suggest that a substantial proportion uses it for income generation purposes occasionally. Our subsequent statistical analysis regresses the occurrence of individual credit delinquency within each month on the monthly average predicted probability of business-like electricity usage, relying on a time-dependent proportional hazards model. Our results show that customers with more business-like electricity usage patterns are significantly less likely to face repayment difficulties, suggesting that using the system to generate additional income can help to alleviate cash constraints and prevent default.

1. Introduction

In economically less developed countries, where access to electricity is still very limited (World Bank 2017), solar panel systems not only provide electricity for private consumption (D'Agostino et al., 2016), but also offer means to generate additional income. The electricity generated by the system can be used to boost an existing business (e.g., by using lights to allow for longer operating hours of shops, bars, or restaurants) or start a new business (such as a phone charging business or a home cinema) and thereby diversify income sources. As solar panel systems are often financed through credit arrangements (World Bank, 2020), the generated income can further help to repay the investment.

However, so far there is little evidence to which extent households use their solar panel systems for business purposes. Indeed, data on this is difficult to obtain. While solar panel owners can be surveyed and asked directly about their usage behavior, surveys are limited in terms of their scale and are less well suited to track changes of usage over time. Backward looking survey data on past usage behavior cannot fully fill this gap due to reporting and recollection biases (Rom et al., 2020).

In this paper, we use high-frequency electricity usage and credit repayment data in order to study the extent to which solar panel systems are used for business purposes and its implications for repayment behavior. The data was provided to us by a clean energy company that sells solar panel home systems through a flexible credit arrangement in several countries in East Africa (Grohmann et al., 2021). We focus on the daily energy consumption behavior of over 70,000 customers located in Tanzania for a time period of 3.5 years. Relying on customer survey data that allows us to identify prospective business users at the time of the purchase, we first predict each customer's likelihood of using the system for business purposes on a daily basis based on their hourly patterns of electricity usage in the first few months after the purchase and installation of the solar panel. For this purpose, we utilize a machine learning

* Corresponding author.
E-mail address: friederike.lenel@uni-goettingen.de (F. Lenel).

https://doi.org/10.1016/j.deveng.2021.100074
Received 12 September 2021; Received in revised form 19 October 2021; Accepted 9 November 2021
Available online 12 November 2021

algorithm, a so-called classifier or supervised learning method, that links the binary outcome of business or non-business use to a large number of statistics that describe the patterns and variability of daily energy use. We can thereby predict the individual probability of business-like energy usage that varies with changing electricity consumption patterns over time. We then study whether customers whose electricity usage patterns suggest business use are better able to repay the loan for their solar panel home system by linking the average monthly predicted probability of using electricity for business purposes to the monthly likelihood of credit non-repayment.

To predict the likelihood of a customer being a business or private user at a daily basis, we use power usage data recorded in real time by a sensor that each system is equipped with. We aggregate this high-frequency data at the hourly level and first generate variables (so-called features) that capture relevant dimensions of electricity usage (among others, its average intensity over time, its variance as well as its hourly dynamics). Our machine learning approach relies on a supervised classifier, XGBoost (Extreme Gradient Boosting (Chen and Guestrin, 2016)), that links customer-day observations of these features to an indicator of potential business usage. We identify possible business users based on large-scale survey data that is collected by the company as part of its due diligence before a customer can be provided with a loan and that elicits a customer's intended purpose of the system. The indicator of intended business usage becomes the pre-defined label for the training dataset, which contains electricity usage data for the first months after the system was installed. Since our target variable, the label, is well-specified, this approach is referred to as supervised learning; in contrast, unsupervised learning methods are applied when no pre-specified target variable (here, no information on possible business usage) exists. Subsequently, we use XGBoost for out-of-sample predictions of the individual likelihood of being a business user at a daily level for all customers and throughout the whole time period. Out-of-sample predictions over a test dataset (observations from the initial time period that were not used for training) provide us with metrics for the quality of the prediction itself. In a last step, we study the implications of business usage for repayment by relating the occurrence of a customer becoming delinquent to the monthly average probability of business usage of electricity in a time-dependent Cox proportional hazards model (Therneau and Grambsch, 2000). An overview of the process is shown in Fig. 1.

We show that such a supervised classification approach to capture electricity usage behavior can be implemented as long as good-quality labelled data exists. Although only less than 8% of the customers in our sample report to intend to use the solar panel for business at the time of its purchase, a substantially larger share of households shows electricity usage behavior at some later point in time that is associated with income generating activities. On average, 23% of a customer's usage days are predicted to be business days with at least 10% probability. This corroborates evidence from smaller customer surveys that show that up to a quarter of all households may operate businesses at a point in time. Furthermore, we show that the predicted business usage probability of each household is statistically significantly related to credit repayment behavior. In particular, we find that the average predicted likelihood of business use within each month is negatively correlated with credit delinquency, conditional on socio-economic characteristics and the average intensity of electricity use. Households that are more likely to use their system to generate additional income thus face less difficulties in repaying their loan.

This study makes three major contributions. First, we show that electricity load profiles from solar panel systems can be used to classify customers into business and non-business users. A number of studies use classification and clustering algorithms in order to investigate customer segmentation based on electricity load profiles. These studies are focused almost exclusively on industrialized countries. For instance, (Zufferey et al., 2012) show how machine learning approaches can be used to detect the type of electrical home appliance used; (Viegas et al., 2016) combine survey data with smart metering data to classify customers according to their electricity consumption; (Beckel et al., 2013) predict socio-economic properties of households, and (Kleiminger et al., 2013) estimate occupancy of households using electricity consumption data. These studies rely on various machine learning methods, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Linear Discriminant Analysis, combined with regression analysis. To the best of our knowledge, we are the first to systematically investigate electricity consumption behavior of solar panel users in a low- or middle-income country context.

Second, we show that indeed solar panel systems are used for business purposes, yet that this varies considerably over time indicating that many households make use of the option in a flexible manner, for instance when in need of additional income. In many low- and middle-income countries, solar panel systems are on the rise as a clean alternative to electricity from the grid (World Bank, 2020). So far, solar panels are studied primarily as an affordable and clean mean for households to access electricity. Its potential as a tool for income diversification—relevant in particular for farmers in times of increasing weather risk—has received less attention. The few studies analyzing usage purposes explicitly are exclusively based on survey data which does not allow studying the intensity of usage nor changes in usage behavior over time (among others, Lemaire (2018); Mondal and Klein (2011); Wassie and Adaramola (2021)).[1] Finally, we provide evidence that households using their system for business purposes are less likely to face repayment difficulties. As the targeted households typically do not have the cash on hand to afford a solar panel system, the systems are often offered as pay-as-you-go systems, where the households only pay for the energy they consume but never own the system, or are purchased on credit with flexible repayment schemes (Barry and Creti, 2020). In both cases, understanding how households use the system and whether payment can be attributed to certain usage behavior can be helpful to further develop the product and payment schemes to be better aligned with the targeted customers' circumstances. Many solar panel systems are already equipped with sensors that measure electricity consumption. Our study shows that leveraging this data can be very insightful both from a researcher's as well as from a practitioner's perspective.

The remainder of the paper is structured as follows. Section 2 presents the data. Section 3 outlines the classification approach with XGBoost, presents the results and discusses the limitations of the classification procedure outlining alternative classification approaches. In section 4, we link a customer's repayment behavior to the predicted probability for business usage. Section 5 concludes and provides suggestions for further research.

2. Context and data

2.1. Setting

Our data stems from a cooperation with a pro-social business that sells solar panel home systems with additional appliances, such as a TV, lights, radio, and a charger for multiple phones, to low-income

[1] The results are mixed. Surveying solar panel users in Ethiopia Wassie and Adaramola (2021), find that less than 10% of the households use their system for income generation, but those that do report a substantial income gain due to the system Harun (2015). comes to similar findings for users in Bangladesh. Indeed the majority of the studies find only limited economic impact, one suggested reason being the lack of know-how and proper business training (see also the review by Feron (2016)). Yet, the systems analyzed are relatively small and most come without additional appliances except for lights. A recent report based on surveys conducted with solar panel system users in East Africa, who bought their system on credit or use pay-as-you-go to pay for the electricity consumed, suggests that nearly one fourth of the customers use their system to support their business (at least at some point in time), with almost half of those having started a new business with the help of the system (GOGLA, 2018).

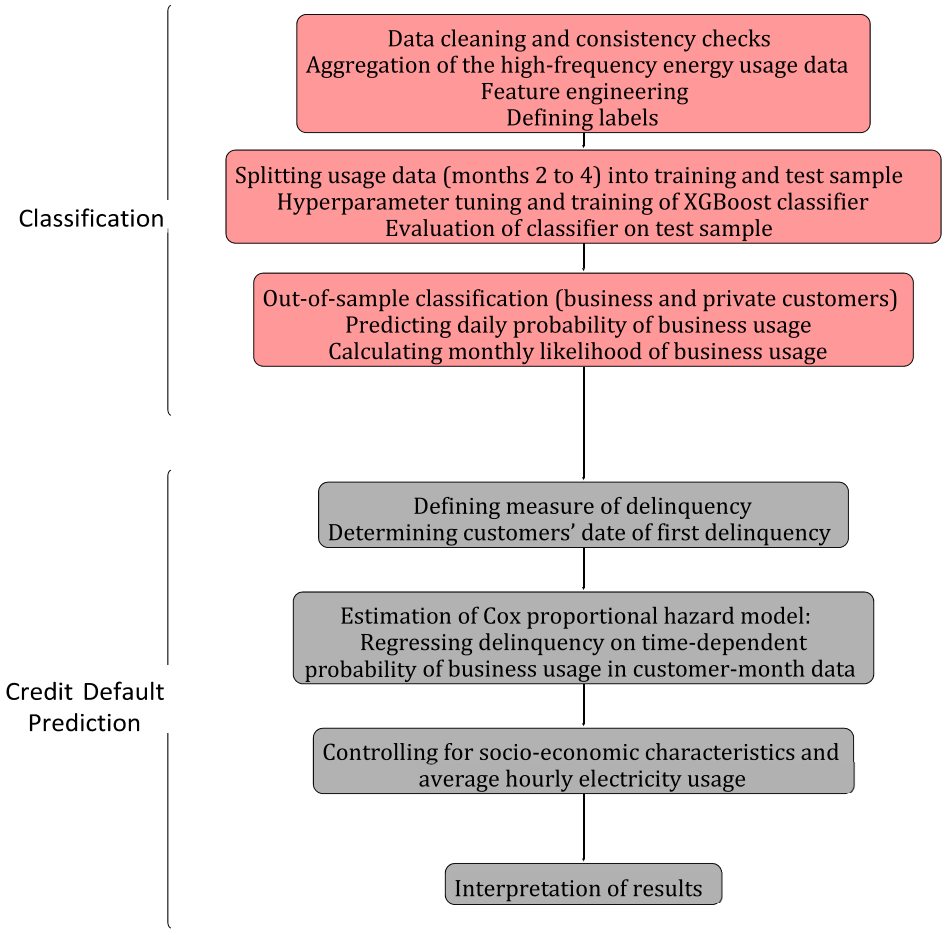

Fig. 1. Process for solar panel user classification and credit default prediction.

households in Tanzania. There are different system types that vary in the system's power (80 W–200 W) and the appliances that come with the system (see Fig. D1 for an example of one of the systems). Customers can purchase additional appliances (e.g., additional lights, a stereo, an electric shaver, or a fan) at any point in time. The systems and the appliances are sold in shops located throughout the country as well as through travelling sales agents in more remote regions. The systems have a four year warranty and there is close customer support. In case of problems, the customers can call a toll-free number; if needed, a technician is sent to resolve the problem.

While primarily designed for private consumption, the system can also be used for business purposes. Small-scale survey data suggests that about one out of four customers may use the system to generate income at some point of time. Households either start new businesses, e.g., by charging phones, opening barber shops or home cinemas (see Fig. D2) or boost their existing businesses. Lights allow for longer opening hours of stores or kiosks, whereas a radio, stereo, or TV equipment can attract additional guests to bars and restaurants.

A system costs between 600 US-$ and 1300 US-$. Nearly all households purchase the solar panel on credit. They have three years to repay the loan. Payment is done via mobile money. Customers are free to decide on the timing and amount of payments. Each payment also charges the solar panel according to the payment amount similar to pay-as-you-go systems.[2] Whenever the panel is not charged sufficiently anymore, it shuts down automatically until the next payment is made. The company allows for a grace period of 30.5 days per year, during which the system can be shut down due to insufficient payments. If this grace period is exceeded, the customer is considered to be delinquent. Households can recover from delinquency by repaying the outstanding payments. A delinquent customer is attended to by a loan field officer who determines (through phone calls and personal visits) whether a household is willing and able to repay the loan. If households cannot provide payments in a timely manner, the system is repossessed by the company.

Each system is equipped with a sensor that tracks electricity generation and consumption in real-time. This data is transmitted every 10 min through an integrated modem. The data allows the company to trace the technical status of the system and check the performance of individual components. The data is also used to send automatic alert messages to the customers, for example in case a battery is nearly fully consumed and needs to be re-charged.

[2] For instance, a 600 US-$ loan to be repaid within 3 years translates into a monthly payment of 16.67 US-$ or a daily payment of 0.55 US-$; in this example, a payment of 20 US-$ would keep the system running for 36 days (independent of the extent of electricity usage).

2.2. Data

We combine (1) high-frequency data on the electricity usage that is directly recorded within the system; (2) survey data collected during the initial loan-eligibility interview, which provides us with information on the system's usage purpose (for business or private consumption) as well as on socio-economic characteristics of the customers; and (3) repayment data that is recorded through the mobile money operators. Our analysis focuses on 73,064 households that purchased the system on credit between June 2014 and January 2018 and their usage and payment behavior from June 2014 to November 2018.

Usage data entails the energy consumption data of each solar panel system, recorded at a 10 min interval. Besides the total energy consumption, to which we refer as the *overall load*, the data distinguishes between energy consumption by small and large devices, to which we refer as *small load* and *big load* respectively. The data is cleaned by removing invalid records, which can occur if a customer tampered with the system or the system is broken. In order to reduce the size and complexity of the data, we aggregate the usage data at an hourly level. We hereby disregard missing values which can be due to interruptions in the signal inhibiting the transmission of the recorded data. We exclude customer-day observations with more than 10% of missing values or invalid records.

Survey data provides us with labels that are used for the training of our supervised models as well as with basic control variables for the proportional hazards model. Our labels are based on the loan eligibility survey that is conducted by the company as part of the due diligence before a customer can be provided with a loan. In the survey, prospective customers are asked about their intended usage of the system, in particular, whether they plan to use the solar panel for consumption, business or both purposes. The survey contains information on the system's purpose for 29,552 households (i.e. roughly 40% of our sample).[3] Of these households, 92.5% report that they plan to use their system for private purposes only, 5% intend to use the system exclusively for business purposes and 2.5% plan to use the system for both business and private purposes. However, non-representative surveys conducted with small subsets of households at a later point in their repayment cycle suggest that the proportion of households using the system for business purposes can increase up to about 20–30% over time. The loan eligibility and survey data also provides basic socio-economic information on the customers, including their gender, household size and the main source of income, which we categorize broadly in self-employment, wage-employment and farming. Furthermore, for each household the exact location is recorded when the system is installed.

Repayment data records the timing and amount of each individual payment. This data allows us to infer whether, when and for how long households are late in their payments (for more detail on the repayment data see Grohmann et al., 2021). We define a household to be delinquent on repaying the credit when the official grace period is exceeded, that is, when the system was shut down due to non-payment for more than 30.5 days within a year. Most of the households (75%) experience delinquency at some point in time. The vast majority of them (90%), however, recovers by paying the outstanding amount.

2.3. Customer characteristics

Table A1 shows the main customer characteristics in the total customer sample. Most of the customers are male (82%) and live in rural areas (78%). The majority are either farmers (45%) or operate their own business (32%); only few are wage-employed.[4] On average, households consume 8 W of electricity per hour. As a comparison: if the multiple phone charger, which can simultaneously charge up to ten phones, is fully used, the charger can consume up to 40 W; a TV consumes on average, depending on screen size and brightness, between 11 and 24 W, while a light consumes just around 1–3 W.

Fig. 2 shows the energy usage profiles for the average load over a day for a randomly selected day of four randomly selected customers. Customers 1, 2 and 4 seem to use the generated electricity for lighting (also over night), as their energy usage goes down after 6 a.m. and then goes up again at or after 6 p.m.[5] Customer 1 experiences a further usage spike at lunchtime and after that a second one in the evening. Customer 4 uses more electricity than the others, but with most usage concentrating in the evening hours. Increases in the afternoon and evening hours could potentially reflect that customers come home from work and watch TV or listen to the radio. In contrast to customers 1, 2 and 4, the usage profile of customer 3 has a more distinct usage pattern during the day: energy consumption sharply increases in the morning, stays up until lunch time and then reduces in the afternoon. Potentially, this household also uses the system for business purposes, e.g., by operating a shop that closes in the late afternoon. However, the system's purpose cannot be inferred unambiguously solely by observing the usage profiles.

In addition, load profiles differ from day to day. Fig. 3 depicts the daily average load for a randomly sampled week for two randomly drawn customers. For customer A, there are clear peaks in electricity consumption in the morning, around noon and in the evening. This stays more or less consistent throughout the week. Customer B, by contrast, uses the system consistently only in the evening, yet during daytime electricity consumption varies strongly from day to day. Indeed, also customers that use their system to generate additional income (e.g., through phone charging or a village cinema) presumably do not run this business necessarily every day. Business usage should thus be classified on customer-day and not solely on customer level. Aggregating the daily likelihood of business usage into monthly patterns will subsequently reflect the overall intensity of business-like energy usage within any month.

3. Supervised classification of business users

The goal of our classification exercise is to detect daily electricity usage patterns that describe usage for business purpose as compared to private consumption. For this purpose, we predict the daily probability of business usage. We do not aim for a binary classification as households might only use the solar system to a certain extent for business purposes; furthermore a binary classification would require the choice of an arbitrary cut-off. For the prediction, we rely on a supervised classification approach, utilizing labels that are based on information on whether the system was originally planned to be used for business or for non-business purposes.

In order to reduce the dimensionality of the data and to increase the interpretability of our predictions, we first derive a set of relevant features from the electricity usage data which then form the basis for the classification procedure.

[3] The question on system purpose was included in the loan eligibility interview only from mid of 2016 onward and the information is therefore not available for customers who have purchased the system before. The sample of households for which this information exists is, however, roughly comparable with households for which this information does not exist in terms of socio-economic characteristics (see Tables A2 and A3).

[4] For a more detailed description of the customer profiles and how they compare to the Tanzanian population see Grohmann et al. (2021).

[5] In Tanzania sunrise generally happens between 6:15 a.m. and 6:45 a.m. through the year and sunset usually happens between 6:30 p.m. and 7:00 p.m. Most regions experience 8–9 h of daily sunshine on average, which reduces by 1–2 h during the rainy season. Typically, the system can operate on optimal capacity throughout the year except for very cloudy days when energy production reduces to 10–25% of optimal capacity. However, this does not seem to affect usage. In our data, we observe no systematic difference in usage across seasons.

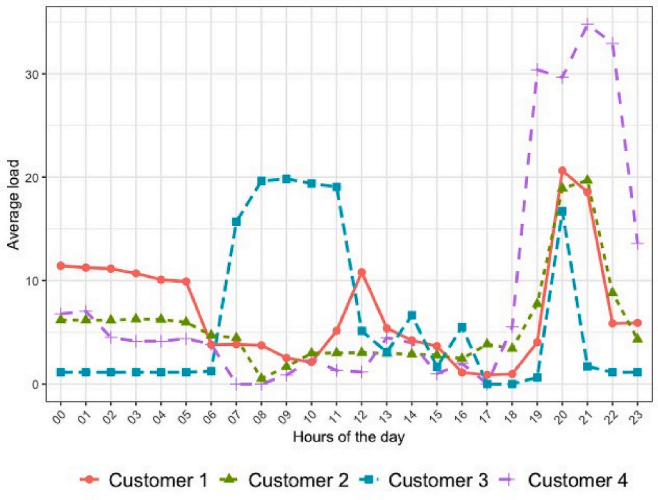

Fig. 2. The graph displays four randomly selected electricity usage profiles for the average load over a day. The unit of measurement is Watt.

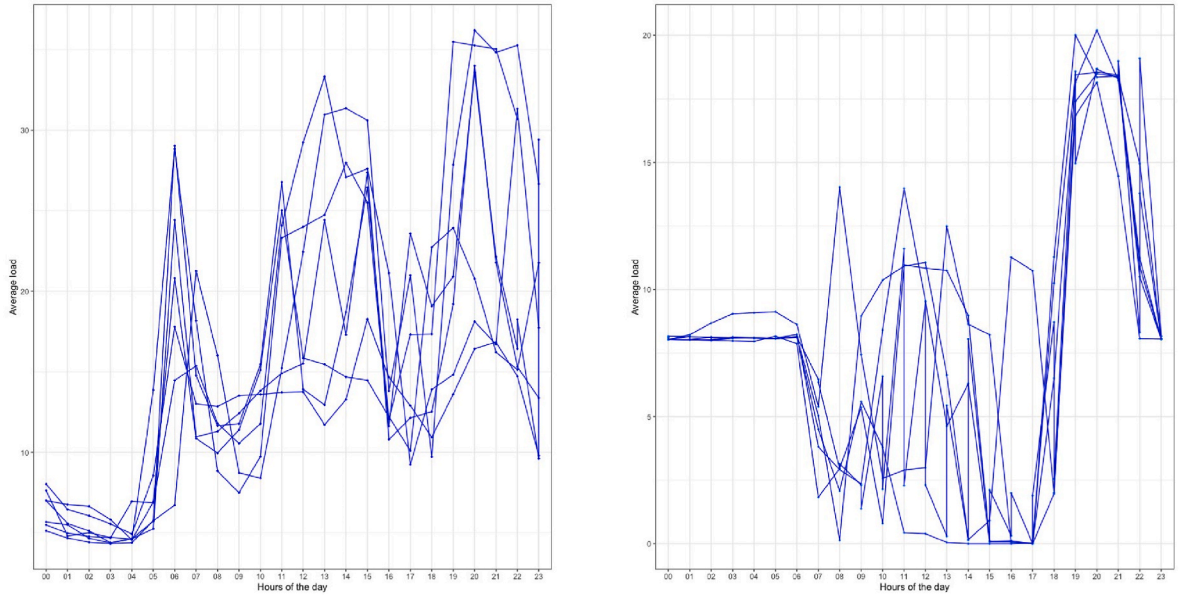

Fig. 3. The figure shows the daily average load for a randomly sampled week for two randomly drawn customers. The unit of measurement is Watt.

3.1. Feature generation

After aggregating the raw electricity usage data into average hourly usage in terms of total, small and big load, we generate a total of 84 features that describe the temporal dynamics of electricity usage of each customer-day observation. These features can be grouped into four main categories:

1. *Daily usage metrics*: daily mean and daily standard deviation of total, small and big load [6 features];
2. *Count metrics of daily usage*:Number of hours with low usage (below the 25th percentile), number of hours with intensive usage (above the 75th percentile), number of hours with zero usage for total, small and big load [9 features];
3. Within-day usage metrics: Average usage during 7 time intervals of the day (early morning 5–8 am, late morning 8–11 a.m., noon 11 a.m.–2 pm, afternoon 2–5 pm, early evening 5–8 pm, late evening 8–11 p.m., night 11 p.m.–5 am), for total, small and big load [21 features];
4. Metrics of usage changes over time:
   a) First order difference in usage from each hour to the previous hour (excluding the hours from 0 a.m. to 4 a.m.), for total, small and big load.[6] [38 features];
   b) Difference between big load and small load calculated at the 7 time intervals outlined above [7 features];
   c) Difference between the cumulative usage at prime time (8 a.m.–11 p.m.) and non-prime time (11 p.m.–8 am), for average, small and big load [3 features].

These features reflect not only average electricity use but also the

[6] For example, the first difference from 6 a.m. to 7 a.m. is calculated as the usage from 6 a.m. to 7 a.m. minus the usage from 5 a.m. to 6 a.m.

overall variability of usage as well as how strongly usage is increasing or decreasing at certain time periods.

3.2. Classification with Extreme Gradient Boosting (XGBoost)

XGBoost is one of the most powerful machine learning classifiers for structured data. It constructs a random forest for prediction based on the regularized objective function

$$l_{\text{pen}}(f(\boldsymbol{x})) = \sum_{i=1}^{n} l(y_i, f(\boldsymbol{x}_i)) + \text{pen}\big(f(\boldsymbol{x})\big),$$

where $(y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n$ are observations on a response variable y and features $\boldsymbol{x}$, $l(y_i, f(\boldsymbol{x}_i))$ is a convex loss function quantifying the deviation between the response y_i and the prediction $f(\boldsymbol{x}_i)$ (in our case the log-likelihood of a binary logistic regression model). The regularization penalty for the random forest pen $(f(\boldsymbol{x}))$ is given by

$$\text{pen}(f(\boldsymbol{x})) = \gamma T + \frac{1}{2}\,\lambda\,\|\mathbf{w}\|^2 .$$

where T is the size of the tree (number of terminal leaves), $\boldsymbol{w}$ is the vector of leaf weights and $\gamma > 0$ and $\lambda > 0$ are regularization parameters. Minimization is achieved greedily in a gradient-based boosting approach where the estimate $\widehat{f}(\boldsymbol{x})$ is iteratively updated as

$$\widehat{f}^{(v)}(\boldsymbol{x}) = \widehat{f}^{(v-1)}(\boldsymbol{x}) + \widehat{g}^{(v)}(\boldsymbol{x}),$$

where v denotes the iteration index and $\widehat{g}^{(v)}(\boldsymbol{x})$ is the random forest update determined in the v-th iteration of the boosting procedure. To quickly optimize the objective function, a second order Taylor expansion of the loss function is employed (Chen and Guestrin, 2016). The implementation of XGBoost as a supervised machine learning classifier and the required hyperparameter tuning is more complex and requires more computing resources than the training of a more simplistic classifier, such as a logistic regression, but optimizes the classification performance. Using a supervised machine learning approach such as XGBoost is usually easier in terms of the implementation and interpretation of the results than the application of unsupervised machine learning models. If labelled data exists in comparable settings, we therefore suggest the training and optimization of an XGBoost model (or comparable classifiers) as demonstrated in this paper.[7]

We train the classifier with the usage data of those households that were asked about their prospective use of the solar panel home system. If customers indicate that they intend to use the generated electricity for business or mixed (partially business) purposes, we classify them as prospective business users whereas all others are considered as non-business users. For the training data, we rely on the electricity consumption behavior in month 2 to month 4 after the solar panel was installed. Restricting the training data to this time period is an ad-hoc choice owing to the specific setting. With this restriction, we assume that those customers who indicated to use the system for business purposes had sufficient time to establish such a business (i.e. one month), while those who indicated to use the system for private purposes have unlikely already changed their minds and switched to business use. Our approach allows us to generate a large training sample for our classifier, even though the labelled data set refers to a limited time-range of individual observations. This approach can thus be also implemented in panel data settings where only partial samples of labelled data are available, but there is a long time series of individual data.

In order to train the classifier, we sample 1,588,750 customer-day observations in total. We retain 80% of the customer-day observations for training the classifier and 20% customer-day observations for testing purposes. Note that we take a random sample of all customer-day observations within our target period instead of sampling business and private customers first and including subsequently all days within our target period in the sample as we find that the sampling of customer-day observations leads to better classification results.

Fig. 4 displays the average daily electricity usage profile belonging to business and non-business users in our training sample. It shows that the electricity usage of households that report to operate a business is somewhat higher on average but also follows distinct time patterns over the day. Business users consume relatively more electricity during daytime but are barely distinguishable from purely private users during the peak evening hours. When further distinguishing between small and big load (see Fig. 5), we see that the difference is driven primarily by heavy load appliances. Whereas these average differences are already suggestive, the supervised classification exercise relies on the substantially more extensive set of features to capture the various dimensions of usage dynamics throughout a day.

3.3. Classification results

To evaluate the performance of our classification approach, we rely in particular on the Receiver Operating Characteristic (ROC) curve and corresponding Area under the Receiver Operating Characteristic (AUROC) as standard performance metrics in the machine learning literature for supervised classification.[8] The particular advantage of the ROC curve is that it evaluates the performance for a range of thresholds, which circumvents the need to define one single threshold in a binary classification and allows us to use the predicted probabilities directly instead (Gron, 2017; James et al., 2014). We obtain an AUROC of 0.784. A random classifier would achieve an AUROC of 0.5. This provides an indication that the classifier performs reasonably well given the data structure and the challenging classification task of classifying business usage on a daily basis. As a further means to evaluate the classification performance, Fig. 6 displays the predicted probabilities of business usage within the test sample comparing customers that indicated to plan to use the system for business purposes with customers that indicated to use the system for private purposes only. It shows that our classifier indeed distinguishes between business and private users considerably well. Note that the figure also shows that the predicted probability of business-like usage is widely spread among those customers that reported that they intend to use their solar panel for business purposes. By contrast, business probabilities are more skewed towards zero among customers that reported no intentions for business usage. As a result of the very specific classification task and data structure, a comparison with benchmark results in the literature is not reasonably possible. However, the considered performance metrics and visualisation of the predicted probabilities for the test sample show that our classifier performs sufficiently well in predicting the daily probability of business usage.

Figs. A1 and A2 in the appendix depict the distribution of the out-of-sample predicted probabilities on a daily and monthly level respectively. The vast majority of daily observations cluster at relatively low predicted business usage probabilities, reflecting that most of the households use the produced electricity of their solar panel primarily for private purposes. The average daily predicted probability of business usage lies around 8.3% (see Table A5). As shown in Fig. 6, a cut-off of 10% of predicted business usage probability already discriminates reasonably well between private and business usage; when using this cut-off to determine a day as "businessday," on average 23% of a household's usage days can be defined as business days, i.e. as days on

[7] For the implementation, we use the mlr package (Bischl et al., 2016) and XGBoost implementation in R (Chen et al., 2020). The hyperparameter tuning is presented in Appendix B.

[8] A short introduction to the performance measures is provided in appendix E.

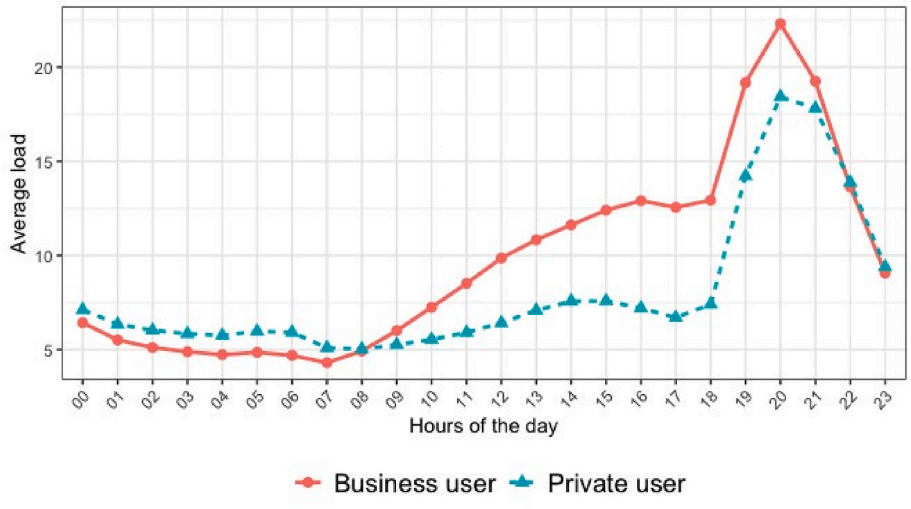

Fig. 4. The graph displays the average daily electricity usage profile belonging to business and non-business customers in our training sample for average load. The unit of measurement is Watt.

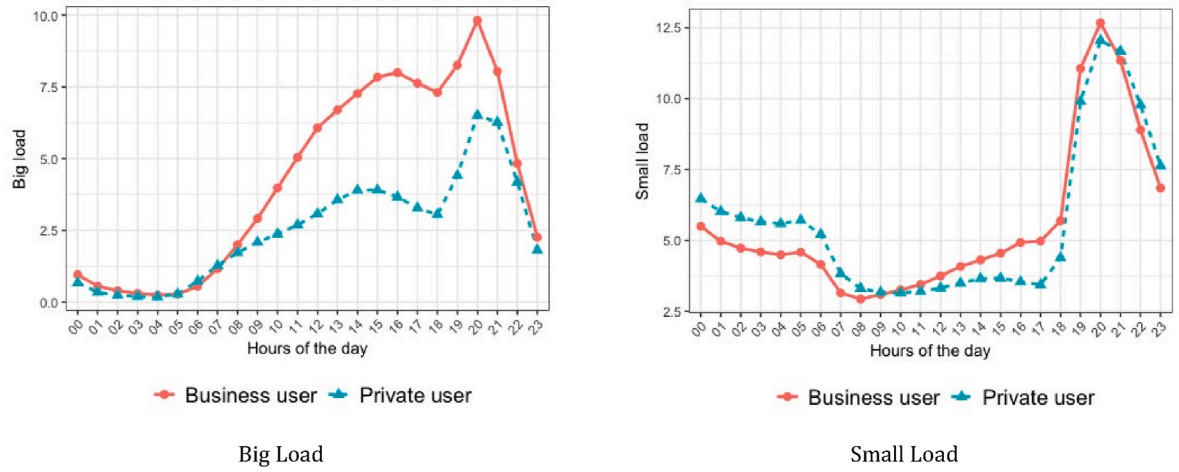

Fig. 5. The graph displays the average daily electricity usage profile belonging to business and non-business customers in our training sample separately for big and small load. The unit of measurement is Watt.

which customers use the system presumably for business purposes (see also Fig. A3). Increasing the threshold to 25% business probability naturally reduces this proportion, still on average 6% of a household's usage days would be defined as business days. The distribution, however, is highly skewed; only very few households show such extreme business-like behavior on most of their days (see Fig. A4).

On average, the predicted business probability does not change much over time (see Fig. A5 in the appendix, which depicts average predicted business probability in the first 12 months after system purchase over the whole sample). However, this average masks considerable variation across customers. Fig. A6 shows the monthly predicted business probability for a random sample of five customers. While for customers 1, 2 and 3, the probability remains more or less stable, for customers 4 and 5 there are notable changes over time in the extent to which the system is predicted to be used for business purposes. These customers seem to make use of this option according to circumstances, e.g., when in need of additional income.

From the originally specified 84 features, Fig. C1 in appendix C displays the 20 most important features that predict business-like electricity usage by XGBoost according to the Gain metric. The Gain metric is a conventionally used measure for XGBoost to evaluate the relative importance of features. It measures the improvement in classification accuracy that results from splits on a feature (Chen and Guestrin, 2016; Chen et al., 2020). Hence, the features in Fig. C1 lead to the largest improvements in classification accuracy in relative terms and are therefore most important to distinguish between business and non-business users for XGBoost. However, the Gain metric does not provide information on the direction of the effects to evaluate whether a

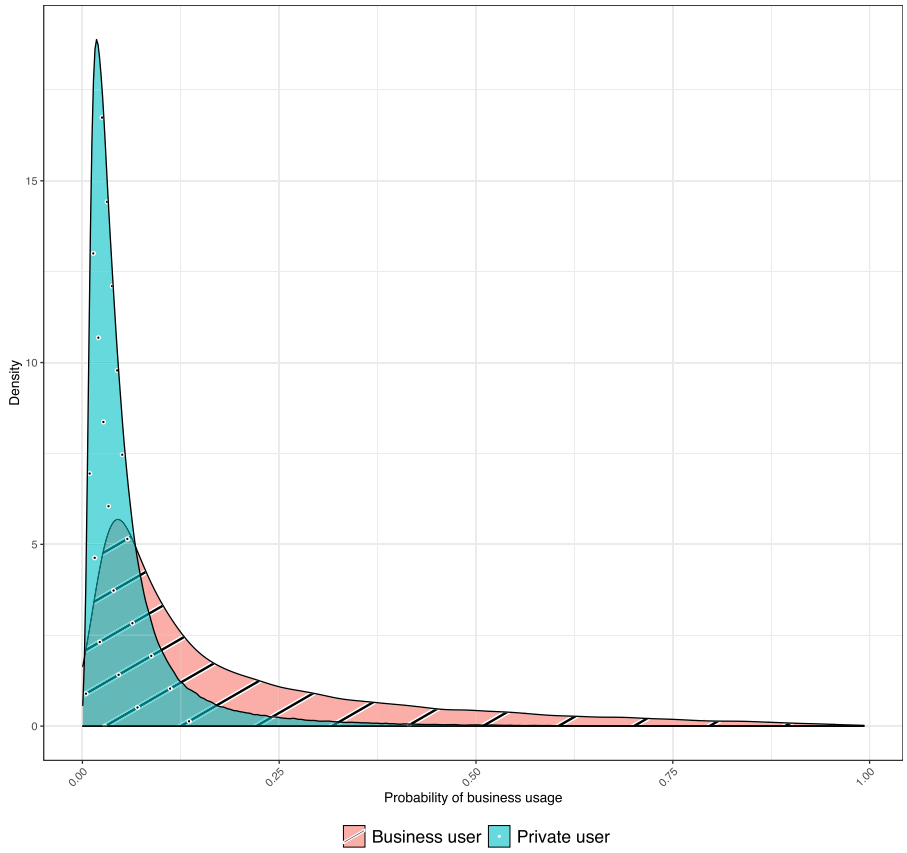

Fig. 6. The graph displays densities of the predicted probabilities of business usage for the business and private users in the test sample.

feature is positively or negatively associated with business usage. According to our knowledge there is no established metric in the literature to evaluate the direction of effects for XGBoost. As an ad-hoc remedy, we compute the correlation between the most important features and the binary business usage labels. Table C1 reports the Point-Biserial Correlation between features with the largest predictive power and the binary business usage label in the XGBoost training sample. Fig. C1 shows that according to the Gain metric, especially the volatility of electricity usage is important for the classifier to discriminate between private and business usage as well as electricity usage in the early evening. According to Table C1, high volatility and a rather high electricity consumption in the early evening hours are both significantly positively correlated with business use. This is reasonable given that the most prominent business related use of the system is charging phones followed by operating a home cinema. Charging the phones of others results in a volatile usage pattern, while a home cinema is typically frequented in the early evening hours. Other important features include the first difference in usage both for the morning and the evening hours, i.e. the steeper an increase or decline in usage in the morning and evening, the more likely is usage predicted to be linked to business. This might reflect the fact that businesses such as small shops switch all their appliances on (off) when opening (closing) their stores. Yet, the fact that the highest partial correlation coefficient does not exceed 0.2 clearly indicates that a combination of a number of different features is needed for predictive purposes rather than just one or two specific features.

3.4. Discussion

Using a labelled dataset for classification purposes provides us with the unique opportunity to identify time-variant patterns of electricity use for business purposes. Our classification, however, comes with two important limitations. First, the information on business usage is collected at the time of the purchase of a solar panel. It thereby only reflects planned use and could furthermore suffer from strategic misreporting by prospective users.[9] Second, we train the XGBoost classification algorithm based on early usage data, i.e., during the first months after the system was installed. If the electricity usage patterns of business users change substantially over time, restricting the training period to early usage can limit our ability to predict business usage for a later point in time.

Alternatively, one could derive the labels based on surveys conducted with a sample of existing customers who are asked about their usage behavior. The drawback of survey data, however, is that the number of observations is typically substantially smaller and the customers reached are rarely representative for all customers. Furthermore, the information is only reported for a specific point in time, i.e., the period when the survey is conducted, and might thereby not be representative for usage behavior over a longer time horizon; recalling past usage behavior instead can be prone to reporting errors (Rom et al., 2020).

If labelled data is not available, supervised classification approaches cannot be applied. As a remedy, unsupervised clustering methods, such as Gaussian mixture models (GMM), could be applied to cluster daily load profiles in order to discover distinct behavior groups. The average

[9] While the reported purpose of the solar panel is not used as an eligibility criterion by the company, a prospective user might nevertheless (wrongly) believe that prospective use affects loan eligibility. However, the direction of the bias is a priori unclear as customers might just as likely assume that business or that private users would be preferentially treated.

load profiles during a day can then be visualised for the different clusters and—based on contextual evidence on typical usage patterns—the clusters can be labelled as describing predominantly business or private use. Finally, to derive a probability for business usage for each customer-day observation, the probabilities for each cluster k for the customer-day observations i can be accumulated. For such an unsupervised learning approach, contextual information is crucial. Yet, the ex-post labelling of business clusters is likely arbitrary so that supervised approaches should be preferred if sufficient labelled data is available.

4. Business use and repayment

Using the solar panel system to generate income can relieve cash constraints and help borrowers to repay their loan. In order to examine the implications of business usage for repayment, we regress the time until first credit delinquency on the predicted probability that a household had used the system for business purposes.[10]

This statistical analysis illustrates whether the predicted probability of business usage contains relevant information on the households' economic decisions and circumstances. For the estimations, we only rely on out-of-sample predictions of the probability of business usage and exclude data from the customer-months that we used for training and testing the classifier.

More specifically, we implement a Cox proportional hazard model (Cox, 1972; Therneau, 2020) with the time-dependent business probability as explanatory variable in the following form:

$$h(t, b_i(t), \boldsymbol{x}_i, u_i(t)) = h_0(t)\exp\left[\delta_1 b_i(t) + \boldsymbol{x}_i'\boldsymbol{\beta} + \delta_2 u_i(t)\right], \quad (1)$$

where $h(t, b_i(t), \boldsymbol{x}_i, u_i(t))$ denotes the hazard, i.e., the risk of first delinquency, of household i in month t and $h_0(t)$ is the time-dependent baseline hazard function, which describes how the risk of first delinquency varies in response to the monthly predicted average business probability, $b_i(t)$.[11] More specifically, $b_i(t)$ describes household i's predicted business probability averaged over all days during which the system was used in month t.[12] As we are interested in the first delinquency, for this analysis we treat all households that become delinquent once as permanently delinquent, irrespective of whether they recover through new payments or not. Our main coefficient of interest is δ_1, where $\exp(\delta_1)$ reflects the multiplicative difference in rates of delinquency (hazard ratio) between business and non-business users.

We control for a vector of time-invariant explanatory socio-economic variables, $\boldsymbol{x}_i$ namely the gender of the buyer, household size, a set of indicators for the main source of income (wage employment, self employment or farming) and an indicator for households living in urban areas. Additionally, we control for the system type of the solar panel, distinguishing between system sizes of 80 W, 120 W and 200 W. Finally, we include as a further time-variant control the average electricity usage within a month, $u_i(t)$, in order to ensure that our classification results on business use provide additional information beyond being simply correlated with a higher intensity of electricity usage.

Table 1 reports the outcomes of the regression analysis. Coefficients are reported as hazard ratios. We run three different specifications: we first include only the predicted probability of business use (column 1) and then successively add controls for socioeconomic characteristics (column 2) as well as for average electricity usage (column 3). The results show a robust negative association between the risk of delinquency and the predicted probability of being a business user $b_i(t)$ within any given month. Households that are more likely to have used electricity for business purposes during a given month experience a lower risk of delinquency.

Table 1
Cox Model with time-dependent business probability.

Dependent:	Month of first delinquency		
	(1)	(2)	(3)
Prob. of business use per month	0.571***	0.532***	0.585***
Male		1.094***	1.091***
Household size		0.983***	0.984***
Self employed		1.003	1.005
Wage employed		0.823***	0.827***
Farmer		1.012	1.016
Urban		1.235***	1.242***
System with 120 W		1.008	0.966*
System with 200 W		1.042	0.951
Average hourly usage per month			1.014***

Notes: $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Number of observations is 441,837. Number of delinquency events is 28,241. Coefficients are reported in form of hazard ratios (HR) by using the exponential function.

Results are robust to controlling for basic socio-economic characteristics of the household, and more importantly, also to controlling for average electricity use directly (column 3). This implies that our measure of predicted business probability is able to detect additional patterns of usage that go beyond simply the average intensity of use. The estimated effect size is substantial: switching the average probability of using electricity for small-scale business from 0 to 1 decreases the risk of delinquency by 42–47 percent, depending on the specification.

These findings could be interpreted as evidence that using the solar panel for business purposes can help households to repay their loan; it enables them to generate additional income and diversify their income sources. However, there are also other channels that could explain our results. For example, households that use their system for business purposes are arguably more dependent on their system; they therefore might have stronger incentives to keep their system running and repay on a more regular basis than customers that use the system for private purposes only. In addition, our analysis is purely correlational and one should be cautious not to interpret these results as causal. Wealthier customers, who should face less difficulties in repaying the loan, might be more likely to use the system for business purposes, as they have the resources to make the required investments. Furthermore, running an own business requires a certain level of financial literacy, which in turn can also affect repayment. Our data cannot speak to the underlying channel. This is an avenue for future research. Nevertheless, our results show that the derived predicted business usage probability is a meaningful indicator that can help predict repayment difficulties.

5. Conclusion

Solar panel systems can provide a clean and cost-effective alternative to extend electricity coverage, in particular in countries where access to electricity is limited. We show that such systems are also used as means to generate income and can thereby help to relieve cash constraints. Combining customer interviews at the time of the purchase of solar panels in Tanzania with high-frequency electricity usage data, we rely on supervised classification to predict the time-variant likelihood of customers belonging to the group of small-scale business users. While

[10] Note that we analyze delinquency, i.e. whether a household exceeded the contractually allowed grace period, instead of eventual default. Once the grace period is exceeded, a customer is attended to by a loan field officer who determines whether the customer is able to repay the loan or whether the system needs to be repossessed. As a large part of this decision is at discretion of the loan field officer, there are a number of unobservable factors affecting the timing of eventual default, making it a less suitable proxy for the purpose of our analysis.

[11] See Table A4 for the summary statistics of all variables included in the model.

[12] Days where the system was shut off due to insufficient payments are treated as missing to preclude any mechanical correlation between business usage and non-repayment.

the average predicted business probability is low, it shows considerable variation over time. Our results suggest that a substantial proportion of customers use their system for income generation occasionally and that the hazard of not being able to repay the system is significantly lower when customers use the system for business purposes. We find a robust negative association between the likelihood of credit delinquency and the predicted probability of being a business user within any given month even after controlling for individual socio-economic characteristics and the average intensity of electricity use.

Being able to use the solar panel system for income generation is a highly valuable feature for the users. They can thereby not only boost their existing business but also expand into new ones. In times of increasing climate variability, having additional means to generate income can be particularly helpful for farmers to reduce their reliance on farming related activities (Gao and Mills, 2018; Mathenge and Tschirley, 2015). In addition, our findings suggest that using the system to generate income can help households to repay the substantial investment that a solar panel home system presents for most. Firms should thus be encouraged to offer solar panel home systems that allow for business usage, e.g., by providing the relevant appliances. Furthermore, business and financial literacy training could be offered for the prospective business owners through complementary programs. Such business training is of particular importance to prevent long-term negative consequences for the households. For example, households should be familiarized with the additional investments that a business requires over time as well as the risks involved with running certain businesses. Households should also be cautioned to not completely focus on the new business and neglect their other income sources. The viability of certain businesses crucially depends on the electrification rate in the area; as more households have their own solar panels or gain access to the grid, the demand for certain electricity-related services, such as phone charging, declines. Other, more specialized services, on the other hand, can remain viable (such as sewing, barber-shops, or bars).

To the best of our knowledge, this is the first study that systematically investigates the consumption pattern of electricity generated by solar panels in a low and middle income context. There are a number of avenues for future research. Our data cannot speak to the underlying channel that explains the strong relationship that we find between business usage and repayment. More information is needed, e.g. from survey data or by exploiting exogenous events. For example, linking electricity usage and repayment data with information on extreme weather events could inform us whether the use of the solar panel systems for income diversification can help farmers to overcome negative income shocks resulting from harvest loss. Moreover, combining usage and repayment data with administrative or survey data covering local electricity access over time might allow investigating how viable the use of solar panels for business purposes is in the long run. Finally, and on a very different topic, the high-frequency electricity usage data gathered from solar panels can also be used to capture the presence of household members during daytime as well as time usage patterns within each household, complementing other sources of information on the local labor market and consumption dynamics.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to an unnamed company for providing access to their proprietary data and several employees of the company for their support throughout the project as well as to Henry Stemmler for useful comments and discussions. Financial support by the German Research Foundation (RTG 1723) as well as by the Open Access Publication Funds of the Göttingen University is gratefully acknowledged.

Appendix A. Descriptive Statistics

Table A1
Descriptive Statistics for all customers

Variable	Mean	Std. Dev.	Min	Max
Male	0.821	0.383	0	1
Household size	4.397	2.022	1	30
Self employed	0.322	0.467	0	1
Wage employed	0.229	0.420	0	1
Farmer	0.450	0.497	0	1
Urban	0.225	0.394	0	1
Average hourly usage	7.759	4.077	0.006	61.402
Delinquent (at least once)	0.750	0.433	0	1

Notes: Summary statistics for all customers included in the analysis. On customer level (N = 73,064); not all characteristics available for all customers.

Table A2
Descriptive Statistics for customers without business and private labels

Variable	Mean	Std. Dev.	Min	Max
Male	0.831	0.375	0	1
Household size	4.742	2.105	1	30
Self employed	0.322	0.467	0	1
Wage employed	0.228	0.419	0	1
Farmer	0.450	0.497	0	1
Urban	0.244	0.429	0	1
Average hourly usage	7.320	4.140	0.006	34.397
Delinquent (at least once)	0.778	0.415	0	1

Notes: Summary statistics for customers without business and private labels. On customer level (N = 44,671); not all characteristics available for all customers.

Table A3
Descriptive Statistics for customers with business and private labels

Variable	Mean	Std. Dev.	Min	Max
Male	0.809	0.393	0	1
Household size	3.945	1.810	1	30
Self employed	0.321	0.467	0	1
Wage employed	0.230	0.421	0	1
Farmer	0.450	0.497	0	1
Urban	0.198	0.399	0	1
Average hourly usage	8.503	3.856	0.120	61.402
Delinquent (at least once)	0.722	0.448	0	1

Notes: Summary statistics for customers with business and private labels. On customer level (N = 28,393); not all characteristics available for all customers.

Table A4
Descriptive Statistics for Cox proportional hazards model

Variable	Mean	Std. Dev.	Min	Max
Male	0.805	0.396	0	1
Household size	4.458	1.931	1	30
Self employed	0.288	0.453	0	1
Wage employed	0.235	0.424	0	1
Urban	0.198	0.316	0	1
System with 80 W	0.658	0.474	0	1
System with 120 W	0.342	0.474	0	1
System with 200 W	0.001	0.001	0	1
Average hourly usage	6.970	3.688	0	129.705
Delinquent (at least once)	0.064	0.245	0	1
Predicted prob. of business use	0.074	0.068	0.003	0.898

Notes: Summary statistics for the variables included in the Cox proportional hazards model. On customer-month level (N = 441,837).

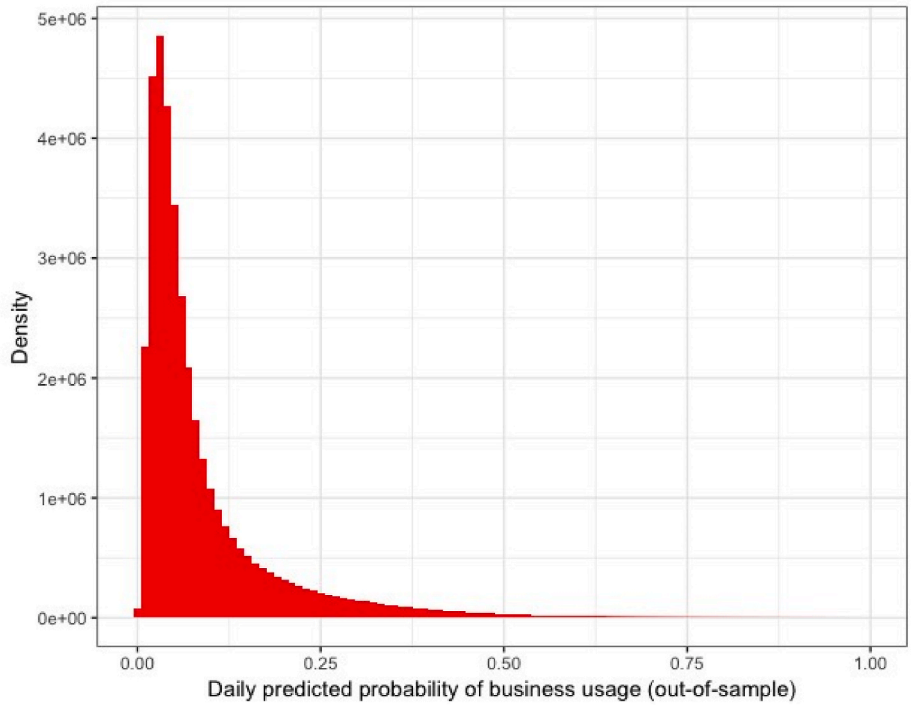

Fig. A1. The graph displays a histogram of the daily out-of-sample predicted probabilities of business usage.

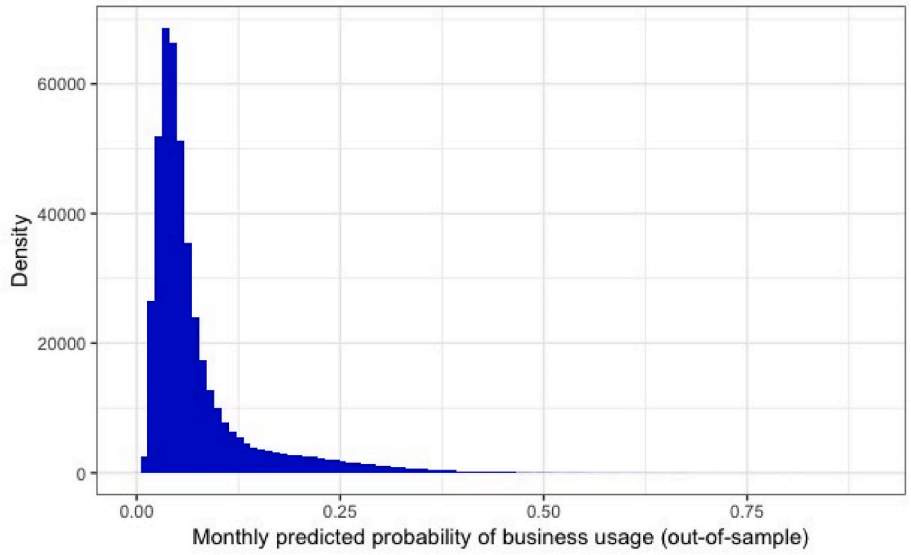

Fig. A2. The graph displays a histogram of the monthly out-of-sample predicted probabilities of business usage.

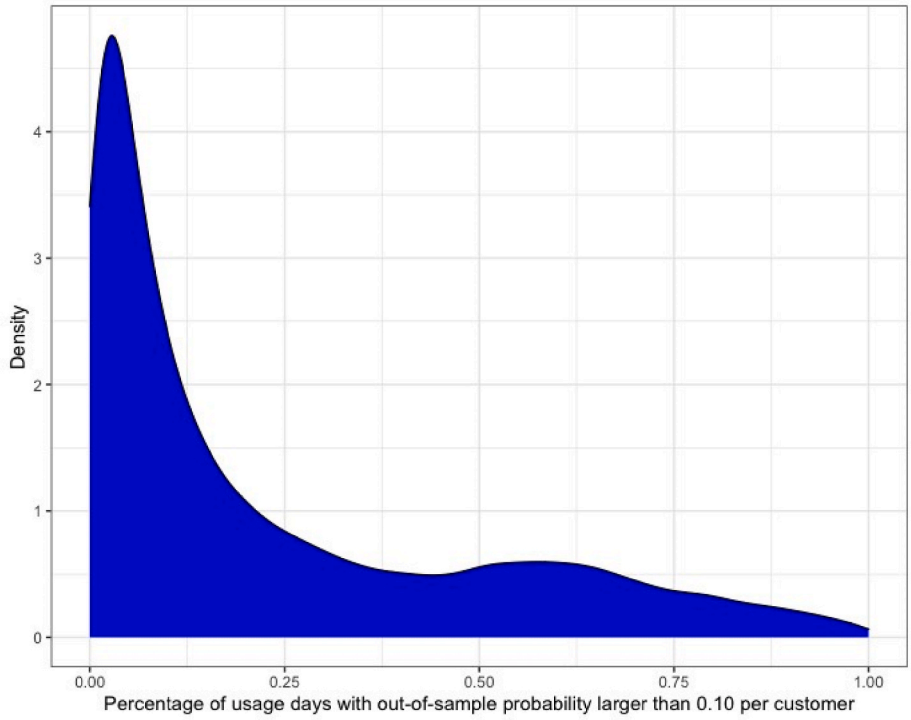

Fig. A3. The graph displays a histogram of the percentage of usage days with P(Business) $\geq$ 0.1 on the customer level.

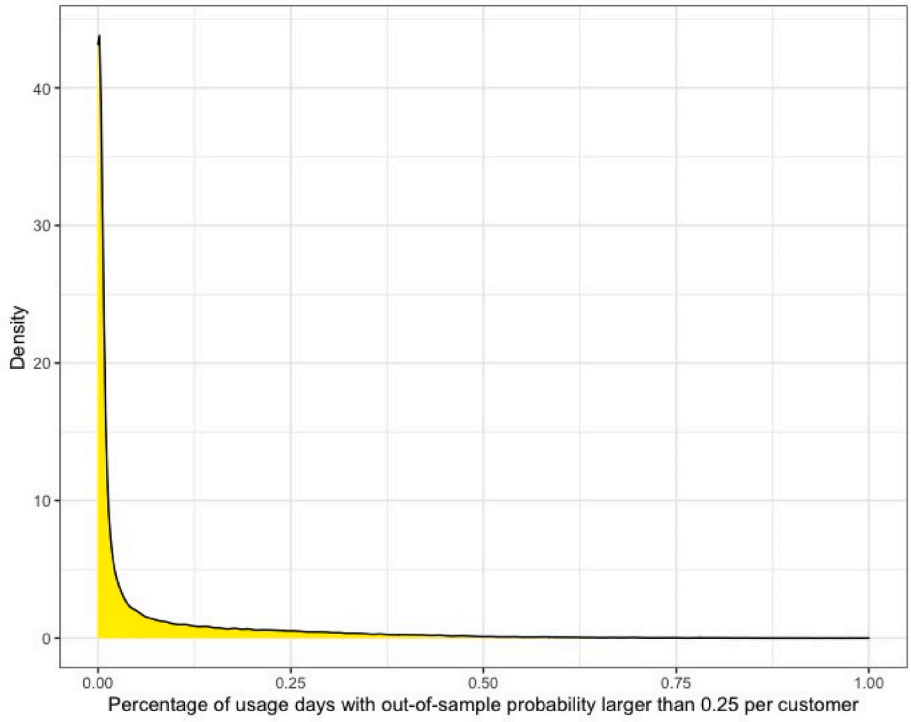

Fig. A4. The graph displays a histogram of the percentage of usage days with P(Business) $\geq$ 0.25 on the customer level.

Table A5
Descriptive Statistics: out-of-sample predicted probabilities of business usage

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
P(Business)	0.007	0.042	0.058	0.083	0.106	0.778
Percentage of usage days with P(Business) $\geq$ 0.10	0.000	0.037	0.118	0.236	0.389	1.000
Percentage of usage days with P(Business) $\geq$ 0.25	0.000	0.000	0.005	0.064	0.060	1.000

Notes: Summary statistics for out-of-sample predicted probabilities on customer level.

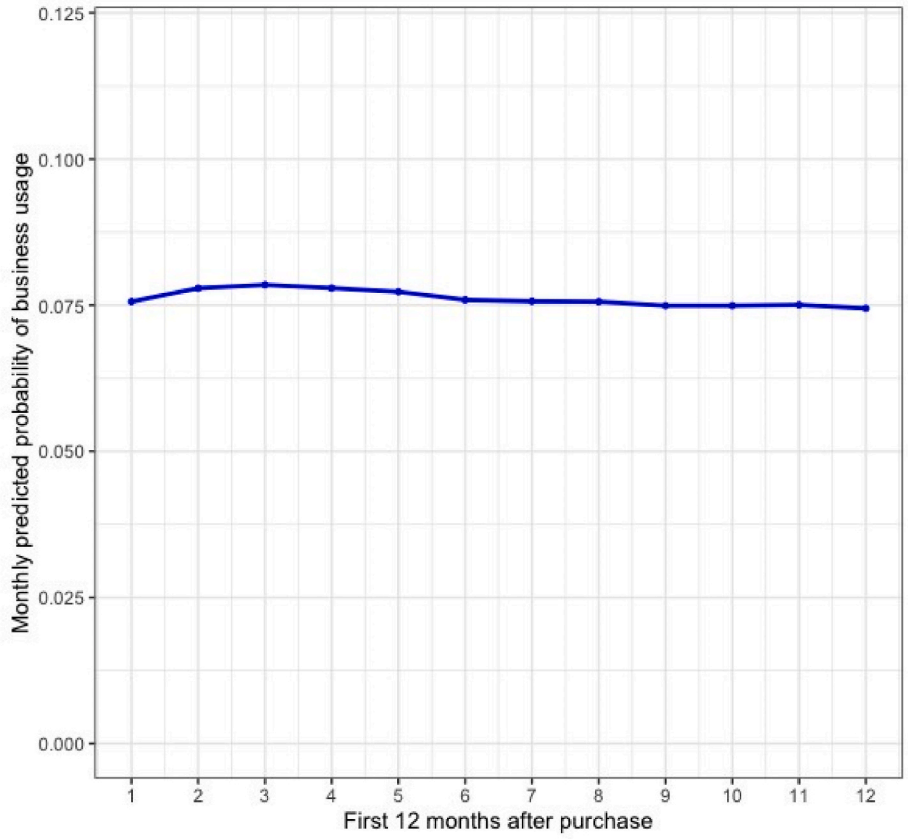

Fig. A5. The graph displays the average monthly predicted probability of business usage for customers that use the system for at least 12 months without a delinquency.

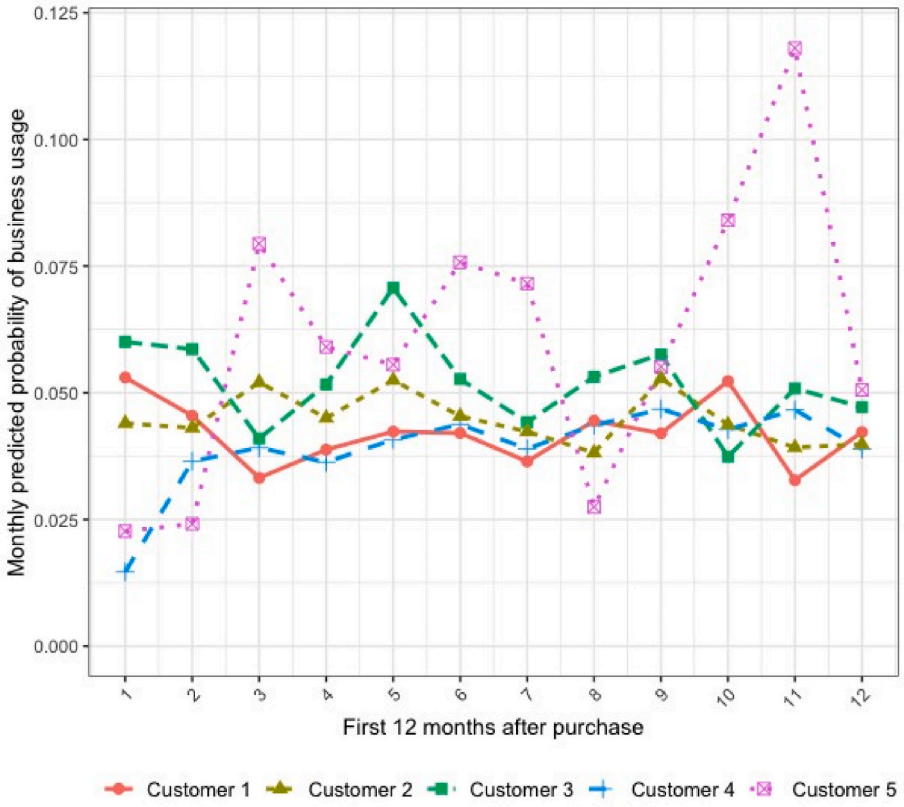

Fig. A6. The graph displays the monthly predicted probability of business usage for a random sample of customers that use the system for at least 12 months without a delinquency.

Appendix B. Hyperparameter tuning of XGBoost

We select the following ranges of hyperparameters for XGBoost. The learning rate $\eta \in (0,1)$ is set to 0.1. For the maximum depth of a tree we set a range of 3–12. For the minimum number of observations in the terminal node we set the range of 1–10. We use stochastic boosting, for which a sample of the data is selected in the construction of a tree, and set the range for the *subsample* as 0.5 to 1. For the sampling of variables in the growing of each new tree, we choose the range from 0.5 to 1. We apply k-fold cross-validation with k as 5. For the maximal number of boosting iterations, we choose a range of 100–500 number of iterations. Several trials show that a larger range only leads to extremely marginal performance improvements.

Appendix C. Variable importance

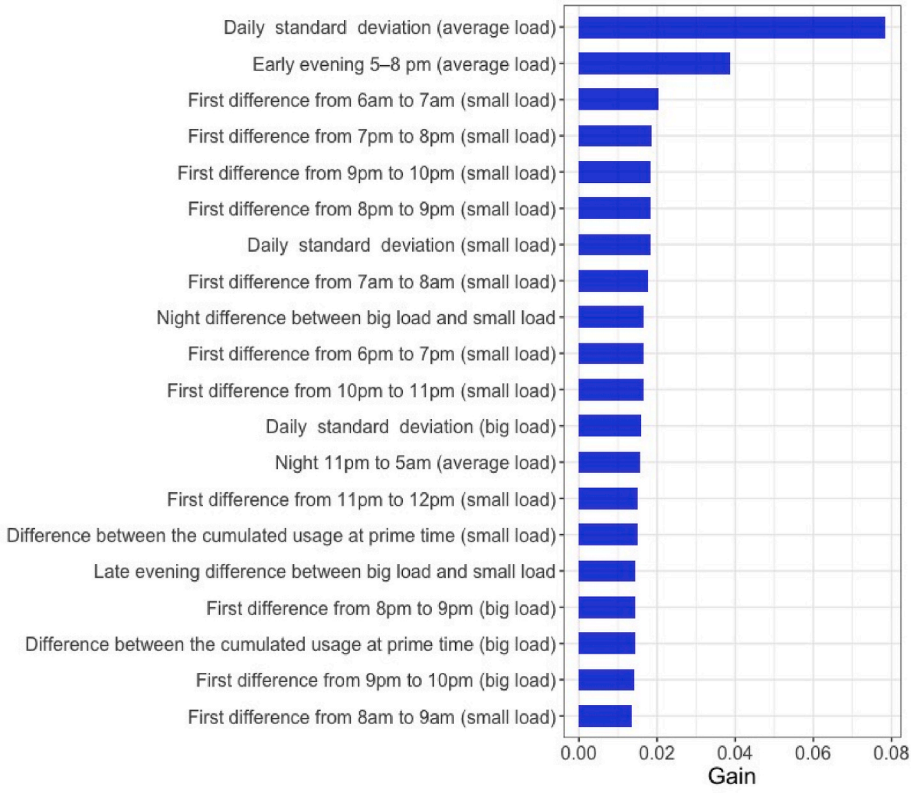

Fig. C1. Variable importance for the 20 features with the largest predictive power in XGBoost according to the Gain metric.

Table C1
Point-Biserial Correlation between features with the largest predictive power and binary business usage label in XGBoost training sample.

Variable	Point-Biserial Correlation
Daily standard deviation (average load)	0.185***
Early evening 5–8 pm (average load)	0.191***
First difference from 6am to 7am (small load)	0.008***
First difference from 7pm to 8pm (small load)	−0.017***
First difference from 9pm to 10pm (small load)	−0.026***
First difference from 8pm to 9pm (small load)	−0.032***
Daily standard deviation (small load)	0.081***
First difference from 7am to 8am (small load)	0.035***
Night difference between big load and small load	0.047***
First difference from 6pm to 7pm (small load)	0.007***
First difference from 10pm to 11pm (small load)	−0.008***
Daily standard deviation (big load)	0.143***
Night 11pm to 5am (average load)	0.039***
First difference from 11pm to 12pm (small load)	−0.021***
Difference between the cumulated usage at prime time (small load)	0.021***
Late evening difference between big load and small load	0.043***
First difference from 8pm to 9pm (big load)	−0.022***
Difference between the cumulated usage at prime time (big load)	0.039***
First difference from 9pm to 10pm (big load)	−0.036***
First difference from 8am to 9am (small load)	0.023***

Notes: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Point-Biserial Correlation between the 20 features with the largest predictive power in XGBoost according to the Gain metric and the binary label business or private customers in the training sample. Note that business usage is coded as 1 and private usage as 0.

Appendix D. The Product

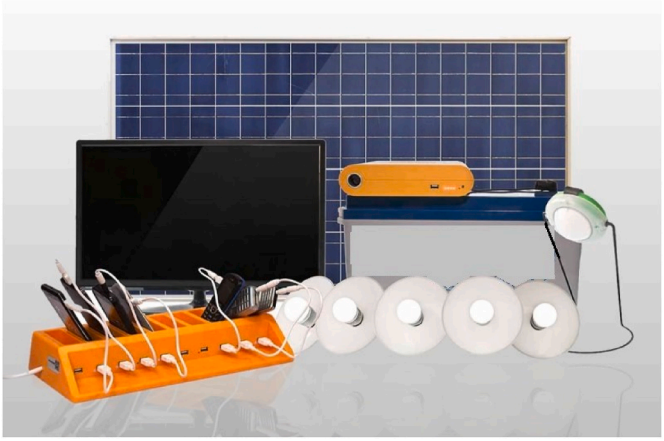

Fig. D1. One version of the solar panel home systems sold in Tanzania. Source: provided by the company.

Fig. D2. Customer using the solar panel home system to operate a village cinema. Source: private.

Appendix E. Performance measures and Receiver Operating Characteristic curve

We provide a short introduction to the main performance measures that are commonly used in the supervised machine learning literature. Detailed information on these measures can be found for instance in (Gron, 2017) and (James et al., 2014). This section is in particular intended to introduce the reader to Receiver Operating Characteristic (ROC) and Area under the Receiver Operating Characteristic curve (AUROC), which we use to evaluate the performance of our classification approach with XGBoost. We refer to the customer-day observations that are correctly classified as a business usage days as *True Positives* (T_P) and those that are falsely classified as positive as *False Positives* (F_P). Correspondingly, customer-day observations that are correctly classified as negative are *True Negatives* (T_N) and those that are falsely classified as negative *False Negatives* (F_N). The following performance measures are conventionally used in the literature and are essential to understand the computation and interpretation of the ROC and AUROC.

Accuracy measures the percentage of customer-day observations that are correctly classified, i.e.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}.$$

With a classification threshold of 0.5 applied to the predictive probabilities on the test sample. However, the Accuracy metric is not suitable for highly imbalanced data sets, such as our data with business user days as the underrepresented class.

Recall measures the proportion of customer-day observations that are correctly classified as business user days (*True Positives*) to all observations that are in fact business user days (*True Positives* and *False Negatives*). Note that Sensitivity or True Positive Rate is also used to refer to Recall such that

$$True\ Positive\ Rate = \frac{T_P}{T_P + F_N}.$$

Correspondingly, the False Positive Rate is defined as

$$False\ Positive\ Rate = \frac{F_P}{F_P + T_N}.$$

We use the Receiver Operating Characteristic (ROC) and Area under the Receiver Operating Characteristic curve (AUROC) as the main performance evaluation approach. The ROC curve combines the Recall with the False Positive Rate by plotting both for a grid of threshold probabilities, which we denote by c. Hence, the ROC curve is given by pairs of the True Positive Rate and False Positive Rate that are computed for different values of c as denoted below:

$$ROCcurve = \left\{ \left(\frac{T_P(c)}{T_P(c) + F_N(c)}, \frac{F_P(c)}{F_P(c) + T_N(c)} \right), c \in (0,1) \right\}$$

The AUROC is given by the area under the ROC curve. A classifier that is completely random (e.g. tossing a coin) has an AUROC score of 0.5, while a perfect classifier has an AUROC score of 1. The AUROC is particularly well suited to evaluate our classification approach, because we do not decide on a specific threshold c to distinguish between business and private users, but use the predicted daily probabilities as a more precise measure for the intensity of business usage. The ROC for the test sample is shown in Fig. E1. The corresponding AUROC is 0.784, which is an improvement against the benchmark of a random classifier with an AUROC of 0.5.

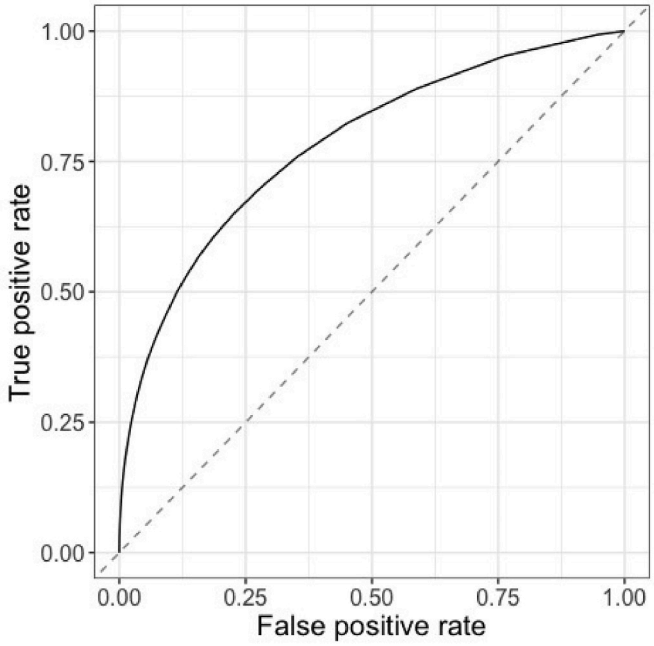

Fig. E1. The graph displays the Receiver Operating Characteristic (ROC) for the test sample.

References

Barry, M.S., Creti, A., 2020. Pay-as-you-go contracts for electricity access: bridging the "last mile" gap? a case study in Benin. Energy Econ. 90, 104843.

Beckel, C., Sadamori, L., Santini, S., 2013. Automatic socio-economic classification of households using electricity consumption data. In: Proceedings of the Fourth International Conference on Future Energy Systems, vol. 13. Association for Computing Machinery, eEnergy, pp. 75–86.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M., 2016. mlr: machine learning in r. J. Mach. Learn. Res. 17, 1–5.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, KDD '16, New York, NY, USA, pp. 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., contributors, X., 2020. R Package: Xgboost. https://cran.r-project.org/web/packages/xgboost/index.html.

Cox, D.R., 1972. Regression models and life-tables. J. Roy. Stat. Soc. B 34, 187–220.

D'Agostino, A.L., Lund, P.D., Urpelainen, J., 2016. The business of distributed solar power: a comparative case study of centralized charging stations and solar microgrids. WIREs Energy Environ. 5, 640–648.

Feron, S., 2016. Sustainability of off-grid photovoltaic systems for rural electrification in developing countries: a review. Sustainability 8. https://www.mdpi.com/2071-1050/8/12/1326.

Gao, J., Mills, B.F., 2018. Weather Shocks, Coping Strategies, and Consumption Dynamics in Rural Ethiopia, vol. 101. World Development, pp. 268–283.

Gogla, opportunity, Powering, 2018. The economic impact of off-grid solar. Tech. Rep. Global Association for the Off-grid Solar Energy Industry.

Grohmann, A., Herbold, S., Lenel, F., 2021. Repayment under flexible loan contracts: evidence from high frequency data. Tech. Rep. Available at: https://ssrn.com/abstract=3917712.

Gron, A., 2017. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, first ed. O'Reilly Media, Inc.

Harun, M.A., 2015. The Role of Solar Home System (Shs) in Socio-Economic Development of Rural bangladesh, Dissertation. BRAC University. Available at: https://core.ac.uk/download/pdf/61807642.pdf.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2014. An Introduction to Statistical Learning: with Applications in R. Springer.

Kleiminger, W., l, C., Staake, T., Santini, S., 2013. Occupancy detection from electricity consumption data. In: Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings. Association for Computing Machinery, BuildSys'13, pp. 1–8.

Lemaire, X., 2018. Solar Home Systems and Solar Lanterns in Rural Areas of the Global South: What Impact?, vol. 7. Wiley Interdisciplinary Reviews: Energy and Environment, e301.

Mathenge, M.K., Tschirley, D.L., 2015. Off-farm labor market decisions and agricultural shocks among rural households in Kenya. Agric. Econ. 46, 603–616.

Mondal, A.H., Klein, D., 2011. Impacts of solar home systems on social development in rural Bangladesh. Energy Sustain. Dev. 15, 17–20.

Rom, A., Günther, I., Borofsky, Y., 2020. Using sensors to measure technology adoption in the social sciences. Dev. Eng. 5.

Therneau, T.M., 2020. A Package for Survival Analysis in R. https://CRAN.R-project.org/package=survival, R package version 3.2-3.

Therneau, T.M., Grambsch, P.M., 2000. Modeling Survival Data: Extending the Cox Model. Springer, New York.

Viegas, J., Vieira, S., Melicio, R., Mendes, V., Sousa, J., 2016. Classification of new electricity customers based on surveys and smart metering data. Energy 107, 804–817.
Wassie, Y.T., Adaramola, M.S., 2021. Socio-economic and environmental impacts of rural electrification with Solar Photovoltaic systems: evidence from southern Ethiopia. Energy Sustain. Dev. 60, 52–66.
World Bank, 2017. Sustainable energy for all — progress toward sustainable energy. Tech. Rep., International Energy Agency (IEA) and the World Bank.
World Bank Group, 2020. Off-grid solar market trends report 2020. Tech. Rep., International Finance Corporation.
Zufferey, D., Gisler, C., Khaled, O.A., Hennebert, J., 2012. Machine learning approaches for electric appliance classification. In: 2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA), pp. 740–745.