Center for European, Governance and Economic Development Research | cege

cege Discussion Paper

No. 428

September 2021

Using Solar Panels for Business Purposes: Evidence Based on High-Frequency Power Usage Data

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

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

ISSN: 1439-2305

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] and 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 HISTORY
Compiled September 6, 2021

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.

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

Contact: Friederike Lenel. Email: friederike.lenel@uni-goettingen.de.

1. Introduction

In economically less developed countries, where access to electricity is still very limited [22], solar panel systems not only provide electricity for private consumption [7], 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 [23], 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 [17].

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 [11]. 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 train a supervised classifier to predict each customer's likelihood of using the system for business purposes on a daily basis. 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 generate features that capture relevant dimensions of electricity usage (among others, its average intensity over time, its variance as well as its hourly dynamics). We train a supervised classifier, XGBoost (Extreme Gradient Boosting, [4]), based on the first months of individual electricity usage data. The labels for business use are derived from 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 covers a customer's intended purpose of the system. Subsequently, we use XGBoost for out-of-sample predictions of the individual likelihood of being a business user at a daily level throughout the whole time period. In order to study the implications of business usage for repayment, we relate 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 [19]. An overview of the process is shown in Figure 1.

We show that the supervised classification approach to capture electricity usage behavior can be implemented relatively easily, as long as some labelled data exists, and performs reasonably well. 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, [24] show how machine learning approaches can be used to detect the type of electrical home appliance used; [20] combine survey data with smart metering data to classify customers according to their electricity consumption; [2] predict socio-economic properties of households, and [13] 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 [23]. 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, [14, 16, 21]).[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 on credit with flexible repayment schemes [1]. 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

[1]The results are mixed. Surveying solar panel users in Ethiopia, [21] 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. [12] 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 [8]). 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 [10].

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.

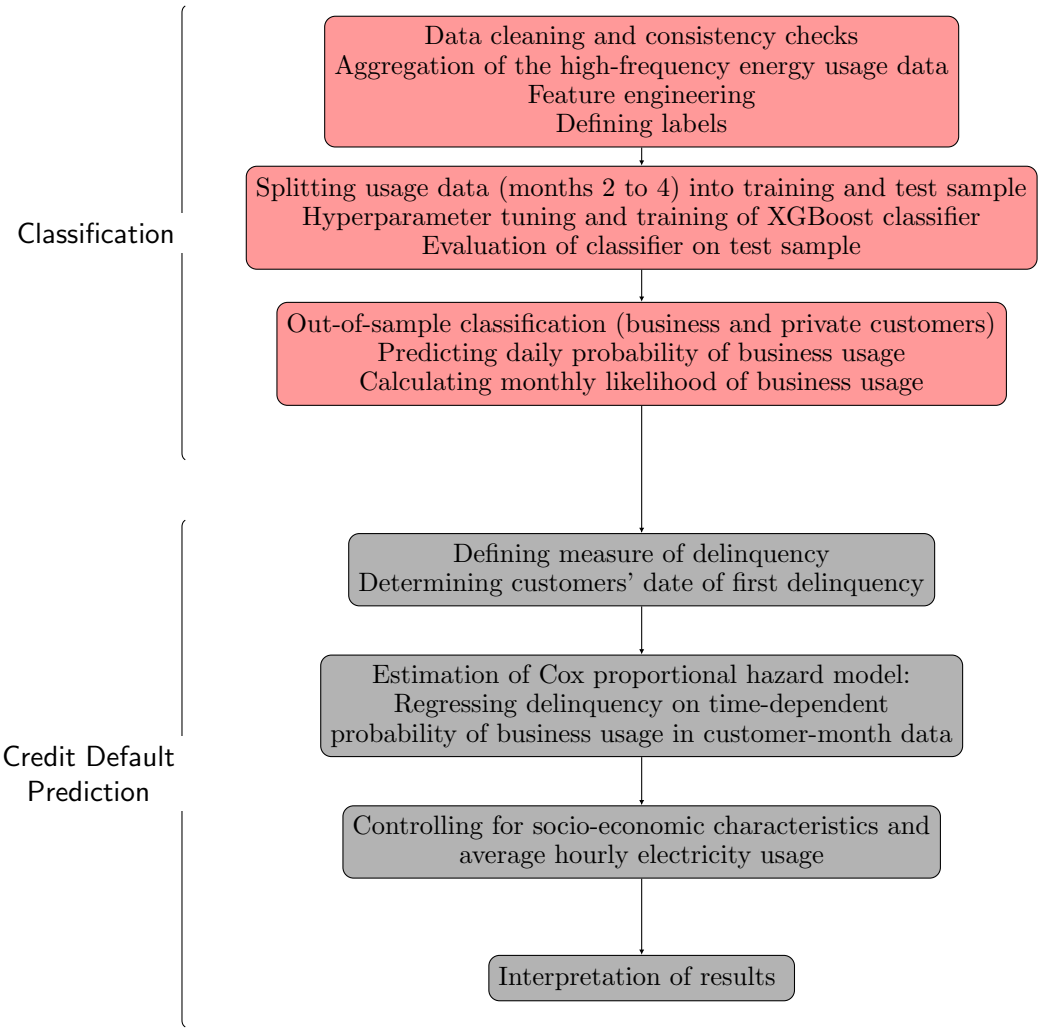


Figure 1. Process for solar panel user classification and credit default prediction.

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 households in Tanzania. There are different system types that vary in the system's power (80W-200W) and the appliances that come with the system (see Figure 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 Figure 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 1,300 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. 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. If households are unable or unwilling to 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 ten minutes 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.

2.2. 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 ten minute 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).[2] 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 to 30% over time. The loan eligibility 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-employed, wage-employed 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 [11]). 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 (74%) 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 (81%) and live in rural areas (87%). The majority are either farmers (47%) or operate their own business (30%); only few are wage-employed.[3] On average, households consume 7W 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 40W; a TV consumes on average, depending on

[2]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 representative of the complete sample we are analysing in terms of socio-economic characteristics.

[3]For a more detailed description of the customer profiles and how they compare to the Tanzanian population see [11].

screen size and brightness, between 11 and 24W, while a light consumes just around 1 to 3W.

Figure 2 shows the energy usage profiles for the average load over a day for a randomly selected day of four randomly selected customers. Presumably customers 1 and 2 use the generated electricity primarily for lighting as their energy usage goes down after 6 am and then goes up again at 6 pm.[4] Customers 1, 3 and 4 experience a clear usage spike in the evening, potentially when they come home from work and watch TV or listen to the radio. In contrast to customers 1, 2 and 3, the usage profile of customer 4 has a more distinct usage pattern during the day. Potentially, this household uses the system for business purposes, e.g., by operating a shop that closes during lunch time. However, the system's purpose cannot be inferred unambiguously solely by observing the usage profiles.

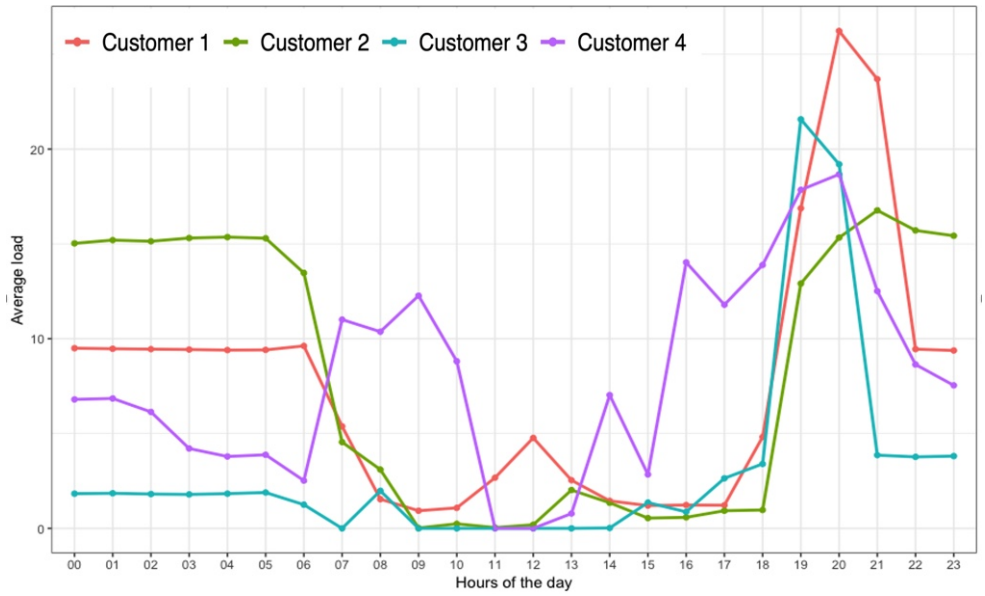
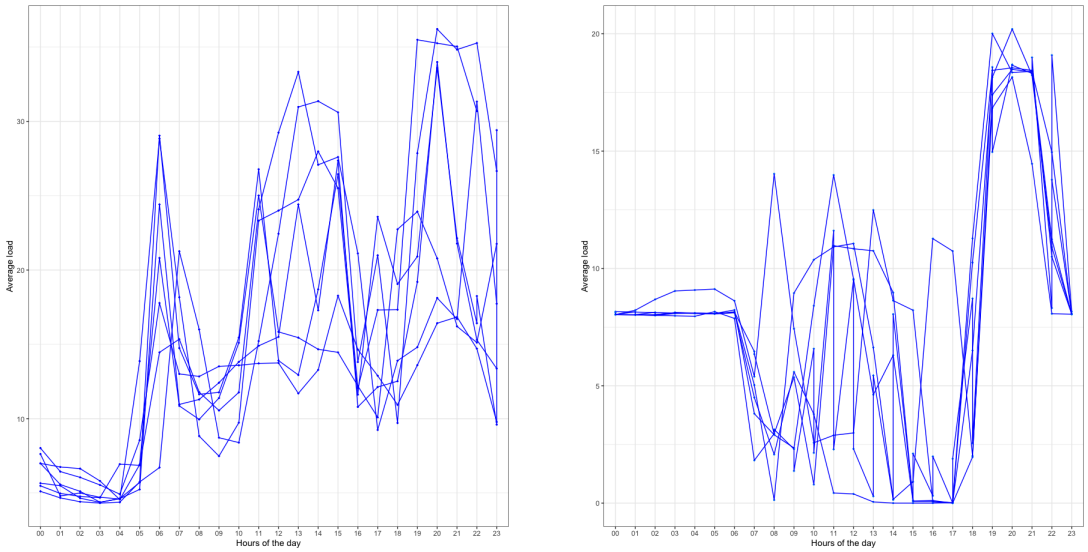


Figure 2. The graph displays four randomly selected electricity usage profiles for the average load over a day. The unit of measurement is Watt.

In addition, load profiles differ from day to day. Figure 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.

[4]In Tanzania sunrise generally happens between 6:15 am and 6:45 am through the year and sunset usually happens between 6:30 pm and 7:00 pm. Daily sunshine duration lasts usually about 12 hours.



Customer A Customer B

Figure 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. 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. We rely on a supervised classification approach for this purpose, 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.

3.1. 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 am, noon 11 am–2 pm, afternoon 2–5 pm, early evening 5–8 pm, late evening 8–11 pm, night 11 pm–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 am to 4 am), for total, small and big load.[5] [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 am–11 pm) and non-prime time (11 pm–8 am), for average, small and big load [3 features].

These features reflect not only average electricity use but also the 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

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

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 regularisation penalty for the random forest $\text{pen}(f(\boldsymbol{x}))$ is given by

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

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 $\hat{f}(\boldsymbol{x})$ is iteratively updated as

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

where v denotes the iteration index and $\hat{g}^{(v)}(\boldsymbol{x})$ 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 [4].[6]

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 provides those customers who indicated that they plan to use the system for business purposes with sufficient time to establish such a

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

[6] For the implementation, we use the the mlr package [3] and XGBoost implementation in R [5]. The hyperparameter tuning is presented in appendix B.

business, 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. 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.

Figure 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 Figure 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.

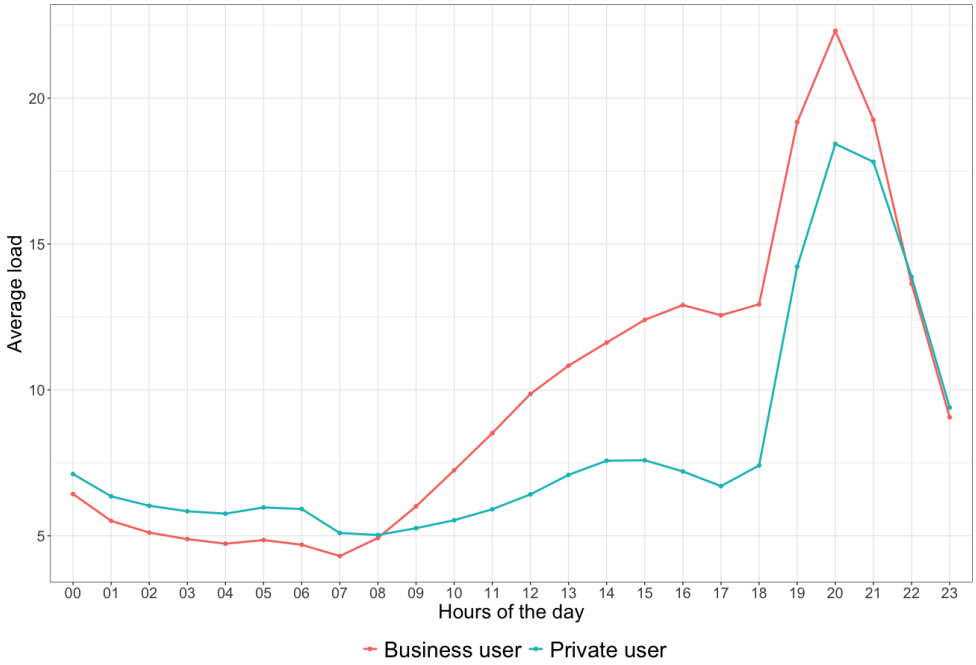
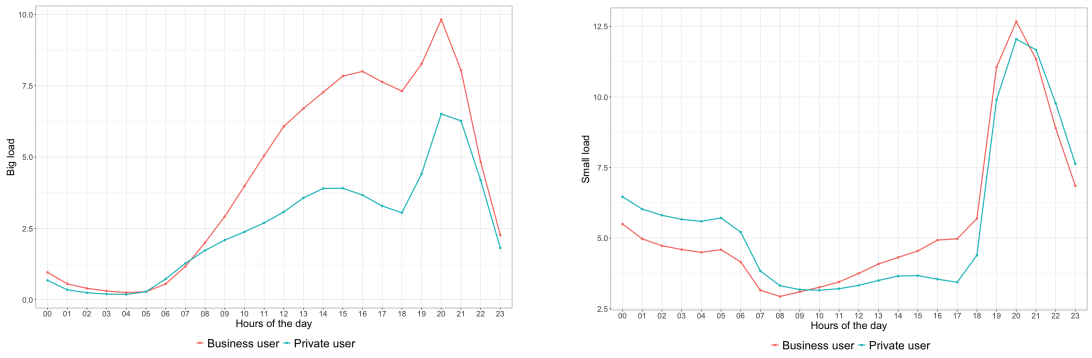


Figure 4. The graph displays the average daily electricity usage profile belonging to business and non-business customers in our training sample for average load.



Big Load Small Load

Figure 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.

3.3. Classification Results

Figure 6 displays the predicted probabilities of business usage within our training and 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.[7] 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. A threshold of 10% business probability can already clearly discriminate between business and private users.

Figures 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 A3). As shown above, a 10% cut-off of business usage probability already discriminates reasonably well between private and business usage; when using this cut-off to determine a day as "business-day," on average 23% of a household's usage days can be defined as business days, i.e. as days on which customers use the system presumably for business purposes (see also Figure 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 Figure A4).

On average, the predicted business probably does not change much over time (see Figure 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. Figure 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

[7]Standard performance metrics such as the Area under the Receiver Operating Characteristic (ROC) curve (AUC) (0.784) and Area under the Precision-Recall (PR) curve (AUCPR) (0.325) suggest that the classifier performs reasonably well given the data structure and the classification task.

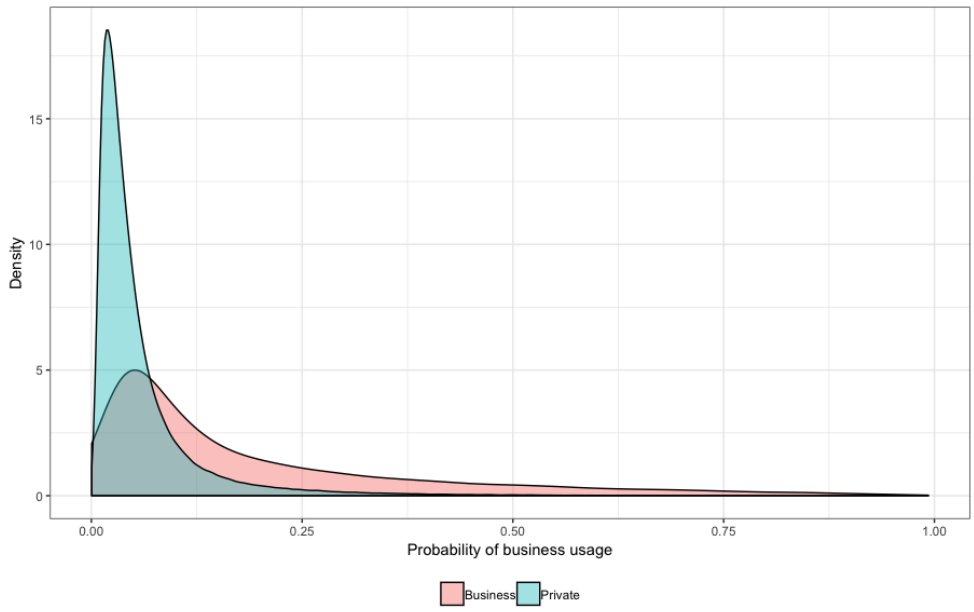


Figure 6. The graph displays densities of the predicted probabilities of business usage for the business and private users in the test sample.

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, Figure C1 in appendix C displays the 20 most important features that predict business-like electricity usage by XGBoost according to the Gain metric.[8] In addition, Table C1 reports the correlation between the most important features and the binary business usage label. We find that 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. Days that show high volatility and a rather high electricity consumption in the early evening hours are more likely to be classified as business usage days. 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.

3.4. Discussion

Using a labeled 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 mis-reporting by prospective users. 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

[8]The Gain metric measures the total gain in the classification performance that results from splits in the trees for the respective feature. It is a conventional metric for measuring feature importance with XGBoost [4, 5].

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 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 the year; recalling past usage behavior instead can be prone to reporting errors [17].

If labeled 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 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. 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 [6, 18] with the time-dependent business probability as explanatory variable in the following form:

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

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, $b_i(t)$.[9] More specifically, $b_i(t)$ describes household i's predicted business probability averaged over all days the system was used in month t.[10] 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

[9] See Table A2 for the summary statistics of all variables included in the model.

[10] 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.

$\exp(\delta_1)$ 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 80W, 120W and 200W. 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 socio-economic 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.

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 to 47 percent, depending on the specification. While this is only a correlation analysis and one should be cautious interpreting these results as causal,[11] the results show that the derived predicted business usage probability is a meaningful indicator that can help predict repayment difficulties.

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 watt		1.008	0.966*
System with 200 watt		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 28,249. Coefficients are reported in form of hazard ratios (HR) by using the exponential function.

[11] For example, wealthier customers, who should have 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.

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 the average predicted business probability is low, there is 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 [9, 15]. 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.

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. Linking electricity usage and repayment data with information on extreme weather events would allow investigating whether the use of the solar panel systems for income diversification can help farmers to overcome negative income shocks resulting from harvest loss. Moreover, the high-frequency electricity usage data gathered from solar panels can 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.

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.

References

[1] M.S. Barry and A. Creti, *Pay-as-you-go contracts for electricity access: Bridging the "last mile" gap? a case study in benin*, Energy Economics 90 (2020), p. 104843.

[2] C. Beckel, L. Sadamori, and S. Santini, *Automatic Socio-Economic Classification of Households Using Electricity Consumption Data*, in *Proceedings of the Fourth International Conference on Future Energy Systems*. Association for Computing Machinery, e-Energy 13, 2013, pp. 75–86.

[3] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z.M. Jones, *mlr: Machine learning in r*, Journal of Machine Learning Research 17 (2016), pp. 1–5.

[4] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA. ACM, KDD '16, 2016, pp. 785–794.

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

[6] D.R. Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society. Series B. 34 (1972), pp. 187–220.

[7] A.L. D'Agostino, P.D. Lund, and J. Urpelainen, *The business of distributed solar power: a comparative case study of centralized charging stations and solar microgrids*, WIREs Energy and Environment 5 (2016), pp. 640–648.

[8] S. Feron, *Sustainability of off-grid photovoltaic systems for rural electrification in developing countries: A review*, Sustainability 8 (2016).

[9] J. Gao and B.F. Mills, *Weather shocks, coping strategies, and consumption dynamics in rural ethiopia*, World Development 101 (2018), pp. 268–283.

[10] GOGLA, *Powering opportunity. the economic impact of off-grid solar*, Tech. Rep., Global Association for the Off-grid Solar Energy Industry, 2018.

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

[12] M.A. Harun, *The role of solar home system (shs) in socio-economic development of rural bangladesh*, dissertation, BRAC University, 2015. Available at https://core.ac.uk/download/pdf/61807642.pdf.

[13] W. Kleiminger, C. l, T. Staake, and S. Santini, *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, 2013, pp. 1–8.

[14] X. Lemaire, *Solar home systems and solar lanterns in rural areas of the Global South: What impact?*, Wiley Interdisciplinary Reviews: Energy and Environment 7 (2018), p. e301.

[15] M.K. Mathenge and D.L. Tschirley, *Off-farm labor market decisions and agricultural shocks among rural households in kenya*, Agricultural Economics 46 (2015), pp. 603–616.

[16] A.H. Mondal and D. Klein, *Impacts of solar home systems on social development in rural bangladesh*, Energy for Sustainable Development 15 (2011), pp. 17–20.

[17] A. Rom, I. Günther, and Y. Borofsky, *Using sensors to measure technology adoption in the social sciences*, Development Engineering 5 (2020).

[18] T.M. Therneau, *A Package for Survival Analysis in R* (2020). Available at https://CRAN.R-project.org/package=survival, R package version 3.2-3.

[19] T.M. Therneau and P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer, New York, 2000.

[20] J. Viegas, S. Vieira, R. Melicio, V. Mendes, and J. Sousa, *Classification of new electricity customers based on surveys and smart metering data*, Energy 107 (2016), pp. 804–817.

[21] Y.T. Wassie and M.S. Adaramola, *Socio-economic and environmental impacts of rural*

electrification with Solar Photovoltaic systems: Evidence from southern Ethiopia, Energy for Sustainable Development 60 (2021), pp. 52–66.

[22] World Bank, *Sustainable energy for all — progress toward sustainable energy*, Tech. Rep., International Energy Agency (IEA) and the World Bank, 2017.

[23] World Bank Group, *Off-grid solar market trends report 2020*, Tech. Rep., International Finance Corporation, 2020.

[24] D. Zufferey, C. Gisler, O.A. Khaled, and J. Hennebert, *Machine learning approaches for electric appliance classification*, in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. 2012, pp. 740–745.

6. Appendices

Appendix A. Descriptive Statistics

Table A1. Descriptive Statistics for all customers

Variable	Mean	Std. Dev.	Min	Max
Male	0.812	0.391	0	1
Household size	4.347	1.880	1	30
Self employed	0.299	0.458	0	1
Wage employed	0.218	0.413	0	1
Farmer	0.469	0.499	0	1
Urban	0.128	0.334	0	1
Average hourly usage	7.289	3.318	0.010	50.883
Delinquent (at least once)	0.739	0.439	0	1

Notes: Summary statistics for all customers included in the analysis.

Table A2. 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	0.112	0.316	0	1
System with 80 watt	0.658	0.474	0	1
System with 120 watt	0.342	0.474	0	1
System with 200 watt	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.

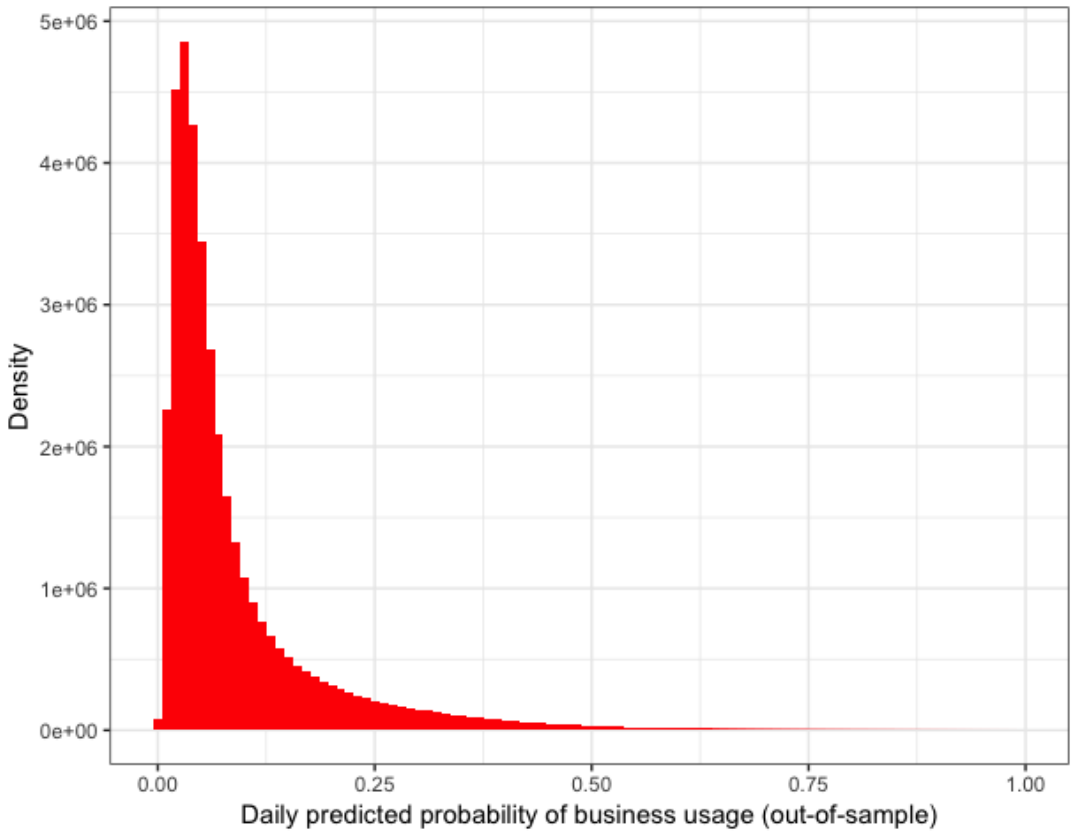


Figure A1. The graph displays a histogram of the daily out-of-sample predicted probabilities of business usage.

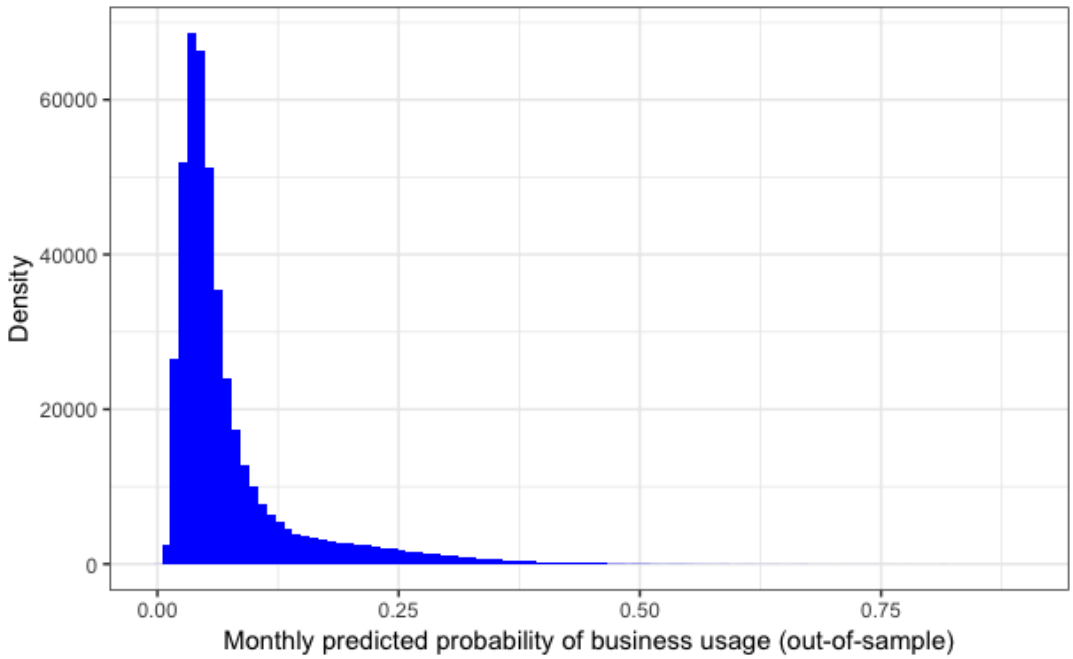


Figure A2. The graph displays a histogram of the out-of-sample predicted probabilities of business usage that are used in the Cox model aggregated on the monthly level.

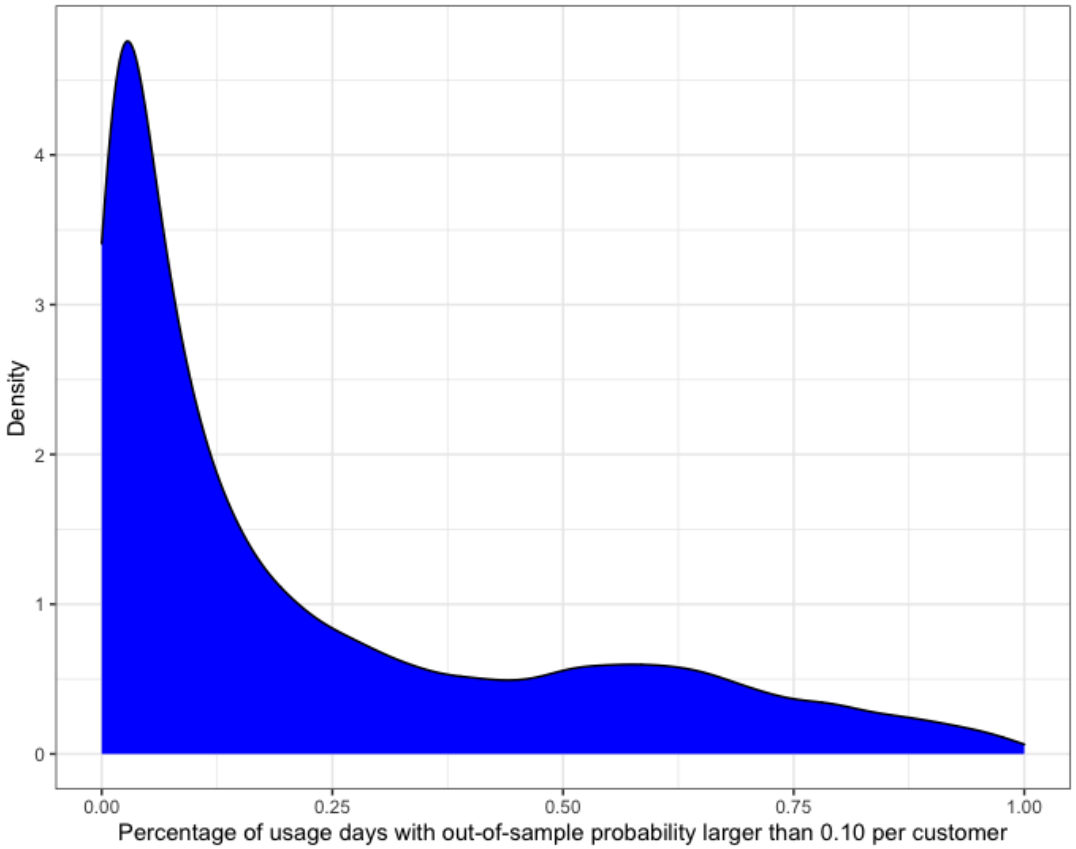


Figure A3. The graph displays a histogram of the Percentage of usage days with P(Business) ≥ 0.1 on the customer level.

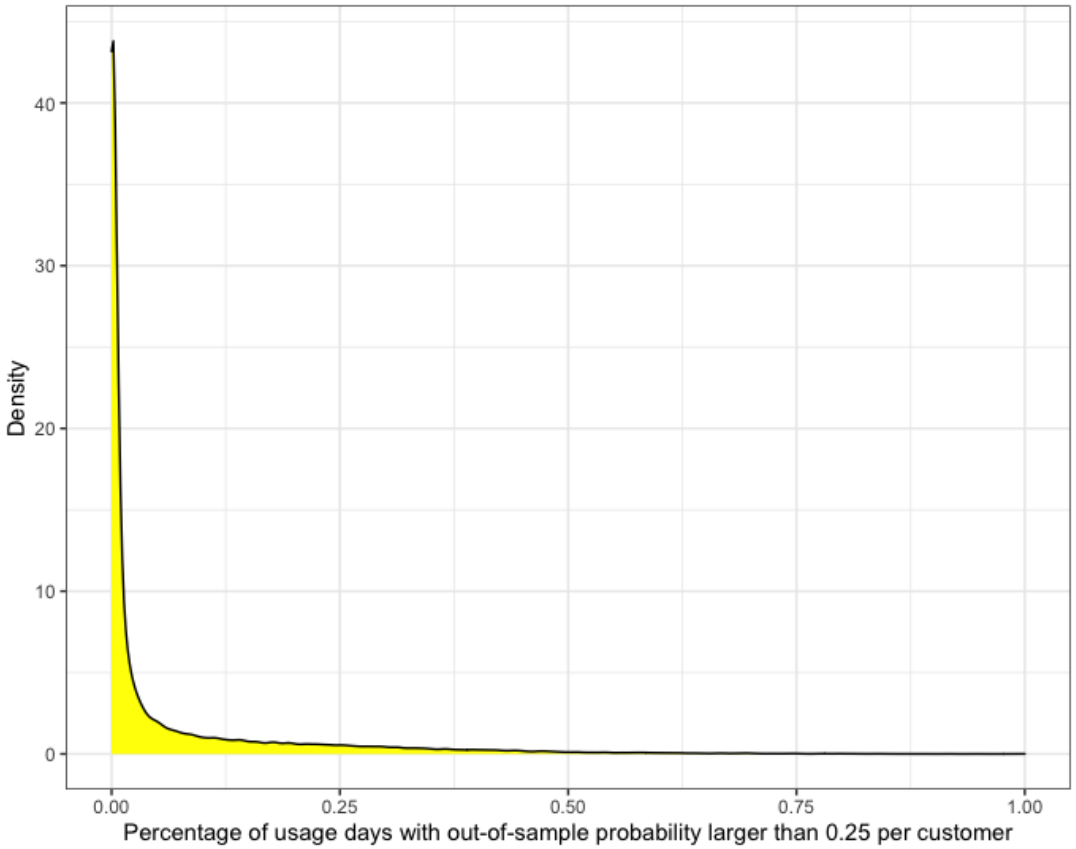


Figure A4. The graph displays a histogram of the Percentage of usage days with P(Business) ≥ 0.25 on the customer level.

Table A3. 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) ≥ 0.10	0.000	0.037	0.118	0.236	0.389	1.000
Percentage of usage days with P(Business) ≥ 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.

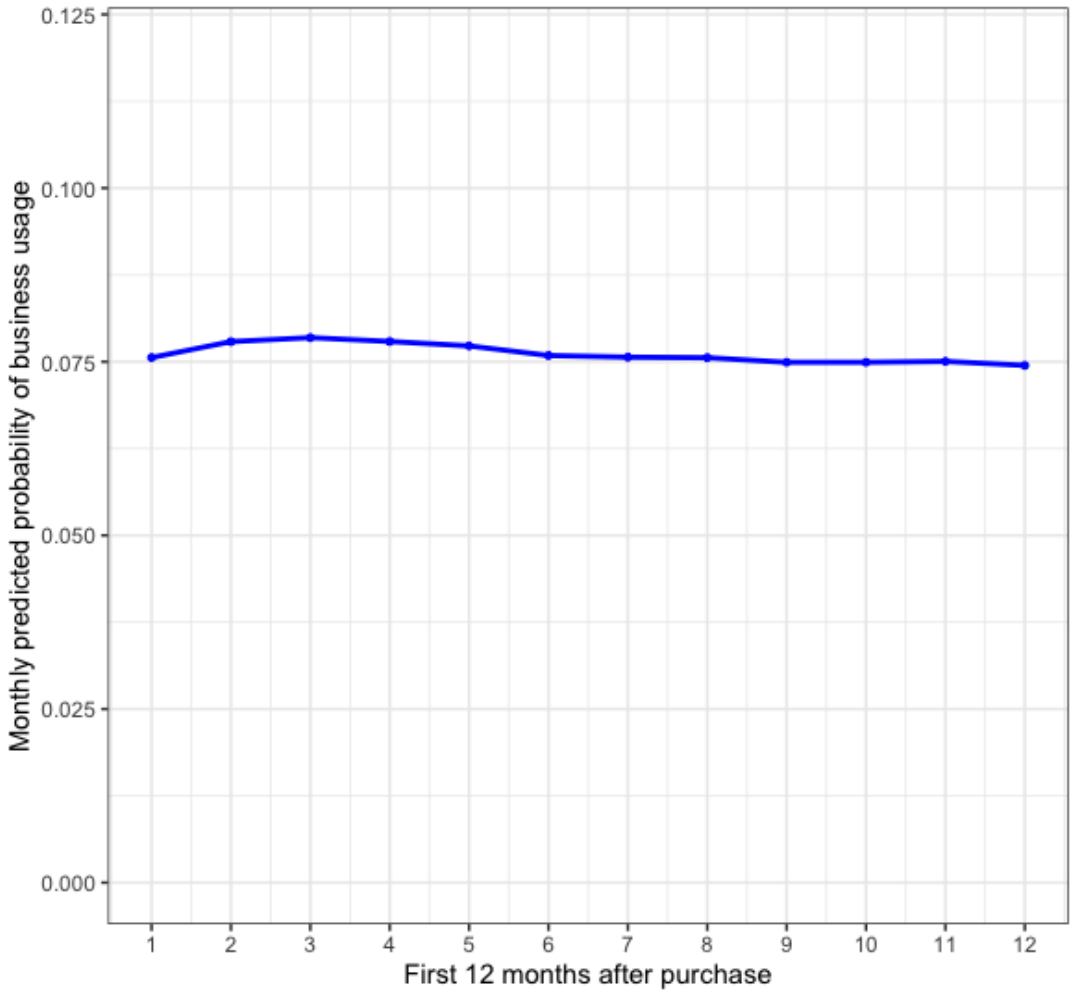


Figure A5. The graph displays the average monthly predicted probability of business usage for customers that use the system for at least 12 months without a delinquency.

Figure 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

B.1. 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 to 12. For the minimum number of observations in the terminal node we set the range of 1 to 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 to 500 number of iterations. Several trials show that a larger range only leads to extremely marginal performance improvements.

Appendix C. Variable importance

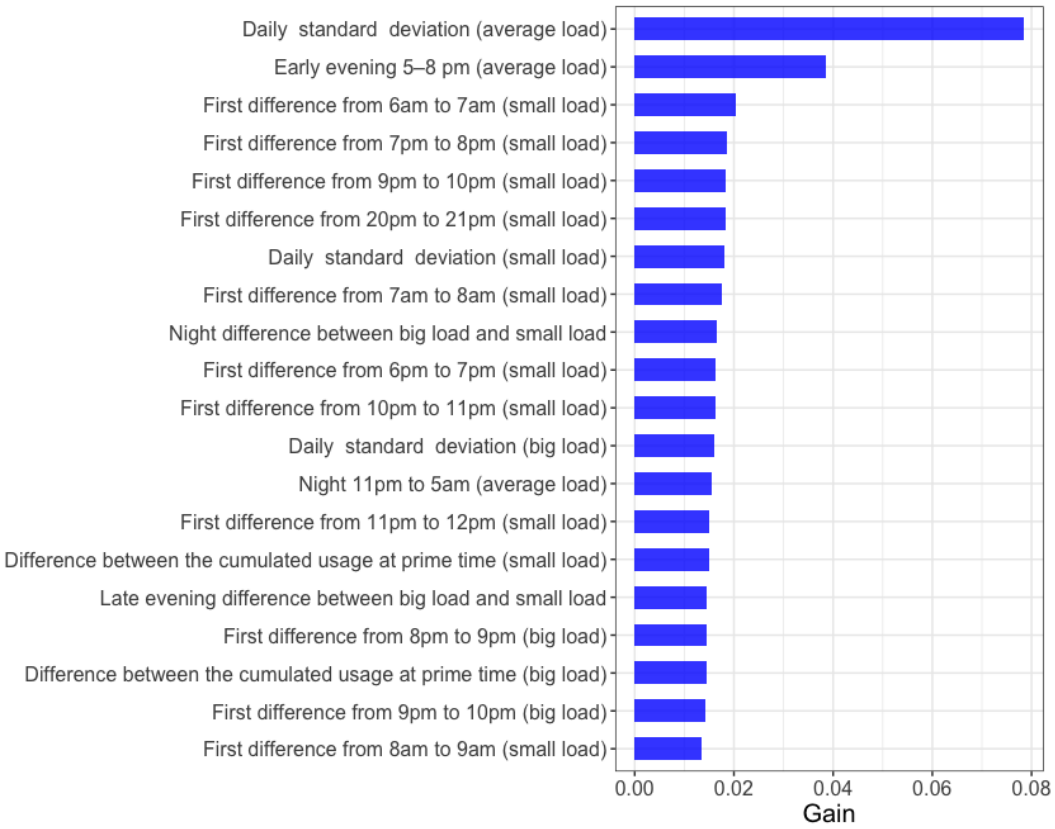


Figure C1. Variable importance for the 20 features with the largest predictive power in XGBoost according to the Gain metric.

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 20pm to 21pm (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

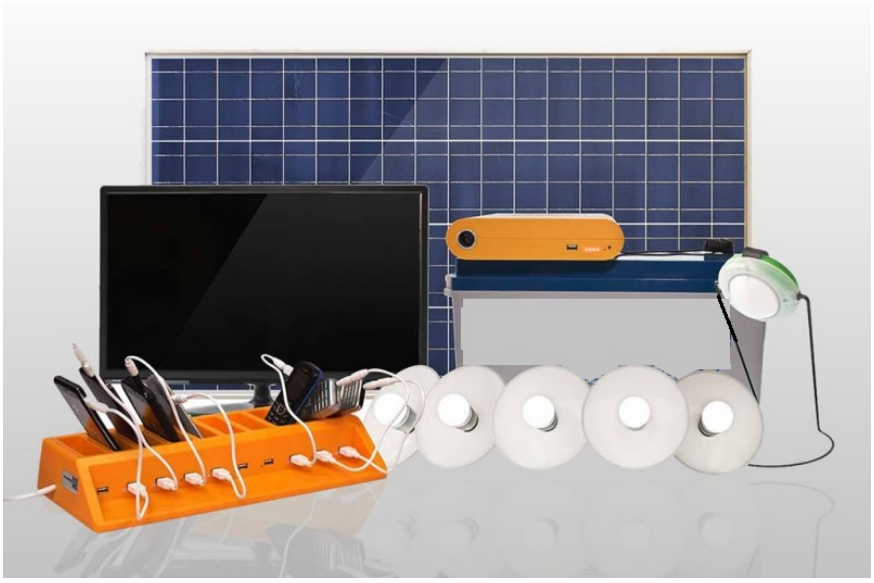


Figure D1. One version of the solar panel home systems sold in Tanzania. Source: provided by the company.

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