Predictive Modeling of Customer Repayment for Sustainable Pay-As-You-Go Solar Power in Rural India

Hugo Gerard*
Innovations for Poverty Action
New Haven, CT
hugorgerard@gmail.com

Kamalesh Rao*
MasterCard Advisors
Purchase, NY
kamalesh@gmail.com

Mark Simithraaratchy*
Coach
New York, NY
mark.simithraaratchy@gmail.com

Kush R. Varshney*
IBM T. J. Watson Research Center
Yorktown Heights, NY
krvarshn@us.ibm.com

Kunal Kabra
Simpa Networks
Noida, UP, India
kunal.kabra@simpanetworks.com

G. Paul Needham
Simpa Networks
Noida, UP, India
paul@simpanetworks.com

ABSTRACT
In this paper, we describe a data science project analyzing the repayment behavior of customers of pay-as-you-go (PAYG) solar power systems in Indian villages that experience severe cuts in grid power. The innovative PAYG paradigm allows people that cannot afford the capital expense of a solar panel at any one time to finance their acquisition with a small down payment and affordable 'recharge' payments. In particular, we examine data from the social enterprise Simpa Networks and develop a logistic regression model to predict the risk of a customer failing to make recharge payments until the system is fully paid for. This prediction is to be made at the point in time when a person applies to be part of the program using attributes solicited on an application form. The task is made difficult because repayments take place over two to three years and the sample size of customers with that much history is small, because application form data is noisy, incomplete and contains many free text fields, and because the customer population characteristics are changing over time. Nevertheless, we are able to obtain classification performance results with the promise for business impact.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning; J.4 **[Social and Behavioral Sciences]**: Economics

General Terms
Algorithms

Keywords
customer analytics; poverty economics; social good

1. INTRODUCTION
Pay-as-you-go (PAYG) is an emerging model in sub-Saharan Africa and South Asia for providing solar power to the rural poor in villages that are either not connected to the power grid or experience severe power cuts [1]–[3]. Clean and safe sources of light and electricity can profoundly improve the lives of residents of such places in many ways. In the status quo, people rely on unhealthy and unsafe kerosene lamps and candles, one reason being because both commodities can be purchased in small quantities as funds become available. Members of this population, approximately 1.6 billion people worldwide [4], cannot afford the healthier, safer and more useful option of solar panels because of the large capital expense required at one time. PAYG is an innovative business model and financing scheme that puts solar power systems within reach of the rural poor by allowing customers to make small payments at a rate similar to how they pay for kerosene.

Figure 1. Simpa Networks' solar photovoltaic system.

Simpa Networks is a social enterprise pursuing the PAYG strategy to provide solar-as-a-service to households and small businesses in villages in Uttar Pradesh, India. Simpa customers make a small initial down payment for a high-quality solar photovoltaic system, shown in Fig. 1, and then pre-pay for the energy service, topping up their systems in small user-defined increments via mobile phone or with Simpa-affiliated merchants. Each payment for energy also adds toward the final purchase price. Once fully paid, the system unlocks permanently and produces energy, free and clear.

For Simpa to be financially sustainable, most customers in its portfolio should, on balance, continue to make payments until they have attained an unlocked system. Therefore, Simpa must enroll mostly 'good' customers and limit the number of 'bad' customers who stop paying and whose systems must be repossessed. This leads to a data science problem of trying to predict whether a customer will be good or bad, appropriately defined, at the point in time when he or she is applying for a solar power system. This problem can be viewed as one of credit scoring.

*The project described herein was conducted on a pro bono basis under the auspices of the DataKind DataCorps, New York, NY. Views expressed in this paper are those of the authors and not necessarily those of the authors' organizations.

Credit scoring is one of the early successful examples of predictive analytics and has been primarily approached through supervised machine learning techniques [5], [6]. However, it should be noted that most of the literature has focused on applications in which: 1) the customers are either from developed countries or from non-poor urban sections of developing countries, 2) the credit score does not depend on how a loan will be utilized, and 3) data is available on historical repayments of individuals from a range of sources. Herein, we examine the unique setting of the rural poor in India with a specific purpose for the credit: a photovoltaic system, and we only have access to data solicited in an application form. Nevertheless, we also approach the problem through supervised learning.

The remainder of the paper is organized as follows. In Section 2, we give more details on Simpa's business processes as well as how model predictions of good and bad customers can be used in various ways. In Section 3, we analyze customer repayment data and develop a concrete definition of good and bad customers to be used as a response variable for a supervised classification task. Section 4 discusses characteristics of the application form data as well as how we clean that data and derive new features from the raw application form attributes. Section 5 discusses the machine learning algorithm and presents empirical results. We conclude in Section 6 with a summary and several recommendations for the future.

2. BACKGROUND AND OBJECTIVE

As mentioned in Section 1, Simpa uses an application form to screen potential customers. The form is quite extensive and contains a large number of free-text fields. The information is entered into a web-based form, shown in Fig. 2, resulting in data known as digital customer application form (DCAF). The form is often completed by Simpa's urja mitra (sales agent) because of illiteracy or limited literacy in English of potential customers. The English proficiency of urja mitras is also sometimes less than fluent. Simpa's credit team also telephones the applicant, and in certain cases visits the applicant, to verify the information and ensure that he or she understands the risks and benefits of the loan.

The credit team approves or rejects these applications using its existing credit scoring model and some simple criteria that were developed during the early phases of the business. Potential customers are approved if it is believed that they are likely to make their repayments and rejected otherwise. It is up to the team to decide on the level of risk they want in their portfolio of customers, which determines the implicit threshold between good and bad applicants.

Once an application is approved, a system is delivered and installed at the customer's site. A typical system can power two or three lights and a fan. The physical device does not allow electricity to be drawn unless the customer has prepaid for energy days through their mobile device. A prepayment is known as a 'recharge,' to mirror the term used for prepayments of mobile phone minutes. The system enters a locked state once the paid for energy days have been exhausted. The typical prepayment amount is 30 energy days and customers are given a 5 day grace period. Once the system is fully paid for, typically in 2 or 3 years, it remains unlocked to deliver electricity without further payment. If a customer fails to make payments for a long time, the system is repossessed. Each activity, including delivery, activation, recharge, maintenance, unlocking, and repossession, are recorded with a timestamp in a revenue management system (RMS).

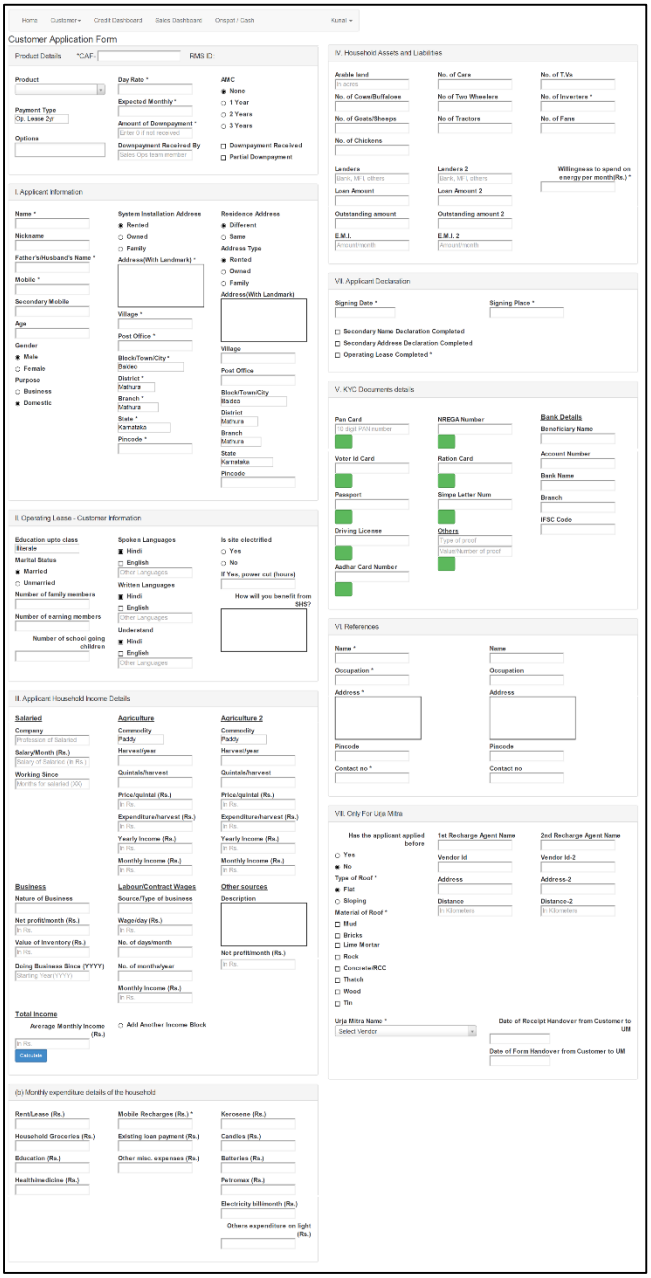

Figure 2. Digital customer application form.

In our work, we aim to use historical DCAF and RMS data from Simpa customers to develop a parsimonious predictive model of good and bad repayment behavior based on information in applications that generalizes to future potential customers. The model should be a binary classifier with probabilistic outputs to allow Simpa to choose an appropriate threshold according to the desired risk.

Such a model and its predictions can improve Simpa's business processes in a few different ways. First, the algorithmic predictions can help the credit team be more accurate, objective and rapid when they make approval decisions. Second, Simpa will be developing a mobile app to collect information from applicants to improve on the user experience for both applicants and urja mitras; a sparse predictive model can highlight which pieces of information are extraneous in DCAF allowing the app to be as simple as possible.

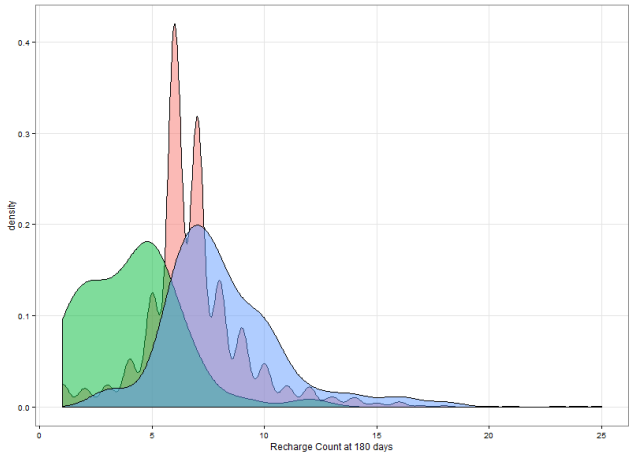

Figure 4. The distribution densities of the final status of accounts relative to the number of recharges in the first 180 days of the account. Unlocked accounts are in blue, accounts that are still making payments are in light red and repossessed accounts are in green. The final status of the accounts was determined at a snapshot in late January 2015.

Third, the model-based predictions can help Simpa better understand the risks of their customer portfolio as they move towards securitizing their loans.

3. CUSTOMER REPAYMENT ANALYSIS

The first step in developing the predictive model is understanding the patterns by which customers recharge their systems leading ultimately to either unlocking or repossession. Ideally we would have liked to use this final status as a binary class label for the machine learning task. Unfortunately, however, since Simpa only started operations in Uttar Pradesh under the current financing model in April 2011 (at a limited scale initially), there are very few customers in the available data who have been either unlocked or repossessed. Moreover, although unlocking is quite automatic and gives us a clean label of good customers, repossession is a biased indicator of a bad customer. Some bad customers' systems may be repossessed at different times simply due to the logistics involved with repossession. Also, sometimes systems are not repossessed from customers engaged in agriculture who are suffering a poor crop yield to give them a second chance. Therefore, we must define some other indicator of good and bad customers.

In lieu of using the final customer state as the class label, we desire a variable that can be calculated early in the customer relationship to ensure enough samples and that is indicative of the final state. There were a number of such 'intermediate' variables that would serve as good proxies for customer payment performance, many of which were highly correlated with one another. To determine which variable to use, we examine the relationship between the candidate intermediate variables and the final status of accounts, with the hope to find a variable that can distinguish the worst performing customers (those that had their units repossessed) from the rest (those customers who were able to keep up with their payments or payed off the cost of their unit entirely).

We settle on using the number of recharges in the first 180 days as our intermediate variable. As is demonstrated in Fig. 3, this measure of payment activity does a fairly good job of distinguishing weaker accounts from the rest. The data suggests that an account that has made more payments in its first 180 days is more likely to continue to make payments or pay for the entire

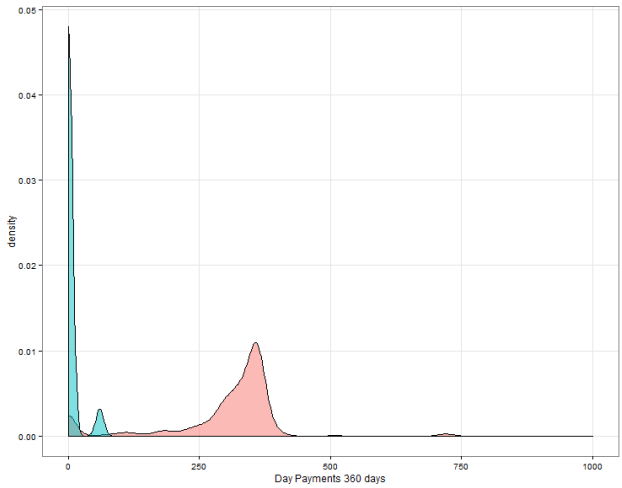

Figure 3. The distribution densities of the day payments. Accounts that had fewer than three recharge payments in the first 180 days are in cyan; those that had more are in light red.

unit than be repossessed. In particular, an account that made three or fewer payments in its first 180 days has a 60% chance of being repossessed (again, an imperfect measure of poor performance).

Another way to test the effectiveness of the intermediate variable is to see how well it determines the status of the account at the end of a year, which though not a proper measure of a 'final' status, gives a longer-term measure of performance than 180 days. And instead of using the three categories of 'unlocked,' 'repossessed' and 'ongoing,' a better proxy for performance, the account's 'day payments,' can be used. Simpa measures day payments as the total amount of recharge payments divided by the daily rate of the loan. This calculation then yields a day payment of around 360 if an account has kept up with the loan payment schedule over the course of the year, but less than 360 if an account has fallen behind.

As shown in Fig. 4, again, accounts that had fewer than three recharge payments in the first 180 days were much less likely to have kept up with payment schedule at the end of the year than those who had three or greater recharge payments. Note though that the separation is not perfect—there are customers that managed to catch up, reaching 300 or more day payments by the end of a year. Similarly, there are customers that managed to fall behind. One feature of the data is that customers do have a tendency to fall behind their payment schedule, which makes sense given the cash-poor nature of the local economy.

4. FEATURE EXTRACTION

With the class labels defined, the next step in setting up the machine learning problem is to extract features. DCAF, shown in Fig. 2, is the source of the raw data. It includes questions about demographics, sources of income, assets, expenses, reasons for wanting a system, etc.

A number of variables in the DCAF data are missing, in some cases due to data entry error and in others due to certain questions not being asked. We considered a few alternatives for dealing with missing values but settled on assigning a new label in the case of a missing categorical variable (such as, roof type or languages understood, for example). For numeric variables, we replace missing observations with the mean of non-missing values from the training set, and include a dummy variable equal to 1 if the observation is missing, and 0 otherwise.

Table 1. Sample of recoding for nature of business variable.

Nature of Business	Recoded Value
Animal husbanded	Livestock Business
Animals business	Livestock Business
Ara Machine	Skilled Labour/Driver
Arif Vairayte pailesh	Restaurant and Hotel
Ata Chakki	Grains/Fruits
Atta Chaki	Grains/Fruits
Atta Chakki	Grains/Fruits
Auto	Vehicle Sales and Repair
auto parts shop	Vehicle Sales and Repair
Auto Dealer	Vehicle Sales and Repair
Auto driver	Skilled Labour/Driver
auto mobile shop	Vehicle Sales and Repair
Auto parts	Vehicle Sales and Repair
Auto Parts Repairing	Vehicle Sales and Repair
Auto parts shop	Vehicle Sales and Repair
Auto spare parts	Vehicle Sales and Repair
Auto(Transport)	Skilled Labour/Driver
Automobiles shop	Vehicle Sales and Repair
Autoparts	Vehicle Sales and Repair
B.K satring Matarial Shop	Construction Business
Bafallo Sale Business	Livestock Business
Bajaj Agency	Vehicle Sales and Repair
Bajaz Dealer	Vehicle Sales and Repair
band	Tent House and Band
Band Business	Tent House and Band
Band Shop	Tent House and Band
banke bihari general store.	General Business
Baraber	Beauty
Barber	Beauty
bartan ki shop	Jewelry and Pots

There are also a number of fields where nearly all observations are missing, which are discarded. Recoding is done for free-text fields relating to commodities, nature of business and salaried company variables, and education into sensible categories. Two sample extracts of recodings are given in Table 1 and Table 2. Some other text fields are so non-standardized that we have to discard them.

For some numeric variables, such as number of goats owned, we recode these to categorical variables indicating 0, 1, or more than 1. Finally, we also construct new features, including normalizing by family size in a number of cases, and creating new binary indicator features for different income sources and sectors. Overall, after all processing, we arrive at 172 features.

We also make a distinction between features relating to the customer themselves (such as their demographics or employment) and variables relating to the rollout of the Simpa product (such as day rate, branch, and down payment). These latter variables, 6 in total, we term exogenous variables. These are important as they can be correlated with other variables of interest and excluding them could give misleading interpretations. But they are also not feasible to include when forecasting new regions or products. Therefore, in the following section we consider two feature sets, including and excluding these exogenous variables.

Table 2. Recoding for agricultural commodity variable.

Commodity	Recoded Value
Barley	Other
Dhaan	Paddy
Dhan	Paddy
Jo	Other
Maize	Other
Makka	Other
Millet	Other
mustard	Other
Others	Other
Paddy	Paddy
Patato	Potato
Peace	Other
peas	Other
Potato	Potato
Sarson	Other
Sorghum	Other
Sugarcan	Other
Sugarcane	Other
Sunflower	Other
Urad Dal	Other
weat	Wheat
Whaet	Wheat
Wheat	Wheat
Wheet	Wheat
Whhet	Wheat

5. PREDICTIVE MODEL

In Section 3 and Section 4, we have presented the two main ingredients for learning the predictive model: the outcome variable and the predictors. In this section, we discuss the final ingredient: the machine learning algorithm and give empirical performance results.

5.1 LASSO-Regularized Logistic Regression

As mentioned in Section 2, we would like to learn a binary classifier with probabilistic outputs in order to allow for different operating points corresponding to different classifier thresholds and levels of risk for the credit team to choose among. Also, we would like to have some level of human interpretability for the learned model to give intuition and understanding to Simpa about their portfolio.

Table 3. Class frequencies in the training and test sets.

	N	Proportion Good	Proportion Bad
Training Set	2456	0.87	0.13
Test Set	1903	0.82	0.18

Decision trees and decision lists, although very interpretable, do not have probabilistic outputs. Ensemble methods and neural networks are not very interpretable. Logistic regression and support vector machines (with linear kernel) can produce scores between zero and one, and their coefficients can be meaningfully examined by people. Therefore such methods are appropriate for the problem at hand and either would suffice. Both algorithms tend to produce models with similar generalization performance; we go with logistic regression.

In addition, to achieve the parsimony required for simplifying the DCAF form and improving user experience, we would like to learn a sparse model containing few features with non-zero coefficients. To achieve this, we regularize the logistic regression with the LASSO penalty. The penalty also serves to prevent overfitting.

5.2 Training and Test Sets

Ultimately in operation, the model will be trained on all available labeled data, but for the purpose of evaluating the performance of the classifier, we temporally divide our data into two sets. The first set contains customers activated between April 1, 2011 and July 25, 2014, which we use as the training set. The second set contains customers activated between July 26, 2014 and November 21, 2014, which we use as the test set. Recall that it takes 6 months after activation for a customer to acquire a good or bad label, so the remaining customers in our data set, which was pulled on May 21, 2015, cannot be used for evaluation purposes. The sample sizes and class proportions for the training and test sets are given in Table 1. The test set contains a bit more bad customers.

5.3 Learned Models

We apply the LASSO-regularized logistic regression algorithm on the training set with two different sets of the features. First, we model the raw attributes and derived features strictly from DCAF, and second, we additionally include the exogenous variables like branch and day rate discussed earlier. We sweep over different values of the regularization parameter and rank the importance of the variables by the order in which they enter into learned models as the regularization parameter decreases. We use the so-called 'one standard error rule' with the area under the curve (AUC) performance metric to determine the regularization parameter value for prediction. The ranked features for the DCAF model are given in Table 2 and for the DCAF with exogenous variables in Table 3. The last column (Imp.) indicates the impact of a feature, i.e. whether a positive change in that feature increases or decreases the probability of a customer being labeled as good.

We note that the top DCAF features and their ranks in both models are very similar to each other, indicating a level of stability in the learning. For example, we can see that males are less likely to be good customers than females in both models and that a person whose nature of business is skilled labor or driver is also less likely to be a good customer. Many of the features included in the models

Table 4. Top features in DCAF-only logistic regression model ranked by model entry with different regularization parameters and an indication of direction of impact. An asterisk indicates features included in the final one standard error rule predictive model. Other lower-ranked features are also included in the model, but are omitted here due to space.

Rank	Feature	Cat.	Imp.
1	*gender = male	demo.	–
2	*nature of business = skilled labour/driver	bus.	–
2	*(value of inventory)/(family size)	bus.	–
4	*arable land	agri.	–
5	*nature of business = grains/fruits	bus.	–
5	*power cut	demo.	+
5	*spoken languages	demo.	+
5	*understand languages	demo.	+
5	*written languages	demo.	+
10	*commodity 2 = potato	agri.	–
11	agriculture = true	agri.	–
11	*price per quintal	agri.	–
11	*salaried company = milk	salaried	–
14	*number of cows ≥ 3	asset	+
14	*age	demo.	+
14	*commodity 2 = paddy	agri.	+
14	*battery expense	expense	–
14	*distance to first recharge agent	demo.	–
14	*(loan repayment)/(total expense)	expense	–
20	*(candle expense)/(family size)	expense	–
20	*(other expense)/(family size)	expense	–
20	*nature of business = beauty	bus.	–
20	*(loan repayment)/(total expense)	expense	+
24	*nature of business = tent house and band	bus.	–
24	*price per quintal	agri.	–
24	*(total monthly expense)/(family size)	expense	–

are demographic, related to the profession, or related to expenses. The model with exogenous variables is more compact, having only 12 features in the final predictive model, and as we shall see in the next section also more accurate on the test set. It seems that many DCAF features are needed to capture the information in the exogenous variables, and that too imperfectly. Only this small set of features can be elicited in an updated DCAF mobile app.

Table 5. Logistic regression model with both DCAF features and exogenous features ranked by model entry with different regularization parameters and an indication of direction of impact. An asterisk indicates features included in the final one standard error rule predictive model.

Rank	Feature	Cat.	Imp.
1	*day rate	exog.	–
2	*gender = male	demo.	–
3	*branch = Mathura	exog.	–
4	*arable land	agri.	–
4	*branch = Bareilly-1	exog.	+
4	*down payment = 2500	exog.	–
4	*nature of business = skilled labour/driver	bus.	–
4	*spoken languages	demo.	+
4	*(value of inventory)/(family size)	bus.	–
10	*understand languages	demo.	+
10	*written languages	demo.	+
12	*nature of business = grains/fruits	bus.	–
13	agriculture = true	agri.	–
13	commodity 2 = potato	agri.	–
13	(candle expense)/(family size)	expense	–
13	price per quintal	agri.	–
13	salaried company = milk	salaried	–
18	age	demo.	+
19	battery expense	expense	–
19	distance to first recharge agent	demo.	–
19	monthly labour income	labour	+
19	address type = owned	demo.	+
23	number of cows ≥ 3	asset	+
23	family size	demo.	+
23	nature of business = beauty	bus.	–
23	(loan repayment)/(total expense)	expense	+

5.4 Performance Results

In this section, we present the classification accuracies of the models given in Section 5.3. We plot the receiver operating characteristic (ROC) and use the area under the curve (AUC) of the ROC as the performance metric of interest because of the different operating points envisioned in the application of interest. We examine the accuracy through tenfold cross-validation within the training set to understand the performance in an in-sample setting and by applying the models to the test set to understand the performance in the typical use case for the model: future prediction.

In addition to the two logistic regression models, we also examine classification accuracy of two benchmark models. The first is a basic scalar score variable that Simpa currently uses in its operations as discussed earlier, and the second is this score plus the exogenous variables. In each benchmark model, both the score and its square are included in a simple logistic regression. It is worth emphasizing, however, that these benchmarks will at best only proxy the process Simpa uses to assess new customers. In practice, the Simpa credit team uses a combination of the score variable and other criteria.

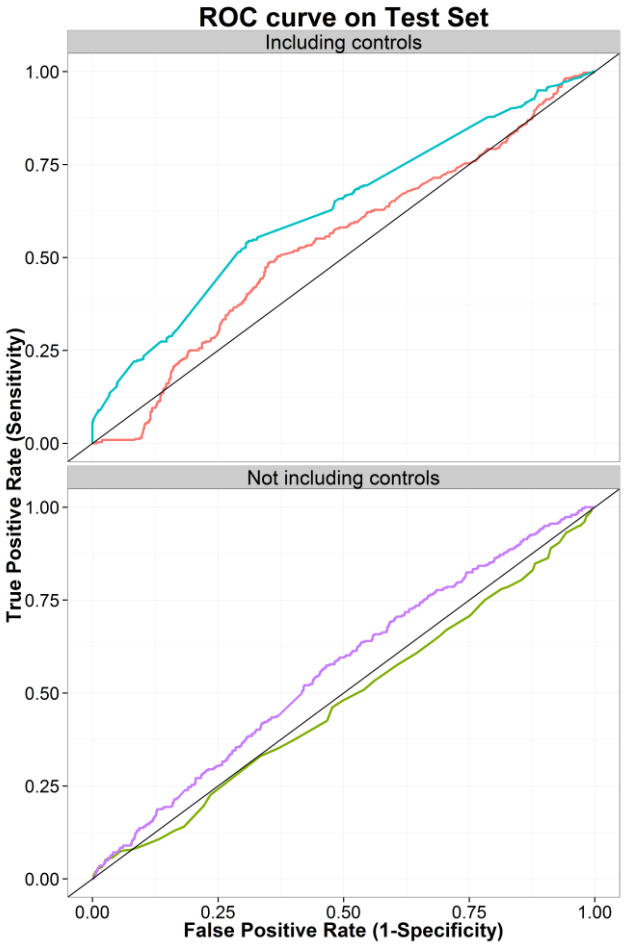

Figure 5. Receiver operating characteristic on the test set. Top panel is with exogenous variables and bottom panel is without exogenous variables. The blue and purple lines are the DCAF-based logistic regression. The red and green lines are the benchmarks.

Table 6. Area under the curve results.

Feature Set	Cross-Validation	Future Test
Benchmark	0.546	0.477
Benchmark + exogenous variables	0.642	0.537
DCAF	0.722	0.563
DCAF + exogenous variables	0.699	0.628

The AUC values for the different models in the cross-validation and out-of-sample testing settings are given in Table 6. We can see that the future prediction performance is worse than the in-sample cross-validation performance, and that the performance generally increases with a greater number of variables. On this test set, the future prediction performance of the score variable benchmark by

itself is actually worse than random guessing. The best performance on the test set is achieved with the DCAF features plus exogenous variables. Examining the ROCs for the test set in Fig. 5, we see that this model is best for nearly all operating points.

The future test performance is worse than the cross-validation performance because of non-stationarity in the customer population distribution. Machine learning methods assume identically distributed data in the training and test sets, but this is not true in our data. Since Simpa, as a startup company, is expanding its customer base rapidly this year (corresponding to the test set), it is entering new markets and taking applications from people with slightly different characteristics than before (the population in the training set). This makes our study an interesting application of forecasting in a real-world setting where a number of assumptions often made in machine learning do not hold.

Another way to interpret the results is shown in Fig. 6, which plots the proportion of bad customers as a function of the classifier threshold using predictions from the DCAF plus exogenous variables model. By using this model to accept those customers most likely to be good, Simpa could have reduced the rate of bad customers accepted by almost one third (18% to 12.5%) while still accepting around 70% of customers. Of course, whether implementing this strategy is sensible also depends on other business factors.

The overall accuracy we see is not as high as is achieved in other classification problems related to credit scoring [6] or in other application domains. We hypothesize this could be for a number of reasons. First, our data set only contains customers that were approved and does not contain applicants rejected by the credit team. The bad customers in our data set were accepted and had systems installed, but then were delinquent in payments. The fact that we were not able to achieve a very high rate of accuracy suggests that Simpa's existing process was doing a fairly good job of vetting customers. Second, future repayment behavior is hard to predict without features on past repayment behavior. As discussed previously, individuals in developing countries often do not have the same financial history available when building credit scores as in developed country applications. Additionally, there are external factors that affect repayment behavior that cannot be known at the time of application, such as if the photovoltaic system needs constant repair. Third, the population characteristics are changing over time, as previously discussed. Fourth, it could be that DCAF data is simply not that predictive of future repayment behavior, and other data sources (discussed more in Section 6) could be more informative. Nevertheless, we think the 0.628 AUC on the test set using DCAF features and exogenous variables is of sufficient quality to have value for Simpa's operations going forward, certainly as a supplementary model to the currently used benchmark score, which we saw has no predictive power for the class label in our test set.

6. CONCLUSION

We have analyzed Simpa Networks' DCAF and RMS data to develop a classifier that predicts whether a potential customer of a solar photovoltaic system is likely or unlikely to keep up with payments. We have performed a good amount of cleaning and preparation on the DCAF data to transform it into a state amenable to statistical modeling and have defined an indicator of good and bad customers from the records in RMS data using the number of recharges in the first 6 months after activation.

We have used LASSO-regularized logistic regression and obtained a sparse model containing only 12 features that generalizes to new

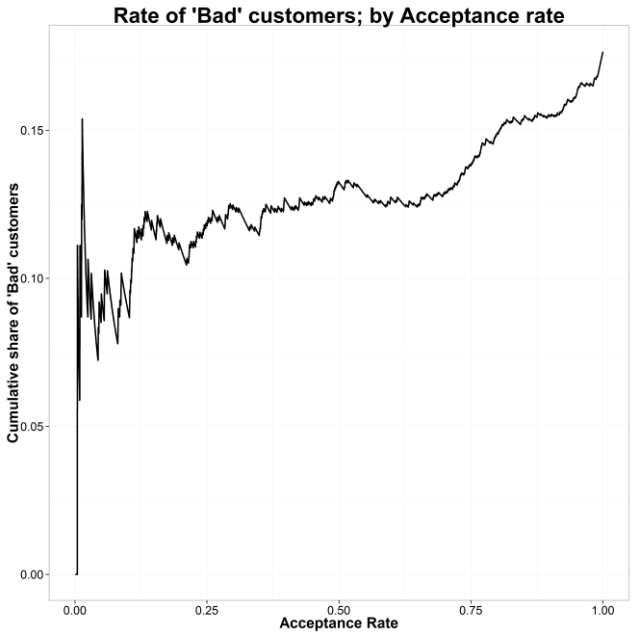

Figure 6. Proportion of bad customers as a function of the classifier threshold.

unseen test samples in the future. The classification problem is difficult because the signal to be estimated is weak and the classification performance on future test samples is worse than on in-sample cross-validation testing due to changes in the customer characteristics over time. However, the accuracy is sufficiently high that we plan on integrating the predictive modeling into Simpa's operations soon.

In conducting this project, we have noted several recommendations and discovered several directions for future work. Simpa can consider prioritizing variables listed in Table 4 and Table 5 as they are the primary drivers of the model. These variables would benefit from accurate capture in the future; therefore a focus for training Simpa urja mitras on their level of importance in effectively capturing would be helpful. In addition to this, these variables can be prioritized in an updated mobile app for data capture as well.

Given the output of the model, if Simpa wishes to keep a similar risk profile (current data shows approximately 18% of customers being bad), they should consider keeping the current threshold to maintain that level of risk. The business may benefit from this guideline being relaxed in certain situations, such as when rapid acquisition (and a resulting network effect of word of mouth) would be of higher importance. Situations such as this could arise in expansions to entirely new regions or otherwise sparse areas.

Through the data cleansing process, several variables could have been important to the model, but had too much missing data or were too non-standardized to be effectively leveraged, for example the reason for no electricity and the benefit. These variables' data capture should be refined going forward as they may have future benefit on future iterations of the model. Moreover, the DCAF form can be updated to have discrete choices in the fields we recoded from free text in Section 4. In addition to this, there are several contextual elements which aren't included within the data such as mobile phone data [7], as well as previous loan or credit information, which could be useful to the model going forward. Partnerships with other organizations may be needed to obtain such information.

Retraining the current model is imperative to capture shifting customer behaviors. Approximately every month, or as frequently as it makes sense logistically, a new test/train split should be built and the model should be retrained based on this new information.

Several other machine learning problems and certain enhancements to the algorithms can be considered in the future. In the third business use for the predictions mentioned in Section 2, understanding the portfolio better, a more useful prediction than the binary classification problem is to predict the total revenue to be collected from the customer over the life of the relationship. Such a problem can be approached in the future when data on more customers with longer histories is available. Another problem that may be of interest is to predict the final state some months after the customer's activation, which would enable us to use those months of repayment information as features. A possible enhancement to the current model to account for the changing population distribution over time is to use importance sampling and covariate shift [8].

Outside of data capture and predictive analytics, another data science project that can be considered in future work is the following. Simpa tries to increase repayment via reminder phone calls and text messages, but has not yet studied the causal effects of these interventions. It would be quite interesting to conduct an A/B test on the interventions.

7. ACKNOWLEDGMENTS

The authors thank Sharat Thakur, Shyam Patro, Navneet Kumar, and Piyush Mathur of Simpa Networks for discussions and support, and Craig Barowsky, Peter Darche, Shubha Bala, and Jake Porway of DataKind for the same.

8. REFERENCES

[1] Alstone, P., Gershenson, D., and Kammen, D. M. 2015. Decentralized energy systems for clean electricity access. *Nature Climate Change* 5 (Apr. 2015), 305-314.

[2] Nique, M. and Smertnik, H. 2015. The synergies between mobile phone access and off grid energy solutions. In *Decentralized Solutions for Developing Economies: Addressing Energy Poverty Through Innovation*. Springer, Cham, Switzerland.

[3] Ladd, T. 2015. Affordability through business models for distributed energy at the base of the pyramid. *SAIS Eur. J. Glob. Aff.* (Mar. 2015).

[4] Olla, P. and Onwudinjo, N. 2014. Productive use of renewable energy (PURE) for economic development in developing countries. In *Sustainable PracticesL Concepts, Methodologies, Tools, and Applications*. Information Science Reference, Hershey, PA.

[5] Hand, D. J. and Henley W. E. 1997. Statistical classification methods in consumer credit scoring: A review. *J. Royal Stat. Soc. Ser. A* 160:3 (Sep. 1997), 523-541.

[6] Lessmann, S., Seow, H.-V., Baesens, B., and Thomas, L. C. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *Eur. J. Oper. Res.* 247:1 (Nov. 2015), 124-136.

[7] Kumar, K. and Muhota, K. 2012. Can digital footprints lead to greater financial inclusion? World Bank, Washington, DC.

[8] Wei, D., Ramamurthy, K. N., and Varshney, K. R. 2015. Health insurance market risk assessment: Covariate shift and k-anonymity. In *Proc. SIAM Int. Conf. Data Min.*, Vancouver, Canada (Apr.–May 2015), 226–234.