Program Targeting with Machine Learning and Mobile Phone Data: Evidence from an Anti-Poverty Intervention in Afghanistan [*]

Emily L. Aiken[†] Guadalupe Bedoya[‡] Joshua E. Blumenstock[†] Aidan Coville[‡]

This version: September 14, 2021
First version: April 10, 2020

Abstract

Can mobile phone data improve program targeting? By combining rich survey data from a "big push" anti-poverty program in Afghanistan with detailed mobile phone logs from program beneficiaries, we study the extent to which machine learning methods can accurately differentiate ultra-poor households eligible for program benefits from ineligible households. We show that supervised learning methods leveraging mobile phone data can identify ultra-poor households nearly as accurately as survey-based measures of consumption and wealth; and that combining survey-based measures with mobile phone data produces classifications more accurate than those based on a single data source.

Keywords: Targeting; Machine Learning; Mobile Phone Data; Afghanistan
JEL: I32, I38, O12, O38, C55

[*] We thank Seungmin Lee, Maria Camila Ayala and Thomas Escande for excellent research assistance. This work was supported by DARPA and NIWC under contract N66001-15-C-4066, the NSF under grant IIS-1942702, and by the World Bank's Knowledge for Change Program. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense, the U.S. Government, the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

[†] School of Information, University of California, Berkeley

[‡] Development Impact Evaluation Department, World Bank

1 Introduction

Each year, hundreds of billions of dollars are spent on targeted social protection programs. The importance of these programs increased dramatically in the past year: In 2020, global extreme poverty increased for the first time in two decades, and most countries expanded their social protection programs, with more than 1.1 billion new recipients receiving government-led social assistance payments (Gentilini et al., 2020).

Determining who should be eligible for program benefits — *targeting* — is a central challenge in the design of these programs (Hanna & Olken, 2018; Lindert et al., 2020). In high-income countries, targeting frequently relies on tax records or other administrative data on income. In low- and middle-income countries (LMICs), where a large fraction of the workforce is informal, programs often require primary data collection. The difficulty and cost of collecting data, and the variable quality of what gets collected, can introduce significant errors in the targeting process (Deaton, 2016; Jerven, 2013; Grosh et al., in press). These issues are exacerbated in fragile and conflict-affected countries, where two thirds of the world's poor are expected to reside by 2030 (Corral et al., 2020).

This paper evaluates the extent to which non-traditional administrative data, processed with machine learning, can be used for program targeting. Specifically, we match call detail records (CDR) from a large mobile phone operator in Afghanistan to household survey data from an impact evaluation of the Afghan government's Targeting the Ultra-Poor (TUP) anti-poverty program. Eligibility for the TUP program was determined through a combination of a community wealth ranking (CWR) and a short follow-up survey (we refer to this combination as the *hybrid targeting method*). We then assess the accuracy of three counterfactual targeting approaches at identifying the actual beneficiaries of the TUP program: (i) our *CDR-based method*, which applies machine learning to data from the mobile phone company; (ii) an *asset-based wealth index*, which uses asset ownership to approximate poverty in a spirit similar to a proxy-means test (PMT); and (iii) *consumption*, a common benchmark for measuring poverty in LMICs.

Our analysis produces three main results. First, by comparing errors of inclusion and exclusion using the program's hybrid method as a benchmark, we find that the CDR-based method is nearly as accurate as the asset and consumption-based methods for identifying the phone-owning ultra-poor households. Second, we find that methods combining CDR data with measures of assets and consumption are more accurate than methods using any single data source to identify the ultra-poor. Third, we find that when non-phone-owning households are included in the analysis, the CDR-based method remains accurate if non-phone-owning households are classified as ultra-poor and therefore program-eligible; however, targeting performance is quite poor if households without phones are ineligible for benefits.

These results connect two distinct strands of prior work. The first is a rich literature on program targeting, which studies the effectiveness of different mechanisms for identifying program beneficiaries. In LMICs, research has focused in particular on the performance of proxy means tests (PMTs), (Grosh & Baker, 1995; Filmer & Pritchett, 2001; Brown et al., 2018), community-based targeting strategies (Alatas et al., 2012; Fortin et al., 2018), and related approaches (Banerjee et al., 2007; Karlan & Thuysbaert, 2019; Premand & Schnitzer, 2020). A meta-analysis by Coady et al. (2004), which includes 8 PMTs and 14 community-based programs, finds little difference in targeting accuracy between the two methods — but notes that targeting is regressive in a quarter of programs reviewed. In addition to issues with targeting accuracy, the current methods available for poverty targeting in LMICs are time- and resource-intensive, and may be infeasible in fragile or conflict-affected areas or in contexts when mobility and social interaction is limited, such as during a pandemic.

The second body of work explores the extent to which non-traditional sources of data, in conjunction with machine learning, might help address data gaps in LMICs (e.g. Blumenstock, 2016; Burke et al., 2021). Much of this work focuses on estimating the geographic distribution of wealth and poverty at fine spatial granularity, using data from satellites (Jean et al., 2016; Engstrom et al., 2017), mobile phones (Blumenstock et al., 2015; Hernandez et al., 2017), social media (Fatehkia et al., 2020; Sheehan et al., 2019), or some combination of these data sources (Steele et al., 2017; Pokhriyal & Jacques, 2017; Chi et al., 2020). Most relevant to our current analysis, two prior papers investigate whether the wealth of individual mobile subscribers can be accurately estimated using mobile phone data. Blumenstock et al. (2015) show that CDR data are predictive of an individual-level asset-based wealth index among a nationally representative sample of 856 Rwandan mobile phone owners (cross-validated $r = 0.68$). Blumenstock (2018b) finds similar results with a sample of 1,234 male heads of households in the Kabul and Parwan districts of Afghanistan.

Our paper connects these two distinct literatures by rigorously assessing the extent to which phone-based estimates of poverty can help with program targeting (Blumenstock, 2020; Aiken et al., 2021).[1] The context of our empirical analysis – identifying ultra-poor households in Afghanistan – is a particularly challenging environment for data collection and program targeting, as 62% of the households classified as not *ultra*-poor still fall below

[1] The anti-poverty program implemented and described by Aiken et al. (2021) in Togo was based on the methods developed and evaluated in this paper. Due to the more time-sensitive nature of the COVID-19 response described in Aiken et al. (2021), the two academic articles are in circulation concurrently.

the national poverty line. The fact that these methods show promise in this context suggests that they may be relevant to a broad class of targeting applications. We therefore conclude by discussing the important ethical and logistical considerations that may influence how CDR methods are used to support targeting efforts in practice.

2 Data and Methods

Our main analysis evaluates the extent to which machine learning and mobile phone data can accurately differentiate between ultra-poor and non-ultra-poor households in rural Afghanistan. This section describes the study population, the key datasets, and methods used to perform the evaluation.

2.1 Household Survey Data

The ground-truth data that we use to evaluate this new approach to program targeting were collected as part of the Targeting the Ultra-Poor (TUP) program implemented by the government of Afghanistan with support from the World Bank. The TUP program included an impact evaluation of a "big push" anti-poverty program that provided multifaceted benefits to ultra-poor households (Bedoya et al., 2019). Our analysis is centered on a baseline survey that was collected for the TUP program, which contains well-being measures for 2,852 households in 80 of the poorest villages in Afghanistan's Balkh province, surveyed prior to the TUP launch (between November 2015 - April 2016).[2] These data include surveys of nearly all of the 1,173 ultra-poor households in the villages deemed eligible for the program, and a random sample of 1,679 non-ultra-poor households.[3] Baseline surveys were conducted in two in-person interviews, one with the primary woman of each household, and one with the primary man.

Ultra-Poor Designation Eligibility for the TUP program was determined based on based on geographic criteria,[4] followed by a two-step process including a community wealth ranking (CWR) and a follow-up in-person survey. CWRs were conducted separately in each village,

[2] Our analysis restricts to 2,814 households for whom consumption and all asset data are non-missing.

[3] The response rate for ultra-poor households was 96%. Approximately 20 households in each of the study villages were randomly drawn (excluding TUP-eligible households), to provide a representative benchmark for the TUP sample.

[4] The poorest villages in the province were identified subject to having availability of veterinary services, financial institutions, and social services, and being relatively accessible.

coordinated by a local NGO and village leaders, in collaboration with the government's Microfinance Investment Support Facility for Afghanistan (MISFA). CWRs divided households into four categories: well-off (6%), better-off (18%), poor (33%), and extreme-poor (43%). The CWR was followed by an in-person survey to determine whether nominated households met a set of qualifying criteria, coordinated by the NGO and MISFA representatives, and based on a measure of multiple deprivation.

For a household to be designated as *ultra-poor*, and therefore eligible for program benefits, it had to be considered extreme-poor in the CWR, and also meet at least three of six criteria:[5]

1. Household is financially dependent on women's domestic work or begging.
2. Household owns less than 800 square meters of land or is living in a cave.
3. Targeted woman is younger than 50 years of age.
4. There are no active adult men income earners.
5. Children of school age are working for pay.
6. Household does not own any productive assets.

Ultimately, 11% of the households classified as extreme-poor in the community wealth ranking step — 6% of the total population in the study villages — were classified as ultra-poor and eligible for TUP benefits. Of the 2,852 households surveyed for the TUP project, 1,173 (41%) were designated as ultra-poor, and 1,679 (59%) were non-ultra-poor.

Consumption The consumption module of the TUP survey contains information on food consumption for the week prior to the interview and non-food expenditures for the year prior to the interview. These are used to construct monthly *per capita* consumption values, as detailed in Bedoya et al. (2019). While consumption data are reported for the household as a whole, the survey questions were asked of the primary woman of the household. Based on these data, we construct as an outcome measure the logarithm of *per capita* monthly consumption, consistent with the approach used by the Afghanistan government to determine the national poverty line.

[5] While these were the official criteria used to guide selection, they were not always strictly enforced — see Bedoya et al. (2019).

Asset Index We construct an asset-based wealth index to assess the relative socioeconomic status of surveyed households. The asset questions, which describe the household as a whole, were asked of the primary woman of the household. The asset index is calculated as the first principal component of variation in household asset ownership for sixteen items detailed in Table S1. The principle component analysis (PCA) is calculated over the dataset of 2,814 households not missing any asset data, after standardizing each asset variable to zero mean and unit variance. This wealth index explains 25.3% of the variation in asset ownership. Figure S1 shows the distribution of the underlying asset index components and Table S1 shows the direction of the first principal component.

Other Variables The TUP surveys collected several other covariates that we use in subsequent analysis. These include a food security index (composed of variables relating to the skipping and downsizing of meals, separately for adults and children), a financial inclusion index (composed of access to banking and credit, knowledge of banking and credit, and savings), and a psychological well-being index for the primary woman (standardized weighted average scores on the 7-item Center for Epidemiological Studies Depression scale, the World Values Survey happiness and satisfaction questions, and Cohen's 4-item stress scale). The construction of each index is documented in Bedoya et al. (2019). Crucially, the survey also collected data from each household on mobile phone ownership. Nearly all (99%) households with a cell phone provided their phone numbers and consented to the use of their call detail records for this study.

Sample Representativity Portions of our analysis are restricted to the 535 households from the TUP survey with phone numbers that match to our CDR (see Section 2.2). While the 2,852 households in the TUP survey are representative of the 80 study villages, they are not nationally representative of Afghanistan as a whole, and the 535-household subsample is not representative of the overall sample from the TUP survey. Table 1 compares characteristics of households included and excluded from the 535-household subsample; Figure S2 compares the distributions of these characteristics. There are some systematic differences: the 535-household sample we analyze is richer on average than households surveyed in the TUP study, which is consistent with households in the subsample being required to own at least one phone. For instance, while 88% of non-ultra-poor households in the TUP survey own at least one phone, only 72% of ultra-poor households own at least one phone.

Summary Statistics As shown in Table 1 and Figure S3, the three measures of well-being in our dataset are only weakly correlated with one another: for example, the correlation between the asset index and consumption measure is 0.37 in the full survey and 0.34 in the matched subsample. It is particularly important to note the characteristics of the ultra-poor: while the ultra-poor population makes up 27% of the overall sub-sample, less than half of the ultra-poor fall into the bottom 27% of the sample by wealth index or consumption.

Sample Weights Since the TUP survey oversampled the ultra-poor (by a factor of roughly 12), portions of our analysis use sample weights to adjust for population representativity. When sample weights are applied, it is explicitly noted; if not mentioned, no weights are applied. The sample weights are derived from the population of the village, and the household's ultra-poor designation.[6] After sample weights are applied, the ultra-poor make up 5.98% of the overall population, and 4.63% of our matched subsample.

2.2 Mobile Phone Metadata

In a follow-up survey conducted in 2018, we requested informed consent from survey respondents to obtain their mobile phone CDR and match them to the survey data collected through the TUP project. CDR contain detailed information on:

- **Calls:** Phone numbers for the caller and receiver, time and duration of the call, and cell tower through which the call was placed
- **Text messages:** Phone numbers for the caller and recipient, time of the message
- **Recharges:** Time and amount of the recharge

For participants who consented, we match baseline survey data (collected November 2015 - April 2016) to CDR covering that same period, obtained from one of Afghanistan's main mobile phone operators. For households with multiple phones and a designated household head (N=65), we match to CDR for the phone belonging to the household head. For households where the household head does not have a phone and someone else does (N=17), we match to CDR for one of the households' phones selected at random. In total, for the 535 households in our sample, 629,543 transactions took place in the months of November 2015 to April 2016, broken down into 310,883 calls, 305,756 text messages, and 12,904 recharges.

[6] A census listing was conducted in each village to facilitate the CWR exercise. In some cases, large villages were split into smaller units and weights are based on the sub-village.

From these CDR, we compute a set of 797 behavioral indicators that capture aggregate aspects of each individual's mobile phone use (de Montjoye et al., 2016). This set includes indicators relating to an individual's communications (for example, average call duration and percent initiated conversations), their network of contacts (for example, the entropy of their contacts and the balance of interactions per contact), their spatial patterns based on cell tower locations (for example, the number of unique antennas visited and the radius of gyration), and their recharge patterns (including the average amount recharged and the time between recharges). Each indicator is computed separately for weekdays, weekends, daytime, and nighttime activity. The distributions of a sample of these indicators are shown in Figure S4.

2.3 Machine Learning Predictions

CDR-based Method Extending the approach described in Blumenstock et al. (2015), we test the extent to which ultra-poor status can be accurately predicted from CDR. This analysis uses the 535 TUP households who match to CDR to train a supervised machine learning algorithm to predict ultra-poverty status from the mobile phone features. The intuition — also highlighted in Figure S4 — is that ultra-poor individuals use their phones very differently than non-ultra-poor individuals, and machine learning algorithms can use those differences to predict ultra-poor status.

Our main analysis uses a forest of gradient boosted decision trees (hereafter referred to as the "gradient boosting model"), which generally out-performs several other common machine learning algorithms for this task (including a standard logistic regression, a regularized logistic regression with L1 penalty, and a random forest). The feature importances for the trained model are shown in Table S2. For comparison, results using other machine learning algorithms are provided in Table S3.

Probabilistic predictions are generated via 10-fold cross-validation with each model, with folds stratified to preserve class balance. We tune hyperparameters using five-fold cross-validation for each prediction fold separately, optimized over a wide grid of hyperparameters for each model. For the linear models and random forest, features are standardized to zero mean and unit variance and missing values are mean-imputed, separately for each prediction fold. Additional details on the machine learning methods are provided in Appendix A.

Combined Methods We also evaluate several approaches that use data from multiple sources to predict ultra-poor status. Our main *combined method* trains a logistic regression to classify the ultra-poor and non-ultra-poor households using the predicted probability from the CDR-based method (i.e., the output of the gradient boosting algorithm described above), as well as asset and consumption data collected in the TUP survey. For comparison, we similarly evaluate the performance of methods that combine only two of the available data sources (i.e., assets plus consumption, assets plus CDR, and consumption plus CDR). Predictions for each of the combined methods are pooled over 10-fold cross-validation.

2.4 Evaluation

Evaluation on Matched Subsample Our main analysis focuses on the 535 households for which we observe both CDR and survey data, and evaluates whether machine learning methods leveraging CDR data can accurately identify households designated as ultra-poor by the TUP program (using the two-step hybrid approach described in Section 2.1). We compare the performance of the CDR-based method to the performance of methods based on the wealth index, consumption data, and combinations of these different data sources.[7] Each targeting method is evaluated based on classification accuracy, errors of exclusion (ultra-poor households misclassified as non-ultra-poor) and errors of inclusion (non-ultra-poor households misclassified as ultra-poor). We focus on the ultra-poor designation as the 'ground truth' status of the household, against which other methods are evaluated, since it is the most carefully vetted measure of well-being for this population, and the proxy that the government decided to use in targeting TUP benefits.

To evaluate the performance of the CDR-based and combined methods, we pool out-of-sample predictions across the ten cross-validation folds, so that every household in our dataset is associated with a CDR-based predicted probability of ultra-poor status that is produced out-of-sample. To account for class imbalance, we evaluate model accuracy using a "quota method", by selecting a cut-off threshold for ultra-poor qualification (a maximum wealth index, maximum consumption, and minimum CDR-based predicted probability of being ultra-poor) such that each method identifies the proportion of ultra-poor households in our subsample; this cut-off also balances inclusion and exclusion errors. In our 535-household matched dataset this threshold is 27%, as 27% of households are ultra-poor; in

[7] An important difference between the CDR-based method and the asset- and consumption-based approaches is that the CDR-based method uses machine learning to model the ultra-poverty outcome, whereas the asset- and consumption-based approaches do not. We therefore separately test whether performance improves when machine learning methods are applied to the original survey data to model the ultra-poverty outcome. In results shown in Table S4, we find that a machine-learned asset predictor provide only marginal improvements (AUC=0.73) on the standard asset-based wealth index and consumption measures (AUC=0.73 and AUC=0.71, respectively) – see Panel A of Table 2.

other samples (see following subsection), the percentage is different. We evaluate each model at this threshold for precision (positive predictive value) and recall (sensitivity). To capture the trade-off between inclusion and exclusion errors for varying values of this threshold, we also construct receiver operating characteristic (ROC) curves for each method and consider the area under the curve (AUC) as a measure of targeting quality. For each evaluation metric (precision, recall, and AUC), we bootstrap 1,000 samples from the original dataset to calculate the standard deviation of the mean of the accuracy metric. Each bootstrapped sample is of the same size as the original dataset, drawn with replacement.

Accounting for Households Without Phones Our main results assess the performance of different targeting methods on the sample of 535 households for whom we have both survey data and mobile phone data. We also present results that show how performance is affected when the analysis includes TUP households for whom we do not have mobile phone data (typically because they do not have a phone or because they use a different phone network than the one who provided CDR). For such households, it is straightforward to assess the performance of asset-based and consumption-based targeting. To evaluate households without CDR, we assume the CDR-based targeting would target such households (1) before households with CDR, or (2) after households with CDR. More details on this procedure are provided in Section 3.4.

We present results based on three different samples:

1. *Matched Sample:* The 535 households for whom we were able to match survey responses to CDR.

2. *Balanced Sample:* This sample includes the 535 matched households as well as the 472 households in the TUP survey who report not owning any phone. It excludes households that own a phone on a different phone network than the one who provided CDR. The motivation for this sample is to provide an indication of targeting performance in a regime in which CDR can be used to target all phone-owning households. In addition to applying sample weights from the survey, households that do not own a phone are downweighted so that the balance of phone owners to non-phone-owners (with sample weights applied) is the same as in the baseline survey as a whole (with sample weights applied, 84% phone owners).

3. *Full Sample:* All 2,814 households in the TUP baseline survey for which asset and consumption data are available, with sample weights applied.

Note that the quota used to evaluate targeting changes for each sample, based on the number of households that are ultra-poor in the sample. For the matched sample, the targeting quota is 27.29%; for the balanced sample and full sample the quotas are 5.47% and 6.02%, respectively.

3 Results

3.1 Performance of Targeting Methods

Our first set of results evaluate the extent to which different targeting methods can correctly identify ultra-poor households. This analysis compares the performance of CDR-based targeting methods to asset-based and consumption-based targeting, using the sample of 535 households for which survey data and CDR data are both available.

An overview of these results is provided in Figure 1. Figure 1a, shows the distribution of assets and consumption, as well as the distribution of predicted probabilities of being non-ultra-poor generated by the CDR-based and combined methods, separately for the ultra-poor (pink) and non-ultra-poor (blue). The dashed vertical line indicates the threshold at which point 27% of households are classified as ultra-poor; we use this quota because 27% of households in this sample were designed as ultra-poor by TUP. Figure 1b provides confusion matrices that compare the true status (rows) against the classification made by each method (columns). These confusion matrices are also used to calculate the measures of precision and recall reported in Table 2 Panel A.

We find that the CDR-based method (precision and recall of 42%) is close in accuracy to methods relying on assets (precision and recall of 49%) or consumption (precision and recall of 45%). To evaluate the trade-off between inclusion errors and exclusion errors resulting from selecting alternative cut-off thresholds, Figure 1c shows the ROC curve associated with each classification method, and the associated Area Under the Curve (AUC). AUC scores are comparable among methods, with assets (AUC=0.73) slightly superior to consumption (AUC=0.71) and the CDR-based method (AUC=0.68).

3.2 Comparison of Errors Across Methods

To better understand where the targeting methods are making mistakes, Panel A of Table 3 indicates how the ultra-poor misclassified as non-ultra-poor (errors of exclusion, or false negatives) compare to the correctly classified ultra-poor (true positives). Panel B shows how the non-ultra-poor misclassified as ultra-poor (errors of inclusion, or false positives) compare to the correctly classified non-ultra-poor (true negatives).

We find that, broadly speaking, the classification errors made by all three methods tend to be sensible: when these methods make mistakes, they are generally not egregious. Across methods, false negatives tend to score higher on food security, financial inclusion, and psychological well-being than true positives – that is, all three targeting methods misclassify ultra-poor households as non-ultra-poor when those ultra-poor households are better-off, according to other observable characteristics not used in the targeting per se. Likewise, false positives (non-ultra-poor misclassified as ultra-poor) tend to score lower than true negatives across these same measures. The CDR-based method in particular tends to prioritize households that score low on these alternative measures of well-being.

To test for systematic misclassification of certain types of households, Table 4 displays the overlap in errors of exclusion and inclusion between methods. Our results suggest that the three classifiers misidentify the same households at a rate only slightly above random.[8]

3.3 Combining Targeting Methods

Since the different targeting methods are identifying different populations as ultra-poor, there may be complementarities between asset, consumption, and CDR data. We therefore test a set of methods that integrate multiple data sources into a single classification. As shown in Panel A of Table 2, we find that this *combined method*, which takes as input the wealth index, total consumption, and the output of the CDR-based method, performs better (AUC = 0.78) than methods using any one data source (AUC = 0.68 - 0.73). As shown in Table S5), the full method also outperforms methods based on any two data sources (AUC = 0.75 - 0.76). However, it is worth noting the strong performance of a method that combines CDR data and the asset index (AUC = 0.76); this two-component method may be more practical than the combined method, since consumption data can be difficult to collect for large populations.

[8] The rates of overlap should be interpreted relative to the expected overlap in errors for random classifiers with the same cut-off threshold for ultra-poor classification. Based on our selection of thresholds such that 27% of the sample is identified as ultra-poor, our three classifiers misidentify 15-27% of the non-ultra-poor and 51-65% of the ultra-poor. If these classifiers were random, we would expect approximately 20% overlap in errors of inclusion and 55% overlap in errors of inclusion.

3.4 Targeting Households Without Phones

An important limitation with CDR-based targeting is that households without phones do not generate CDR. This is a conceptual issue that we revisit in Section 4; for now, we present results that show how predictive performance is impacted by the inclusion of these households in the analysis.

This analysis uses two additional samples of TUP households to evaluate targeting performance: (i) the *balanced sample*, which adds all of the 472 households without phones to the sample of 535 for whom we have matched CDR; the balanced sample is intended to illustrate the performance of CDR-based targeting if CDR were available from all operators in Afghanistan — though it relies on the assumption that phone-owners observed on our mobile network are representative of all phone owners in Afghanistan (an assumption that is not fully satisfied, as shown in Table 1); and (ii) the *full sample*, which includes all 2,814 households surveyed in the TUP baseline; this sample includes an additional 1,807 households who report owning a phone, but whose number does not match to any number in the CDR provided to us by the single mobile operator.[9]

Results in Panels B and C of Table 2 show the performance of each targeting approach on the balanced and full sample, respectively. Note that as described in Section 2.4, different targeting quotas are applied for each panel based on the proportion of each sample that is ultra-poor. In the CDR-based and combined approaches, we report performance when the households without CDR are targeted first (i.e. households without CDR are targeted in a random order and then the households predicted to be poorest are targeted until the quota is reached) as well as when households without CDR are targeted last (i.e., after the 535 households with phones are targeted, households without phones are included in a random order until the quota is reached).

Unsurprisingly, these results suggest that CDR-based targeting is not particularly effective when a large portion of the target population does not own a phone. This is particularly true in Panel C of Table 2, where only 16% of the sample (with sample weights applied) has matching CDR. However, when we simulate more realistic levels of phone ownership in Panel B (84% of the households, based on our survey data), we note that CDR-based targeting is once again comparable to asset- or expenditure-based targeting, particularly when households without phones are targeted first (AUC = 0.72, 0.70, 0.68 for assets, consumption, and CDR, respectively). On the other hand, if the CDR-based method is used and

[9] These 1,807 households include households that report owning a phone on a different network (this network is estimated to have around 30% market share in Afghanistan), as well as phones on our network that were not active during the six-month period of CDR that we analyze.

households without phones are targeted last (for example, if program administrators base targeting wholly on CDR and provide no benefits to any household without a phone), the CDR-based method only improves marginally on random targeting.

3.5 Additional tests and simulations

Our main analysis considers the household head to be the unit of analysis. As described in Section 2.2, this analysis is based on matching survey-based measures of well-being to phone data from the household head – to the best of our ability. This approach is most consistent with the design of the TUP program and the TUP sample frame. An alternative approach that we explore matches survey data reported by the household head to all phone numbers associated with the household. As shown in Table S6, the predictive accuracy of these models is slightly attenuated relative to the benchmark results (Table S3), particularly for the more flexible machine learning models.

We also explore the extent to which CDR can be used to predict other measures of socioeconomic status. The preceding analysis focuses on the household's TUP's ultra-poor designation as the ground truth measure of poverty, since this was a carefully curated label and the actual criteria used to determine TUP eligibility. In Table S7, we report the accuracy with which CDR (obtained from the household head, who is typically male) can predict consumption and asset-based wealth (elicited from the primary woman of each household).[10] In general, these machine learning models trained to directly predict consumption or asset-based wealth from CDR do not perform well. This contrasts with prior work documenting the predictive ability of CDR for measuring asset-based wealth (e.g. Blumenstock et al., 2015). We suspect a key difference in our setting – aside from the fact that we are matching CDR to socioeconomic status at the *household* rather than the *individual* level – is the homogeneity of the beneficiary population: whereas Blumenstock et al. (2015) uses machine learning to predict the wealth of a nationally-representative sample of Rwandan phone owners, our sample consists of 535 individuals from the poorest villages of a single province in Afghanistan, where even the relatively wealthy households are quite poor.

[10] Due to the design of the TUP survey, which interviewed women in the household, we cannot avoid this mismatch between the survey respondent and the phone owner.

4 Discussion

Our key finding is that, in a sample of 535 phone-owning households in a set of poor villages in one province of Afghanistan, machine learning methods leveraging behavioral indicators computed from CDR are nearly as accurate as standard asset- and consumption-based methods for identifying ultra-poor households. Further, we find that methods combining survey data with CDR perform better than any of the methods using a single data source. In contexts like Afghanistan where standard targeting benchmarks are unavailable or of questionable quality, methods that integrate CDR may create new options for program targeting.

However, as we demonstrate empirically, low rates of phone ownership — or the inability to access data from all operators — can quickly undermine the value of CDR-based targeting. While mobile phone penetration rates continue to rise in LMICs (GSMA, 2020), we expect that, for the forseeable future, CDR-based methods may be best deployed in conjunction with alternative approaches. In our specific setting, the CDR-based approach still works well if households without phones are targeted before the CDR-based algorithm then selects the poorest households with phones. However, this approach may not be appropriate in other contexts where phone ownership is less predictive of wealth, or where potential beneficiaries have the ability to strategically under-report phone ownership (Björkegren et al., 2020).

Our analysis also highlights several broader considerations that we believe are worth deeper investigation in future work. These include:

Tradeoffs in data privacy and predictive accuracy CDR contain sensitive and personally identifying information, including phone numbers, contact networks, and location traces (De Montjoye et al., 2013; Taylor, 2016). Informed consent can help ensure participant autonomy, but also creates significant logistical complications. Differential privacy and related methods can provide formal privacy guarantees on CDR and other data (Hu et al., 2015), but there is an inherent trade-off between privacy and data utility when such privacy guarantees are introduced.

Algorithmic transparency and strategic behavior Using CDR to determine program eligibility may introduce incentives for people to manipulate their phone use. This consideration is not unique to CDR, as varying degrees of manipulation have been documented in social programs that use proxy means tests and other traditional targeting mechanisms (Camacho & Conover, 2011; Banerjee et al., 2018). Indeed, complex and non-linear machine learning algorithms, like the one presented in this paper, may obfuscate the logic behind targeting decisions and thereby reduce the scope for manipulation. However, society often demands transparency in algorithmic decision-making, as black-box decisions are difficult to audit or hold to account. There is therefore a tension between the goals of increasing transparency and reducing manipulation, though recent advances in machine learning explore mechanisms for pursuing both objectives at once (Björkegren et al., 2020).

Centralized vs. local knowledge CDR-based methods enable a top-down, centralized and standardized approach to program targeting, rather than a bottom-up approach that prioritizes local knowledge that can be elicited, for example, through community wealth rankings. While the empirical results in this paper indicate that the efficiency gains from CDR-based targeting are significant, it may reinforce existing power structures (Taylor, 2016; Blumenstock, 2018a; Abebe et al., 2021). Efficiency gains should also be considered within the context of evidence suggesting that participating communities may prefer community-based approaches (Alatas et al., 2012), but also may perceive them as less legitimate (Premand & Schnitzer, 2020).

To summarize, our results suggest that there is potential for using CDR-based methods to determine eligibility for economic aid or interventions, significantly reducing program targeting overhead and costs. Our results also indicate that CDR-based methods may complement and enhance existing survey-based methods. We note, however, that the practical and ethical limitations to CDR-based targeting are significant. We emphasize the need to consider these limitations and the constraints of specific local contexts alongside the efficiency gains offered by CDR-based targeting.

References

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S. L., & Sadagopan, S. (2021, March). Narratives and Counternarratives on Data Sharing in Africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 329–341). New York, NY, USA: Association for Computing Machinery. Retrieved 2021-03-04, from https://doi.org/10.1145/3442188.3445897 doi: 10.1145/3442188.3445897

Aiken, E., Bellue, S., Karlan, D., Udry, C. R., & Blumenstock, J. (2021, July). *Machine Learning and Mobile Phone Data Can Improve the Targeting of Humanitarian Assistance*

(Working Paper No. 29070). National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w29070 doi: 10.3386/w29070

Alatas, V., Banerjee, A., Hanna, R., Olken, B., & Tobias, J. (2012). Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, *102*(4), 1206-1240.

Banerjee, A., Duflo, E., Chattopadhyay, R., & Shapiro, J. (2007). Targeting efficiency: How well can we identify the poor? *Institute for Financial Management and Research Centre for Micro Finance, Working Paper Series No. 21*.

Banerjee, A., Hanna, R., Olken, B. A., & Sumarto, S. (2018, December). *The (lack of) Distortionary Effects of Proxy-Means Tests: Results from a Nationwide Experiment in Indonesia* (Working Paper No. 25362). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w25362 doi: 10.3386/w25362

Bedoya, G., Coville, A., Haushofer, J., Isaqzadeh, M., & Shapiro, J. (2019). No household left behind: Afghanistan targeting the ultra poor impact evaluation. *World Bank Policy Research Working Paper*, *8877*.

Björkegren, D., Blumenstock, J. E., & Knight, S. (2020). Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

Blumenstock, J. (2016). Fighting poverty with data. *Science*, *353*, 753-754.

Blumenstock, J. (2018a). Don't forget people in the use of big data for development. *Nature*, *561*, 170-172.

Blumenstock, J. (2018b). Estimating economic characteristics with phone data. *American Economic Review: Papers and Proceedings*, *108*, 72-76.

Blumenstock, J. (2020, May). Machine learning can help get COVID-19 aid to those who need it most. *Nature*. Retrieved 2020-05-15, from https://www.nature.com/articles/d41586-020-01393-7 doi: 10.1038/d41586-020-01393-7

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone data. *Science*, *350*, 1073-1076.

Brown, C., Ravallion, M., & van de Walle, D. (2018). A poor means test? econometric targeting in Africa. *Journal of Development Economics*, *134*, 109-124.

Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, *371*(6535).

Camacho, A., & Conover, E. (2011, May). Manipulation of Social Program Eligibility. *American Economic Journal: Economic Policy*, *3*(2), 41–65. Retrieved 2017-09-25, from https://www.aeaweb.org/articles?id=10.1257/pol.3.2.41 doi: 10.1257/pol.3.2.41

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2020). Micro-Estimates of Wealth for the Developing World. *Revise and Resubmit, Nature*.

Coady, D., Grosh, M., & Hoddinott, J. (2004). Targeting outcomes redux. *The World Bank Research Observer*, *19*(1).

Corral, P., Irwin, A., Krishnan, N., & Mahler, D. G. (2020). *Fragility and conflict: on the front lines of the fight against poverty.* World Bank Publications.

Deaton, A. (2016, June). Measuring and understanding behavior, welfare, and poverty. *American Economic Review*, *106*(6), 1221-43. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.106.6.1221 doi: 10.1257/aer.106.6.1221

de Montjoye, Y., Rocher, L., & Pentland, A. (2016). bandicoot: a python toolbox for mobile phone metadata. *Journal of Machine Learning Research*, *17*, 1-5.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376.

Engstrom, R., Hersh, J. S., & Newhouse, D. L. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being. *World Bank Policy Research Working Paper*(8284).

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., & Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, *9*(1), 22.

Filmer, D., & Pritchett, L. (2001). Wealth effects without expenditure data—or tears: An application to educational enrollments in states of India. *Demography*, *39*, 115-132.

Fortin, S., Kameli, Y., Kone, K., Belem, B., Sangho, H., & Savy, M. (2018). Targeting vulnerable households in rural mali: Effectiveness of a community-based methodology, with or without addition of a proxy-mean test, 2016. *Revue d'Épidémiologie et de Santé Publique*, *66*, S353. Retrieved from https://www.sciencedirect.com/science/article/pii/S0398762018310174 (European Congress of Epidemiology “Crises, epidemiological transitions and the role of epidemiologists”) doi: https://doi.org/10.1016/j.respe.2018.05.317

Gentilini, U., Almenfi, M., Orton, I., & Dale, P. (2020). Social protection and jobs responses to covid-19 : A real-time review of country measures. *The World Bank, Washington, DC*. Retrieved from https://openknowledge.worldbank.org/handle/10986/33635 (License: CC BY 3.0 IGO)

Grosh, M., & Baker, J. L. (1995). Proxy means tests for targeting social programs. *Living standards measurement study working paper*, *118*, 1–49.

Grosh, M., Leite, P., & Wai-Poi, M. (in press). *A New Look at Old Dilemmas: Revisiting Targeting in Social Assistance*. The World Bank.

GSMA. (2020). *Mobile economy.* https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Global.pdf.

Hanna, R., & Olken, B. (2018). Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, *32*, 201-226.

Hernandez, M., Hong, L., Frias-Martinez, V., & Frias-Martinez, E. (2017). Estimating poverty using cell phone data: evidence from Guatemala. , *World Bank Policy Research Working Paper Series No. 7969*.

Hu, X., Yuan, M., Yao, J., Deng, Y., Chen, L., Yang, Q., ... Zeng, J. (2015). Differential privacy in telco big data platform. *Proceedings of the VLDB Endowment*, *8*(12), 1692–1703.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016, August). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. Retrieved 2016-10-07, from http://science.sciencemag.org/content/353/6301/790 doi: 10.1126/science.aaf7894

Jerven, P. (2013). *Poor numbers.* Cornell University Press.

Karlan, D., & Thuysbaert, B. (2019). Targeting ultra-poor households in Honduras and Peru. *The World Bank Economic Review*, *33*(1), 63-94.

Lindert, K., Karippacheril, T. G., Caillava, I. R., & Chávez, K. N. (2020). *Sourcebook on the foundations of social protection delivery systems.* World Bank Publications.

Pokhriyal, N., & Jacques, D. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, *114*, E9783–E9792.

Premand, P., & Schnitzer, P. (2020, 09). Efficiency, Legitimacy, and Impacts of Targeting Methods: Evidence from an Experiment in Niger. *The World Bank Economic Review*. Retrieved from https://doi.org/10.1093/wber/lhaa019 (lhaa019) doi: 10.1093/wber/lhaa019

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Lobell, D., ... Ermon, S. (2019). Predicting economic development using geolocated wikipedia articles. *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Steele, J., Sundsøy, P., Pezzulo, C., Alegana, V., Bird, T., Blumenstock, J., ... Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, *14*.

Taylor, L. (2016). No place to hide? the ethics and analytics of tracking mobility using mobile phone data. *Environment and Planning D: Society and Space*, *34*(2), 319–336.

Tables and Figures

Table 1: Summary statistics for different samples of survey respondents

Outcome	(1) **Full sample** **(all observations)**	(2) **Matched** **Subsample**	(3) **Unmatched** **Owns Phone**	(4) **Unmatched** **No Phone**
Panel A: Balance of Covariates				
Ultra-Poor	0.42 (0.49)	0.27 (0.45)	0.40 (0.49)	0.66 (0.47)
Asset Index	0.00 (2.01)	1.36 (2.60)	-0.05 (1.76)	-1.35 (0.79)
Log Expenditures	4.43 (0.71)	4.64 (0.70)	4.46 (0.70)	4.12 (0.65)
# Phones	1.35 (1.18)	1.72 (1.33)	1.59 (1.04)	0.00 (0.00)
Food Security Index	0.30 (0.90)	0.35 (0.74)	0.34 (0.93)	0.10 (0.89)
Financial Inclusion Index	0.15 (1.27)	0.34 (1.39)	0.15 (1.32)	-0.05 (0.79)
Psychological Well-being Index	0.35 (1.01)	0.38 (1.00)	0.43 (0.97)	-0.02 (1.07)
CWR Group	0.62 (0.90)	0.89 (1.02)	0.62 (0.88)	0.26 (0.66)
Panel B: Correlations Between Outcomes				
Ultra-Poor ⟷ Asset Index	-0.32	-0.30	-0.27	-0.14
Ultra-Poor ⟷ Consumption	-0.39	-0.30	-0.39	-0.26
Asset Index ⟷ Consumption	0.37	0.34	0.34	0.15
N	2,814	535	1,807	472

Notes: Table reports average characteristics, with standard deviations in parentheses, of TUP survey respondents. Each column represents a different sample of respondents: (1) all respondents in the TUP survey; (2) Just those respondents who own a phone, where the phone number matches to the CDR obtained from the mobile phone operator; (3) Respondents who report owning a phone, but whose phone number does not match to the CDR obtained from the operator; (4) Respondents who report they do not own a phone.

Table 2: Targeting simulation results

	(1)	(2)	(3)	(4)
Targeting Method	**AUC**	**Accuracy**	**Precision**	**Recall**
Panel A: Matched Sample (N=535) - for whom we have survey and CDR data				
Random	0.50 (0.028)	0.60 (0.025)	0.27 (0.038)	0.27 (0.038)
Asset Index	0.73 (0.024)	0.72 (0.020)	0.49 (0.041)	0.49 (0.041)
Consumption	0.71 (0.026)	0.69 (0.023)	0.45 (0.038)	0.45 (0.038)
CDR	0.68 (0.027)	0.69 (0.021)	0.42 (0.042)	0.42 (0.042)
Combined	0.78 (0.022)	0.75 (0.020)	0.55 (0.039)	0.55 (0.039)
Panel B: Balanced Sample (N=1,007) - as above, plus households without phones				
Random	0.50 (0.017)	0.90 (0.006)	0.05 (0.010)	0.05 (0.010)
Asset Index	0.72 (0.026)	0.90 (0.006)	0.10 (0.013)	0.10 (0.013)
Consumption	0.70 (0.028)	0.90 (0.006)	0.15 (0.025)	0.15 (0.025)
CDR (Target Phoneless First)	0.68 (0.030)	0.90 (0.006)	0.11 (0.035)	0.11 (0.035)
CDR (Target Phoneless Last)	0.51 (0.028)	0.90 (0.006)	0.12 (0.033)	0.12 (0.033)
Combined (Target Phoneless First)	0.74 (0.026)	0.90 (0.006)	0.11 (0.046)	0.11 (0.046)
Combined (Target Phoneless Last)	0.57 (0.022)	0.90 (0.006)	0.18 (0.007)	0.18 (0.007)
Panel C: Full Sample (N=2,814) - as above, plus households with phones on other networks				
Random	0.50 (0.009)	0.89 (0.005)	0.06 (0.007)	0.06 (0.007)
Asset Index	0.65 (0.017)	0.89 (0.005)	0.07 (0.014)	0.07 (0.014)
Consumption	0.69 (0.015)	0.89 (0.006)	0.08 (0.031)	0.08 (0.031)
CDR (Target Phoneless First)	0.52 (0.008)	0.89 (0.005)	0.06 (0.008)	0.06 (0.008)
CDR (Target Phoneless Last)	0.48 (0.008)	0.89 (0.005)	0.08 (0.010)	0.08 (0.010)
Combined (Target Phoneless First)	0.52 (0.008)	0.89 (0.005)	0.06 (0.008)	0.06 (0.008)
Combined (Target Phoneless Last)	0.49 (0.008)	0.89 (0.005)	0.09 (0.009)	0.09 (0.009)

Notes: Four different measures of performance (columns) reported for different targeting methods (rows), using different samples of survey respondents (panels). Standard deviations, calculated using 1,000 bootstrap samples, in parentheses. Panel A: The 535-household subsample that is matched to CDR. Panel B: The 535-household matched sample, plus the 472 households that do not have a phone; this is meant to approximate targeting performance if CDR from all mobile networks were available. Sample weights are applied as described in Section 2.4. Panel C: All 2,814 observations from the TUP survey, including households matched to CDR, households that own phones not matched to CDR, and households without phones, with sample weights applied. For Panels B and C, we simulate two types of CDR-based targeting: targeting households without phones first and targeting households without phones last.

Table 3: What types of households are misclassified?

Panel A: Ultra-Poor Households (Differences Between True Positives and False Negatives)

	Asset Index			**Consumption**			**CDR**		
	TP	**FN**	**Diff.**	**TP**	**FN**	**Diff.**	**TP**	**FN**	**Diff.**
Ultra-Poor	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Asset Index	-1.03	1.18	-2.21	-0.34	0.47	-0.81	-0.09	0.25	-0.34
	(0.49)	(1.34)	(0.17)	(1.09)	(1.69)	(0.23)	(1.16)	(1.70)	(0.24)
Consumption	4.21	4.40	-0.19	3.78	4.74	-0.96	4.29	4.32	-0.02
	(0.70)	(0.62)	(0.11)	(0.32)	(0.56)	(0.07)	(0.60)	(0.71)	(0.11)
# Phones	0.89	1.63	-0.74	1.02	1.48	-0.46	1.18	1.33	-0.16
	(0.68)	(1.12)	(0.15)	(0.73)	(1.14)	(0.16)	(0.61)	(1.21)	(0.15)
Food Security Index	-0.59	-0.51	-0.08	-0.83	-0.32	-0.51	-0.51	-0.58	0.07
	(1.13)	(1.10)	(0.18)	(1.19)	(0.99)	(0.18)	(1.14)	(1.09)	(0.19)
Financial Inclusion Index	-0.00	0.29	-0.29	0.10	0.19	-0.09	0.16	0.14	0.02
	(0.79)	(1.02)	(0.15)	(0.80)	(1.02)	(0.15)	(0.98)	(0.88)	(0.16)
Psychological Well-being Index	-0.35	-0.13	-0.22	-0.37	-0.12	-0.24	-0.31	-0.17	-0.14
	(0.92)	(0.94)	(0.15)	(0.86)	(0.98)	(0.15)	(0.81)	(1.02)	(0.15)
CWR Group	0.09	0.01	0.07	0.02	0.08	-0.06	0.06	0.04	0.03
	(0.44)	(0.12)	(0.05)	(0.12)	(0.41)	(0.05)	(0.40)	(0.24)	(0.06)

Panel B: Non-Ultra-Poor Households (Differences Between True Negatives and False Positives)

	Asset Index			**Consumption**			**CDR**		
	TN	**FP**	**Diff.**	**TN**	**FP**	**Diff.**	**TN**	**FP**	**Diff.**
Ultra-Poor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Asset Index	2.53	-1.08	3.61	2.06	0.94	1.12	1.94	1.43	0.51
	(2.62)	(0.50)	(0.16)	(2.92)	(1.75)	(0.26)	(2.87)	(2.27)	(0.30)
Consumption	4.82	4.57	0.25	4.97	3.98	0.99	4.78	4.74	0.04
	(0.66)	(0.65)	(0.08)	(0.58)	(0.23)	(0.04)	(0.68)	(0.61)	(0.08)
# Phones	2.11	0.96	1.15	1.98	1.52	0.46	1.91	1.80	0.11
	(1.43)	(0.76)	(0.12)	(1.49)	(0.92)	(0.13)	(1.44)	(1.24)	(0.16)
Food Security Index	0.24	-0.16	0.40	0.24	-0.14	0.37	0.15	0.18	-0.02
	(0.87)	(1.03)	(0.13)	(0.88)	(0.99)	(0.12)	(0.91)	(0.94)	(0.12)
Financial Inclusion Index	0.80	-0.01	0.82	0.77	0.18	0.59	0.78	0.17	0.61
	(4.92)	(0.82)	(0.29)	(4.94)	(1.24)	(0.31)	(4.98)	(1.10)	(0.31)
Psychological Well-being Index	0.69	0.21	0.47	0.62	0.49	0.13	0.62	0.49	0.13
	(0.97)	(0.75)	(0.10)	(0.98)	(0.80)	(0.11)	(0.95)	(0.93)	(0.12)
CWR Group	1.30	0.84	0.46	1.23	1.13	0.10	1.26	1.01	0.25
	(1.00)	(0.96)	(0.12)	(1.03)	(0.94)	(0.12)	(1.01)	(0.98)	(0.12)

Notes: Table shows the average characteristics (with standard deviations in parentheses) of households that are correctly classified (True Positives [TP] and True Negatives [TN]) and incorrectly classified (False Negatives [FN] and False Positives [FP]), as well as the difference in average characteristics between correctly and incorrectly classified households (Diff.). Panel A: Differences between ultra-poor households correctly classified as such and those misclassified as non-ultra-poor (errors of exclusion). Panel B: Differences between non-ultra-poor households correctly classified as such and those misclassified as ultra-poor (errors of inclusion).

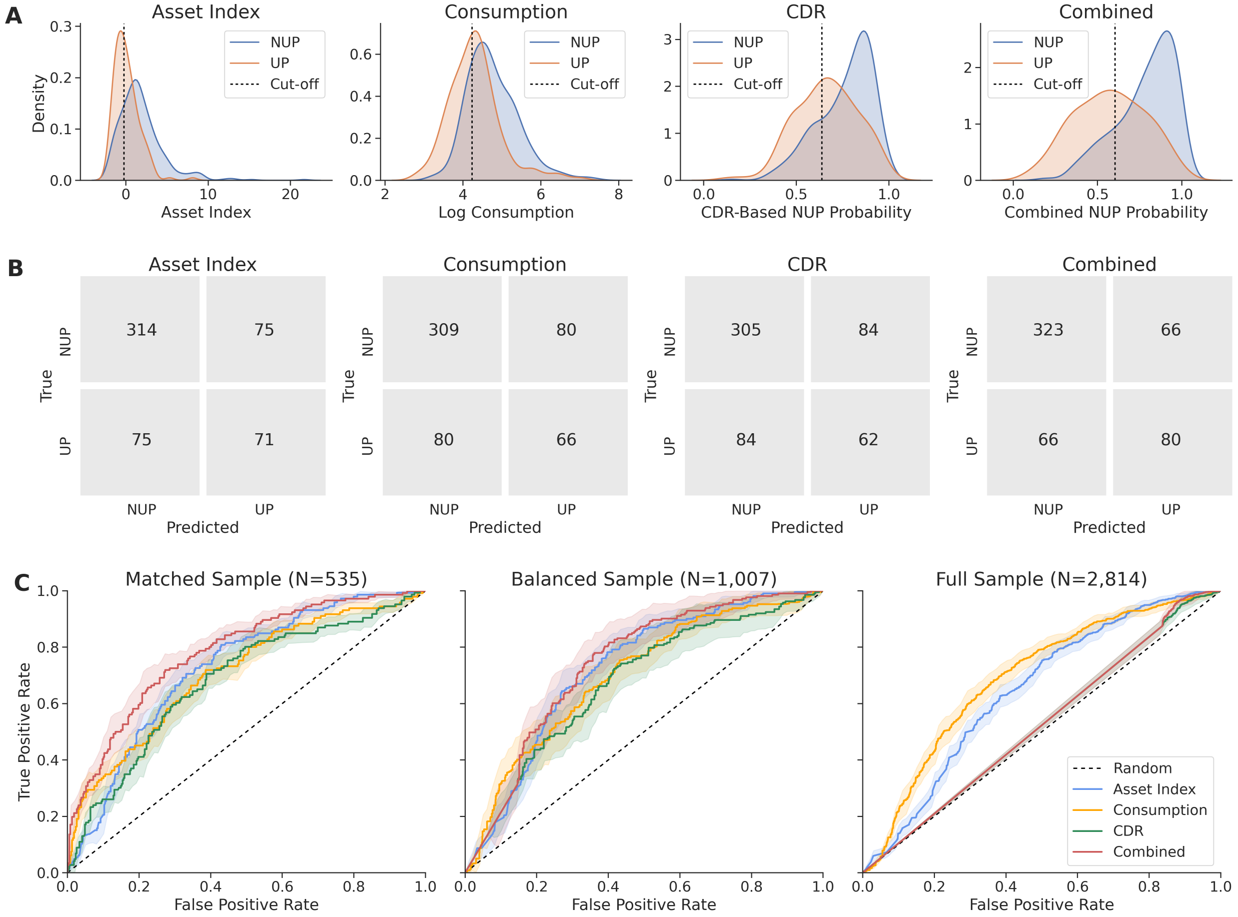
Table 4: Overlap in targeting errors between methods

	Asset Index	**Consumption**	**CDR**	**Combined**
Panel A: Overlap in Errors of Exclusion				
Asset Index	100.00%	65.33%	57.33%	66.67%
Consumption	61.25%	100.00%	56.25%	62.50%
CDR	51.19%	53.57%	100.00%	63.10%
Combined	75.76%	75.76%	80.30%	100.00%
Panel B: Overlap in Errors of Inclusion				
Asset Index	100.00%	26.67%	22.67%	48.00%
Consumption	25.00%	100.00%	16.25%	37.50%
CDR	20.24%	15.48%	100.00%	46.43%
Combined	54.55%	45.45%	59.09%	100.00%

Notes: Table measures the extent to which the targeting errors produced by each pair of targeting methods overlap. Evaluation is performed on the matched sample of 535 TUP respondents. Panel A: Overlap between ultra-poor households that are misclassified as non-ultra-poor (errors of exclusion) for each targeting method. Panel B: Overlap between non-ultra-poor households that are misclassified as ultra-poor (errors of inclusion).

Figure 1: Predicting ultra-poor status from CDR

.



Notes: Panel A: Comparing the predictive accuracy of assets, consumption, and CDR-based methods for identifying the ultra-poor in our 535-household sample. To adjust for class balance, thresholds for classification (shown in dashed black vertical lines) are selected such that the correct number of households are identified as ultra-poor. Panel B: Confusion matrices showing the targeting accuracy of each method shown in Panel A. Panel C: ROC curves for each of the four targeting methods. In the third subplot, the CDR-based and combined methods target non-phone-owning households first as described in Section 2.4

.

Online Appendix

A Machine learning methods and hyperparameters

Although our paper is focused on identifying the ultra-poor with CDR, we experiment with predicting four measures of ground-truth welfare with CDR features: ultra-poor status (binary), below the national poverty line (binary), asset index (continuous), and log consumption (continuous). For the binary measures, we experiment with four classification models: logistic regression (unregularized), logistic regression with L1 penalty, a random forest, and a gradient boosting model. For the continuous measures, we experiment with four regression models: linear regression, LASSO regression, a random forest, and a gradient boosting model. The linear models and random forest are implemented in Python's scikit-learn package. The gradient boosting model is implemented with Microsoft's LightGBM.

In each case, we produce predictions out-of-sample over 10-fold cross validation. We use nested cross-validation to tune the hyperparameters of each model over 5-fold cross-validation within each of the outer folds to avoid any information leakage between folds. We report both the mean score across the 10 folds as well as the overall score when data from all folds is pooled together. For the linear models and random forest, missing data is mean-imputed and each feature is scaled to zero mean and unit variance before fitting models (these transformations are done separately for each fold, with parameters fitted only on the training data for each fold). For the gradient boosting model missing values are left as-is and features are not scaled. We re-fit the model on the entire data, again tuning hyperparameters over 5-fold cross validation, to report selected hyperparameters and feature importances. We also report the top 5 features for each model, determined by the magnitude of the coefficient for the linear models, and by and by maximum impurity reductions for the tree-based models.

Hyperparameters are selected from the following grids for each model:

Linear/Logistic Regression

- Drop columns with missingness over: {50%, 80%, 100%}
- Drop columns with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}

LASSO Regression

- Drop columns with missingness over: {50%, 80%, 100%}
- Drop columns with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}
- L1 penalty: {0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100}

Random Forest

- Drop columns with missingness over: {50%, 80%, 100%}
- Drop columns with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}
- Number of Trees: {20, 50, 100}
- Maximum Depth: {1, 2, 4, 6, 8, 10, 12}

Gradient Boosting Model

- Drop columns with missingness over: {50%, 80%, 100%}
- Drop columns with variance under: {0, 0.01, 0.1}
- Winsorization limit: {0%, 1%, 5%}
- Number of Trees: {20, 50, 100}
- Minimum data in leaf: {5, 10}
- Number of leaves: {5, 10, 20}
- Learning rate: {0.05, 0.075}

B Abbreviations in Feature Names

Figure S4 and Tables S7, S2, and S6 use a set of abbreviations in CDR feature names. This appendix lists the relevant abbreviations.

- BOC: Balance of contacts
- CD: Call duration
- IPC: Interactions per contact
- IT: Interevent time
- NOI: Number of interactions
- PPD: Percent pareto durations (percentage of call contacts accounting for 80% of call time)
- PPI: Percent pareto interactions (percentage of contacts accounting for 80% of subscriber's interactions)
- RD: Response delay
- RR: Response rate
- WD: Weekday
- WE: Weekend

Supplementary Tables and Figures

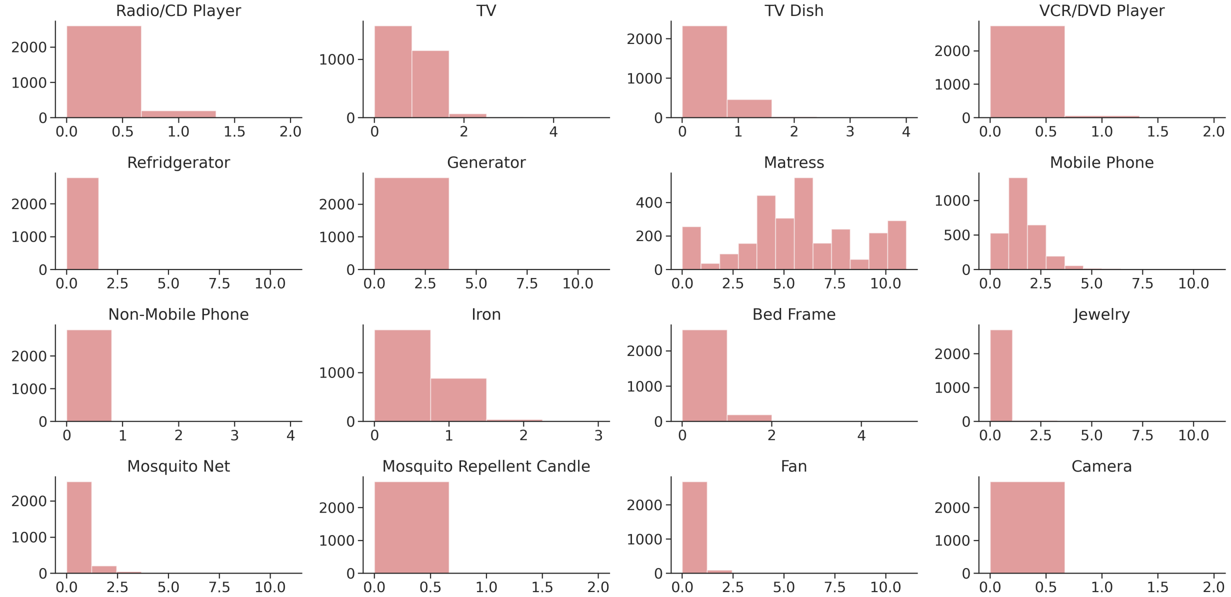


Figure S1: Histograms showing the distribution of each underlying asset used to construct the asset index.

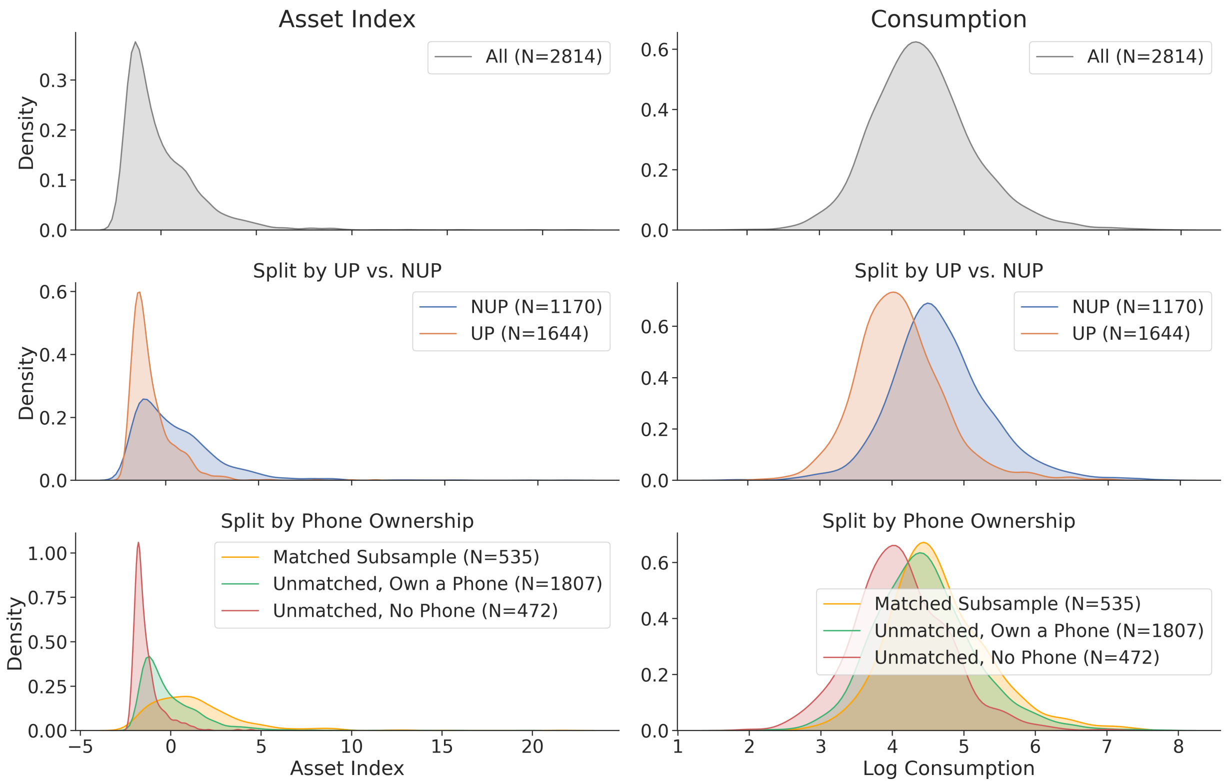


Figure S2: Distributions of asset index and log-transformed consumption, for the entire survey sample, separately for ultra-poor and non-ultra-poor households, and again separately for households in the subsample matched to CDR, households outside of the matched subsample that report owning at least one mobile phone, and households outside of the matched subsample that report not owning a mobile phone.

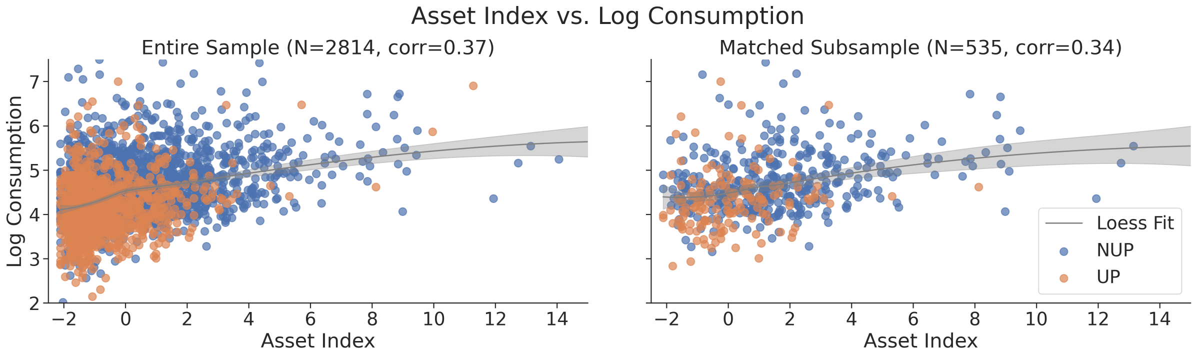


Figure S3: Correlation between asset index and log-transformed consumption, separately for the entire survey sample and the matched subsample. We include the LOESS fit, along with a 95% confidence interval.

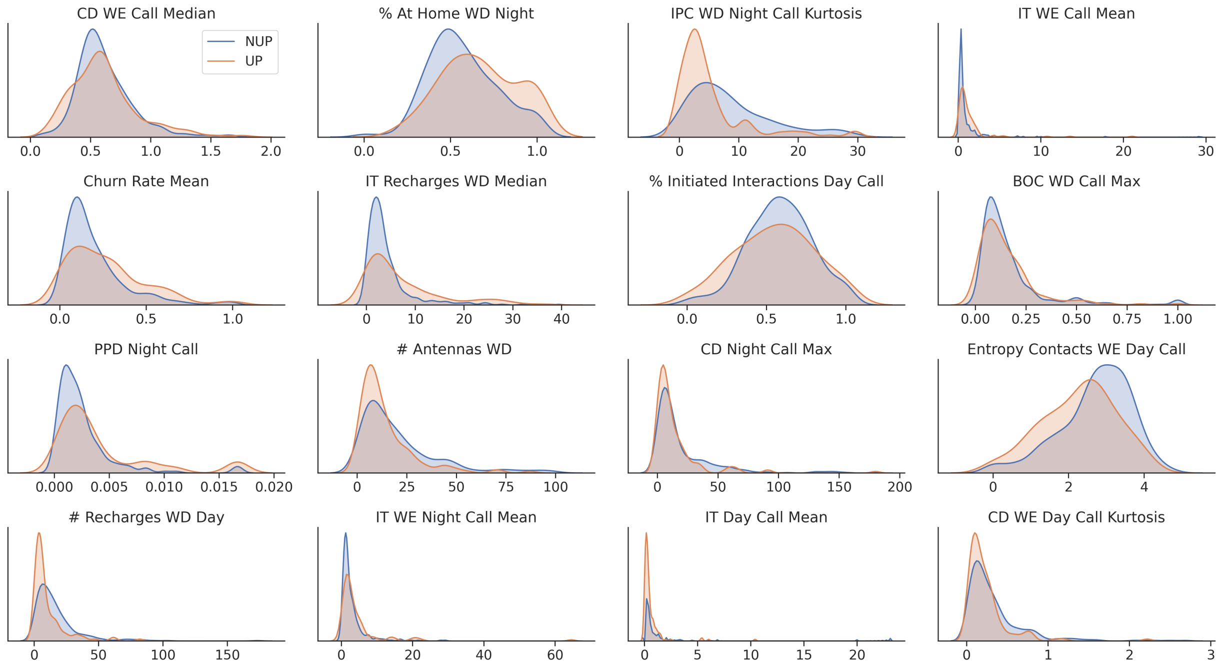


Figure S4: Kernel density estimates for 16 of the most important features for predicting ultra-poor status from CDR, with density estimates shown separately from UP and NUP households. Since many features are near-redundant, rather than showing the raw top 16 features from the table above, we show 16 selected features from the top 50. See Appendix B for abbreviations in feature names.

Table S1: Direction of first principal component of asset ownership

Asset	**Magnitude**
Radio/CD Player	0.04
TV	0.37
TV Dish	0.29
VCR/DVD Player	0.15
Refridgerator	0.25
Generator	0.11
Matress	0.24
Mobile Phone	0.31
Non-Mobile Phone	0.06
Iron	0.36
Bed Frame	0.29
Jewelry	0.27
Mosquito Net	0.26
Mosquito Repellent Candle	0.08
Fan	0.37
Camera	0.16

Notes: The asset index is calculated over the entire 2,814 household sample, without sample weights. We standardize each of the features to zero mean and unit variance before decomposition. The first principal component accounts for 25.28% of the variation in these standardized features.

Table S2: Feature importances (gradient boosting model)

Feature	Importance	Feature	Importance
CD WE Call Median	8	IT Recharges Night Min	3
% At Home WD Night	7	BOC WD Call Median	3
IPC WD Night Call Kurtosis	7	IT WE Call Min	2
CD Day Call Median	6	BOC WD Night Call Kurtosis	2
IT WE Call Mean	6	IPC Day Call Kurtosis	2
Churn Rate Mean	5	% Nocturnal WD Call	2
IPC Day Call Skew	5	IT WD Day Text Mean	2
IT Recharges WD Median	5	CD Night Call Max	2
% At Home Day	4	IT WE Call Skew	2
% Initiated Interactions Day Call	4	IPC WE Day Call Kurtosis	2
% Initiated Interactions WD Day Call	4	IT WE Text Median	2
BOC WD Call Max	4	% At Home WE Night	2
% Initiated Interactions WD Night Call	3	Entropy Contacts WE Day Call	2
PPD Night Call	3	# Recharges WD Day	2
IT Recharges WD Night Min	3	Entropy Antennas WD	2
IT Recharges Night Median	3	IPC Night Call Skew	2
IPC WD Night Text Mean	3	IT WE Night Call Mean	2
IT Recharges Day Kurtosis	3	# Contacts Day Call	2
IT Night Text Min	3	CD WD Call Max	2
IT WE Day Text Median	3	IT Day Call Mean	2
# Antennas WD	3	IT WD Night Text Min	2
CD WD Night Call Kurtosis	3	Entropy Antennas Day	2
IPC Night Call Kurtosis	3	% Initiated Interactions WE Day Call	2
IPC WE Night Call Kurtosis	3	CD WE Day Call Kurtosis	1
IPC WE Night Call Skew	3	IPC Day Call Std	1

Notes: For our selected machine learning model – the gradient boosting model used to predict ultra-poor status from CDR features – we display feature importances for the top 50 features. Feature importances for the gradient boosting model represent the total number of times the feature is used for a split in the entire ensemble of decision trees. We report feature importances when the model is trained on all 535 observations (rather than over cross validation). See Appendix B for abbreviations in feature names.

Table S3: Details of machine learning models

Model	**AUC**	**Top Five Features**
Logistic (No Penalty)	0.53	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Logistic (L1 Penalty)	0.66	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Random Forest	0.68	NOI Out Day Call, NOI Out WD Day Call, Nois Call, Entropy Contacts Night Call, NOI Out WE Call
Gradient Boosting	0.68	CD WE Call Median, % At Home WD Night, IPC WD Night Call Kurtosis, CD Day Call Median, IT WE Call Mean

Notes: Each row indicates performance (AUC) of a different machine learning algorithm, trained to predict ultra-poor status on the sample of 535 matched households. AUC is reported as the mean AUC score over 10-fold cross validation. See Appendix B for details of features.

Table S4: Machine learning an asset index

Model	**AUC Score**	**Top Five Features**
Logistic (L1 Penalty)	0.60	TV, TV Dish, Fridge, Mattress, Mobile Phone
Random Forest	0.73	Fridge, Iron, Bedframe, Mattress, TV Dish
Gradient Boosting	0.74	Mattress, Bedframe, Fridge, Mobile Phone, TV Dish

Notes: The asset index benchmark we used is constructed following standard procedures based on principal comnponent analysis (see Table S1). However, it is possible that an alternative asset-based predictor, trained using machine learning to predict ultra-poor status directly from the 16 underlying components, could perform better. We test this hypothesis by adapting our machine learning pipeline for identifying the ultra-poor from CDR to the task of identifying the ultra-poor from asset possession. As with the CDR-based prediction, we evaluate the model over nested cross validation: the model's predictions are evaluated out-of-sample over 10-fold cross validation, and within each fold hyperparameters are tuned over 5-fold cross validation. We retrain the model on the entire dataset to report hyperparameters and feature importances. Hyperparameters are chosen from the same grid as for the CDR-based models. We display the AUC score and top features for each model.

Table S5: Performance using one, two or three predictor datasets

Data Sources	**AUC**
Assets	0.73
Consumption	0.71
CDR	0.68
Assets + Consumption	0.76
Assets + CDR	0.76
Consumption + CDR	0.75
Assets + Consumption + CDR	0.78

Notes: AUC scores for targeting methods using a single data source, pair of data sources, and all three data sources together

Table S6: Matching household to multiple phone numbers

Model	**AUC**	**Top Five Features**
Logistic (No Penalty)	0.50	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Logistic (L1 Penalty)	0.65	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Random Forest	0.67	NOI call, NOI Out WE Call, IPC WD Night Call Kurtosis, IPC Night Call Kurtosis, IT Recharges WD Day Min
Gradient Boosting	0.66	Churn Rate Std, CD WE Call Median, IPC WD Night Call Kurtosis, IPC Day Call Skew, % Initiated Interactions Day Call

Notes: In our main analysis, for multi-phone households we use only the phone number belonging to the household head (or to a random household member, where no household head is specified), leaving 535 household-level observations. Here we consider instead using machine learning methods to predict individual-level ultra-poverty, with a dataset of 634 individual phone numbers matched to the ground-truth wealth measures for the associated households. We find that the individual-level models are slightly less accurate than the household-level models presented in the main paper, but we focus on the household-level models in the main paper since the household was the unit of targeting in the TUP program. See Appendix B for abbreviations in feature names.

Table S7: Predicting other measures of poverty from CDR

Model	**R^2 or AUC**	**Top Five Features**
Panel A: Predicting below poverty line (binary)		
Logistic (No Penalty)	0.53	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Logistic (L1 Penalty)	0.53	Reporting # Records, Active Days, Active Days Day, Active Days Night, Active Days WD
Random Forest	0.56	NOI Out Night Call, BOC Night Call Kurtosis, CD Day Call Skew, Nois Night Call, IT Night Call Kurtosis
Gradient Boosting	0.55	IT Night Call Kurtosis, IT Text Max, Radius Gyration WE Night, Entropy Antennas, NOI Out WD Call
Panel B: Predicting consumption (continuous)		
Linear Regression	-0.21	% Pareto Recharges WE Night, % Pareto Recharges WE, % Pareto Recharges Night, Entropy Contacts WD Day Text, PPI WE Night Text
LASSO Regression	-0.00	Reporting # Records, PPI Text, PPI Day Text, PPI Night Call, PPI Night Text
Random Forest	-0.02	Churn Rate Mean, IPC WE Night Call Kurtosis, IT Recharges WE Day Skew, IPC WE Night Call Skew, CD WE Call Median
Gradient Boosting	-0.03	CD WD Night Call Skew, IPC WD Day Text Skew, IT WD Night Call Min, IT WD Night Call Max, IT WE Night Call Max
Panel C: Predicting asset index (continuous)		
Linear Regression	-0.06	IPC Text Min, IPC WD Text Min, IPC WD Day Text Min, BOC WD Text Min, % Initiated Conversations WD
LASSO Regression	0.00	Active Days WE Day, Active Days WD, Active Days WE, Active Days, Active Days WD Day
Random Forest	0.00	IT Night Call Skew, IPC Text Min, IT WE Day Call Median, IT WE Call Median, Entropy Contacts WE Night Call
Gradient Boosting	-0.02	IT Text Median, Entropy Antennas WE, Entropy Antennas WD Night, Entropy Contacts WE Night Call, IT Recharges Night Min
Panel D: Predicting CWR group (continuous)		
Linear Regression	0.01	PPI Night Text, IT Recharges Day Skew, IPC WE Call Min, Active Days WE Night, IT Recharges WD Day Skew
LASSO Regression	0.05	PPI Night Text, Active Days WE Day, Active Days WE Night, IT Recharges WD Day Skew, IT Recharges Day Skew
Random Forest	0.04	# Contacts WE Day Call, Entropy Contacts WD Night Call, IPC Night Call Kurtosis, # Contacts WE Call, IT Call Kurtosis
Gradient Boosting	0.03	IT Call Kurtosis, IT Recharges Day Skew, # Contacts WE Day Call, IT Recharges Day Kurtosis, IPC WD Night Call Kurtosis

Notes: Machine learning results for predicting: (A) Below-poverty-line status, using consumption data and based on Afghanistan's national poverty line; (B) Total consumption (log-scale); (C) Asset index; and (D) Community Wealth Ranking. Performance is evaluated on the sample of 535 matched households. Binary metrics (A) are evaluated using the mean AUC score over 10-fold cross validation; Continuous metrics (B-D) are evaluated using the mean R^2 score over 10-fold cross validation. See Appendix B for details of features.