Enhancing the Distribution of Social Services in Mexico

Maria Fernanda Alcala-Durand
Instituto Tecnológico Autónomo de México
mfalcalad@gmail.com

Mobin Javed
UC Berkeley
mobin.javed@berkeley.edu

Diego Garcia-Olano
University of Texas at Austin
diegoolano@gmail.com

Kris Sankaran
Stanford University
kriss1@stanford.edu

Adolfo De Unánue
Instituto Tecnológico Autónomo de México
adolfo.deunanue@itam.mx

Paul van der Boor
McKinsey
pvboor@gmail.com

Eric Potash
University of Chicago

Luis Iaki Alberro
Secretaría de Desarrollo Social

Roberto Sánchez Avalos
Secretaría de Desarrollo Social

Lauren Haynes
University of Chicago
lnhaynes@uchicago.edu

Rayid Ghani
University of Chicago
rayid@uchicago.edu

ABSTRACT

The Government of Mexico's social development agency, SEDESOL, actively collects and analyzes rich data on on the day-to-day operation of its many services in order to more effectively achieve its mission of lifting families out of poverty. We describe two specific applications of using this data to enhance the distribution of social services, implemented in collaboration with SEDESOL. We detail the problem context, available data, our machine learning formulation, experimental results, and a characterization of effective feature sets. We view this work as a step towards facilitating the delivery of social services based on readily available historical data.

KEYWORDS

poverty, prediction, data science, social good

ACM Reference format:
Maria Fernanda Alcala-Durand, Mobin Javed, Diego Garcia-Olano, Kris Sankaran, Adolfo De Unánue, Paul van der Boor, Eric Potash, Luis Iaki Alberro, Roberto Sánchez Avalos, Lauren Haynes, and Rayid Ghani. 2017. Enhancing the Distribution of Social Services in Mexico. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 8 pages.
DOI:

1 INTRODUCTION

Here, we describe an effort to quantitatively inform decision-making related to two specific problems encountered by the Secretería de Desarrollo Social (SEDESOL), the Government of Mexico's social development agency. The first problem is related to systematic underreporting on applications for social services – we ask whether it is possible to automatically flag suspicious applications in order to facilitate investigation of those attempting to defraud the system and hence ensure that services go to those in real need.

The second problem, also related to the distribution of social services, is whether it is possible to approximate the multidimensional poverty profile of a household – usually computed using detailed surveys which are time and labor intensive to collect – using internal transactional and publicly available census data as a more scalable alternative.

We report progress on both problems. We begin with a brief review of the problem context, and then describe our general machine learning formulation. We propose different feature sets and evaluation strategies for each problem and explain contrasting situations where our machine learning pipeline either succeeds or fails in delivering meaningful improvements over natural baselines.

1.1 Partner Description

The organization in charge of quantifying poverty and evaluating social policy in Mexico is the Consejo Nacional de Evaluación de la Polítca de Desarrollo Social (CONEVAL). They have developed a multidimensional measure of poverty, which includes six Basic Needs Indicators, as well as Welfare Income Lines.

The six Basic Needs Indicators are defined in [5] as follows,

(1) Education: lacking the ability to attend school if in relevant age, or basic education if not.
(2) Health Services: lacking affiliation to any health institution, via work benefit, government program or voluntary enrollment.
(3) Social Security: lacking social security, for all employed, unemployed and elderly population.
(4) Quality of the Dwelling: lacking of proper structure or enough space for inhabitants.

Conference'17, Washington, DC, USA

DOI:

(5) Basic Services on the Dwelling: lacking of water, draining or electricity.
(6) Food: lacking of food security on any degree (low, moderate or severe).

On the other hand, the Welfare Income Lines are divided into four categories: first, they differ depending on the location type (urban or rural); second, there are two levels of welfare. The Minimum Welfare Line (LBM, in Spanish: Línea de Bienestar Mínimo) represents the income level a person needs to have in order to cover their food needs, while the Welfare Line (LB, in Spanish: Línea de Bienestar) includes other expenses related to basic goods and services.

SEDESOL operates to fight poverty in Mexico [1]. They aim to break the poverty cycle for communities and individuals by empowering them both through nutrition, education, social security, training and employment programs that make joining and staying in the workforce easier. With the aid of these efforts, families can come out of poverty by generating their own income employing the skills they have learned.

In order to receive many of these programs, potential beneficiaries take a standardized survey in their households called the Single Questionnaire of Socioeconomical Indicators (CUIS, in Spanish: Cuestionario Único de Indicadores Socioeconómicos). This survey contains very detailed information about the household concerning characteristics, services and household supplies, as well as about every individual inhabiting the household, such as their education, work situation, income, etc.

1.2 Problem Description

The responses of this survey are then used to estimate the income level of the household, which must be below the LBM in order to be eligible for receiving assistance. This creates a considerable incentive for underreporting the household situation and supplies, as well as their self-reported income and expenses (these two variables are in fact not taken into account while estimating the eligibility conditions) [12]. However, for a subset of the beneficiaries, there is a second part of the survey, called the Home Verification Module. This part consists of a surveyor going inside the house of a potential beneficiary who has just taken the survey and visually verifying the self-reported answers concerning observable housing variables.

In an effort to avoid program redundancy and centralize transactional information, SEDESOL collects all the assistance information from a wide range of social service programs, into the Single Register of Beneficiaries (PUB, in Spanish: Padrón Único de Beneficiarios) [10]. Nevertheless, not all of the beneficiaries in PUB have taken the CUIS, so the only known information about them is they programs they are receiving assistance from and – to some extent – the spatial context that surrounds them.

Using the CUIS, there is enough information to determine the presence or absence of each of the six types of poverty. This information can help SEDESOL better understand the needs of the population it serves. However, since most beneficiaries do not take a CUIS, a natural question is whether the six indicators can be estimated directly from PUB and their spatial context. We call this problem the imputation of poverty indicators.

2 EXISTING SOLUTIONS AND RELATED WORK

Several systems are in place to address both underreporting and imputation of poverty indicators, though none so far are both reliable and scalable, which leaves room for data-driven techniques.

The most direct existing solution is to require that surveyors attempt to complete a home verification module. While this provides the ground truth for a number of questions, and therefore an indication of whether a respondent underreported, it can only practically be applied for a small subset of SEDESOL's beneficiaries – currently about 400,000 of the 6.8 million households who have taken the CUIS were covered by a home verification module.

An alternative approach involves comparing self-reported incomes with those estimated by responses from other questions. Since no ground truth is available for SEDESOL beneficiaries, this approach relies on surveys developed by CONEVAL to understand regional poverty profiles [5]. However, it is unclear whether those surveyed by CONEVAL are approximately representative of the population of SEDESOL beneficiaries, and even if they were, estimating income from limited household characteristics can be challenging.

For the imputation problem, the most direct solution is increasing the number of individuals who are given comprehensive CUIS. This however is not scalable. An approximate solution is provided by CONEVAL, which estimates the prevalence of the six poverty indicators at the municipal level every five years and at the state level every two years [5], based on general population surveys. Hence, for a finer-resolution view of poverty indicators tailored to SEDESOL beneficiaries, further work is needed.

3 APPROACH

3.1 Data

We compiled various beneficiary enrollment data sources and complementary socioeconomic and geospatial data sets in order to provide the basis for modeling both underreporting and imputation of poverty indicators.

For this project, we extracted transactions from PUB associated with the last quarter of 2015. This comprises 120 million payments across households. In addition to beneficiary IDs and payment amounts, the data includes the program providing the payment and the home address of the recipient, both potentially related to households' poverty profiles.

We obtain responses to the CUIS through a database called SIFODE [7]. In addition to responses for every survey, SEDESOL has estimated six binary variables corresponding to the multidimensional poverty profile. We use these indicators as ground truth, and the intersection between SIFODE and PUB is the basis for training our imputation models. In addition to these surveys, we obtained the home verification module, which are available for a subset of programs.

To supplement these raw program data, we collected related geospatial and socioeconomic information. As geospatial data, we generated latitude-longitude coordinates based on the addresses recorded in PUB, using an open source geocoding library [9]. Naive implementations were unsatisfactory, since addresses in Mexico are often not standardized [2]. Best performance was achieved by combining two approaches – one based on street address searches

constrained to known localities, and another using information on known side streets – where evaluation was based on the proportion of geocoded addresses in the correct (a priori known) sublocality.

For socioeconomic context, we retrieved census data from INEGI [11]. This data describes demographics and development across Mexico, at the level of street blocks.

3.2 Formulation

There are several ways to frame both the underreporting and imputation problems, here we describe the reason behind our particular machine learning formulation. For the underreporting problem, the primary factors potentially useful as model responses are (1) distance between self-reported and estimated incomes, (2) general discrepancies on the home verification module, (3) a binary overall "potentially underreporting" question within the module, also filled out by the surveyor.

The first approach has the advantage of enabling training on the CUIS that are not associated with a home verification visit; however, estimating income can itself be a difficult problem [4]. The second and third approaches restrict training to those households visited by home verification inspectors, but have the advantage of being ground truth, and so were preferred. Between these options, we choose use the full set of verification questions and summarize them according to whether there was a discrepancy on any of them. This allows us to retrospectively identify which questions are mostly often underreported.

Regardless of the response, a household's distance from the poverty line is also relevant when operationalizing results, because this threshold is used to determine program eligibility. Therefore, even if a household is predicted to be underreporting, if it is far from this threshold, it may not be worth flagging. Rather than incorporating this at the training level, we account for this fact in the interpretation stage, see Section 3.5.

For the imputation problem, we treat the six indicators as independent binary responses, training using those households present in both SIFODE and PUB. This assumes that the relationship between features and responses is comparable across both populations. To assess the quality of this approximation would require on-the-ground experiments. Further, while treating responses independently prevents improvements based on potential correlation between responses, we preferred maintaining access to the much larger class of modeling techniques designed for single responses, though multitask methods are a potentially interesting area of future investigation.

3.3 Features

We use four primary types of features: spatial, transactional, socioeconomic, and survey features. All four are available in the underreporting problem, but survey responses cannot be used for imputation, since they require completion of the CUIS. In both cases, features are generated at the household level.

For spatial features, we consider the raw geographic coordinates resulting from geocoding. Also in this category are averages, computed over training folds, of each poverty indicator over street blocks and localities.

For transactional features, we build summaries of a household's PUB transaction history. For example, we consider which programs they are enrolled in, the number, rate, and total amount of payments coming from these programs programs, and the initial enrollment date. These are natural choices of features for the imputation problem, because certain programs are directed towards specific poverty indicators, and enrollment to these programs often occurs during campaigns.

Socioeconomic features are derived from summaries of socioeconomic development and are reported by INEGI at the locality and street-block levels [11]. For example, these data include the estimated proportion of households within a street-block that have access to electricity. Such data complements the raw spatial features, giving some description of the development status across different neighborhoods.

Survey features are those available from the CUIS. These data include features like the program recipients' ages, occupations, and needs.

We applied basic preprocessing to all features before inputting them to models. For missing values in numeric features, we use median imputation, while for categorical features we impute the most common class. Further, all categorical features are converted to their dummy variable equivalent before training.

3.4 Models

For the underreporting problem, our baseline predicts that each household is not underreporting – this is the majority class. Across feature sets, we applied random forests (RF), varying the number of trees and choice of splitting criterion [3]. While other models could potentially improve performance, we preferred the generic applicability of random forests across various feature sets without the need for extensive tuning. This allowed greater focus on feature engineering.

For the imputation problem, our baseline predicts the majority class for each of the six indicators. We considered nearest neighbors (kNN), gradient boosting machines (GBM), and random forests (RF) – nonlinear methods were more appropriate for drawing decision boundaries between spatial coordinates and identifying subsets of time indicative of certain programs [6, 8]. To account for scale during both training and prediction steps, we trained different models across the thirty two states in Mexico, with Mexico City, the state with the most beneficiaries, further split into subregions. This parallelization also accounts for heterogeneity in the regression function across regions.

3.5 Evaluation

For both the underreporting and imputation problems, we use nested cross-validation to estimate out-of-sample precision and recall; we further construct visualizations to compare model results with external factors.

We first describe evaluation in the underreporting problem. To properly simulate the situation in which a model encounters a new household that is not part of the training data, we nest our cross-validation folds at the household level. That is, every household is contained entirely within a single cross-validation fold.

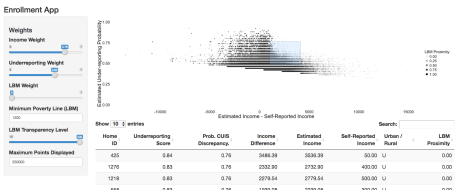

Figure 1: A Shiny application places predicted underreporting probability in context. Each point is a household, those further to the top right have high predicted probability of underreporting as well as large discrepancies between self-reported and estimated incomes. More transparent points are less relevant, because they are far from the income threshold that determines program eligibility.

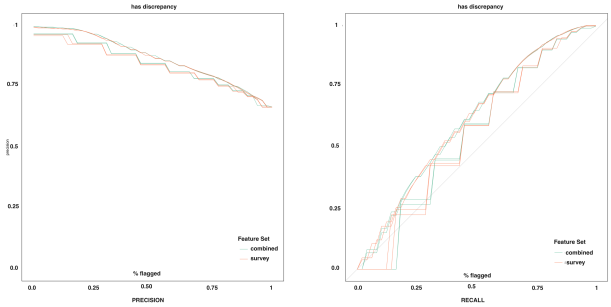

Figure 2: Precision (left) and Recall curves (right) obtained from various models applied to the `any_discrepancy` outcome in the underreporting problem using Survey and Combined (Survey + Demographic) features.

As our response is binary, and since each of our classifiers provides a predicted probability for each class label, we are able to calculate precision-recall curves. Briefly, for a fixed decision threshold, precision measures the number of flagged underreporters who were actual underreporters, while recall measures the proportion of actual underreporters who were identified. The curves are created by varying the threshold on a grid from 0 to 1. This metric is appropriate, considering that in practice, following-up on flagged underreporters takes resources – precision-recall curves can help navigate the trade-off.

Further, to guide interpretation of model results, we implemented a Shiny application to sort samples according to a user-adjustable loss-function [13]. Specifically, it is useful to place a household's model-estimated underreporting probability in context of its distance from the poverty line and the estimated discrepancy between self-reported and estimated incomes. This is because underreporting is only problematic for households just above the poverty line, where underreporting could affect program enrollment. Further, a difference between estimated and self-reported incomes would corroborate any suspected underreporting identified by our models. In our application, predicted underreporting probability and income discrepancies are plotted against each other on a scatterplot, and households far from the poverty line are faded into the background. Brushed points are printed in a table whose rows are sorted according to a tunable weighting these three factors, see Figure 1.

Evaluation for the imputation problem is more straightforward. Again, we nest cross-validation folds at the household level. As in our modeling, we split evaluation across states, since we expect prediction to be more difficult in some than others. Here, our response is a multidimensional binary vector, one for each poverty indicator. Instead of combining error across indicators, we simply compute precision and recall curves for each indicator separately. Hence, we base evaluation by viewing a grid of precision and recall curves across state and indicator combinations, see Figures 6 and 7.

Finally, note that we consider imputation a purely descriptive exercise. That is, our goal is to impute poverty indicators on historical data in PUB, without necessarily attempting to forecast the value of indicators in the future. Hence, we do not pursue temporal cross-validation.

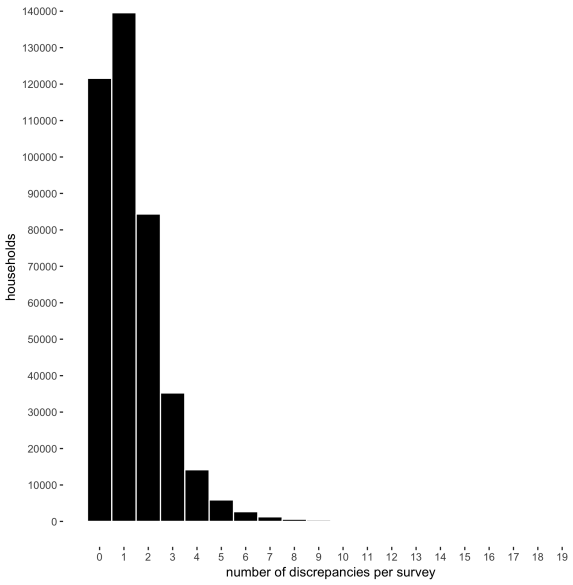

Figure 3: Number of discrepancies found per survey

3.6 Experiments

The results of the underreporting experiment are summarized in Figure 2 and show how different models perform using survey features and a combination of survey and demographic data. Similar to the imputation task, both feature sets use spatial data. Of the 409,000 families for whom we have home verification information, 70% of those visits contain a discrepancy on at least one survey question which in our formulation includes both under and over reporting counts. Of those surveys with discrepancies, the majority of them, 91% contain 3 or fewer discrepancies, see Figure 3. This finding among other things can be due to a respondent's misunderstanding of a question, a data collection error where the surveyor does not mark an initial question correctly or due to misrepresentation by the person surveyed. The high baseline for discrepancy likelihood is reflected in the levels seen in the precision graph. Because of this

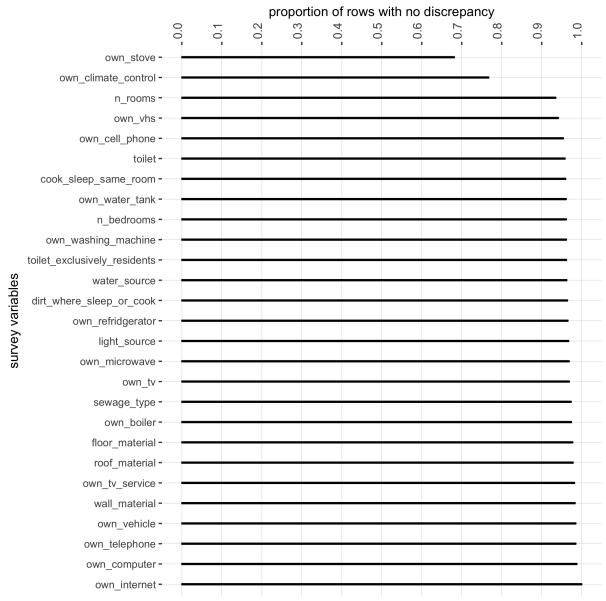

Figure 4: Proportion of survey questions without discrepancies.

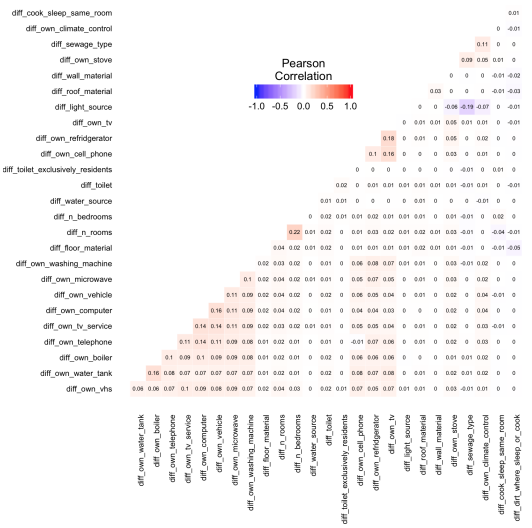

Figure 5: Correlation between survey discrepancies.

likelihood imbalance and due to the utility value for SEDESOL, future modeling of this task will also handle individual question level responses. This task is more challenging however since most individual questions, outside of stove and air conditioning ownership, have a discrepancy level well beneath 5%, see Figure 4 and there exists very little correlation between the discrepancies themselves, see Figure 5 .

The results of an imputation experiment are summarized in Figures 6 and 7. Here, we use kNN (12 and 25 neighbors), GBM (100 and 150 estimators), RF (50 and 100 trees) across four feature sets –

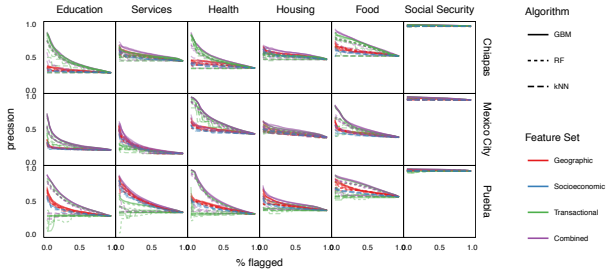

Figure 6: Precision curves across models and feature sets, for the imputation problem. The columns give different poverty indicators, sorted from least to most prevalent, and rows correspond to different subregions (only 3 of 34 shown). Different colors correspond to different feature sets, and line types correspond to algorithms.

geographic, socioeconomic, transactional, and all combined – split across 34 regions. To save space, only three are shown. Chiapas is an example of a rural state, while both Puebla and Mexico City[1] are urban.

We note two filtering steps used to simplify the problem. First, for some poverty indicators, a small fraction of samples in SIFODE are missing a label, we discard these in both training and evaluation. Further, while we use the potentially noisy geocoding results generated on samples with partially missing addresses, we discard those on which no geographic information is available (not even locality labels). For the remaining households – between 30,000 and 80,000 for each state – we use the nested CV approach described before, with $K = 2$, visualizing precision and recall on every held out data set.

As expected, baselines are quite different across indicators and states. For example, almost all beneficiaries lack access to social security, and so our models provide little benefit over flagging everyone with the deprivation. On a related note, when a state has relatively fewer people with an indicator, our model has more value; see the difference between prediction of access to services between Mexico City, a more urban area where lack of access to utilities is rarer, and Chiapas, a state with relatively less developed infrastructure.

Interestingly, different feature sets are more or less informative, depending on the indicator being analyzed. For example, program enrollment features are useful for predicting access to education, health, and food deprivations, while these same features are not as useful as geographic information when predicting access to services and adequate housing. In retrospect, this relationship is natural – access to housing and services like electricity, water, and garbage disposal would be expected to be associated with geography, while education, health, and food access tend to be targeted by specific programs within SEDESOL.

Finally, among the three models, GBM and RF deliver comparable performance, and are consistently more effective than kNN.

[1]This is actually one of several subregions of the state of Mexico City on which we parallelized the imputation task. We are showing the most densely populated subregion

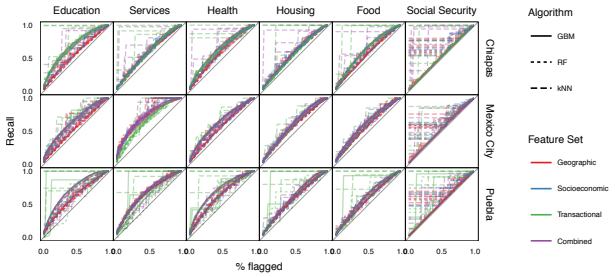

Figure 7: Recall curves corresponding to the precisions in Figure 6

4 DISCUSSION

To better understand how the models are generating predictions, we inspect the feature importances of successful models, and then visualize the interclass variation for the more important features.

Feature importances for a collection of RFs with either 10 or 50 trees each using a combination of survey, socioeconomic and spatial features associated with the underreporting problem are displayed in Figure 8. Air conditioning and stove ownership, money spent on food, age of the person surveyed, number of rooms reported, the frequency of consumption of vegetables, milk and fruit, and meals per day were the most important predictive features.

Of the approximately 130,000 cases where there was a discrepancy with stove ownership, 127,000 were cases of overreporting, i.e. where the person being surveyed said they owned a functioning stove when they did not. Thus, the true underreporting frequency for stove ownership is actually less than 2.5%! Similarly for air conditioning ownership, only in 1000 out of 95,000 cases where a discrepancy had been found was it due to under reporting.

These "dignity" discrepancies, where a respondent misreports their true living situation, was something mentioned by surveyors as anecdotal stories during a site visit to the country. Being born in the Federal State of Mexico is also curiously included in the group of important features, though that seems more a consequence of it being the most populous state in the country.

Feature importances associated with prediction of access to education in the Chiapas subset based on the combined feature set are displayed in Figures 9 and 10. Only the top 100 features are displayed. The GBMs concentrate on a smaller subset of features, compared to the RFs, but there is high overlap between the top features between the two models. We know from Figures 6 and 7 that transactional features are likely most informative, but we can now identify specific programs and benefits associated with access to education. Further, most socioeconomic variables seem only weakly informative, suggesting limited utility of census date alone. Finally, even though geographic features on their own are relatively poor predictors for education in Chiapas, the households' street block coordinate (`manzana_latitude` and `manzana_longitude`) appears among the top ten predictors for both the GBM and RF.

The programs and benefits appearing in these top variables are displayed in Figures 11 and 12, respectively. From Figure 11, we see specific programs whose beneficiaries are either more or less likely to have the education deprivation, compared to a random

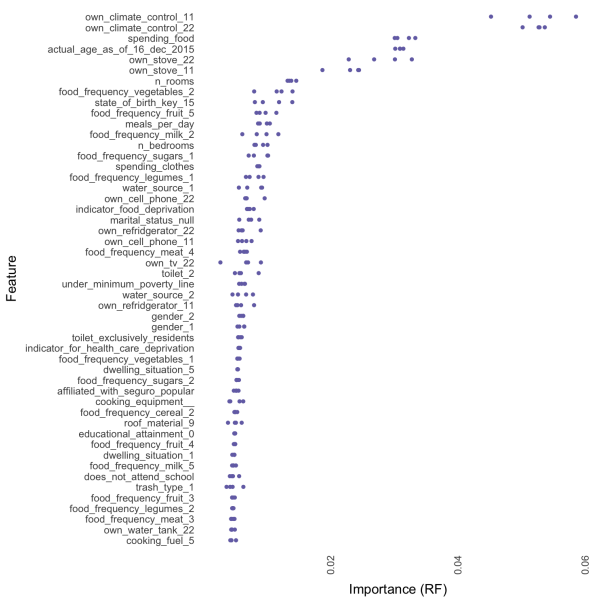

Figure 8: Feature importances from the RF applied to the underreporting problem using survey, socioeconomic and spatial data.

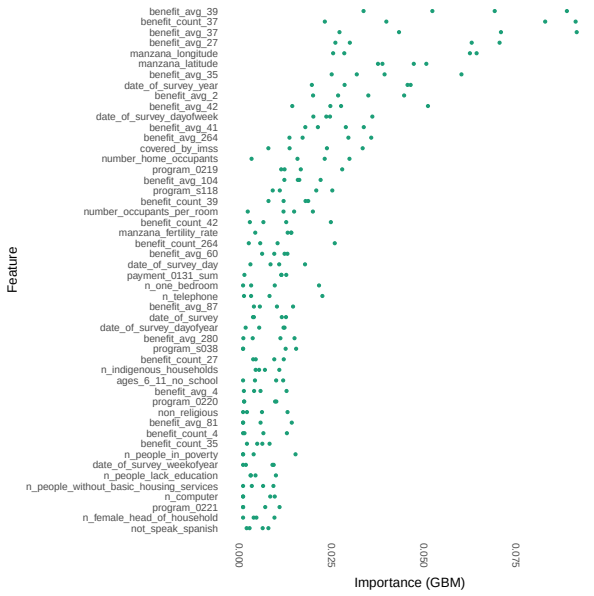

Figure 9: Feature importances from the GBMs, across multiple folds and parameter settings, trained on the combined feature set, when predicting access to education in Chiapas.

household in the data. For example, program S176 corresponds to Pensión para Adultos Mayores, a pension program for senior citizens. Those who are in the program are much more likely to have

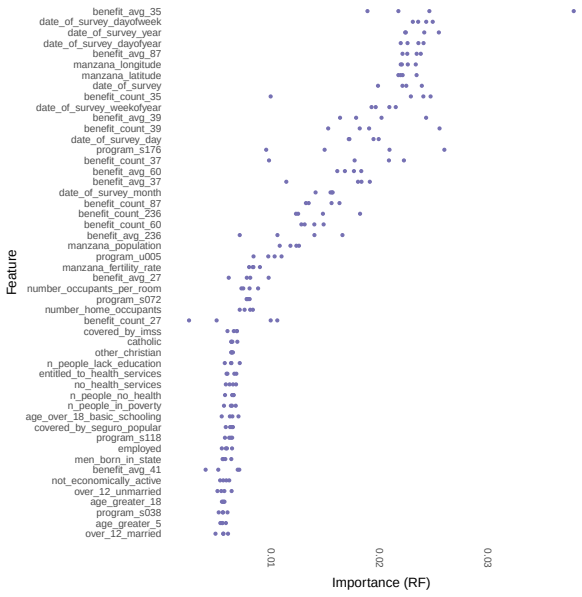

Figure 10: The analog of 9 computed from the corresponding random forest models.

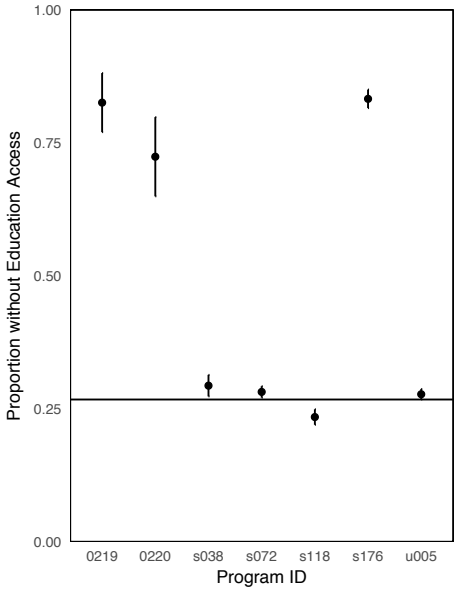

Figure 11: How does program enrollment relate to education access? From the original transactional data used in the models from Figures 9 and 10, we can estimate the proportion of households that do and do not have access to education, grouping by each program. The vertical bars give a 95% confidence interval on the estimated proportions – some programs have more beneficiaries, and so lend themselves to more precise estimation. The horizontal line is the proportion of households with this indicator, across all samples.

the education deprivation[2]. The program with a lower proportion lacking access is a food program (Programa Apoyo Alimentario), and it is possible that households in this program are targeted for insufficiency on the food dimension of poverty, rather than the education dimension.

In Figure 7, we look at the frequency of benefits from programs with high variable importance, split according to whether the household is flagged as lacking access to education. Since most households do not receive these benefits, there is in fact a large spike at zero for each row – we omit this for clarity. We can see the discriminative potential in the way different colors are not split evenly with the red bars (lacking access) at about 1/2 the height of the green bars (not lacking access), as would be expected by chance, according to the baseline from Figure 11. For example, if a household received any benefits from programs 39 (a renewable energy program), 27 (Apoyo por concepto de beca, a scholarship program), or 2 (Litro de leche, a milk distribution program), 87 (Atencion a la salud, a program focused on health), etc. we would suspect that they do not lack access to education. On the other hand, if they received benefits from programs 27 (Apoyo para adulto mayor, a senior support program), 42 (Servicios educativos de alfabetización, a literacy program), or 60 (Apoyo al ingreso de los productores agricolas, support for agricultural workers), it is likely that they lack access to education in their multidimensional profile.

We have chosen to investigate education in Chiapas because it was a clear example where using transactional features yielded meaningful improvements over baseline. To see an example where geographic features outperform program features – the main alternative regime visible in the precision and recall curves – we consider prediction of services in Mexico City. In Figure 13 we plot a subset of the Mexico City model's training data, split across whether the household lacks access to services and shaded by the predicted probability of lacking access according to a RF trained on geographic features. Since many households can map to the same street block, we make points semitransparent and jitter them slightly. A few dark clumps are associated with errors in geocoding – the center of a neighborhood is sometimes returned when no more higher-resolution location can be found. Nonetheless, it appears that the method has identified a neighborhood, in the bottom right, whose residents seem to more frequently lack access to services. This view gives a sense of the granularity in geocoding that could improve model performance – it is worth investing in a geocoder that can accurately place households in the correct neighborhood, but higher resolution than this only has marginal benefit.

[2]Note that the access to education field asks for the highest education level of different household members. It is possible that seniors enrolled in this program never attained higher levels of education.

5 CONCLUSION

We have systematically analyzed the problems of flagging underreporters and imputing poverty indicators, which has consequences for the design and delivery of services at SEDESOL, the social development ministry in Mexico. We have compiled a range of relevant data sources, applied a machine learning pipeline to a variety of subproblems, and studied the resulting models.

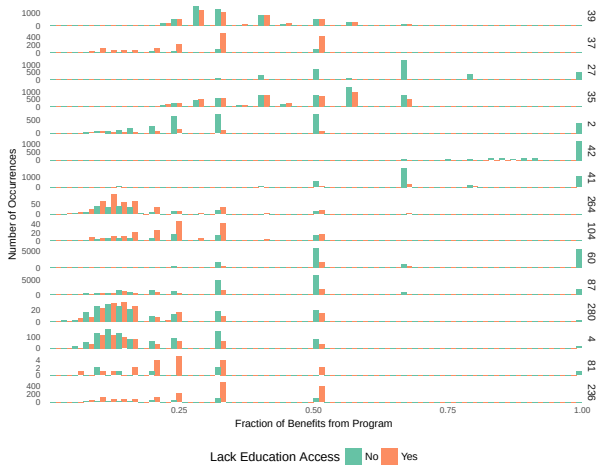

Figure 12: How do program payments relate to education access? Each row here corresponds to a SEDESOL program, and the colors in the histograms correspond to households with and without access education. When all of a household's benefits come from a given program, it falls into the bin at the far right for that program. Generally, a household's benefits will be derived from several programs, and so they will lie in intermediate bins. Spikes at 0, for households never receiving the associated benefit, are omitted for clarity.

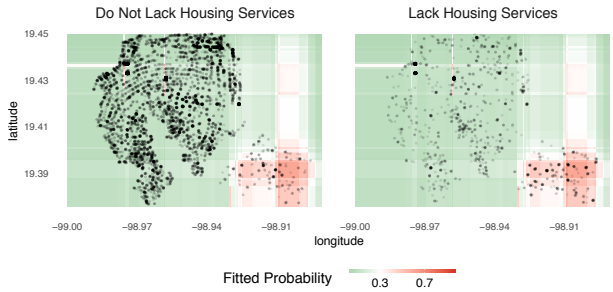

Figure 13: Points here correspond to the street-block coordinates of different households in a neighborhood within the Mexico City region – darker points indicate more households at that block. The two panels separate households with and without access to housing services. The background color is the predicted probability of lacking access using a GBM trained on geographic features, with the color scale adjusted according to baseline prevalence.

We found that survey data alone can suggest potential underreporting, and have conjectured that much of what appears to be underreporting is related to misunderstanding. Further, we found geographic features useful for housing and service related indicators and transactional data informative for other dimensions of poverty.

We are excited by the prospect of analyzing data that already exists at social development organizations in order to facilitate their operation. We hope our approach provides a further example for others seeking to use data to better serve those in need.

ACKNOWLEDGMENTS

Acknowledgements

REFERENCES

[1] 2013. Misin de la SEDESOL. (Feb 2013). http://www.2006-2012.sedesol.gob.mx/es/SEDESOL/SEDESOL

[2] Klaus Ackermann, Eduardo Blancas Reyes, Sue He, Thomas Anderson Keller, Paul van der Boor, Romana Khan, Rayid Ghani, and José Carlos González. 2016. Designing Policy Recommendations to Reduce Home Abandonment in Mexico. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 13–20.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Alfredo Bustos. 2015. Estimation of the distribution of income from survey data, adjusting for compatibility with other sources. *Statistical Journal of the IAOS* 31, 4 (2015), 565–577.

[5] CONEVAL. 2010. Medición de la Pobreza. http://www.coneval.org.mx/Medicion/MP/Paginas/Que-es-la-medicion-multidimensional-de-la-pobreza.aspx. (2010). Accessed: 2016-08-18.

[6] Luc Devroye, László Györfi, and Gábor Lugosi. 2013. *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media.

[7] María del Rosario Cárdenas Elizalde, Fernando Alberto Cortés Cáceres, Agustín Escobar Latapí, Salomón Nahmad Sittón, John Scott Andretta, Graciela María Teruel Belismelis, Gonzalo Hernández Licona, Thania Paola de la Garza Navarrete, Ricardo C Aparicio Jiménez, Edgar A Martínez Mendoza, and others. Investigadores académicos. (????).

[8] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[9] Xianping Ge. 2005. Address geocoding. (Aug. 23 2005). US Patent 6,934,634.

[10] ROgeLIO góMez HeRMOSILLO. 2011. Padrón único de beneficiarios: México. *Integrados de Información Social* (2011), 319.

[11] Censo de Población INEGI. 2011. Vivienda 2010. *Resultados definitivos, México* (2011).

[12] César Martinelli and Susan Wendy Parker. 2009. Deception and misreporting in a social program. *Journal of the European Economic Association* 7, 4 (2009), 886–908.

[13] R Studio. 2014. Shiny: A web application framework for R. (2014).