Responsible Data Science
DS-GA 1017

Costa Rican Household Poverty Level Prediction

Automated Decision System - Nutritional Label

Team Members

Anthony Chen (ac8480)
Sudharsan Asaithambi (sa6149)

Table of Contents

1. Introduction

Developing countries have limited resources for social welfare and want these programs to target-low income households to optimally finance aid. Developing countries often do not have dependable processing, central institutions, and robust infrastructure to collect, validate, and organize income data at scale with the majority of the population working in unorganized sectors.

Proxy Means Test (PMT) is an alternative method to measure the income of a household in developing countries where reliable income data is unavailable. PMT uses observable household characteristics, which can be used as proxy variables that highly correlate with income level. These characteristics would include each households' constructional make up, education, and family composition from surveys.

The Inter-American Development Bank has compiled a dataset of Costa Rican households with 100+ features that refer to the characteristics of their house, household items owned, educational qualifications, and demographics. They have already developed a PMT model with linear regression that classifies the household into 4 income categories from the least poor to the most poor. The IDB has released this dataset in a Kaggle competition, asking the users to help create a robust algorithm that will be critiqued based on the solution's Macro F1 score.

The goal of this ADS is to use more sophisticated means to more accurately identify poverty-stricken households. However, since this ADS has already been implemented, there are multiple trade-offs that emerge.

First, there is a trade-off of inclusion and exclusion errors. Typical PMT models feature an inherent 30-40% of these types of errors. Inclusion error occurs when the model considers that the household was poor when it wasn't, and exclusion error occurs when the model fails to identify a poor household. This results in either higher costs for already constrained social welfare programs or the inaccessibility of needed services for vulnerable households respectively. To address this, more advanced PMT models have a multi-staged approach where the results of one model are plugged into another model. Inevitably, there is usually a higher bias in these models for lower variability in results.

Second, data manipulation for subjective labels such as employment status in developing countries will not make a strong prediction due to fluctuations of adjustments in that feature. A third trade-off of the ADS's goal is fairness because sensitive features such as gender, age, marital status, disability, and geography of the residence are highly prevalent in the dataset. Finally, differential

privacy of individuals is sacrificed because synthetic pre-processing isn't used beforehand to mask re-identification of applicants.

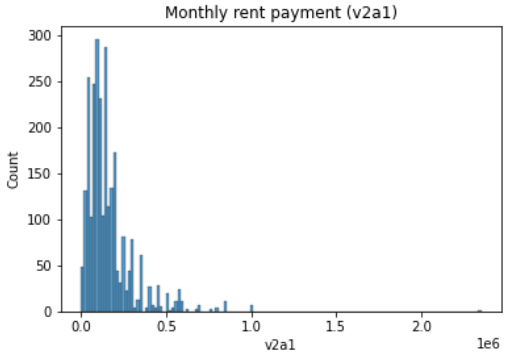
2. Inputs and outputs

According to Table A.5.3 in the appendix of a December 2005 report by the Inter-American Development Bank, the way the data was selected was based on larger categorical groups of demographic variables. These are: 1) dwelling characteristics, which use the floor's, wall's, roof's main material, the type of dwelling, and equipment used, 2) basic services such as the water supply, sewerage & garbage elimination system, electricity, and cooking methods, 3) human capital such as the years of education for household members, and 4) household composition such as gender of the head of household, and the age of children.

Although it is not stated by the IDB on the Kaggle website, data on the Costa Rican population are likely collected via household income and expenditure surveys, which are also known as Household Economic Surveys or Living Standards Measurement Study Surveys. Enumerators, also known as social workers, can visit these households to verify the data for quality assurance, and to also control for fraud. This type of census data is recommended to be recent within the last 5 years due to socioeconomic changes, and it should be nationally representative (World Bank Group). We show in-depth examples of these gathered demographic variables continuing this section.

***Input feature related to dwelling characteristics*:**

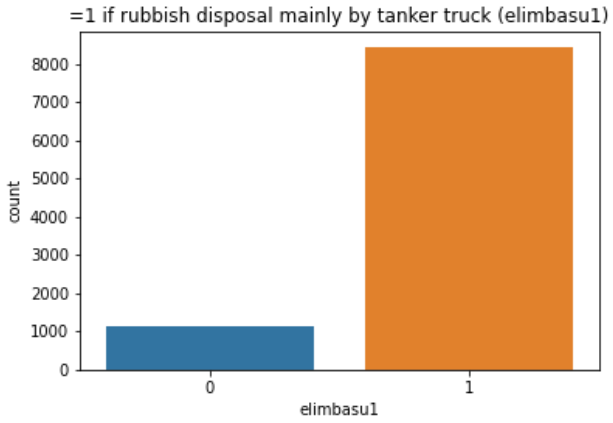
Let's analyze one of the main housing features, the rent of the apartment, which is a float data type. Monthly Rent of a household has a median of 130000 with mean 165000 in their local currency.



***Input feature related to basic services*:**

Some of the features include basic services on how the household disposes their rubbish, what sanitation they have. Below we analyze one of the features described as 'if rubbish disposal is mainly by tanker truck.'

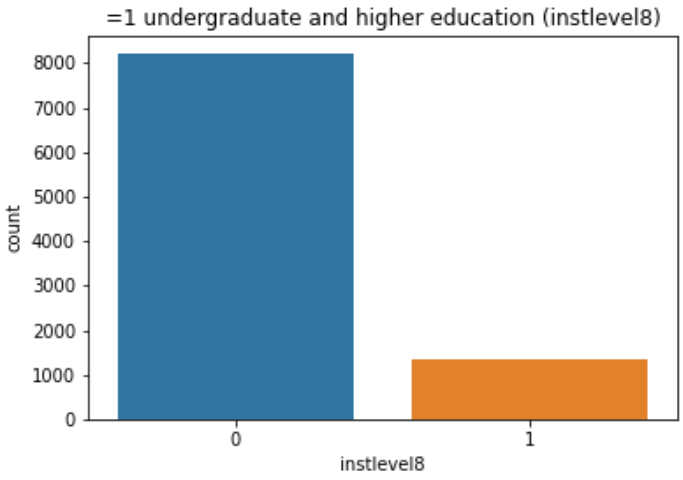
This integer data type feature takes a value of 1 if the rubbish is disposed of by tanker truck, or 0 if by other methods.



***Input feature related to human capital*:**

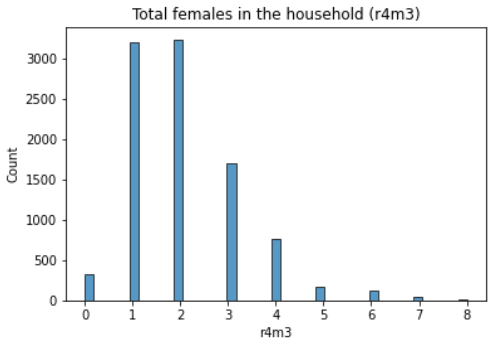
The dataset also contains features on human capital like education, job.

Below we have listed one such feature, which is also an integer datatype. It is a boolean of whether or not an individual has an undergraduate degree.



***Input feature related to household composition*:**

Some of the important features of the dataset are in household composition statistics. It lists how many total people dwell in the house, how many of them are male, female and children. One such feature known more formally as total females in the household is listed below. The feature is an integer data type ranging from 0 to 8.



The output of this ADS is a class label ranging from 1 to 4 indicating poverty levels that are assigned with the following descriptions: 1 = extreme poverty, 2 = moderate poverty, 3 = vulnerable households, and 4 = non-vulnerable households.

3. Implementation & Validation

In order to gain a detailed understanding of the motivation behind the ADS, we are analyzing Mr. Will Koehrsen's Kaggle notebook titled, “A Complete Introduction and Walkthrough”, as the notebook contains comprehensive commentary along with the code.

The author begins by outlining the problem statement and performing detailed exploratory data analysis. Mr. Koehrsen notes that the data is given for each individual in a separate row, while the prediction task is on a household level. Mr. Koehrsen, created a dataset by aggregating the attributes of all the people in a household to a single row with different descriptive statistical functions like sum, min, max, standard deviation, etc. For instance, if a household has three members of age 10, 35 and 38 he would then aggregate them to create age_sum, age_average, age_min, age_max, age_std to adequately represent the age distribution in the household.

Prior to hyperparameter tuning the models, feature selection is further performed on this dataset. The author computed the feature importance using a Random Forest Classifier and found that out of the 200+ features after feature engineering, the top 132 features gave a cumulative feature importance of 95%. Thus, he uses this method to eliminate redundant features.

Mr. Koehrsen performs hyperparameter tuning through tweaking 13 different model parameter configurations in models such as LSVC, GNB, MLP, LDA, Random Forest, etc. He finally submits his predictions with a Random Forest Classifier with the optimized conditions of 100 trees, gini criterion, and selecting maximum features = sqrt(num_features). We analyze his approach and model on various aspects such as performance, fairness, transparency, and privacy. Since the original test set didn't contain true labels due to the set-up of the competition, we split the given training set into 80% training and 20% "validation" as our test set.

4. Performance

According to table A.5.4 in the IDB's report, the bank presents results from its own Income PMT regression analysis with higher coefficients for the model on the rural areas. Examples of broad features used along with specific details are Geographic regions ('del Valle'), Human capital ('spouse's years of education', 'number of literate members of the households older than 14'), Household composition ('Household's members residing abroad'), Dwelling characteristics ('Number of rooms measured by square foot'), Equipment ('Car', 'Air conditioning', 'electric generator', 'washing machine'), Basic Services ('Indoor Plumbing'), Employment ('Head of household employed', 'number of household employed', 'work affiliation in water and electricity and construction sectors', and 'head of household is affiliated to social security'). Thus, these variables impact the overall income classification more for the rural populace than the urban populace. In table A.5.3, the household's demographic variables account for 45% of the overall variance in both the Urban and Rural models.

Their income poverty map (ICV) model generated an ICV estimate for each individual within a household to classify them into one of four welfare categories. Cut-off ranges were then used to determine poverty categories. The following is the table that showcases this method:

Table A.5.1: ICV Cut Off Points

ICV Category	Urban	Rural
ICV-I equivalent to extreme poverty	0.0 – 43.0	0.0 – 32.0
ICV-II equivalent to poverty	43.0 – 58.5	32.3 – 52.5
ICV-III Non poor	58.5 – 75.8	52.5 – 73.9
ICV-IV Non poor	75.8 – 100.0	73.9 - 100

Source: Morrillo, Guerrero and Alcántara (2004)

Our approach instead grouped the given poor categories (1,2, and 3) into the 'poor group', and those belonging to non-vulnerable category 4 into the 'nonpoor group'. By doing this, we can then measure the disparate impact for each selected protected attribute as well as enhance the interpretability of results. We could then further measure the accuracy, precision, recall, false

positive rates, and false negative rates for both the unprivileged and privileged groups within each protected class. Mr. Koehrsen's Random Forest model shows that it performs better on urban households than rural ones.

Since the test dataset did not contain true labels, through using 80% of the aggregated training data to train the model, and 20% of the remaining, the classifier resulted in an accuracy of 0.72. It also performed the best in terms of recall at 0.94, so it identified a majority of the true positives, which are the Costa Ricans who actually need economic assistance. However, it had relatively low precision at 0.71, so roughly 0.29 of positively classified households are identified as poor when they are not. This is further shown by the high false positive rate at 0.69. Overall, since the F1-score, which is the harmonic mean of the recall and precision, is at 0.81 there is a balance of both these metrics, and the model doesn't favor either one of them.

5. Fairness

The IDB dataset contains comprehensive details on dwelling characteristics, household composition, basic services, and human capital. These personal characteristics allow us to investigate the ADS on its fairness. We analyze fairness of the ADS through investigating the impact on true classifications by the protected attributes: Gender, disability, geographical location (urban vs. rural), marital status, and age. We also calculate the ADS's performance metrics and compare them between the privileged and unprivileged classes.

The following are our attempts at "nutritional labels" provided for each of these protected attributes.

Gender

- The survey participants are almost equally distributed between males and females and are representative of the population.
- The disparate impact between males and females is evident with more females in poor households (54%) compared to poor household males (46%). This suggests that men are more privileged compared to women.
- This also suggests the presence of pre-existing bias in the society that could seep into the ADS.
- We also note that the ADS performs differently for the male-dominated vs female-dominated households due to higher accuracy, precision, recall and a lower false negative rate for male dominated households. Females dominated houses have lower false positive rates than male.

Composition

Gender	**Count**
Female	17094 (51%)
Male	16319 (49%)

Disparate Impact

Socioeconomic Status Classification/ Gender	**Male (Privileged Group)**	**Female (Unprivileged Group)**
Poor	46%	54%
Not Poor	50%	50%

Model Performance

Gender / Metrics	**Accuracy**	**Precision**	**Recall**	**FPR**	**FNR**
Male (Privileged)	0.75	0.75	0.95	0.70	0.05
Female (Unprivileged)	0.70	0.70	0.93	0.67	0.06

Disability

- Around 6% of the survey participants are disabled, while 7% of the labeled poor are disabled.
- This suggests that there is a disparate impact based on a person's disability.
- We observe that accuracy and precision for disabled people are much lower than non-disabled people. The false negative rate for non-disabled people is also slightly higher.
- This suggests that the privileged group has better outcomes than the unprivileged group from this ADS.

Composition

Disability Status	**Count**
Disabled	1904 (6%)
Not Disabled	31509 (94%)

Disparate Impact

Socioeconomic Status Classification/ Disability Status	**Not Disabled (Privileged Group)**	**Disabled (Unprivileged Group)**
Poor	93%	7%
Not Poor	95%	5%

Model Performance

Disability Status/ Metrics	**Accuracy**	**Precision**	**Recall**	**FPR**	**FNR**
Not Disabled (Privileged)	0.75	0.76	0.94	0.68	0.06
Disabled (Unprivileged)	0.61	0.55	0.95	0.69	0.05

Geographical Location

- While the majority of the survey participants (72%) are from urban areas, only 66% of the poor population are from urban areas.
- The model performance on the rural survey participants are substantially lower than the urban survey participants creating different outcomes for the two subpopulations.

Composition

Geography	**Count**
Urban	24045 (72%)
Rural	9368 (28%)

Disparate Impact

Socioeconomic Status Classification/ Geography	**Urban (Privileged Group)**	**Rural (Unprivileged Group)**
Poor	66%	34%
Not Poor	75%	25%

Model Performance

Geography/Metrics	**Accuracy**	**Precision**	**Recall**	**FPR**	**FNR**
Urban (Privileged)	0.77	0.77	0.97	0.7	0.05
Rural (Unprivileged)	0.66	0.65	0.95	0.57	0.11

Marital Status

- We find that the marital status of the survey participant does not create a substantial disparate impact or difference in model performance.

Composition

Marital Status	Count
Married	22085 (66%)
Single	11328 (34%)

Disparate Impact

Socioeconomic Status Classification/ Marital Status	Married (Privileged Group)	Single (Unprivileged Group)
Poor	66%	34%
Not Poor	66%	34%

Model Performance

Marital Status/Metrics	Accuracy	Precision	Recall	FPR	FNR
Married (Privileged)	0.74	0.75	0.94	0.68	0.06
Single (Unprivileged)	0.72	0.72	0.94	0.69	0.06

Age

- We find that the survey participants younger than 18 years are more poor than participants older than 18 years.
- This is shown by a substantial difference in the composition of sub-populations in poor and non-poor groups.
- The model performance does not unilaterally favor any sub-population based on age, with the unprivileged sub-population having higher precision and lower false positive rate. However, younger participants have a higher false negative rate.

Composition

Age	**Count**
Senior	24497 (73%)
Junior	8916 (27%)

Disparate Impact

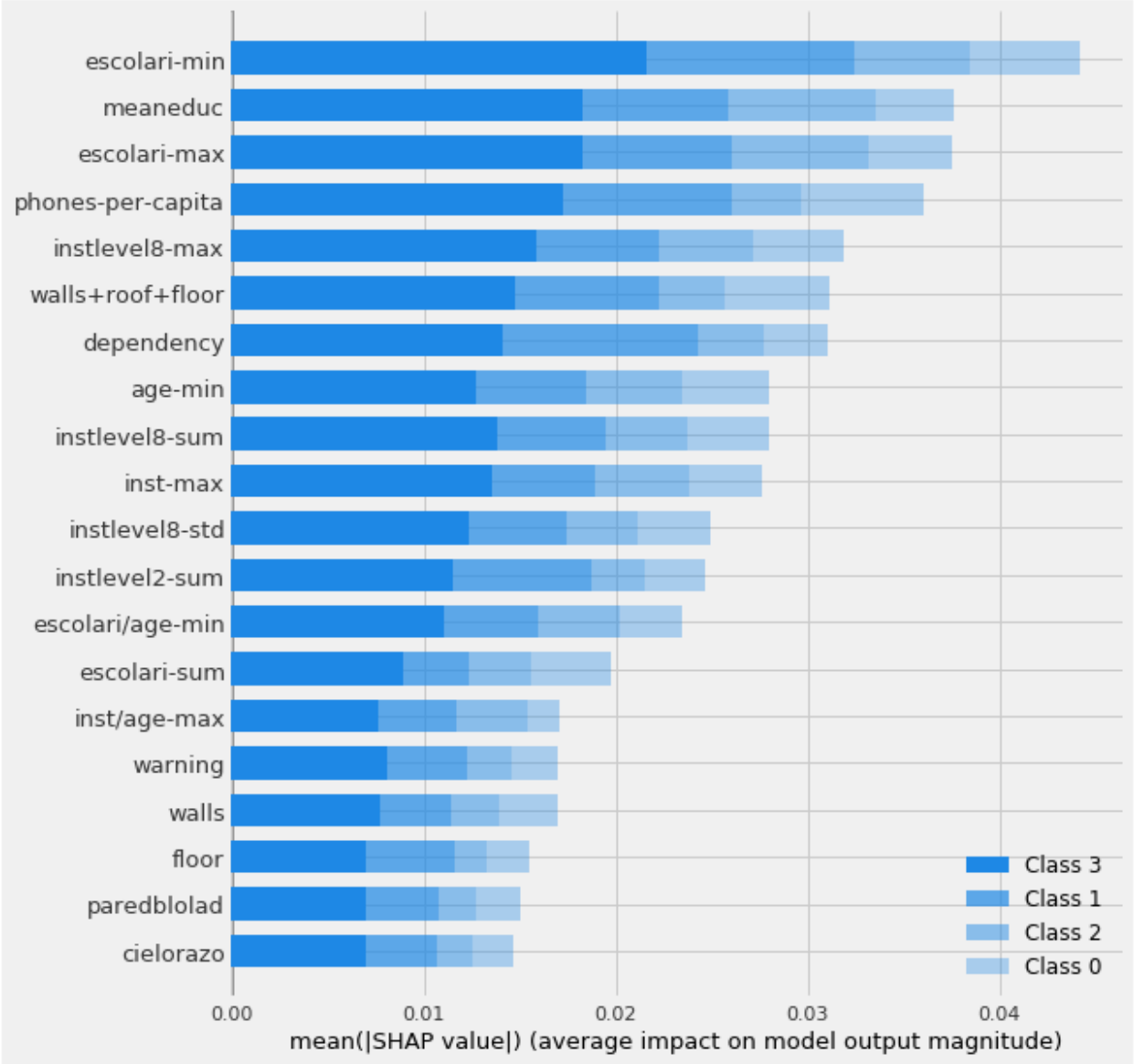
Socioeconomic Status Classification/ Age	**Senior (Privileged Group)**	**Junior (Unprivileged Group)**
Poor	64%	36%
Not Poor	79%	21%

Model Performance

	Accuracy	**Precision**	**Recall**	**FPR**	**FNR**
Senior (Privileged)	0.73	0.71	0.97	0.70	0.03
Junior (Unprivileged)	0.73	0.75	0.92	0.67	0.08

6. Transparency & Privacy

Since the features used by the model are obtained from household provided data, it has a high impact on the economic lives of individuals. Thus, giving Costa Ricans explainability and transparency in how the household survey questions are used to generate this model is very important. The following is a SHAPley Explanation plot of the top 20 features used in the ADS:



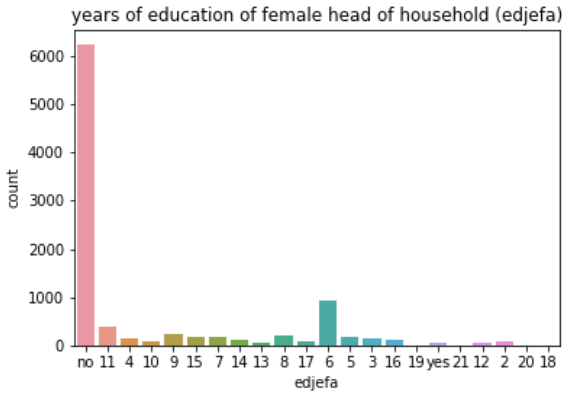
The following are the detailed definitions of the top 20 features:

Escolari-(min/max): years of schooling aggregated on each household
Escolari/age-min: minimum years of schooling / age
Age-min: Minimum age of household

Phones-per-capita: Number of mobile phones / number of persons living in the household
Instlevel8-(max/std): the max or standard deviation of the indicator value of undergraduate and higher education in each household
Inst-max: Max education level ranging from 0-8, which was the column index of each record, in each household
Inst/age-max: Maximum ratio of education level / age in each household
Instlevel2-sum: total number of incomplete primary education for each household
Meaneduc: average years of education for adults who are Costa Ricans aged 18 or older
Walls+roof+floor: Ranges from 0 to 6, and it measures the quality of the house's structure
Dependency: Dependency rate, calculated = (number of members of the household younger than 19 or older than 64)/(number of member of household between 19 and 64)
Warning: The sum quality of the house. It will be a negative value, with -1 point each for no toilet, electricity, floor, water service, and ceiling. The indicators of electricity and ceiling were deliberately set to 0 in the calculation.
Walls: indicator of column index based on wall quality
Floor: indicator of column index based on floor quality
Paredblolad: binary indicator of 0,1 if the predominant material on the outside wall is block or brick in the household
Cielorazo: binary indicator of 0,1 if the household has a roof

According to the SHAP plot, we can see that education level, owning a phone, having quality living space, the dependency rate, and the minimum age of a resident in a household are the most important features that provide the most weight in classifying whether a person is non-vulnerable. The minimum years of schooling of the entire household provides a strong indication if the family needs further social aid.

Furthermore, safe and responsible usage of this ADS requires implementation of autonomy, beneficence, and justice principles in ethics applied to each survey participant.
An example of a potential risk in the dataset is shown by using the following feature, 'edjefa':



After close inspection of the count distribution of ‘edjefa’, which represents the years of education of the head household, only 2 people in the training set have an ‘edjefa’ value of 2. Thus, privacy of Costa Ricans in the dataset remains a concern because of linkage attacks from using other features.

For example, we can further base off of the other feature descriptions: 'v2a1', 'hhsize', 'instlevel1', 'instlevel2', 'instlevel3', 'instlevel4', 'instlevel5', 'instlevel6', 'instlevel7', 'instlevel8', 'instlevel9', 'tipovivi1', 'tipovivi2', 'tipovivi3', 'tipovivi4', ' tipovivi5', 'lugar1', 'lugar2', 'lugar3', 'lugar4', 'lugar5', 'lugar6', 'area1', 'area2', 'male', 'female', 'age', 'parentesco1', 'estadocivil1', 'estadocivil2', 'estadocivil3', 'estadocivil4', 'estadocivil5', 'estadocivil6', 'estadocivil7', and 'Target' that both persons are from a Central region in Costa Rica, and live in urban areas.

One of them is a male aged 52 and the other is a female aged 43, and the female is the head of the household. Both seem to be part of a coupled union, and are non-vulnerable with a target value of 4. The male completed a high school education, whereas the female completed a postgraduate education. Given that the country is still developing, not many have that high of an educational background. They both own a fully paid house, and live in a household size of 2 so we can fully assume that they are a couple who live together. Thus, this shows the importance of differential privacy in order to protect individuals from being easily re-identified through adding noise to the data based on an epsilon budget.

7. Summary

We believe that this data was not appropriate for this ADS due to privacy concerns. The implementation does not show that it is robust in terms of fairness for certain protected attributes. We focused primarily on the disparate impact among different groupings based on each protected attribute because it would be essential for the PMT to accurately represent the households that generate the information. The stakeholders who are impacted by these measures are the IDB who want to boost socioeconomic development, and the impoverished Costa Ricans who need a boost in their livelihood. Optimizing on the false negative rate will allow more people who genuinely need aid receive it. As for the false positive rate, optimizing on this metric would allow the IDB to conserve more of its budget for extremely poor individuals.

From our analysis, for gender, even though both groups are close to an even distribution, there are more females who are poor compared to men in the same poverty class, and they are also more likely to be labeled as non poor when they aren’t due to a higher false negative rate. When we look at the disability status of individuals, there are more disabled who are classified as poor relative to the non-disabled who are also deemed poor. The model also predicts more accurately and has less false positives when the individuals are non-disabled. There is a lower percentage of urban people who are categorized as poor compared to the original percentage

composition of urban and rural populations. The ADS shows worse results on the rural population, especially on the false negative rate, so they would get less access to aid. Marital status was representative throughout both poor and non-poor groupings, and model performance was not impacted as much, although married people had slightly higher accuracy. Finally, for age, there is a disparate impact shown in the poor group, where more young people comprise it. The model does not perform better based on either senior or junior age group, but juniors are more likely to be classified wrongly as not needing economic assistance compared to their counterparts. We would not be comfortable deploying this ADS in the public sector without a preprocessing method such as Data Synthesizer on the household dataset due to a potential in linkage attacks.

One recommendation for the usage of this ADS is having one entity to process and validate the data of a roster of beneficiaries with constant updating from large inflows and outflows regarding household demographic structure or eligibility conditions. This will help ensure the accountability, transparency, efficiency, and equity of these social programs. In the United States, there is the usage of rigorous verification to improve target accuracy. It is also recommended to have well-designed manuals and adequate training for users if the entity is decentralized.

Another recommendation is for the data collection process, which needs to be carefully designed to outreach the poor, be cost-efficient, and have administrative feasibility. The quasi-exhaustive survey approach has the advantage of being cheaper to implement per interview. The on-demand applications approach favors dynamic, on-going registration as well as updating and re-certification because of an extensive network of welfare offices. However, the latter approach can also miss the poor who may be less informed and connected. A mixture of these methods can be a good way to maximize outreach. In addition, micro-area poverty maps can help guide design choices to provide localized poverty density information.

The IDB also needs to be aware that the PMT model may “penalize” households who recently enrolled their children in school under a scholarship program giving them an ineligibility in welfare assistance.

We also note that the fundamental set-up of the labeling is indecisive and unspecific because a household that has extreme poverty could be also vulnerable. Does vulnerable households indicate that they have the potential to be poor, but not there? The IDB never fully described their labeling intentions.

Multi-label classifications also have tradeoffs in any ADS. Things can potentially be done differently as a regression or classification problem through using three labels, such as poor, almost poor, well-established instead.

Other things to consider in the data provided for this ADS are occupational downgrades because of barriers in transferring qualifications or lack of formal employment status, even though they are well-educated. If the model learns that education is an important feature for the poverty level, such as this particular classifier used, this is something that is not addressed.

Another consideration are inclusion errors attributed to the underestimation of poverty level of new immigrants who are building up assets, and so they may eventually reach their native-born income levels.

Moreover, although a subpopulation may not impact the overall PMT model's accuracy, it is still an important group to keep in mind in terms of social equity. By applying these recommendations and taking in the considerations, policy design could be more effective for humanitarian crises.

8. References

1. Regalia, F., & Robles, M. (2005, December). *Social Assistance Poverty and Equity in the Dominican Republic*. Inter American Development Bank. Retrieved May 2022, from https://publications.iadb.org/publications/english/document/Social-Assistance-Poverty-and-Equity-in-the-Dominican-Republic.pdf

2. Social Protection & Labor team. (n.d.). *Measuring income and poverty using Proxy Means Tests*. World Bank Group. Retrieved from https://olc.worldbank.org/sites/default/files/1.pdf

3. Sims, W. (2020, May). *The Accuracy Of Proxy Means Tests For Immigrant Populations: A Case Study In Colombia*. Retrieved May 2022, from https://jpia.princeton.edu/news/accuracy-proxy-means-tests-immigrant-populations-case-study-colombia