RDS Project

10 May 2021 11:55pm

Name: Oviya Adhan, Mina Mohammadi

1. Background: general information about your chosen ADS

a. What is the purpose of this ADS? What are its stated goals?

The Inter-American Development Bank (IDB) gives need-based aid to Costa Rican families. In order to figure out who to distribute aid to, several Latin American social programs, including those in Costa Rica, utilize and develop an algorithm called the Proxy Means Test (PMT). PMT takes into consideration observable household attributes, such as the material of ceilings and walls of the family's shelter, to determine the family's level of need. Although this approach is an improvement from past algorithms, especially from traditional econometrics methods (ie demographic information or geographical location), even IDB itself believes there is a way to improve upon the PMT-based Automated Decision System (ADS) that they use, hence the Kaggle competition three years ago.

With this background, the ADS that we chose is part of a competition created by the Inter American Development Bank to find a more accurate ADS using new machine learning methods in order to determine at need families in Costa Rica to provide aid accordingly (loans, grants etc.). The ADS we chose implements a machine learning model in order to provide IDB with a more accurate ADS. Its goal is to distribute aid more efficiently taking into account factors that have previously been ignored like specific household attributes like material of their walls, ceilings and other related assets. These attributes better help the ADS and thus the policy makers to better directly target most at need groups and provide social goods responsibly. Stakeholders in this ADS that would especially benefit include the Costa Rican government and poor families in need.

b. If the ADS has multiple goals, explain any trade-offs that these goals may introduce.

The ADS here seems to have only one goal of understanding the most at need and at risk groups in poverty. The goal is to tune the model to the right variables to create the most accurate system to target the right people in poverty. However, there could be one potential trade off in this ADS because determining that one variable is more highly weighted than another variable, its real world consequence is that some families may miss out on aid because the algorithm makes their poverty indicator a lower weighted variable.

2. Input and output

a. Describe the data used by this ADS. How was this data collected or selected?

The data for this ADS is from the IDB who, in agreement with the Costa Rican government, uses Costa Rican census data. The IDB and the Costa Rican government have shared interests in awarding aid accurately to citizens and they share information with each other.

b. For each input feature, describe its datatype, give information on missing values and on the value distribution. Show pairwise correlations between features if appropriate. Run any other reasonable profiling of the input that you find interesting and appropriate.

Our data has 142 input variables and for that reason we used the .dtypes function to illustrate the data types into our code. SEE CODE.

There were 5 input variables with missing values listed here: v18q1 (number of tablets household owns) , rez_esc (years behind in school), meaneduc (average number of years of education), v2a1 (monthly rent payment), and SQBmeaned (squared mean years of education of adults in the household). We found these NaN values through the function .isna () as illustrated in THE CODE. We observed that many of these missing values are related to each other, (ie mean education and square mean education years). The possible explanations for the missing values of these input variables is as follows:

1. V18q1 is the number of tablets per household. There exists a separate input variable, v18q, that is a binary variable and indicates whether the household owns tablets or not. Therefore, for v18q1, if a household has no tablets, the value is automatically NaN rather than 0.
2. Rez_esc is the number of years behind in school, which only holds a float value when that particular individual is behind in school. If no one in that household is behind in school, the value is set to NaN.
3. Meaneduc is the average number of years of education for all adults of the household. The NaN values of meaneduc may be because not all adults necessarily had primary education, meaning that there may be NaN values for individual education, which would result in Nan for the average.
4. V2a1 is the monthly rent payment. The NaN values of this variable are most likely due to some of these households living in government housing, meaning a $0 rent payment.
5. SQBmeaned is the squared mean of years of education of adults in a household. This is directly related to meaneduc, so when meaneduc = NaN, SQBmeaned = NaN.

We decided to find the correlation between each floor and wall type and the output variable in order to see which materials had heavier positive or negative effects on how households were labelled at the end. We also did a correlation analysis between gender ratio in each household to the output variable to see if our guess that more females would mark a household as more vulnerable. As for wall material, the highest correlation with the poverty vulnerability output, was 'paredblolad', which was also a positive correlation at 0.9001 meaning that if a house of the household had walls made of block or brick, there was a far less chance that the household was labelled as impoverished or vulnerable. The only two materials with a negative correlation, which means a higher chance of being labelled impoverished/vulnerable, were 'paredes' and 'paredfibras', which are waste material and natural fiber respectively. As for floor material, the highest correlation with the poverty vulnerability output, was 'pisomoscer', which was also a positive correlation at 0.9765 meaning that if a house of the household had floors made of mosaic, ceramic, or terrazzo, the household had a strong chance of being labelled as non-vulnerable. The only two materials with negative correlation, which means a higher chance of being labelled impoverished/vulnerable, were 'pisonatur' and 'pisonotiene', which are natural material and no floor respectively. As for gender ratio, we found that there was a negative correlation b/w female ratio and output versus a positive correlation b/w male ratio and output. This means that a higher percentage of females in a household means that there is a higher chance that the household was labelled as vulnerable or in poverty compared to a household with a higher male ratio.

c. What is the output of the system (e.g., is it a class label, a score, a probability, or some other type of output), and how do we interpret it?

_______The output is an ordinal variable the level of impoverishment of a household. The possible scores given go from 1-4, 1 indicating extreme poverty, 2 indicating moderate poverty, 3 indicating vulnerable households, and 4 indicating vulnerable households. In other words, the lower the score, the more impoverished the household, the more likely IDB would provide aid to that household.

3. Implementation and validation: present your understanding of the code that implements the ADS.

a. Describe data cleaning and any other pre-processing

In the ADS we chose, there was limited data cleaning or pre-processing. There was a substitution of binary (1 and 0) for (yes and no) in order for the data to be better understood. He replaces the NaN values with their category's mean through the imput_na() function. He also dropped other values including unique identifiers, target variables, the unique identifier for household and the variable "elimbasu5" the pollution of a body of water through rubbish disposal. From there, he separated features as target and explanatory values

b. Give high-level information about the implementation of the system

The input of the ADS includes various variables including tangible features of the houses of the households, gender ratios of the households, education levels, etc. The ADS includes all input variables except for IDs. He trained a gradient boosted decision trees model. The model is then used to predict outcomes for the test data. The outcome predicted is a score between 1- 4 to indicate the poverty level of the household.

c. How was the ADS validated? How do we know that it meets its stated goal(s)?

The ADS was validated with these methods:

-prediction made with model
-accuracy of prediction model checked (91.42259414225941)
-created a confusion matrix heat map

The predictions made by the ADS have a pretty high accuracy at about ~91% suggesting that the ADS accurately labels households at their impoverishment level. However, the ADS model was overfitted and uses a grandfathered/deprecated method, making us have to create adjustments when testing accuracy for part 4. SEE PART 4 FOR REVISED MODEL. It meets its single stated goal(the goal is to tune the model to the right variables to create the most accurate system to target the right people in poverty).

4. Outcomes

a. Analyze the effectiveness (accuracy) of the ADS by comparing its performance across different subpopulations.

After making adjustments to the model, we found that the overall model accuracy is actually 77.2489539748954. Amongst subpopulations, we found that the model accuracy for females is about 76.6260162601626 and the model accuracy for males is 77.90948275862068. This would suggest that the model's performance is the same across the populations.

b. Select one or several fairness or diversity measures, justify your choice of these measures for the ADS in question, and quantify the fairness or diversity of this ADS.

Many of the fairness metrics cannot handle a multiclass format, since the ADS does not clean out variables of various types, thus a mix of datatypes still remain, rendering many of fairness metrics methods useless.

We decided to find the model accuracy for both males and females as not only would it double as a measure of effectiveness of the ADS, but it also works as a robust fairness measure, to see that amongst gender (2 very large subpopulations with historical issues and implications), the ADS performs the same suggesting that the accuracy is quantified as fair for both.

c. Develop additional methods for analyzing ADS performance: think about stability, robustness, performance on difficult or otherwise important examples (in the style of LIME), or any other property that you believe is important to check for this ADS.

Given the way that the ADS stands with so many codependent variables, LIME analysis would be essentially useless. Therefore, dimension reduction would need to be conducted. For now, the grading boosting feature ranking is sufficient enough to understand how the ADS is classified. From our feature ranking system, we find that meaneduc is the most important feature in analyzing ADS.

5. Summary

a. Do you believe that the data was appropriate for this ADS?

We would argue that this data is appropriate for the ADS. When looking at the feature selection, we see that a lot of historically overlooked features (overcrowding, possession of mobile phone, sizes of rooms etc.) in traditional econometrics methods, hold great importance in this ADS model. These data give the opportunity for the ADS to better target these impoverished areas and pay better attention to these economic indicators.

b. Do you believe the implementation is robust, accurate, and fair? Discuss your choice of accuracy and fairness measures, and explain which stakeholders may find these measures appropriate.

The implementation appears to be quite fair except for when it comes to seeing education levels. As seen with the feature importance ranking, mean education of adults in households is the most relevant variable when it comes to the XGB model. However, education level may just show a stronger correlation because higher education most likely leads to more income and therefore less impoverishment. Many of the variables are codependent, so there may be compounding effects when the variables are placed together.

However many other subgroups appear to have fair treatment. For example, we tested to see whether females are more likely to be labelled as impoverished, but the accuracy between the genders turned out to be quite evenly distributed. As for overall accuracy, the model is only around 77.248% accurate, but this is most likely far more accurate than the previous econometric-based PMT method.

When it comes to the robustness of our model, we would argue that our methods are not entirely robust, we would suggest that to better handle outliers and missing data, we would run more methods/improvements as we mention in section D.

Generally speaking, the stakeholders in this case, the Costa Rican government and impoverished groups still continue to benefit from the usage of this ADS as compared to pre-existing econometric methods.

c. Would you be comfortable deploying this ADS in the public sector, or in the industry? Why so or why not?

We would still suggest, even with our accuracy being around ~77% that this ADS is still of good use in the public sector. As compared with other more regional indicators of poverty, like province/city GDP etc. the data in this ADS better targets familial needs on a more micro level, giving opportunities for aid to be better directed. Current econometric methods do not provide these sorts of micro level data on familial structure, leaving a lot of impoverished families without aid.

d. What improvements do you recommend to the data collection, processing, or analysis methodology?

We recommend three suggestions for the original ADS:

1) Taking a random sample out of the data in order to train and test the data will make the model more efficient without losing much accuracy especially since outdated methods were used. This would help not only make the model faster, but also avoid running into the problem of overfitting. This would also save space.
2) During pre-processing, many of the binary variables can be consolidated into single, ordinal variables. In other words, dimension reduction can be heavily useful. For example, material of floors can be consolidated into one input variable rather than having 7 different variables for floor material alone. This would also mean ranking the materials from cheapest to more expensive in order to make it ordinal. This method would also help save space while running.
3) Use different models other than XGB. Since XBG is R-based rather than Python, several premade useful packages such as the sklearn metrics do not work. For simplicity sake as well as for heightened accuracy of the model, trying different models such as Random Forest can help.