Big Data and Machine Learning - Problem Set 3

Mariana Bonet De Vivero, Natalia Jaramillo Erazo
Luis C Ricaurte

31 de julio de 2023

Introduction

The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways. La introducción enuncia con claridad y precisión el problema, además explica los antecedentes, datos y pertinencia para responder a dicho problema. Contiene un resumen de los resultados y las principales conclusiones.

Poverty has historically been a central issue worldwide, affecting today individuals all over the world. Recently in Colombia, poverty has declined in urban areas but increased in rural areas, this after an overall spike in 2020 due to COVID-19. Over the years, institutions have tried to collect reliable data and generate research in order to inform on the challenges, lessons and priorities; however understanding poverty is a complex task. Poverty prediction is one of the first and essential steps in poverty reduction. Using machine learning models represents certain advantages to traditional econometric methods. Machine learning models are able to better capture complex and non linearity of poverty dynamics. In Colombia's case, machine learning models may capture the complex relationship between variables related to particular factors that have impacted poverty levels: notorious regional disparities, sustained armed conflict and violence, and high levels of internal displacement and informal economies.

With the objective of measuring poverty at the household level in Colombia, the present study looks to develop a predictive model that is easier, less time consuming and cheaper than traditional methods to measure poverty. We used four databases from DANE and the mission for the "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE". The data contains information at the household and individual level for 2018. Although there are more than 200 variables in the original databases, we selected 17 variables associated only with household property characteristics, individual sociodemographic characteristics, and poverty measures, to name a few: income, gender, education, property ownership, and estimated property price. The only difference between the train databases and the test databases is that the train databases has all variables while the test databases do not have information on certain variables, for example, income. Colombian poverty literature has shown that the variables selected are determinant in whether an individual is considered poor or not, thus we expect that these variables lead to completing our objective: creating the best poverty prediction model with the least amount of variables. More detailed information on the data will be described in the next section as well as a description of the pre-processing and cleaning methods that were used.

We found that to rapidly and cheaply measure poverty the model that includes the variables age, age-squared and sex, based on the databases from DANE and the mission for the "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE". Because of this, when building our model, we aimed to use the minimum number of variables taking into account the total income and the linear regression of age, sex and age-squared, based on a random forest spatial block cost-complexity pruning with bagging, which is slightly better than many other models we tried such as predicting poverty via trees, ada boost, logit and forests, which included the same variables as before but with a different

decision taking models. It is interesting nonetheless than most models have a very similar accuracy and that kaggle shows the same results for most of the models we uploaded, showing accuracy values of around 80 % and kaggle results of around 0.7881. Nonetheless, we consider this 80 % accuracy is a rather good estimate for poverty given the little amount of variables to predict we used to compensate for the cost of collecting poverty variables.

From this, our main takeaway includes that it would be interesting to add other variables such as if the household leases or owns a home, what would they have to pay, or what do they actually pay, what is the highest educational level, which department they are on, and how many rooms the house has, or to try other specifications of the model such as interactions between variables and quadratic independent variables to see if the accuracy becomes larger, but given that measuring poverty is hard, time consuming, and expensive and by building simpler models with a good enough accuracy, we can run surveys with fewer, more targeted questions that rapidly and cheaply measure the effectiveness of new policies and interventions, we consider our model is good enough to predict poverty.

All the information used for this prediction can be accessed through this link to the GitHub Repository.

Data

a. Describe the adequacy of the data to solve the predictive question, the sample construction process, including how the data was cleaned, combined, and how new variables were created.

Data from DANE and the mission for the "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE" is made up of household and individual level surveys applied all over Colombia to try to understand monetary poverty and inequality. The data is divided into four databases, training and testing at household and individual level for 2018. Since the data from individuals have an identification for household, the database is suitable to solve the predictive question once collapsed to be merged with the household data. Nonetheless, some variables are missing in the testing data bases which add an additional level of complexity to the problem. To solve this, we decided to only keep those variables relevant to the estimation of poverty, dropping many that where filled with NAs and thus would not add much to the predictive capacity of the model. Once this was made, we created columns of missing values for the variables that were missing in the testing databases to be then predicted using machine learning, such as *pobreza, indigencia* and *ingreso total*. We also adjusted the total income dividing the sum of the total income at individual level per household by the number of individual per household to get the income per person, and chose to use the representative household leader's age as the one to keep. Finally, we merged variable P5140 into P5130, standing for the amount the household would have to pay for rent if the home wasn't theirs, with how much they actually pay for rent when the home is not theirs, since we noticed that the NAs in both columns were reciprocal to one another.

The original data bases were composed of 23 variables for train_hogares, 16 for test_hogares, 135 for train_personas and 63 for test_personas. The training individual data had 543,109 observations, the testing individual data 291,644, the training household data 164,960 and finally the testing household 66,168. As it was explained before, we combined both databases by aggregating firstly the individual training and testing database by id for total income and the orden variable which represents the number of people per household. With this we divided total income by number of members of the family to get income per person. We also chose the family's head age, sex and highest educational level since more than 60 % of income is brought by the the head of the family. With this variables created we left joined the individual dataframes into a bigger one, and then left joined again this new data base with the training and testing household data since after the transformation they have the same dimension.

To clean the data, we kept only variables that had reasonable number of missing values, dropping those that either didn't make much sense to predict poverty, or those who had many missing values

since they wouldn't add much to the predictive capacity of the model. We then replaced the missings with the mode in *Ingtot* in the training database, since around 17 % of them were NAs. We also changed the sex variable (*P6020*) from 2 to 0 to identify women, to have a dummy variable with 0 and 1s. We did the same procedure for the *Clase* variable from 2 to 0 resto", all the other municipalities in Colombia that are not the main city. We then adjusted all the categorical variables for R to take them as factors, and we checked for imbalance in the data, finding that 79.98 % of the sample is not poor, while around 20 % is.

Finally, prior to running particular models, we chose to escalate the continuous variables (*age, rent, poverty and indigence line, and total income*) so that they could have a zero mean and normal variance.

b. Include a descriptive analysis of the data. At a minimum, you should include a descriptive statistics table with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justi- fication for your data choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.

To calculate aggregated household data we collapse information at the individual-level data and use the characteristic of household head as representative family member. The evidence shows that the household head on average contributes more than 60 % of household income. In this sense it is reasonable to think that the household income is determined by the characteristics of this person (see figure 1). Thus we were able to retrieve information like sex, age, educational level to add it as a household characteristic.

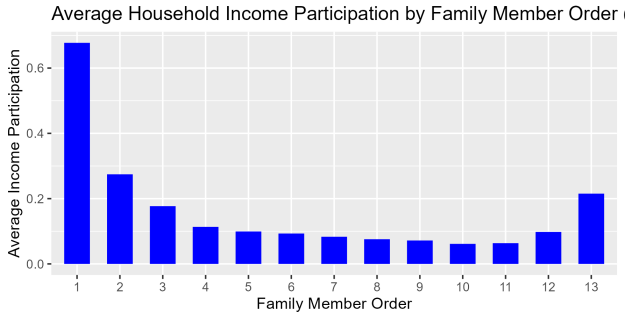

Figura 1: Family Members Income Contribution

With the completed database we then started the cleaning process. The following table shows descriptive statistic for some numerical variables:

Cuadro 1: Numerical Data

	Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
1	Homelessness Line	231128	120337	7267	99545	120001	123716	131126
2	Poverty Line	231128	271368	34003	167222	275482	285767	303817
3	Age	231127	50	16	11	37	61	108
4	Total Income (ingtot)	164960	779121	1166450	0	260583	879291	85833333
5	Rent Estimate	231128	472793	3814139	20	200000	500000	800000000

Table number 1 provides an overview of the central tendency (mean), variability (standard deviation), and distribution of each variable, also it shows minimum and maximum values, along with percentiles (25th and 75th) to understand the data's spread and distribution. We can see that the homeless line

and the poverty line show a robust coefficient of variation given the relationship between their standard deviation and the mean. in the first case this relationship is 6 % while the second is 12.5 %. For the age variable, this relationship is 32 %, however, for the total income and estimated income variables, this same relationship presents a very high variation, which suggests the need to leverage the data used to avoid biases, induced by its measurements or data transformation

Cuadro 2: Categorical Data

	Variable	N	percentage
6	Sex	231127	
7	... No	96525	42 %
8	... Yes	134602	58 %
9	Poor	164960	
10	... No	131936	80 %
11	... Yes	33024	20 %
12	Homelessness	164960	
13	... No	157000	95 %
14	... Yes	7960	5 %
15	Home ownership	231128	
16	... Own, fully paid	87355	38 %
17	... Own, they are paying	7764	3 %
18	... for rent or sublease	89654	39 %
19	... In usufruct	35259	15 %
20	... possession without title	10895	5 %
21	... Other	201	0 %
22	Education Level	231127	
23	... None	12271	5 %
24	... Preschool	18	0 %
25	... Primary school (1st - 5th)	65638	28 %
26	... Secondary school (6th - 9th)	30390	13 %
27	... High School (10th -13o)	60055	26 %
28	... Higher or university	62726	27 %
29	... Does not know, does not inform	29	0 %
30	House Rooms	231128	
31	... 1	15291	7 %
32	... 2	30735	13 %
33	... 3	78554	34 %
34	... 4	73756	32 %
35	... 5	23490	10 %
36	... 6	6672	3 %
37	... 7	1824	1 %

The table number 2 shows descriptive statistics for some of the most interesting categorical variables (including individual-level data and household-level data). The table provides information about the count and percentage distribution of each category within the categorical variables. It includes variables related to Sex, poverty status Poor, homelessness, home ownership, education level of household head, and the number of household rooms. Also shows the number of observations for each variable.

The table allows us to understand the distribution of categories within each variable, providing insights into the composition of the dataset for the specific categorical variables analyzed. So, we can say that almost half (48 %) of the heads of households are women. and as noted above, 20 % of households are poor while 5 % are below the poverty line. Regarding home ownership, 38 % of families are owners, however, a slightly higher percentage (39 %) represents households that live on lease. Another interesting statistic is the distribution by educational level, where a similarity can be observed between Primary

school (28 %), High School (29 %) and Higher or university (27 %).

Model and Results

For our predictive task we used different specifications and methods. We divided the models into two groups: those that directly used a classification approach to predict poverty and those that used an income prediction approach to first estimate income and later indirectly predict poverty. It is important to note that since our restriction was to use the least amount of variables possible, we decided to use three to train all of the models: gender, age and age squared. Throughout all the models, we were sure to convert gender to a factor in order for R to correctly understand the values of the variables and avoid any class issues. As mentioned before, age is a continuous variable, ranging from 11 to 108 years old, and we decided to explore the non-linearity of this variable. What truly varied were the methods used for the predictions, which will be described below. A summary of the models and their characteristics can be seen in Cuadro 3.

Classification models.
We used four classification models: a logit classification, an ada boost classification, a tree classification and a random forest classification. For the first model, the logit classification, we used the train function and the train data, specified the variable "pobre.ªs the dependent variable and the three variables discussed above as the independent variables. We chose the glmnet method, the Generalized Linear Model with Elastic Net Regularization. After predicting out sample, we set a classification rule where 0.5 is the threshold for the classification probabilities. In this sense, The predicted poverty status in the test data was based on whether the predicted probability of poverty was greater than the rule or not. If the probability resulted greater than the rule, the poverty status was set to 1; otherwise, it was set to 0.

Next, we performed an AdaBoost classification with the same model specification as before. We set up the configuration for the Ada Boost classification using the trainControl function, specifically cross validation with 5 folds and the class probabilities were retained. Additionally, the three parameters set for tuning were 50, 100, and 150 as the number of boosting iterations; 1, 2 and 3 as the maximum depth of the tree; and Breiman, the logistic approach, and Freund, the exponential approach, as the boosting coefficient. These are the parameters that will be optimized during training to find the best combination. To train the model we used the train function, with the same dependent and independent variables as before. Finally, we predicted the outsample poverty classification.

Our third model followed a Decision Tree classification. In this model, we set up the configuration using the trainControl as well, using a cross-validation approach with 5 folds and class probabilities retained. We used the rpart method to train the model, with the same dependent and independent variables, and a tuneLength set to 100, indicating the algorithm to explore different hyperparameters. Finally, we predicted the outsample and obtained our results for poverty prediction.

The last model of this section employed a Random Forest algorithm. The configuration set up was the same as the Decision Tree model and the AdaBoost model, cross-validation with 5 folds and class probabilities retained. We trained the model using the ranger function, specified the same dependent and independent variables, and set three different parameters for the algorithm to combine and optimize: 8 variables randomly sampled as candidates at each split, the gini as the splitting rule for the nodes, and 15, 30, 45, and 60 as the minimum size of terminal nodes. After being trained, the model predicted the outsample poverty classification using the test data.

Income regression models.

We estimated three income regression models in order to indirectly predict poverty: a logit regression, a random forest regression, and a continous tree regression. For the first model, the logit regression, we

used the train function and the train data, specified the variable ïngtot.ªs the dependent variable and the three variables discussed above as the independent variables. We made sure to remove rows with zeros as ingtot to avoid problems with the regression. We chose the glmnet method, the Generalized Linear Model with Elastic Net Regularization, and after predicting the income both in sample and out sample, we constructed the dummy "pobre". If the income is below the variable "lp", the variable is set to 1; otherwise, it is set to 0.

Our next model, which is also our final model, uses the Random Forest Algorithm to predict the income based on three independent variables: gender, age, and age squared. In this model, we used cross-validation with 10 folds and train the model with the ranger function. The model was tuned with different values of hyperparameters: 1, 2 and 3 as the number of predictors randomly sampled as candidates at each split; the variance as the rule for node splitting; and 5, 10, and 15 as the minimum number of observations in a node. After the outsample prediction was carried out, the classification process was the same as the logit regression before. We constructed the dummy "pobre"that is 1 if the income is below the variable "lp", and 0 otherwise.

Finally, we used a continuous tree or Decision Tree regression with the same specification as before. We also used cross-validation, this time with 5 folds, used the function rpart to train our model and adjusted the tuning of the model to 100, which indicates different values of hyperparameters should be used during cross-validation. After making the income predictions, the procedure was the same as the two other models in this section.

The table below shows the summary of all the models used in the study. As can be seen, the accuracy among models remained relatively similar, except the last model, Tree Continuous. Our chosen final model has an accuracy of 0.8049 which is relatively high and indicates the model has a modest performance.

Cuadro 3: Model Comparison

	Model	Method	Variables	Accuracy
1	Model 1	Logit Regression	3	0.8049
2	Model 2	Logit Classification	3	0.7998
3	Model 3	Ada Boost Classification	3	0.8048
4	Model 4	Tree Classification	3	0.7998
5	Model 5	Spatial Block Cost Complexity Prunning - Bagging	3	0.8049
7	Model 7	Bosque Classification	3	0.7998
8	Model 8	Tree Continuous	3	0.0000

Conclusions and recommendations

In this section, you briefly state the main takeaways of your work.

Based on our findings, a very simple model such as one that only includes age, age squared and sex has a relatively high predictive capacity in terms of accuracy, reaching over 80 %. All in all, our predictions with different model specifications and estimating practices showed similar variables for accuracy, with 0.8049 and 0.7998. Particularly among the highest are models 1, 3, and 5 models, which in kaggle submissions all of them have a score of 0.78810. Therefore, given the capacity to deal with more

complex models, the model we chose as the the best of all was v5, which was trained via spatial block cost complexity pruning and bagging. This model took into account the variables age, age-squared and sex.

The advantage of this model versus the other ones we built and tested was that the spatial blocks allowed for correcting for potential correlation, while cost complexity pruning allowed us to build a tree with all its branches and then prevent over-fitting the regression tree by removing some of the leaves until we find the version of the pruned tree with the lowest sum of square residuals. Finally, bagging was chosen to reduce variance and improve overall predictive performance. The result of the model was an accuracy of 0.80489, which means that the predictive model accurately estimates poverty in 80 % of the sample.

Because of this accuracy, our main takeaway include even though the model could become even better by including other variables such as if the household leases or owns a home, what would they have to pay, or what do they actually pay, what is the highest educational level, which department they are on, and how many rooms the house has, or to try other specifications of the model such as interactions between variables and quadratic independent variables to see if the accuracy becomes larger.

Therefore, even if it would be interesting to include into the model additional variables, measuring poverty is hard, time consuming, and expensive and by building simpler models with a good enough accuracy, we can run surveys with fewer, more targeted questions that rapidly and cheaply measure the effectiveness of new policies and interventions. Because of this, we consider our model with the accuracy described above, successfully fulfills the objective of the study, to predict poverty in Colombia while spending as little as possible, by building an accurate model with the least amount of variables, so that we can target interventions and iterate on policies, maximizing the impact and cost-effectiveness of strategies.

Bibliography

(2021, December 13). InsideBigData - Your Source for AI, Data Science, Deep Learning Machine Learning Strategies. The $500mm+ Debacle at Zillow Offers – What Went Wrong with the AI Models?. Retrieved from http://insidebigdata.com/2021/12/13/the-500mm-debacle-at-zillow-offers-what-went-wrong-with-the-ai-models/