Costa Rican Household Poverty Prediction

COMP3354 Statistical Learning Group Project Report

Anushka Vashishtha, Oka Courtney Julin, Rohak Singhal,
Aman Johar, Tarun Sudhams, Varun Vamsi Saripalli

University of Hong Kong

3rd December 2018

Abstract - **It is extremely difficult for social programs such as this to gauge the right amount of aid that needs to be given to the right people. This problem is made exponentially more difficult when that program is dealing with the least fortunate portion of the population. This is because they cannot provide the necessary details of their income, asset or expense records to justify that they need the aid to qualify.**

Hence, this paper's defining question is: how to determine a method to effectively gauge the right amount of aid to be given to each household given the multitude of variables present in the vast dataset?

MOTIVATION

The income inequality rate in Costa Rica has been steadily increasing from 2000 with the poorest 20 percent receiving an income which is around 19 times lower than the richest 20 percent. There are more than 20 government run social spending programs, yet the poverty rates have not been significantly reduced. The extreme poverty rate has also increased to 7.2 percent from around 5.8 percent in recent times. To tackle the issue of growing poverty, the government launched a plan in 2015 called Bridge to Development. The aims of this program are to uphold human rights, to reduce the levels of poverty, to link all the different social welfare programs under one administration. Government funds were not being allocated properly because of confusion caused by having too many related but different welfare programs.

To ensure the success of this program and to execute the program in an effective manner, the Mixed Social Welfare Institute came up with a customized Plan of Family Intervention(PIF) to record each family's needs and commitments which was laid out in 4 phases which included:

- Phase 1 – Family Eligibility
- Phase 2 – PIF Formulation
- Phase 3 – PIF Intervention
- Phase 4 – Family Exit

The goals of this project are to help identify what factors are best correlated with poverty. This will be useful in a few ways. Firstly, understanding what factors correlated with poverty will lead to understanding the circumstances around how families become impoverished. While future research will need to be done as to the causality of the factors, it is an insightful starting point. This could help with the allocation of funds to the proper kinds of welfare programs.

This is a problem which needs to be solved as it can have major implications on the economy of the country and can help in bettering the lives of more than a million people. The potential impact of this project is that this analysis can be further implemented to analyze the same problems in other Latin American countries and improve the economic conditions of the country as a whole by uplifting the poorest segments and giving them access to a better life.

METHODS

The following paragraphs would go over the entire analysis process which includes, data cleansing, feature engineering, machine learning modelling and finally using gradient boosting to see any improvements over the outcome achieved by using modelling. The target of our analysis would be to predict poverty on a household level as defined in the dataset.

Given that dataset is on an individual level in terms of the data points, however we will only include the head of the

household to stick with our plan of conducting the analysis on a household basis.

The following would be the target variable values:

- 1 = extreme poverty
- 2 = moderate poverty
- 3 = vulnerable households
- 4 = non vulnerable households

Ultimately we want to build a machine learning model that can predict the integer poverty level of a household. Our predictions will be assessed by the Macro F1 Score.

I. Data Source

The dataset used in the project originated from the website Kaggle, which is an online community for data scientists to participate in machine learning competitions. Kaggle's public datasets are not that transparent in how the data is collected, but it is a company that is currently owned by Google, Inc., which deals with vast amounts of data and has a reputation of being fiercely data driven

II. Variables of Interest

TABLE I
VARIABLES OF INTEREST

Variable Name	Description
idhogar	HOUSEHOLD LEVEL IDENTIFIER
v2a1	MONTHLY RENT
hacdor	OVERCROWDING BY BEDROOMS
v14a	BATHROOM IN THE HOUSEHOLD
v18q	OWNS A TABLET
r4h1	MALES YOUNGER THAN 12 YEARS
r4m1	FEMALES YOUNGER THAN 12 YEARS
parentesco1	HOUSEHOLD HEAD
instlevel	LEVEL OF EDUCATION
refrig	HOUSEHOLD HAS REFRIGERATOR
cielorazo	HOUSE HAS CEILING
sanitario1	NO TOILET IN DWELLING
qmobilephone	NO OF MOBILE PHONES
age	AGE IN YEARS

Over the course of the methods, certain components will be deemed less important, but to start with, we are equally interested in what all these variables might tell us about poverty in Costa Rica. Therefore, to dig deeper, we did exploratory data analysis.

III. Exploratory Data Analysis

The next aspect of the exploratory data analysis focused on two parameters which are *males younger than 12* as well as *males older than 12*. We tried establishing a relationship on how these two parameters are distributed among various household types ranging from extreme poverty to non-vulnerable.

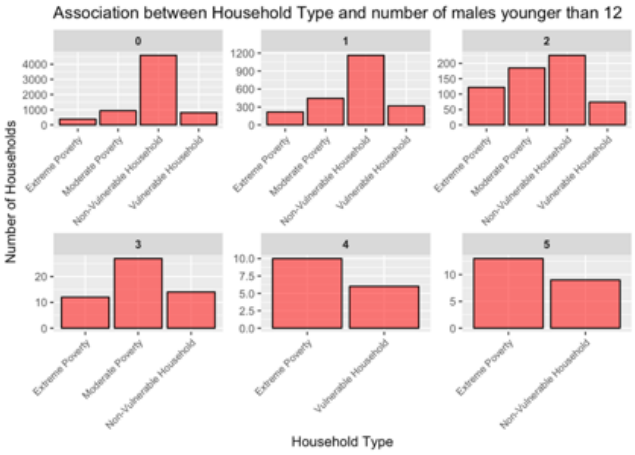

Figure 1: Males younger than 12 distributions with household type

First of all, we studied how many households contain what number of males younger than 12 as illustrated in Figure 1, and the results indicated that households with higher number of males younger than 12 tend to be either in extreme poverty or are in a non-vulnerable state.

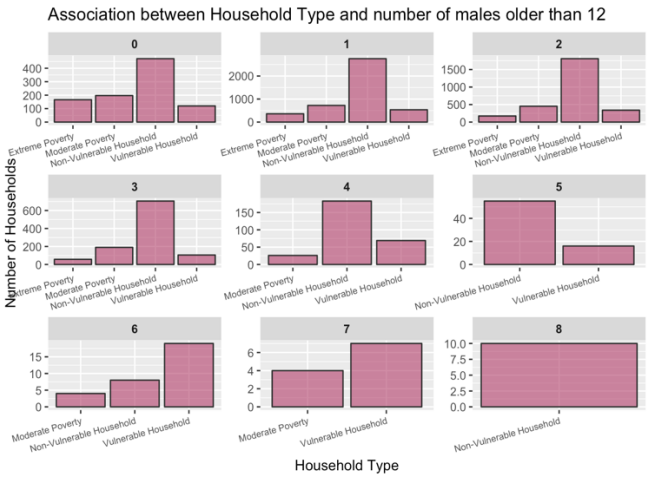

Figure 2: Males older than 12 distributions with household type

Next, for males older than 12, as illustrated in Figure 2, the conclusions were quite different. According to the analysis, we found that if the number of males older than 12 in a household is large, then that household tends not to be in extreme poverty.

After performing the same procedure on parameters *females younger than 12* and *females older than 12*, we found that if higher numbers of *females younger than 12* are present, then the household is most probably a moderate poverty household, while on the other hand, if higher number of females older than 12 is present, then the household is most probably a non-vulnerable household. An explanation which can be given to such an outcome is that males and females older than 12 are more capable in contributing to the family income thereby alleviating poverty significantly.

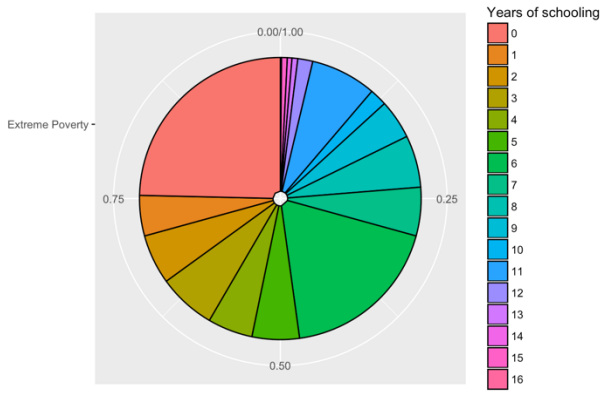

Figure 3: Years of Schooling in extreme poverty households

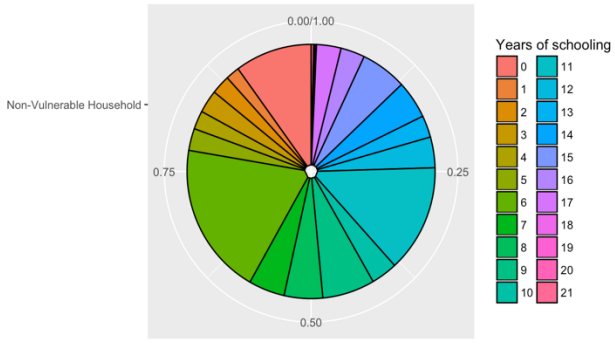

Figure 4: Years of Schooling in non-vulnerable households

The last feature which we concentrated on is the *years of schooling*. After doing a detailed analysis of the distribution of number of years of school done by people belonging to each household type, we came to the conclusion that a large percentage of people belonging to an extreme poverty household have no schooling and this percentage decreased as we moved on to a moderate poverty household, then to a vulnerable household and lastly to a non-vulnerable household. An explanation for such a result can be that people belonging to an extreme poverty household, generally lack in resources which can support their education.

IV. Data Cleaning

The next step after EDA involved cleansing the data to make it ready for the application of various statistical models which involved two steps.

First, we applied a common target label to each member of a household. As described in the initial background, we concluded that the analysis of only the head of a household is done in order to predict the target label for that particular household. Consequently, each member of a household should have the same target label as that of the head as they all live in the same household. This should also hold if there is no head for a particular household. Henceforth, we went through the training data and applied the correct target label to those members of a household, for whom the condition did not hold.

The second step to clean the data, involved filling up missing values. To achieve this, we narrowed down the features like *number of tablets*, *monthly rent payment* and *years behind in school* as they were the ones with missing values. For the *number of tablets* feature, all those families which do not own a tablet have *NaN* as a predefined value. We replaced it with 0. Following this, we moved on to the *monthly rent payment* feature and here we put 0 as a value in a missing value row, in case the respective household owns the property. For the other households, we can leave the missing values to be imputed later, but for now we added a flag (Boolean) column indicating that these households have missing values. Lastly, we moved on to the feature *years behind in school* and for this we first identified that this variable is defined for the age group of 7-19 year olds only. Hence, in case of missing values, we put 0 as a value for those data points which do not fall into the 7-19 age group and otherwise we flagged it.

V. Feature Engineering

a. *Removing Squared Variables:* There are some squared variables included in the dataset to help with feature engineering in case of nonlinear data modelling. But since we are using more complex models, these squared variables are redundant and are removed. Including these variables in the features may hurt the result as they are highly correlated with their non-squared version.

b. *Remove highly correlated household and individual variables:* We saw the correlations between different variables to see which ones are highly correlated. If there are any that are too highly correlated, then we might want to remove one of the pair of highly correlated variables. We considered pairs with correlation more than 0.95 as fit for removing.

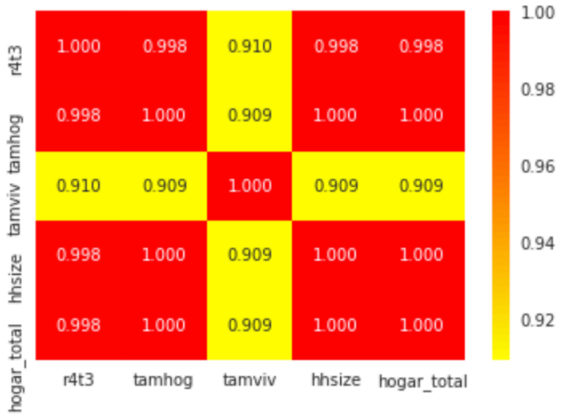

Figure 5: Heat map of highly correlated household variables

For example, the above figure represents a heat map of highly correlated household variables. We can see that *hhsize* is almost entirely correlated with *tamhog*, *r4t3* and *hogar_total* therefore we keep hhsize and remove the other variables.

c. *Creating Ordinal Variable:* We tried to construct some ordinal variables by compressing several related Boolean variables. We mapped the description based on the data description given to us in the problem. For example, in case of electricity: -

- 0: No electricity
- 1: Electricity from cooperative
- 2: Electricity from CNFL, ICA, ESPH/JASEC
- 2: Electricity from *privateplant3*

An ordered variable has an inherent ordering, and for this we choose our own based on the domain knowledge. After we create this new ordered variable, we can drop the four others.

d. *Feature Aggregation:* In order to incorporate the individual data into the household data, we need to aggregate it for each household. The simplest way to do this is to group by the *family id*, *idhogar* (family ID) and then aggregate the data.

VI. Statistical Learning Modelling

In order to find out what are the most important features of the dataset, a simple baseline Random Forest Classifier model will be used. Upon building the model, the most important features of the model were analyzed and the F1 score for this model was 0.352 which is rather quire poor. However, this model gives us a baseline to improve our newer models.

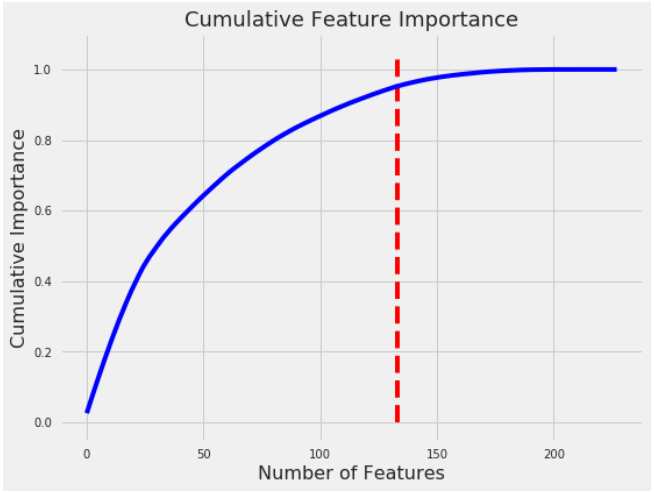

Figure 6: Cumulative importance of features

Figure 7 shows us that, in the dataset, about 133 features correlates to 95% of the data available. So our modelling efforts would be based on these features to extract the maximum information from the data.

As the Random Forest Classifier that we built didn't perform very well, we decided to build more classifier models to understand and analyze the how bad was our current classifier model and how can we make it better.

In order to do that, we built 6 other models to compare their performance with each other. Those classifiers are:

- Linear Support Vector Classifier
- Multilayer Perceptron
- Linear Discriminant Analysis
- RIDGE Classifier
- KNN with 5, 10 and 20 neighbors
- Extra Trees Classifier

The graph in figure 7 gives us a comparative view of all the models that we built for the specific problem in terms of F1 score.

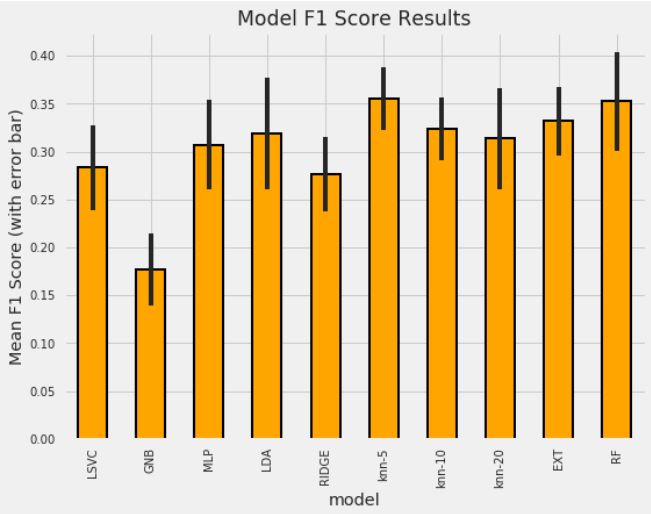

Figure 7: Comparative view of F1 score

To our surprise, we found that the model we considered to be the worst performing and baseline model, turned out to be the best model out of all the models we tried. This helped us conclude that modelling is not a good approach to this problem and hence we needed ways to either fix the current method or adopt a new method of analysis. One possible attempt would be to make us of hyper-optimization of parameters. However, the optimization so conducted would not help us gain any significant upgrade in the F1 score. This means that we would have to change our approach from modelling to gradient boosting to see what treasures our data is hiding and then we can see how boosting machines and the models that we built compare with each other, again in terms of F1 score.

V. Gradient Boosting Machine

The Random Forest Classifier trained on our dataset achieved the highest out of sample accuracy out of all the models - about 35%. In order to further increase this accuracy, we explored ensemble learning methods such as Gradient Boosting Machine (GBM), which are known to greatly improve the performance of weak learners.

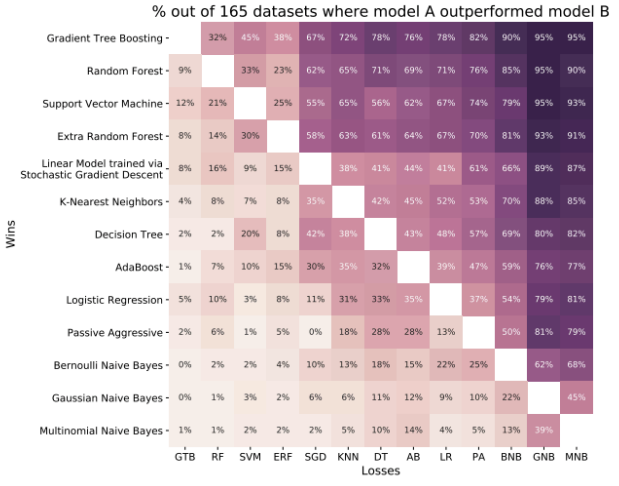

Fig. 2. Heat map showing the percentage out of 165 datasets a given algorithm outperforms another algorithm in terms of best accuracy on a problem. The algorithms are ordered from top to bottom based on their overall performance on all problems. Two algorithms are considered to have the same performance on a problem if they achieved an accuracy within 1% of each other.

Figure 8: Data-driven advice for applying machine learning to bioinformatics problems. Randal S. Olson, William La Cava, Zairah Mustahsan, Akshay Varik, and Jason H. Moore

The figure above (Olsen R. et al.) demonstrates how GBM compares against 12 other commonly-used statistical learning methods. When these 13 methods were used to train on 165 datasets, GBM outperformed majority of the methods, most of the times. GBM is a form of Ensemble Learning, which uses weak learners together in a certain order to improve the accuracy of the model. In contrast to ordinary statistical learning, where models try to learn one hypothesis from training data, ensemble methods construct a set of hypotheses and combine them. (Zhou, 2009).

In specific, GBM is a form of Boosting. In boosting, weak learners are combined sequentially to compensate the shortcomings of existing weak learners in a certain way. Since GBM is a form of boosting (combines models sequentially), it tries to identify shortcomings in a model by computing the error residual.

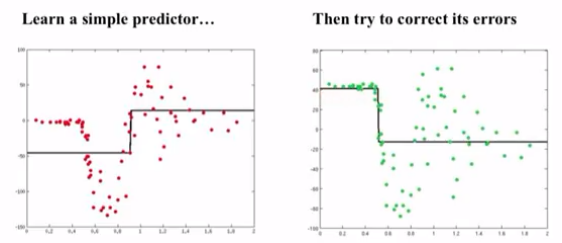

Figure 9: (a) and (b)

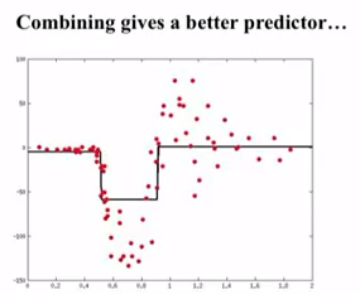

Figure 9: (c)

The figure above shows the three main steps of the algorithm. A weak model is fitted on the set of observations (part a). As the model poorly predicts the function, error residuals are calculated, which are actual values - predicted values (part b). A new weak model is fitted on these residuals and combined with the old model, to get a better fitting on the original dataset (part c).

Thus, the algorithm:

- Learn a predictor
- Compute the error residual
- Learn to predict the residual
- Combine predictors, back to step 2

Consider observations (x1, y1),(x2, y2), ...,(xn, yn), model F(x) to be fit on the observations to minimize square loss. For GBM, we wish to combine models sequentially so that the overall model fits better (minimum square loss).

Thus, we wish to achieve:

$$F(x_1) + h(x_1) = y_1$$
$$F(x_2) + h(x_2) = y_2$$
$$\dots$$
$$F(x_n) + h(x_n) = y_n$$

Or, equivalently:

$$h(x_1) = y_1 - F(x_1)$$
$$h(x_2) = y_2 - F(x_2)$$
$$\dots$$
$$h(x_n) = y_n - F(x_n)$$

yi − F(xi) are called residuals. The role of h is to compensate the shortcoming of existing model F. The goal is to keep adding weak learners until F+h becomes a satisfactory model. In Gradient Descent, we minimize a function by moving in the opposite direction of the gradient:

$$\theta_i := \theta_i - \rho \frac{\partial J}{\partial \theta_i}$$

Our loss function:

$$L(y, F(x)) = (y - F(x))^2/2$$

And, we want to minimize:

$$J = \sum_i L(y_i, F(x_i))$$

by adjusting F(x1), F(x2), ..., F(xn).

Taking derivative w.r.t F(xi):

$$\frac{\partial J}{\partial F(x_i)} = \frac{\partial \sum_i L(y_i, F(x_i))}{\partial F(x_i)} = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} = F(x_i) - y_i$$

$$y_i - F(x_i) = -\frac{\partial J}{\partial F(x_i)}$$

Thus, we can interpret residuals as **negative gradients**.

Since the residuals are actually negative gradients, fitting h to residuals implies that we are fitting h to negative gradient and updating F based on the residuals implies we are

updating F based on negative gradients. Thus, the model is being updated using **Gradient Descent**.

Our Gradient Boosting model is thus using Boosting as an ensemble learning method on Random Forests (weak learners) and combining these learners using Gradient Descent.

```
% Data set (X, Y)
mu = mean(Y); % Constant "mean" predictor
dY = Y - mu; % subtract this prediction away

For k=1:Nboost,
 Learner{k} = Train_Regressor(X,dY); % Can be any weak learner
 alpha(k) = 1;  % alpha: a "learning rate"

 % compute the residual given our new prediction
 dY = dY - alpha(k) * predict(Learner{k}, X)
end;

% Test data Xtest
[Ntest,D] = size(Xtest);
predict = zeros(Ntest,1);

For k=1:Nboost,                    % Predict with each learner
 predict = predict + alpha(k)*predict(Learner{k}, Xtest);
end;
```

Figure 10: Pseudocode for the algorithm

The above figure compares the performance of GBM on our dataset with other methods. GBM clearly outperforms previous methods by a great margin, with an accuracy of about 43%! The performance can be further increased in future work using model optimization.

RESULTS

There 2973 training labels we used for our models after all the data cleaning described earlier. The distribution of labels for the training data is represented in the table next to figure 6 showing the distribution. Where 4 is "non-vulnerable," 3 is "vulnerable," 2 is "moderate," and 1 is "extreme."

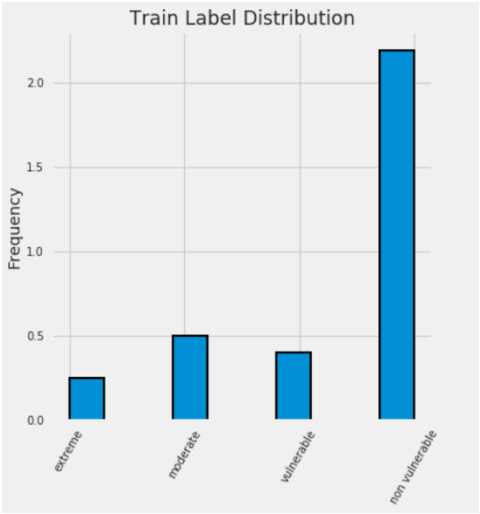

Figure 11: Train Label Distribution

TABLE 2
DISTRIBUTION OF LABELS

Label Values	Distribution
4	1954
3	442
2	355
1	222

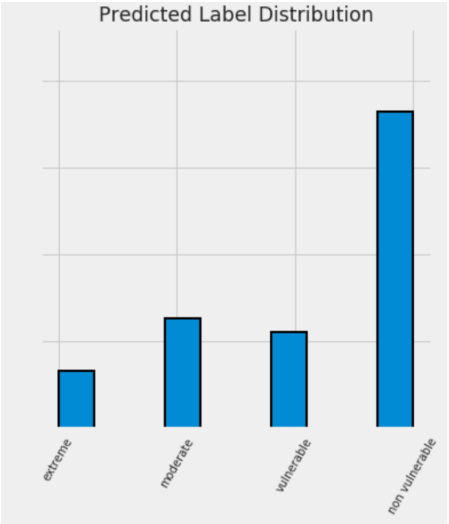

Figure 12: Predicted Label Distribution

As a first attempt at looking into our model, we can visualize the distribution of predicted labels on the test data. We would expect these to show the same distribution as on the training data. Since we are concerned with household predictions, we'll look at only the predictions for each house and compare with that in the training data. The above histograms in figure 6 and 7 are normalize meaning that they show the relative frequency instead of the absolute counts. This is necessary because the raw counts differ in the training and testing data. Furthermore, the predicted distribution looks close to the training distribution although there are some differences

Now, after normalizing our confusion matrix (figure 8) we derived for the poverty level, we can conclude that our model really does not do that well for classes other than Non Vulnerable. It only correctly identifies 15% of the Vulnerable households, classifying more of them as moderate or non-vulnerable.

Overall, these results show that imbalanced classification problems with relatively few observations are very difficult. There are some methods we can take to try and counter this such as oversampling or training multiple models on different sections of the data, but at the end of the day, the most effective method may be to gather more data.

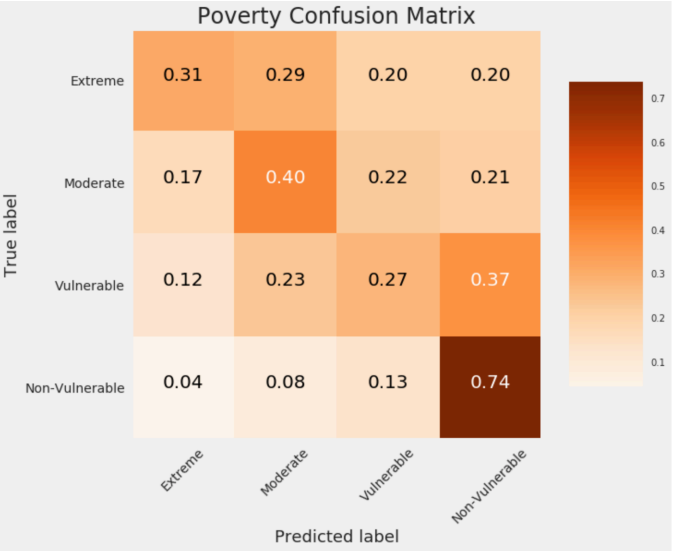

Figure 13: Normalized Poverty Confusion Matrix

DISCUSSION

We showed the best model for this data set for predictive power was the gradient boosting machine. We were able to classify household poverty, and identity which factors were most correlated with identifying it without the use of household income, which can be difficult to procure. This can help guide further research in how to combat household poverty by looking into which families require more assistance and how to help them. Being able to identify them is just the first step in how to assist the systematic problem.

While we have explored the best ways to correlated components to poverty, there are two main limitations of this project. First, understanding the correlations is important and can lead to meaningful follow up, but it cannot explain causality. In all our models, the most explanatory variable for poverty was the average amount of education for adults in the household, but this could be a symptom or a cause. We cannot know for sure that improving the average amount of education would result in lower poverty for the household. The second main limitation, is that in the dataset, most of the data points are for non-vulnerable families, as a result, the most power of the model was to predict whether a family would be non-vulnerable. This is would likely lead to insufficient conclusions. The solution to this would be to collect more data, which can be difficult and costly.

REFERENCES

[1] Reports, Staff. "What's Causing Poverty in Costa Rica?" *BORGEN*, 1 Oct. 2017, www.borgenmagazine.com/poverty-in-costa-rica/.

[2] "SAI Costa Rica, IDI Assess 'Bridge to Development.'" *IJGA International Journal of Government Auditing*, 21 Apr. 2018, intosaijournal.org/sai-costa-rica-idi-unite-assess-bridge-development/.

[3] *Costa Rican Household Poverty Prediction Challenge | Kaggle*, www.kaggle.com/willkoehrsen/a-complete-introduction-and-walkthrough.

[4] Cs.nju.edu.cn. (2018). [online] Available at: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/springerEBR09.pdf [Accessed 3 Dec. 2018].

[5] (n.d.). Cheng Li's Homepage. Retrieved from http://www.chengli.io/tutorials/gradient_boosting.pdf

[6] (n.d.). SLI | Main / Homepage. Retrieved from http://sli.ics.uci.edu/Classes/2012F-273a?action=download&upname=10-ensembles.pdf