Comparing of FACT properties of baseline and top models for prediction of household poverty level in Costa Rica

# Abstract

This project is dedicated to comparison of models of different complexity in context of fairness and transparency. Prediction of the family’s poverty level in Costa Rica was chosen as the use case for consideration. The corresponding Kaggle competition resulted in several baseline models as well as top performing ones. In this work baseline and sophisticated solutions are investigated from the point of view of responsible data science. This creates a precedent of model selection and tuning based not only on quality metrics, but meta conditions, caused by developers responsibility. In addition, application of existing tools, in particular SHAP, to ease decision making for both types of models is discussed. Results and conclusions of the investigation are described in this report.

# Introduction

Currently the main focus of Data Science and Machine Learning research lays into the field of efficiency. General approaches, algorithms and specific tools are developed to produce results of the best quality according to the formally defined quality metrics, while utilizing reasonable amount of resources. In particular, space and performance requirements, prediction quality and clever use of an available training data are considered. Interpretability, transparency, fairness and other aspects of responsible development are trending, but still not widely practiced.

At the same time, best models for many applied tasks are sophisticated. Intuitively and practically this leads to complicated inpretability. Moreover, models efficient on the available data are doing their best in reproducing biases present in the data. Thus question of widening quality requirements for DS projects should be raised.

# Motivation

First applications of Machine Learning were pretty technically oriented and applied to simple practical problems. This attempts were successful, in particular because subject field constraints were technical as well; there were no special regulations or requirements from the ethical point of view. Thus, previous results are mostly not touching ethical aspects of the development.

However, impressive results of this applications lead to constantly increasing business value, as well as growing trust to such solutions. Because of that, to stay competitive, companies of any kind are trying to make use of Data Science and Machine Learning. Consequently, applications in human-related fields are appearing, e.g. healthcare, insurance, social services etc. In this fields transparency, accuracy, fairness and confidentiality are of much higher importance.

Hence, bringing these aspects into a spotlight of model tuning and selection, will result in better development process. It will be more consistent and reliable, as well as solutions will be. As typical Data Science project development process goes from a simple baseline solution to a sophisticated one, it is important to understand how solution changes during mentioned transformation from the point of view of responsible development.

Providing insights on Data Science project development that can ease application of responsible development aspects in future projects is expected to be the impact of this report.

# Problem Formulation

As it was discussed earlier, main focus of this work is to analyze and compare baseline solution with top performing one for a certain use case from the point of view of responsible development. In particular, how biases in the data are consumed, how fair predictions are and how flaws can be fixed.

To this end, certain use case should be considered. Socially important problem should be chosen, such that fairness and transparency of a solution are of crucial importance for the real-world applicability. This way results of the work can bring practical benefits, but not only be toy example for educational purposes.

Thus, prediction of households poverty level in Costa Rica was chosen. This problem was released to the Kaggle community by The Inter-American Development Bank [<https://www.kaggle.com/c/costa-rican-household-poverty-prediction>], the largest source of development financing for Latin America and the Caribbean. Classical instruments of determining financial need don’t work well, when considering the poorest segment of the population, as limited data about their income and expense is available. Thus, organizations that make use of such a classification use Proxy Means Test as a best practice. It is a prediction based on the family’s observable household attributes.

The formulated problem was to determine level of financial vulnerability for the household, based on the data about its’ members. Extensive dataset was provided with 140 features and ca. 29 000 entries, with 4 possible values of the target variable - poverty of the household - Extreme, Moderate, Vulnerable, Non-Vulnerable. The target quality metric was macro F1-score, i.e. an arithmetic average of F1-scores for each of the 4 possible classes.

Solving this problem with modern methods can improve current classification results and consequently have a positive influence on many lives in Costa Rica. Moreover, common problems are appearing all the time in social, public and community services all around the globe, as resources are limited and one should make sure that right people receive appropriate amount of aid. So, solution can be used for other cases.

# Approach to solution

## Use case analysis

Among many results created by Kaggle community to solve this problem the most outstanding one was released by Will Koehrsen [<https://www.kaggle.com/willkoehrsen/a-complete-introduction-and-walkthrough>]. It is well-structured, clear and concise description of the solution development process. What is even more important it reflects all steps of the development of the Data Science solution according to CRISP methodology: problem understanding, data understanding, data preparation, solution development and evaluation, iteration over previous steps to reach the desired performance. I.e. it provides extensive data analysis with cleaning and feature engineering steps, baseline model and stepwise improvements until the top performing one.

On the other side, focus in the model selection and tuning was totally on the target quality metric. Taking into account all the facts above it was chosen as an ideal fit for the purpose of our analysis.

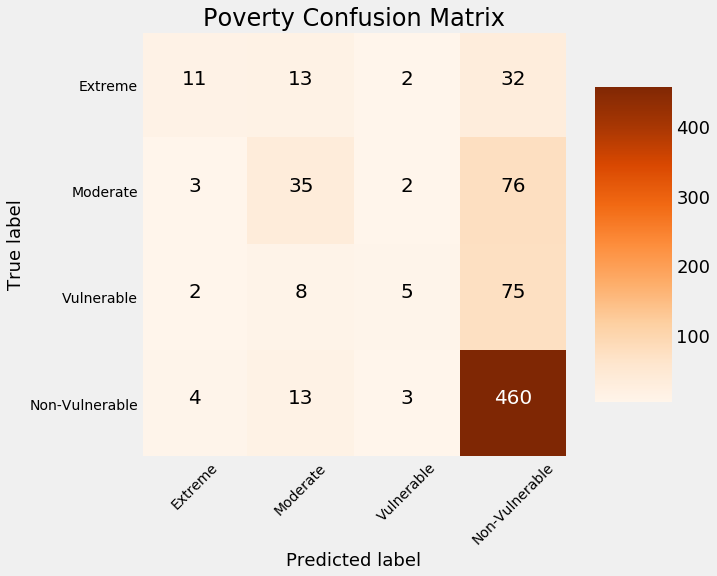
## SHAP

SHAP library was utilized as this new tool have become a swiss knife for machine learning model analysis

# Results

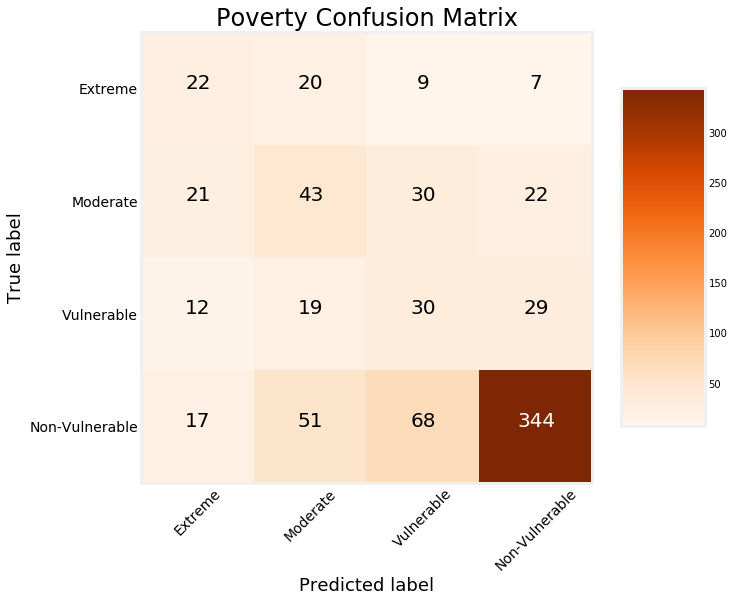
Baseline model is Random Forest, top performing one is XGBoost. Exact values of parameters can be found in the source code [TBD]. For the following experiments model with the parameters for baseline and best case were taken and trained on the subsample of the labeled data. Confusion matrix build for predictions for the remaining labeled data. Train part is 75% of the labeled data.

The following confusion matrix was received for the baseline model:



Bias of the model is observed, it tends to classify households as non-vulnerable. Most likely, as classes are unbalanced in the data.

The following confusion matrix was received for the top model:



In this case reversed shift is observed. Model tends to classify less entries as non-vulnerable, however, there is significant increase in erroneously classified objects of the Non-Vulnerable class. Taking into account the original problem this can be even worse, as leads to inefficient distribution of funds, while still not protecting significant part of poor citizens.

On the other side, baseline model is good enough in identifying households of classes other than Non--Vulnerable. So with human supervision on cases, where prediction is Non-Vulnerable, such system could work and reduce workload on social government workers.

Thus, we see that simpler models can be superior to better performing ones considering application and ethical aspects. Let’s investigate where and why models are making mistakes.

## Non-vulnerable vs Extreme

### Baseline model

### Top model

## Non-vulnerable vs Moderate

### Baseline model

13 Non-Vulnerable households were classified as Moderate poor. 67 Moderate poor household were classified as Non-Vulnerable. All 13 observations from the first category and 20 randomly sampled observations from the second category was analyzed using SHAP library plots. Experiment was done to find the minimal changes in the most influential features for the positive change in the prediction.

Insights for the false Moderate:

* The features, which contributes most to the misclassification in the case of false Moderate are high dependency, low escolari-min, low meaneduc, walls+roof+floor, low phone-per-capita, low escolari-max, high eviv1, high warning, low rooms-per-capita, low floor, high escolari/age-min, low instlevel8-max.
* Observations # 4, 11, 12 from the sample had higher chance to be classified as false Extreme poor rather than true Non-Vulnerable.
* 25% of the sample predictions could be corrected by 20% increase of escolari-max (Maximum years of schooling of the household members).
* In case of observation # 5 the increase of escolari-max starting from 40% causes the even worse misclassification as false Extreme poor - which seems to be the example of biased ML application

Insights for the false Non-Vulnerable:

* False Non-vulnerable prediction in the case are explained by low dependency, high phones-per-capita, high escolari-min, hight meaneduc, high instlevel8-max, low warning, high walls+roof+floor, high escolari-max, high instlevel8-sum, high escolari-sum, high cielorazo, high escolari/age-min, high inst-max, low intslevel2-sum.
* 14% of the sample predictions could be corrected by 60% decrease of phones-per-capita (Number of mobile phones per household members).
* 9% of the sample predictions could be corrected by 70% increase of dependency (number of members of the household younger than 19 or older than 64 divided by the number of member of household between 19 and 64).

### Top model

51 Non-Vulnerable households were classified as Moderate poor. 22 Moderate poor household were classified as Non-Vulnerable. 5 randomly sampled observations from the both categories was analyzed using SHAP library plots. Experiment was done to find the minimal changes in the most influential features for the positive change in the prediction.

Insights for the false Moderate:

* False Moderate prediction in the case are explained by low walls+roof+floor, high dependency, high age-sum, low escolari-min, high escolari/age-max, low floor, low rooms-per-capita, low phones-per-capita.
* 25% of the sample predictions could be corrected by 20% increase of escolari/age-sum (years of schooling divided by sum of age of household members).
* 25% of the sample predictions could be corrected by 70% decrease of dependency (number of members of the household younger than 19 or older than 64 divided by the number of member of household between 19 and 64).

Insights for the false Non-Vulnerable:

* False Non-vulnerable prediction in the case are explained by high phones-per-capita, high rent-per-capita, high v2a1, high escolari-min, high walls+roof+floor, low dependency, low escolari/age-min, high bonus, hight meaneduc, high edjefe.
* Two obs. with extreme high influence of high escolari-max feature, and one obs. with extreme high influence of rent-per-capita feature.
* 33% of the sample predictions could be corrected by 40% decrease of phones-per-capita (years of schooling divided by sum of age of household members).

## Non-vulnerable vs Vulnerable

### Baseline model

3 Non-Vulnerable households were classified as Vulnerable. 75 Vulnerable households were classified as Non-Vulnerable.

To take a closer look at these misclassified cases household data and prediction for it were investigated with the help of SHAP library. All 3 samples from the first category and 5 randomly drawn samples from the second one were taken for consideration.

The following insights were found:

* Misclassified cases are edge cases. Even for human it is hard to classify this cases.
* Features related to education are important. Their influence can be both strongly negative or strongly positive for the probability of certain class. Seems like, model might overfit for certain values of these features, as in some cases influence on prediction is exactly opposite to the expected for a proper label of the entry.
* Difference between predicted probabilities is tiny, 1-3% for many cases. In simple words, this means that model is not sure about prediction.

### Top model

68 Non-Vulnerable households were classified as Vulnerable. 29 Vulnerable households were classified as Non-Vulnerable.

To take a closer look at these misclassified cases household data and prediction for it were investigated with the help of SHAP library. 5 randomly drawn samples were from both categories for consideration.

These are insights found:

* Features related to education are important. However, their influence changes from case to case, can be both strongly negative or strongly positive for the probability of certain class.
* Difference between predicted probabilities are big. So, model misclassified households with much more confidence.
* Features created during feature engineering steps are of strong importance and working properly, as they are decreasing predicted probability for somen classes, while increasing it for others

Comparison of both models leads to the following insights:

* SHAP utilization helps to understand why certain predictions were made and ease decision making for human supervisor.
* Unbalanced сlasses in the data harmfully influence prediction quality both in formal and application aspects.
* Seeing the predicted probabilities of classes (or proxy score for XGBoost) is extremely useful for understanding of model nature and overfitting level.
* Overfitting negatively influence both prediction quality and how model can be used as a supportive tool for decision making.
* Some features have very strong influence: features related to education of household members; features created during feature engineering stage - aggregating signs of poor/prosper household. These features require more attention during model development and results analysis.

# Conclusions & Future work

Results of this work satisfied our interest according to the aim of the project, as following insights were found:

1. Fighting overfitting is very important when model is expected to be used as a supportive tool in decision making. Simpler model can be more useful for such purpose.
2. Balancing classes in the training data is a must.
3. Making decision based on such a big number of features is hard. Supportive tool can be very beneficial.
4. Expert based feature engineering and automated solutions for it can strongly help in solving real-life problems, even though for a supportive tool explaible features made by human expert are much easier to utilize.

However, there are following options for improvements and further work:

1. Generalize approach of comparison baseline and sophisticated models.