Machine Learning Engineer Nanodegree

Capstone Proposal

David Jansen
July 29, 2018

Proposal

Costa Rican Household Poverty Level Prediction

Domain Background

Accurately assessing social needs to ensure the poorest people our planet get the help they need is a difficult task. The Inter-American Development Bank (IADB) is an organization which focuses to improve the lives of those who live in Latin America and the Caribbean. Publications made in December 2016 by the IADB reveal that in 2015 between 8.2% for Chili to 68.7% of the total population of those countries live of less than 5 USD a day. Extreme cases of poverty range between 2.7% for Chili and a shocking 32.6% for Guatemala, where the income a day is less than 3.1 USD. In a study in 2015 a proposition was made to apply machine learning to poverty targeting (McBride & Nichols 2015). The study revealed that they were able to improve on the then standardized method for targeting applied by the USAID by 2 to 18 percent using a Random Forest algorithm. For this reason, the IADB is looking for new ways to reach people who are in need of help. They have reached out to the Kaggle community in order to find new ways to help identify vulnerable households who may need help.

Problem Statement

In Latin America, a Proxy Means Test (PMT) is used to asses the level of need a household needs. Despite this assessment being an improvement, a need for a model which more accurately classifies these households is present. The IADB has asked to Kaggle Community to create such a model. They have provided a dataset containing multiple characteristics of a Costa Rican household (See datasets and inputs).

Datasets and Inputs

There are two files the IADB has provided. A training dataset containing multiple features including a target label feature and a test set containing the same features, but without the target label. The datasets respectively contain approximately nine thousand and twenty-four thousand data points.
Due to the nature of the Kaggle competition, external data is not allowed.
The dataset contains the following core data fields (from Kaggle):

- Id - a unique identifier for each row.
- Target - the target is an ordinal variable indicating groups of income levels.
    - 1 = extreme poverty
    - 2 = moderate poverty
    - 3 = vulnerable households
    - 4 = non vulnerable households
- idhogar - this is a unique identifier for each household. This can be used to create household-wide features, etc. All rows in a given household will have a matching value for this identifier.
- parentesco1 - indicates if this person is the head of the household.
- This data contains 142 total columns. (Appendix A)

An initial look at the data reveals that the data points are heavily skewed towards non-vulnerable households which are overrepresented in the data. Due to this class imbalance, care has to be taken when splitting the data for training and cross-validation. Using stratified splitting, the data can be split without losing the class imbalance.

class label	count
1	755
2	1597
3	1209
4	5996

Solution Statement

To accurately predict the ‘Target’ variable, a deep neural network will be created. The choice of this approached is made partially due to the fact we cannot assume that the data is linearly separable. Additionally, neural networks can outperform traditional machine learning algorithms, but they need some more finetuning to get there. Due to the way the data is presented, some feature engineering may be needed to get a good F1-score. The ‘idhogar’ feature is stated to be a good baseline for this.

Benchmark Model

For benchmarking, the macro F1-score of the solution will be compared against two other models. Firstly, an oversimplified predictor will be created by simply assuming the modus of the ‘Target’ variable present in the data, as a prediction value for all that ‘Target’ variables. Secondly, a Random Forest model will be created to challenge the performance.

Evaluation Metrics

Due to the nature of the competition, all submissions for the project will be evaluated by their macro F1-score. Both the benchmark model and both the solution model will be evaluated based on this score.

Project Design

Explore the data
To get a general feel for the data, some time is needed to explore what kind of data is actually in the dataset. Some questions that are central at this point are: what type of data is there, what is the range of the data, is there missing data, what features seem promising for feature engineering. The goal is to get familiar enough with the data to be able to start with preprocessing the data.

Preprocess the data
Findings in the exploration phase will be implemented here. This can include but is not limited to data cleanup, normalization of the data, feature engineering and one-hot encoding. Here the data will also be split so there is a validation set available. The goal is to have a dataset which is usable for building the neural network on.

Building and training the network
The goal is to create a model architecture which is capable of accurately predicting the target variable. A multi-layer perceptron will be created from scratch to challenge the benchmarking models. Due to the nature of the Kaggle competition, it is not allowed to use any pre-trained model, omitting transfer learning as an option. For familiarity, the Keras framework will be used for constructing the model Taking this into account, a initial model would look contain at least an input layer with n input nodes, where n is the amount of features present in the data after pre-processing. The output layer will consist of 4 nodes, 1 for each of the output classes. The hidden layers will contain at least dropout layers to counter overfitting the data and activation layers. Tuning and experimenting will take place here based on the results in the evaluation stage. Additionally, early stopping will be employed to also help reduce overfitting.

An additional challenge in this phase will be accounting for possible performance issues to the amount of features in the data. Adding too many layers may cause the duration of training to get unacceptably long. Additionally, as the kernels will have to run on Kaggle, there is an unknown factor of what the limits of the kernel will be. After building the model, it will be trained using the training data made available. This stage will most likely be revisited multiple times to create a better model.

Evaluate model performance
At this stage, the model is able to make predictions. Validation loss and the F1-score will be used to determine the accuracy of the

current model. The F1-score of the neural network will at this point be compared against the F1-score of the oversimplified model and the Random Forest model. If not satisfactory, the model can be tuned by revisiting the previous stage. The goal is to get an F1-score which is better than the other models.

Test the model
If the model has reached a point of satisfactory result, the model will be used to make predictions on the test dataset. Per requirement, a file will be written and will be submitted in the Kaggle environment for evaluation. Though this should be the final stage, this stage may be revisited in order to optimize the final result.

References

McBride, L., & Nichols, A. (2015). Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools. Unpublished manuscript. Available at: http://www.econthatmatters.com/wp-content/uploads/2015/01/improvedtargeting_21jan2015.pdf.

STATISTICS ON POVERTY AND INCOME INEQUALITY IN LAC (18 COUNTRIES). (2016, Dec) Retrieved from https://www.iadb.org/en/research-and-data/poverty%2C7526.html.

Appendix A

List of features available in the dataset

Variable name	Variable description
v2a1	Monthly rent payment
hacdor	=1 Overcrow ding by bedrooms
rooms	number of all rooms in the house
hacapo	=1 Overcrow ding by rooms
v14a	=1 has bathroom in the household
refrig	=1 if the household has refrigerator
v18q	ow ns a tablet
v18q1	number of tablets household ow ns
r4h1	Males younger than 12 years of age
r4h2	Males 12 years of age and older
r4h3	Total males in the household
r4m1	Females younger than 12 years of age
r4m2	Females 12 years of age and older
r4m3	Total females in the household
r4t1	persons younger than 12 years of age
r4t2	persons 12 years of age and older
r4t3	Total persons in the household
tamhog	size of the household
tamviv	number of persons living in the household
escolari	years of schooling
rez_esc	Years behind in school
hhsize	household size
paredblolad	=1 if predominant material on the outside w all is block or brick
paredzocalo	zinc or absbesto"
paredpreb	=1 if predominant material on the outside w all is prefabricated or cement
pareddes	=1 if predominant material on the outside w all is w aste material
paredmad	=1 if predominant material on the outside w all is w ood
paredzinc	=1 if predominant material on the outside w all is zink

paredfibras	=1 if predominant material on the outside w all is natural fibers
paredother	=1 if predominant material on the outside w all is other
pisomoscer	terrazo"
pisocemento	=1 if predominant material on the floor is cement
pisoother	=1 if predominant material on the floor is other
pisonatur	=1 if predominant material on the floor is natural material
pisonotiene	=1 if no floor at the household
pisomadera	=1 if predominant material on the floor is w ood
techozinc	=1 if predominant material on the roof is metal foil or zink
techoentrepiso	mezzanine "
techocane	=1 if predominant material on the roof is natural fibers
techootro	=1 if predominant material on the roof is other
cielorazo	=1 if the house has ceiling
abastaguadentro	=1 if w ater provision inside the dw elling
abastaguafuera	=1 if w ater provision outside the dw elling
abastaguano	=1 if no w ater provision
public	ESPH/JASEC"
planpri	=1 electricity from private plant
noelec	=1 no electricity in the dw elling
coopele	=1 electricity from cooperative
sanitario1	=1 no toilet in the dw elling
sanitario2	=1 toilet connected to sew er or cesspool
sanitario3	=1 toilet connected to septic tank
sanitario5	=1 toilet connected to black hole or letrine
sanitario6	=1 toilet connected to other system
energcocinar1	=1 no main source of energy used for cooking (no kitchen)
energcocinar2	=1 main source of energy used for cooking electricity
energcocinar3	=1 main source of energy used for cooking gas
energcocinar4	=1 main source of energy used for cooking w ood charcoal
elimbasu1	=1 if rubbish disposal mainly by tanker truck
elimbasu2	=1 if rubbish disposal mainly by botan hollow or buried
elimbasu3	=1 if rubbish disposal mainly by burning

elimbasu4	=1 if rubbish disposal mainly by throw ing in an unoccupied space
elimbasu5	creek or sea"
elimbasu6	=1 if rubbish disposal mainly other
epared1	=1 if w alls are bad
epared2	=1 if w alls are regular
epared3	=1 if w alls are good
etecho1	=1 if roof are bad
etecho2	=1 if roof are regular
etecho3	=1 if roof are good
eviv1	=1 if floor are bad
eviv2	=1 if floor are regular
eviv3	=1 if floor are good
dis	=1 if disable person
male	=1 if male
female	=1 if female
estadocivil1	=1 if less than 10 years old
estadocivil2	=1 if free or coupled uunion
estadocivil3	=1 if married
estadocivil4	=1 if divorced
estadocivil5	=1 if separated
estadocivil6	=1 if w idow /er
estadocivil7	=1 if single
parentesco1	=1 if household head
parentesco2	=1 if spouse/partner
parentesco3	=1 if son/doughter
parentesco4	=1 if stepson/doughter
parentesco5	=1 if son/doughter in law
parentesco6	=1 if grandson/doughter
parentesco7	=1 if mother/father
parentesco8	=1 if father/mother in law
parentesco9	=1 if brother/sister
parentesco10	=1 if brother/sister in law

parentesco11	=1 if other family member
parentesco12	=1 if other non family member
idhogar	Household level identifier
hogar_nin	Number of children 0 to 19 in household
hogar_adul	Number of adults in household
hogar_mayor	# of individuals 65+ in the household
hogar_total	# of total individuals in the household
dependency	calculated = (number of members of the household younger than 19 or older than 64)/(number of member of household betw een 19 and 64)
edjefe	yes=1 and no=0
edjefa	yes=1 and no=0
meaneduc	meaneduc,average years of education for adults (18+)
instlevel1	=1 no level of education
instlevel2	=1 incomplete primary
instlevel3	=1 complete primary
instlevel4	=1 incomplete academic secondary level
instlevel5	=1 complete academic secondary level
instlevel6	=1 incomplete technical secondary level
instlevel7	=1 complete technical secondary level
instlevel8	=1 undergraduate and higher education
instlevel9	=1 postgraduate higher education
bedrooms	number of bedrooms
overcrow ding	# persons per room
tipovivi1	=1 ow n and fully paid house
tipovivi2	paying in installments"
tipovivi3	=1 rented
tipovivi4	=1 precarious
tipovivi5	borrow ed)"
computer	=1 if the household has notebook or desktop computer
television	=1 if the household has TV
mobilephone	=1 if mobile phone
qmobilephone	# of mobile phones

lugar1	=1 region Central
lugar2	=1 region Chorotega
lugar3	=1 region PacÃƒÂfico central
lugar4	=1 region Brunca
lugar5	=1 region Huetar AtlÃƒÂ¡ntica
lugar6	=1 region Huetar Norte
area1	=1 zona urbana
area2	=2 zona rural
age	Age in years
SQBescolari	escolari squared
SQBage	age squared
SQBhogar_total	hogar_total squared
SQBedjefe	edjefe squared
SQBhogar_nin	hogar_nin squared
SQBovercrowding	overcrowding squared
SQBdependency	dependency squared
SQBmeaned	square of the mean years of education of adults (>=18) in the household
agesq	Age squared