API 222 Prediction Competition

Machine Learning and Big Data Analytics

Due Before 1:15 pm on November 15, 2018

Under the Kennedy School Academic Code, this assignment is a Type II assignment. You are encouraged to work in a study group, but must submit your own hand- or type-written solutions. It is not acceptable to work on one electronic document as a group and submit identical, or nearly identical versions.

1 Overview

The following is a real problem posted as a competition on kaggle.com by the Inter-American Development Bank. It represents a real way in which policy makers are leveraging the tools of machine learning to gain insights to improve policy.

The Inter-American Development Bank is asking the Kaggle community for help with income qualification for some of the world's poorest families. Are you up for the challenge?

Here's the backstory: Many social programs have a hard time making sure the right people are given enough aid. It's especially tricky when a program focuses on the poorest segment of the population. The world's poorest typically can't provide the necessary income and expense records to prove that they qualify.

In Latin America, one popular method uses an algorithm to verify income qualification. It's called the Proxy Means Test (or PMT). With PMT, agencies use a model that considers a family's observable household attributes like the material of their walls and ceiling, or the assets found in the home to classify them and predict their level of need.

While this is an improvement, accuracy remains a problem as the region's population grows and poverty declines.

To improve on PMT, the IDB (the largest source of development financing for Latin America and the Caribbean) has turned to the Kaggle community. They believe that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT's performance.

Beyond Costa Rica, many countries face this same problem of inaccurately assessing social need. If Kagglers can generate an improvement, the new algorithm could be implemented in other countries around the world.

2 Assignment

Your objective is to develop a model that accurately predicts poverty. As with the online version of the competition, your score will be determined using the macro F_1 Score on a holdout set of data. Poverty in this data is measured by a categorical but ordinal variable called `Target`, so this is a classification problem. Prof. Saghafian will announce in class the first, second, and third place students who have been able to achieve the highest macro F_1 scores using a Machine Learning algorithm.

The macro F_1 score is an aggregate measure for classification errors. It calculates four F_1 scores (because the target can take four unique values) and then averages them. For example, the first F_1 score it calculates will treat 1's as positives and 2s, 3s, and 4s collectively as negatives. The F_1 score for a binary classification problem is defined as:

$$F_1 = \frac{2 \cdot \text{ true positives}}{2 \cdot \text{ true positives} + \text{ false negatives } + \text{ false positives}} \tag{1}$$

Unlike with the online version, your score will also be determined by an accompanying write up that clearly explains the process you went through of choosing a model and describes your final approach.

You will be required to submit working code (we must be able to run it by changing only one file path line), a CSV file of predictions for a hold out set of data that will be released 48 hours before the submission deadline, and a write-up between 1.5 and 2.5 pages single spaced, size 12 font Times New Roman. Your writeup should be geared toward a member of the IDB staff who has some familiarity with Machine Learning but who is not an expert. The goal of the written portion of this assignment is to get you familiar with explaining the process of model selection to a broad audience in a clear way. This will be an important skill in facilitating the adoption of high-performing yet new or unfamiliar methods in the types of organizations where many of you will work after graduation.

3 Prediction Competition Rules

You may not use any data other than the data provided by the course instructors in developing your model. **Anyone who uses any additional data will receive zero credit for the assignment.** However, you may do whatever you like with the data provided, such as generating new features through interactions, non-linear transformations, etc.

4 Grading

This competition is worth 15% of your course grade. Therefore, the assignment will be worth 15 points, which will be broken down into three evenly weighted components (e.g. 5 points each):

1. The write-up, which has
   (a) A thorough description of the process you took to arrive at your final model
   (b) A clear description of your final model, including any data manipulation or feature engineering
   (c) A discussion of your approach as it pertains to algorithmic bias and transparency. This section should contain some numbers illustrating how your model performance varies along salient characteristics, such as demographic and geographic characteristics. It should also concretely discuss the tradeoffs your model makes between predictive performance and interpretability / transparency.
2. Clean code that:
   (a) Trains your model
   (b) Produces a CSV of predictions for the holdout data

   The teaching staff must be able to successfully run the code by changing only one line of the file path. **Code that we cannot run without further edits will receive at most 1 of the 5 possible points.**
3. A CSV file of predictions. We will order students in terms of predictive accuracy on the holdout data.
   (a) Students in the top one-fifth of the class on this measure will receive 5 out of 5 points on this component.
   (b) Students in the second-to-top one-fifth will receive 4 out of 5 points on this component.
   ⋮
   (c) Students in the bottom fifth will receive 1 out of 5 points on this component.

   We will provide you with a sample submission CSV, which will have two columns:
   (a) `Id` - An ID column that maps to the holdout data released 48 hours before the submission deadline

(b) `Prediction` - A column that contains your predicted poverty level for each observation in the holdout sample

You should submit a CSV file with the same two columns, and those columns should be named `Id` and `Prediction`. The filename should be `lastname_prediction.csv`, where you replace `lastname` with your actual last name (for example Prof. Saghafian's file would be `saghafian_prediction.csv`). **Any submissions that do not include these two columns (with the correct column names) or have the wrong file name will receive zero points on this section.**

5 Data Description

Number of Observations	7646
Number of Predictors	140
Number of Id Variables	2
Response Variable	`Target`

5.1 Core Data Fields

- `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.

5.2 All Data fields

Variable name, Variable description

```
v2a1, Monthly rent payment
hacdor, =1 Overcrowding by bedrooms
rooms, number of all rooms in the house
hacapo, =1 Overcrowding by rooms
v14a, =1 has bathroom in the household
refrig, =1 if the household has refrigerator
v18q, owns a tablet
v18q1, number of tablets household owns
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 wall is block or brick
paredzocalo, =1 if predominant material on the outside wall is socket (wood, zinc or absbesto
paredpreb, =1 if predominant material on the outside wall is prefabricated or cement
pareddes, =1 if predominant material on the outside wall is waste material
paredmad, =1 if predominant material on the outside wall is wood
paredzinc, =1 if predominant material on the outside wall is zink
paredfibras, =1 if predominant material on the outside wall is natural fibers
paredother, =1 if predominant material on the outside wall is other
pisomoscer, =1 if predominant material on the floor is mosaic, ceramic, 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 wood
techozinc, =1 if predominant material on the roof is metal foil or zink
techoentrepiso, =1 if predominant material on the roof is fiber cement, 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 water provision inside the dwelling
abastaguafuera, =1 if water provision outside the dwelling
abastaguano, =1 if no water provision
public, =1 electricity from CNFL, ICE, ESPH/JASEC
planpri, =1 electricity from private plant
noelec, =1 no electricity in the dwelling
coopele, =1 electricity from cooperative
sanitario1, =1 no toilet in the dwelling
sanitario2, =1 toilet connected to sewer 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 wood 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 throwing in an unoccupied space
elimbasu5, =1 if rubbish disposal mainly by throwing in river, creek or sea
elimbasu6, =1 if rubbish disposal mainly other
epared1, =1 if walls are bad
epared2, =1 if walls are regular
epared3, =1 if walls 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 union

estadocivil3, =1 if married
estadocivil4, =1 if divorced
estadocivil5, =1 if separated
estadocivil6, =1 if widow/er
estadocivil7, =1 if single
parentesco1, =1 if household head
parentesco2, =1 if spouse/partner
parentesco3, =1 if son/daughter
parentesco4, =1 if stepson/daughter
parentesco5, =1 if son/daughter in law
parentesco6, =1 if grandson/daughter
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, Dependency rate, calculated
= (number of members of the household younger than 19 or older than 64)/(number of
member of household between 19 and 64)
edjefe, years of education of male head of household, based on the interaction of escolari
(years of education), head of household and gender, yes=1 and no=0
edjefa, years of education of female head of household, based on the interaction of
escolari (years of education), head of household and gender, yes=1 and no=0
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
overcrowding, persons per room
tipovivi1, =1 own and fully paid house
tipovivi2, =1 own, paying in installments
tipovivi3, =1 rented
tipovivi4, =1 precarious
tipovivi5, =1 other(assigned, borrowed)
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 Pacifico central
lugar4, =1 region Brunca
lugar5, =1 region Huetar Atlantica
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