✔ COMPETITION COMPLETE

DAT102x: Predicting Poverty Around the World

HOSTED BY MICROSOFT

Problem Description

- About the Data
- Target Variable
  - Submission Format
  - Performance Metric
- Features
  - Example Row
- References

HOME (/COMPETITIONS/15/PREDICTING-POVERTY/PAGE/46/)

PROBLEM DESCRIPTION (/COMPETITIONS/15/PREDICTING-POVERTY/PAGE/47/)

EXPORT SCORE FOR EDX (/COMPETITIONS/15/PREDICTING-POVERTY/PAGE/48/)

LEADERBOARD RESULTS (/COMPETITIONS/15/PREDICTING-POVERTY/LEADERBOARD/)

DATA DOWNLOAD (/COMPETITIONS/15/PREDICTING-POVERTY/DATA/)

About the Data

Your goal is to predict the probability that individuals across seven different countries live below the poverty line at the $2.50/day threshold, given other socioeconomic indicators. The probability of being in poverty was calculated using the Poverty Probability Index (PPI), which estimates an individual's poverty status using 10 questions about a household’s characteristics and asset ownership. The remaining data comes from the Financial Inclusion Insights household surveys conducted by InterMedia.

Target Variable

We're trying to predict the variable `poverty_probability` for each row of the test data set. `poverty_probability` is a positive floating point number (e.g. 0.84) between 0.0 and 1.0 inclusive.

Your job is to:

1. Train a model using the inputs in train_values.csv and the labels train_labels.csv
2. Predict floats for each row in test_values.csv for which you don't know the true probability of poverty.
3. Output your predictions in a format that matches submission_format.csv **exactly**.
4. Upload your predictions to this competition in order to get a score.
5. Export your grading token (click the "Export Score for EdX" tab) and paste it into the assignment grader on edX to get your course grade.

Submission Format

The format for the submission file is two columns with the `row_id` and the `poverty_probability`. The data type of `poverty_probability` is a float between 0.0 and 1.0, **so make sure there is a decimal point in your submission**. For example `0.66` is a valid float representing 66%, but `66` is *not*.

If you predicted all individuals had a 100% change of being in poverty, a probability of 1.0, the first few lines of the `.csv` file that you submit would look like:

```
row_id,poverty_probability
0,1.0
1,1.0
2,1.0
3,1.0
4,1.0
⋮
```

SUBMISSIONS (/COMPETITIONS/15/PREDICTING-POVERTY/SUBMISSIONS/)

OFFICIAL RULES (/COMPETITIONS/15/PREDICTING-POVERTY/RULES/)

Performance Metric

We're predicting a probability (continuous float), so this is a regression problem. To measure regression, we'll use a metric called r-squared, also known as the coefficient of determination. It is a statistical measure of the goodness-of-fit of a regression model, so a higher value is better (as opposed to an error metric, where a lower value is better).

$$R^2 = 1 - \frac{SSres}{SStot}$$

Where $SSres$ is the sum of squares of residuals and $SStot$ is the total sum of squares. The best possible score is 1, but the worst possible score can be negatively infinite.

Features

There are 58 variables in this dataset. Each row in the dataset represents an individual in a given country. Each `country` is represented by a unique letter. There are seven countries in total. In addition to country, `religion` is also obscured, with each religion corresponding to a unique letter.

The variables are as follows:

DEMOGRAPHICS

- `country` - Unique identifier for each country
- `is_urban` - Urban vs. rural area of residence
- `age` - Age
- `female` - Sex (True=female, False=male)
- `married` - Marital status
- `religion` - Unique identifier for religion
- `relationship_to_hh_head` - Respondent's relationship to the head of the household

EDUCATION

- `education_level` - Highest level of education (0=no education, 1=primary education, 2=secondary education, 3=higher education)
- `literacy` - Ability to read and understand
- `can_add` - Ability to add
- `can_divide` - Ability to divide
- `can_calc_percents` - Ability to calculate percents
- `can_calc_compounding` - Ability to calculate compounding interest

EMPLOYMENT

- `employed_last_year` - Whether the respondent was employed in the last year
- `employment_category_last_year` - Category of employment last year (e.g. employed, retired)
- `employment_type_last_year` - Type of employment last year (e.g. salaried, seasonal)
- `share_hh_income_provided` - Share of household income provided
- `income_ag_livestock_last_year` - Whether the respondent received income from agriculture or livestock in the last year
- `income_friends_family_last_year` - Whether the respondent received income from friends or family in the last year
- `income_government_last_year` - Whether the respondent received income from the government in the last year
- `income_own_business_last_year` - Whether the respondent received income from their own business in the last year
- `income_private_sector_last_year` - Whether the respondent received income from the private sector in the last year
- `income_public_sector_last_year` - Whether the respondent received income from the public sector in the last year

ECONOMIC

- `num_times_borrowed_last_year` - Number of times the respondent borrowed money in the last year
- `borrowing_recency` - Recency of last borrowing activity
- `formal_savings` - Has savings at a formal institution
- `informal_savings` - Has savings at an informal institution

- `cash_property_savings` - Has savings in cash or property
- `has_insurance` - Has at least one form of insurance
- `has_investment` - Has at least one form of investment
- `bank_interest_rate` - Interest rate paid on a bank loan
- `mm_interest_rate` - Interest rate paid on a mobile money loan
- `mfi_interest_rate` - Interest rate paid on a Microfinance Institution (MFI) loan
- `other_fsp_interest_rate` - Interest rate paid to another financial service provider (FSP)
- `num_shocks_last_year` - Number of financial shocks experienced in the last year
- `avg_shock_strength_last_year` - Average strength of shocks experienced in the last year
- `borrowed_for_emergency_last_year` - Borrowed money for an emergency in the last year
- `borrowed_for_daily_expenses_last_year` - Borrowed money for daily expenses in the last year
- `borrowed_for_home_or_biz_last_year` - Borrowed money for home or business expenses in the last year

PHONE

- `phone_technology` - Sophistication of phone type (0=no phone, 1=basic phone, 2=feature phone, 3=smartphone)
- `can_call` - Ability to make a phone call
- `can_text` - Ability to text
- `can_use_internet` - Ability to use internet on one's phone
- `can_make_transaction` - Ability to make a financial transaction on one's phone
- `phone_ownership` - Phone owership (0=no phone, 1=shares phone, 2=owns phone)
- `advanced_phone_use` - Ability to do advanced tasks on a phone

FINANCIAL INCLUSION

- `reg_bank_acct` - Has a bank account in their own name
- `reg_mm_acct` - Has a mobile money account in their own name
- `reg_formal_nbfi_account` - Has an account at a non-banking financial institution (NBFI) in their own name
- `financially_included` - Financially included, which is defined as having at least one of the following: a registered bank account, a registered mobile money account, or a registered NBFI

account

- `active_bank_user` - Has used their bank account in the last 90 days
- `active_mm_user` - Has used their mobile money account in the last 90 days
- `active_formal_nbfi_user` - Has used their formal NBFI account in the last 90 days
- `active_informal_nbfi_user` - Has conducted financial activity at an informal NBFI in the last 90 days
- `nonreg_active_mm_user` - Has used a mobile money account in someone else's name in the last 90 days ('over-the-counter' use)
- `num_formal_institutions_last_year` - Number of formal financial institutions used in the last year
- `num_informal_institutions_last_year` - Number of informal financial institutions used in the last year
- `num_financial_activities_last_year` - Number of different types of financial activities conducted in the last year

Example Row

Here's an example of one of the rows in the dataset so that you can see the kinds of values you might expect in the dataset. Many are boolean, some are numerical, a few are categorical, and there can be missing values.

	0
country	C
is_urban	False
age	18
female	True
married	True
religion	P
relationship_to_hh_head	Other

	0
education_level	1
literacy	True
can_add	True
can_divide	True
can_calc_percents	True
can_calc_compounding	True
employed_last_year	False
employment_category_last_year	housewife_or_student
employment_type_last_year	not_working
share_hh_income_provided	1
income_ag_livestock_last_year	False
income_friends_family_last_year	False
income_government_last_year	False
income_own_business_last_year	False
income_private_sector_last_year	False
income_public_sector_last_year	False
num_times_borrowed_last_year	0
borrowing_recency	0
formal_savings	False
informal_savings	False

	0
cash_property_savings	False
has_insurance	False
has_investment	False
bank_interest_rate	NaN
mm_interest_rate	NaN
mfi_interest_rate	NaN
other_fsp_interest_rate	NaN
num_shocks_last_year	0
avg_shock_strength_last_year	0
borrowed_for_emergency_last_year	False
borrowed_for_daily_expenses_last_year	False
borrowed_for_home_or_biz_last_year	False
phone_technology	0
can_call	True
can_text	True
can_use_internet	False
can_make_transaction	False
phone_ownership	1
advanced_phone_use	False
reg_bank_acct	True

	0
reg_mm_acct	False
reg_formal_nbfi_account	False
financially_included	True
active_bank_user	True
active_mm_user	False
active_formal_nbfi_user	False
active_informal_nbfi_user	False
nonreg_active_mm_user	False
num_formal_institutions_last_year	1
num_informal_institutions_last_year	0
num_financial_activities_last_year	1

References

- Innovations for Poverty Action (IPA). About the PPI: A Poverty Measurement Tool. https://www.povertyindex.org/about-ppi (https://www.povertyindex.org/about-ppi)
- InterMedia. Financial Inclusion Insights. http://finclusion.org/ (http://finclusion.org/)

Hosted by (https://www.drivendata.org/)