Predicting the Poverty Probability Index

Justin Sharber

September 10, 2019

Executive Summary

The Poverty Probability Index (PPI) estimates the probability that a household will fall below the poverty line, which is defined as $2.50 per day.[1] The goal of this project is to estimate the probability given by that index, based on data taken mainly from the Financial Inclusion Insights household surveys.[2]

As the goal is the estimation of a probability, this is a regression problem. However, the data contains almost no usable numeric features, making regression more difficult.

The main work of the project was analyzing the data and carefully managing and classifying numeric features. The metric for the project is the R^2 score, with a goal of 0.41. Several algorithms were used, and the boosted decision tree regressor proved the most predictively accurate at 0.422.

This report has two main conclusions. First, the top R^2 scores achieved are fairly low. This shows that the majority of variation in the dataset was not able to be explained on the basis of these features. In other words, if the PPI is predictable from other data, this dataset simply do not contain enough information to do it.

Second, there is a fair amount of information overlap in the data. The starkest indicator of the overlap is in the dramatic difference of feature importance between similar features, explored in the last section of this report.

[1] Innovations for Poverty Action (IPA). About the PPI: A Poverty Measurement Tool. https://www.povertyindex.org/about-ppi

[2] Financial Inclusion Insights. http://finclusion.org/

Description of the Data

The dataset consists of 12,600 data with 58 features and one label (the PPI, *poverty_probability*). The data are summarized below. A complete description of the features is available online.[3]

Category	Feature examples
Demographics	Country of residence, age.
Education	Highest level of education, mathematical ability.
Employment	Whether the individual was employed last year, whether the individual received income from friends or family in the last year.
Economic	The number of times the individual borrowed money last year, whether the individual has a savings at a formal institution.
Phone	Level of phone technology.
Financial Inclusion	Whether the individual has a bank account in their own name, whether the individual has used their bank account recently.

Most of these features are Boolean (True-False). This point anticipates the main challenges in the data, a lack of usable numeric features.

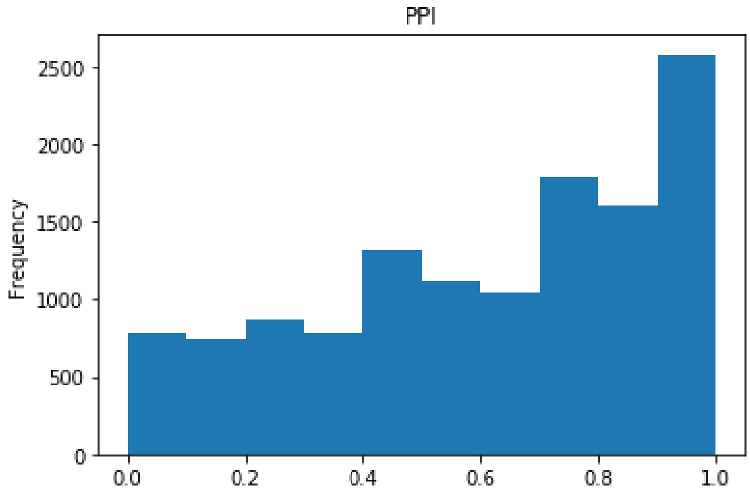
While most of these general categories are intuitive choices to include, it is a surprise (at least to the author) that a whole category of “Phone” features is included. There are seven features in this set, some of which are informative.

[3] DAT102x: Predicting Poverty Around the World. Problem Description. https://datasciencecapstone.org/competitions/15/predicting-poverty/page/47/#features

Comments on the data

The Label: Poverty Probability Index

- PPI is skewed. Only 36% of those in the dataset have a PPI < 50%.

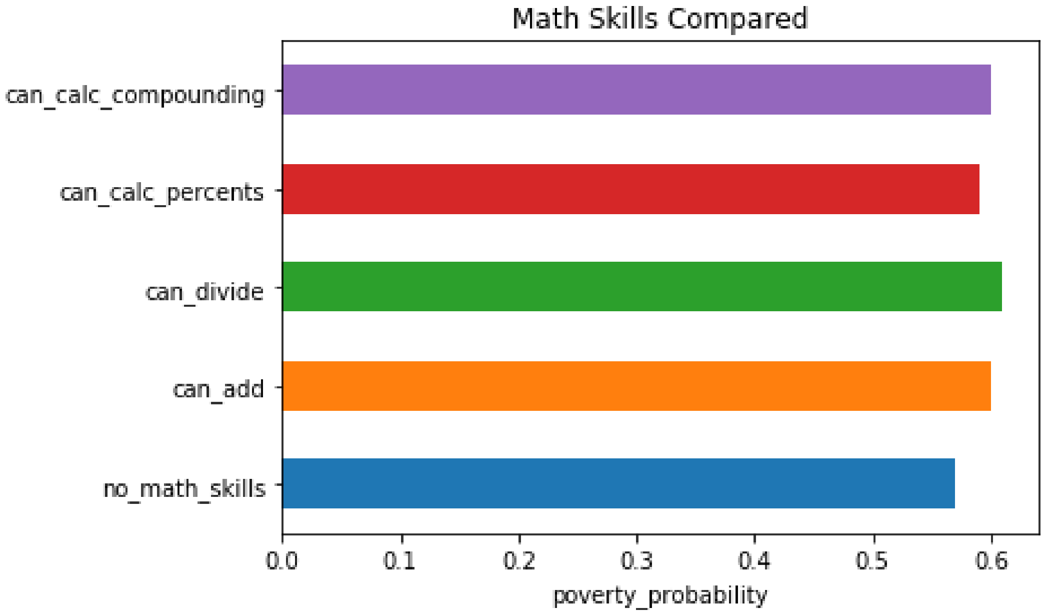


Demographics

- Seven countries and five religions are represented. Both show a fair range in average PPI of about 25%.
- Urban people have a lower average PPI (49%) than non-urban (67%).
- Women have a slightly higher PPI (63% vs 59%).
- Married people have a slightly higher PPI (63% vs 57%).

Education

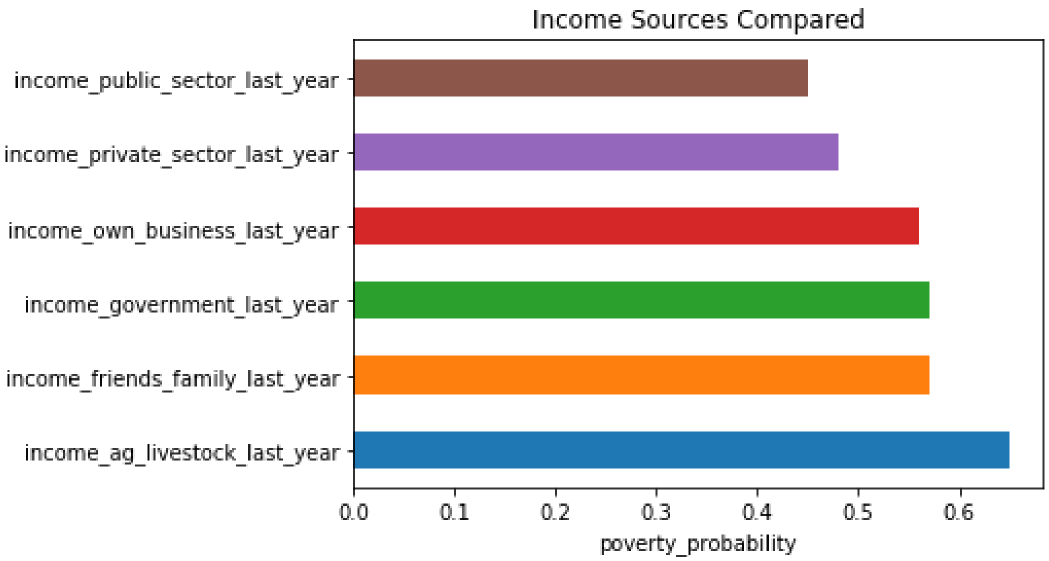
- Education is a major factor. Those with no education have an average PPI of 71%. Those with higher (college-level) education have an average PPI of 39% (which may still be considered high). Only 10% of individuals have higher education.
- Literacy does not protect against poverty as much as one might expect (57% vs. 68%).
- Math skills make little difference to PPI. The education category includes 4 math skills: the ability to add, divide, calculate percents, and calculate compound interest. A fifth category was created for comparison, representing having none of these skills.



Description of the Data

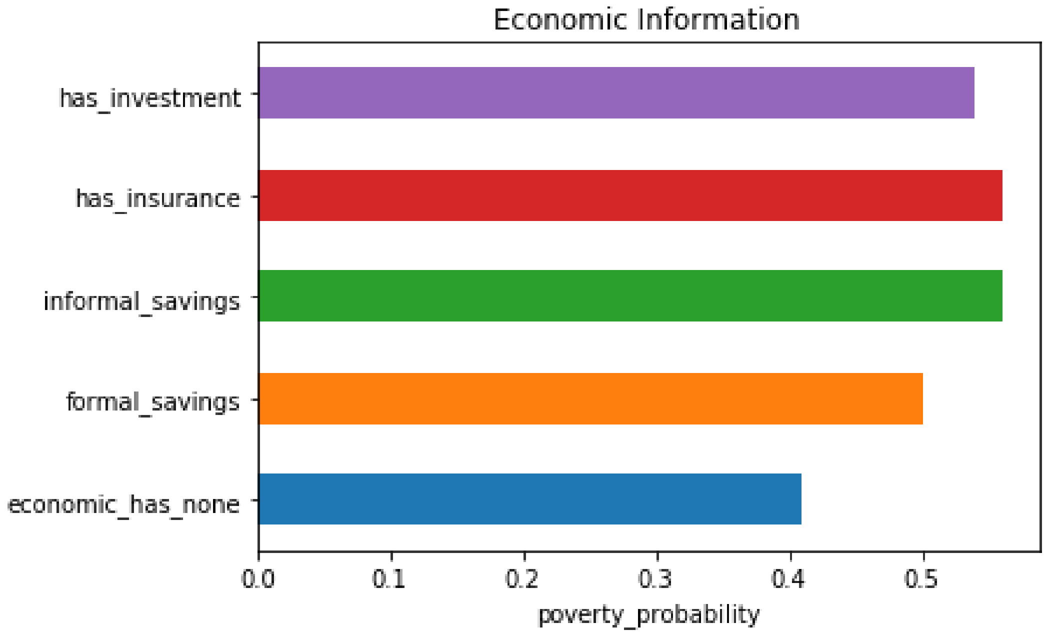
Employment

- Whether the individual was employed in the previous year did not make a significant difference in PPI. However, the type of employment for the previous year did. Salaried individuals had the lowest PPI (though still high at 51%). Seasonal-irregular workers had the highest at 70%.
- Similarly, the individual's sources of income were informative about PPI. The plot below shows average PPI for when the individual had the corresponding income.



Economic Information

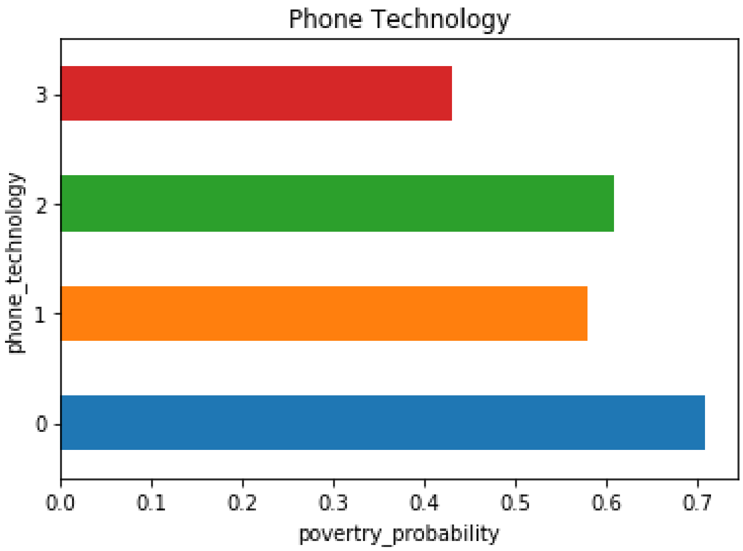
- The number of times the individual borrowed last year and the recency of borrowing were uninformative.
- Having formal savings, informal savings, an investment, or insurance of some kind made a substantial difference to PPI. To show this, a new column was constructed, *economic_has_none*, defined as having none of these other four forms of economic holding.



Description of the Data

Phone Technology

- The individual's level of phone technology was nearly as indicative of PPI as any other category. Individuals with a smartphone (coded as "3") had the lowest PPI; those with no phone ("0") had the highest.



Financial Inclusion

- Financially included individuals (having a registered bank account, mobile money account, or NBFI account) had a PPI 18 percentage points lower than those not included.
- Active bank users had a PPI roughly 12 percentage points lower than non-active bank users.

The project is to estimate the PPI from the given the Financial Inclusion Insights survey data. This survey data is not the same as the data used to construct the PPI in the first place. The PPI is an index that estimates the probability that a person lives in poverty, based on certain information (not available here). There is no guarantee that the dataset used here has adequate information to predict the PPI.

By way of analogy, consider the Body Mass Index (BMI), which is proportional to a person's weight divided by their height-squared. If one has weight, height, and some examples of BMI in a training set, one should be able predict BMI without much trouble. But instead, imagine a data analyst is trying to predict BMI using a different set of data from household surveys. Imagine the analyst knows how many times a person gets fast food per week, the person's shoe size, their family size, education, geographic region, and religion, as well as BMI for training cases. The analyst's chances of predicting BMI accurately from this information are not good. Like the BMI case, the data for this project are largely categorical and coarse-grained, so predicting PPI on this basis will probably not be very accurate.

Challenges

There are major challenges in attempting to estimate the PPI from the given dataset. While the target variable is continuous, the vast majority of the data is categorical. Unfortunately, the few usable continuous variables that exist in the dataset are uninformative about the target variable, meaning they correlate weakly or not at all.

The target variable is a probability with a range of possible values. But the data is more basic: Boolean or categorical. This is a weak basis for an inference to a continuous variable. Mathematically, the inference is still possible, depending on one's desired level of precision. But prospects are not good.

The data came in four data types. The numeric features consisted of floats (continuous variables) and integers. Natural categorical features are Booleans (True / False variables) and "objects" (text entries, usually denoting categorical values).

Data Type	Features
Boolean	37
Object	5
Float	8
Integer	8

Float-type features should represent continuous variables, and should be the best for regression. Integer-type variables can be valuable as well, particularly if they cover a wide range of values. There appear to be plenty of numeric features for regression—16 in total. But the analysis of these features tells another story.

Challenges

Gappy features

Of the eight float-type features, four of them were *mostly empty*. Recall that the total number of data was 12600. Here are the counts of null data for those columns.

Feature	Null counts
bank_interest_rate	12311
mm_interest_rate	12449
mfi_interest_rate	12399
other_fsp_interest_rate	12361

As a result, these features were dropped. At this point, there were still 4 floats, and 12 numeric features in total.

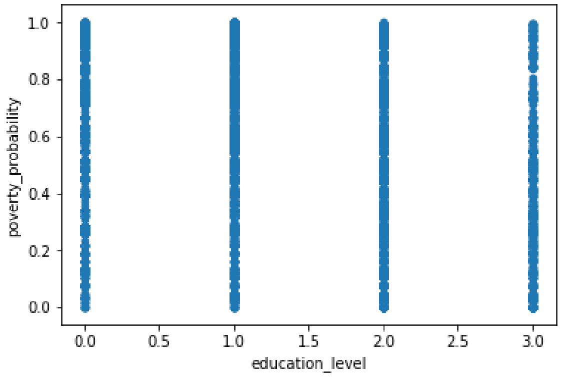
Misconstrued categoricals

Three features were encoded as numerics, but are actually categorical features. One example is *phone_technology*, coded as follows.

phone_technology values	
0	No phone
1	Basic phone
2	Feature phone
3	Smartphone

Although there is a natural ordering to this feature, it is better classified as a category than a number. (Compare this to paradigm numeric variables such as like “number of children” or “years in school.”)

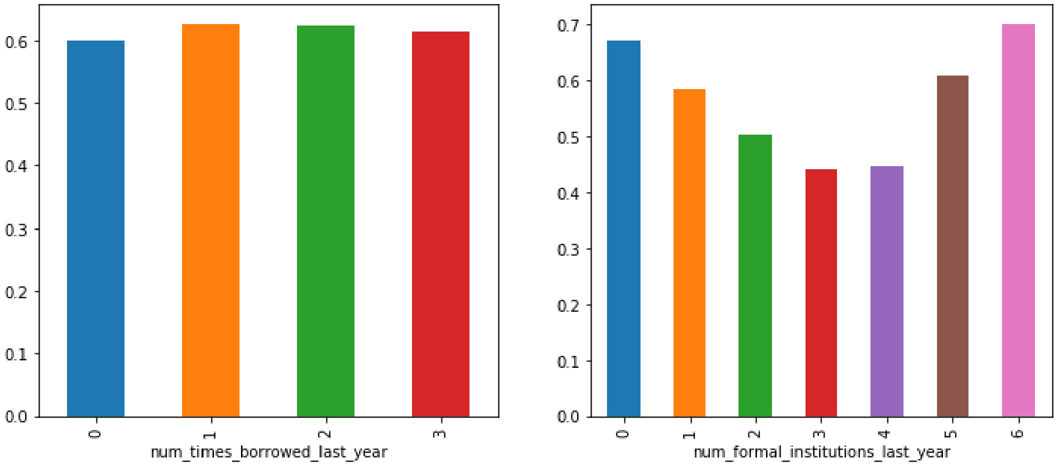
Additionally, two of the remaining features coded as floats were actually integers. A simple scatter-plot of *education_level* makes obvious that these entries do not fall on a continuum.



After accounting for these features, 7 numeric features were left.

Badly correlated integer-type features

Finally, many of the integer-type features were either uninformative, small in value count, or otherwise badly correlated with mean PPI. See the two features below, plotted as the mean PPI for each value.

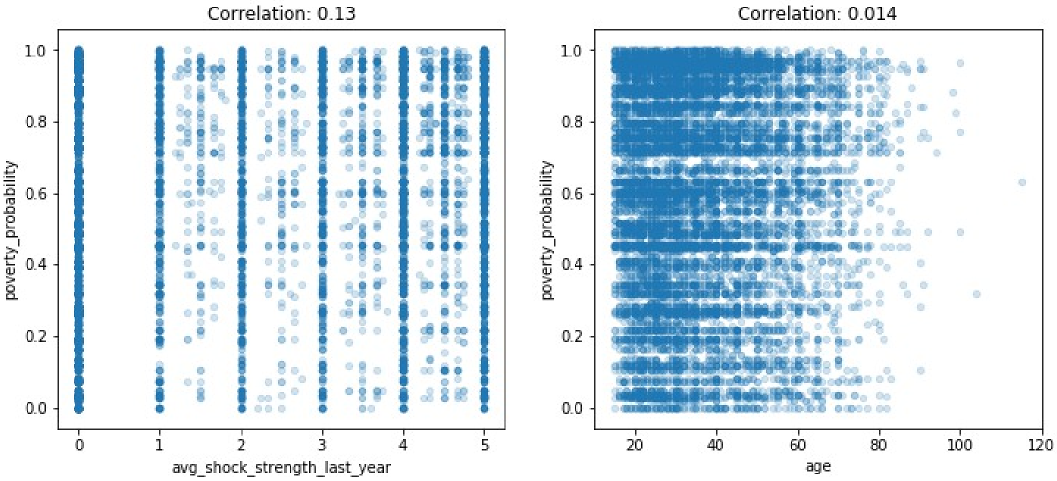


As the reader can see, *num_times_borrowed_last_year*, on the left, simply does not correlate with PPI—it is uninformative. *num_formal_institutions_last_year*, right, also does not correlate with PPI. It *is* informative, but not as a *numeric feature*.

Uninformative features like *num_times_borrowed_last_year* were eliminated on the basis of PPI differentiation; any feature for which PPI varied less than 5 percentage points were dropped. In addition, several badly correlated or non-monotonic integer type features were recoded as categorical features to capture the information they contained.

Result

After adjusting for these issues, only 2 float-type and 3 integer-type features remained. Neither float correlated well with the PPI. As a result, the data set provides *almost no numerical basis for regression.*



(It is clear from the plot above that *avg_shock_strength_last_year* is more like a discrete variable with some in-between values, but this does not effect the overall analysis.)

Process and Techniques

Initially, an in-depth analysis in Python was attempted. Yet Azure ML Studio proved the better source for testing a variety of high-powered algorithms. After the analysis of the data, there were only three parts to the process: cleaning, trying different algorithms, and fine-tuning the algorithms. A train-test split of 70-30 was used.

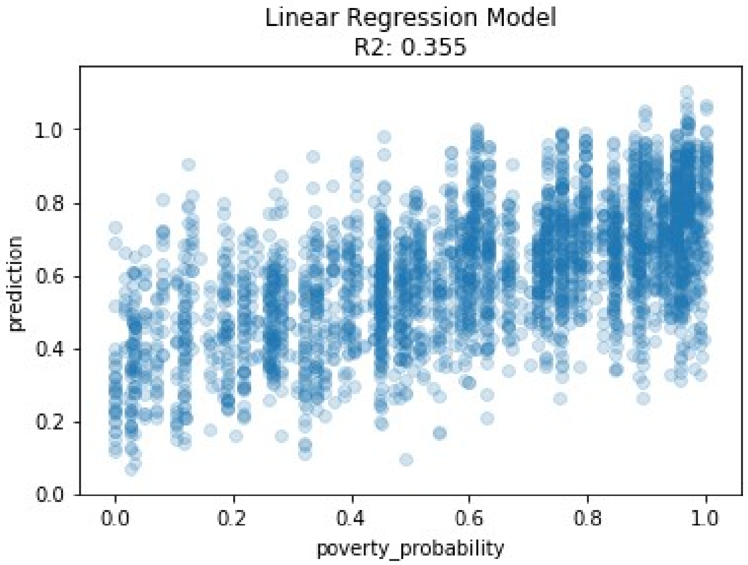
Process and Techniques

Data cleaning was done simply by removing the “gappy features,” or float-type features which were mostly empty (covered in an earlier section). Except for these columns, most were complete. Other rows with null values were dropped.

Algorithms Used

Linear Regression

Basic linear regression was attempted, yielding an R^2 of 0.355 (shown below). Ridge regression was applied to the linear regression model. After sweeping, an ideal *alpha* value of 1 was found, barely nudging the score to 0.357.



Linear regression has a poor degree of accuracy in this case. Yet it captures most of the predictive power in the dataset. The best model increases the score only by 20%, in relative terms.

Decision Tree Regressor

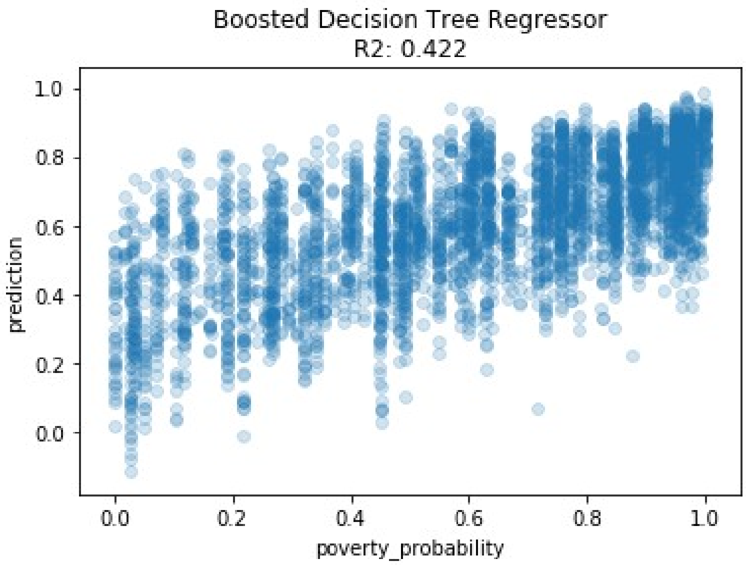
A decision tree regressor was tried on the data. A parameter sweep was tried on the decision tree across the maximum depth of the tree. The depth of 7 yielded the highest R^2 on test data. But the decision tree regressor was outperformed by the linear regression model.

Process and Techniques

Tree Max Depth	R^2
5	0.310
6	0.327
7	0.331
8	0.317
9	0.300
10	0.275

Boosted Decision Tree Regressor

A boosted decision tree regressor was tried in Azure ML Studio. The key to improving this algorithm's score proved to be the learning rate parameter. The top R^2 score was a 0.422, beating the goal. However, the trained model only achieved a 0.403 on brand new data.



The boosted decision tree regressor's predictions correlate just a little better than the linear regressor's, although visually the difference is subtle. In short, the performance of the best model was still weak.

Feature importance

Which features were most important in the final model? Of the large set of features, only seven accounted for a value greater than 0.01. Of these, only one, the individual's country is particularly indicative.

Feature	R^2 in model
country	0.285
education_level	0.075
is_urban	0.048
phone_technology	0.032
age	0.024
religion	0.022
num_shocks_last_year	0.016
married	0.011

The final set of feature importances yielded some interesting comparisons. *phone_technology* is an important feature at 0.048, but the related feature *can_call* is completely uninformative at 0.000. Whether the individual is married (0.011) is roughly five times as important as whether the individual is female (0.002). And *age* (0.024) is roughly ten times more important than *literacy*, or *can_add*, or *can_text* (all 0.002).[4]

Comparing R^2 values for individual features against their contribution in the final predictive model shows that the dataset has a high degree of information overlap. Features did not carry the same R^2 value in the predictive model as they did individually.

The table below focuses on the Phone features to compare these values. Note the first few rows. Practically speaking, *phone_technology*, or the type of phone the individual has, effectively determines the values for *can_call* and *can_text*. The importance values bear this out: while *can_call* and *can_text* are somewhat predictive individually, they have no predictive power once *phone_technology* has

[4] The fact that age has an R2 value in the final model greater than its correlation coefficient in linear regression shown earlier is a bit mysterious. It may be that its combined use with other features makes it more informative in the final mode.

already been incorporated. *phone_technology* itself has a much lower R^2 score in the model; it must covary with other features in the dataset in turn, such as *country*.

Phone feature	R^2 alone	R^2 in model
phone_technology	0.11	0.03
can_call	0.02	0
can_text	0.07	0
can_use_internet	0.08	0
can_make_transaction	0.05	0.01
phone_ownership	0.07	0.01
advanced_phone_use	0.06	0

Since the various features of phone capability are clearly related anyway, the information overlap here may be a special case. But the dataset as a whole clearly exhibits a fair amount of informational overlap or redundancy.

Key Insights

An R^2 score quantifies how much of the variation in the data, or the shape of the data, is captured by the predictions. At 42%, most of that variation is unexplained by these data. This supports the following conclusions:

1. Poverty (or at least the full *probability of poverty*) cannot be predicted well from this dataset here.
2. The dataset contains a fair number of features that are either uninformative or overlapping in information.
3. Performing regression with primarily Boolean and categorical data is unlikely to yield strong results. Usually, such data will not be rich enough in information for the job.

Key Insights

The question remains whether this data could predict actual poverty with fair accuracy, with “actual poverty” understood as a simple binary “True / False” variable. This goal would be less ambitious and may be achievable.