Analysis of Poverty Probability

ancai, July 2019

Executive Summary

With the advance of society, human basic needs are evolving and differ greatly between countries and regions. Because of this, examining poverty by only looking at income is no longer sufficient, as incorporating other indicators can give a better representation of the real living conditions of an individual (OPHI n.d.)[1].

This report presents an analysis of multiple socioeconomic indicators and their influence in determining which individuals have a higher probability of living below the poverty line, set at $2.50/day. The data used is comprised of 12,600 observations for people from seven countries, each containing the probability of living below the poverty line (which will be referred to as 'poverty probability' in this report) and 58 other variables, which represent information about their demographics, education, employment and economic situation.

The data was explored with the use of statistical methods, and visualisations were created for further insights. Several relationships between the socioeconomic indicators and poverty probability were identified, which were then used to create a regression model to predict the poverty probability for a new data set.

On the basis of this analysis, the following features can be highlighted for their high impact on the poverty probability of an individual:

- **Country**: one of seven countries, identified with letters. There is a large variance between poverty probabilities across countries, with countries A and D being poorer on average, in contrast to countries F and G, where people have the lowest poverty probability on average
- **Education_level**: the highest level of education of an individual. The poverty probability decreases significantly for people who have completed higher levels of education
- **Is_urban**: urban versus rural area of residence. Living in urban areas tends to indicate a lower poverty probability compared to people living in rural areas
- **Employment_type_last_year**: the type of employment in the past year. Salaried workers have the lowest poverty probability, while people conducting irregular, seasonal work have a much higher probability of being poor
- **Phone_technology** and **phone_ownership**: sophistication of phone type and the level of phone ownership for an individual. People who do not have access to a

1. OPHI n.d., *Policy – A Multidimensional Approach*, OPHI, accessed 15 July 2019, <https://ophi.org.uk/policy/multidimensional-poverty-index/>

phone have a high poverty probability, while owning a more sophisticated phone indicates a lower poverty probability

- **Num_financial_activities_last_year**: how many different types of financial activities conducted by an individual in the last year. This variable can be a proxy for financial knowledge, and indicates that the more financial activities a person conducted in the last year, the least likely it is that they are poor
- **Regular_bank_acct** and **active_bank_user**: The people in these categories tend to be less likely to be poor, compared to people who use non-banking financial institutions or over-the counter infrequent activities

Data Exploration

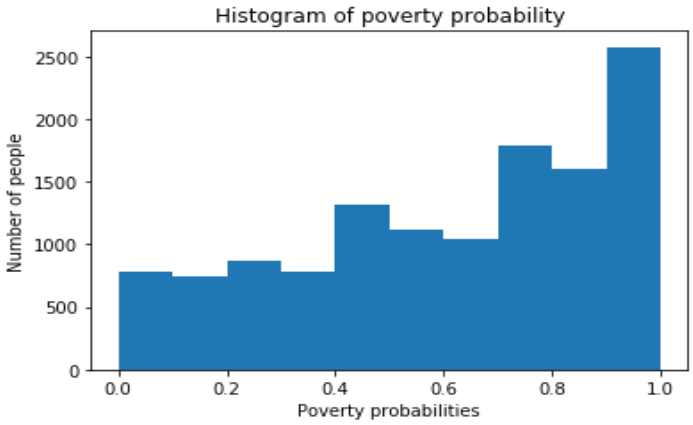
An initial look at the data showed 12,600 rows of observations, each containing 58 variables and a corresponding poverty probability. The variable types found were boolean (predominantly), with some numerical and categorical features.

Missing data was addressed at this stage:

- rows with missing entries for *share_hh_income_provided* and *education_level* were deleted as they represented a very small proportion (~450 missing entries), and attempts to fill with mean or other values led to contradicting insights which could then lead affect predictions
- the four interest rate columns (*bank_interest_rate, mm_interest_rate, mfi_interest_rate, other_fsp_interest_rate)* were dropped from the analysis due to the small number of entries filled (only ~200 out of 12,600 for each)

Then, the dataset was explored in detail with a focus on identifying relationships and patterns which could be used to build a predictive model for poverty probability.

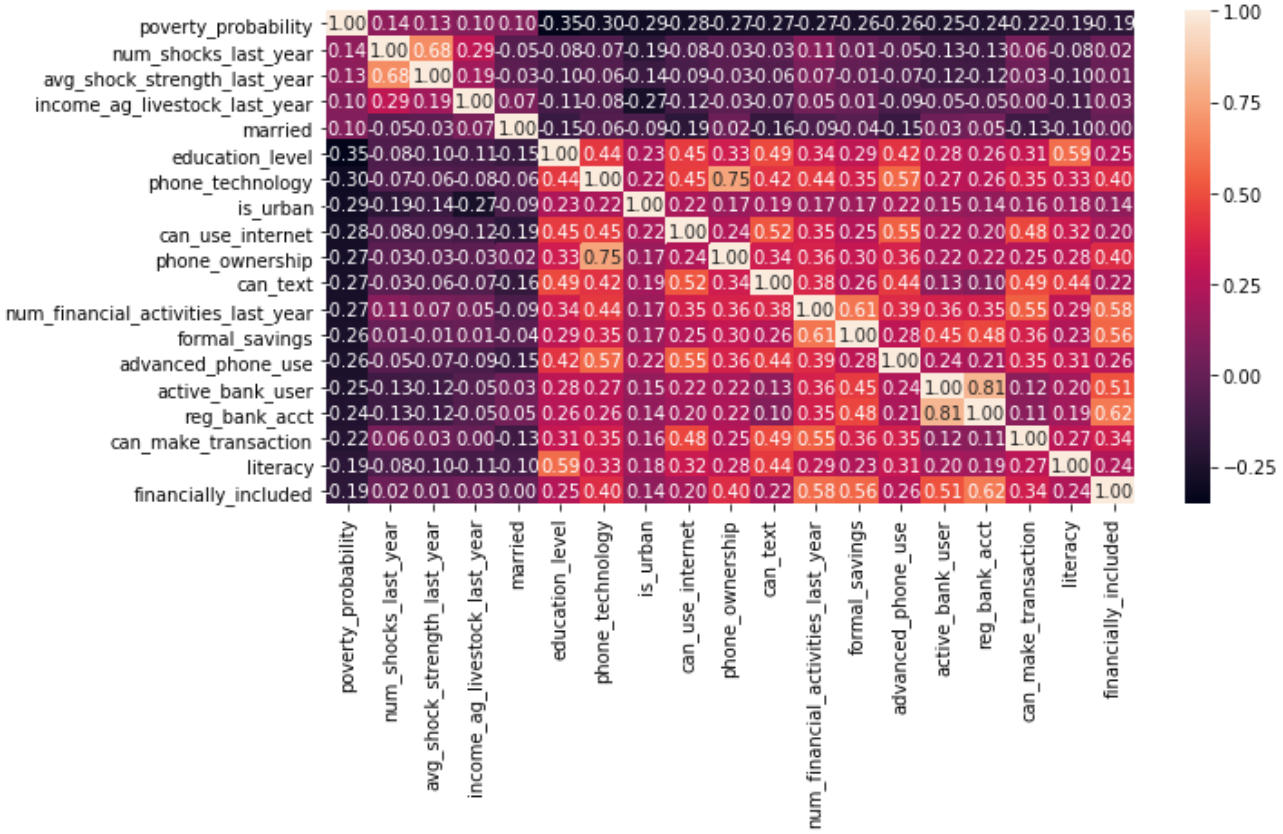
Starting with the variable of interest, it was noted that *poverty_probability* contains values in the range of 0 to 1, with the mean and median close in value, at 0.61 and 0.63 respectively. The standard deviation of 0.29 indicates a large variance between the observations, and it is noticeable that the histogram shows a left-skew, as a significant percentage of people have a poverty probability in the range of 0.7 to 1.



The other variables were explored in relationship with the poverty probability data and with each other. A very small number of outliers was found, and they were removed in order to obtain a more generalised picture of the data. The most important observations will be discussed in the following subsections, which will look at the variables grouped by type.

Initial correlation analysis with matrices and heatmap

Using a Pearson correlation matrix and heatmap in Python, the variables with the highest positive and negative correlations with *poverty_probability* were identied (used only for numerical and boolean features):



The heatmap indicates the following:

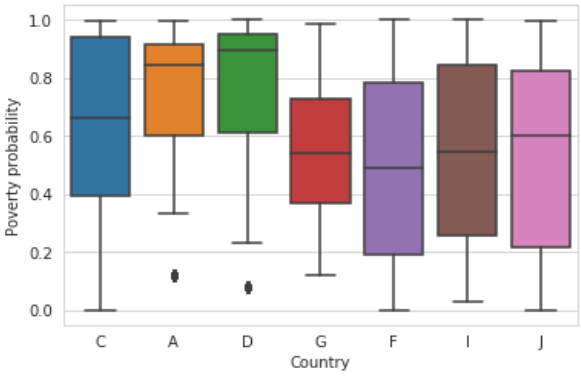
- ***education_level*** has the highest negative correlation with ***poverty_probability*** of -0.35, followed by ***phone_technology*** (-0.30) and ***is_urban*** (-0.29)
- key variables which have a direct relationship with ***poverty_probability*** are ***num_shocks_last_year***, ***income_ag_livestock_last_year*** and ***married***
- significant correlations were detected between some variables, for example between ***education_level*** and ***literacy*** (0.59), ***phone_technology*** and ***phone_ownership*** (0.75), and between the financial indicators ***reg_bank_acct*** and ***active_bak_user*** (0.81)

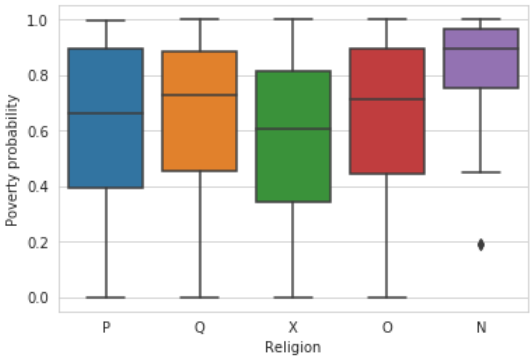
Variables with mid-range correlations were also examined individually to identify those that bring the most additional insight into predicting poverty probability, while avoiding multicollinearity problems that arrise with the use of many (correlated) variables.

Categorical Features Analysis

Country

There are seven countries coded with letters, with almost equal fequency in the observation set. Countries A and D show the highest poverty probabilities on average (over 0.8) and some outliers, while country F has the lowest average poverty probability of 0.49.





Religion

There are 5 religion types, which exhibit large variance in poverty probability.Religions Q and X account for almost 10,000 of the people, while the other religions have much lower frequencies

Relationship_to_hh_head (household head)

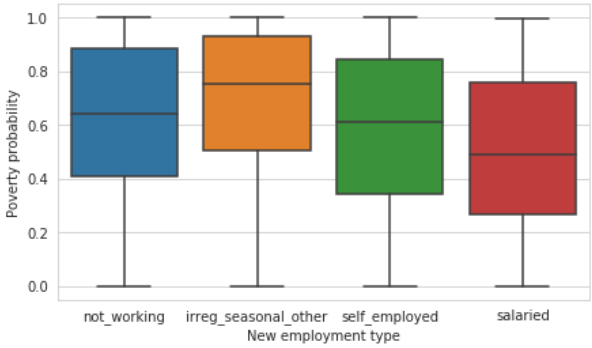
Most people in the observation set are either the head of the family or a spouse (75%). The children tend to have the same poverty probability as the household head, while the spouses show a higher poverty probability on average of 0.72, compared to 0.63 for the head of the family.

Employment_type_last_year

This is the most differentiating employment feature and shows a clear separation between categories in terms of the poverty probability.The categories 'iregular seasonal' and 'other' have the lowest frequencies, and they also display similar poverty distributions. It was decided to merge them into a new category named 'irregular seasonal and other', resulting in fewer categories.

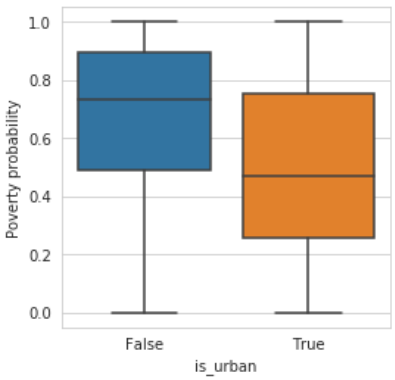
```
people['employment_type_last_year'].value_counts()
```

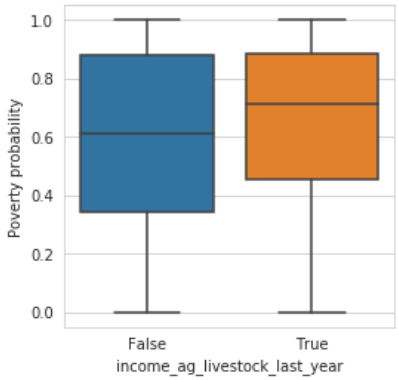
```
not_working           4513
self_employed         3113
irregular_seasonal    2349
salaried              1960
other                  665
Name: employment_type_last_year, dtype: int64
```

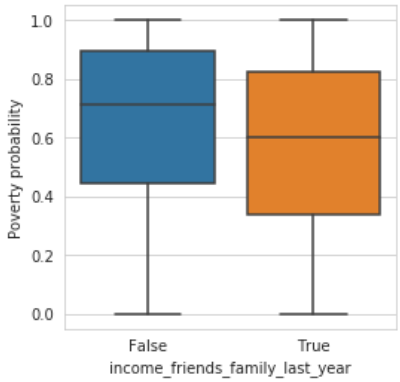


Boolean Features Analysis

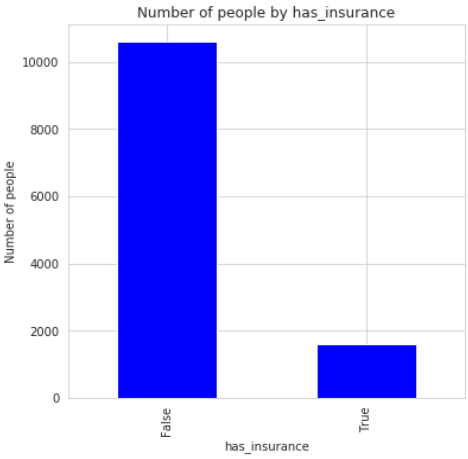
Boolean features were examined with box-plots and bar charts for frequency, and the following observations were made:

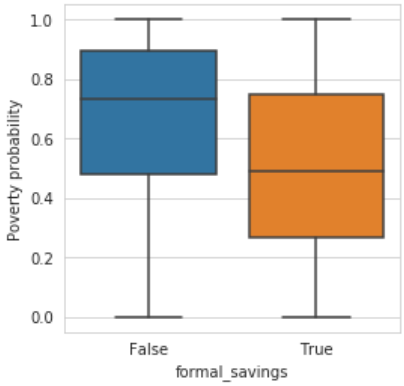


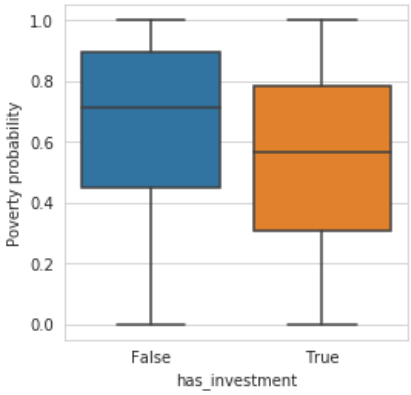




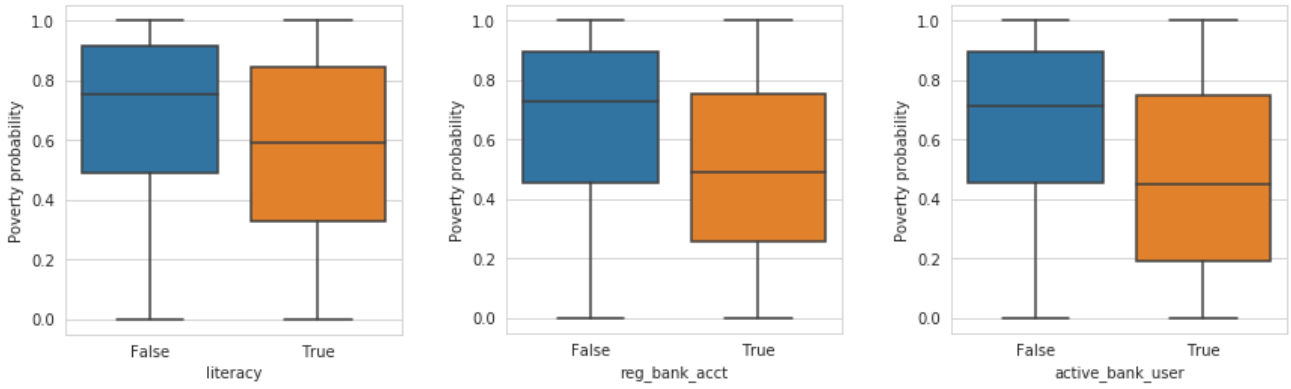
- Twice as many people live in *non-urban areas*, and on average people living in *urban* areas have a much lower poverty probability
- Almost twice as many people are *married*, and it appears they have a higher poverty probability on average compared to single people
- For *income* indicators: people who received income from friends/family, the private sector, their own business, the government or the public sector are less likely to be poor. The contrasting feature is income from agriculture and livestock
- People who *borrowed for emergency* or *daily expenses* are more likely to be poor on average, while *borrowing for home or business* expenses indicates a lower poverty probability







- The *savings* indicators show that more people keep their savings in cash and property rather than with a formal institution, and people without a *formal savings* account are much more likely to be poor
- More than 10,000 people (87%) do not have any form of *insurance*, while those who do have a lower poverty probability
- Having at least one form of *investment* is an indicator for a lower poverty probability



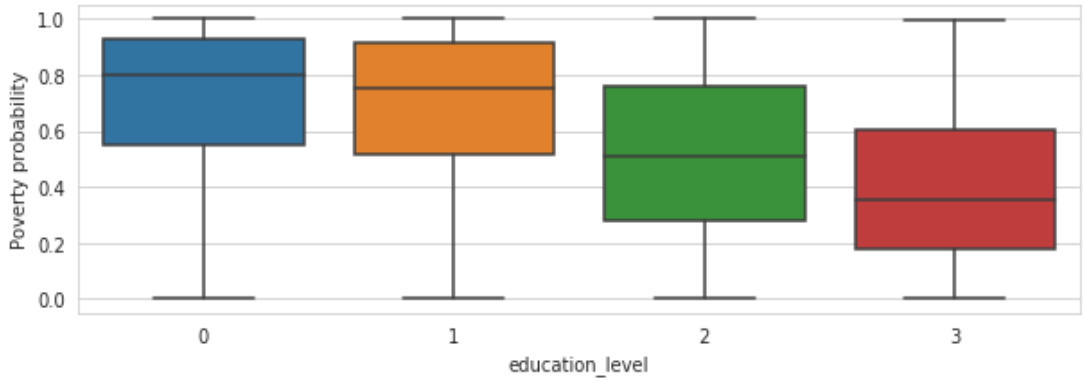
- Most people are *literate* (67%), almost all can add and divide, and around 40% can calculate percentages and compounding - all of these features indicate a lower poverty probability
- *Financial inclusion* is 50% across the people in the observation set, though other variables show large variance in poverty probability depending on the type of institutions used and frequency of the financial activities conducted
- Less than a 30% of the people has a *bank account* in their own name, and even less have used it in the last 90 days. These indicators show a high gap in terms of poverty probability between people who are (active) bank users and those who are not
- Similarly, about 30% of the people has a *mobile money account* in their own name, with fewer people who use it actively; the poverty probability for this category is also lower on average

Numerical Feature Analysis

The following variables represent numerical observations or rankings, and they have been considered numerical for the purpose of this analysis.

Education level

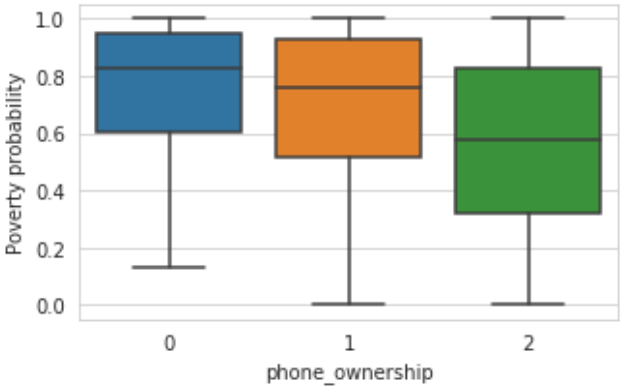
There are four levels of education: 0=no education, 1=primary education, 2=secondary education, 3=higher education. The highest level of education for most people is primary (37%), followed by secondary (32%); 20% of the people have no education, and 10% have completed higher education. There is a clear inverse relationship between poverty probability and the education level on average, which can be seen in the box plot analysis:

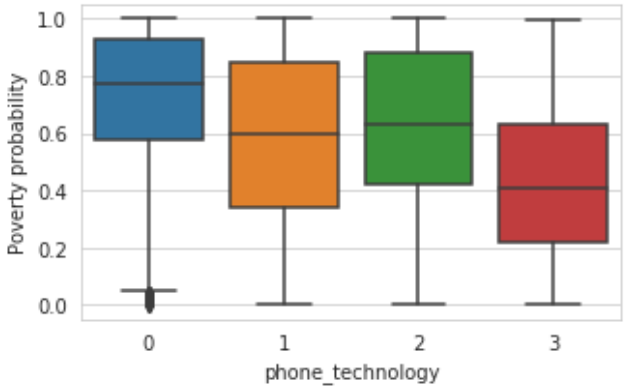


Phone technology and ownership

People with no phone (label 0) represent 35% of the sample data and have the highest poverty probability on average, of 0.77. There is an inverse relationship between *phone_ownership* and *poverty_probability*, as sharing (label 1) or owning a phone (label 2) decreases the poverty probability on average.

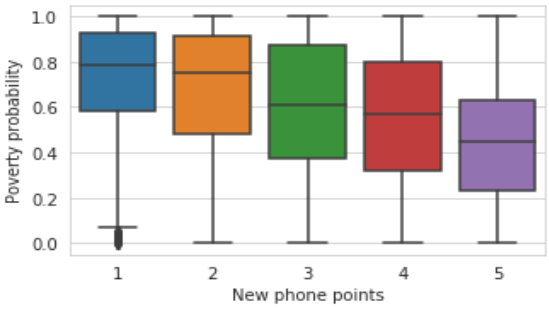
Smartphone technology (label 3) indicates a significantly low poverty probability on average. The remaining people represent half of the observation set, and own either a basic (label 1) or a feature (label 2) phone. It is worth noting that a feature phone shows a slight increase in poverty probability compared to people with basic phones, therefore the relationship between *phone_technology* and *poverty_probability* is not linear.





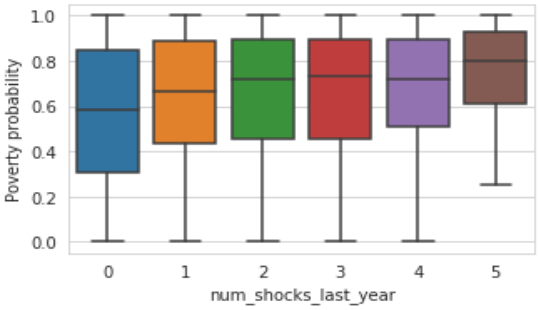
Phone points

There is a significant correlation between *poverty_probability* and the features *can_call, can_text, can_make_transaction* and *can_use_internet*, and these have been combined into a new field which adds how many of these functions can a person complete with their phone, leading to a new variable named phone points.



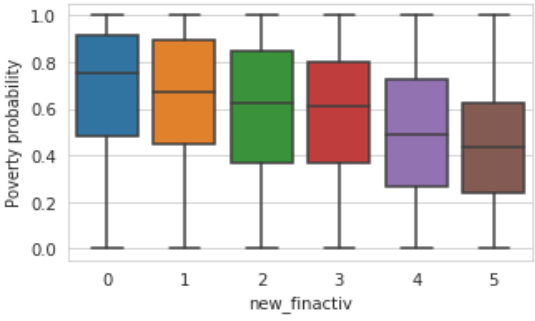
Num_shocks_last_year

There is a positive correlation between the number of financial shocks experienced by a person and their poverty probability, although the number of people who have experienced more than 3 shocks is much smaller (around 15%).



Num_financial_activities_last_year

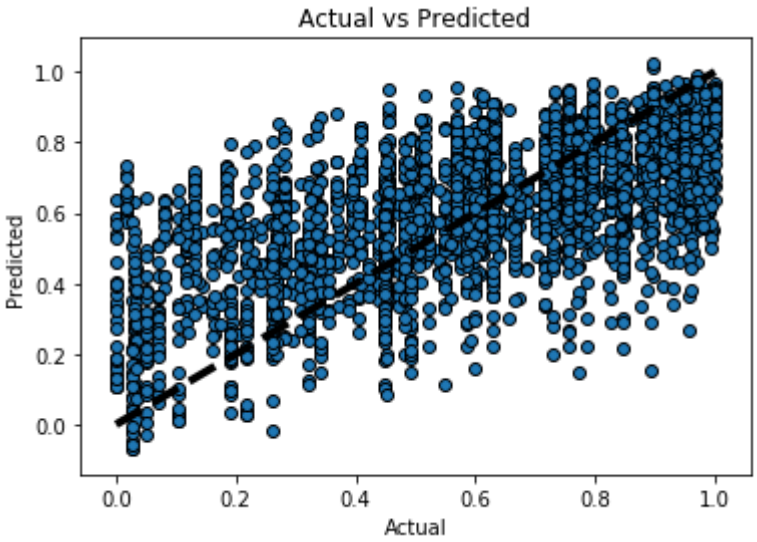
As the number of financial activities increases, the poverty probability decreases, however the frequency of the data becomes significantly smaller for the 5-10 financial activities range. To improve this, the number of financial activities over 5 have been grouped, leading to a more linear and balanced category.



Predictive Model for Poverty Probability

After exploring the relationships between the variables in the dataset and cleaning the data, several predictive models to calculate the poverty probability for a new data set were created. Regularised linear models such as Lasso were used to compare the features selected during the data exploration stage with the features selected by the model, and this further improved results. To obtain the best predictive power on the dataset, a gradient boosting model was used.

The model was trained with a random split of 75% of the data, and tested with the remaining 25%. The results obtained with the gradient boosting regressor were good, but slightly lower on the public leaderboard. Cross-validation was also performed to ensure the model accuracy is consistent and the parameters of the model were adjusted, which led to further increases in the public score in the mid 0.40-0.41 range.




```
R2 sq train:  0.6009305730257197
Mean squared error: 0.05
Test Variance score: 0.43
```

It can be seen that the residuals follow the trend line loosely, as there is moderate prediction power for *poverty_probability* from the features provided alone. This could mean that there are other factors which influence whether or not a person lives below the poverty line, which are not captured in the model or the data.

Conclusion

The analysis has examined the relationship between poverty probability and several socioeconomic features, for a data set of observations for individuals across seven countries. A machine learning model was used to predict poverty probability using the data provided, which achieved moderate prediction confidence. Several key features have been identified for their high impact on the poverty probability of an individual, in particular their country and area of residence, education level, employment type, and technological and financial status. Secondary features such as marrital status, income sources and economic information provide further insights which can improve prediction and define a more complete context to understand a person's probability of living in poverty conditions.