Predicting Poverty Around the World

Sandipan Bhaumik, July 2019

Executive Summary

This document presents an analysis of the data concerning poverty around seven different countries and the probability of an individual living below the poverty line at $2.50 per day threshold given various socioeconomic indicators. The goal of this exercise is to predict a probability variable called `poverty_probability` (between 0 and 1) for 8400 individuals based on observations made from 12600 known cases.

On performing some descriptive data exploration on the given dataset using various statistical calculations and data visualisations, several potential relationships between the indicators (or features) and the chance of an individual of living below the poverty line, i.e. `poverty_probability` were discovered. Finally, a regression model was trained on the known cases using the relevant features to predict the `poverty_probability` of 8400 individuals provided in the test dataset.

After performing the analysis, the author presents the following conclusions:

While many factors can help predict the `poverty_probability` value, the significant features found in this analysis were (25 features were used to train the model, the features below had the maximum score calculated using the *Permutation Feature Importance* component):

- **country** – the country the respondent lives in. There is a clear difference in poverty stats for different countries.
- **education_level** - Highest level of education. Individuals with a higher education were at lower risk of falling under the poverty threshold as opposed to the ones with no education.
- **is_urban –** Area of residence. Individuals from rural areas are more likely to be poor.
- **phone_technology** - Sophistication of phone type. Smart phone users are less likely to be poor.
- **employment_type_last_year** - Type of employment last year. Salaried individuals are safer economically as they hold a permanent source of income.
- **avg_shock_stregth_last_year** - Average strength of economic shocks experienced in the last year. People with higher shocks were economically unstable, thus being more likely to be poor.
- **active_bank_user** - Has used their bank account in the last 90 days. Active bank users seem to be financially included and at lower risk.
- **literacy** - Ability to read and understand. Individuals who could read and understand are at lower risk of poverty than the ones who couldn’t.
- **gender** – gender seemed to have played a part with the females facing a higher risk of poverty than the males.
- **num_of_times_borrowed_last_year -** Number of times the respondent borrowed money in the last year.

Initial Data Exploration

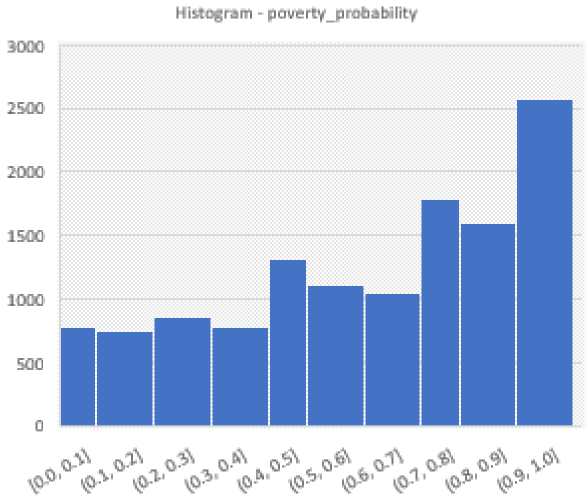
The initial exploration of data revealed some columns with missing values. The missing values in some columns were replaced either with ‘0’ or with a value derived from deeper analysis of multiple features. The TRUE/FALSE flags were converted to numerical 0/1 values. The data exploration then continued with some summary and descriptive statistics.

Individual Feature Statistics

While most of the columns in the observational dataset are of categorical nature, the summary statistics with mean, median, standard deviation, minimum and maximum were calculated for some of the numerical features from the 12600 observations that were made.

Column	Mean	Median	Std	Min	Max	Dcount
age	36.28	33	15.15	15	115	84
num_times_borrowed_last_year	0.66	0	0.92	0	3	4
borrowing_recency	0.87	0	0.96	0	2	3
bank_interest_rate	0.23	0	2.71	0	100	30
mm_interest_rate	0.11	0	1.78	0	100	33
mfi_interest_rate	0.17	0	1.89	0	100	33
other_fsp_interest_rate	0.16	0	1.84	0	100	25
num_shocks_last_year	1.10	1	1.19	0	5	6
avg_shock_strength_last_year	2.11	2	2.02	0	5	41
num_formal_institutions_last_year	0.71	1	0.81	0	6	7
num_informal_institutions_last_year	0.19	0	0.47	0	4	5
num_financial_activities_last_year	1.56	1	2.04	0	10	11

A histogram of the `poverty_probability` column shows that the probability values are left-skewed, which infers larger number of individuals are likely to be poor given the socioeconomic indicators.



In addition to the numerical variables, the observation data contains several categorical and binary features. The following features were treated as categorical features in training model.

- **country** – each country is represented by a letter
- **religion** – there were 5 religions in the dataset, each represented by a letter
- **relationship_to_hh_head** – this column represents the relation of the individual to the head of the household
- **education_level** – ordinal value representing the highest level of education (0=no education, 1=primary education, 2=secondary education, 3=higher education).
- **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** – the share of household income provided by the individual. There were missing values in this feature which were replaced by 0.
- **phone_technology** - Sophistication of phone type (0=no phone, 1=basic phone, 2=feature phone, 3=smartphone)
- **phone_ownership** - Phone ownership (0=no phone, 1=shares phone, 2=owns phone)

Bar charts were created to analyse the frequency of the categorical features across the dataset and the following observations were made:

Demographics -

- Country J had the highest number of individuals in the observation data, while countries C and F had the least.
- Around 68.6% of the respondents were from rural areas.
- The data collected were distributed uniformly between age groups 25 – 85 years.
- Religion Q had the highest number of individuals followed closely by X. These two are the religions with most of the individuals in the observation dataset.
- Females responded in more numbers than the males and around 65% of the total were married.
- Majority of the respondents were the head of the household themselves, followed closely by the spouse.

Education -

- While majority of the respondents were literate, a good number (~30%) of them were not.
- Females were found to be less literate than males.
- A smaller number of individuals pursue higher education, while most of them complete primary education in these countries.
- A good number of people who can add, can divide as well. But cannot calculate percentage or compounding.

Employment -

- Most of the respondents were employed the previous year and a good number of them were self-employed. The head of the household contributes to the income mostly.

Economic –

- Majority of the respondents do not have a form of formal, informal or cash property saving.
- Majority of the respondents do not have any insurance or investment.
- Around 31% faced at least one financial shock the previous year.

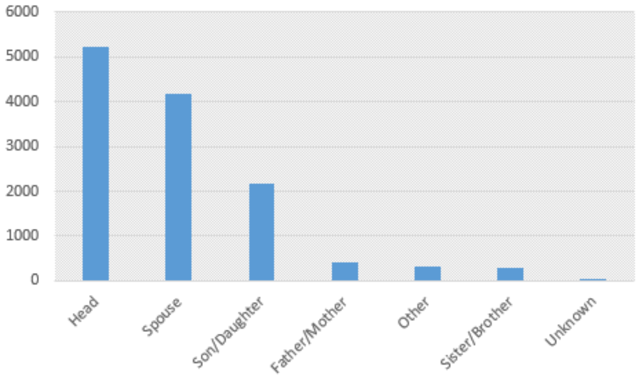
Phone –

- Of the people who own phone, 25% use a smartphone.
- Only 49% of the smartphone users can do the advanced tasks on the phone.
- While most of the phone owners can call, only a small number can either text, use the internet or make transactions.

Financial Inclusion –

- 51% of the respondents are not included financially, i.e. they do not have registered bank account, a registered mobile money account, or a registered NBFI account.

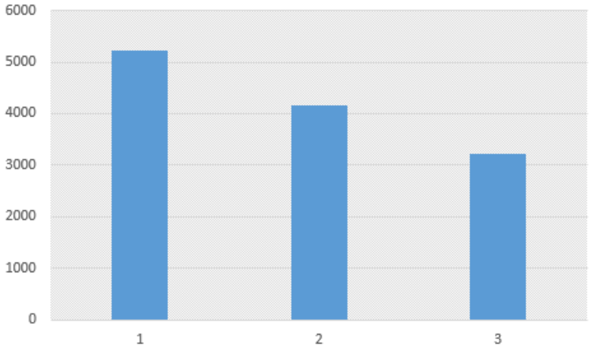
A couple of categorical features were engineered to reduce the number of categories in order to help the model function better.
In the relationship_to_hh_head column, 'Head, 'Spouse' and 'Son/Daughter' have significant records, the other values have low frequency.



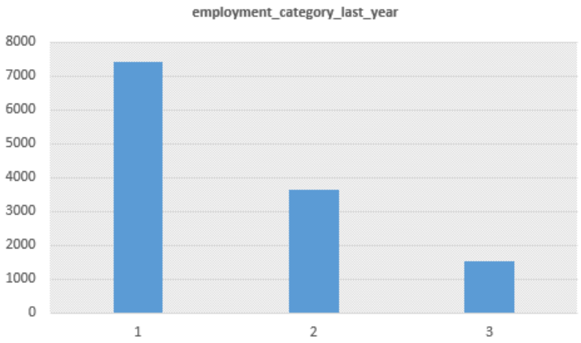
It was decided to reduce the categories to three numerical groups:

1 – Head
2 – Spouse, Son/Daughter
3 – All other categories

This resulted in a smaller group of categories, as below:



A similar approach was taken with the column employment_category_last_year where 5 categories were reduced to three categories.



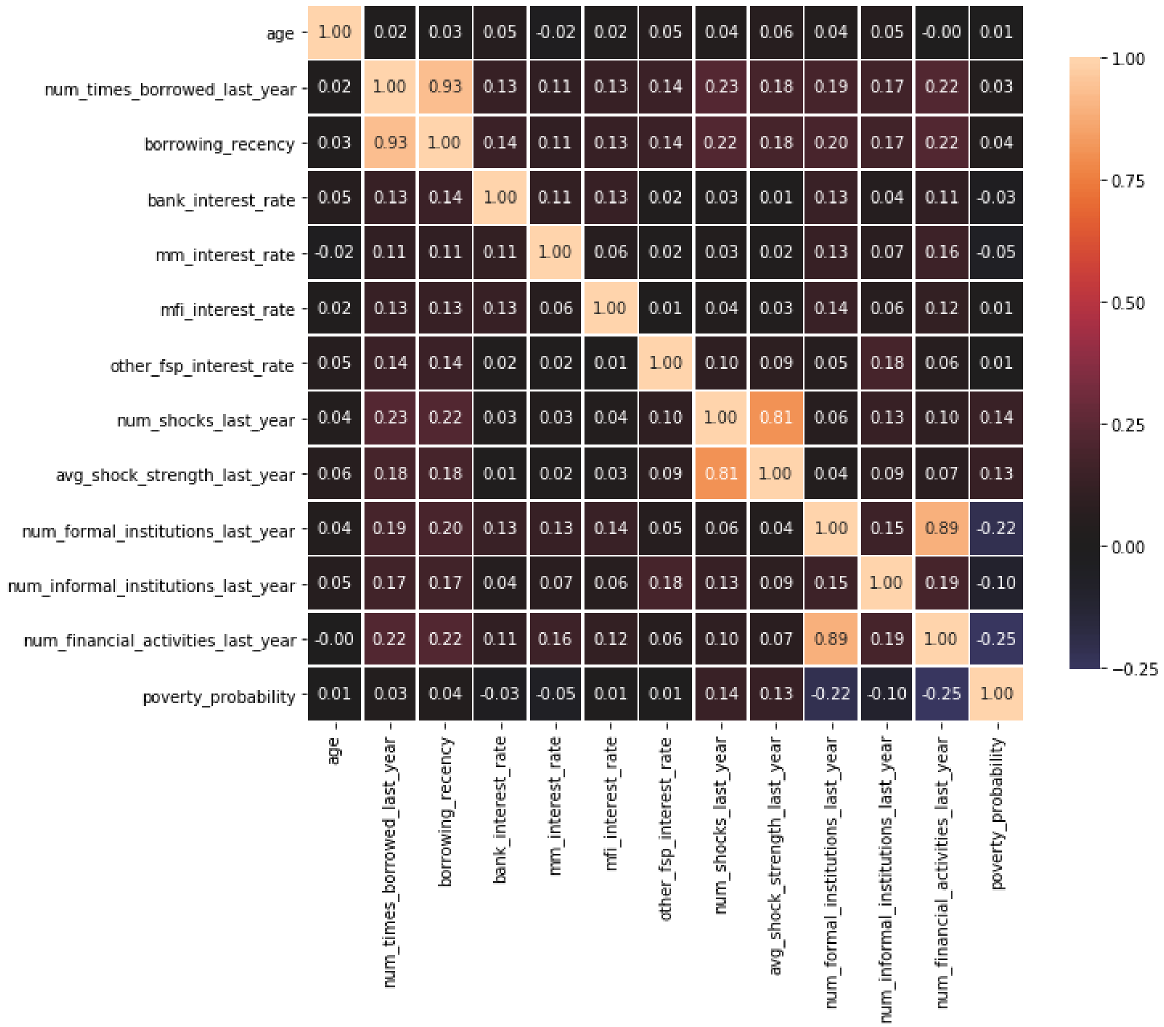
Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between the features of the data – in particular, between poverty_probability and other features.

Numerical Relationships

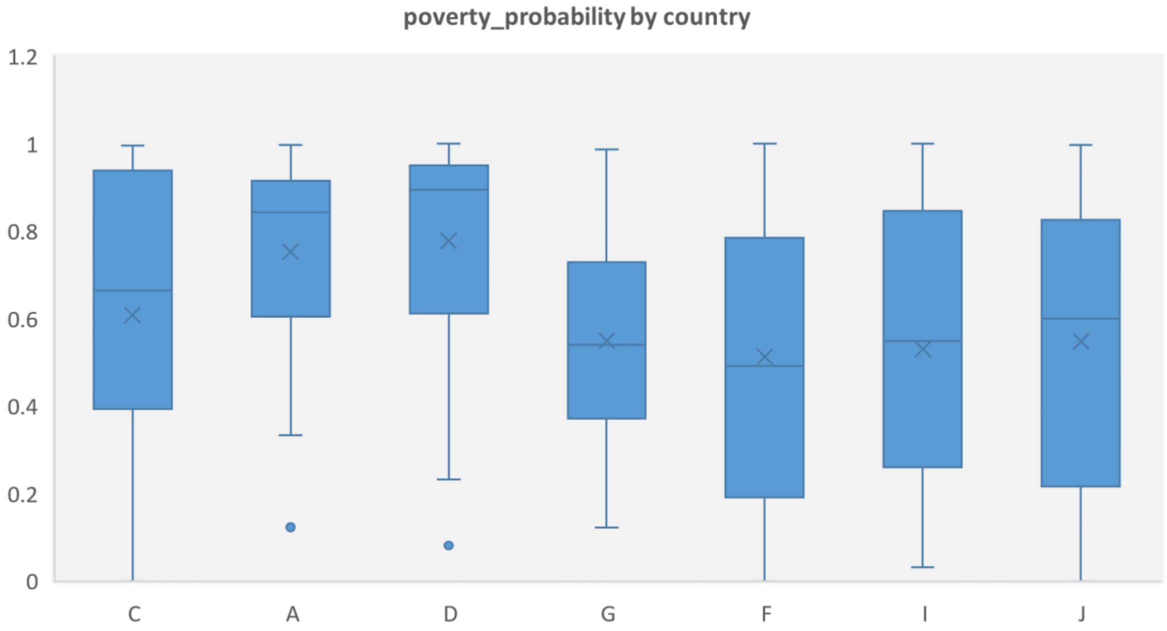
A correlation matrix heat map was generated to compare the numerical features with one another. The heat map below reveals that there is a low correlation between the numerical columns in general. A few columns are highly correlated, like borrowing_recency and num_times_borrowed_last_year, num_shocks_last_year and avg_shock_strength_last_year, num_financial_activities_last_year and num_formal_institutions_last_year.

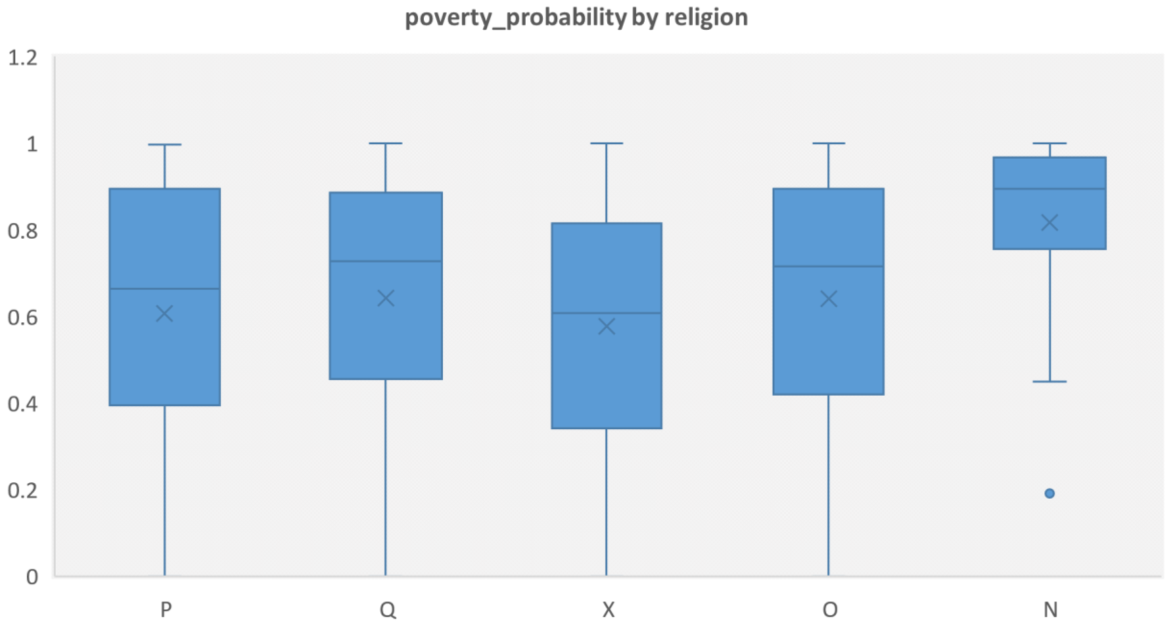
It can also be seen the following columns have a negative correlation with **poverty_probability**:
num_financial_activities_last_year
num_formal_institutions_last_year
num_informal_institutions_last_year

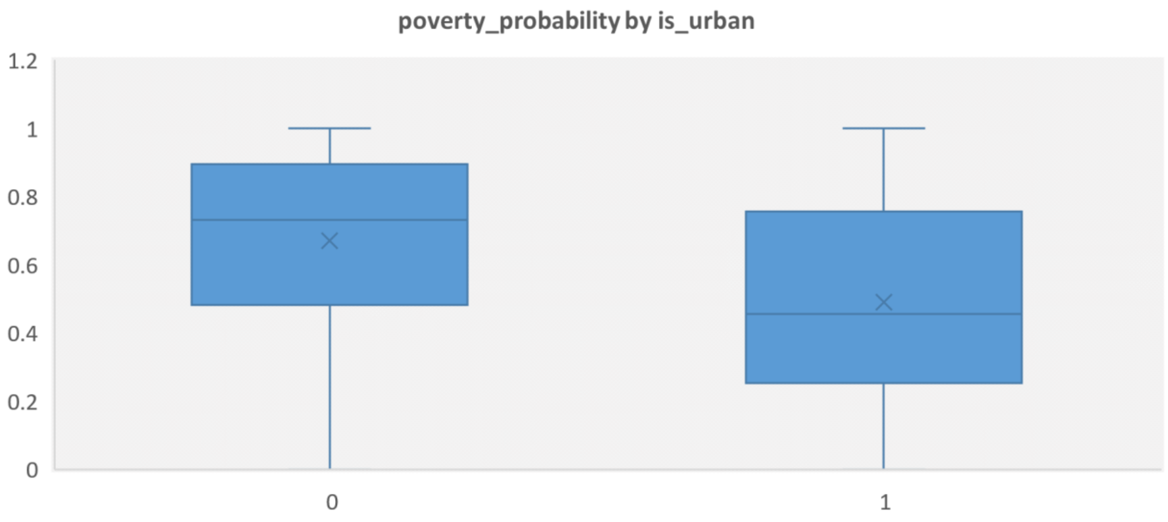


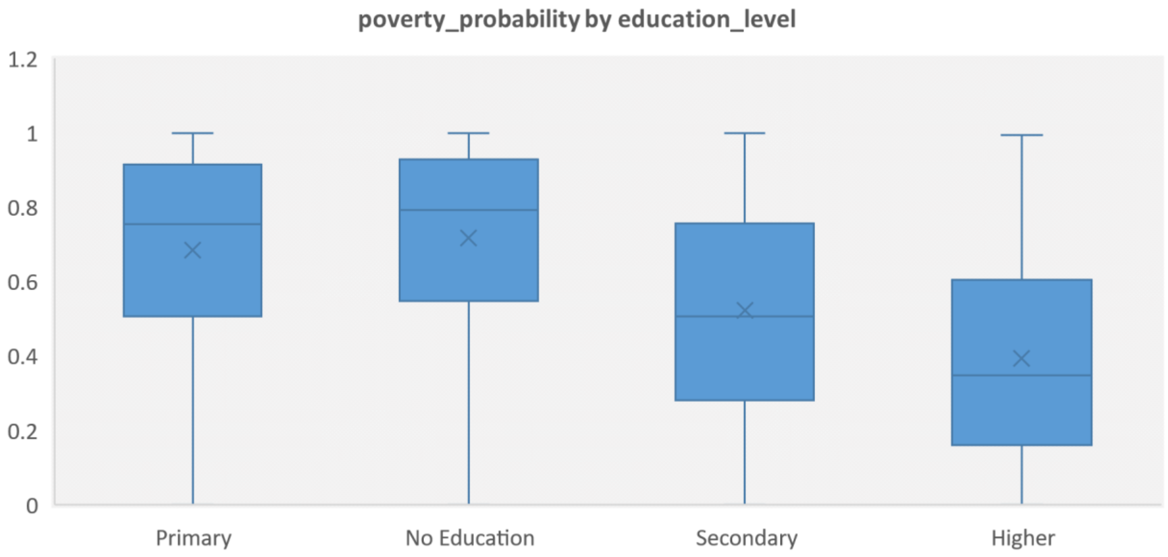
Categorical Relationships

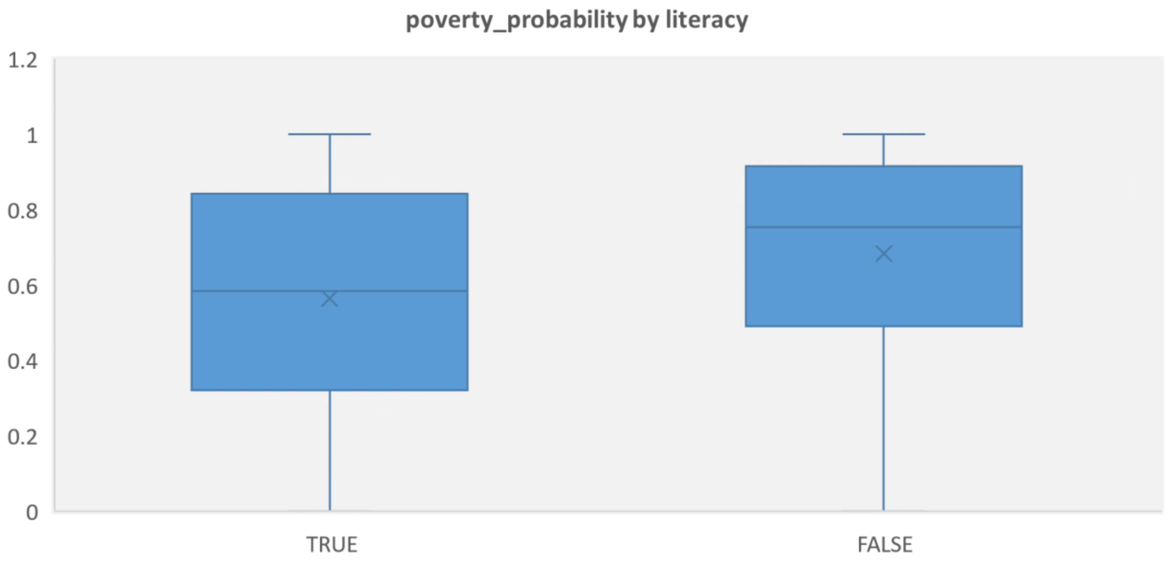
After exploring the relationship between the numerical columns, an attempt was made to identify any apparent relationship between the categorical feature values and poverty probability. The following box-plots show the categorical columns that seem to exhibit a relationship with the probability values.

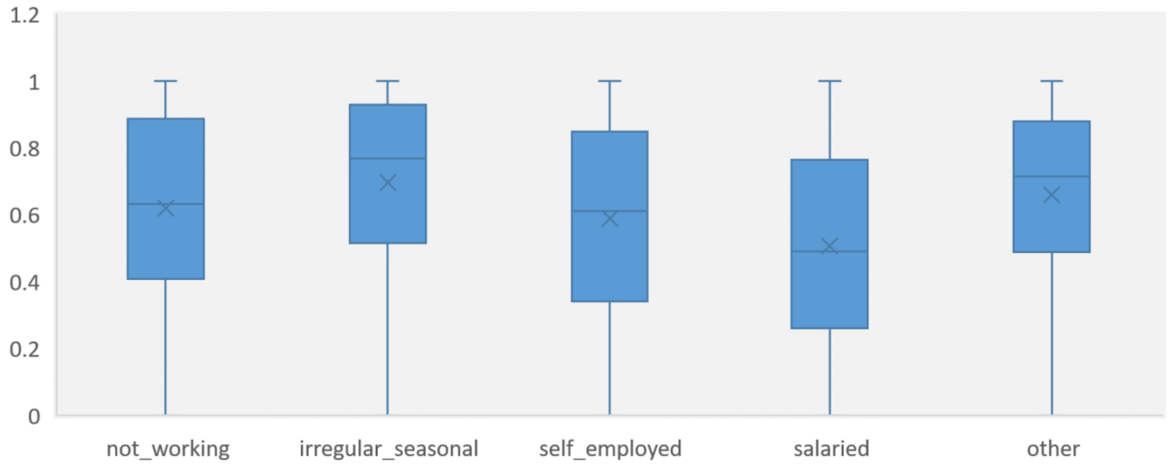




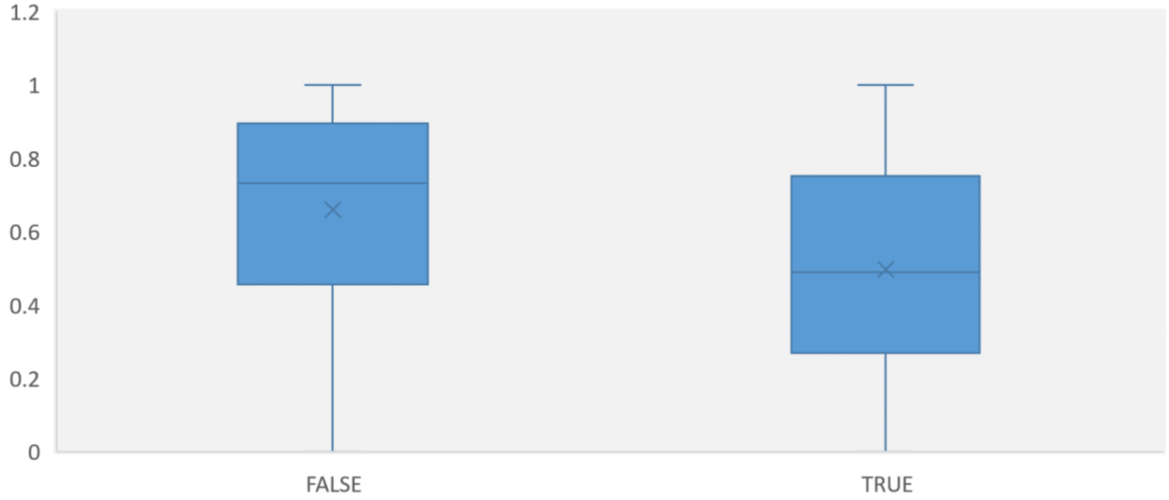

poverty_probability by is_urban
1.2
1
0.8
0.6
0.4
0.2
0
0
1


poverty_probability by education_level
1.2
1
0.8
0.6
0.4
0.2
0
Primary
No Education
Secondary
Higher

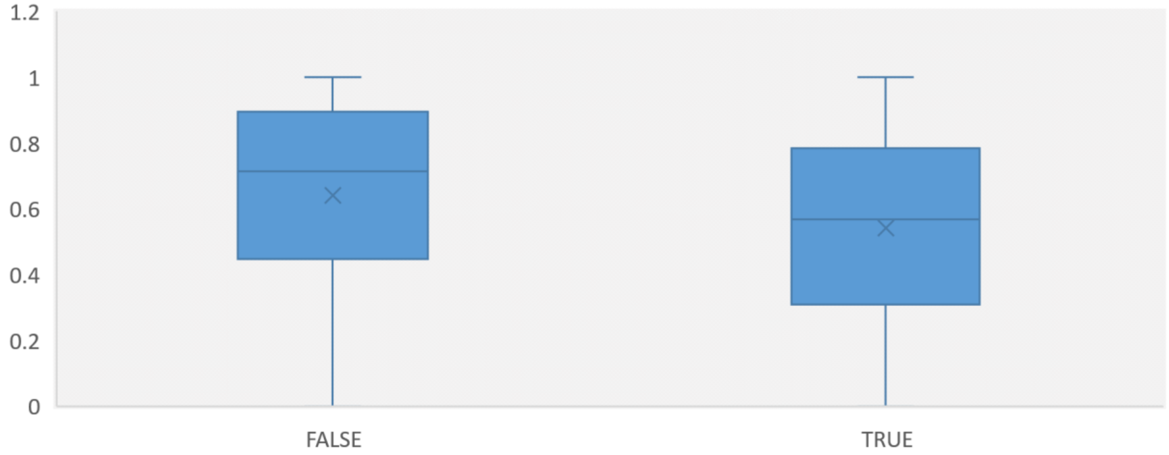

poverty_probability by literacy
1.2
1
0.8
0.6
0.4
0.2
0
TRUE
FALSE



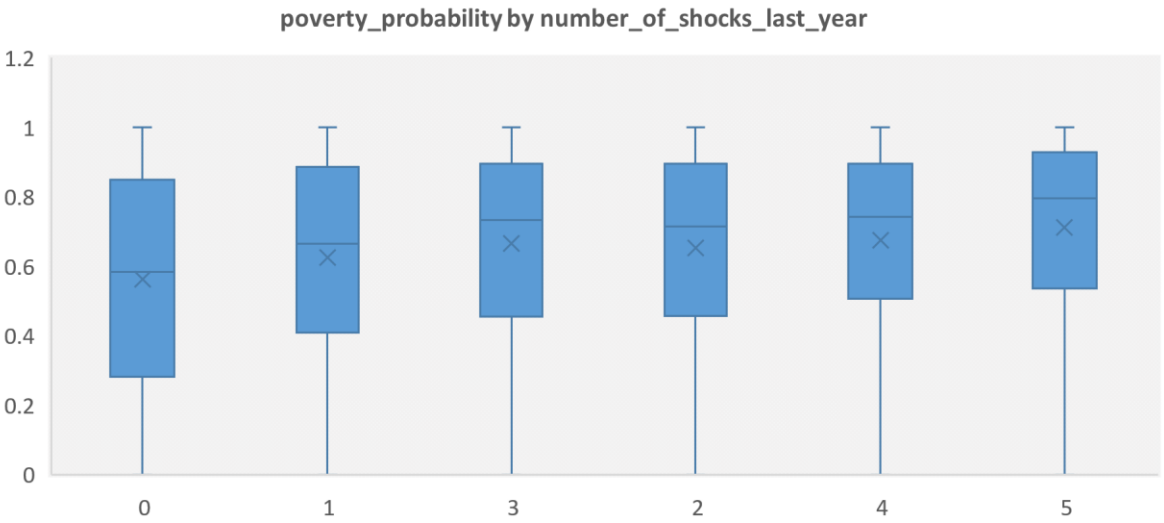
poverty_probability by employment_type_last_year

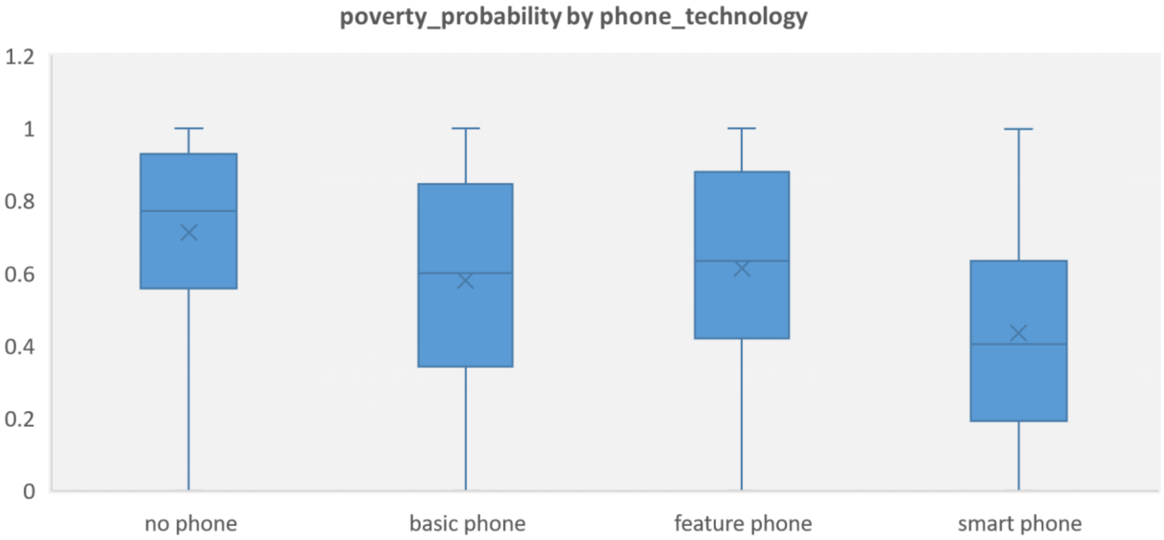


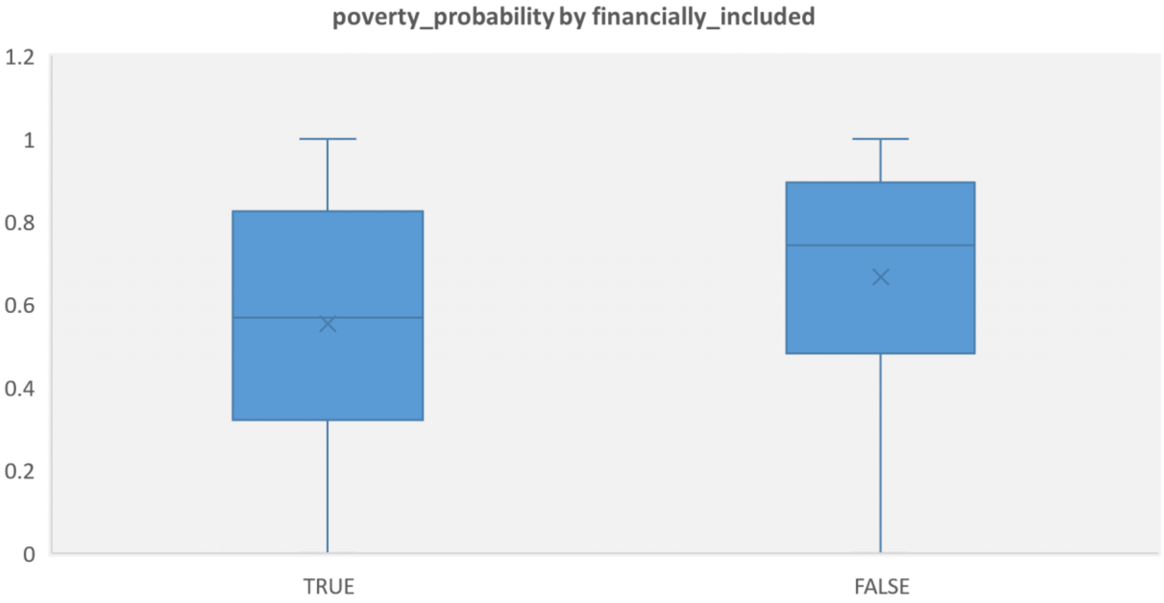
poverty_probability by formal_savings

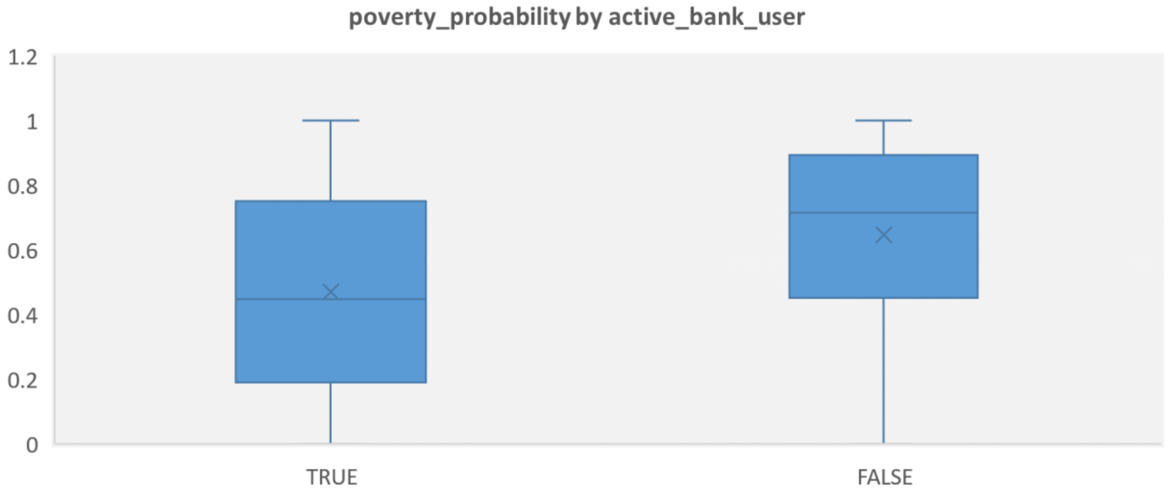


poverty_probability by has_investment


poverty_probability by number_of_shocks_last_year
1.2
1
0.8
0.6
0.4
0.2
0
0
1
3
2
4
5


poverty_probability by phone_technology
1.2
1
0.8
0.6
0.4
0.2
0
no phone
basic phone
feature phone
smart phone


poverty_probability by financially_included
1.2
1
0.8
0.6
0.4
0.2
0
TRUE
FALSE



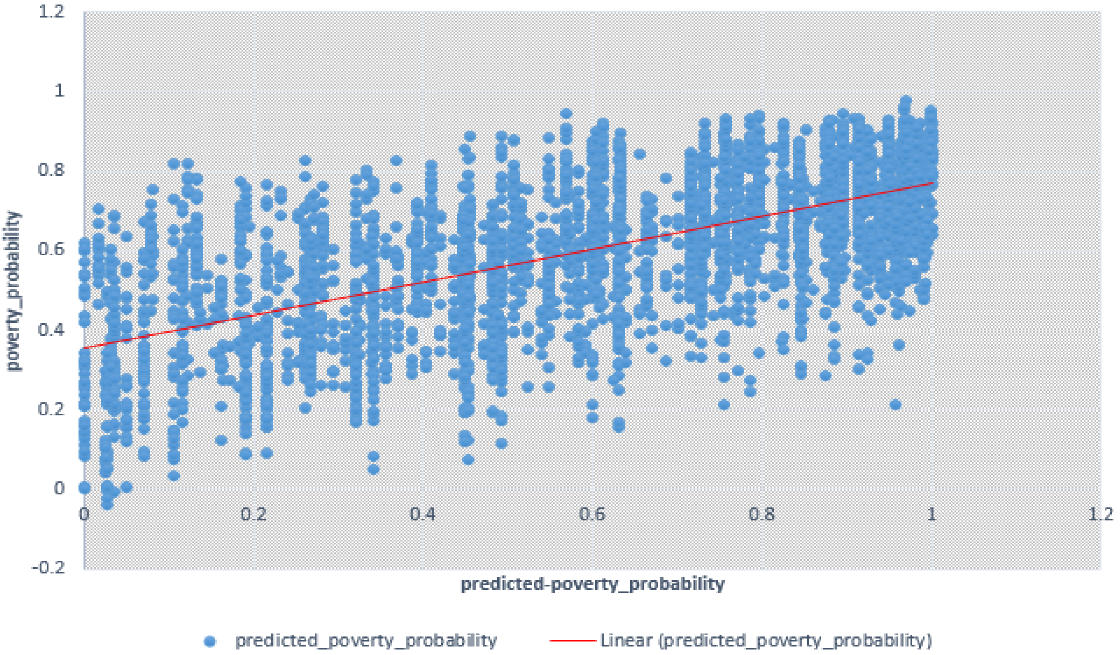
The box plots show some clear differences in terms of median and range of probability values for different categorical features. For example:

- Some of the **demographics** features clearly demonstrate the differences in distribution by coutry and religion. It was also observed that the respondents from the rural areas were more likely to be poorer than the ones who live in urban areas.
- **Education** plays an important role in one's economic prosperity as can be seen in the distribution by education_level and literacy. Individuals with a secondary or higher education seem to do well economically when compared to the ones who had no education.
- It is evident from the **employment** features that individuals with a salaried job last year had the least chance of falling under the poverty line.
- **Economic** stability definitely contribute to one's well being and some of the features rightly depict that. People with some form of formal savings, who have made some investment and faced lower economic shocks in the previous year, are less likely to be poor.
- Use of a **phone technolgy** relates to poverty, invividuals with phones, especially a smart phone, tend to be less poorer than the ones who do not own a phone. It shows how use of technology can contribute to financial health.
- **Financial Inclusion** is another aspect that contributes to one's well being. The data correctly reveals that – an individual who is financially included, e.g. has an active bank account and is an active bank user is less likely to be poor.

Regression

A regression model to predict the **poverty probability** of respondents was created, since the value to be predicted is of continuous nature. Based on the apparent relationships identified when analysing the data, a **Boosted Decision Tree Regression** model was created to predict the probability. A few other regression models were used in parallel during testing, but the boosted decision tree regression model achieved the best fit.

The model was trained with 70% of the data, and tested with the remaining 30%. A scatter plot showing the predicted probability and the actual probability is shown below:



Although there are outliers, yet the plot shows a linear relationship between the actual values and the predicted values for poverty_probability.

The model achieved a Coefficient of Determination value of 0.42 in training (cross-validated, with slight over fit). The test version of the model achieved a value of **0.41**.

Conclusion

The analysis has shown that the **poverty_probability** values could be predicted using the identified features. A low co-efficient of determination value does not necessarily indicate a bad fit. The graph above shows that even noisy, high-variability data can have a significant trend. It also depicts that larger number of individuals are more likely to be poor. The trend line indicates that the predictor variable still provides information about the response even though several data points fall further from the regression line.