Analysis of World Poverty Data

Microsoft Professional Capstone: Data Science

Simon Cox, July 2019

Executive Summary

This document presents the results of an analysis of data collected on individuals concerning more than 50 different demographic, economic, education and employment features of populations in different countries, and the probability that the individual is living in poverty (where poverty is defined as earning less than USD 2.50 per day). The analysis is based on data collected from 12,600 individuals across 7 different countries.

The goal of the analysis was to use the features provided to create a predictive supervised machine learning model to determine the probability that an individual is living in poverty. To evaluate the effectiveness of the model, an R^2 score (aka coefficient of determination) was calculated when the model was applied against a given set of test data containing 8,400 records. The score achieved for the created regression model was 0.4123, which was significantly higher than the 0.39 benchmark set for the challenge.

In order to create the machine learning model a number of steps were undertaken. The data was cleansed and then explored with the use of descriptive statistics and visualisations of the data. Through this process identification of the potential key features in the dataset were identified. This was then confirmed by working through different variations of parameters and features for different regression algorithms to produce predictive machine learning models that determine the probability that an individual is living in poverty. The model producing the highest R^2 score was selected and used to make predictions for the set of test data.

While many features in the dataset were found to be useful in predicting the poverty probability for an individual, the following features were shown to be the most important:

- Country – In the dataset this was represented by a unique identifier (actual country was masked). Some countries ('A' and 'D') ranked significantly higher than others in importance in predicting poverty probability. The data shows that individuals from these countries in particular ('A' and 'D') were more likely to be in poverty than individuals from other countries.
- Education Level – The highest level of education. In the dataset this was represented on a numeric scale from 0 (no education) to 3 (higher education). Individuals who have a higher Education Level have a lower probability of being in poverty.
- Is Urban – Indicator that determines whether the individual lives in an urban area or a rural area. In the dataset this was represented as a Boolean (true or false). Individuals living in rural areas are more likely to be living in poverty than those living in urban areas.
- Registered Bank Account – Indicator that determines whether the individual has a registered bank account in their name. Individuals with no registered bank account have a higher probability of being in poverty.
- Can use internet – The ability to use the internet on one's phone. In the dataset this was represented as a Boolean (true or false). Those who cannot use the internet on their phone have a higher probability of being in poverty.
- Phone Technology – The sophistication of one's phone type. In the dataset this was represented as a numeric scale from 0 (no phone) to 3 (smartphone). The more sophisticated the phone the lower the probability of being in poverty.
- Active Bank User – Whether the individual has used their bank account in the last 90 days. In the dataset this was represented as a Boolean (true or false). Active users have a lower probability of being in poverty.
- Phone Capability – An amalgamated score of different capabilities on one's phone including the ability to call, text, use the internet, make a transaction and perform advanced tasks. The higher an individual's phone capability, the lower the probability of being in poverty.

Data Analysis Process

Data Cleansing

Training data containing a total of 58 features for 12,600 individuals was provided, along with labels containing the poverty probability for each individual (determined by the PPI – see https://www.povertyindex.org/about-ppi). The nature of the PPI is that it consists of no more than 10 questions to determine one's poverty probability, which immediately indicates that we should expect a relatively small number of features to provide proportionately large importance to the final model.

There were several features with missing information and these were dealt with as follows:

- Education Level - 236 blank values. It is logical that the education level may significantly impact the poverty probability and therefore an effort was made to infer values for the missing records. This was done with the help of an additional feature that indicated Literacy (a Boolean feature). The median value for education level was calculated for individuals who were literate giving a value of 2. Similarly the median value for education level was calculated for individuals who were not literate giving a value of 1. These medians were then applied to the records where the education level was blank, depending upon whether the individual was literate.
- Share of Household Income Provided – 305 blank values. The nature of this categorical variable was not fully known, but it consisted of whole numbers ranging from 1 to 5. An arbitrary measure of the median for this variable across the whole dataset was calculated giving a value of 2. This was applied to the records where this variable was blank.
- Bank Loan Interest Rate – 12,311 blank values. Given the small number of records where this data is provided it was deemed that this would not provide any significant contribution to the predictive model and therefore this feature was removed.
- Mobile Money Interest Rate – 12,449 blank values. Given the small number of records where this data is provided it was deemed that this would not provide any significant contribution to the predictive model and therefore this feature was removed.
- MFI Loan Interest Rate – 12,399 blank values. Given the small number of records where this data is provided it was deemed that this would not provide any significant contribution to the predictive model and therefore this feature was removed.
- Other FSP Interest Rate – 12,361 blank values. Given the small number of records where this data is provided it was deemed that this would not provide any significant contribution to the predictive model and therefore this feature was removed.

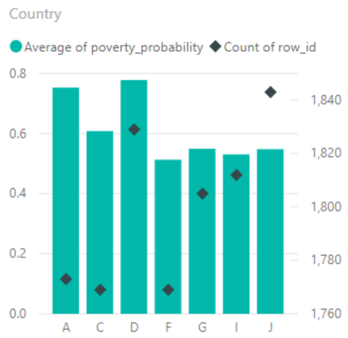
Data Exploration

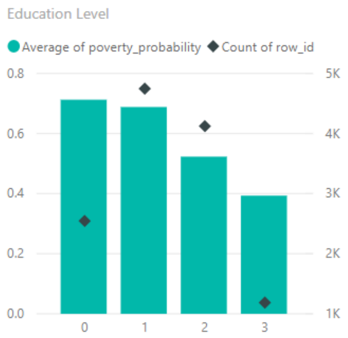
The features were broken up into categorical, Boolean and numeric features as follows:

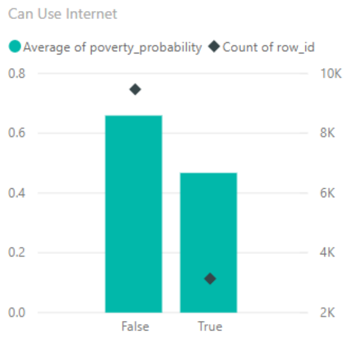
Categorical	Boolean		Numeric
country	is_urban	has_investment	education_level
employment_type_last_year	female	borrowed_for_emergency_last_year	num_times_borrowed_last_year
religion	married	borrowed_for_daily_expenses_last_year	borrowing_recency
employment_category_last_year	literacy	borrowed_for_home_or_biz_last_year	num_shocks_last_year
relationship_to_hh_head	can_add	can_call	avg_shock_strength_last_year
share_hh_income_provided	can_divide	can_text	phone_technology
	can_calc_percents	can_use_internet	phone_ownership
	can_calc_compounding	can_make_transaction	num_formal_institutions_last_year
	employed_last_year	advanced_phone_use	num_informal_institutions_last_year
	income_ag_livestock_last_year	reg_bank_acct	num_financial_activities_last_year
	income_friends_family_last_year	reg_mm_acct	age
	income_government_last_year	reg_formal_nbfi_account	
	income_own_business_last_year	financially_included	
	income_private_sector_last_year	active_bank_user	
	income_public_sector_last_year	active_mm_user	
	formal_savings	active_formal_nbfi_user	
	informal_savings	active_informal_nbfi_user	
	cash_property_savings	nonreg_active_mm_user	
	has_insurance		

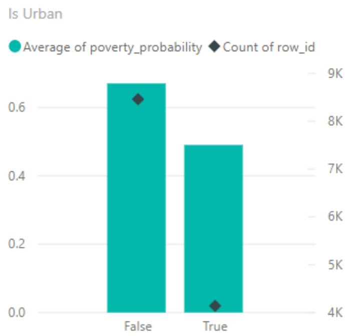
Note – education level, phone technology and phone ownership may be considered as categorical, but were treated as numeric for the purpose of the analysis. This is because the values for these variables are numeric and represent a scale – e.g. the higher the value for phone technology the greater the sophistication of the phone.

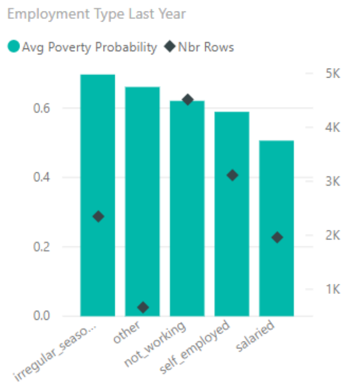
Bar charts were created for each feature to graph the feature values against the average poverty probability and to highlight where there is an imbalance of records for the given feature. The following bar charts show the results for the features that show the greatest variance in the raw data:

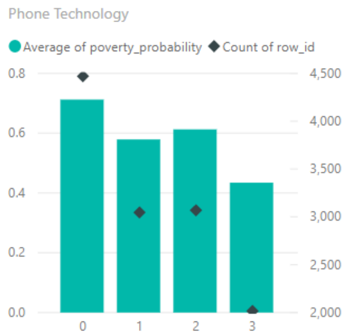


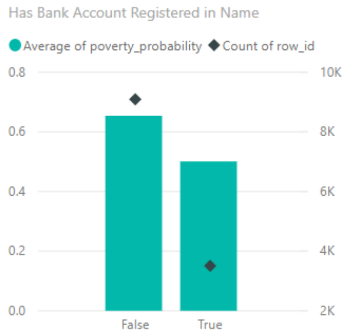


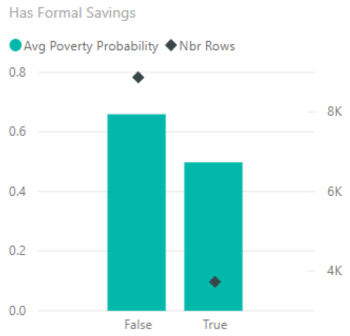


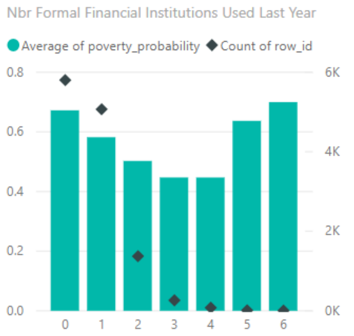


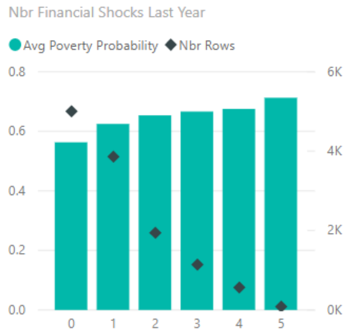


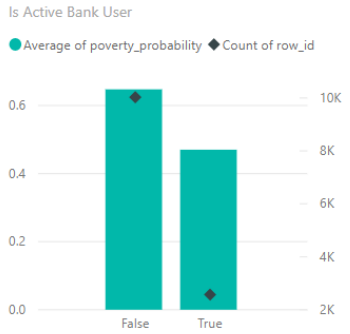


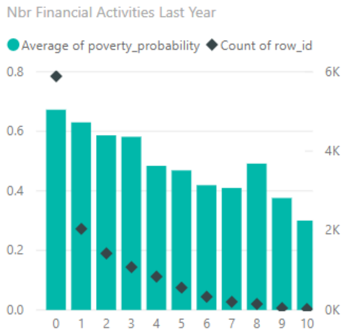








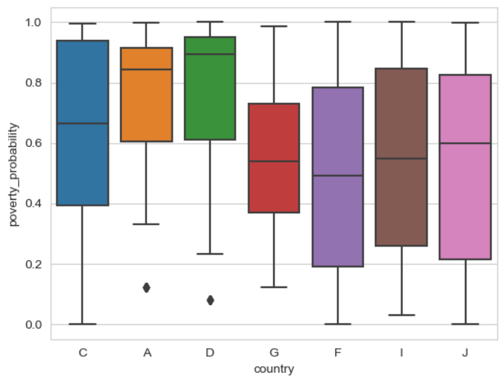




By looking at these charts we can surmise:

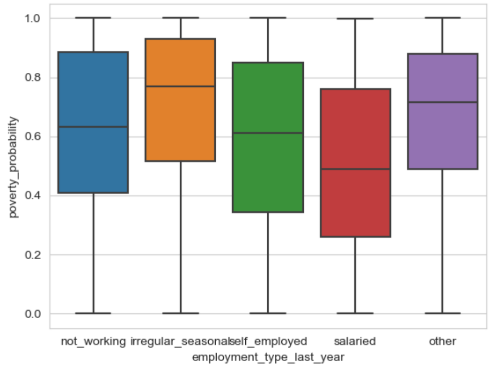
- The average poverty probability (APP) is higher for individuals in countries A and D than those in other countries
- The greater the level of education reached the lower the APP
- Individuals who can use the internet on their phone have a lower APP
- Individuals living in urban areas have a lower APP than those living in rural areas
- Salaried individuals have the lowest APP in comparison with other employment types
- The greater the sophistication of one's phone the lower the APP
- Individuals with a registered bank account have a lower APP than those without one
- Individuals with formal savings have a lower APP than those without savings
- The greater the number of financial institutions used (up to 4), the lower the APP. However those that use more than 4 tend to have a higher APP.
- The greater the number of financial shocks, the greater the APP
- Individuals that regularly use their bank accounts have a lower APP
- The greater the number of financial activities undertaken the lower the APP

Further analysis of categorical variables can be undertaken with box plots as below:



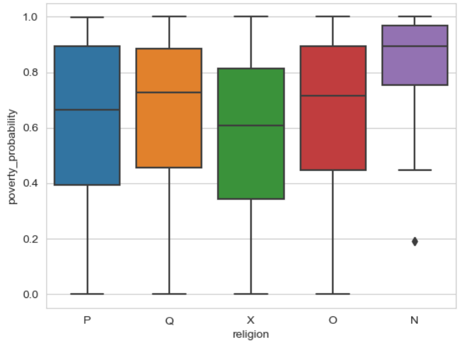
The distribution of poverty is similar between countries G, F, I and J.

We already know that countries A and D had the highest APP, however the boxplot for country C shows a wide range of poverty probability values with a large proportion experiencing relatively high poverty probability.



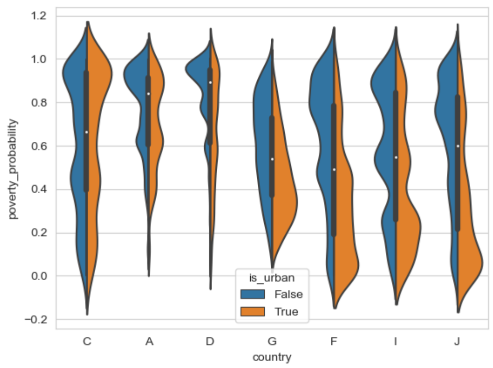
Whilst the Salaried category has the lowest poverty probability the distribution of employment types is almost identical for the following pairs:

- Not working and Self-employed
- Irregular Seasonal and Other



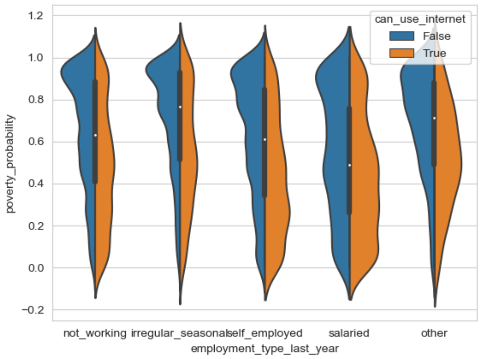
This box plot shows that individuals in religion N have a rather narrow distribution of poverty probability, which is significantly higher than for other religions.

Finally, split violin plots can show where Boolean variables have an impact on the poverty probability within the different category classes:



Here we see that the distribution of poverty probability is similar for individuals living in urban and rural areas for countries C and A.

However we see significant differences in the distributions for all other countries, with individuals living in urban areas having a much lower poverty probability.



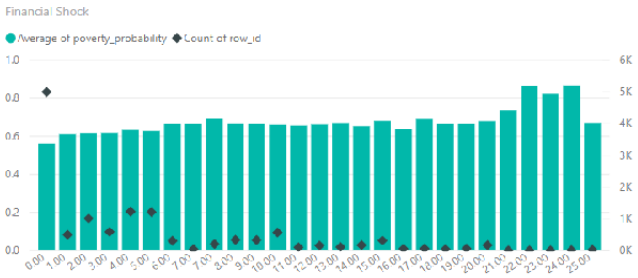
The ability to use the internet on the phone has an impact on poverty probability by employment category.

For all employment categories the poverty probability is lower for individuals who can use the internet on their phone.

Feature Engineering

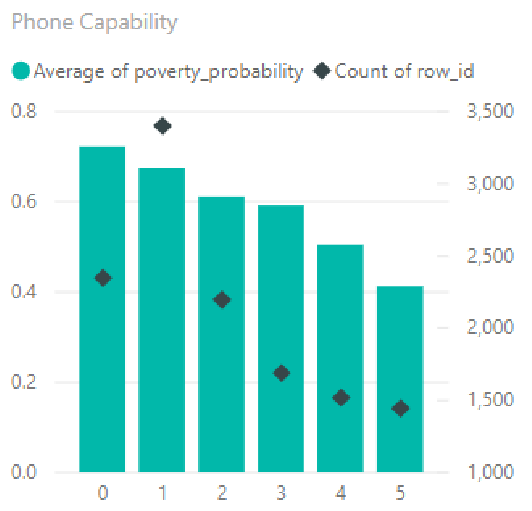
A review of the features made available in the training set resulted in the creation of 2 new features that used a combination of existing features. These were created in order to help any future predicted models:

- Financial Shock – This was created as the product of num_shocks_last_year and avg_shock_strength_last_year and represents the impact of the financial shock experienced. A bar chart showing the distribution of the newly created feature is below.



This bar chart shows that there was not a lot of variability in the new feature and therefore it may not add a lot of extra value to the predictive machine learning model

- Phone Capability – This was created as the sum of the Boolean features (where True = 1 and False = 0) related to an individual's phone use. These features were can_call, can_text, can_use_internet, can_make_transaction and advanced_phone_use. The higher the value the more sophisticated the individual's phone usage is. A bar chart showing the distribution of this newly created feature is below.



This bar chart shows significant variability in the new feature and therefore it may add extra value to the predictive machine learning model

Scaling and Splitting

Numeric features in the training data were scaled using the min / max method of scaling. The standard (Z-score) scaling method was not used since none of the numeric features exhibited a normal distribution.

The training data was then split as follows:

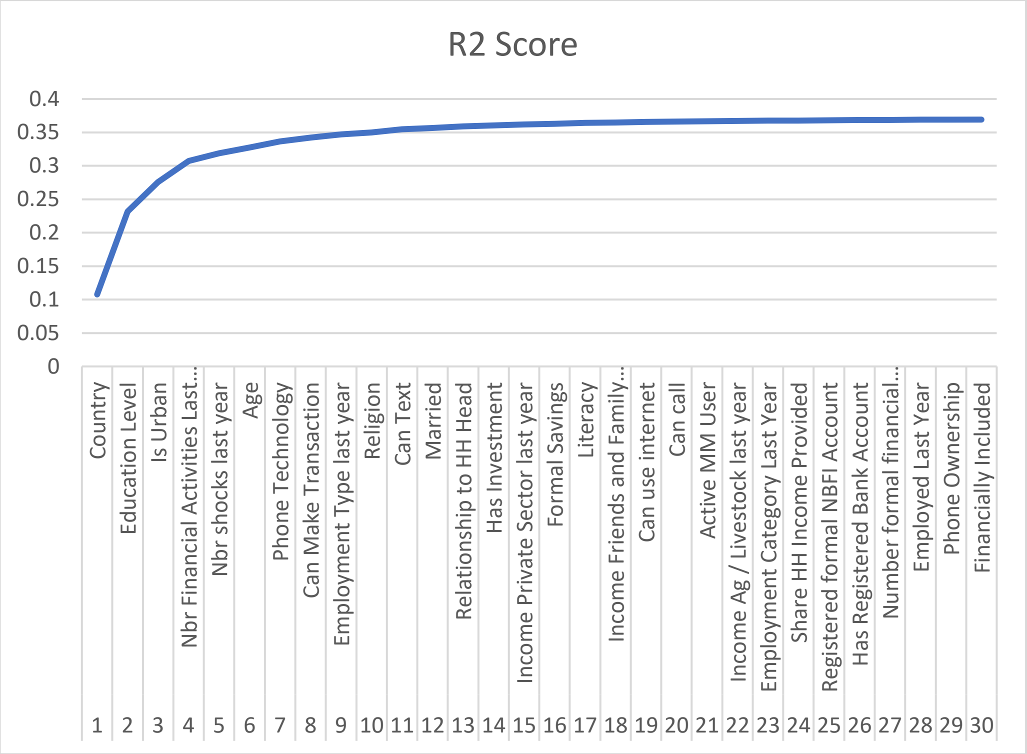
- 9,600 records to train the model
- 3,000 records to test the trained model's effectiveness

Linear Regression Model

Several different types of regression model algorithms were used to try and improve the R^2 score. The first model was Linear Regression.

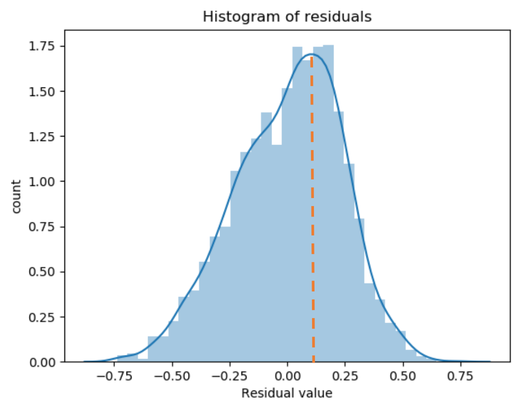
For the Linear Regression model a process of “greedy forward selection” was undertaken. This involved firstly running a regression model with a single feature and measuring its effectiveness via the R^2 score. Each feature was tested, and the feature that scored the highest R^2 score was added to the model. This process was then repeated using the feature from the first iteration, and testing the model after adding each feature. Again, the feature with the highest score was added to the model and the process repeated.

A chart showing the features added to the model and the resulting R^2 scores is shown below:



This indicated that there was very little gain in model performance beyond 20 or so features for the linear regression model. This is an important consideration to avoid overfitting.

The summary statistics and histogram of residuals are below:



```
Mean Square Error       = 0.05367

Root Mean Square Error = 0.23166

Mean Absolute Error     = 0.18872

Median Absolute Error   = 0.16719

R^2                     = 0.36169

Adjusted R^2            = 0.34999
```

This shows a relatively good histogram of residuals, but the R^2 score against the training data is well below the 0.39 benchmark, therefore other regression algorithms needed to be explored.

Alternative Regression Models

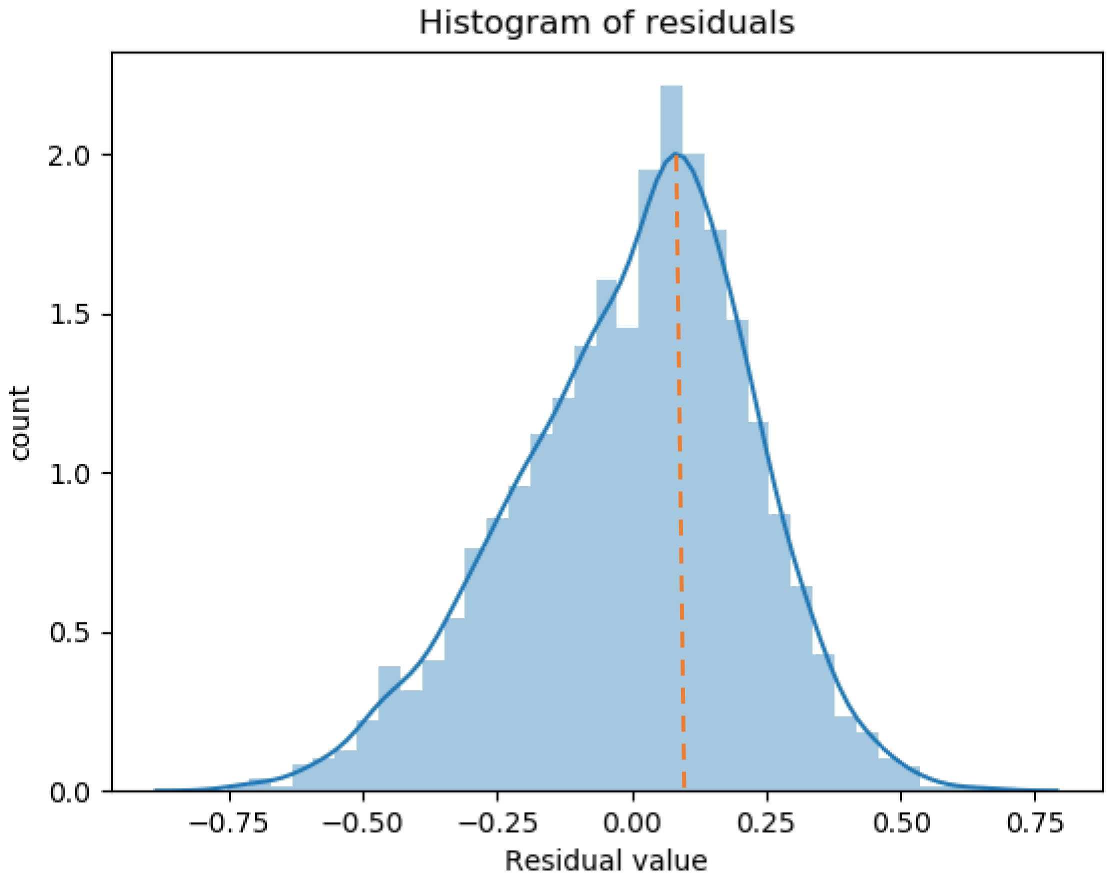
Several alternative models were attempted and scored against the test data (using cross validation techniques) as described below:

Model	**R^2 Score (cross validation)**
Ridge Regression (linear with L2 regularisation)	0.36219
Lasso Regression (linear with L1 regularisation)	0.36085
SVM	0.35145
Adaboost (boosted decision tree)	0.27492
Random Forest	0.41139

The final algorithm showed by far the most promise of all those that were attempted. This model was used as a predictor for the test data and a R^2 of 0.3904 was returned, which was above the benchmark.

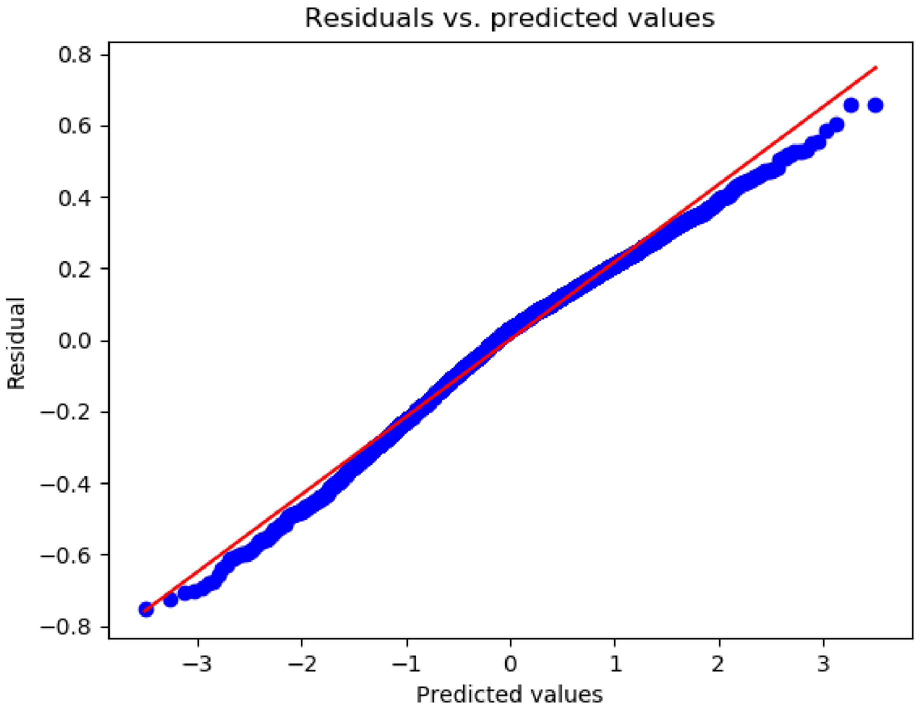
However it was felt that this could be improved upon and therefore the XGBoost library was used to implement the gradient boosting decision tree algorithm. Parameters for this algorithm were tuned and the final model produced the following statistics and plots:

1. Histogram



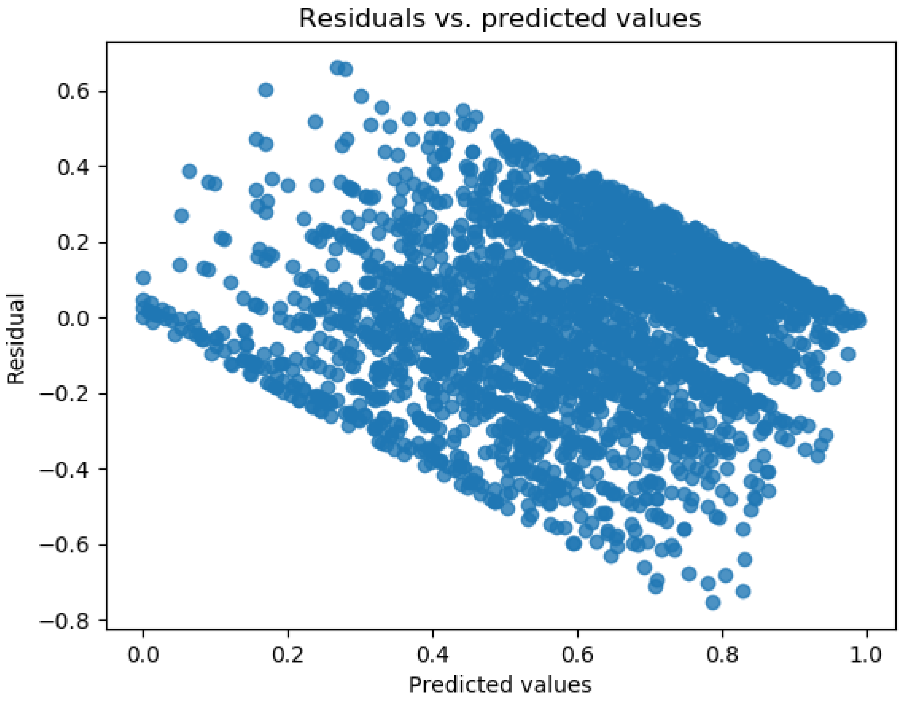
This shows a more “normal” distribution than for the linear regression model, although it is noted that the mean residual is slightly greater than 0.

2. Residual QQ Plot



This shows a good correlation between predicted values and residuals.

3. Residual Scatter Plot



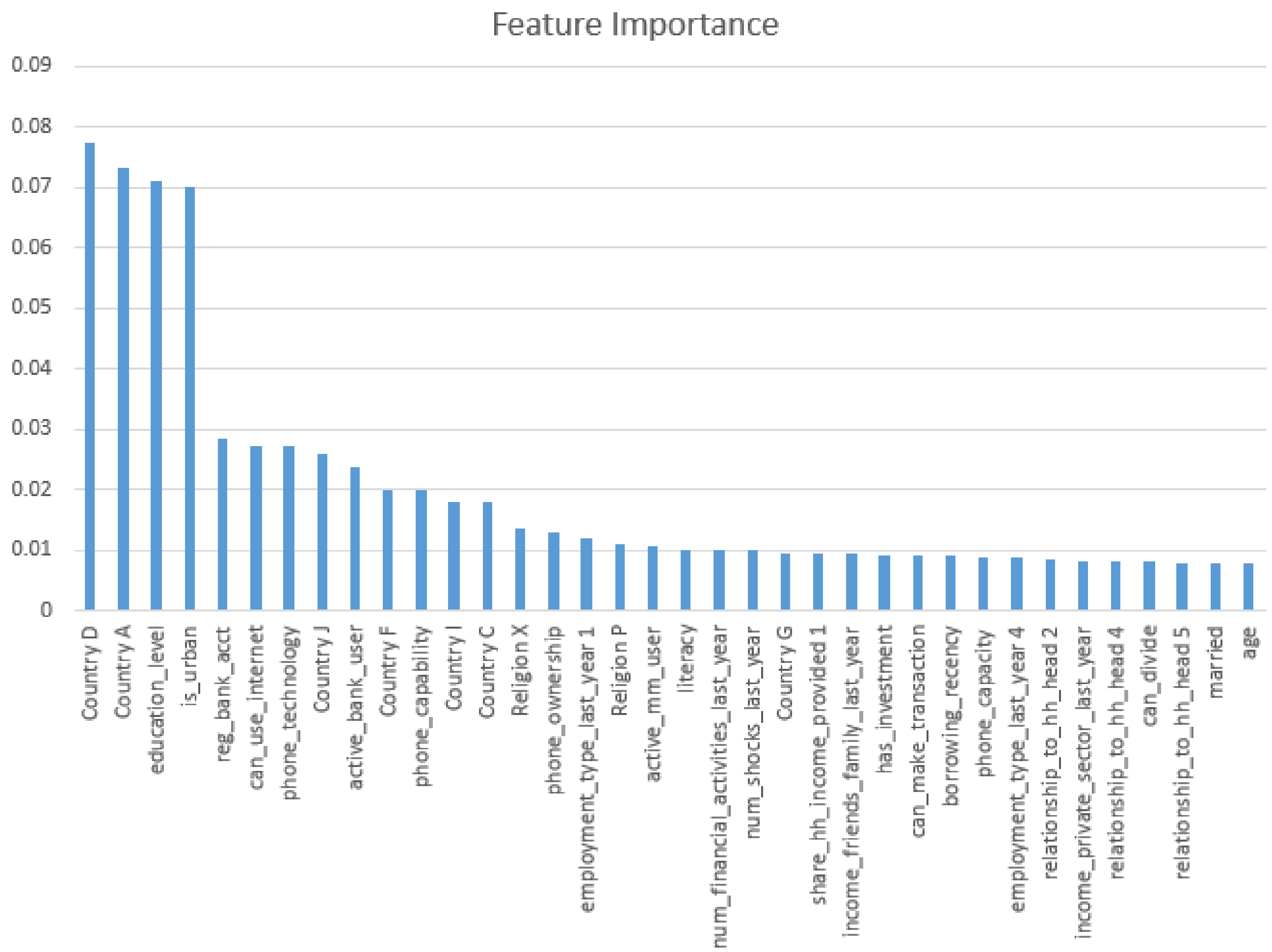
This shows that although the residuals are generally centred around zero, there is a definite shape to the plot indicating some level of bias in the final model.

4. Summary Statistics

```
Mean Square Error       = 0.04742

Root Mean Square Error = 0.21778

Mean Absolute Error     = 0.17475

Median Absolute Error  = 0.14801

R^2                     = 0.43592

Adjusted R^2            = 0.42518

Cross Validation R^2    = 0.42420
```

These statistics, particularly the R^2 score achieved after cross-validation show that the model is much improved on the linear regression model.

5. Feature importance



This shows consistency in the features that were previously identified as being significant.

It can be concluded from these plots and statistics that the model is a good fit for the data.

This model was used as a predictor for the test data and a R^2 score of 0.4110 was returned, which was well above the benchmark of 0.39.

Conclusion

This analysis has shown that the poverty probability for an individual can be predicted from a collection of the individual’s demographic, educational, economic, financial and employment characteristics. In particular, the country, education level and urban/rural indicator all have a significant impact on an individual’s poverty probability. Secondary features such as registered bank account indicator, phone internet usage indicator, level of phone technology, active bank user indicator and overall level of phone capability are also important indicators of poverty probability.

The collection of features was used with a gradient boosting decision tree algorithm to create a predictive machine learning model that whilst by no means perfect still provides significant ability to predict an individual’s poverty probability.