Analysis of Poverty Probability Index

Swagata Sarkar, July 2019

Executive Summary

This document presents a summary analysis of the observations and most relevant features to predict the Poverty Probability Index (PPI) which estimates an individual's poverty status. All analysis was based on a dataset containing 12600 observation of poverty probability index and given 58 variable features.

An exploratory data analysis was done to understanding statistical relationships between PPI and the variable features to identify any potential relationships between them. Then, those relationships are plotted to visually represent the data and uncover more stories hidden in the data set. After having a good understanding of the dataset, a predict model is created to estimates poverty probability index from its features was created.

The general relationship between the dependent variable (PPI) and all of the top ranked features were assessed to identify any strong relationship between PPI and the features. Significant information was found in this analysis were:

- ✓ Except country and urban population, other demographic features have less capability to predict probability index. Poverty possibility is considerable higher in rural areas than same in urban areas.
- ✓ Features related to education are good in estimating PPI and it is apparently seen that observation with higher/positive educational features have lower poverty.
- ✓ Except employment type (salaried or others) other features related to Employment does not have strong relationship with poverty.
- ✓ Some of the economy features show potential relationship with poverty but so many categories have a very limited number of cases. So, some feature engineering was applied to make the features more useful for predicting price.
- ✓ According to some features related phone, it is apparently seen that poor people have less access on phone technology.
- ✓ As can be seen in some features related to Financial Inclusion, poor people have less tendency to use financial and banking facilities.

Initial Exploratory Data Analysis

Before creating analytical models, an understanding of the properties and relationships in a dataset is very vital. Initial exploration of the data began with the application of simple and common data analysis principles.

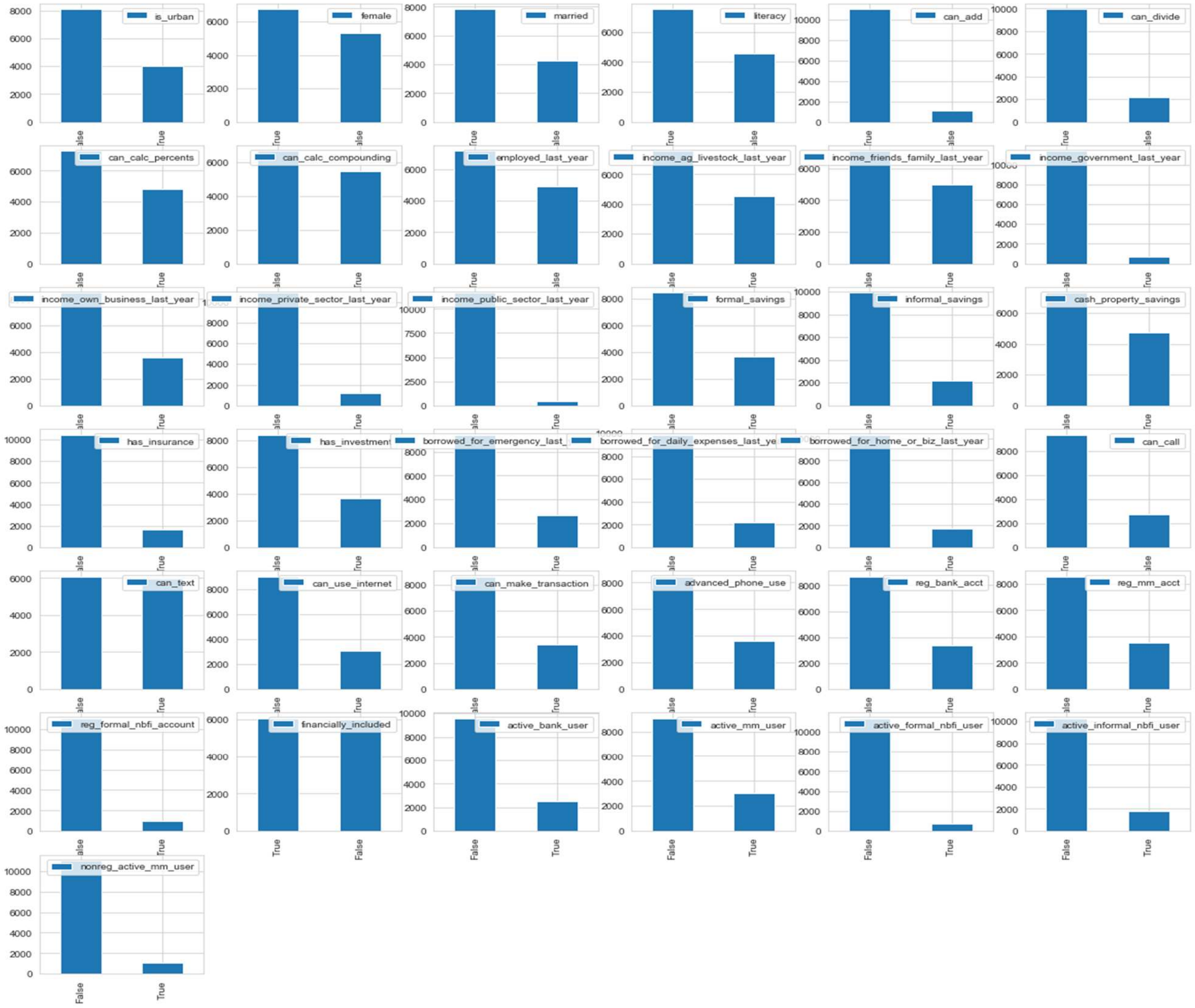
General Overview from Initial Analysis

A quick survey of the provided data revealed a starting point of 59 features.

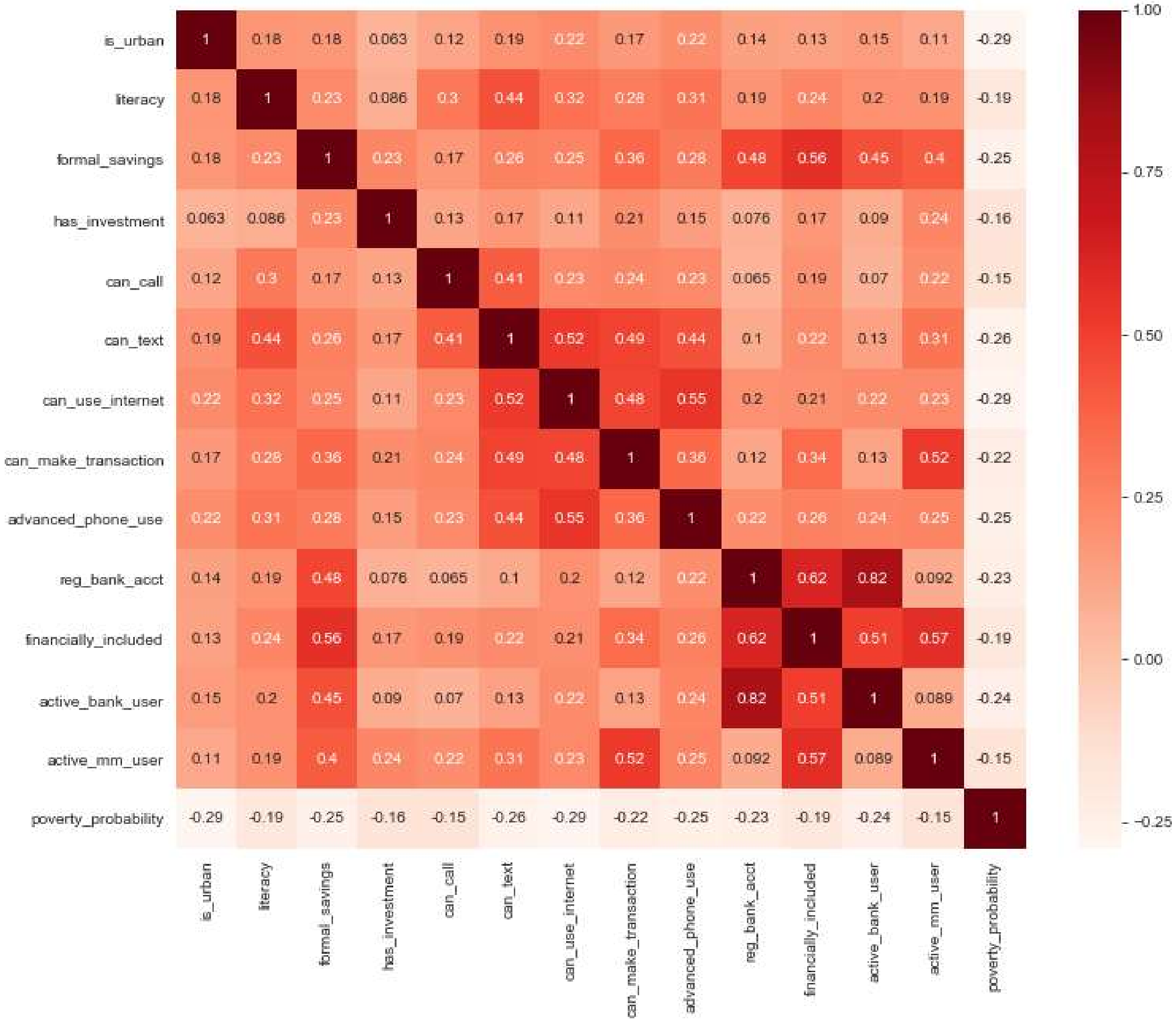
The bank_interest_rate , mm_interest_rate, mfi_interest_rate , other_fsp_interest_rate columns have a significant number of missing values and were removed from the dataset. Education_level and share_hh_income_provided columns have some missing values and so, the corresponding rows with any missing values were removed from the train dataset. Same row from level dataset(poverty_probability) set were also dropped to keep the consistency. The roq_id colum from the feature list was also deleted .Finally, the modified dataset contains 54 features and 12068 observations.

In modified dataset, there were 42 categorical variable and 12 numeric variables. Among 42 categorical variables, 37 are binary variables.

The Fig:1 depicts that there are very few variations in some features with binary variables. They would be less effective for prediction.

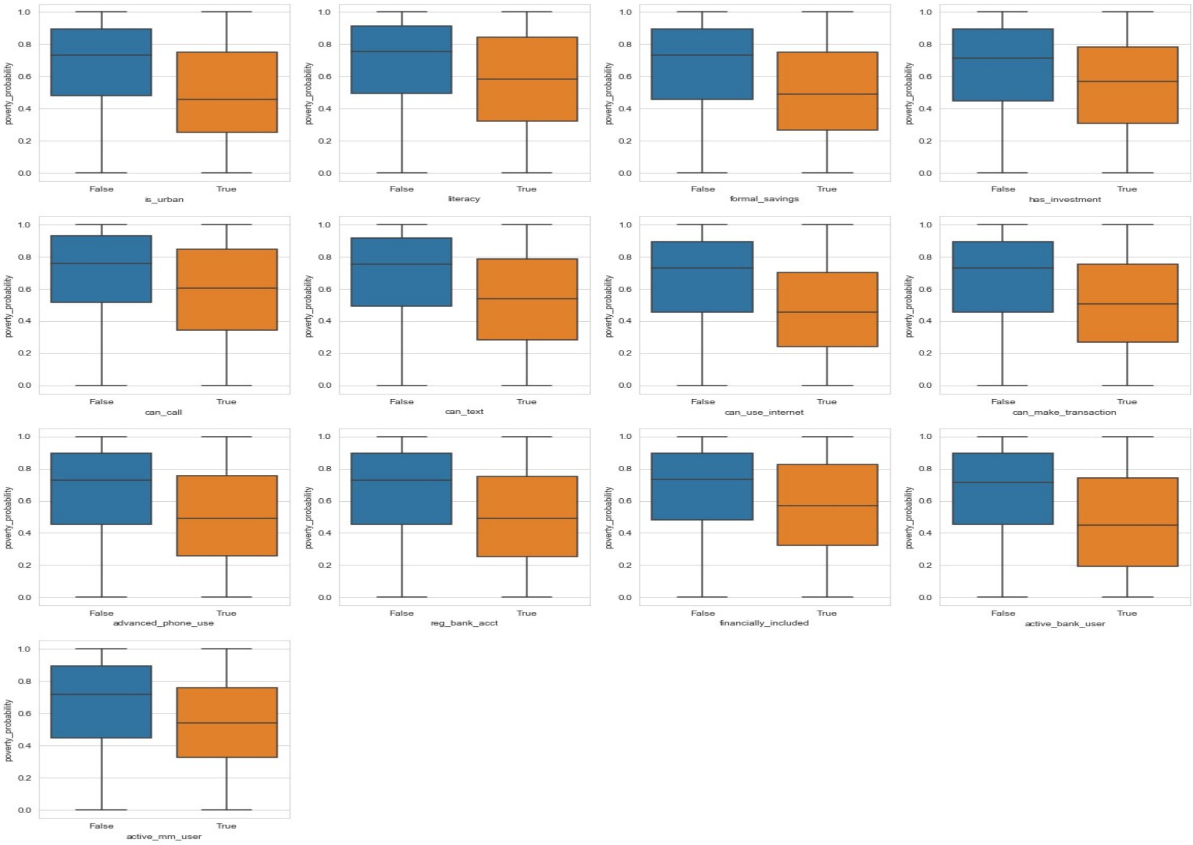


A correlation matrix was developed to find out the relation between the 37 binary variables with poverty probability index. It was found that only few features have considerable R2 value with poverty probability index. These following fourteen features which have correlation of above 0.15 (taking absolute value) with the output variable were selected as effective features. If the threshold limit of R2 reduced to 0.1, then total effective feature number would be 21. But, initially these 14 features were kept in train dataset.



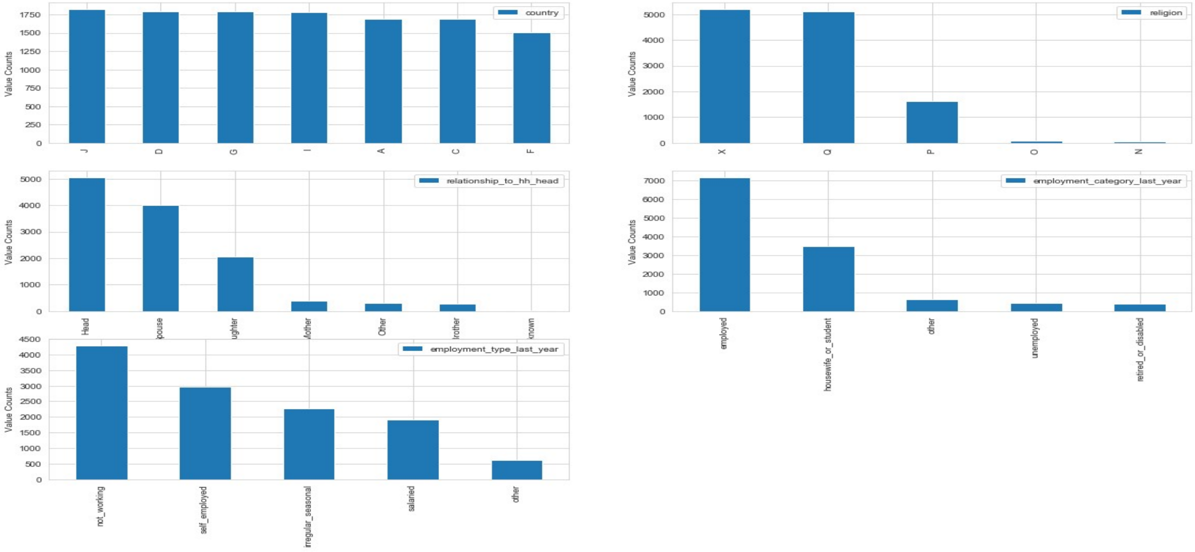
	is_urban	literacy	formal_savings	has_investment	can_call	can_text	can_use_internet	can_make_transaction	advanced_phone_use	reg_bank_acct	financially_included	active_bank_user	active_mm_user	poverty_probability
is_urban	1	0.18	0.18	0.063	0.12	0.19	0.22	0.17	0.22	0.14	0.13	0.15	0.11	-0.29
literacy	0.18	1	0.23	0.086	0.3	0.44	0.32	0.28	0.31	0.19	0.24	0.2	0.19	-0.19
formal_savings	0.18	0.23	1	0.23	0.17	0.26	0.25	0.36	0.28	0.48	0.56	0.45	0.4	-0.25
has_investment	0.063	0.086	0.23	1	0.13	0.17	0.11	0.21	0.15	0.076	0.17	0.09	0.24	-0.16
can_call	0.12	0.3	0.17	0.13	1	0.41	0.23	0.24	0.23	0.065	0.19	0.07	0.22	-0.15
can_text	0.19	0.44	0.26	0.17	0.41	1	0.52	0.49	0.44	0.1	0.22	0.13	0.31	-0.26
can_use_internet	0.22	0.32	0.25	0.11	0.23	0.52	1	0.48	0.55	0.2	0.21	0.22	0.23	-0.29
can_make_transaction	0.17	0.28	0.36	0.21	0.24	0.49	0.48	1	0.36	0.12	0.34	0.13	0.52	-0.22
advanced_phone_use	0.22	0.31	0.28	0.15	0.23	0.44	0.55	0.36	1	0.22	0.26	0.24	0.25	-0.25
reg_bank_acct	0.14	0.19	0.48	0.076	0.065	0.1	0.2	0.12	0.22	1	0.62	0.82	0.092	-0.23
financially_included	0.13	0.24	0.56	0.17	0.19	0.22	0.21	0.34	0.26	0.62	1	0.51	0.57	-0.19
active_bank_user	0.15	0.2	0.45	0.09	0.07	0.13	0.22	0.13	0.24	0.82	0.51	1	0.089	-0.24
active_mm_user	0.11	0.19	0.4	0.24	0.22	0.31	0.23	0.52	0.25	0.092	0.57	0.089	1	-0.15
poverty_probability	-0.29	-0.19	-0.25	-0.16	-0.15	-0.26	-0.29	-0.22	-0.25	-0.23	-0.19	-0.24	-0.15	1

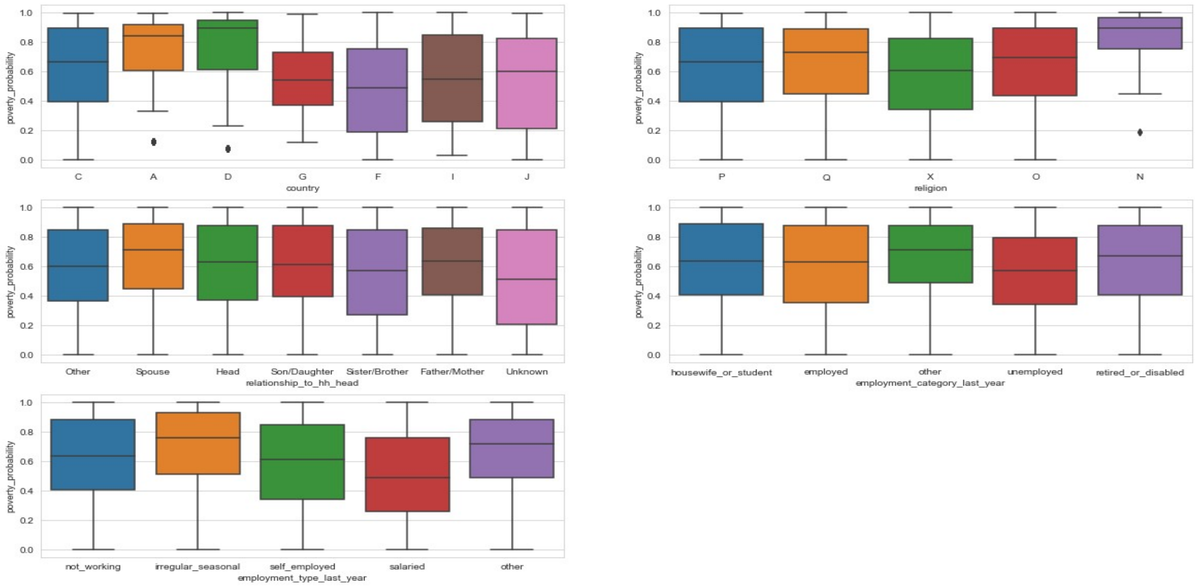
The following box-plots show selected categorical columns that seem to exhibit a relationship with poverty probability index. Among them ‘is_urban’, ’formal_savings’, ‘can_text’, ’can_use_internet’, ‘advanced_phone_use’, ‘active_bank_user’, ‘reg_bank_acct’ features have good separation between their categories.



Categorical Relationships:

Brar charts were created for categorical values to check the frequency for the all unique values and box charts were created to find any apparent relationship between categorical feature values and poverty probability index.





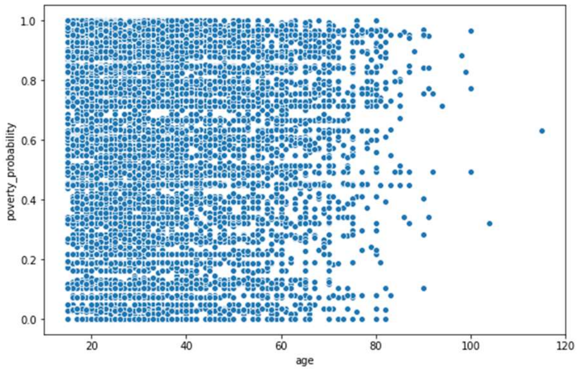
The above two charts show some important behaviours for different categorical features. For example:

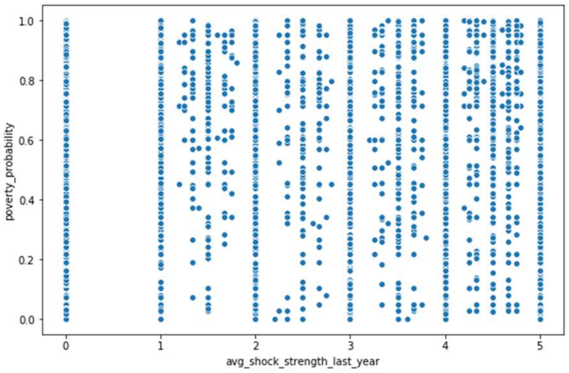
- ✓ In country column, each category has enough samples and it is likely that these countries have different poverty probability than others.
- ✓ In religion columns, O and N have a limited number of cases with higher poverty probability.

- ✓ In relationship_to_hh_head columns, four of these categories have a limited number of samples. These categories can be aggregated to increase the number of cases.
- ✓ In employment_category_last_year,three of these categories have very few cases. It is likely that all of these categories will not have statistically significant difference in poverty probability
- ✓ In employment_type_last_year column, poverty probability index of salaried category can be easily separable with the rest of the categories.

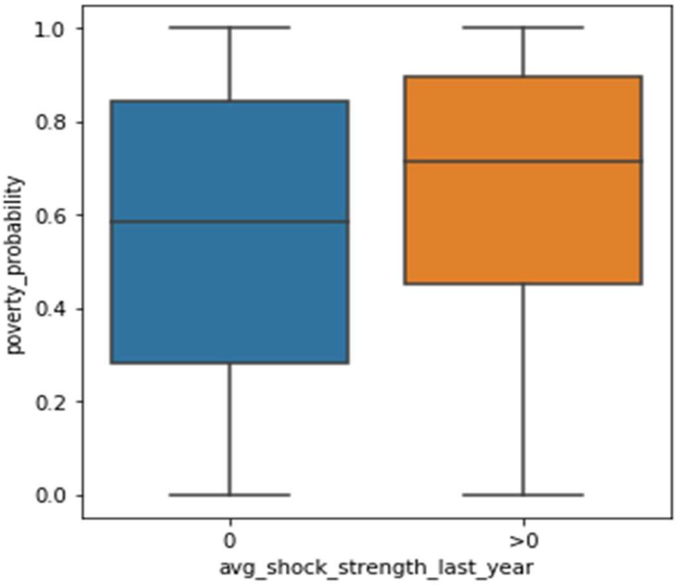
Numeric Relationships

There are twelve numeric variables in the trained dataset and among them ten have a limited and usually fixed number of values; which can be treated as categorical features. The rest two age and 'avg_shock_strength_last_year' have numeric variable, and scatter plots were plotted and found that there was not potential relationship with these numerical values with poverty probability.

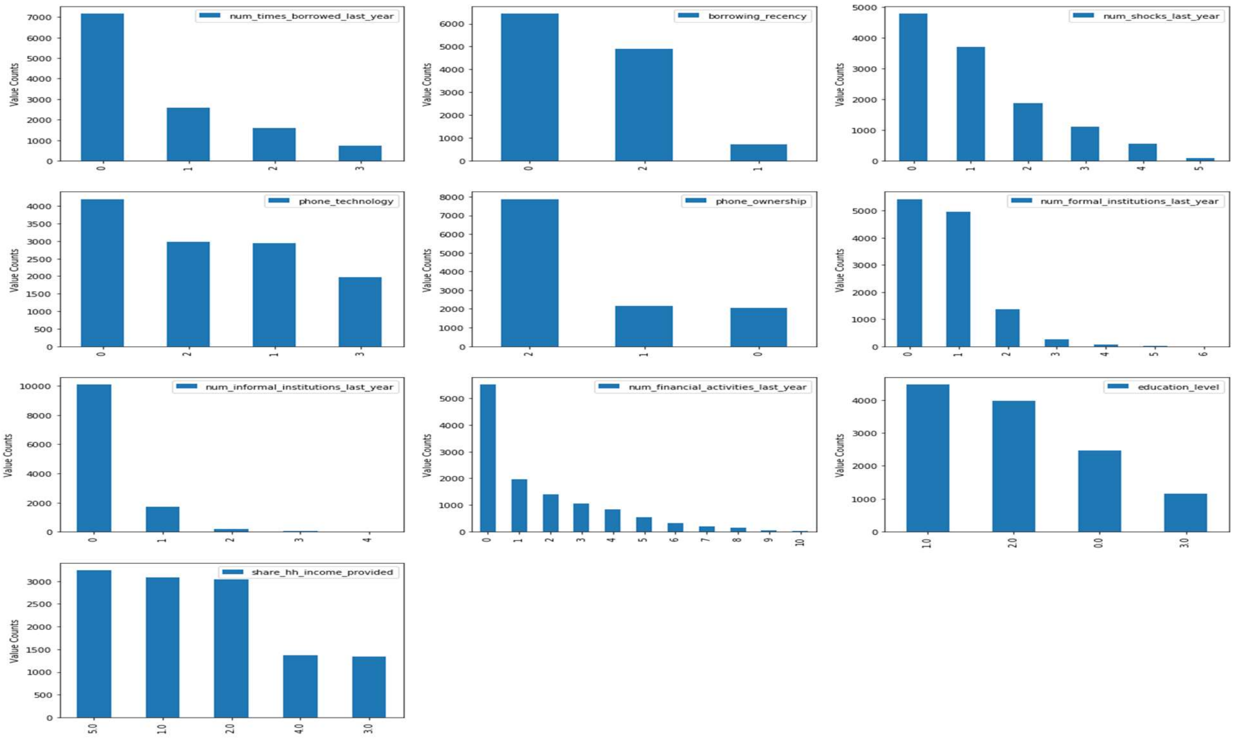


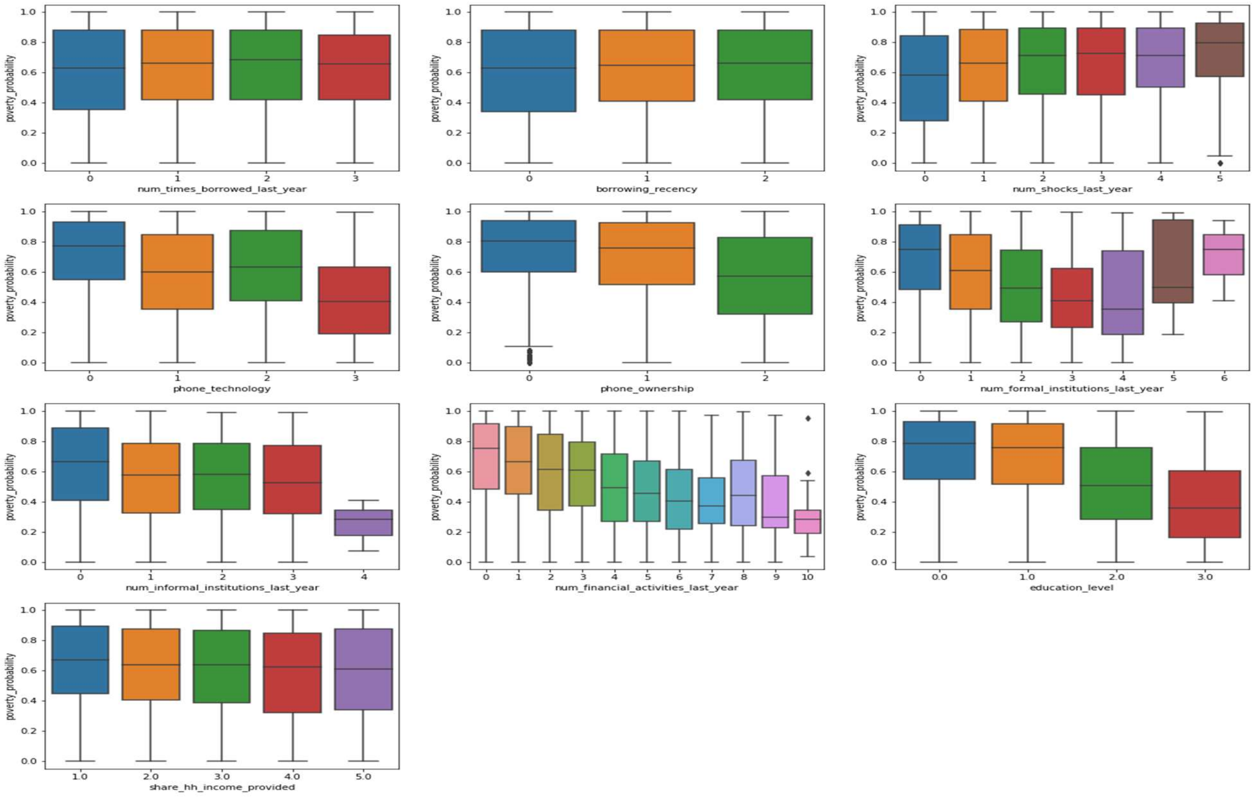


But after classifying the 'avg_shock_strength_last_year' feature as greater and less than zero, the poverty probability index of these categories became distinctive.



For the rest of ten numeric variable ,as they have very limited number of categories, were treated as categorical features. The bar charts were created to check the frequency for the all unique values and box charts were created to find any apparent relationship between categorical feature values and poverty probability index.





The above two category charts show how these features relate with the output poverty probability index. According to the graphs,

num_times_borrowed_last_year' , 'borrowing_recency' and 'share_hh_income_provided' features do not have statistically significant difference in predicting poverty probability index.

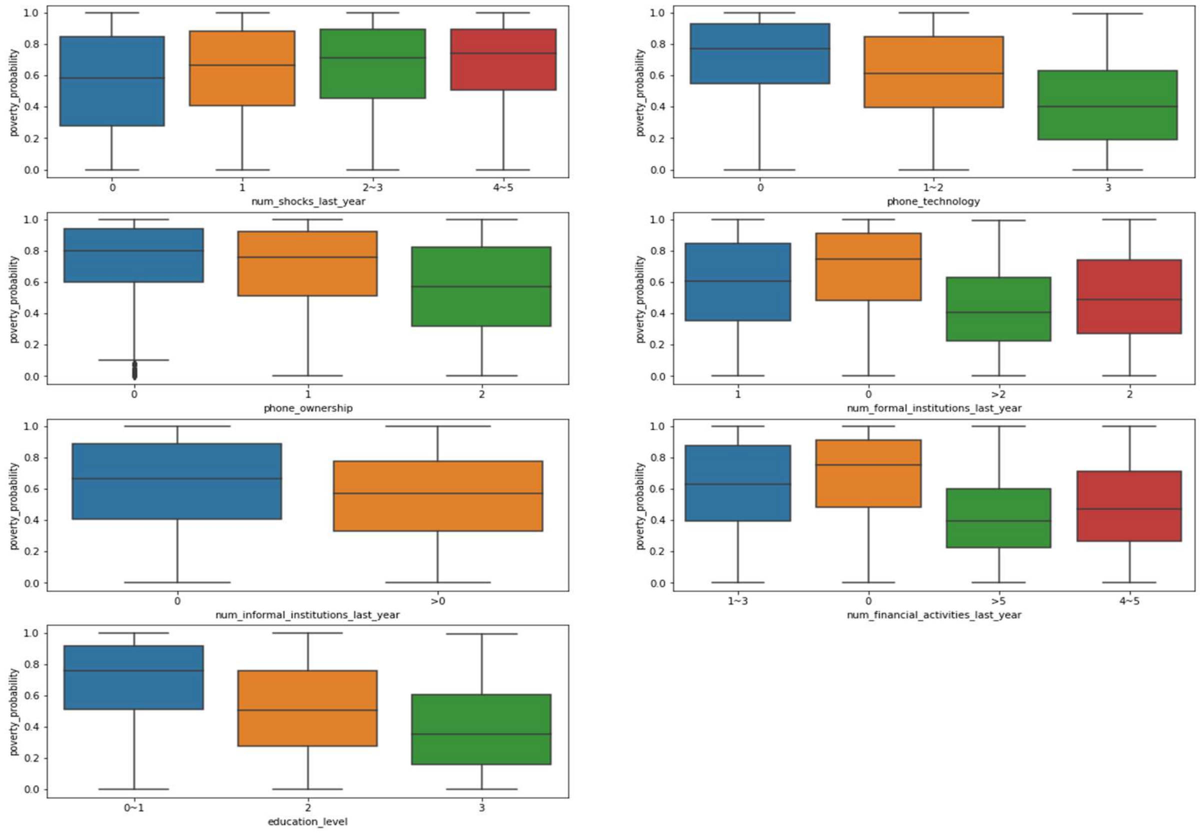
PPI is considerably high with higher number of 'num_shocks_last_year' but as some of the features have very limited number of cases, these can be aggregated.

In ‘ Phone_technology’ feature, as can be seen , category 1 and 2 can be merged as the PPI range of these categories are almost similar.

All of the categories in 'num_formal_institutions_last_year' , 'num_informal_institutions_last_year', 'num_financial_activities_last_year' and 'share_hh_income_provided' do not have statisticclay significance capabilities in prediction PPI and some of their frequency are quite low. But after the combining the categories to increase their cases, these new categories will be useful in predicting PPI.

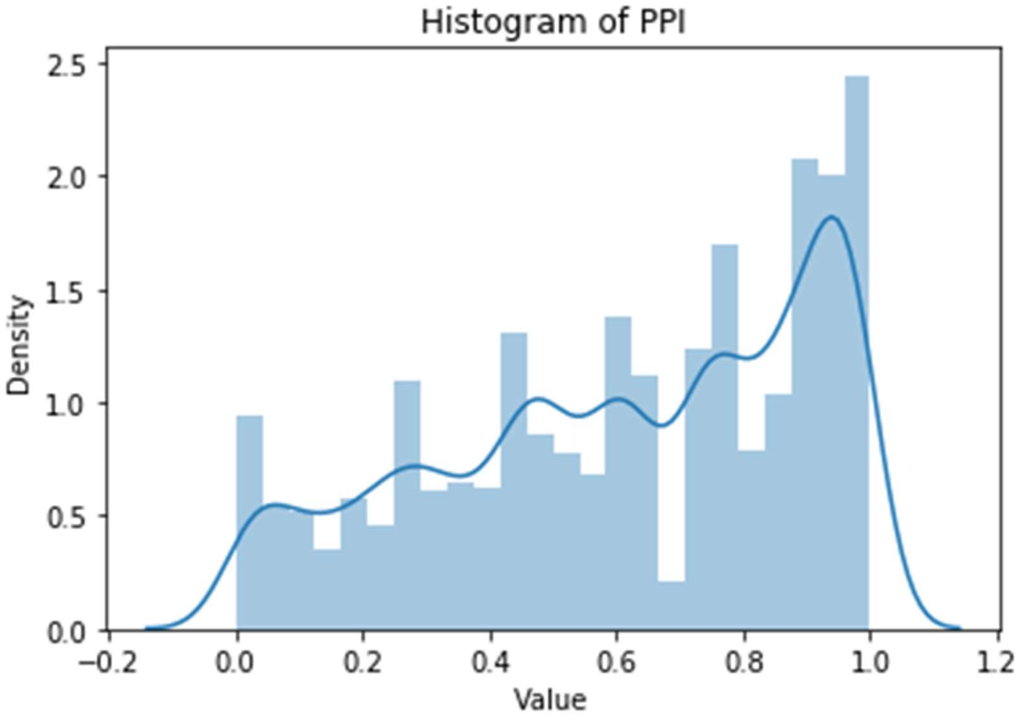
In ‘education_level’ feature, it is clear that higher education lowers the poverty. But the poverty range with education level 0 and 1 is almost identical, they can be merged to prepare a separate category to make it more separable from other two categories.

After adjusting the categories , the following box charts were plotted to asses relationship between the dependent variable and all of these selected features.



Dependent Variable Analysis

The poverty probability index dataset is negatively skewed (skw -0.45/kur -0.96) with a range from 0 to 1.0. Though the mean and median values are not so significantly different , the comparatively large standard deviation indicates that there is considerable variance in the poverty probability index. A distribution plot of the poverty price index shows that the PPI values are left-skewed – in other words, most observation are at the upper end of the range.



Model Selection

The goal of model selection is to find the best performing model for the problem at hand. After the data preparation and selecting the beat features, a cross validation model was created to select the best model to predict the poverty probability index. Here k-fold cross validation module with k value 10 was used to trained and tested the dataset ten times with both Linear regression and Boosted Decision Tree Diagram model. The result by Boosted Decision Tree Diagram module was more impressive than the Linear regression module. Here, below the 10 folds results matrix of the model run with Boosted Decision Tree Diagram module

Fold Number	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
0	0.18	0.22	0.71	0.58	0.42
1	0.18	0.22	0.72	0.60	0.40

2	0.18	0.22	0.71	0.58	0.42
3	0.17	0.22	0.70	0.56	0.44
4	0.17	0.22	0.70	0.57	0.43
5	0.18	0.22	0.72	0.57	0.43
6	0.18	0.22	0.72	0.59	0.41
7	0.17	0.22	0.68	0.53	0.47
8	0.18	0.23	0.72	0.60	0.40
9	0.18	0.22	0.71	0.59	0.41
Mean	0.18	0.22	0.71	0.58	0.42
Standard Deviation	0.00	0.00	0.01	0.02	0.02

The results show that model developed by this dataset and the Boosted Decision Tree Diagram module will have Coefficient of Determination range from 0.40 to 0.47 which is good enough to get our target score for this assessment. Here, the Mean Coefficient of Determination is 0.42 and its deviation is negligible. So, it is obvious that a model with this module will give an expected result.

Regression

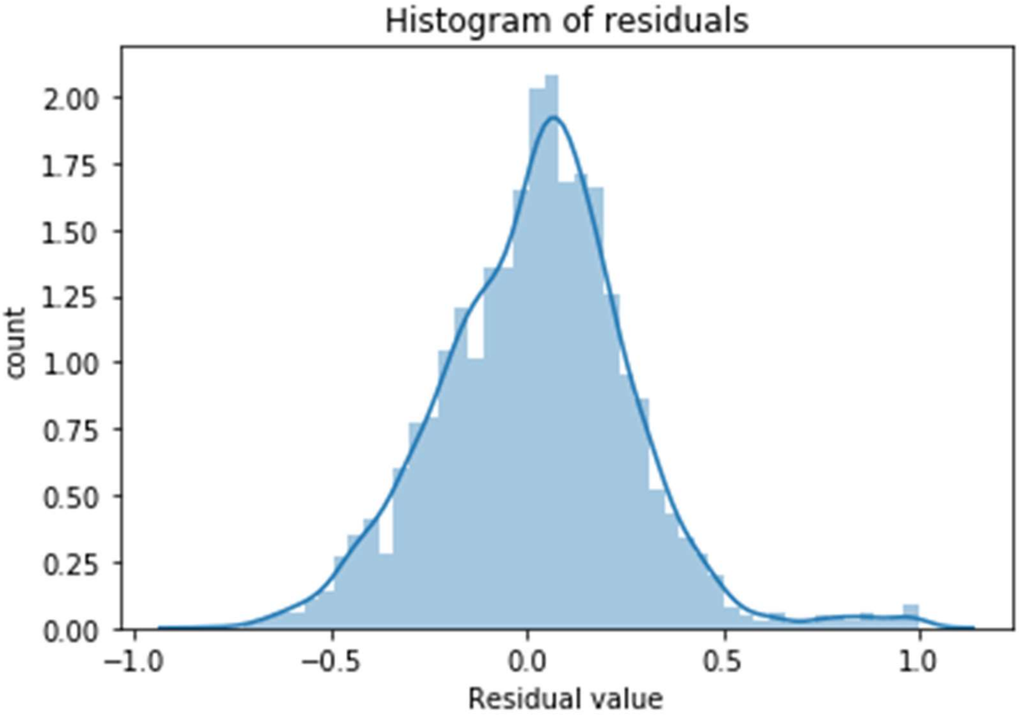
After selecting the model, the Boosted Decision Tree Diagram module was trained with 70% of the data, and tested with the remaining 30%.Here the same random seed was used to split the data as it was used in cross validation model to get a steady result.

The evaluation matrix for the model is given below

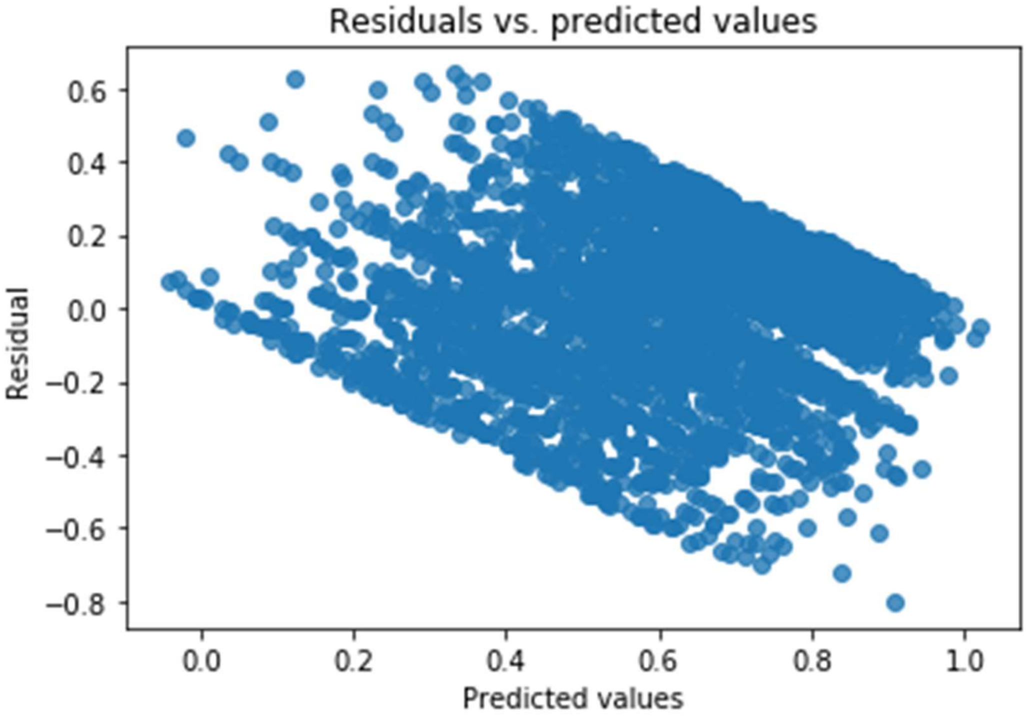
Mean Absolute Error	0.177227
Root Mean Squared Error	0.221557
Relative Absolute Error	0.707067
Relative Squared Error	0.576305
Coefficient of Determination	0.423695

At first glance, these metrics look not so promising but as per the target set, the model goes well.

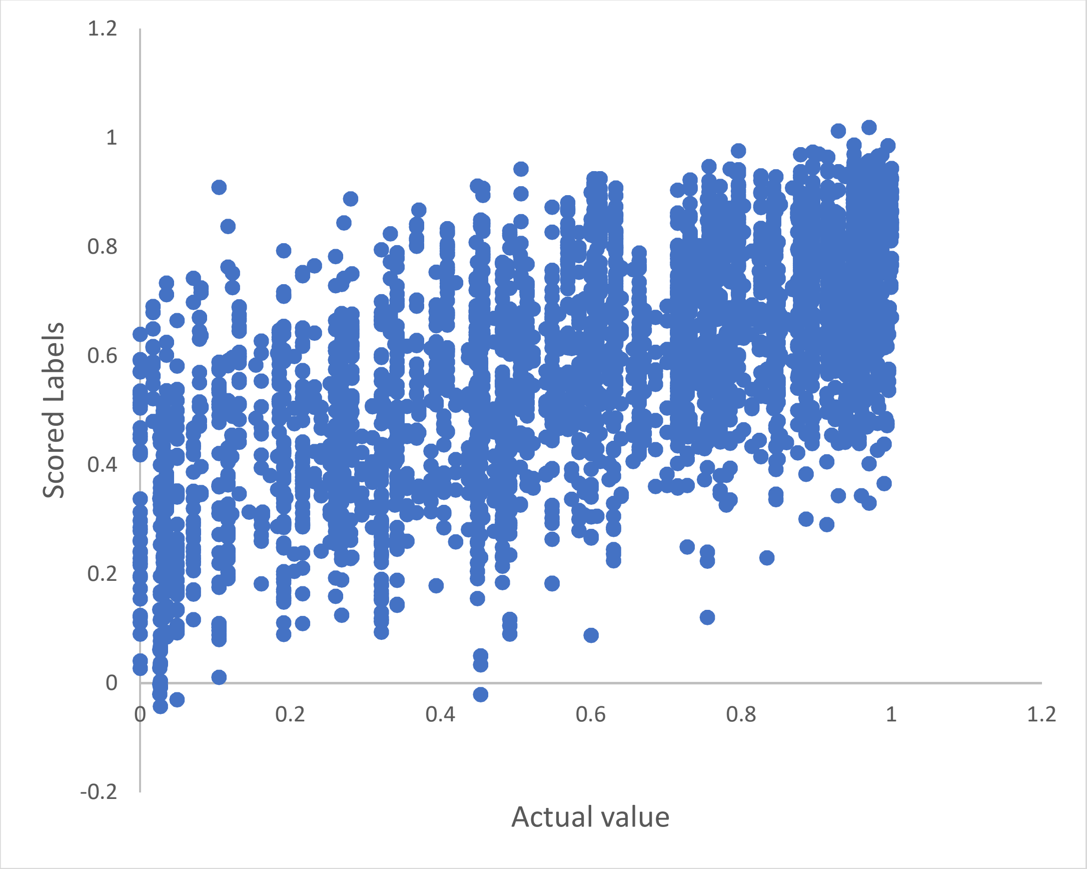
The following histogram of the residuals depicts that most of the errors are in a quite big range but the skewness of distribution is less and the centre of the distribution is very near to zero.



The below plot with residuals vs. predicted values describes that the model generates more error during predicting higher poverty probability index. In mid-range, the performance of the model is quite better. The regression model seems to do a good job of predicting mid-range poverty than lower and higher end poverty.



A scatter plot showing the predicted poverty probability index and the actual poverty probability is shown below



This plot shows a clear linear relationship between predicted and actual values in the test dataset but with a wide range of error. And the error is considerable high in both lower and higher end.

Conclusion

The analysis has shown that it is very hard to predict the poverty probability index confidently with these large number of given features. But it is worth noticing that some features have strong relationship with poverty. Overall, it is apparently seen that poor people have less access of better education, finance and banking facilities and current technologies. So, some of these features are good to predict PPI. It is also observed that poverty level is noticeably higher in some specific countries, and the rural people suffer it most.