ANALYSIS OF POVERTY DATA

Microsoft Professional Capstone: Data Science
Reinhard Streicher, July 2019

EXECUTIVE SUMMARY

The document presents an analysis of data on individuals from seven countries that details the likelihood of these individuals living in poverty. The data consists of 12600 observations of individuals in a given country containing characteristics of demographics, education, employment, economical, phone and financial inclusion data. The probability of being in poverty is calculated using the Poverty Probability Index (PPI).

The challenge set forth is to predict the poverty probability of individuals based on the training data described above. This is a regression problem to predict poverty probability a continuous float between 0 and 1. The metric that the model is measured on is r-squared:

$$R^2 = 1 - \frac{SSres}{SStot}$$

Where SSres is the sum of squares of residuals and SStot is the total sum of squares.

After exploring the data by calculating summary statistics, descriptive statistics, and by creating visualizations of the data, several relationships between poverty probability and the characteristics that influence it is identified. The author presents the following conclusions:

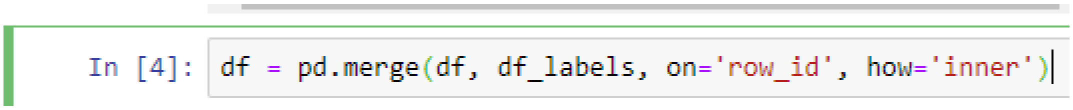
While many features influence the likelihood of individuals living below the poverty line, significant features in the analysis are listed below:

- **Country** – The analysis indicates that depending on the country individuals are living in, the likelihood of being in poverty increases or decreases. It is the strongest of the features in predicting the likelihood of living in poverty. The names of the countries are obscured in the data set.
- **Education level** – Education has a positive impact on reducing poverty in the individuals in the data set. As the education of individuals increases from no education to higher education, there is a reduction in the likelihood of living in poverty.
- **Being urban or not** – Individuals living in rural areas are more likely than individuals living in urban areas to live in poverty.
- **Religion** – Religion has an impact on poverty probability and from the data shows that depending on the type of religion poverty probability is either increased or decreased. The religion name is obscured in the data; therefore, it is not possible to state which religion by name is more likely to reduce poverty compared to others.
- **Married** – Separating the data into male and female categories for married, shows that females have a higher likelihood of living in poverty compared to males if unmarried.

Finally, a predictive regression model is created based on data exploration and feature selection. The model uses the boosted decision tree algorithm and results in an r-squared score of 0.42.

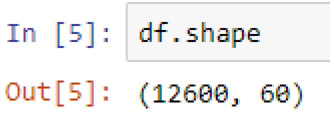
Initial Data Exploration and Wrangling

The data are in two data sets the training labels and training values. Firstly, the author explored the data sets to identify a common field on which to combine the data sets. The row_id field is identified and used to combine the data sets using the python merge function.

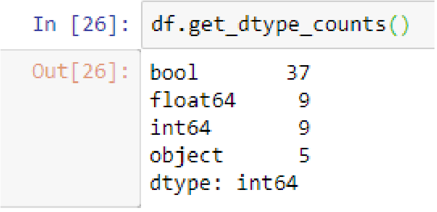


```
In [4]: df = pd.merge(df, df_labels, on='row_id', how='inner')
```

Determining the shape and type of data:



```
In [5]: df.shape
Out[5]: (12600, 60)
```



```
In [26]: df.get_dtype_counts()
Out[26]: bool       37
         float64     9
         int64       9
         object      5
         dtype: int64
```

The combined data set consists of 12600 observations and 18 numerical and 42 categorical features, including the label poverty probability, before any data wrangling.

The official data dictionary list the data organized into the following feature categories (the significant features of each category is listed below it as identified through data exploration and the permutation feature importance module in azure ml studio):

- **DEMOGRAPHICS**
  - country - Unique identifier for each country
  - is_urban - Urban vs rural area of residence
  - age - Age
  - female - Sex (True=female, False=male)
  - married - Marital status
  - religion - Unique identifier for religion
  - relationship_to_hh_head - Respondent's relationship to the head of the household
- **EDUCATION**
  - education_level - the Highest level of education (0=no education, 1=primary education, 2=secondary education, 3=higher education)
- **EMPLOYMENT**
  - employment_type_last_year - Type of employment last year (e.g. salaried, seasonal)
  - share_hh_income_provided - Share of household income provided
- **ECONOMIC**
  - num_shocks_last_year - Number of financial shocks experienced in the last year
- **PHONE**

    - 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)
- **FINANCIAL INCLUSION**
    - num_financial_activities_last_year - Number of different types of financial activities conducted in the last year

The initial exploration of the data began with a summary and descriptive statistics of the training data set.

DESCRIPTIVE STATISTICS AND VISUALIZATION

Numeric features

Summary statistics for count, mean, standard deviation, minimum and maximum of the numeric features are stated here:

Feature	count	mean	std	min	max
row_id	12600.0	6.299.500.000	3.637.451.031	0.0	12599.0
age	12600.0	36.280.714	15.145.945	15.0	115.0
education_level	12364.0	1.316.241	0.905442	0.0	3.0
share_hh_income_provided	12295.0	2.888.166	1.564.284	1.0	5.0
num_times_borrowed_last_year	12600.0	0.657698	0.924598	0.0	3.0
borrowing_recency	12600.0	0.866429	0.960866	0.0	2.0
bank_interest_rate	289.0	9.843.080	15.033.089	0.0	100.0
mm_interest_rate	151.0	9.021.026	13.620.161	0.0	100.0
mfi_interest_rate	201.0	10.909.204	10.353.298	0.0	100.0
other_fsp_interest_rate	239.0	8.216.736	10.649.538	0.0	100.0
num_shocks_last_year	12600.0	1.100.159	1.190.072	0.0	5.0
avg_shock_strength_last_year	12600.0	2.112.765	2.019.239	0.0	5.0
phone_technology	12600.0	1.208.730	1.093.060	0.0	3.0
phone_ownership	12600.0	1.468.254	0.776638	0.0	2.0
num_formal_institutions_last_year	12600.0	0.714127	0.805878	0.0	6.0
num_informal_institutions_last_year	12600.0	0.188968	0.473696	0.0	4.0
num_financial_activities_last_year	12600.0	1.559.683	2.043.831	0.0	10.0
poverty_probability	12600.0	0.611272	0.291476	0.0	1.0

- **Data types**
    - Most of the numerical features have a small range being between 0 to a maximum of 10 — the exception being, age, that has a minimum value of 15 and a maximum value of 115. The numerical features will need to be scaled if using a linear regression model. However, the final algorithm used is a boosted decision tree that does not require scaling.
    - The numerical features contain categorical features, encoded as numerical features.
        - Education level, avg_shock_strengh_last_year, phone_technology, phone_ownership

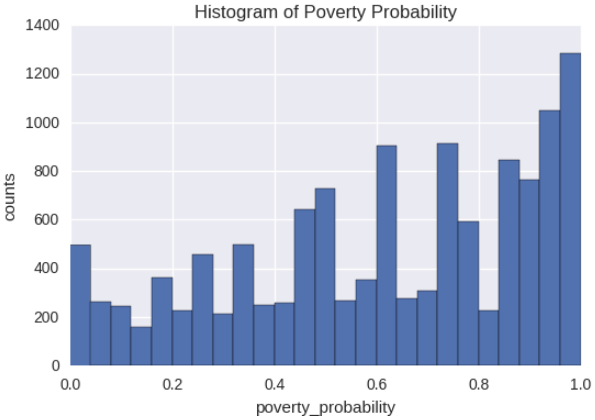
        - These are included as part of further analysis with the categorical data exploration and labelled as categorical data in Azure ML studio.

- **Poverty probability** is the label to predict, and from the summary statistics, this label has a float with a range between 0 and 1 with a mean of 0.6112. A regression machine learning model will, therefore, be created instead of a classification model to predict the label.

- **Missing values**
    - The descriptive statistics identified missing values based on the count field. Further analysis using the Pandas IsNull().sum() function identified all missing values.
    - education level and share_hh_income_provided
        - These features consist of 2% of missing values, and therefore, the missing rows are removed from further analysis.
    - bank interest rate, mm_interest_rate, mfi_interest_rate and other_fsp_interest_rate.
        - These features consist of 98% and more missing values, and therefore, the columns are dropped from further analysis.
    - After removing the missing data, there are 12068 rows and 56 features in the data set remaining.

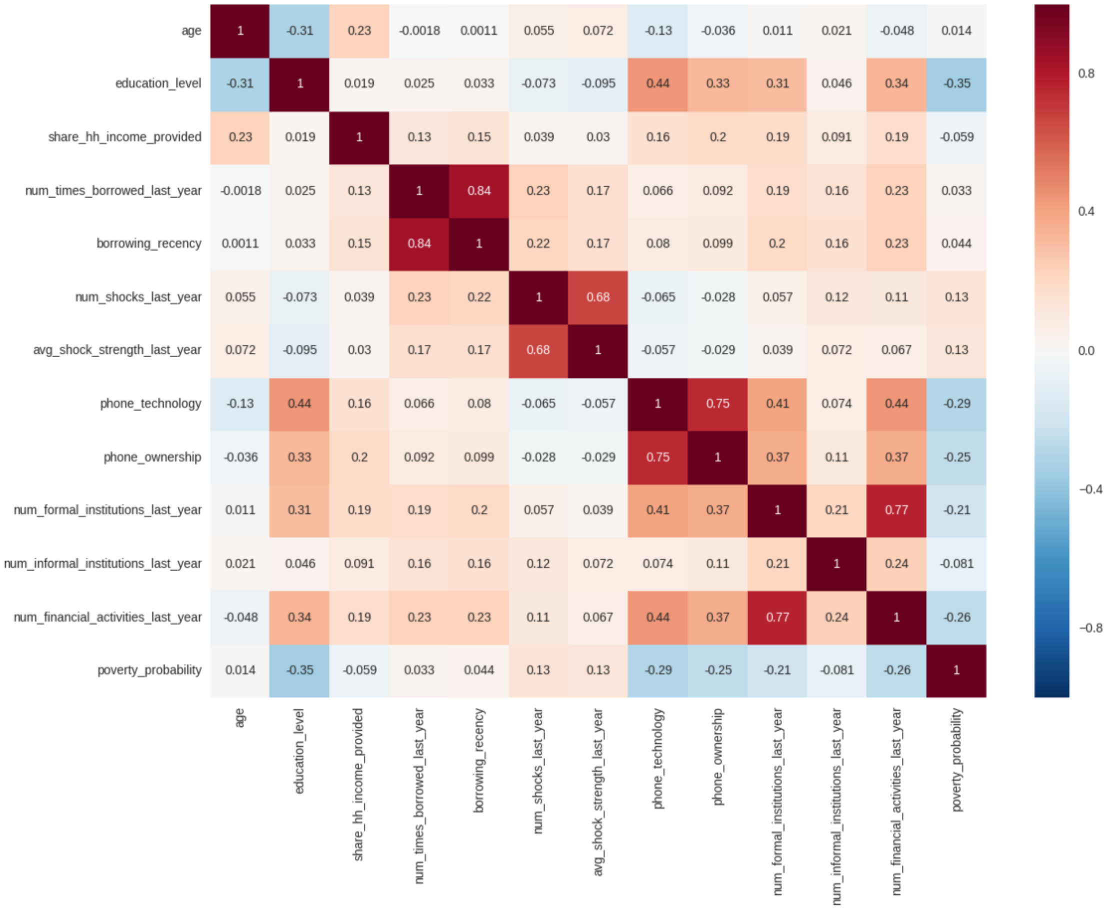
Poverty probability

A histogram is created to understand the distribution of the label poverty probability. The histogram is left-skewed, and there are several modes with the largest **mode** occurring at or near the one value range. Meaning that the largest proportion of poverty probabilities in the data set is near one.



Further analysis in the report mainly focuses on the predictive influence of the features on the poverty probability label to identify features for use in the regression model.

Numeric feature correlation



	age	education_level	share_hh_income_provided	num_times_borrowed_last_year	borrowing_recency	num_shocks_last_year	avg_shock_strength_last_year	phone_technology	phone_ownership	num_formal_institutions_last_year	num_informal_institutions_last_year	num_financial_activities_last_year	poverty_probability
age	1	-0.31	0.23	-0.0018	0.0011	0.055	0.072	-0.13	-0.036	0.011	0.021	-0.048	0.014
education_level	-0.31	1	0.019	0.025	0.033	-0.073	-0.095	0.44	0.33	0.31	0.046	0.34	-0.35
share_hh_income_provided	0.23	0.019	1	0.13	0.15	0.039	0.03	0.16	0.2	0.19	0.091	0.19	-0.059
num_times_borrowed_last_year	-0.0018	0.025	0.13	1	0.84	0.23	0.17	0.066	0.092	0.19	0.16	0.23	0.033
borrowing_recency	0.0011	0.033	0.15	0.84	1	0.22	0.17	0.08	0.099	0.2	0.16	0.23	0.044
num_shocks_last_year	0.055	-0.073	0.039	0.23	0.22	1	0.68	-0.065	-0.028	0.057	0.12	0.11	0.13
avg_shock_strength_last_year	0.072	-0.095	0.03	0.17	0.17	0.68	1	-0.057	-0.029	0.039	0.072	0.067	0.13
phone_technology	-0.13	0.44	0.16	0.066	0.08	-0.065	-0.057	1	0.75	0.41	0.074	0.44	-0.29
phone_ownership	-0.036	0.33	0.2	0.092	0.099	-0.028	-0.029	0.75	1	0.37	0.11	0.37	-0.25
num_formal_institutions_last_year	0.011	0.31	0.19	0.19	0.2	0.057	0.039	0.41	0.37	1	0.21	0.77	-0.21
num_informal_institutions_last_year	0.021	0.046	0.091	0.16	0.16	0.12	0.072	0.074	0.11	0.21	1	0.24	-0.081
num_financial_activities_last_year	-0.048	0.34	0.19	0.23	0.23	0.11	0.067	0.44	0.37	0.77	0.24	1	-0.26
poverty_probability	0.014	-0.35	-0.059	0.033	0.044	0.13	0.13	-0.29	-0.25	-0.21	-0.081	-0.26	1

Perfect correlation is values of 1 or -1, with the positive correlation being a positive change in the one variable resulting in a positive change in the other variable and negative correlation being a positive change in the variable resulting in a negative change in the other variable.

Within the data set, there are moderate,0.68, to strong,0.84, positive correlations:

- num_time_borrowed last year and borrowing_recency
- and phone_technology and phone_ownership
- num_formal_institutions_last_year and num_financial_activities.

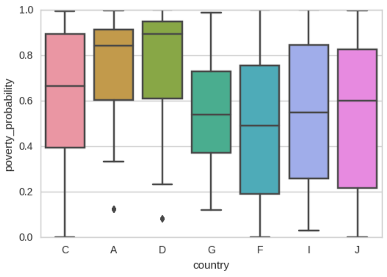
The author theorises that the moderate to strong correlations identified above are indications that the data are very similar and will not add significant strength to the predictive ability of the model used if included in the prediction model. As an example, num_of_times_borrowed_last_year and borrowing_recency is similar in nature. Keeping these features within the data set can result in overfitting of the model.

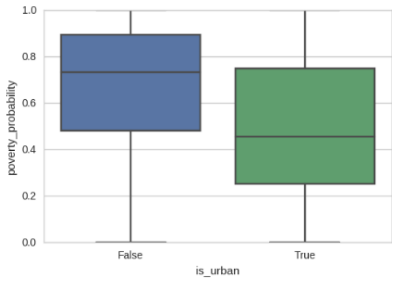
Also, multicollinearity in the features selected will influence the algorithm used, for example, logistic regression, and this will impact linear regression, but decision trees and boosted decision trees will not. For impacted algorithms, highly correlated features should be excluded from the model.

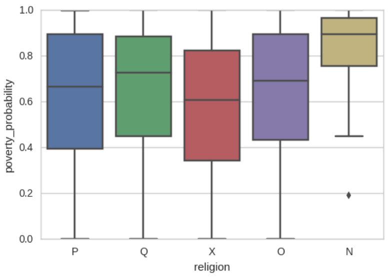
The education level is the largest negative correlation with poverty probability, which means that as the education level increases the poverty probability decreases. Also, as phone technology increases, poverty probability decreases. However, no specific features are moderate to strongly correlated with the label poverty probability.

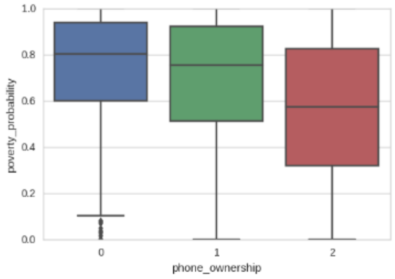
Categorical features

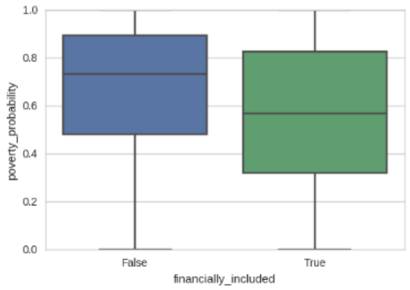
To understand the distribution of the categorical feature data, compared to the poverty probability label, the author used the seaborn package with Python for exploring the data through the use of box plots and violin plot visualisation. The distribution and relationships of the data in the plots support feature selection in the predictive regression model created. The following box plots are identified, highlighting the most influential relationships between the features and the poverty probability label:

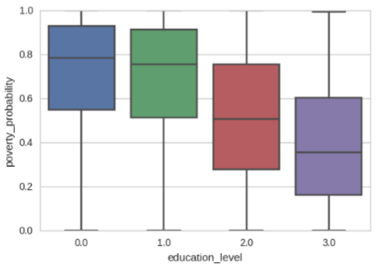


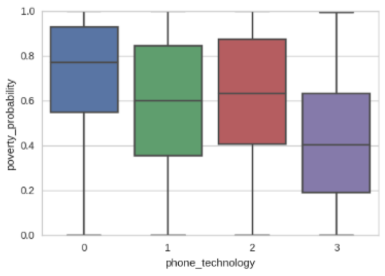


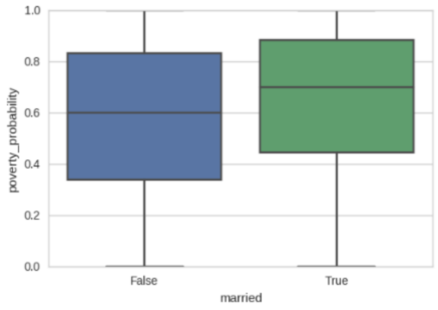


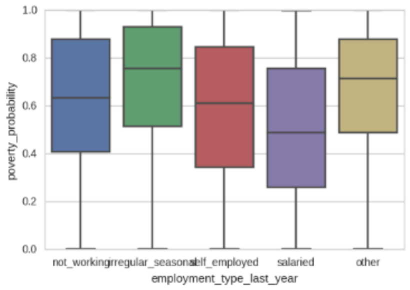








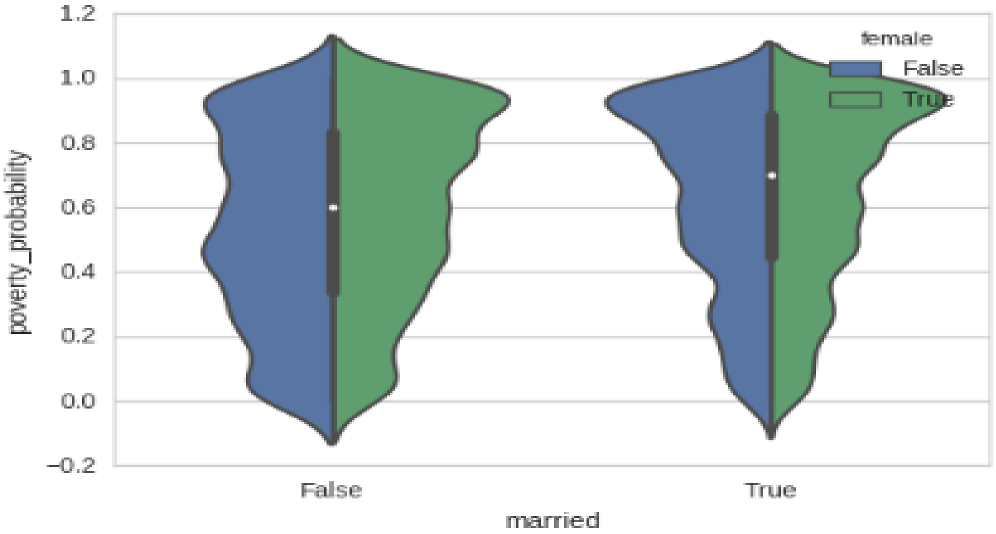




The categorical feature box plots show differences in terms of the maximum, minimum, median and IQR (interquartile range) on the poverty probability values. For example:

- **Country**
  - Based on the plot, there are seven countries represented by letters in the data set; the country names are obscured in the data set. Depending on the country as depicted, there is either a higher or lower IQR of poverty probabilities that the individuals in the data set are in based on the country. Countries A and D's median values are above 0.8 poverty probability compared to countries G and F that have medians below 0.8. Also, the IQR of A and D are closer to 1 compared to, and F. Therefore A and D countries tend to have a higher probability of being in poverty compared to G and F.
- **Urban**
  - Individuals living within the urban area represented as True, have a lower IQR on the poverty probability scale compared to individuals living in rural areas. Individuals that are living in urban areas are more dispersed, based on the IQR on the poverty scale, compared to rural individuals that have a concentrated IQR higher on the poverty probability scale. This indicates that rural individuals are more likely to live in poverty compared to urban individuals in the data set.
- **Phone ownership**
  - Individuals with either no phone or share a phone are more likely to be living within poverty compared to individuals that own a phone. Further analysis is needed to understand if phone ownership results in lower poverty levels or if there is a hidden variable explaining the relationship.
- **Religion**
  - Similar to the country feature, religion names are obscured and replaced with letters representing the religions. From the religion box plot, we can see that the N religion is indicating a higher IQR of poverty probability compared to the X religion.
- **Married**
  - Being married shows a slightly higher IQR for poverty probability compared to not being married per the box plot.

- A **violin plot** with the colours representing green for female and blue for a male is created to identify the relationship between being married or not on sex.



- The violin plot indicates that being female and unmarried has a higher frequency of being in poverty compared to being male, represented by the larger green than blue in the left violin plot near the 1 level. The above is not consistent with being married, where the frequency is similar for males and females.

Similar observations can be made for the other categorical data depicted. However, the above are the most significant findings.

PREDICTIVE MODEL

The label poverty probability is a float between 0.0 and 1.0 as identified through the descriptive statistics and visualization analysis. Various machine learning algorithms were tested during the predictive experiment. However, the boosted decision tree regression algorithm is used, due to the class imbalances identified, potential multicollinearity and providing the best r-squared metric.

For feature selection, the following methods are used:

Data exploration

The data exploration, as noted above-identified relationships between the features and the label. These were taken as the first indication of features to include in the predictive model. Feature selection reduces the model overfitting to the training data and supports the model to generalize well with new data sets.

Starting with the data exploration features and using greedy forward selection to add features based on the Azure ML model filter-based feature selection, additional features are added to the predictive model. These features were used as a starting point to create the experiment. The permutation feature importance model from Azure ML is used to identify features contributing to the metric and negative features were removed from the selection.

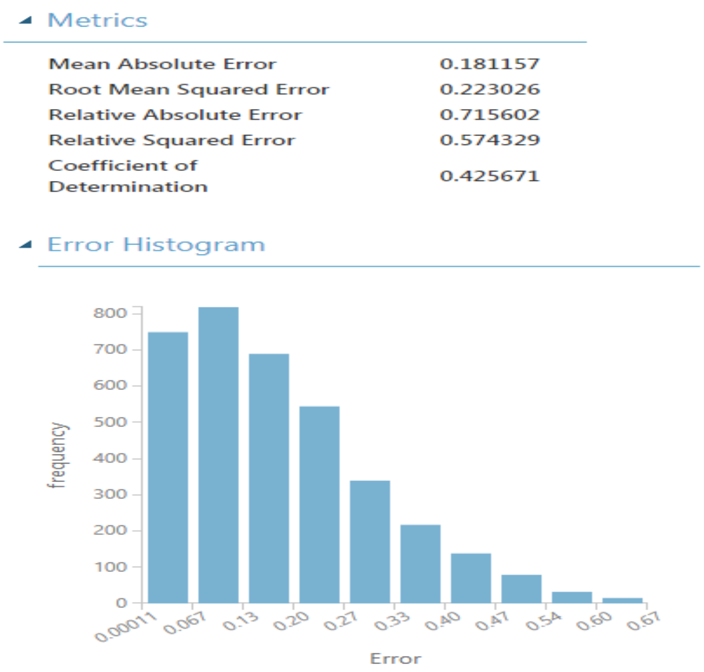
The top 10 features used in the model are:

Importance	Feature	Score
0	country	0.432473
1	education_level	0.128053
2	age	0.125748
3	phone_ownership	0.081178
4	is_urban	0.075994
5	employment_type_last_year	0.059463
6	religion	0.057826
7	num_shocks_last_year	0.053207
8	relationship_to_hh_head	0.044440
9	share_hh_income_provided	0.040968
10	num_financial_activities_last_year	0.038949

The model is trained using the cross-validate module, with ten folds. The first predictive experiment yielded a low score for the r-squared metric(coefficient of determination). The hyperparameters are adjusted to increase the score using the hyperparameter selection model within azure ml. The following hyperparameters are selected as a result of the tuning:

Number of leaves	Minimum leaf instances	Learning rate	Number of trees
13	16,00	0.029474	406,00

Hyperparameter tuning increased the results and yielded the following metrics:



◢ Metrics

Mean Absolute Error	0.181157
Root Mean Squared Error	0.223026
Relative Absolute Error	0.715602
Relative Squared Error	0.574329
Coefficient of Determination	0.425671

◢ Error Histogram

Conclusion

The analysis has shown that the poverty probability can be predicted using the features in the data. Particularly country, urban, religion, married, phone ownership, urban or rural and education, among others are contributors to the poverty probability likelihood.

The predictive model created on the test data resulted in an r-squared metric of 0.42. Also, the model is tested on new data, and an r-squared score of 0.4119 is achieved. The model, therefore, generalized well with new data provided.