2019

Microsoft Professional Capstone: Data Science Report (DAT102x)

Report By
Rahul Palaniappan Kanthabhabha Jeya

Executive Summary:

This report briefly explains the analysis performed on the global poverty data to determine the probability of poverty with respect to the numerous contributing features. This entire project is part of the Microsoft Professional Data Science Capstone project and all the data used in the analysis are provided as part of the course. The total data is split and provided as three csv files – train_values, test_values, train_labels, in addition to submission format file (to submit data for evaluation). The train_values and train_labels files contain 12600 observations with 60 and 2 columns (variables) respectively.

A preliminary or initial basic analysis of the data is performed by exploring the data in R using the summary and descriptive statistics, in excess to understanding the relationship between different variables and label using visualizations methods. This quick overview gave some insights on the relationship between variables and the chances of poverty. After understating the nature of analysis to be made, several regression machine learning models were implemented and tested against the test_values dataset to predict the poverty probability the test data. Out of the all the tested regression model, Random Forest Regression model out performed all the other models and gave a coefficient of determinants (r^2) of 0.413.

After performing the analysis, the author provides the following inferences. Even though most of the variable features have say in the prediction of the poverty probability, some the top features are highlighted below,

- **Country** – It is one of the most important factors in predicting the probability of poverty, where some the countries have a higher significance to chances of poverty of a person.
- **Education Level** – This variable is second the country variable in terms of predicting the poverty status of an individual.
- **Is Urban** – The place of habitat is the next most dominating feature in predicting the label.
- **Age** – As expected the age a person has noteworthy say in determining the poverty level of a subject.
- **Relationship to the household head, religion, phone technology and number of financial activities last year**.

Preliminary Data Exploration:

For this purpose of study, the initial exploration of the train_values data set is performed in a combination of MS Excel and R programming. To simplify the analysis, the poverty_probability values were imported from the train_labels file to train_values file. Using the structure (str) function, it is understood that there are 30+ variables as logical, 10+ varaibles as numerical and few variables as factor and categorical data type. Out the 60 variables, row id variable (no significance on the label) and 4 interest rate variables (due to lack of data on them) are removed from the data set used in the analysis.

Statistics of Contributing Features

The summary statistics of the all the features in the train_values-labels data is performed using the summary function in R programming language and they are provided in the following table.

Table 1: Summary statistics of numerical variables.

Variable Feature	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Age	15.0	25.0	33.0	36.3	45.0	115.0
Num. of shocks la st year	0.0	0.0	1.0	1.1	2.0	5.0
Ave. shock strength last year	0.0	0.0	2.0	2.1	4.0	5.0
Num. of formal insti. last year	0.0	0.0	1.0	0.7	1.0	6.0
Num. of informal insti. last year	0.0	0.0	0.0	0.2	0.0	4.0
Num. of financial act. last year	0.0	0.0	1.0	1.6	3.0	10.0
Poverty Probability	0.0	0.39	0.63	0.61	0.88	1.0

Only summary of few variables is provided in the above table 1. In this study, the probability of poverty is the label that we will predict by training the feature variables using the regression model. It is interesting to see that there is not much significant difference in the mean the median value of the probability of poverty in the given train dataset. This points out that more people in the given data are in under poverty or more susceptible to poverty, if we consider a threshold of 0.5 as the chance of poverty. A histogram of the poverty probability shows that it is left skewed or in simpler terms, more people are under poverty. Nearly, 70% of the population of the data are under poverty.

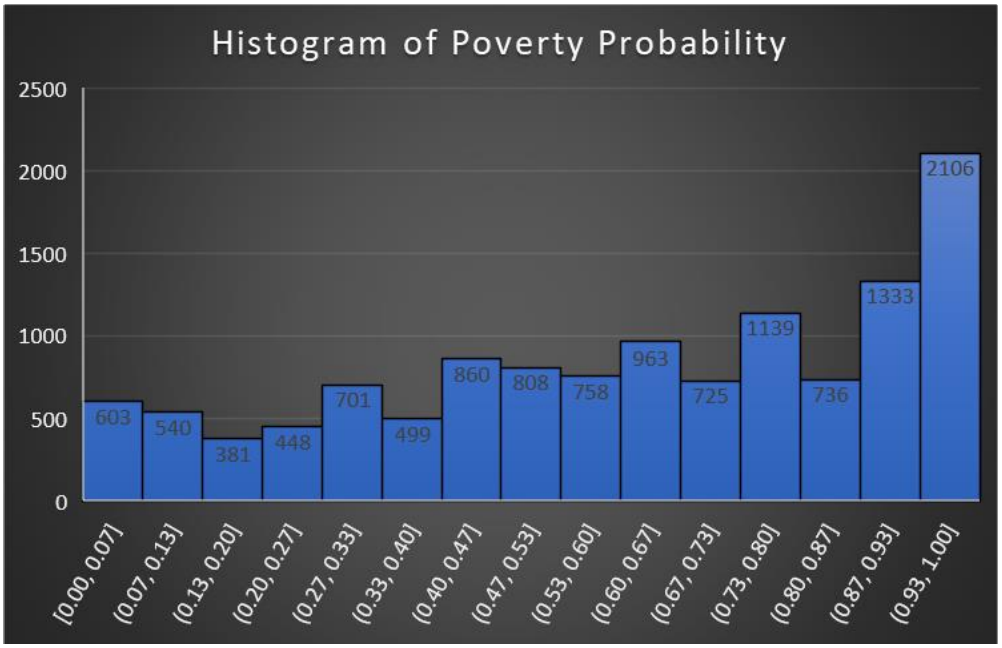

Figure 1: Histogram of Poverty Probability

Apart from the numerical variables, some of the other type of variables that influences the probability of poverty are included below,

- **Country** – a total of 7 countries
- **Is_Urban** – True or False
- **Female -** True or False
- **Married -** True or False
- **Religion** – a total of 5 religion
- **Relationship to the household head** – a total of 7 categories
- **Can add -** True or False
- **Can divide -** True or False
- **Can calculate percentage -** True or False
- **Can calculate compounding -** True or False
- **Employed last year -** True or False
- **Employment category last year -** a total of 5 categories
- **Employment type last year -** a total of 5 categories
- **Income ag &livestock LY -** True or False
- **Income friends & family LY -** True or False
- **Income gov LY -** True or False

- **Income own business LY -** True or False
- **Income private sec LY -** True or False
- **Income public sec LY -** True or False
- **Formal savings -** True or False
- **Informal savings -** True or False
- **Has insurance -** True or False
- **Has investment -** True or False
- **Borrowed for Em. LY -** True or False
- **Borrowed for DE. LY -** True or False
- **Borrowed for home LY -** True or False
- **Financially included -** True or False

In order to understand the frequency of data under each variable or column feature, bar charts can be used to analyze them. For this purpose, the MS Excel PivotChart is used to create the charts. Even though, the summary function in R language provides this information numerically, a graphical representation this information would be easier to follow and more appealing to the readers.

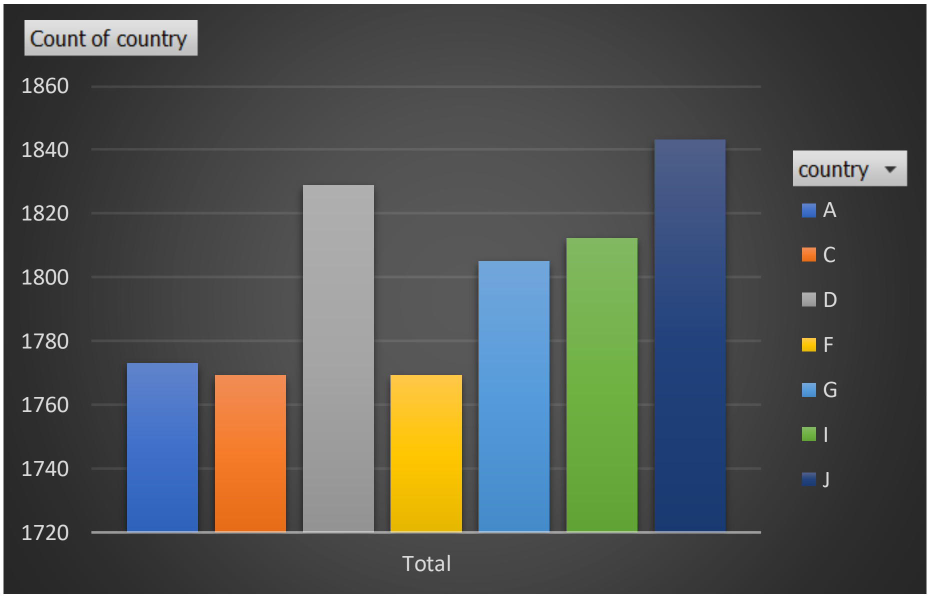

Figure 2: Frequency of countries

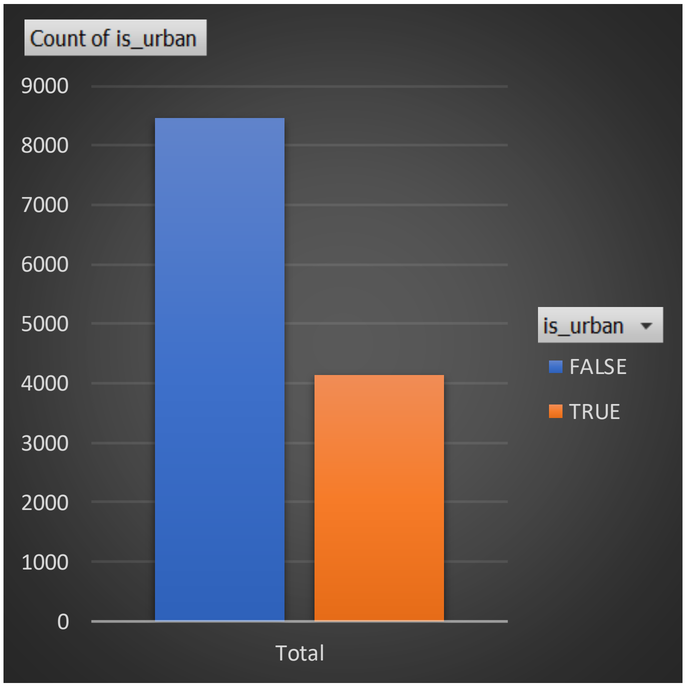

Figure 3: Frequency of Urban Dwellers

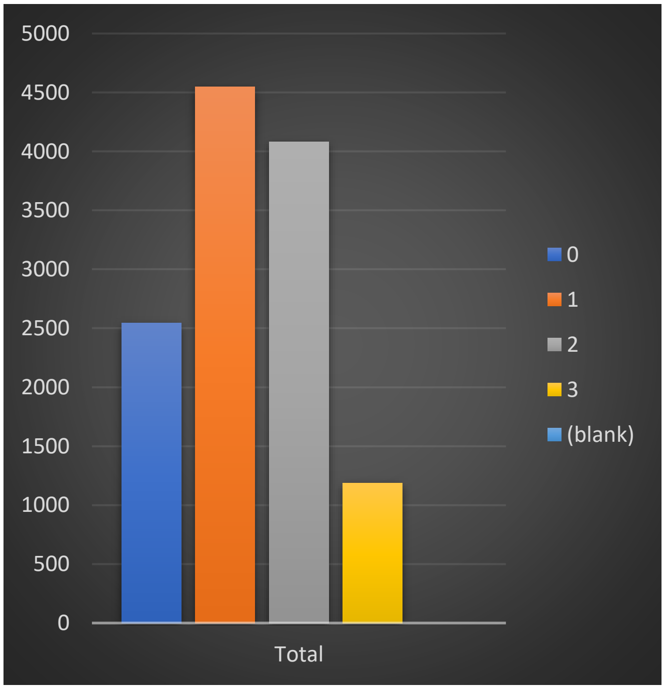

Figure 4: Frequency of females and males

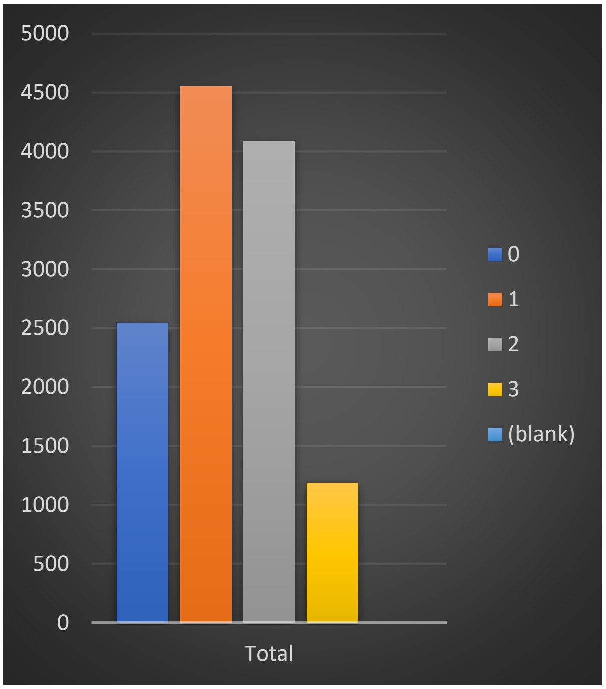

Figure 5: Distribution of level of education where 0 = no education, 1= primary education, 2 = secondary education and 3 = higher education.

Due to the large quantity of the independent variables only few of the important variables are graphically presented in this report. From the figures 2 to 5, it is clear that the surveyed data has nearly two-thirds of females with zero to little education and most of them are not city dwellers.

Regression Model:

After initial exploration of the data through visualization and summary statistics, several types of regression models are tried and tested to obtain a coefficient of determinants that is greater than 0.41. Some the tried and tested regression models are Linear Multiple Regression, Logistic Regression, Support Vector Machine Regression, Decision Tress Regression and Random Forest Regression. Out of all the tested regression models, the Random Forest Regression model fit the data with highest accuracy. To validate the regression model, all the regression models are tested on the train_values-labels data set. This data set is split into train and test data groups by using the createPartitionData function in the ratio of 3:1 in the train to test groups. Subsequently, regression analysis is performed, starting from the simplest Multiple Linear regression to Ridge and Lasso regression. Later more advanced regression techniques like Decision Tree and Random Forest regression are utilized.

The Random Forest regression model made a significant leap in terms of performance as compared to the other regression model. Once the regression model to use is fixed, the existing split of dataset is removed and the new test_values data set is used as the test group, while the entire train_values-labels dataset is used the train group. The same procedure is used to compute Random Forest regression with the new train data and then predict function from the caret package is used to predict the probability of poverty of the test dataset. The applied Random Forest model gave a var % of 43.46 and a mean of sq. residuals = 0.048. By plotting the random forest, it is easy to understand the number of trees required to predict accurate result. Form figure 6, it is obvious that after 175 trees, the reduction in the error is very minimal. This in turns means, that there is no significant improvement in the performance of the model with increase of trees over 175 trees. Less number of trees requires lesser computing power and time.

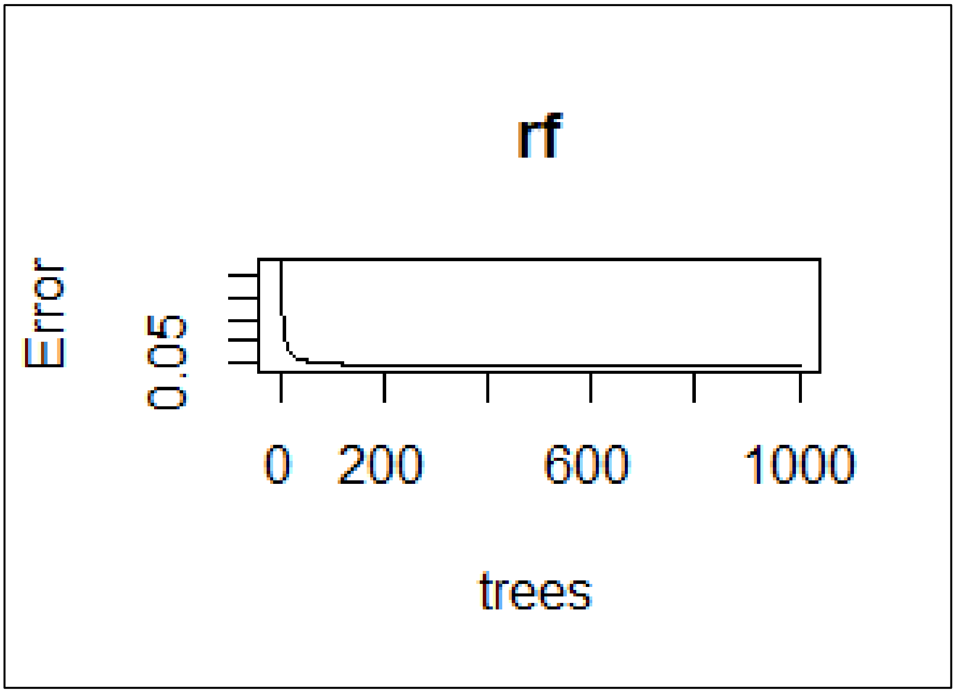

Figure 6: Number of trees vs Error of Random Forest Regression Model

Using the functions – varImpPlot(), importance(), varUsed(), it is easier to understand the importance of the feature variables with respect to the label. From the figure 7, it can be seen that the top five variables have a greater impact on the probability of poverty over the rest of the item. The values of %IncMSE and IncNodePurity for all the variables in the dataset is given in the table 2.

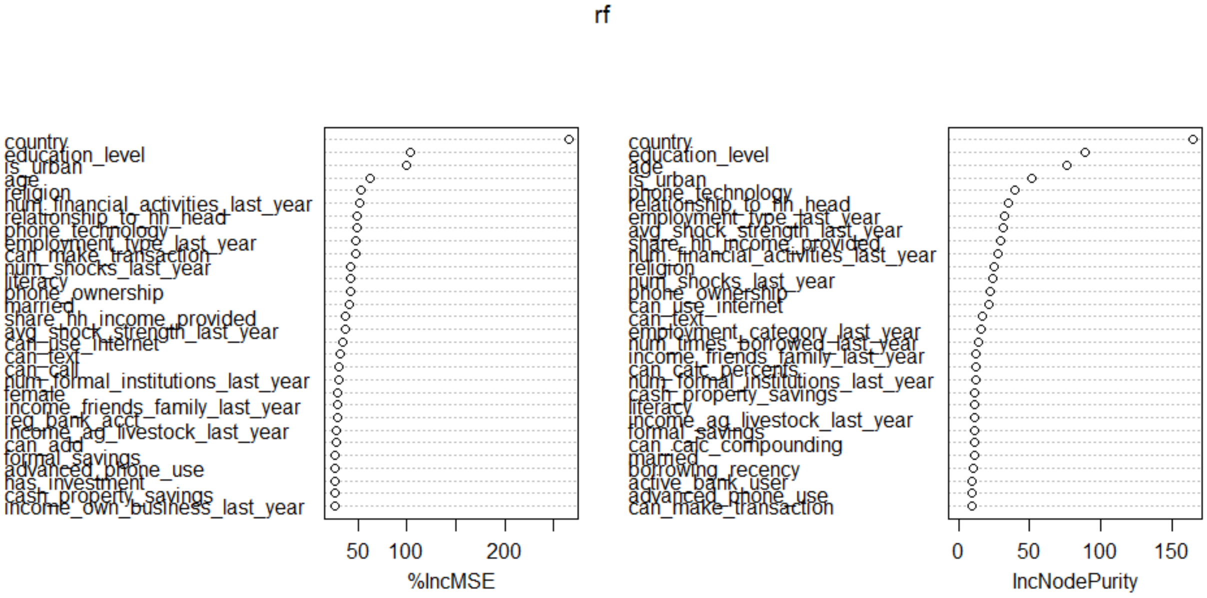

Figure 7: Importance of each variable feature in the data set.

Table 2: Random Forest Regression – variable importance

Variable or feature column	%IncMSE	IncNodePurity	varUsed
country	265.5613	164.8467	130287
is_urban	99.5518	51.67984	28394
age	62.71652	76.49581	433330
female	29.67262	9.316028	71263
married	40.8336	10.9589	72273
religion	52.83706	24.83188	90195
relationship_to_hh_head	48.62384	34.70762	178120
education_level	103.5195	88.76126	92909
literacy	42.21806	11.55385	59158
can_add	27.44653	7.214078	39679
can_divide	16.24833	8.046799	57960
can_calc_percents	14.19447	12.10818	94798
can_calc_compounding	11.99086	11.06021	97146
employed_last_year	17.05415	4.125182	39628
employment_category_last_year	25.31201	16.37195	100864
employment_type_last_year	48.02012	32.06116	169554

share_hh_income_provided	37.62948	29.79081	203869
income_ag_livestock_last_year	28.09336	11.49437	81559
income_friends_family_last_year	29.63989	12.23567	79229
income_government_last_year	9.248902	5.35363	29991
income_own_business_last_year	25.80487	7.96958	63184
income_private_sector_last_year	19.73575	7.236506	39083
income_public_sector_last_year	18.26045	3.868124	16757
num_times_borrowed_last_year	22.90504	13.99293	109813
borrowing_recency	20.89584	10.5948	84020
formal_savings	26.81895	11.39371	43878
informal_savings	16.63977	5.989173	49256
cash_property_savings	25.82655	11.56176	91403
has_insurance	22.93742	7.267211	44334
has_investment	25.8847	9.561383	61268
num_shocks_last_year	42.22946	23.74852	153716
avg_shock_strength_last_year	36.94458	31.21931	208505
borrowed_for_emergency_last_year	22.80617	6.927038	54037
borrowed_for_daily_expenses_last_year	13.71272	5.878098	49750
borrowed_for_home_or_biz_last_year	14.73653	4.895718	39013
phone_technology	48.49997	39.68509	95712
can_call	30.85244	9.457647	63196
can_text	31.95028	16.54623	56353
can_use_internet	33.91371	20.97977	38282
can_make_transaction	47.02646	9.746074	50974
phone_ownership	41.88283	22.38093	57714
advanced_phone_use	26.80723	9.752677	50540
reg_bank_acct	29.21397	9.711704	30607
reg_mm_acct	23.91615	3.987234	30296
reg_formal_nbfi_account	10.39842	3.05565	22749
financially_included	21.36153	5.459677	35069
active_bank_user	22.41589	9.956005	29391
active_mm_user	24.78641	4.186021	29625
active_formal_nbfi_user	9.026277	2.333921	16766

active_informal_nbfi_user	16.70069	4.89723	37626
nonreg_active_mm_user	5.884565	3.565345	29095
num_formal_institutions_last_year	30.30261	11.865	75441
num_informal_institutions_last_year	19.99646	6.956657	49817
num_financial_activities_last_year	51.92455	27.54173	140241

The table 2 gives the values of the variables used in the regression model, the purity of the nodes of the tress in the model and importance of each variable with respect to the label. It is clear that all the variables have an effect on the poverty probability but the magnitude of the effect is stand out for some features like country, is_urban, age, religion, phone technology, number of financial activities last year. In addition, I tried tuning of Random Forest Model with a stepFactor = 0.5 and ntrees = 300 to find the convergence of mtry. It can be seen from the figure 8 that the mtry converge at 18 (mtry – 18). It also indicates that the OOB error decreases a little.

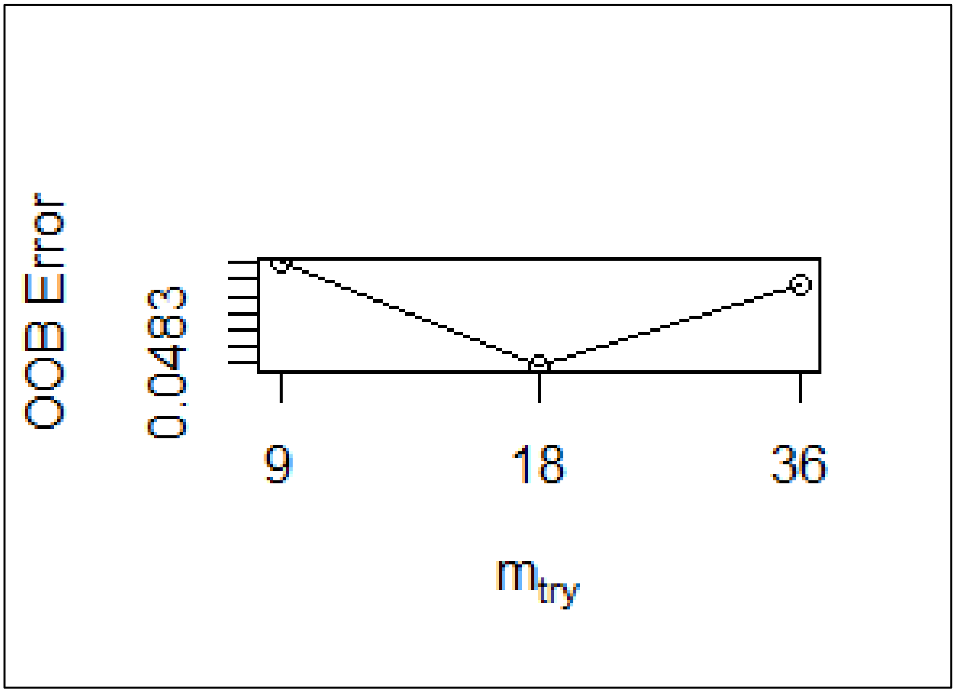

Figure 8: Tuning of RF

Conclusion:

This project gave an opportunity to perform regression machine learning techniques to predict the probability of poverty with the data provided. From this study, it is evident that the Random Forest regression model is best for poverty prediction with R^2 of 0.413. The primary variable that contribute the prediction of label are country, is urban, age, phone technology, number of financial activities, religion, relationship to the household head and employment type last year. All the other variables are secondary that aide in making the model more accurate. The challenge 1 of the capstone project is solved using MS Excel and R programming language, while the challenge 2 is completely done in R language.