MICROSOFT DATA SCIENCE PROFESSIONAL CAPSTONE REPORT

Word Count : 1547

Predicting Poverty Around the World

Sanil Ahan Purryag

July 13, 2019

1 EXECUTIVE SUMMARY

This document presents an overview of features and relevant findings pertaining to the prediction of poverty around the world, carried out for the DAT102x June 2019 Capstone machine learning competition. All the analysis was based on obfuscated data derived from PPI surveys and household surveys conducted by InterMedia. The aim of the competition was to predict the probability of an individual living below the poverty threshold of $2.50 per day across 7 different countries. To this effect, 58 features capturing different socio-economic dimensions were provided to help construct a regression model, which would be evaluated using the R-Squared .

The analysis was conducted following the established CRISP-DM framework for data mining using R version 3.6.1. After conducting an exploratory data analysis and feature engineering; different iterations of Supervised Random Forest models, trained on different samples of the data, were used to predict the required probability. The best performing model scored a R-squared 0.3986 on the undisclosed holdout set.

2 Exploratory Data Analysis

2.1 Summary Statistics

This phase included generating summary statistics and visualisations to understand the nature of the data sets provided. The training data, as well as the testing data, encompassed 58 features which captured socio-economic dimensions relating to poverty, and had 12,600 and 8,400 observations respectively. The summary statistics for these data sets can be found in Table 6.1 and Table 6.2 of the Appendix. The latter illustrate the mean, number of observations, standard deviation, percentiles, minimum and maximum for the related dataset.

Additionally, the response variable for the analysis ('Poverty_probability') appears to have a relatively low standard deviation (0.3), as well as a close mean (0.61) and median (0.63) . It thus does not appear to exhibit a high variance in its distribution. The latter distribution is also found to be left-skewed, indicating that most individuals are on the high side of 'Poverty_probability' range. The associated frequency and density plot can be found in figure 6.1 and 6.2 of the Appendix.

2.2 Data Types

Based on the range of the variables observed, it is noted that the data is predominantly composed of categorical variables. 42 variables have been automatically recognised as boolean/binary variables, with an additional 9 variables which can be considered as categorical, on the basis of their range. Thus, there are only 7 remaining variables which can be formatted as numeric. The complete list of variables and how they have been formatted can be found in table 6.3 of the Appendix.

2.3 Missing value Analysis

The analysis on the train and test dataset reveal that the following variables have high percentages of missing values: 'mm_interest_rate', 'mfi_interest_rate', 'other_fsp_interest_rate' and 'bank_interest_rate'. Additionally, 'share_hh_income_provided' and 'education_level' have been found to have lower percentages of missing values. The diagrams illustrating the latter can be found in figures 6.3 and 6.4 of the Appendix. For ease of modelling, the aforementioned variables have been dropped in the later stages of this project.

2.4 Correlation analysis

To understand the interactions between the different variables, for later feature engineering, a correlation analysis has been made for the numerical and categorical variables.

2.4.1 Correlation for Numerical Variables

The Pearson correlation coefficient was computed for the numerical variables and used as a measure of association. The latter is illustrated in figure 6.5 of the Appendix. The Pearson correlation coefficient is computed as follows for two variables X and Y, where the covariance of the latter variables is divided by their respective standard deviations :

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y} \tag{2.1}$$

Where:

- σ_x is the standard deviation of X.
- σ_y is the standard deviation of Y.
- $cov(X, Y)$ is the covariance between X and Y.

The Pearson correlation coefficient ranges from 0 to 1. The only variables with high levels of correlation are ‘num_formal_institutions_last_year’ and ‘num_financial_activities_last_year’ with a value of 0.7745437.

2.4.2 Correlation for categorical Variables

To measure the association between categorical variables, Cramer's V was used. The latter measure ranges from 0 to 1 and is calculated as follows:

$$\phi = \sqrt{\frac{\chi^2}{N(K-1)}} \tag{2.2}$$

Where:

- χ^2 is the Pearson chi-square statistic.
- N is the sample size involved in the test.
- K is the lesser number of categories of either variable.

It is noted that 22 variables have relatively high levels of positive association, as shown in table 6.4 of the Appendix.

3 Feature Engineering

To obtain additional data for the regression, some features have been engineered based on their levels of association discovered in the correlation analysis. The following features have thus been created:

- 'freq_formal_finance' is a categorical variable that indicates the frequency of using 'num_formal_institutions_last_year'. This variable takes a value of 0 if the 'num_formal_institutions_last_year' is less than 3 and 1 if 'num_formal_institutions_last_year' is greater than 3 but less than 5.
- 'tech_proficiency' is a categorical variable that captures how proficient with technology the individual is. This variable takes a 0 if one of 'can_use_internet' or 'can_text' is a 0. This variable takes a 1 if both 'can_use_internet' and 'can_text' are 1.
- 'personal_investment' is a categorical variable that illustrates whether an individual has made investments while having a personal business. This variable takes a 0 if one of 'has_investment' or 'income_own_business_last_year' is a 0. This variable takes a 1 if both 'has_investment' and 'income_own_business_last_year' are 1.

- ‘country_offering’ is a categorical variable that captures whether a country offers access to mobile services. This variable takes a 1 if a country in the dataset has ‘active_mm_user’ with a 1, otherwise it will have a value of 0.
- ‘both_mm_bank’ is a categorical variable that captures if a country has both official banking and mobile services. It takes a 1 if both ‘reg_mm_acct’ and ‘reg_bank_acct’ are 1, otherwise it shall have a value of 0.

After creating the above features, the categorical variables in the data were all converted to dummies (one-hot encoded) to improve the performance of the considered machine learning algorithm in the next phase. This lead to the number of variables increasing to 123 for the training data.

4 MODELLING

Prior to fitting a machine learning model to predict the probability of being under the poverty threshold, the training data provided was divided into a training set and development set with an 80/20 split ratio. A Random Forest model was used to fit the aforementioned data sets and the R-squared was used to monitor the performance of the algorithm on the different data sets. As proposed by Breiman et al. (1984), Random Forests are an ensemble model composed of multiple randomized base regression trees which are combined to form an aggregated regression estimate. The different steps involved in the Random Forest are described in Algorithm 1 of the Appendix, in addition to its different parameters of interest in table 6.5 of the Appendix.

It should also be outlined that a seed of 12345 was used throughout the procedure.Two versions of the Random Forest were trained in this analysis. The first one was trained on the original sized training data (12,600 observations) and the second on the split training data (which is 80% of the original sized training data or 10,081 observations).

4.1 VARIABLE IMPORTANCE

To improve the performance of the algorithm, feature selection was done using the mean decrease in the Gini coefficient. The resulting variable importance analysis indicates that three variables, namely: 'religionN', 'employment_category_last_yearother' and 'employment_category_last_yearunemployed have little to no bearing on the performance of the algorithm. They have thus been removed from both training data sets for later stages. The tables containing the Mean decrease in Gini coefficient can be found in tables and the variable importance plots can be found in figures.

4.2 TUNING USING OUT OF BAG ERROR

Based on Janitza & Hornung (2018), it is important to tune the value of $mtry$ to allow for an adequate number of predictors to be considered at each split, prior to training the Random Forest. It is noted that a too low $mtry$ would lead to the situation where variables with no predictive ability are selected for a split, as a result of the exhaustive search algorithm present in the classification trees and a too high $mtry$ would allow for predictors with the highest predictive ability to selected first consistently, leading to similar trees being ensembled across the Random Forest. This particular parameter was consequently tuned using Out of Bag Error, which led to an $mtry$ of 20 having the lowest error of 0.05175.

5 Conclusion

This analysis has shown that the probability of an individual living under the poverty threshold can be predicted using socio-economic variables, using this methodology, to achieve a R-squared of 0.3986. More specifically, based on the computed variable importance: ‘age’, ‘countryD’ and ‘is_urbanFalse’ have been revealed to be the most prominent contributors to the regression. Additionally, it can be noted that the best implementation of the Random Forest algorithm (trained on the split training data) overfits the training data, as compared to the development and test dataset. This is evidenced by R-squared values obtained on each data set as shown below:

- Training data - 0.8641
- Development data - 0.3787
- Test Data - 0.3986

Therefore, for further research, alternative cross validation methods such as K-fold cross validation could be used for parameter tuning. Also, alternative models could be considered or stacked with the Random Forest to improve the predictive accuracy.

6 APPENDIX

Table 6.1: Summary statistics for Training Data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
row_id	12,600	6,299.5	3,637.5	0	3,149.8	9,449.2	12,599
is_urban	12,600	0.3	0.5	0	0	1	1
age	12,600	36.3	15.1	15	25	45	115
female	12,600	0.6	0.5	0	0	1	1
married	12,600	0.6	0.5	0	0	1	1
education_level	12,364	1.3	0.9	0.0	1.0	2.0	3.0
literacy	12,600	0.6	0.5	0	0	1	1
can_add	12,600	0.9	0.3	0	1	1	1
can_divide	12,600	0.8	0.4	0	1	1	1
can_calc_percents	12,600	0.4	0.5	0	0	1	1
can_calc_compounding	12,600	0.4	0.5	0	0	1	1
employed_last_year	12,600	0.6	0.5	0	0	1	1
share_hh_income_provided	12,295	2.9	1.6	1.0	1.0	5.0	5.0
income_ag_livestock_last_year	12,600	0.4	0.5	0	0	1	1
income_friends_family_last_year	12,600	0.4	0.5	0	0	1	1
income_government_last_year	12,600	0.1	0.2	0	0	0	1
income_own_business_last_year	12,600	0.3	0.5	0	0	1	1
income_private_sector_last_year	12,600	0.1	0.3	0	0	0	1
income_public_sector_last_year	12,600	0.03	0.2	0	0	0	1
num_times_borrowed_last_year	12,600	0.7	0.9	0	0	1	3
borrowing_recency	12,600	0.9	1.0	0	0	2	2
formal_savings	12,600	0.3	0.5	0	0	1	1
informal_savings	12,600	0.2	0.4	0	0	0	1
cash_property_savings	12,600	0.4	0.5	0	0	1	1
has_insurance	12,600	0.1	0.3	0	0	0	1
has_investment	12,600	0.3	0.5	0	0	1	1
bank_interest_rate	289	9.8	15.0	0.0	1.0	14.0	100.0
mm_interest_rate	151	9.0	13.6	0.0	2.8	10.0	100.0
mfi_interest_rate	201	10.9	10.4	0.0	5.0	15.0	100.0
other_fsp_interest_rate	239	8.2	10.6	0.0	2.2	10.0	100.0
num_shocks_last_year	12,600	1.1	1.2	0	0	2	5
avg_shock_strength_last_year	12,600	2.1	2.0	0	0	4	5
borrowed_for_emergency_last_year	12,600	0.2	0.4	0	0	0	1
borrowed_for_daily_expenses_last_year	12,600	0.2	0.4	0	0	0	1
borrowed_for_home_or_biz_last_year	12,600	0.1	0.3	0	0	0	1
phone_technology	12,600	1.2	1.1	0	0	2	3
can_call	12,600	0.8	0.4	0	1	1	1
can_text	12,600	0.5	0.5	0	0	1	1
can_use_internet	12,600	0.2	0.4	0	0	0	1
can_make_transaction	12,600	0.3	0.4	0	0	1	1
phone_ownership	12,600	1.5	0.8	0	1	2	2
advanced_phone_use	12,600	0.3	0.5	0	0	1	1
reg_bank_acct	12,600	0.3	0.4	0	0	1	1
reg_mm_acct	12,600	0.3	0.5	0	0	1	1
reg_formal_nbfi_account	12,600	0.1	0.3	0	0	0	1
financially_included	12,600	0.5	0.5	0	0	1	1
active_bank_user	12,600	0.2	0.4	0	0	0	1
active_mm_user	12,600	0.2	0.4	0	0	0	1
active_formal_nbfi_user	12,600	0.1	0.2	0	0	0	1
active_informal_nbfi_user	12,600	0.1	0.4	0	0	0	1
nonreg_active_mm_user	12,600	0.1	0.3	0	0	0	1
num_formal_institutions_last_year	12,600	0.7	0.8	0	0	1	6
num_informal_institutions_last_year	12,600	0.2	0.5	0	0	0	4
num_financial_activities_last_year	12,600	1.6	2.0	0	0	3	10
poverty_probability	12,600	0.6	0.3	0.0	0.4	0.9	1.0

Table 6.2: Summary statistics for Testing data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
row_id	8,400	4,199.5	2,425.0	0	2,099.8	6,299.2	8,399
is_urban	8,400	0.3	0.5	0	0	1	1
age	8,400	36.5	15.3	15	25	45	117
female	8,400	0.6	0.5	0	0	1	1
married	8,400	0.6	0.5	0	0	1	1
education_level	8,251	1.3	0.9	0.0	1.0	2.0	3.0
literacy	8,400	0.6	0.5	0	0	1	1
can_add	8,400	0.9	0.3	0	1	1	1
can_divide	8,400	0.8	0.4	0	1	1	1
can_calc_percents	8,400	0.4	0.5	0	0	1	1
can_calc_compounding	8,400	0.4	0.5	0	0	1	1
employed_last_year	8,400	0.6	0.5	0	0	1	1
share_hh_income_provided	8,207	2.9	1.6	1.0	1.0	5.0	5.0
income_ag_livestock_last_year	8,400	0.4	0.5	0	0	1	1
income_friends_family_last_year	8,400	0.4	0.5	0	0	1	1
income_government_last_year	8,400	0.1	0.2	0	0	0	1
income_own_business_last_year	8,400	0.3	0.5	0	0	1	1
income_private_sector_last_year	8,400	0.1	0.3	0	0	0	1
income_public_sector_last_year	8,400	0.04	0.2	0	0	0	1
num_times_borrowed_last_year	8,400	0.7	0.9	0	0	1	3
borrowing_recency	8,400	0.9	1.0	0	0	2	2
formal_savings	8,400	0.3	0.5	0	0	1	1
informal_savings	8,400	0.2	0.4	0	0	0	1
cash_property_savings	8,400	0.4	0.5	0	0	1	1
has_insurance	8,400	0.1	0.3	0	0	0	1
has_investment	8,400	0.3	0.5	0	0	1	1
bank_interest_rate	222	9.5	10.2	0.0	2.0	14.0	100.0
mm_interest_rate	94	9.3	7.7	0.0	4.2	14.0	40.0
mfi_interest_rate	98	12.3	10.3	0.0	7.0	15.0	75.0
other_fsp_interest_rate	170	8.8	12.0	0.0	3.0	10.0	100.0
num_shocks_last_year	8,400	1.1	1.2	0	0	2	5
avg_shock_strength_last_year	8,400	2.1	2.0	0	0	4	5
borrowed_for_emergency_last_year	8,400	0.2	0.4	0	0	0	1
borrowed_for_daily_expenses_last_year	8,400	0.2	0.4	0	0	0	1
borrowed_for_home_or_biz_last_year	8,400	0.1	0.3	0	0	0	1
phone_technology	8,400	1.2	1.1	0	0	2	3
can_call	8,400	0.8	0.4	0	1	1	1
can_text	8,400	0.5	0.5	0	0	1	1
can_use_internet	8,400	0.2	0.4	0	0	0	1
can_make_transaction	8,400	0.3	0.5	0	0	1	1
phone_ownership	8,400	1.5	0.8	0	1	2	2
advanced_phone_use	8,400	0.3	0.5	0	0	1	1
reg_bank_acct	8,400	0.3	0.5	0	0	1	1
reg_mm_acct	8,400	0.3	0.4	0	0	1	1
reg_formal_nbfi_account	8,400	0.1	0.3	0	0	0	1
financially_included	8,400	0.5	0.5	0	0	1	1
active_bank_user	8,400	0.2	0.4	0	0	0	1
active_mm_user	8,400	0.2	0.4	0	0	0	1
active_formal_nbfi_user	8,400	0.1	0.2	0	0	0	1
active_informal_nbfi_user	8,400	0.1	0.3	0	0	0	1
nonreg_active_mm_user	8,400	0.1	0.3	0	0	0	1
num_formal_institutions_last_year	8,400	0.7	0.8	0	0	1	6
num_informal_institutions_last_year	8,400	0.2	0.5	0	0	0	3
num_financial_activities_last_year	8,400	1.5	2.0	0	0	3	10

Table 6.3: Chosen types for variables

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

Table 6.4: Cramer's V values for Categorical Variables of Training Data

Var1	Var2	value
active_mm_user	reg_mm_acct	0.8967659
active_formal_nbfi_user	reg_formal_nbfi_account	0.8266498
active_bank_user	reg_bank_acct	0.8151593
employment_type_last_year	employment_category_last_year	0.7071068
financially_included	reg_mm_acct	0.6383604
financially_included	reg_bank_acct	0.6300175
relationship_to_hh_head	female	0.6201747
reg_mm_acct	country	0.5947367
active_mm_user	financially_included	0.5724599
income_own_business_last_year	employment_type_last_year	0.5686458
financially_included	formal_savings	0.5665713
religion	country	0.5641571
active_mm_user	country	0.5638867
relationship_to_hh_head	married	0.5547647
advanced_phone_use	can_use_internet	0.5447834
has_investment	income_own_business_last_year	0.531636
active_mm_user	can_make_transaction	0.5246402
can_use_internet	can_text	0.5241122
active_informal_nbfi_user	informal_savings	0.5234157
reg_mm_acct	can_make_transaction	0.5211743
borrowed_for_daily_expenses_last_year	borrowed_for_emergency_last_year	0.5163308
active_bank_user	financially_included	0.5135646

Table 6.5: RF Parameters of Interest

Parameter name	Description
$mtry$	Number of variables randomly sampled at each tree split
$nodesize$	Minimum size of terminal nodes
$maxnodes$	Maximum size of terminal nodes
$ntree$	Number of trees to grow
$replace$	boolean value to confirm if sampling of cases should be done with replacement
$localImp$	boolean value to confirm if variable importance should be computed

Algorithm 1 RF Classification Algorithm adapted from Liaw & Wiener (2002)

INPUT: Dataset of Interest, $ntree$, $mtry$

OUTPUT: Predicted Probability for each observation, Out of Bag Error

procedure RANDOM FOREST REGRESSION ALGORITHM
  Start
  Compute number of Bootstrap samples from Dataset of interest = $ntree$
  for i in $sample_1$ to $sample_{ntree}$ **do**
    Grow unpruned classification tree
    Identify Out of Bag observations by comparing sample observations against original data
    Randomly sample number of considered variables for split = $mtry$
    Choose best split from $mtry$ predictor
    Save prediction for $sample_i$
      Compare against OOut of Bag observations and obtain Out of Bag error
      Save Out of Bag error
  Return average of predictions (predicted probabilities) from $sample_1$ to $sample_{ntree}$
  Return Out of Bag error associated with $mtry$ value
  End

Table 6.6: Top 10 variables with lowest Mean Decrease in Gini coefficient from original sized training data

Variables	Mean Decrease Gini
employment_category_last_yearunemployed	-2.915501453
religionN	-0.428309114
relationship_to_hh_headFather.Mother	-0.202928534
employment_type_last_yearother	-0.003020263
relationship_to_hh_headUnknown	1.4515079
employment_category_last_yearother	1.582549611
religionO	1.586944598
nonreg_active_mm_userFALSE	3.038840256
nonreg_active_mm_userTRUE	4.056726823
reg_formal_nbfi_accountTRUE	4.444789643

Table 6.7: Top 10 variables with lowest Mean Decrease in Gini coefficient from split sized training data

Variables	Mean Decrease Gini
religionN	-2.627406415
employment_category_last_yearother	-0.823749076
employment_category_last_yearunemployed	-0.671251002
employment_type_last_yearother	0.789184171
religionO	1.447273796
relationship_to_hh_headFather.Mother	2.008282617
relationship_to_hh_headUnknown	2.349196776
active_formal_nbfi_userTRUE	3.570361105
nonreg_active_mm_userFALSE	3.990813507
reg_formal_nbfi_accountFALSE	4.110167246

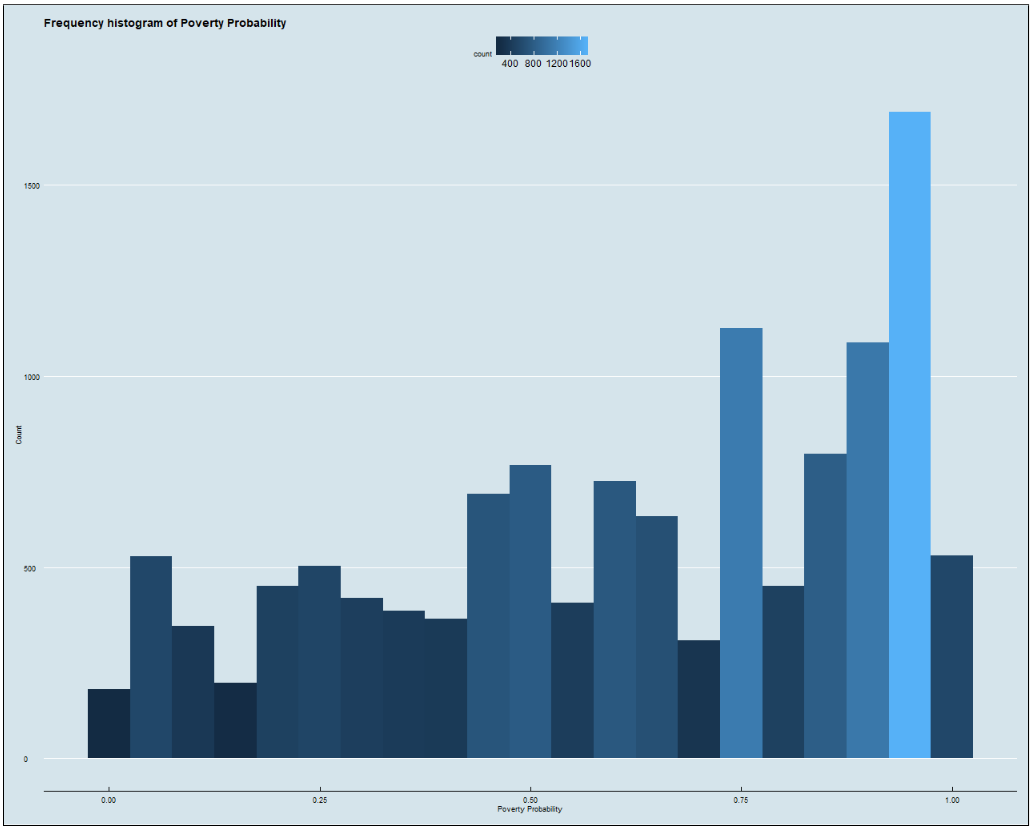

Figure 6.1: Frequency Histogram of Poverty Probability

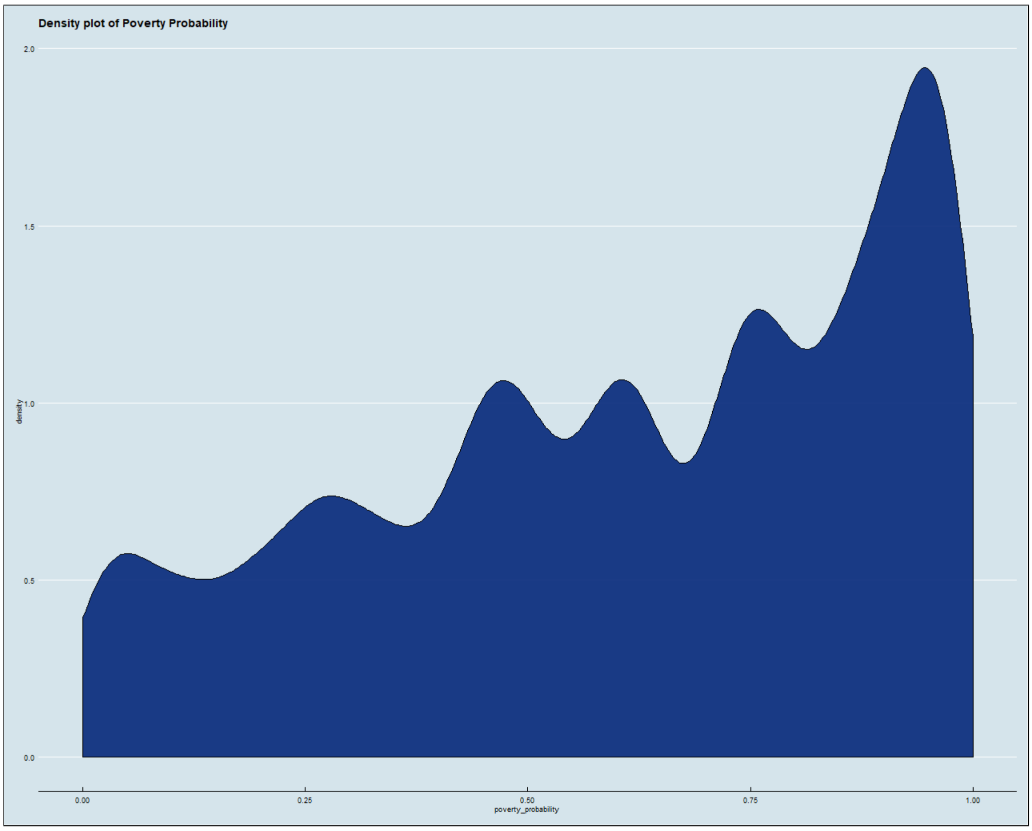

Figure 6.2: Density plot of Poverty Probability

Figure 6.3: Count of Missing values in training data

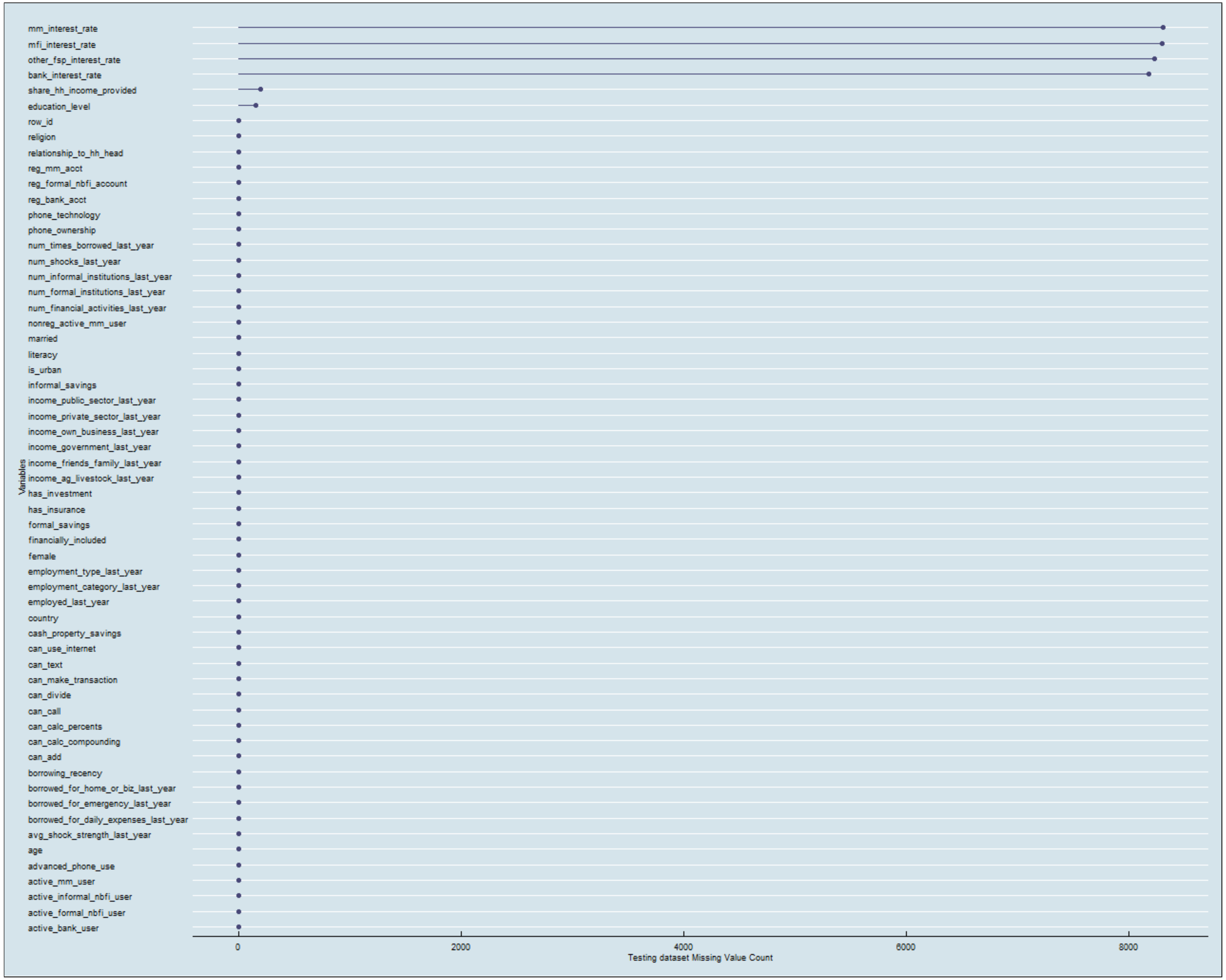

Figure 6.4: Count of Missing values in testing data

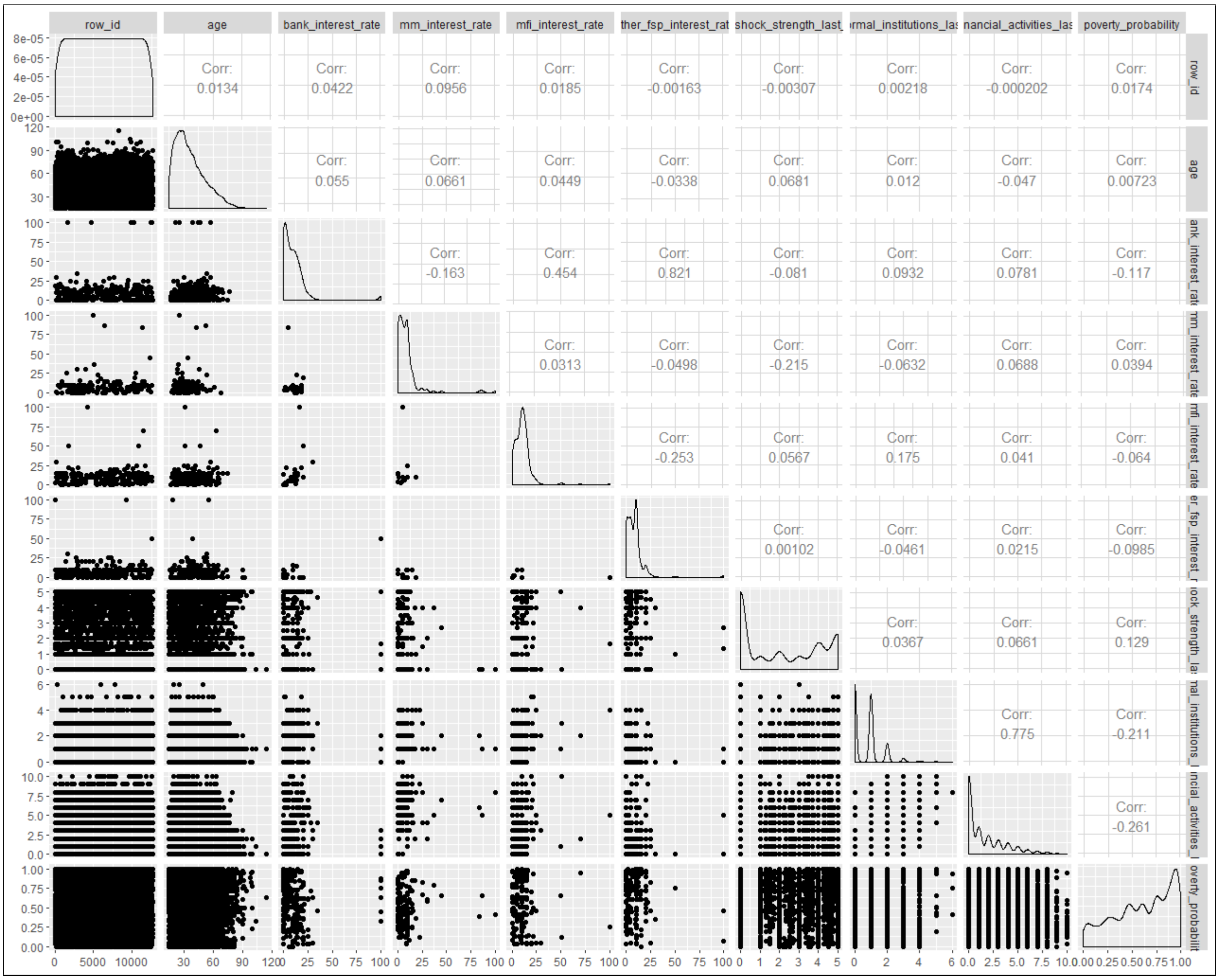

Figure 6.5: Pairs plot for numerical variables of Training Dataset

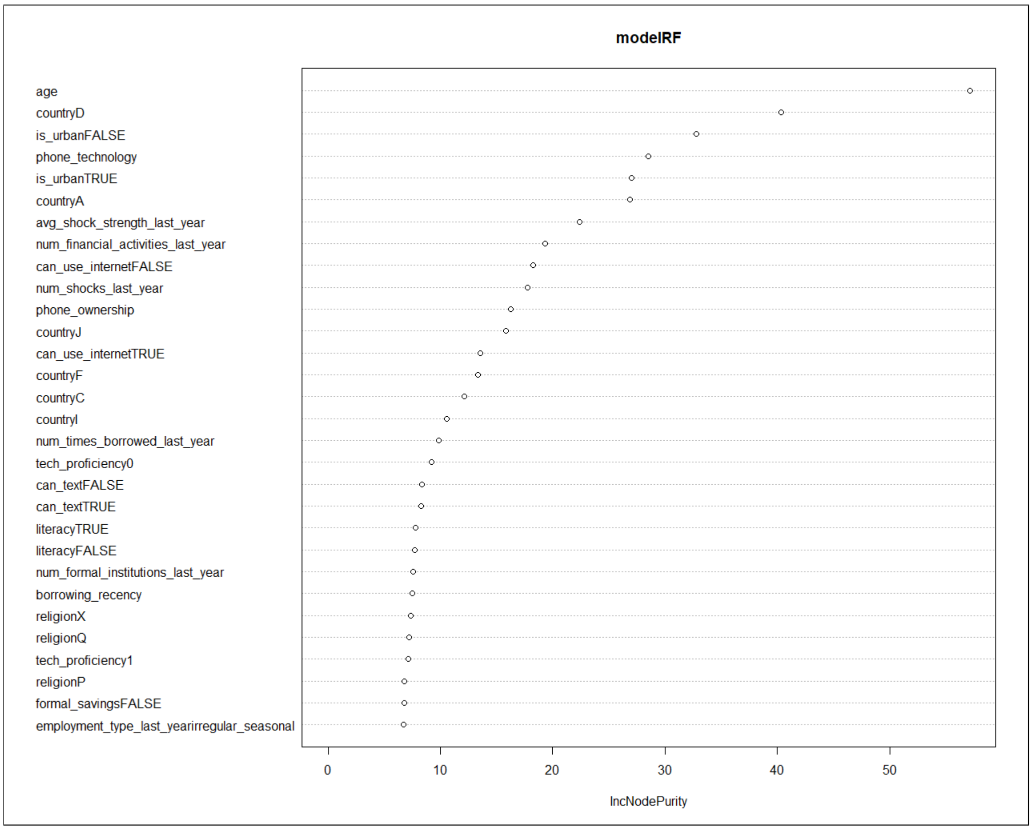

Figure 6.6: Mean decrease in node impurity on split training data

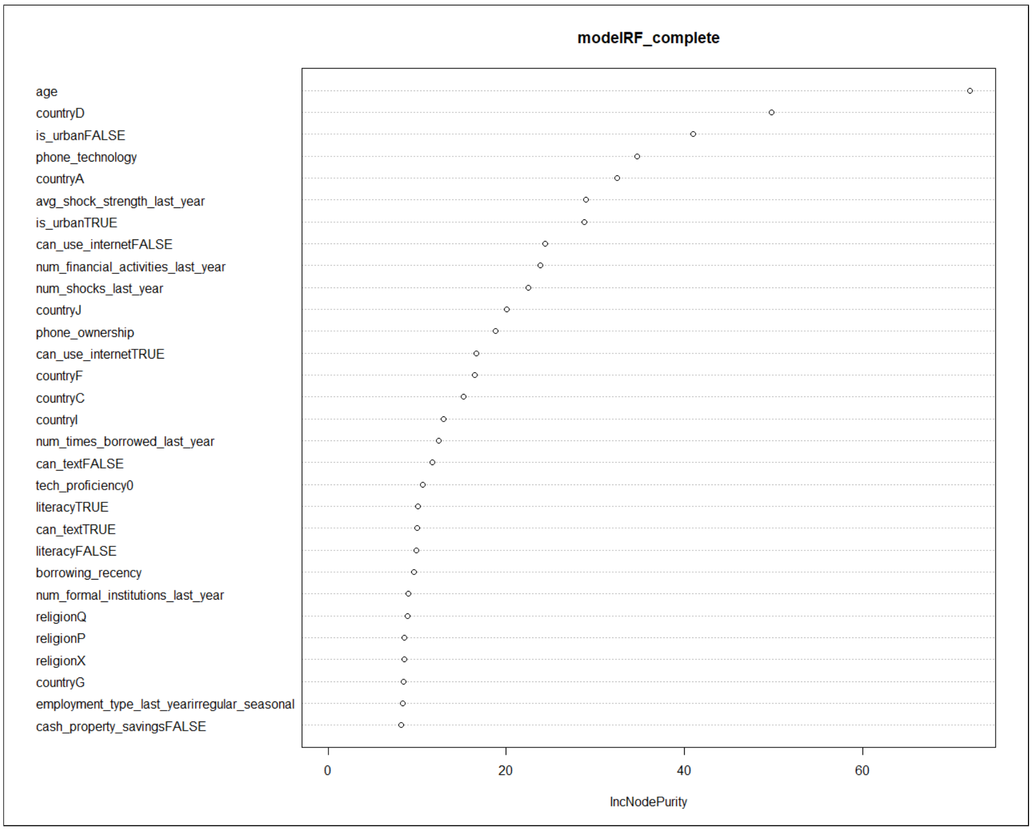

Figure 6.7: Mean decrease in node impurity from original sized training data

REFERENCES

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. (1984), *Classification and regression trees Regression trees*, number June.

Janitza, S. & Hornung, R. (2018), 'On the overestimation of random forest's out-of-bag error', *PLOS ONE* **13**(8), e0201904.
URL: *http://dx.plos.org/10.1371/journal.pone.0201904*

Liaw, A. & Wiener, M. (2002), 'Classification and Regression by Random Forest', *R news* **2**, 18–22.