International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

---

# Random Forests for Poverty Classification

Ruben Thoplan[a]*

[a]*Department of Economics and Statistics, Faculty of Social Studies and Humanities, University of Mauritius, Réduit, Mauritius*

[a]*Email: r.thoplan@uom.ac.mu*

**Abstract**

This paper applies a relatively novel method in data mining to address the issue of poverty classification in Mauritius. The random forests algorithm is applied to the census data in view of improving classification accuracy for poverty status. The analysis shows that the numbers of hours worked, age, education and sex are the most important variables in the classification of the poverty status of an individual. In addition, a clear poverty-gender gap is identified as women have higher chances to be classified as poor as compared to men.

***Keywords:*** data mining; classification; poverty; random forests; poverty-gender gap

## 1. Introduction

According to [1], the international poverty line, which is $1.25 or less a day, represents the point at which the level of income and consumption become insufficient to support a good quality of life. The gap between rich and poor in some developing countries continues to widen. Although extreme poverty has been declining over the years as a percentage of the global population, in 2010 the number of people affected by extreme poverty was more than 1.2 billion. The World Bank has in fact set up a goal to terminate extreme poverty by 2030, [2]. Poverty can have a severe impact on family functioning in different dimensions like communication, behavior control and family role [3].

Poverty is a multidimensional problem which does not have a standard definition for all countries of the world. According to [4], natural and conditional individual characteristics, social causes and inherent properties of social system are the explanations to poverty.

---

* Corresponding author.

E-mail address: r.thoplan@uom.ac.mu.

Poverty is partly linked to health, economic and education problems [5]. On an individual basis, poverty can have a serious influence on the brain development of young children [6]. Unfortunately, not all families have the opportunity of leaving the poverty trap easily. This poverty trap can be sustained for a poor family because of the lack of finance to invest in the education of its children [7].

On a country level basis, there is no uniformity in the definition of poverty as the characteristics of countries are not similar across the world. The United States of America, for example, defines poverty by considering a set of threshold values which depends on the family size and composition [8]. This implies that the poverty threshold is not uniform in the country itself and varies according to the family characteristic. For the case of Mauritius, there is no poverty line as such, but Statistics Mauritius (SM) has a measure of relative poverty instead. At SM, the relative poverty definition is based on the median household incomes.

In this paper, the random forest classification method is used to categorize people below and above the relative poverty line. The year 2000 has been considered because the new census data for Mauritius does not record the income variable. Therefore, based on the threshold set for the poverty line in 2001/2002 from the Household Budget Survey, we consider a person whose income is below Rs 2800 to be poor in the year 2000 [9]. This paper focuses mainly on the strength of the random forest method in distinguishing the poor from the non-poor. The next section discusses the random forest method and its suitability in classifying the poor and the non-poor. Section 3 analyses the year 2000 census data by using the random forest. Section 4 attempts some discussions and recommendations.

## 2. The Method

Random forests are an ensemble of unpruned decision trees where a number of trees are trained by using bootstrap samples and the class with the majority vote is obtained. The random forests algorithm is a powerful classification tool and has many benefits. For instance, it runs fast and is considered to have relatively high accuracy compared to other classification algorithm [10]. Leo Breiman is the first to formally introduce the random forests after the bagging method which is a combination of models in view of increasing classification accuracy. It is important to note that random forests do not over-fit, because for large number of trees, the generalization error converges to a limiting value under the strong law of large number [11]. This implies that even if a very large number of trees in the forest are built, the generalization error will still converge towards a limiting value implying that over-fitting cannot occur for large number of trees.

The random forests algorithm is outlined as follows:

- First, a random sample of observations is taken and subsequent bootstrap samples for other trees are taken.
- A subset of m variables much less than the total number of variables in the dataset is randomly selected and using the Gini score, the best split is determined.
- The out-of-bag (OOB) prediction is obtained through a majority vote across trees whose observation was not included in the bootstrap sample.

One of the features of the random forests algorithm is that it can provide a ranking of variable importance. This is somewhat analogous to sensitivity analysis. To evaluate the importance of a variable, [11, 15] propose to evaluate, for all trees in the forest, the average of an impurity decrease measure for all nodes where the variable is concerned. The variable with the largest decrease in impurity will be considered as the most important variable. This can be achieved through the Mean Decrease Gini (MDG) or the Mean Decrease Accuracy (MDA). In this paper, we focus mainly on the MDG as a measure of variable importance.

Using the notations from [12], any mean decrease impurity measure can be mathematically represented as follows:

$$\mathrm{Im}\,p(X_m) = \frac{1}{N_T}\sum_{T}\sum_{t\in T:\upsilon(s_t)=X_m} p(t)\Delta i(s_t, t) \qquad (1)$$

From equation (1), $X_m$ represents the $m^{th}$ variable, $N_T$ is the number of trees in the forest, $\upsilon(s_t)$ is the variable at split $s_t$, $p(t)$ is the proportion of records at node t out of the total number of records in the data and

$$\Delta i(s, t) = i(t) - p_L\, i(t_L) - p_R\, i(t_R) \qquad (2)$$

$p_L$ represents the number of records in the left child node of t out of the total number of records at node t. For this study, we shall consider the impurity measure i(t) as the Gini index. The Gini index, $i(t)$ is defined as follows for a node t:

$$i(t) = 1 - \sum_{j} p(j\,|\,t)^2 \qquad (3)$$

where j= 1, 2 for our case representing poverty class.

Before applying the random forests algorithm to the Census data of Mauritius, some data preprocessing is required like for any data mining project. The Census data is rich in terms of variables and number of cases. In the first instance, we remove all cases where individuals do not derive any income. Secondly, variables that are of no interest to the study are eliminated from the dataset to form a smaller dataset. After these two steps have been considered, some cleaning of this new dataset is required. The cleaning involves mainly dealing with missing values, outliers and inconsistent values.

Variables like “age at marriage”, “employment status” and “length of service” have been removed from the new dataset due to a very high proportion of missing values. Therefore the variables considered for analysis are described in table 1. However, some of the variables mentioned in table 1 still need some cleaning because of a few missing values, outliers and inconsistent data. Since the missing values are at random and occur in very few instances, the cases are completely deleted instead of going through imputation. For variables like age, income and hours worked, any outlier in the data is checked and removed if necessary. It is essential to highlight that the decision tree is not affected by outliers.

Another procedure is to check the consistency of the data by applying the filtering function to compare records among different variables. For example, it is not expected to have an individual whose age is below 16 to be civilly married. If such cases are found, these cases are scrutinized and imputation is carried out if enough information is available. Otherwise, the record is deleted. After the necessary data pre-processing is applied, an indicator variable named N_Poverty is derived from income as follows:

$$\text{N\_Poverty}=\begin{cases}\text{Poor}, & \text{if income} < 2800\\ \text{Not Poor}, & \text{if income} \geq 2800\end{cases}$$

In the next section, the results from the random forests on the census data are presented. It is attempted to determine the importance of the different variables mentioned in table 1in poverty classification for the case of Mauritius. The out-of-bag (OOB) errors are also produced for each class of poverty. The R Statistical Language is used for the purpose of analysis in this paper.

**Table 1: Variables Used In Analysis**

| Variable Names | Levels | | | | | Type |
|---|---|---|---|---|---|---|
| Sex | 1: M | | | 2: F | | Nominal |
| Age | | | | | | Numerical |
| Marital Status | 0: W | 1: D | 2: SEP | 3: MRC | 4: MR | Nominal |
| | 5: MC | 6 : C | 7: S | 8: UP | 9: OTHER and NOT STATED | |
| Religion | 0: No religion | | 1: Buddhist/Chinese | | 2: Christian | Nominal |
| | 3: Hindu | | 4: Muslim | | 5: Other | |
| Education | 0: < Prim | | 1: Prim | | 2: Low Sec | Nominal |
| | 3: Upp Sec | | 4: Post-Sec Non Tert | | 5: Short-cycle Tert | |
| | 6: Bachelor/Equi | | 7: Master/Equi | | 8: Doc/Equi | |
| Hours Worked | | | | | | Numerical |
| Income | | | | | | Numerical |

### 3. Empirical Results

In this section, the results from the application of the random forests algorithm are presented. First, 500 trees in the forest are run and the OOB errors are determined as a whole and also for each class. As from approximately the 25$^{th}$ tree, the errors converge to a limit as a whole and also for each class. For a small number of trees in the forest, the out-of-bag errors are relatively large but decrease as more trees are added to the forest. This indicates the power of the random forests to improve on classification accuracy compared to a single tree, for example. Besides, the convergence of the error rates towards a limit confirms the theory of [11], which states that the random forests do not over-fit. The OOB errors are presented in figure 1.

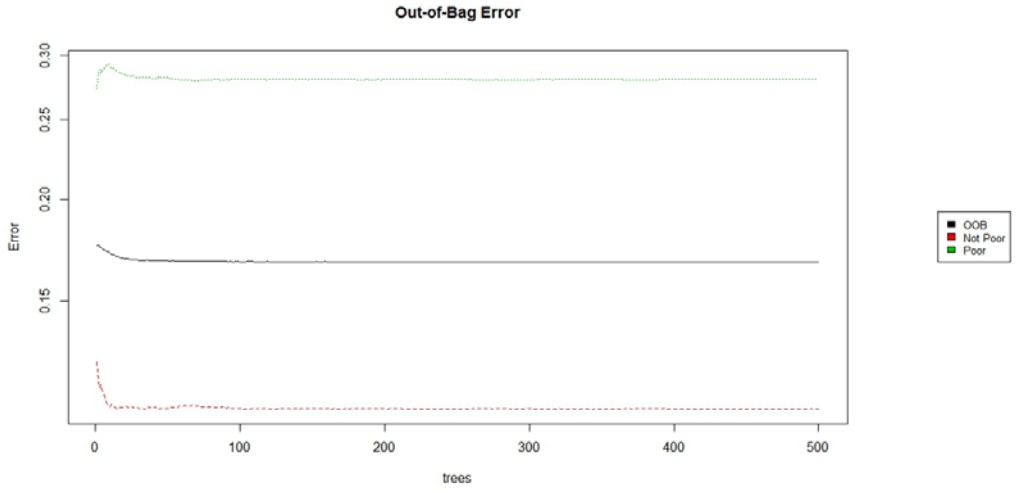


**Figure 1: Out-of-Bag Error Estimates for 500 Trees in The Forest.**

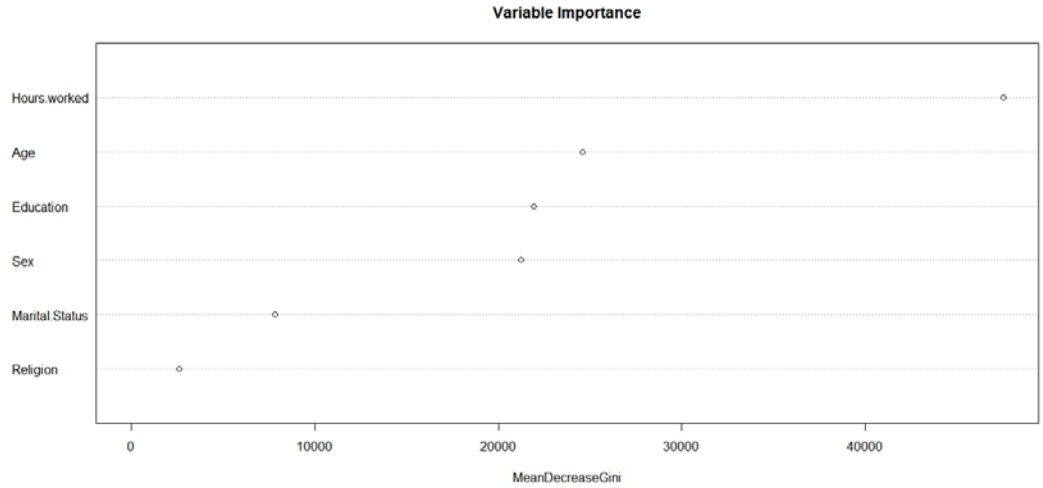


**Figure 2: Variable Importance by Considering Mean Decrease in Gini**

Figure 2 depicts the variable importance by measuring the decrease in mean Gini. A higher mean decrease in Gini will imply a higher importance. We observe that the number of hours worked in a week is a key classifier for poverty. This is to be expected, because since poverty is defined using income, it is predictable that a person who works very little hours in a month would be on average receiving a very small salary. In fact, [13] argues that apart from income, the time allocated to work is a vital determinant of the poverty status of an individual. This study indeed confirms his statement as the most important variable to classify poverty is number of hours worked per week. Three variables in descending order of importance are age, education and sex. But it should be noted that these three variables' importance measure are very close. However, marital status and religion are the two least important variables in poverty classification.

Since it is not statistically sensible to pick out a specific tree from a random forest and plot same, it becomes relatively difficult to understand the internal structure of the tree. But, fortunately some rules can be extracted

from the random forest to obtain some interpretations of the interdependencies among the variables in the forest. The rules are presented in table 2. They have been sorted in descending order of support multiplied by confidence. From the results, we can observe that there is still a gender gap regarding poverty as males tend to be classified as not generally poor whereas females have strong chances of being classified as poor. It clearly seems that the concept of feminization of poverty in the Mauritian context is still present. However, more investigation is required to understand the dynamics of this concept over time. Besides, people who work less one day or less per week have higher chance of being classified as poor.

**Table 2: Rules for Random Forests**

| Support*Confidence | Condition | Prediction |
|---|---|---|
| 0.132204 | Sex = {Male} | Not Poor |
| 0.121542 | Sex = {Female} | Poor |
| 0.105203 | Hours.worked<=24.5 | Poor |
| 0.104834 | Hours.worked>24.5 | Not Poor |
| 0.063758 | Education = {Less than Primary ,Primary} | Poor |
| 0.05782 | Education = {Low Sec,Upp Sec,Post-Sec Non Tert,Short-cycle Tert,Bachelor/Equi,Master/Equi,Doc/Equi} | Not Poor |
| 0.044482 | Age>89.5 | Not Poor |
| 0.037089 | Age>59.5 | Poor |
| 0.035758 | Hours.worked<=25.5 | Poor |
| 0.034848 | Education = {Upp Sec,Post-Sec Non Tert,Short-cycle Tert,Bachelor/Equi,Master/Equi,Doc/Equi} | Not Poor |
| 0.034488 | Sex = {Female} & Hours.worked<=24.5 | Poor |
| 0.033084 | Sex = {Female} & Education = {Less than Primary ,Primary} | Poor |
| 0.030888 | Education = {Less than Primary } | Poor |
| 0.026825 | Religion ={No religion,Buddhist/Chinese,Other} | Not Poor |
| 0.025 | Sex = {Male} & Hours.worked>24.5 | Not Poor |
| 0.024489 | Age<=19.5 | Poor |
| 0.023544 | Age<=59.5 | Not Poor |
| 0.022218 | Education = {Less than Primary ,Primary} & Hours.worked<=24.5 | Poor |
| 0.019845 | Education = {Primary} | Poor |

## 4. Conclusions

This paper applied random forests to study the characteristics of the poor as defined by the SM. The main classifier of the status of poverty for an individual is the number of hours worked per week. It was also observed that there is still a poverty-gender gap in the case of Mauritius in year 2000 as a female has a higher chance of being classified as poor as compared to a man who has a higher chance of being classified as not poor. We also observed that the random forest is an accurate data mining classification tool for the poverty status as the out-of-bag error is low overall when more trees are added in the forest.

By identifying the key variables in the poverty status classification, it is hoped that targeted policy and decisions are taken to alleviate poverty in Mauritius. This can be achieved mainly by giving more opportunities to people who are working one day or less in a week. More investigation could be done in view of improving the class specific accuracy as the accuracy of the poor class was lower as compared to the non-poor.

## Acknowledgements

The author would like to show its appreciation towards the Statistics Mauritius for providing the anonymised Census data which has been invaluable for the successful completion of this paper.

## References

[1] World_Bank, "The World Bank Working for a World Free of Poverty," 2014. [Online]. Available: http://www.worldbank.org/en/topic/poverty/overview. [Accessed 16 July 2014].

[2] P. Olinto, K. Beegle, C. Sobrado and H. Uematsu, "The State of the Poor: Where Are The Poor, Where Is Extreme Poverty Harder to End, and What Is the Current Profile of the World's Poor?," POVERTY REDUCTION AND ECONOMIC MANAGEMENT (PREM) NETWORK, 2013.

[3] A. Banovcinova, J. Levicka and M. Veres, "The Impact of Poverty on the Family System Functioning," *Procedia - Social and Behavioral Sciences,* vol. 132, p. 148–153, 2014.

[4] E. O. Wright, "The Class Analysis of Poverty," *International Journal of Health Services* , vol. 25, no. 1, pp. 85 - 100 , 1995.

[5] F. N. Stapleford, "Causes of Poverty," *The Public Health Journal,* vol. 10, no. 4, pp. 157-161, 1919.

[6] S. J. Lipina and J. A. Colombo, Poverty and brain development during childhood: An approach from cognitive psychology and neuroscience., Washington, DC, US: American Psychological Association, 2009.

[7] V. Barham, R. Boadway, M. Marchand and P. Pestieau, "Education and the poverty trap," *European Economic Review,* vol. 39, no. 7, p. 1257–1275, 1995.

[8] C. Hokayem and M. L. Heggeness, "Living in Near Poverty in the United States:1966 - 2012," U.S. Census Bureau, 2014.

[9] H. Bundhoo, "Poverty Analysis 2001/02," Central Statistics Office, Ministry of Finance and Economic Development, Port Louis, 2006.

[10] R. Nisbet, J. Elder and G. Miner, Handbook of Statistical Analysis and Data Mining Applications, Academic Press, 2009.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[12] G. Louppe, L. Wehenkel, A. Sutera and P. Geurts, "Understanding variable importances in forests of randomized trees," Electronic Proceedings, 2013.

[13] C. Vickery, "The Time-Poor: A New Look at Poverty," *The Journal of Human Resources*, vol. 12, no. 1, pp. 22-48, 1977.

[14] L. Breiman, "Out-of-Bag Estimation," *Technical report, Statistics Department, University of California Berkeley, Berkeley CA 94708*, pp. 1-13, 1996.

[15] L. Breiman, "Manual on Setting Up, Using, And Understanding Random Forests V3.1," Technical Report, 2002.

[16] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217-222, 2005.

[17] J. Maindonald and W. J. Braun, Data Analysis and Graphics Using R: An Example-Based Approach, 3 ed., New York: Cambridge University Press, 2010.

[18] M. Kuhn, "Variable Importance Using the Caret Package," 19 March 2012. [Online]. Available: http://www.icesi.edu.co/CRAN/web/packages/caret/vignettes/caretVarImp.pdf. [Accessed 21 July 2014].