Shulei Yang

PTM Study Analysis

Statistical Consulting 7995 Dr. Karen Kafadar

University of Virginia

Apr. 20, 2020

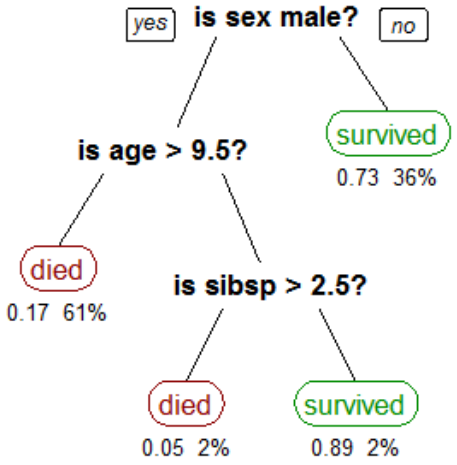
1. Introduction

Rates of malnutrition in Bangladesh are among the highest in the world. More than 56% of preschool-age children are underweighted.[1] Our study intends to determine the effect of PTM202 on environmental enteric dysfunction for children in Bangladesh. The ultimate goal of my analysis is to find a child with what demographic characteristics are more likely to be benefited from the PTM202 nutritional intervention. This report consists of six parts. In 'Method Overview', I will introduce all the technics I used in my analysis. Then in 'Analysis Procedure', I will discuss the exact procedure of data analysis of my data analysis. After that in 'Results', I will present the results from my analysis and carefully show how I get to my conclusion. Finally, I will discuss some potential limitations and conclude my whole analysis.

2. Method Overview

a. Decision Tree

A decision tree is a flowchart-like structure in which each node represents a 'test' on an attribute (e.g whether a coin flip comes up heads or tails), each branch represents the outcome of the 'test'. Below is the famous Titanic Survivals example of a decision tree.



A decision tree is drawn upside down with its root ('is sex male?') at the top. Root is the variable that contributes the most to the final classification of outcomes. Then the next most important variable is 'age'. So, when we have this decision tree plot, we can easily find the variables which are important for us to make the classification decisions.

In our analysis, the final classification decision is whether a child has benefited from the PTM nutritional intervention or not. And with the decision tree method, we can find those variables which are most influential for the final result of whether a child was benefited. Then we can easily use these variables to identify a subset of children who are most likely to have positive changes from the PTM nutritional intervention. And finally, we can develop a targeted intervention on children within the subset.

[1] http://www.fao.org/ag/agn/nutrition/bgd_en.stm

b. Random Forest

One on the extension of the decision tree method is the random forest method. The overall idea of random forest is to aggregate a large number of decision trees and come up with a more robust decision. Random forest 'learns' individual trees with some random perturbation, and then 'averages' these trees. Random forest method can also tell us which variables are most influential for the final outcomes (what demographic characteristic of a child will make the child more likely to have a positive change from the PTM nutritional intervention).

In my analysis, I will use random forest to further confirm the result I gain from the decision tree method.

3. Analysis Procedure

a. How to define 'benefited'?

In order to perform the decision tree method and random forest method, we need to first find out which children are benefited from the intervention. But before everything, we need to first subset all the data sets since we only need the information for children who have received the intervention. I think the most representative indicators of whether a child was benefitted or not are the changes in this child's inflammatory biomarkers. Therefore, I created a new dataset to record the changes in each biomarker for each child. Changes are calculated by subtracting the value of biomarker at 'enrollment' from the value of biomarker at '1 month after intervention'. Since a smaller value of biomarker indicates a more positive intervention, therefore I defined a negative change in biomarker as 'benefited' and a positive change in biomarker as 'not benefited'.

However, determine whether a child has been benefited or not solely on the value of changes in biomarkers is not accurate enough. We need to test whether these changes are significant. I choose to use a t-test to test whether a change is significant. Therefore, children who have a significant decrease in biomarkers will be classified as 'benefited'.

Because we have 4 biomarkers, we need to perform the procedures I discussed above for each biomarker. And finally, we will have four new variables: 1. whether a child was benefited or not based on the change in MPO; 2. whether a child was benefited or not based on the change in Reg1b; 3. whether a child was benefited or not based on the change in sCD14; 4. whether a child was benefited or not based on the change in CRP.

Also, there are many other ways to define 'benefited'. For example, we can let only those children who have significant decreases in all four biomarkers classified as 'benefited'. But in my analysis, I will make analysis based on each biomarker separately.

b. How to perform decision tree method and random forest method?

We need to perform the decision tree method and the random forest method for all the 4 biomarkers. For example, when we decided to use 'MPO' as the indicator of 'benefited', we can create and a new variable to the 'Baseline' dataset. This new variable can take a value of 0 (when this child was not benefited based on the result of MPO) or 1 (when this child was benefited based on the result of MPO). Then we can use this variable as response and all the other variables (except 'sid' and 'trt') as predictors to preform decision tree method and random forest. I exclude 'sid' and 'trt' in the model because these two variables are not predictive for the outcomes. From the result of decision tree and random forest, we can know which variables are most influential for the decrease in MPO. Then we finished the analysis of MPO and we can start to perform the same analysis on the other three biomarkers.

4. Results

a. <u>Results based on MPO</u>

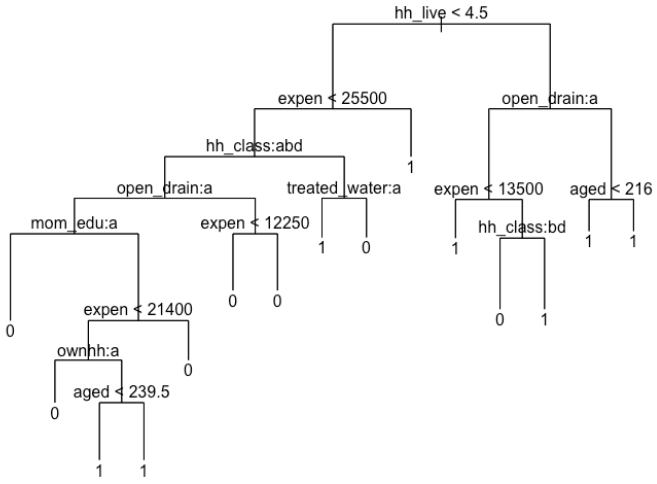


Figure 1: Decision Tree - MPO

Analysis: Decision tree result should read from the top, the top several variables indicate the variables which are most influential in decreasing MPO. And from a node, there will be two branches, the left branch represents 'yes' to the node condition and the right branch represents 'no' to the node condition. The terminal nodes (nodes with no branch) can have two values: 0 or 1. 0 presents the child was not benefited, and 1 represents the child is benefited. From this decision tree, we can see that 'hh_live < 4.5', 'expen > 25500' and 'open_drain = 1' are 'benefited' children's three most important characteristics.

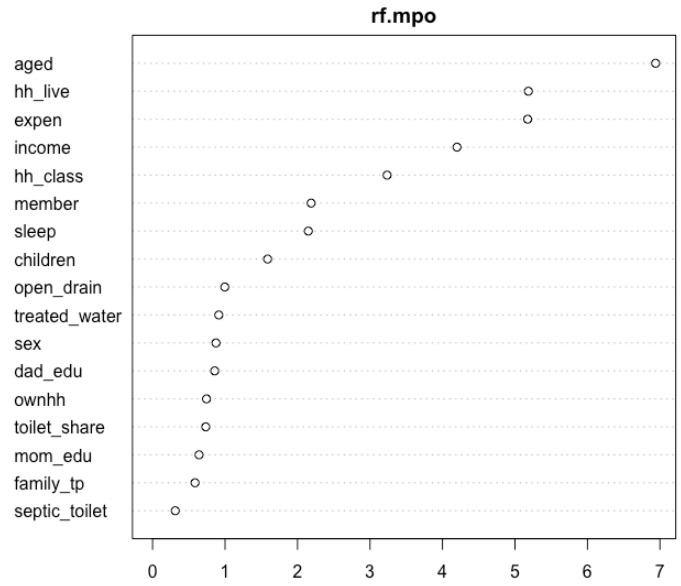


Figure 2: Random Forest Result – MPO

Analysis: The above plot indicates the importance of variables in a random forest model. The variables appear at the top of the plot with a large 'x-value' are more important. The result from a random forest should be more reliable than a single tree model. However, we can only know what variables are important, but we cannot tell how those variables are important. For instance, from this random forest result we know that 'hh_live' is important, but from the above decision tree, we can know a child with 'hh_live < 4.5' are more likely to be benefited from the intervention. But overall, from this random forest result, we know that 'age', 'expen' and 'hh_live' are most important.

Conclusion: 'hh_live', 'expen, 'open_drain, 'age'

b. <u>Results based on Reg1b</u>

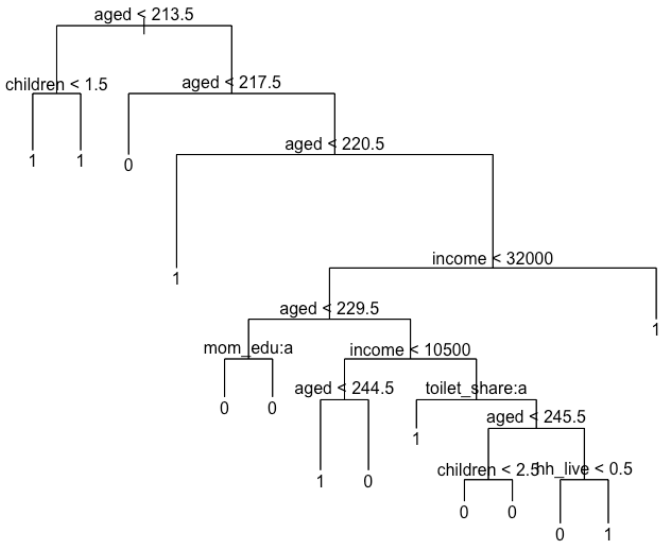


Figure 3: Decision Tree - Reg1b

Analysis: From this decision tree we know that children with 'aged < 213.5" and "children<1.5" are more likely to be benefited from the intervention.

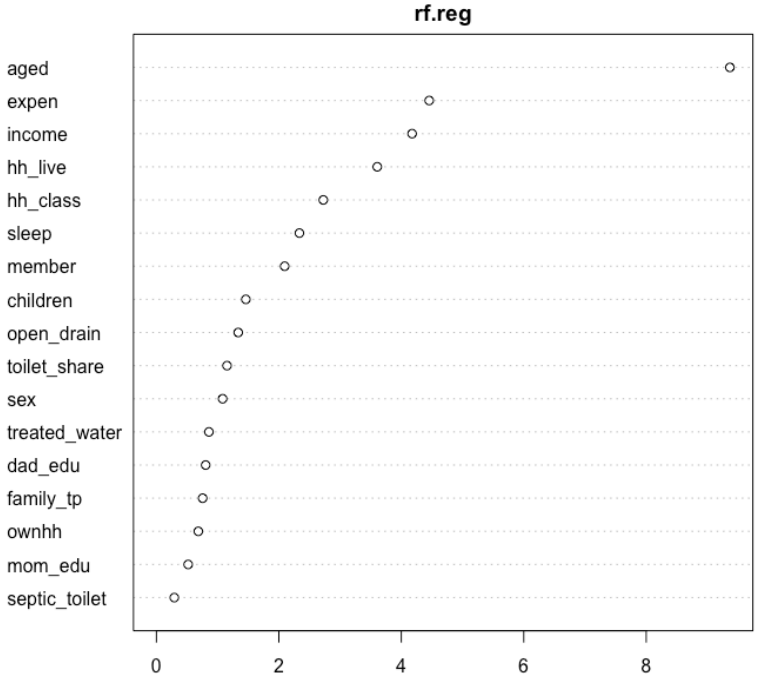


Figure 4: Random Forest Result – Reg1b

Analysis: The result from random forest agrees with decision tree, 'aged' is the most important variable for the decrease in Reg1b

Conclusion: 'aged, 'children'

c. <u>Result based on sCD14</u>

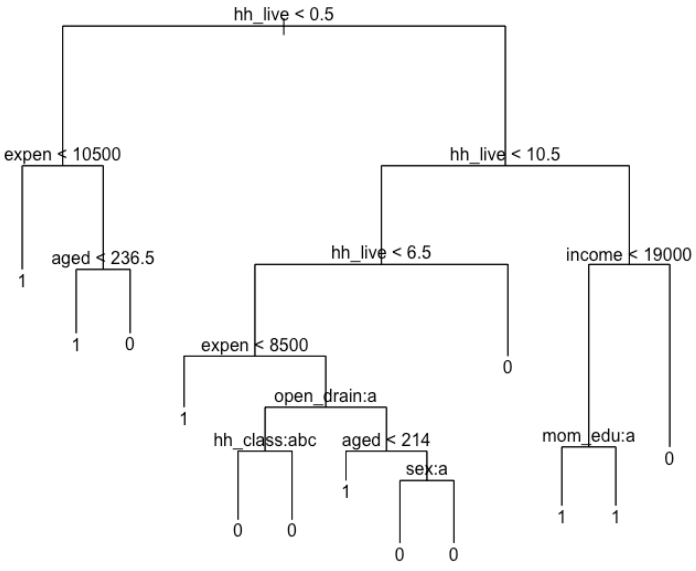


Figure 5: Decision Tree - sCD14

Analysis: From this decision tree we know that 'hh_live < 0.5', 'expen > 10500', and 'aged < 236.5' are the three most important characteristics of children who have been benefited.

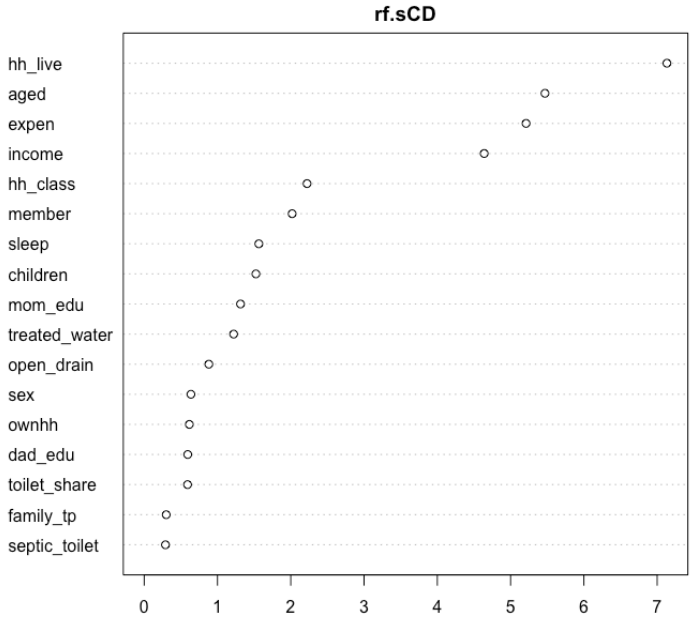


Figure 6: Random Forest Result – sCD14

Analysis: Random forest agrees with decision tree, 'hh_live', 'aged', 'expe' and 'income' are the four most important variables.

Conclusion: 'hh_live', 'aged' , 'expen', 'income'

d. <u>Result based on CRP</u>

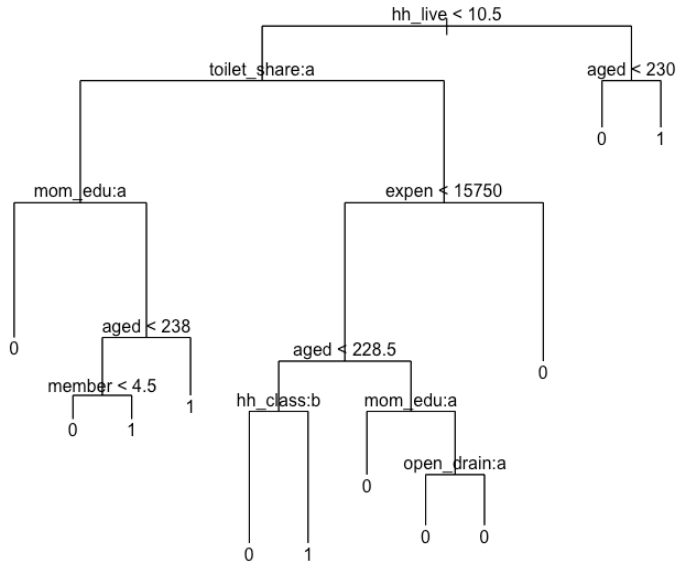


Figure 7: Decision Tree - CRP

Analysis: From this decision tree we know that 'hh_live < 10.5', 'toilet_share = 0', and 'aged < 230' are the three most important characteristics of children who have been benefited.

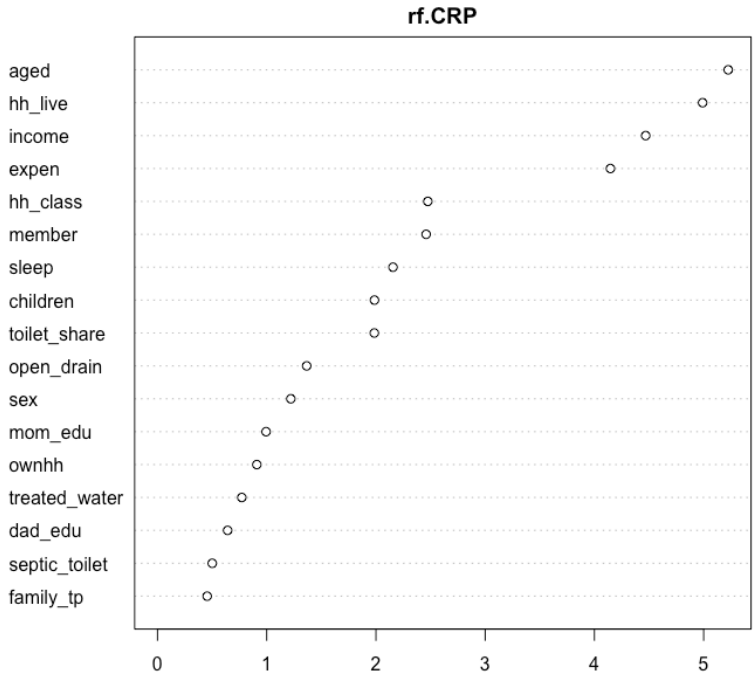


Figure 8: Random Forest Result – CRP

Analysis: From the result of random forest, 'aged', 'hh_live', 'income' and 'expen' are the most influential variables.

Conclusion: 'hh_live', 'toilet_share', 'income', 'expen'

5. Limitations

We should notice that there is a potential limitation of my results. Since there are only 100 children had the nutritional intervention, our sample size is only 100. When dealing with a small sample dataset, both decision tree and random forest could suffer from overfitting. This means that my results might be only statistically meaningful, but may not be practically meaningful. But I have used 5-CV to carefully tune the random forest models, and for most of the time the results from random forests agree with decision tree. Therefore, I believe my results should be useful in practice.

6. Conclusion

Combining the results above, we can have a broad picture of our subset for benefited children. In conclusion, a younger child, who has a family with higher income and higher expense, lives in a newer household and don't share a toilet with others is more likely to be benefited from the PTM202 nutritional intervention.