Using Random Forests, Income Rules, and Linear Regressions to Predict SSI Marginal Tax Rates *

Parker Rogers†

July 2016

Abstract

Using the Current Population Survey, we provide a machine learning method for computing the implicit marginal tax rates (MTRs) pertaining to the Supplemental Security Income Program. We use MTR prediction for the Earned Income Tax Credit program as an example to illustrate the efficacy of the machine learning method. Traditional MTR estimation can be complicated and difficult. Random Forests, a machine learning technique, can provide a simple and effective alternative to MTR calculation.

*Special thanks to Amy Xu at the American Enterprise Institute for her important contributions, and to Matt Jensen, Managing Director of the Open-Source Policy Center at the American Enterprise Institute. All Python code and documentation for the MTR estimation is available at https://github.com/parkerrogers/Benefits

†Brigham Young University, Department of Economics, 121B FOB, Provo, Utah 84602, parker.rogers2@gmail.com.

1 Introduction

Traditionally, implicit marginal tax rates (MTRs) associated with government transfer programs are estimated using programmed income rules or linear regressions. Programming income rules can be very effective, but can become complicated as certain government transfer programs contain numerous requirements for eligibility, income determination, and benefit determination. Even if one is able to program all of the income rules, the data used to calculate the benefit amounts could be lacking the necessary information pertaining to these rules.

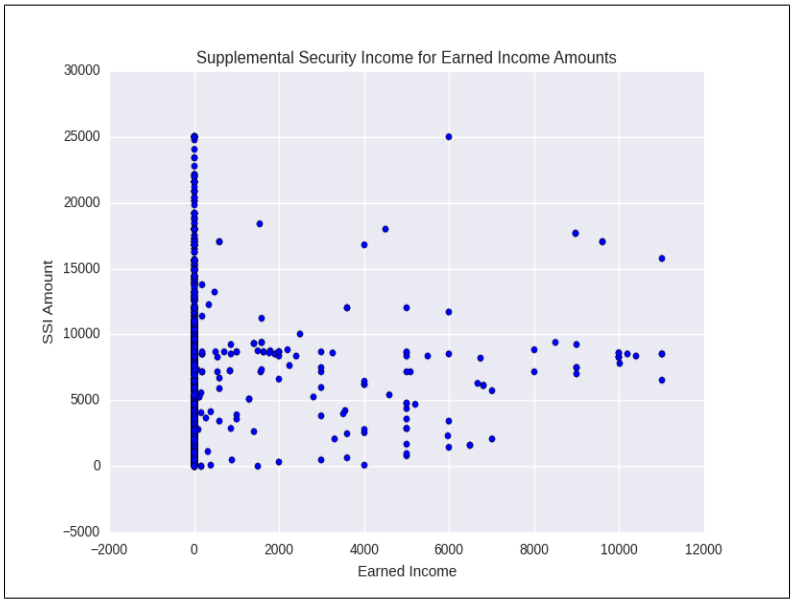
Linear regressions can also be used to define MTRs by defining the model as follows,

$$\text{benefit} = \beta_0 + \beta_1(\text{earned income}) + ...$$

and evaluating the first derivative of the model equation with respect to earned income, which yields an MTR of β_1. However this method can be problematic when the behavior of the program benefit, with respect to earned income, is nonlinear.

This nonlinear behavior is evident in the Social Security Income (SSI) program, as seen in the following plot of SSI amounts over earned income amounts for individuals in the 2014 Current Population Survey,

Figure 1: SSI Amounts for Earned Income Amounts From 2014 CPS



Notice that most individuals who receive SSI do not have positive earned income amounts. Even when earned income is positive, there is not a linear relationship between earned income and SSI benefit amounts.

Thus, we provide an alternative method called Random Forests. Random Forests is one of the many machine learning techniques that may allow for a more effective way to model these complex, nonlinear relationships. Varian (2014) When compared to the other two traditional methods, Random Forests can produce simple, accurate, and robust MTR results.

We first make a case for using Random Forests for Supplemental Security Income (SSI) implicit MTR prediction by using the Earned Income Tax Credit (EITC) program as an example. We provide the results for the EITC MTRs using all three methods of MTR prediction explained above for analysis.

Next, we use Random Forests, in conjunction with programmed income rules and regressions, to calculate the marginal tax rates (MTRs) of the Supplemental Security Income program (SSI).

We provide the results of these three methods for comparison.

2 Random Forests

A Random Forest is method that uses the weighted average of the predictions that come from a collection of subordinate models called prediction trees. In our particular application, the prediction trees used in our Random Forest model are specifically called regression trees.

The main difference between a linear regression, and a regression tree is that a linear regression is a global model, meaning there is a single predictive formula, being the estimated regression equation, that holds for the entire sample space. This can present problems when the variables have complicated nonlinear interactions. To make these complicated interactions more manageable, regression trees partition the sample space into smaller, more manageable regions. Then those partitions are partitioned recursively until the complications can be mitigated enough that a simple model can fit the data within each partition. In summary, regression trees recursively partition the global model until a simple model can be fit to each terminal partition.

Prediction trees, and in our case regression trees use the tree to represent the recursive partition process. A different simple model applies to each terminal node in the tree. Each internal node is split using a clarifying question about the features, and the branches are labeled with the answers to the questions contained at their respective nodes. The process of finding the best partitioning of the data, or the best questions to ask is usually done by maximizing the amount of information the each partition/question gives about the dependent variable.

Next, a simple model is fit to each particular terminal node. A common model fit to each terminal node in regression trees is a constant model, which is the sample mean of all of the dependent variable outcomes in that node.

An advantage of these prediction trees is that variables need not be all continuous or all categorical. One can have a mix of these variables and still obtain accurate predictions. Also, predictions are fast since the models at each terminal node are very simple. Moreover, the regression trees give a nonlinear response, thus they can be effective even when the true regression surface is not smooth. However, if the surface

is smooth, then the piecewise structure of the regression trees can approximate it arbitrarily closely, so long as there are a sufficient amount of terminal nodes. Shalizi (n.d.)

Lastly, the Random Forest repeats the regression tree model a number of times in order to accumulate a collection of regression tree outcomes. It then takes the weighted average of the outcomes of all the regression tree models to determine the final outcome. [1]

3 The Case For Predicting SSI MTRs Using Random Forests

In order to justify our use of Random Forests when predicting SSI MTRs, we provide an example where we compute the marginal tax rates of the Earned Income Tax Credit (EITC) program.

In order to predict EITC MTRs, we first programmed income rules. This process was tedious and assumptions about certain eligibility rules were made due to the lack of information within the data. In order to simplify the process, we also attempted to use linear regression techniques to estimate MTRs. However, due to the nonlinear behavior of the EITC benefit with respect to earned income, we were forced to make complicated piecewise regressions. These complications, is what lead us to consider Random Forests when estimating the SSI.

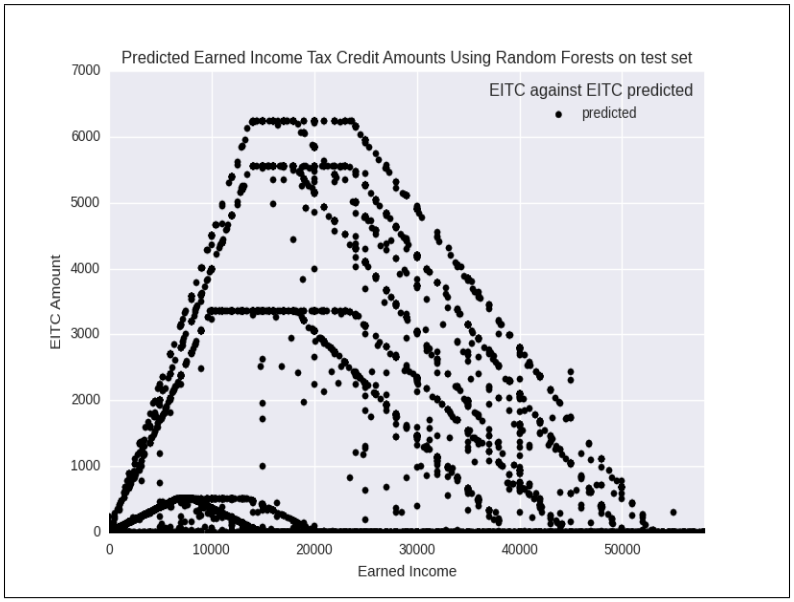
To train the Random Forest model we used the relevant variables used to determine EITC being earned income, adjusted gross income, investment income, number of children, filing status, and the EITC amounts of each individual. We used %80 of the data set, randomly selected, in this training. To measure the accuracy of the fitted model, we cross-validated the fit using the remaining %20 of the data set. The following is a plot of the EITC amounts that each respondent in the CPS received,

[1] For more information on Random Forests visit https://www.stat.berkeley.edu/ breiman/RandomForests/cc_home.htm

Figure 2: EITC Amounts From 2015 CPS

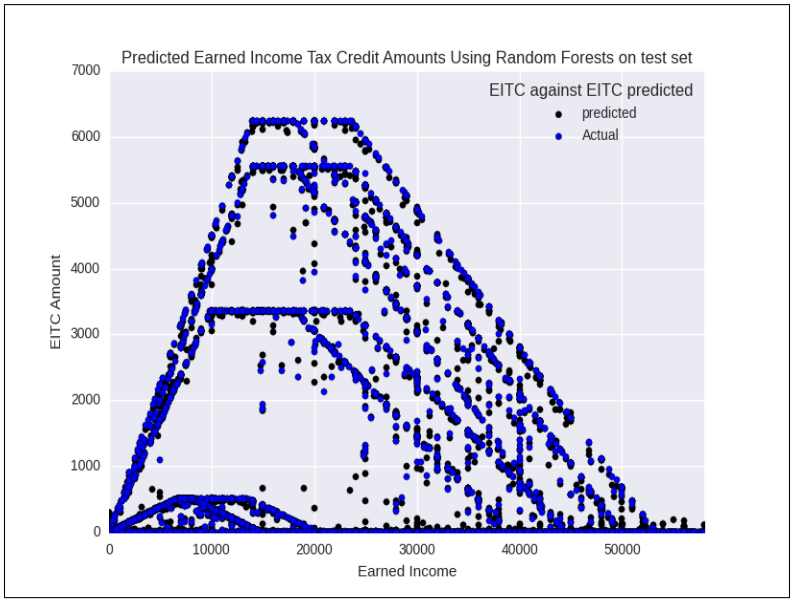
Notice the trapezoidal regions that the EITC amounts follow as a function of earned income. As mentioned earlier, this behavior is difficult to capture using linear econometric methods. However, using a non-linear estimation method, namely Random Forests, we were able to approximate the EITC amounts of the remaining %20 of respondents in the test set with astounding accuracy,

Figure 3: EITC Amounts From 2015 CPS



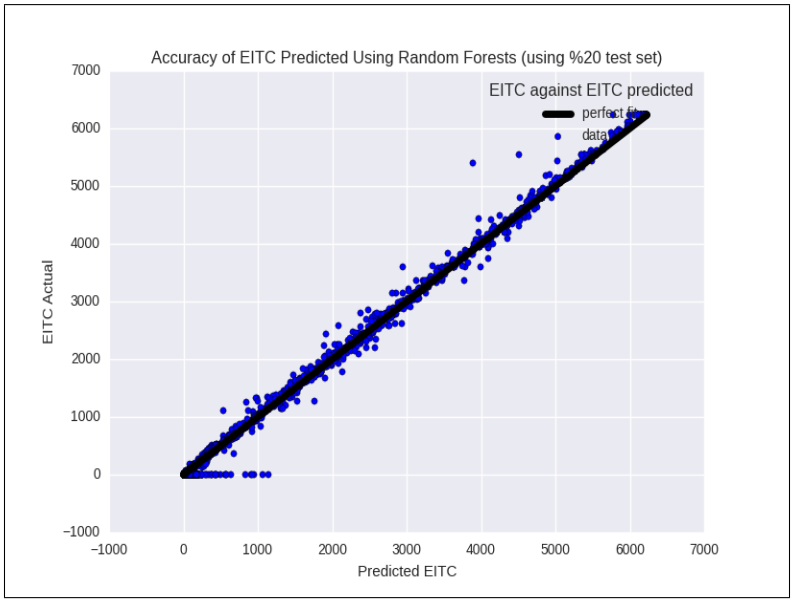
Notice how closely Random Forests predicts the EITC amounts of those in the test set. What is most exciting about this result is that we obtained these predictions without having to manually partition the sample space to account for nonlinearities. Here are the actual EITC amounts of the %20 test set plotted together with the predicted amounts as seen in Figure 3,

Figure 4: EITC Amounts From 2015 CPS



As seen in this figure, the predicted values for EITC are close to the actual values. In addition, the accuracy of this prediction can be seen using a perfect fit plot as follows,

Figure 5: EITC Amounts From 2015 CPS



The predicted values hug tightly the perfect fit line, meaning predicted EITC is very close to the actual EITC amounts. Since the Random Forest technique seems to perform well, this allows us to confidently evaluate the performance of Random Forests when predicting SSI MTRs.

Having seen the performance of Random Forests when predicting EITC amounts, we proceeded to predict EITC MTRs with this method. We trained the model just like before, but instead we used the entire data set. We then added $1,000 to the earned income of each CPS respondent. After the additional earned income, we predicted the new EITC amount. We then took the difference of the EITC amounts before and after the earned income adjustment. By dividing the difference by $1,000, we obtained MTRs for each individual. We used a $1,000 adjustment instead of a $1 adjustment since the model could not accurately capture such a small adjustment. To further illustrate the efficacy of Random Forest MTR prediction, we present the following table, which compares the MTR calculation results using programmed income rules (Program MTR), decile regressions (Reg MTR), Random Forests (RF MTR), and Tax-Calculator EITC MTR amounts (TC MTR). These results MTRs are the weighted average of all MTRs for individuals within the specified income ranges who have one child, and who are married,

Table 1: One Child

MTR for Income Decile				
Income Decile	Program MTR	Reg MTR	RF MTR	TC MTR
$44 \leq$ income < 4930	0.3215	0.3225	0.3395	0.3120
$4930 \leq$ income < 8600	0.3276	0.3257	0.3362	0.3093
$8600 \leq$ income < 10700	0.1061	0.2485	0.0318	0.0950
$10700 \leq$ income < 13000	-0.0079	0.1623	-0.0015	-0.0037
$13000 \leq$ income < 15958	-0.0065	-0.0329	-0.0094	-0.0075
$15958 \leq$ income < 19500	-0.0355	-0.0019	-0.1034	-0.0753
$19500 \leq$ income < 23372	-0.1094	-0.0909	-0.1519	-0.1303
$23372 \leq$ income < 27450	-0.1521	-0.1526	-0.1488	-0.1497
$27450 \leq$ income < 32400	-0.1543	-0.1469	-0.1564	-0.1503
$32400 \leq$ income < 43200	-0.1096	-0.1144	-0.1356	-0.1495

Notice that the Random Forest MTR predictions are similar to the MTRs obtained through programming income rules, and the Tax-Calculator, which are traditional MTR prediction methods. Also, note that Random Forest predictions tend to be just as accurate as the more contrived decile regressions. We refer to the decile regressions as contrived, because these regressions are executed on predetermined decile partitions of the data. These partitions are determined in effort to remove the obvious trapezoidal nonlinearities of the EITC so that the linear regression will produce more accurate results. However, Random Forests automatically partition the sample space without having to manually determine where nonlinearities occur. This is important since possible nonlinear behavior associated with the SSI and other transfer programs may not be as easily identifiable as the EITC is.

4 Predicting MTRs for Supplemental Security Income Program

As Random Forests can produce effective results for the EITC program, we now turn to using Random Forests to predict MTRs for the SSI program. Theoretically, implicit SSI MTRs for all individuals who have no earned income should be 0, since the first $65 of earned income each month is excluded from countable income. Next, for single individuals with earned income greater than zero, the possible values are -.5 and 0, since half of earned income that exceeds $65 is excluded from countable income. Lastly, the possible SSI amounts for married individuals are -.25 and 0, since the SSI benefit for married couples is split between spouses.

We provide MTR calculation results derived from programming income rules, and regressions in order to gage the efficacy of Random Forests. First, we train our Random Forest model to predict the variable "ssi_calc", which is the SSI amount reported by each individual in the CPS. After adding $1 to earned income, we recalculate the SSI amounts. We then take the difference between the SSI amount before and after the adjustment to obtain MTRs for each individual. The table below represents the weighted average of MTRs of all individuals who have no earned income, those who are single with positive income, and those who are married with positive income. The results are as follows,

Table 2: SSI MTR for Income Rules, Random Forests and Regressions

SSI MTR By Earned Income and Filing Status			
Income Decile	Program MTR	RF MTR	Reg MTR
Income = 0	0	0	0
Single & Income > 0	-0.2686	-0.4052	0.0715
Married & Income > 0	-0.1572	-0.1291	-0.0570

We see that the produced Random Forest MTR is close to the programmed income rules MTR result. One possible explanation as to why the random forest MTRs are

slightly different than the Program MTRs is that the Random Forest accounts for both federal and state SSI benefit changes that result from the earned income adjustment, whereas our Program MTR does not. We programmed income rules for only the federal SSI benefit, because the state supplement rules were very difficult to handle. This means that the state benefit is not factored into the MTRs of each individual in Program MTR.

Since the Random Forest and linear regression models were previously accounting for both state and federal SSI benefits, and the Program MTR was not, we provide another set of results where the Random Forest and regression models are only predicting federal SSI benefits. We only predict federal SSI benefits by training our Random Forest and regression models to predict our calculated SSI amount that we derived from programmed income rules. The MTRs associated with each method are as follows,

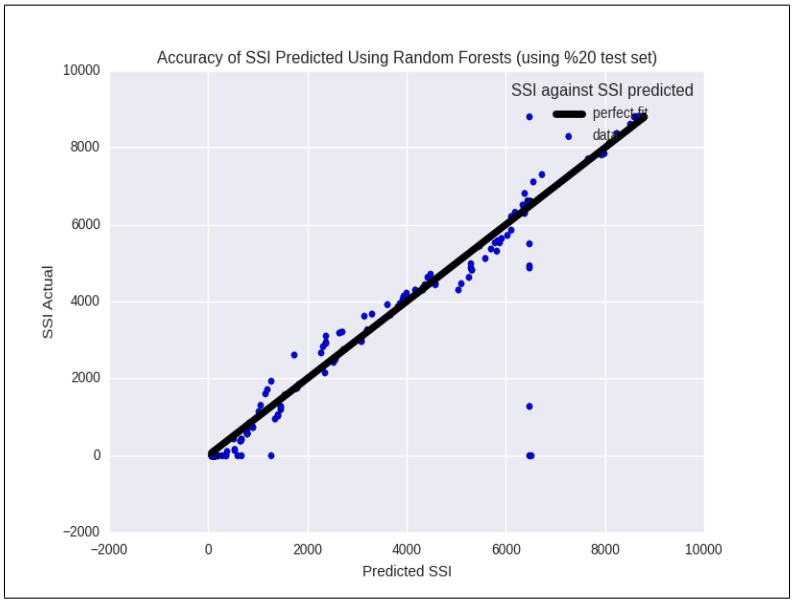
Table 3: SSI MTR for Income Rules, Random Forests and Regressions Using Calculated SSI

SSI MTR By Earned Income and Filing Status			
Income Decile	Program MTR	RF MTR	Reg MTR
Income = 0	0	0	0
Single & Income > 0	-0.2686	-0.2137	-0.1364
Married & Income > 0	-0.1572	-0.1364	-0.1849

Notice that the program MTRs and the Random Forest MTRs are much closer in this example, since both are predicting the same SSI value that was computed using income rules.

To show the accuracy of the Random Forest in predicting our calculated SSI values, we consider a plot of the perfect fit line, where we plot the actual SSI values against the predicted SSI values of the excluded %20 test set,

Figure 6: Predicted SSI Amounts Against Actual SSI Amounts



We see that the Random Forest model predicts the calculated SSI amounts effectively for the respondents in the test set. Specifically, the models explanatory value was $R^2 = 0.97$

A notable characteristic of Random Forest models is their inability to calculate MTRs effectively for all earned income adjustments. However, this effect is not surprising as we will explain below. The three different earned income adjustments we considered were $\$1,000$, $\$250$ and $\$1$. To calculate the MTRs from these adjustments, we take the difference between the predicted SSI benefit before and after the adjustment, and then divided that difference by the adjustment amount. With each variation in income adjustments, we obtain different MTR results. This fluctuation occurs for a couple of reasons.

First, since %90 of individuals who receive SSI do not have any earned income (See Figure 1), there is a lack of information on individuals with positive earned income. This leads to reduced predictive power. In addition, individuals receiving SSI typically have relatively small amounts of positive earned income.

On the contrary, individuals receiving the EITC benefit must have earned income greater than zero to qualify for the benefit. This creates a richness of information

on individuals receiving positive earned income. Thus, for EITC we have the reverse effect. Since all individuals receiving the EITC have positive earned income, a small earned income adjustment may produce too small of a change in benefits for the Random Forest to register since Random Forests work on finite piecewise partitions of the data. Thus, we can obtain accurate implicit MTRs for the EITC by using a larger earned income adjustment, like the $1,000 adjustment we used. Theoretically, this larger adjustment causes a more notable change in benefits, which causes the Random Forest to better detect this change as it is likely to push the new earned income amount into a different partition of regression tree.

Secondly, we must note the effect that large earned income adjustments have on an individual's earned income exclusions. For example, a large adjustment can potentially push an individual's earned income amount past the $65 exclusion threshold, causing even individuals with no earned income to have a MTR less than zero. We see this effect in the table below 4.

The combination of these two factors causes Random Forests to produce inaccurate MTR results for SSI when using larger income adjustments.

We provide the results from our three adjustments of $1,000, $250 and $1 from our first calculations of CPS SSI values and MTRs in order to show this effect,

Table 4: SSI MTR Using Different Adjustment Types For Random Forests

SSI MTR By Earned Income and Filing Status			
Income Decile	RF MTR($1,000)	RF MTR($250)	RF MTR($1)
Income = 0	-0.4009	-0.9005	0
Single & Income > 0	-0.0350	-0.0396	-0.4052
Married & Income > 0	-0.0146	-0.0673	-0.1291

We see that for the $1,000 and $250 adjustments, the Random Forest results are quite different from the $1 adjustment results. As predicted, these two larger adjustments even produced a negative MTR for individuals with no earned income, which is theoretically impossible. Also, as explained, the $1 adjustment produced

MTRs that are similar to both the programmed income rules and the theoretical values.

5 Conclusion

Even when datasets are relatively small, Random Forest models can be a simple and accurate addition to the traditional ways of computing implicit marginal tax rates associated with government transfer programs.

The EITC example shows that we were able to approximate MTRs with as much accuracy as the programmed income rules and linear regressions, without having to go through the complicated task of determining eligibility, income and benefit determination, or manually partitioning the data set in order to tame complex nonlinear relationships.

We later used Random Forests to predict SSI MTRs, and obtained results that were seemingly more accurate than the regression outcomes, and comparable to the programmed income rules outcome.

References

Shalizi, Cosma, "Lecture 10: Regression Trees," http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf.

Varian, Hal R., "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, March./Apr. 2014, *28* (2), 3–28.