Predicting Income with Census Income Data

Benny Chen, Tsz Chan

December 11, 2017

Introduction

Using the data from the Census Income Dataset, we wanted to see which tree-based method had the highest predicitve accuracy and which variables have the strongest predictive power in regards to whether an individual earns over $50000 in income. We compare three methods - classification trees, bagged, trees, and random forest.We begin with preprocessing and exploration of the data. Then, we fit the three models using the training data. With the most accurate model, we fit the test data.

Exploratory Data Analysis

First, we needed to remove missing values that were found in the workclass, native_country, and occupation variables. Removal was appropriate as observations with these missing values made up about 2400 of the 32561 observations. To be used with the models, we needed to aggregate some of the countries of the native_country variable. We chose to aggregate the countries with the 10 fewest occurances in the data into the "other" label, which totaled 137 observations. Based on our prior knowledge, we believed that age, education, and occupation would be important variables. Looking at the proportion of income levels, age and education appear the most different for the two groups. For many of the variables, initial plots were not too insightful due to the fact that the data was not very well balanced. We noticed that the training data had a much larger proportion of incomes greater than $50K, so the AUC metric would be useful.

Our last step for this phase was to remove some of the variables. We felt that "education" and "education_num" variables were very well correlated so we would only use one to fit our models.

Looking at boxplots and histograms, we also saw that "capital_gain" and "capital_loss" contained many outliers and were heavely skewed. This amount of variability just from the fact that these two variables had most occurances around 0 but also very large outlier values would severely bias their importance. Although decision trees and random forest are quite robust, it should be a point of attention.

The last variable of interest was "fnlwgt", which is a stastical metric for each invididual. We felt that this was not relevant to the individual's income level based on the variable's meaning so we decided to not use it in our analysis.

Looking at boxplots and barplots of various variables, it seems that there is a noticeable difference in age and education of the two income groups. The occupation boxplot showed large differences in income distributions of the respective occupations. These can be seen in the provided images.

Analysis

Classificaiton Tree:

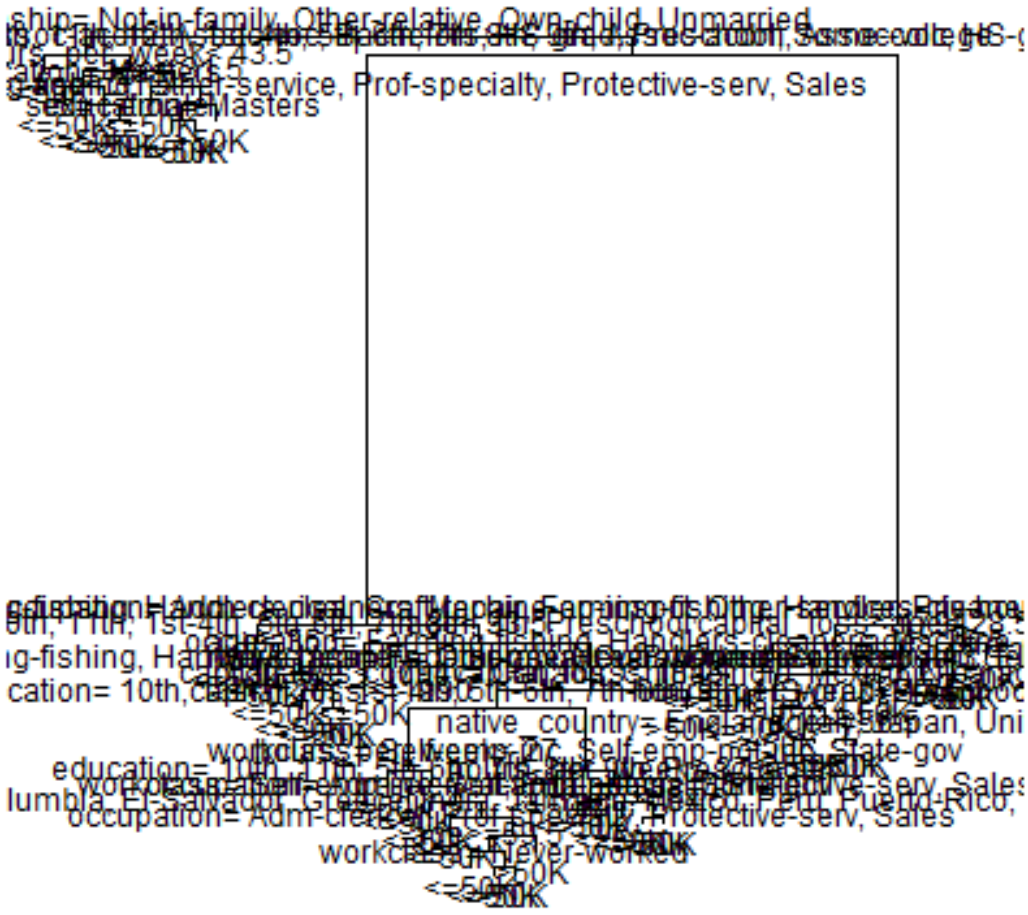
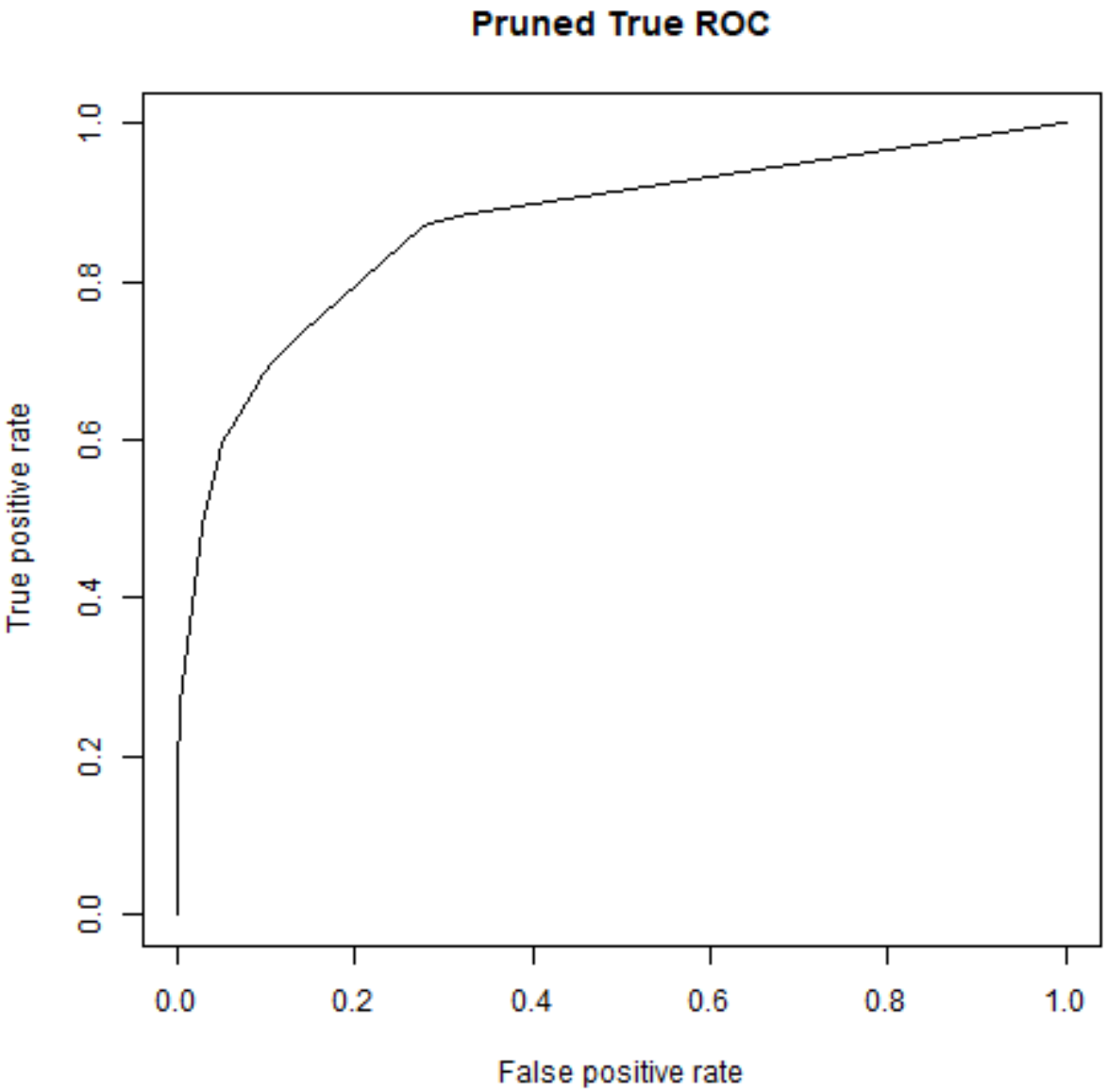

ship= Not-in-family, Other-relative, Own-child, Unmarried
hrs_per_week< 43.5
Masters
<=50K
Prof-specialty, Protective-serv, Sales
native_country=
State-gov
education=
occupation= Adm-clerical
workclass=
<=50K

Figure 1:



We decided to use the "education_num" variable over the "education" variable as it was cleaner for the tree model. 5-fold cross validation was used to tune the complexity parameter and the minimum split size. Decreasing CP and increasing minsplit would help to prevent overfitting by growing too large of a tree as CP decreases. We fit the training data with the combination with the lowest misclassification error, using rpart from the rpart library, then pruned the tree with the CP with the lowest error. This produced a training accuracy of 0.861. The AUC value was 0.872.

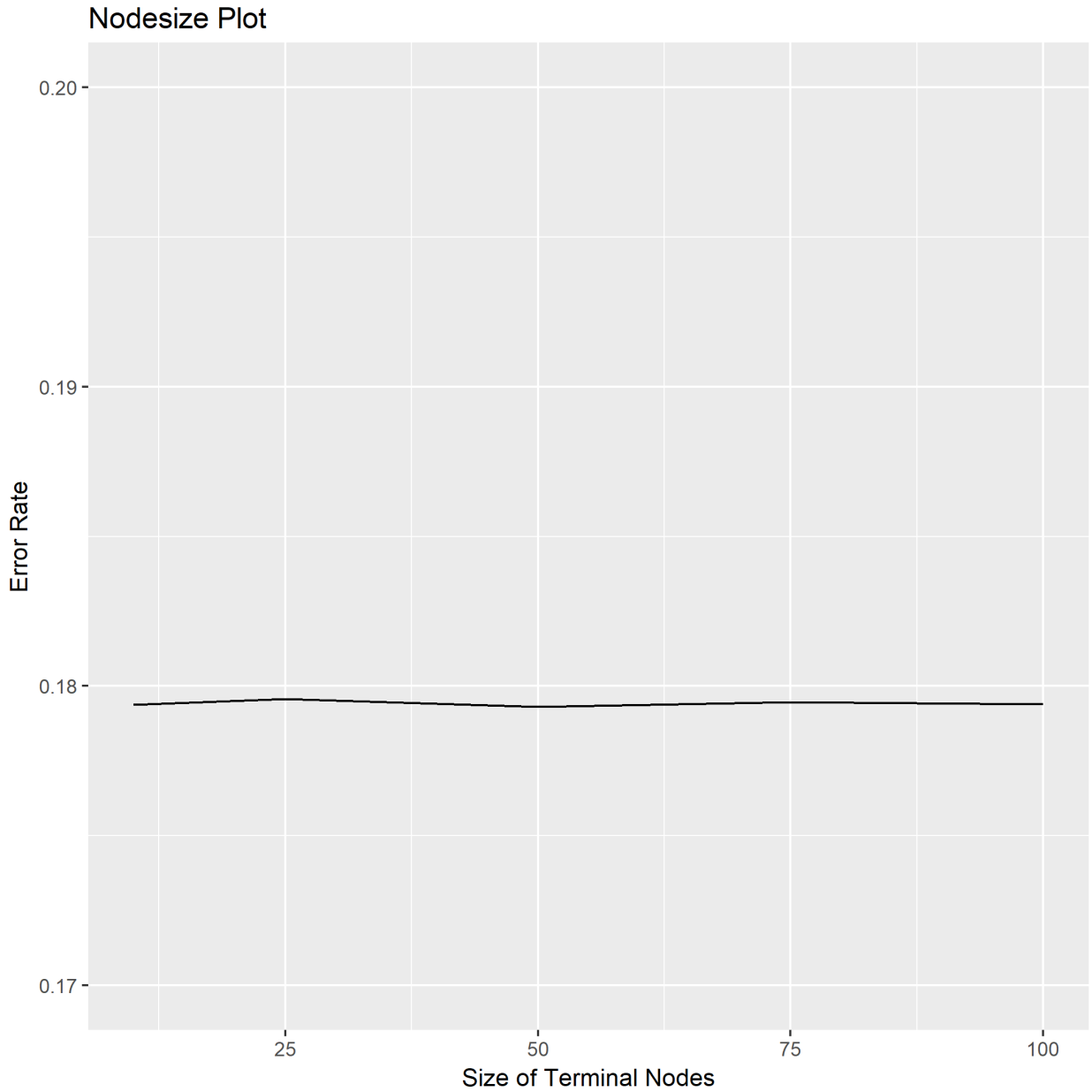
The following are the 5 most important varibales as measured by the average decrease in gini:

```
##         Variable Mean.Decrease.Gini
## 1   relationship          2302.474
## 2 marital_status          2241.866
## 3   capital_gain          1095.988
## 4      education          1016.117
## 5     occupation           939.125
## 6            sex           764.439
```

```
## 7            age            629.203
```

It is interesting that relationship and martial status had more discriminatory power based on this metric than variables such as age and education.

Bagged Tree:



A bagged tree is a special case of a random forest where all the predictors are considered at each split, which is reflected in the mtry parameter. We decided to tune the number of trees and minimum node size. It can be seen from the graph of errors and number of trees that 300 trees stabilizes the error rate. Using CV to calculate the errors for different minimum node sizes, the plot shows that around a minimum of 75 produces the optimal error. Choosing 75 allows for controlling the size of the trees.

Using randomForest from the randomForest library, we fit the training data, getting a training accuracy of 0.813. The AUC was 0.848.

The seven most important variables were:

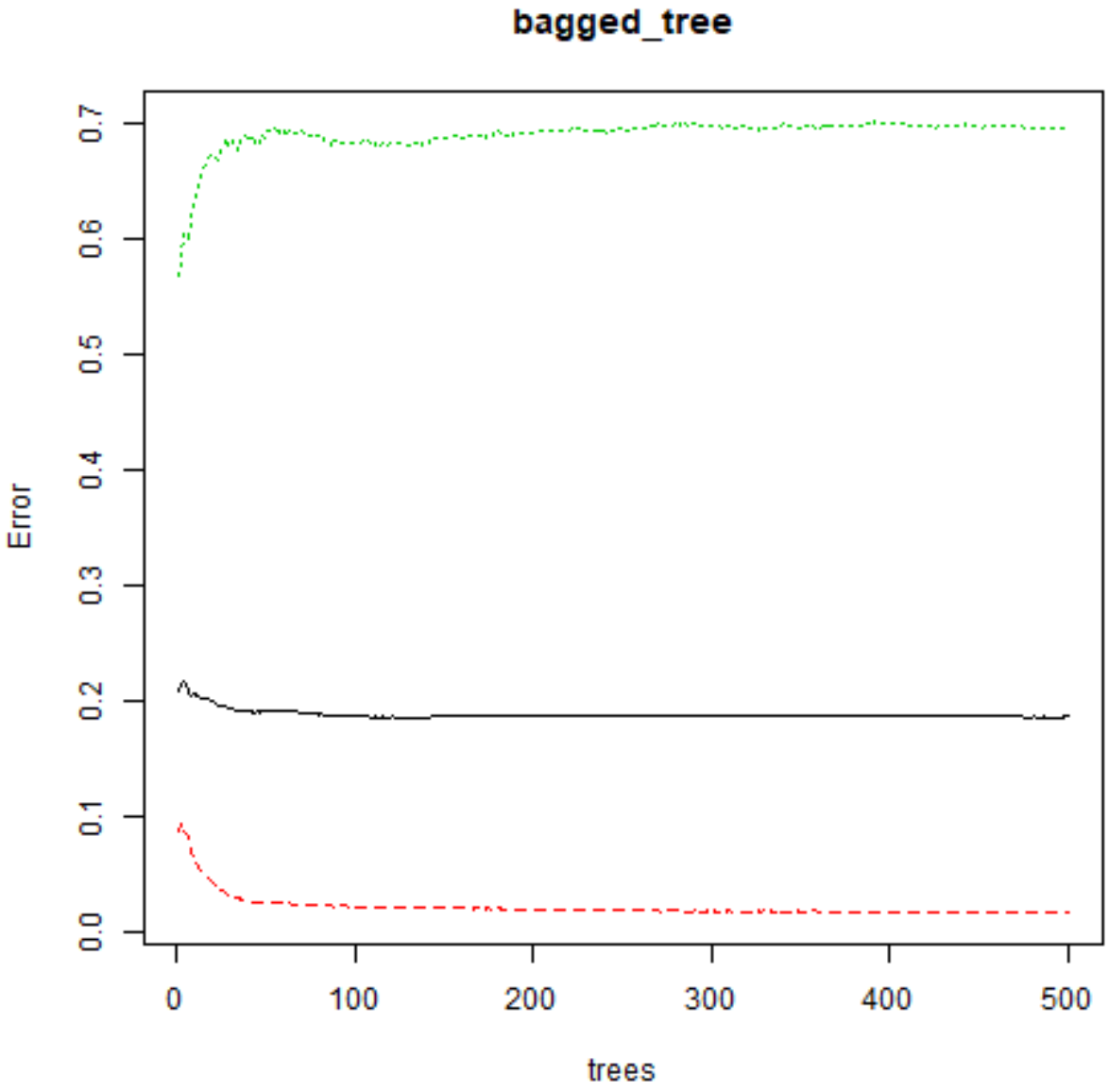


Figure 2: Number of Trees and the Error

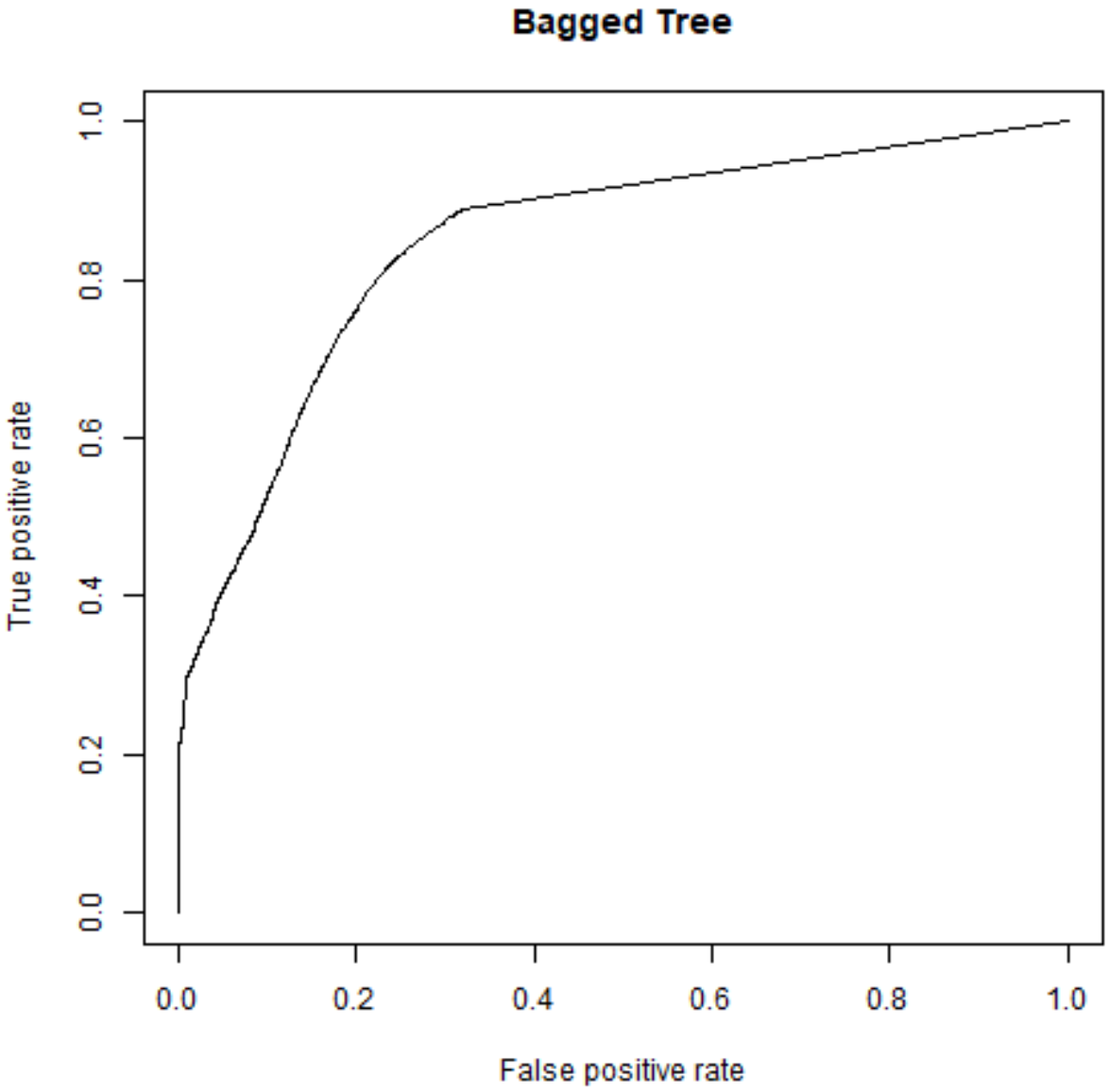

Bagged Tree
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3:

```
##        Variables Mean.Decrease.Gini
## 1   relationship           2302.546
## 2   capital_gain           1099.334
## 3  education_num           1096.084
## 4     occupation            514.019
## 5            age            415.525
## 6   capital_loss            310.240
## 7 hours_per_week            259.663
```

Random Forest:

```
##         Variable Mean.Decrase.Gini
## 1   relationship          1059.079
## 2   capital_gain           883.724
## 3 marital_status           842.542
## 4  education_num           651.851
## 5     occupation           525.925
## 6            age           250.805
## 7   capital_loss           151.114
```

A random forest of size 500 trees was fitted using randomForest from the randomForest library. 500 trees is sufficient to provide our results with robustness.

The number of paramaters to use at each split and the minimum leaf size was tuned using 5 fold CV. The minimum leaf size was chosen because the data is fairly unbalanced. The random forest model has a training accuracy of 0.849, The AUC value was 0.855, lower than that of the tree. It would be useful to look at the false positive rate, which was around 2%.

Mostly the same variables had the most importance as measured by the mean decrease in gini as for the classification tree, although the values of the mean decrease in gini are noticably lower. This could be due to the fact that the mtry parameter restricted the selection of variables for splitting.

Test Data

Based on the training accuracy the tree is the strongest model. After fitting the tree, the test accuracy was 0.784 and the AUC for the ROC curve was 0.868. Based on the plots, this is the highest AUC of the three, which would indicate that this classifier addresses the unbalanced data classes the best. The confusion matrix was (observed is the columns, predicted is the rows):

```
##       <=50K >50K
## <=50K 10759 1553
## >50K    604 2147
```

The true positive rate (senstivity) was:

```
## [1] 0.7804435
```

The true negative rate (specificity) was:

```
## [1] 0.8738629
```

Conclusion

The classification tree is the strongest classifier of the three methods. The accuracy was 0.856 which was lower than the training accuracy of 0.861. This does indicate some overfitting of the training model. All three models indicated that the most discriminative variables were relationship, education, age, and occupation, although there were some differences in ranking. This is in line with our expecations that age, job, and level

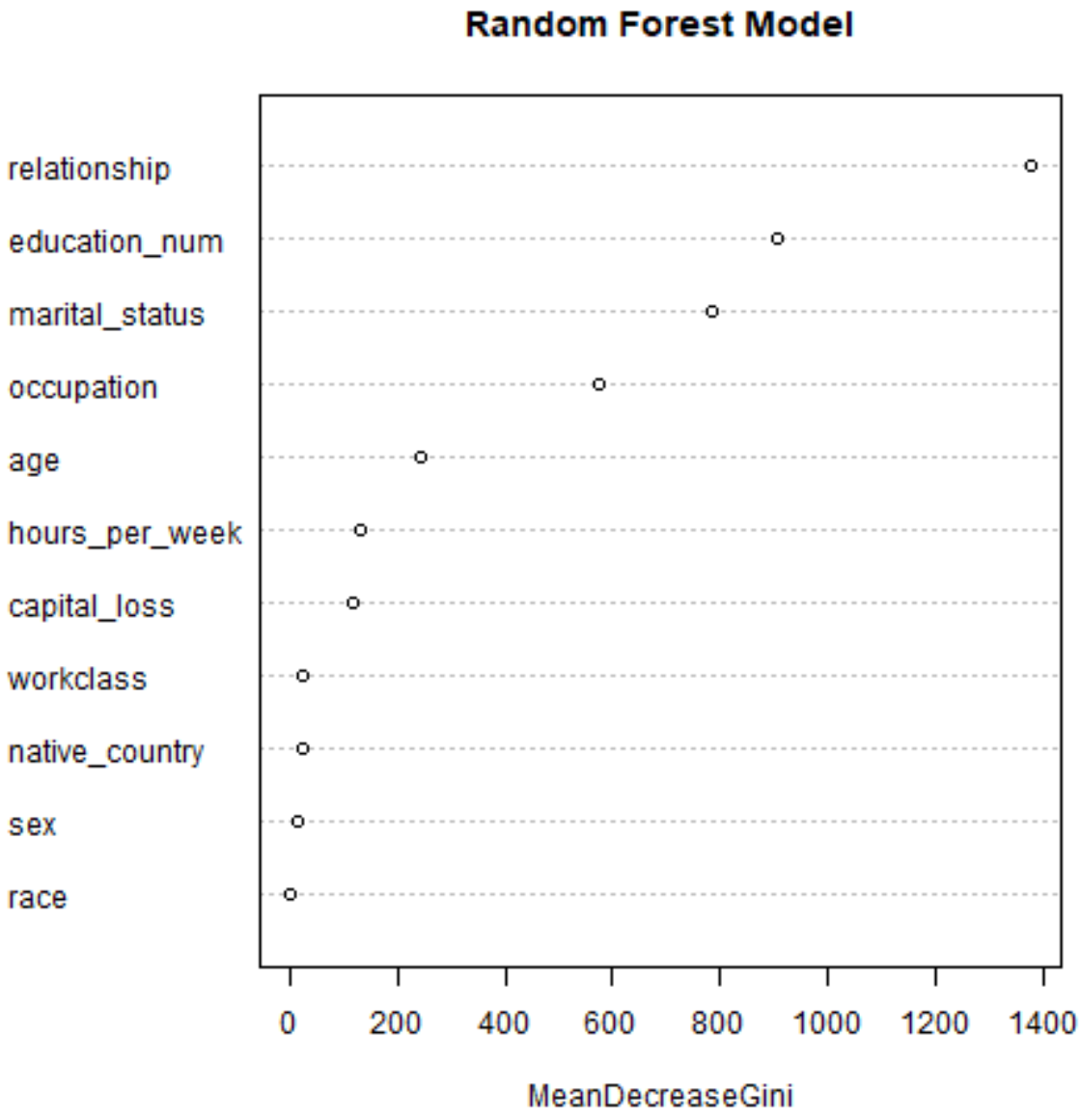

Random Forest Model
relationship
education_num
marital_status
occupation
age
hours_per_week
capital_loss
workclass
native_country
sex
race
0
200
400
600
800
1000
1200
1400
MeanDecreaseGini

Figure 4:

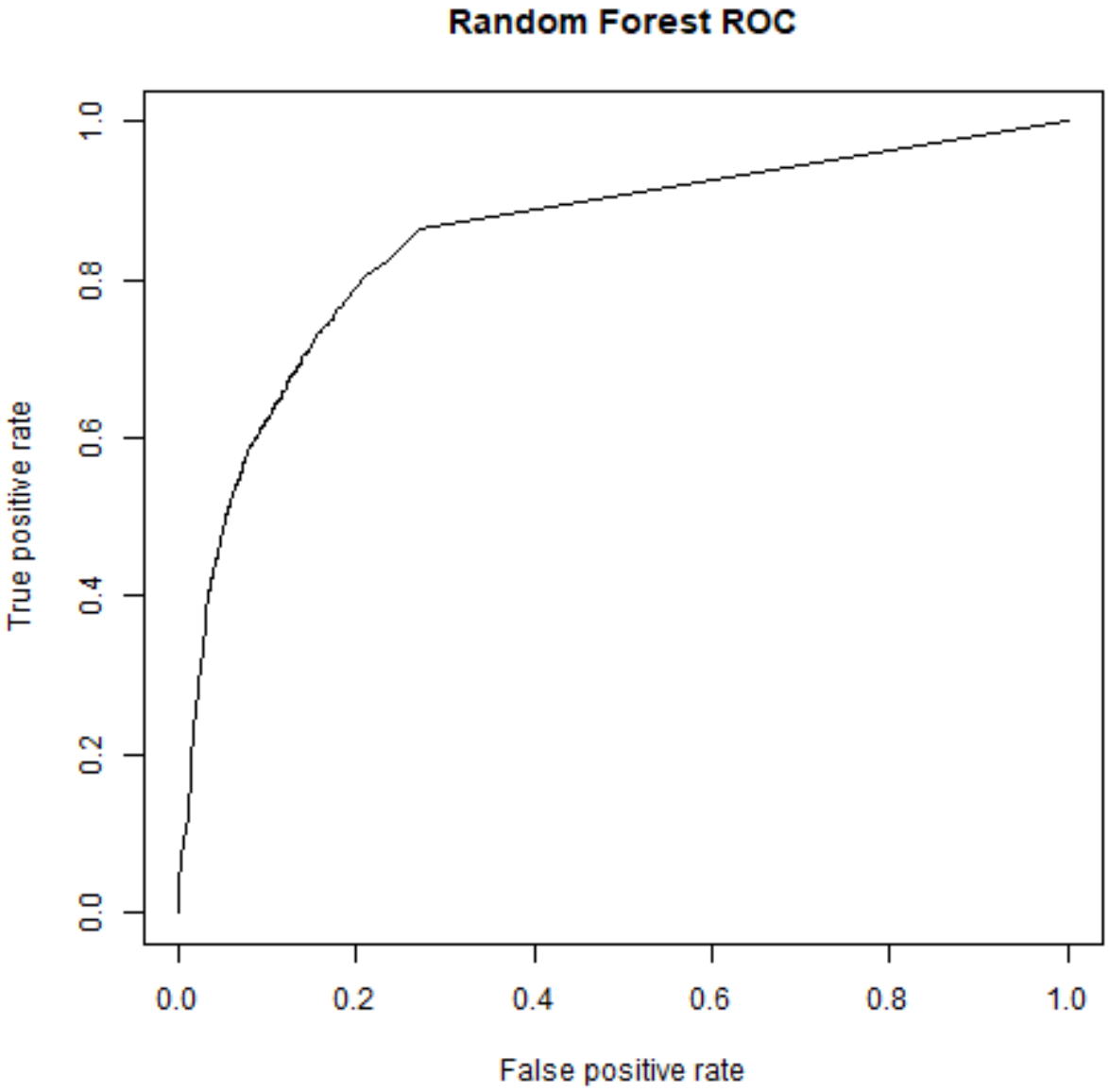

Random Forest ROC
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 5:

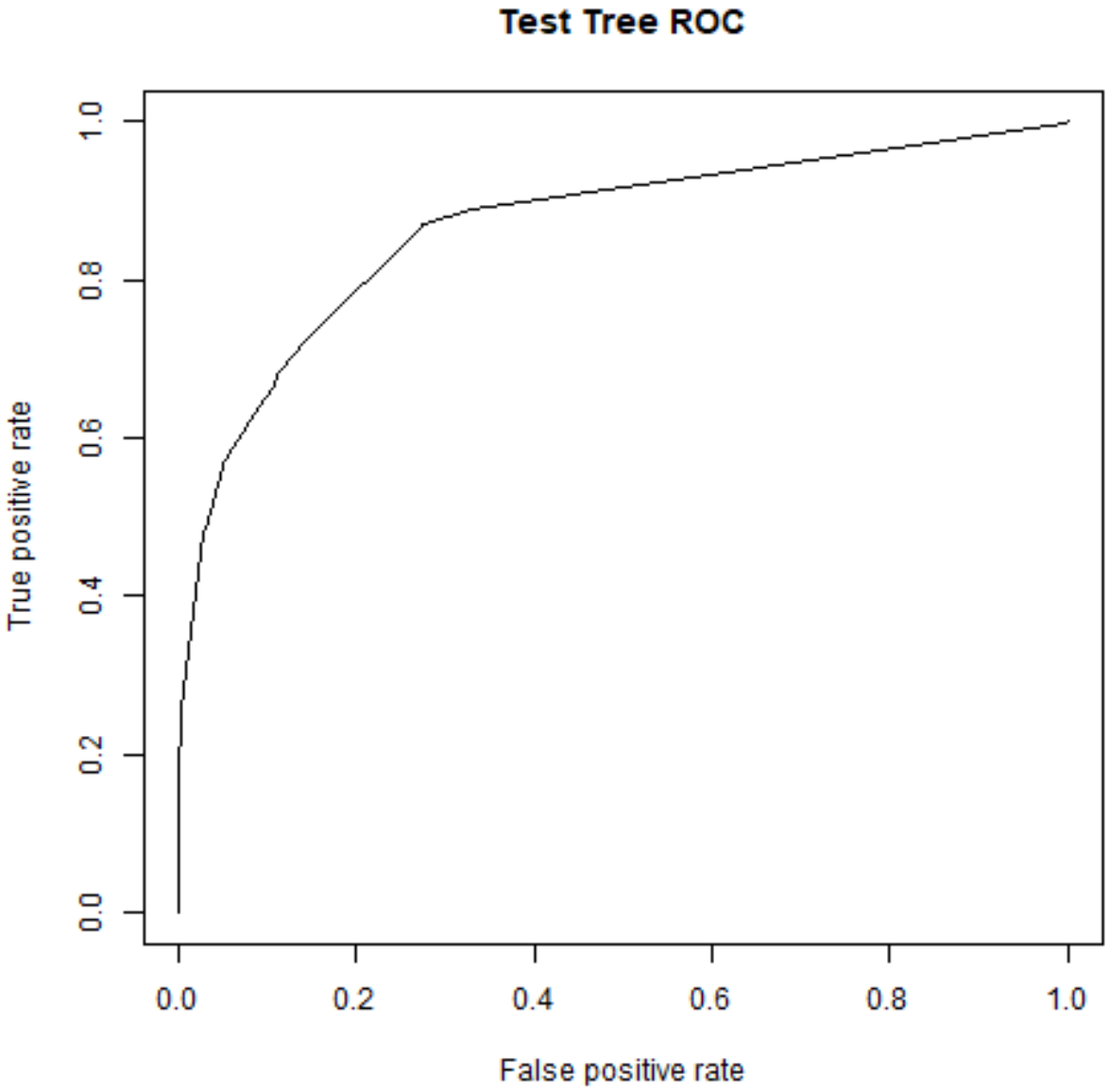

Test Tree ROC
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 6:

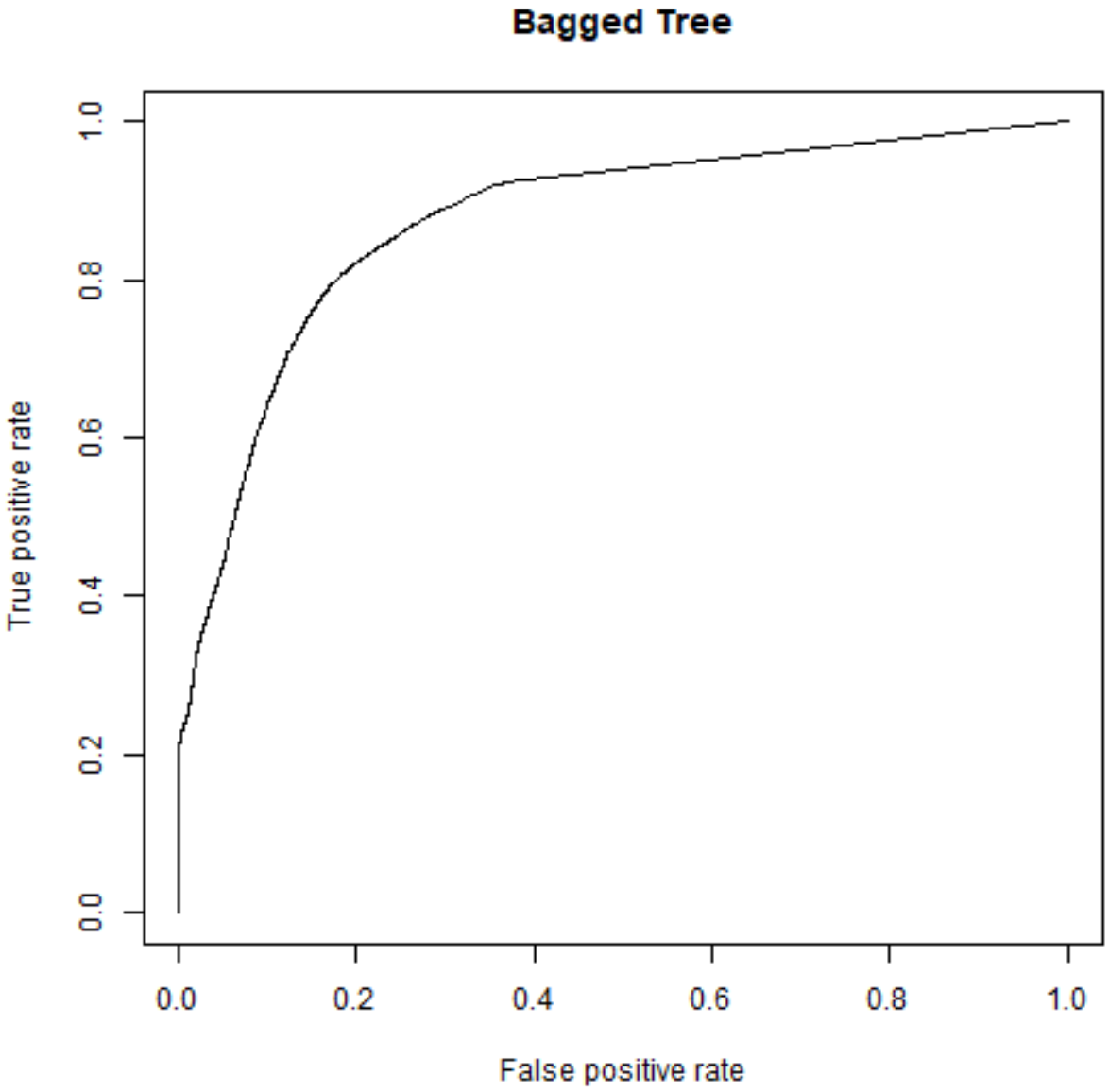

Bagged Tree
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 7:

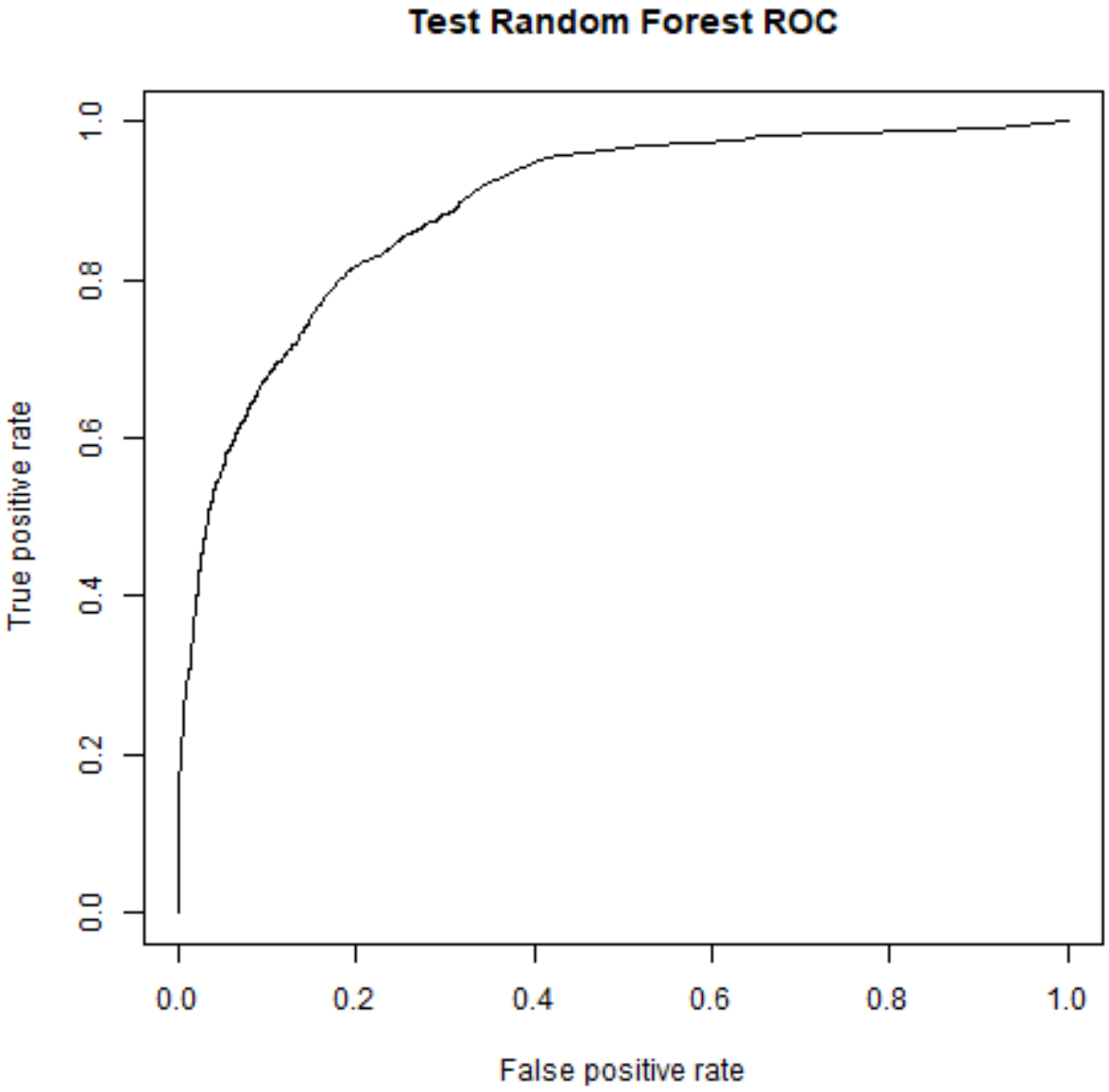

Test Random Forest ROC
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 8:

of education are strong indicators of income. It was interesting that race and gender didn't appear to be significant variables for the three models since these are widely thought of as key factors of income inequality.

Due to computational limits, more granular tuning of parameters could not be done, but it would be useful to tune more of the parameters and more values for each parameter to address the unbalanced nature of the data.