Homework 5

Henry Qiu, Goktug Cengiz, Mehmet Fatih Cagil

May 2019

1 Predictors decision and variable preprocess

The target variable is Adjusted, which is a binary variable. Value of 1 means it's a productive audit, value of 0 means a non-productive audit. Adjustment variable determines the target variable: when a client has adjustment 0, Adjusted value will be 0, this is why adjustment variable will not be a predictor. Our predictors are: Age, Income, Deductions and Hours are numerical variables, while Employment, Education, Marital, Occupation, Gender, Accounts are categorical. We see that there are several missing values in three variables: `Employment`, `Occupation`, `Accounts`.

We split Variable Age, Income and Hours into different categories to make easier their analysis. However, we have not done the same transformation to the variable Deductions. This is because the values are not distributed similarly, the amount of people without deduction is quite big, and the deduction obtained by the people are very varied. However this is not a problem for us because decision trees can work either with categorical or numerical variables.

2 Decision tree to predict "Adjusted" and the cutoff value

We have build our decision model with the function rpart in rpart Package of R. We had to take out the column of the Adjustment variable, because as we stated before, we are not going to use it as predictor.

	CP	nsplit	rel error	xerror	xstd
1	0.150000	0	1.00000	1.00000	0.048732
2	0.034375	2	0.70000	0.81250	0.045209
3	0.028125	3	0.66563	0.78750	0.044673
4	0.006250	5	0.60938	0.66563	0.041806

Figure 1

After we obtain the model, we prune the tree respect to a cutoff value (alpha), calculated as the maximal one up to the minimal crossvalidation error. With this criteria we obtain the optimal tree.

```
1) root 1333 320 0 (0.75993998 0.24006002)
  2) Marital=Absent,Divorced,Married-spouse-absent,Unmarried,Widowed 725  42 0 (0.94206897 0.05793103) *
  3) Marital=Married 608 278 0 (0.54276316 0.45723684)
    6) Occupation=Cleaner,Clerical,Farming,Machinist,Repair,Service,Transport 312  82 0 (0.73717949 0.262820
     12) Deductions< 1708 301  71 0 (0.76411960 0.23588040) *
     13) Deductions>=1708 11   0 1 (0.00000000 1.00000000) *
    7) Occupation=Executive,Professional,Protective,Sales,Support 296 100 1 (0.33783784 0.66216216)
     14) Education=College,HSgrad,Vocational,Yr10,Yr11,Yr5t6,Yr7t8,Yr9 120  57 0 (0.52500000 0.47500000)
       28) Employment=Consultant,Private,PSState 92  37 0 (0.59782609 0.40217391) *
       29) Employment=PSFederal,PSLocal,SelfEmp 28   8 1 (0.28571429 0.71428571) *
     15) Education=Associate,Bachelor,Doctorate,Master,Professional 176  37 1 (0.21022727 0.78977273) *
```

Figure 2

In the above results we can see the rules we can infer from the model, for example one rule we can obtain is that, when a client's aductaion is the typ Associate, Bachelor, Doctorate, Master or Professional, the probability that the audit is a productive audit is a 78.98%.

3 Importance of variables in the prediction.

As you can see from the figure below, when we visualize the importance of the variables, we can see that the most important variable is Marital and the least significant one is Accounts. In fact, if we remove the Accounts variable, we can say that nothing about the result will change. If we sort the variables by importance, we can say Marital > Income > Occupation > Education > Gender > Age > Hours > Deductions > Employment > Accounts.

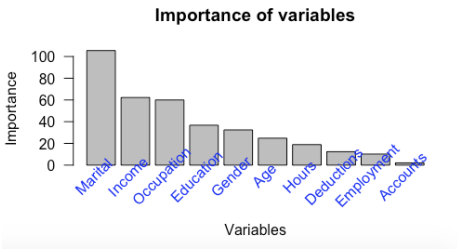


Figure 3: Importance plot

4 Compute the accuracy, precision, recall and AUC on the test individuals.

In order to summarize of prediction results on our classification problem, we created confusion matrix. The number of correct and incorrect predictions are summarized with count values and broken down by each class. So as to calculate it, we used our test dataset (It also could be done with validated data by cross validation). Firstly, We made a prediction for each row in your test dataset by using tree.optimal (as you see in our R script).

`Accuracy` is given by the relation: $(TP + TN)/N$ and our result is `0.826087` `Recall` can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN). Our result is `0.8591549`. Recall is given by the relation: $TP/(TP + FN)$

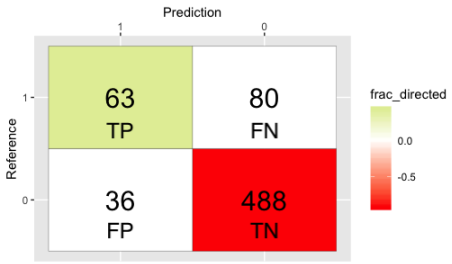


Figure 4: Prediction results

`Precision`, in order to get the value we divide the total number of correctly classified positive examples by the total number of predicted positive examples.High Precision indicates an example labeled as positive is indeed positive (small number of FP). Our result is `0.9312977` Precision is given by the relation: $TP/(TP + FP)$

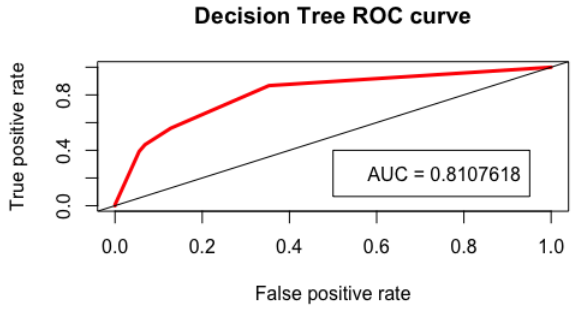


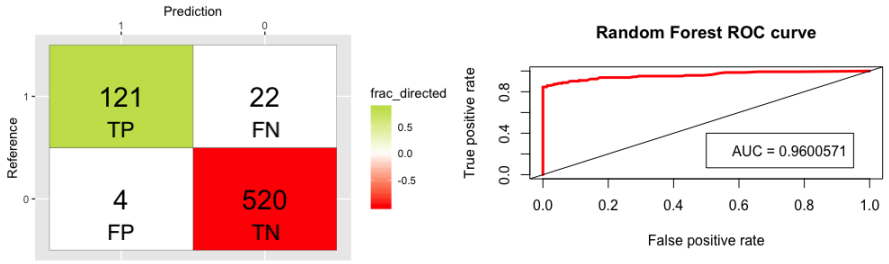
Figure 5

So as to create a complete sensitivity / specificity report and we have drawn the Roc Curve. Before drawing the Roc curve, we calculated the value specified as AUC 'The Area under the Curve' by obtaining the probabilities of the predictions of each class and result is obtained as 0.8107618 which is good.

5 Random Forest

Our data has missing values as we have observed above. Decision tree algorithms ignore missing data and can calculate, because the missing data imputation was unnecessary and therefore did not. However, for Random Forest, this is not valid and we have to do it. Firstly, All the character variables were transformed into factors and then missing values were imputed by Random Forest. As before,

we separated the dataset and changed the data type of the Adjusted variable to factor.



(a) (b)

Figure 6: Figures for RF

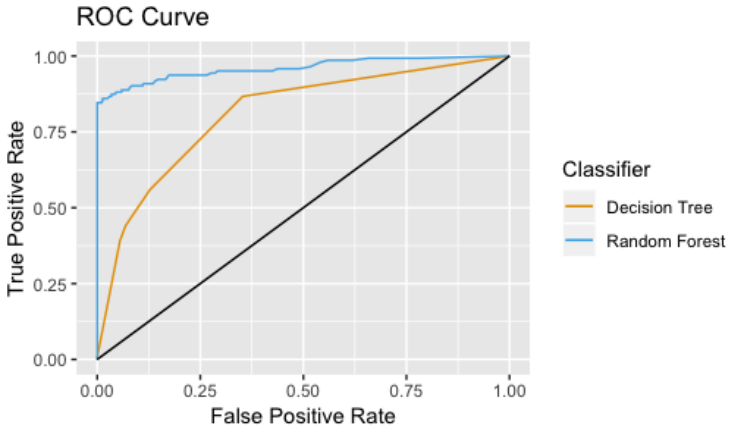


Figure 7

When we compare the random forest results to the previous one, we can see that these results are better. Random forest is more suitable and well resulted in this case.