Ryan Timbrook

Data Science 450, Spring 2017
Date: 04/30/2017
Assignment 2

Description: Decisions Trees and Classification Targeted Marketing Campaign

In this problem we will use historical data from past customer responses to build a classification model. We will apply the trained model to a new set of prospects to whom we may want extend an offer for a PEP. Rather than doing a mass marketing campaign to all new prospects, we would like to target those that are likely to respond positively to our offer (according to our classification model). Use the dataset provided (comma delimited, and the first row contains the field names) - bank-data.csv

A. Open the .CSV file and inspect the attributes, is there one or more attributes that should be removed? If so, do it now.

B. Optional - You can choose to use the ARFF format or use CSV, - Save the file in .ARFF format - Load the .ARFF formatted data file

C. Use Azure Machine Learning or a machine learning tool of your choice to build a classification model. You can choose to use any of the classification algorithms provided.

1. Evaluate your model accuracy

Spend time experimenting with different setting for the treebasses algorithms - such as pruning, binary branching, etc. to see if you can improve the performance of your model.

Consider performing parameter sweeping to optimize for specific evaluation metrics.

3. Review the structure of the tree

      See Table 1 below for tree representation of the models base structure using all parameters as a formula.

      This base models accuracy score is 84% with an AUC of 91%

4. Identify the more important feature information.

      See Table 3.1 below which represents the important feature information associated with its weighting

5. Generate the ROC curve for your final model.

      See table 3.3 below for the final models AUC ROC curve

      The final models accuracy score is 90.5% with an AUC of 91%

6. Show a screenshot of the final decision tree and model accuracy statistics obtained from your model.

      See tables 3 below for final models accuracy statistics and AUC scores.

      Accuracy: 90.5%

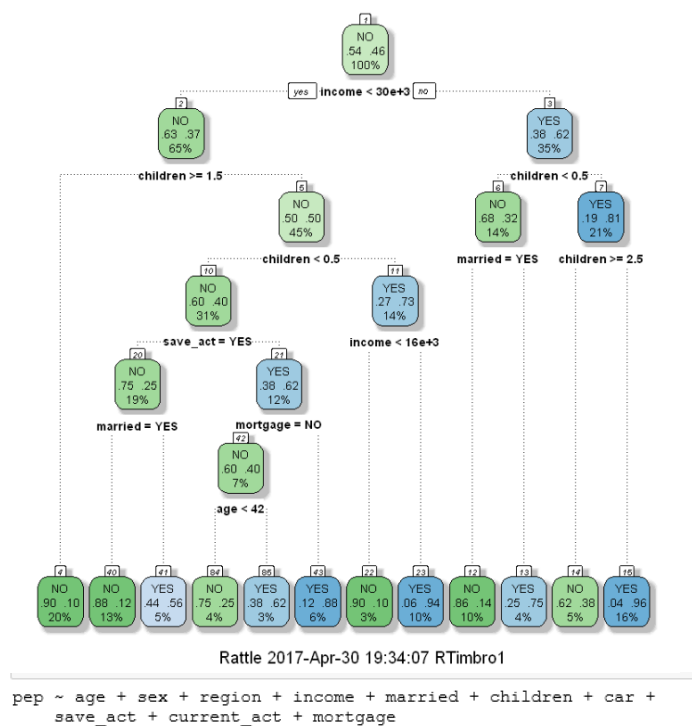   AUC: 91%

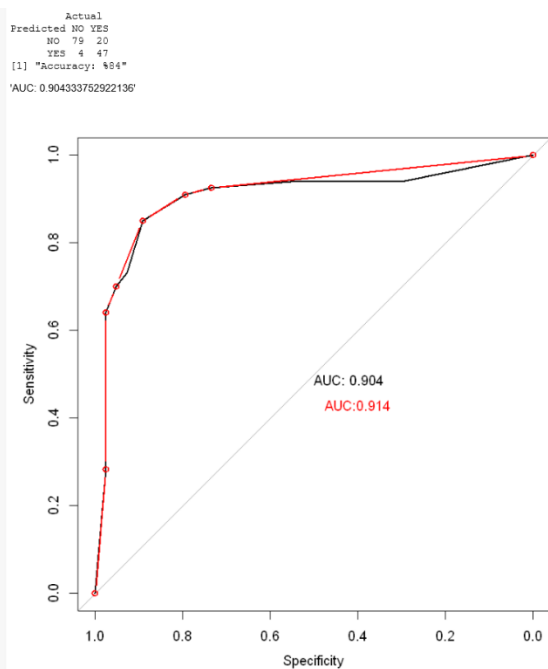# Table: 1.1 Single Tree plot – All base attributes

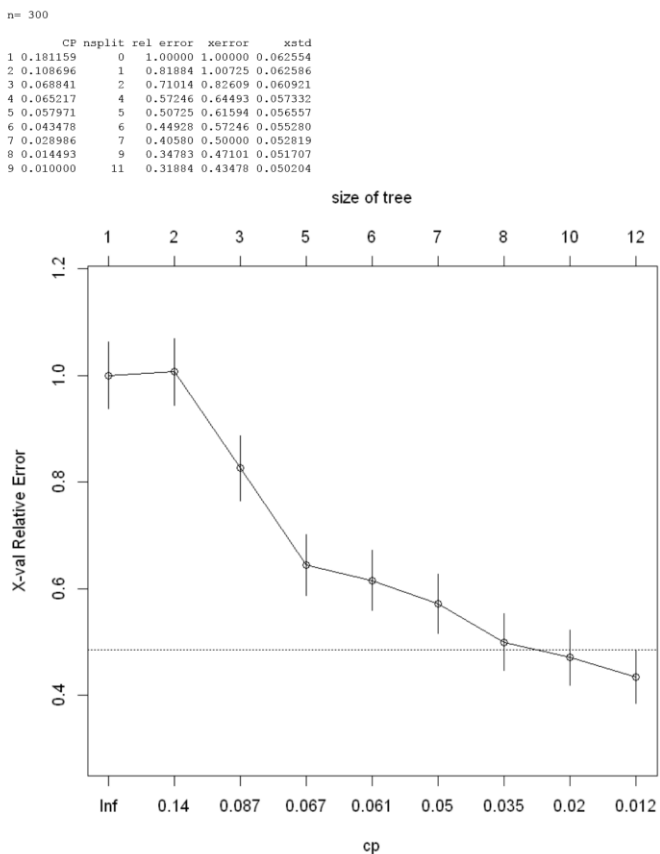Rattle 2017-Apr-30 19:34:07 RTimbro1

```
pep ~ age + sex + region + income + married + children + car +
    save_act + current_act + mortgage
```

Table: 1.2

```
            Actual
Predicted NO YES
      NO  79  20
      YES  4  47
[1] "Accuracy: %84"
'AUC: 0.904333752922136'
```


AUC: 0.904

AUC:0.914

Table: 1.3

```
n= 300

         CP nsplit rel error  xerror    xstd
1 0.181159      0  1.00000 1.00000 0.062554
2 0.108696      1  0.81884 1.00725 0.062586
3 0.068841      2  0.71014 0.82609 0.060921
4 0.065217      4  0.57246 0.64493 0.057332
5 0.057971      5  0.50725 0.61594 0.056557
6 0.043478      6  0.44928 0.57246 0.055280
7 0.028986      7  0.40580 0.50000 0.052819
8 0.014493      9  0.34783 0.47101 0.051707
9 0.010000     11  0.31884 0.43478 0.050204
```

Table: 2.1 Experiment with Pruning	Table: 2.2
Rattle 2017-May-06 16:44:16 RTimbro1	
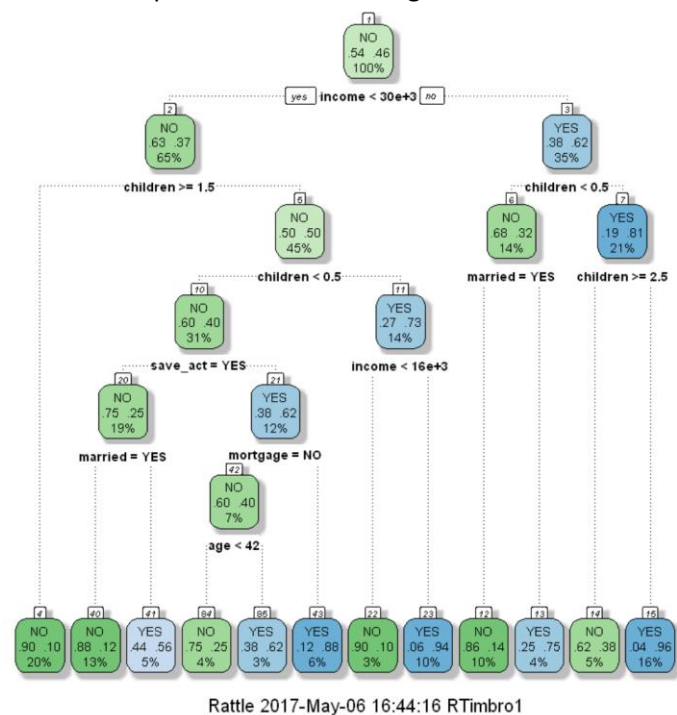
Table 2.3:	

```
Root node error: 138/300 = 0.46

n= 300

        CP nsplit rel error  xerror     xstd
1 0.181159      0  1.00000 1.00000 0.062554
2 0.108696      1  0.81884 1.00725 0.062586
3 0.068841      2  0.71014 0.82609 0.060921
4 0.065217      4  0.57246 0.64493 0.057332
5 0.057971      5  0.50725 0.61594 0.056557
6 0.043478      6  0.44928 0.57246 0.055280
7 0.028986      7  0.40580 0.50000 0.052819
8 0.014493      9  0.34783 0.47101 0.051707
9 0.010000     11  0.31884 0.43478 0.050204
```
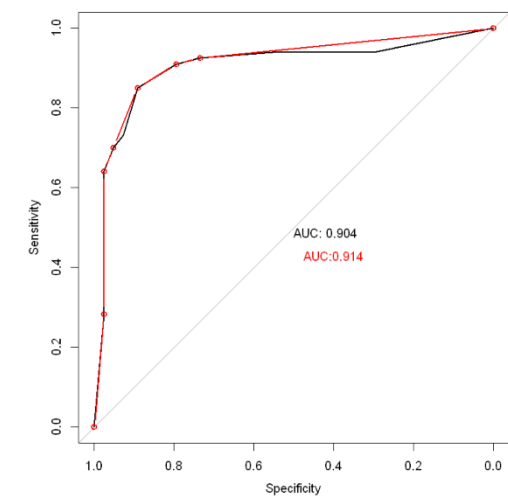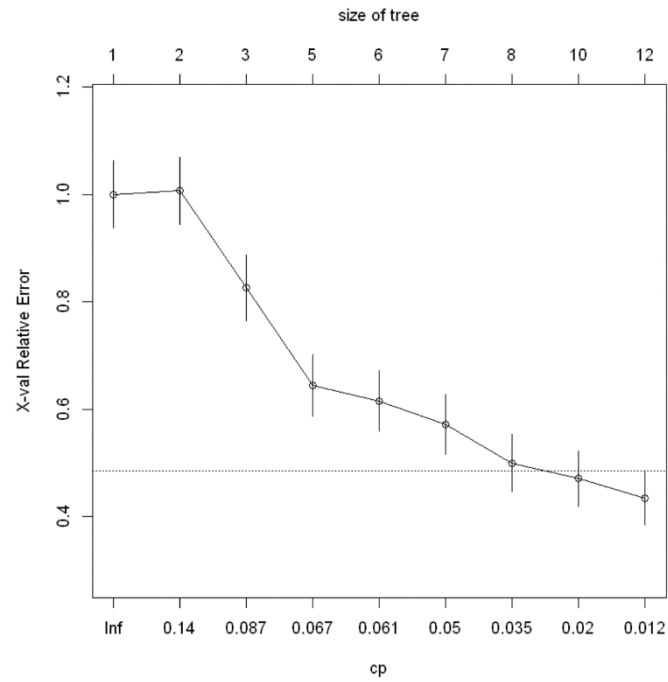
Table 3.1: Parameter Sweep model cp selection

```
CART

600 samples
 10 predictor
  2 classes: 'NO', 'YES'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 539, 541, 541, 541, 540, 540, ...
Resampling results across tuning parameters:

  cp            Accuracy   Kappa
  0.000000000   0.8801681  0.7565883
  0.001824818   0.8868922  0.7702459
  0.018248175   0.8836135  0.7636111
  0.032846715   0.8266287  0.6483315
  0.034671533   0.8199602  0.6344083
  0.040145985   0.8019001  0.5972433
  0.058394161   0.7717542  0.5367789
  0.062043796   0.7667250  0.5269858
  0.098540146   0.6963740  0.3666974
  0.124087591   0.6271390  0.2036250

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was cp = 0.001824818.
```

income	70.4343294945446	
children	54.178777167254	**Final Model feature**
mortgageYES	28.8626761973173	**importance list**
save_actYES	24.6513080259642	
marriedYES	22.0554065499718	
age	21.4636585991491	
regionRURAL	4.36771415189027	
current_actYES	1.3824058287796	
regionSUBURBAN	1.29657996357013	
sexMALE	1.29657996357013	
carYES	1.02050841274655	

Table 3.2: Parameter Sweep, Final Model

Rattle 2017-May-06 16:11:00 RTimbro1
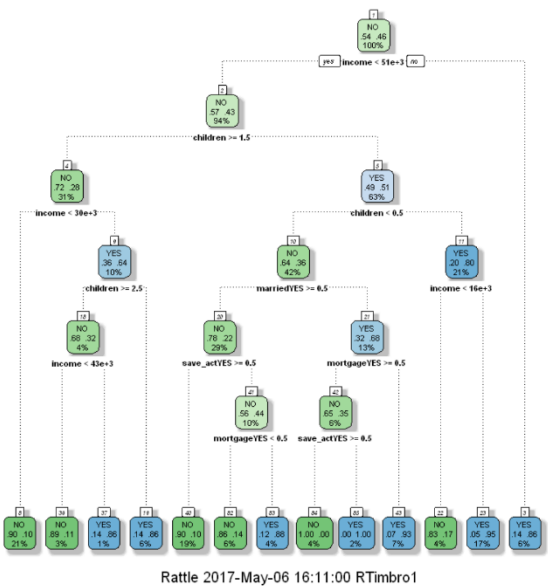
Table 3.3: Parameter Sweep – Final Model AUC

```
              Actual
Predicted   NO  YES
      NO   304   35
     YES    22  239
[1] "Accuracy: 90.5%"

'AUC: 0.920704401952443'
```

AUC: 0.921

AUC:0.916