HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

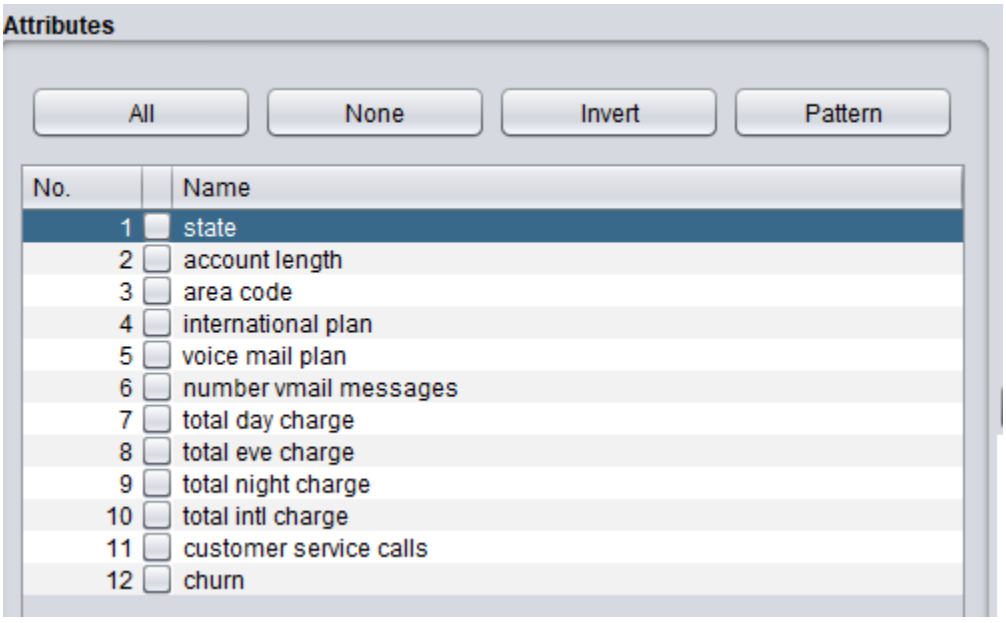
In this exercise, we will be using Weka to build a prediction model using a telecommunications data set looking at customer churn.

Import telecoms_churn.arff into Weka, and answer the following questions:

1. Data Questions

a. Which attributes should be included in the model, and which (if any) attributes should be eliminated from the model and why?

➔ I think that Phone Number can be deleted from the data set since this attribute is unique to each instance, which won't be of much help.
I also think that the attributes such as total minutes and total calls are not of much significance. These attributes can be ignored since we have "total charge" attribute which I think will be more helpful. Therefore, removing the following attributes from the data set:
total day minutes, total day calls, total evening minutes, total evening calls, total Night minutes, total Night calls, total Intl minutes, total Intl calls



HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

b. Some of them might need to be transformed. Which attributes do you feel should be transformation and why? Which transformation do you suggest and why?

➔ Attributes like account length, area code, number vmail messages, total day charge, total eve charge, total night charge, total intl charge, and customer service calls are Numeric. They need to be discretized since for decision tree, Weka can take only nominal values for variables. Therefore, I transformed these values from numeric to nominal by discretizing each attribute into 3 bins.

Next, construct a Decision Trees Model using the J48 Algorithm. Remove any attributes you believe should be removed, and transform the required attributes using the techniques outlined in 1.B

2. Data Model Questions

a. With no tuning, report the Correctly Classified Instances and the % of instances

➔ Without tuning, I am getting 973 correctly classified instances which is 85.8782%

```
=== Summary ===

Correctly Classified Instances         973               85.8782 %
Incorrectly Classified Instances       160               14.1218 %
Kappa statistic                          0.4058
Mean absolute error                      0.1602
Root mean squared error                  0.3434
Relative absolute error                 62.8059 %
Root relative squared error             92.4365 %
Total Number of Instances             1133
```

HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

b. Please describe what each element in the Confusion Matric means in this model

➔ Confusion Matrix tells us about the errors that is made by the model.
Here rows are the actual labels and the columns are the predicted labels. It tells us about the false positive and the false negative in a classification problem.
Therefore, 49 people are the churners, but they are predicted to be non-churners and 111 people are non-churners but are predicted to be churners.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 898  49 |   a = False
 111  75 |   b = True
```

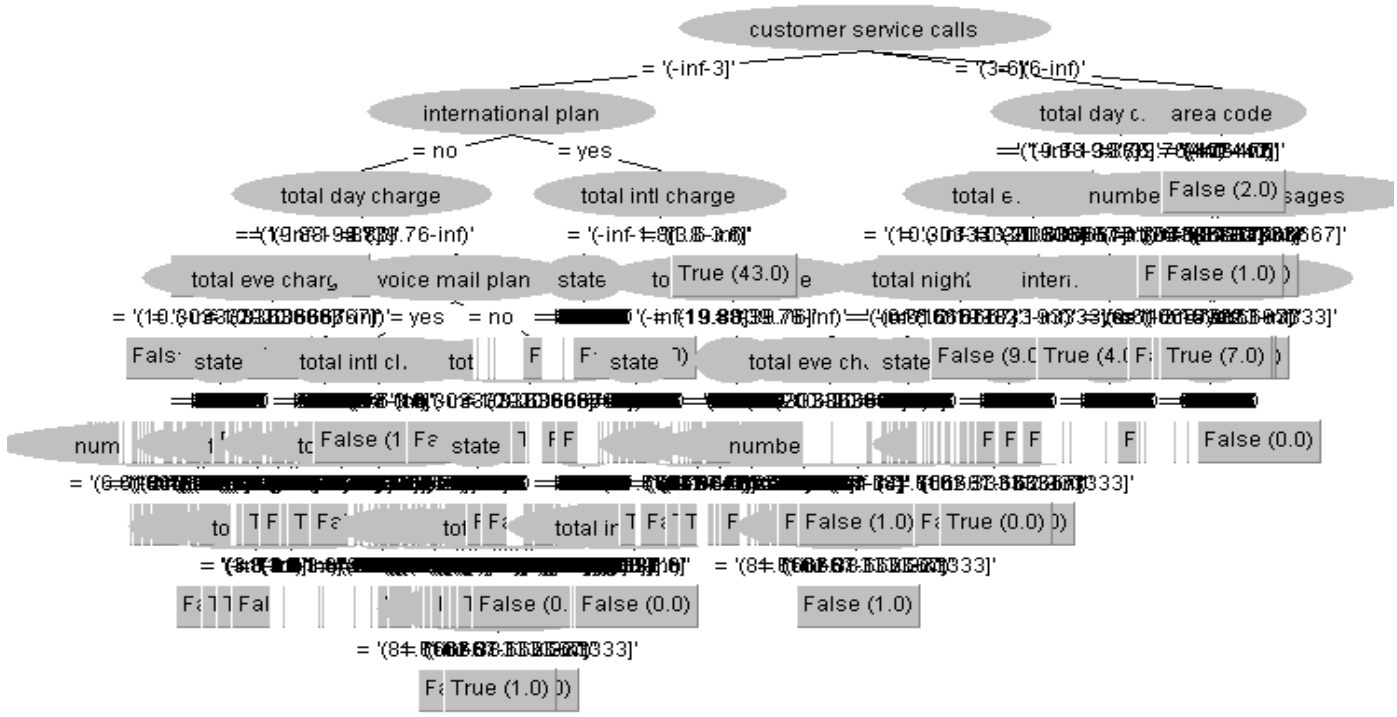
c. Include screen shot of model output.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -U -M 2
Relation:     bigml_59c28831336c6604c800002a-weka.filters.unsupervised.attribute.Remove-R4,8-9,11-12,14-15,17-18-weka.filters.unsupervised.attribute.
Instances:    3333
Attributes:   12
              state
              account length
              area code
              international plan
              voice mail plan
              number vmail messages
              total day charge
              total eve charge
              total night charge
              total intl charge
              customer service calls
              churn
Test mode:    split 66.0% train. remainder test
```

HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

```
Number of Leaves  :     743

Size of the tree :      805


Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         973               85.8782 %
Incorrectly Classified Instances       160               14.1218 %
Kappa statistic                          0.4058
Mean absolute error                      0.1602
Root mean squared error                  0.3434
Relative absolute error                 62.8059 %
Root relative squared error             92.4365 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               0.948     0.597     0.890       0.948     0.918     0.417    0.695     0.888     False
               0.403     0.052     0.605       0.403     0.484     0.417    0.695     0.455     True
Weighted Avg.  0.859     0.507     0.843       0.859     0.847     0.417    0.695     0.817
```



HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

3. Data Model Tuning

a. Experiment with your transformations. Please try 5 different bin sizes, and rerun the model for each. Do different bin sizes effect the model? If so, describe how.
i. Hint: consider accuracy and tree construction

➔ By changing the bin size, we can see that there is a change in the accuracy. The number of correctly classified instances changes. The tree construction also changes. The Number of leaves and the size of the tree increases, making the structure more complex.

Bin = 2

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        1011               89.2321 %
Incorrectly Classified Instances       122               10.7679 %
Kappa statistic                          0.5643
Mean absolute error                      0.131
Root mean squared error                  0.3156
Relative absolute error                 51.349  %
Root relative squared error             84.939  %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.960     0.452     0.915       0.960   0.937      0.572    0.756     0.911     False
                 0.548     0.040     0.729       0.548   0.626      0.572    0.756     0.537     True
Weighted Avg.    0.892     0.384     0.885       0.892   0.886      0.572    0.756     0.849

=== Confusion Matrix ===

   a   b   <-- classified as
 909  38 |   a = False
  84 102 |   b = True
```

Bin = 3

```
Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         973               85.8782 %
Incorrectly Classified Instances       160               14.1218 %
Kappa statistic                          0.4058
Mean absolute error                      0.1602
Root mean squared error                  0.3434
Relative absolute error                 62.8059 %
Root relative squared error             92.4365 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.948     0.597     0.890       0.948   0.918      0.417    0.695     0.888     False
                 0.403     0.052     0.605       0.403   0.484      0.417    0.695     0.455     True
Weighted Avg.    0.859     0.507     0.843       0.859   0.847      0.417    0.695     0.817

=== Confusion Matrix ===

   a   b   <-- classified as
 898  49 |   a = False
 111  75 |   b = True
```

HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

Bin = 4

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         978               86.3195 %
Incorrectly Classified Instances       155               13.6805 %
Kappa statistic                          0.3828
Mean absolute error                      0.1664
Root mean squared error                  0.3491
Relative absolute error                 65.2509 %
Root relative squared error             93.9638 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.965     0.656     0.882      0.965    0.922      0.409    0.687     0.888     False
                 0.344     0.035     0.660      0.344    0.452      0.409    0.687     0.430     True
Weighted Avg.    0.863     0.554     0.846      0.863    0.845      0.409    0.687     0.812

=== Confusion Matrix ===

   a   b   <-- classified as
 914  33 |   a = False
 122  64 |   b = True
```

Bin = 5

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         981               86.5843 %
Incorrectly Classified Instances       152               13.4157 %
Kappa statistic                          0.4439
Mean absolute error                      0.1652
Root mean squared error                  0.3436
Relative absolute error                 64.7731 %
Root relative squared error             92.4898 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.949     0.559     0.896      0.949    0.922      0.453    0.776     0.923     False
                 0.441     0.051     0.631      0.441    0.519      0.453    0.776     0.451     True
Weighted Avg.    0.866     0.476     0.853      0.866    0.856      0.453    0.776     0.845

=== Confusion Matrix ===

   a   b   <-- classified as
 899  48 |   a = False
 104  82 |   b = True
```

HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

Bin = 6

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         996               87.9082 %
Incorrectly Classified Instances       137               12.0918 %
Kappa statistic                          0.5143
Mean absolute error                      0.1486
Root mean squared error                  0.3324
Relative absolute error                 58.2589 %
Root relative squared error             89.4555 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.950    0.484    0.909      0.950    0.929      0.520    0.785     0.925     False
                 0.516    0.050    0.671      0.516    0.584      0.520    0.785     0.492     True
Weighted Avg.    0.879    0.413    0.870      0.879    0.873      0.520    0.785     0.854

=== Confusion Matrix ===

   a   b   <-- classified as
 900  47 |   a = False
  90  96 |   b = True
```

Bin = 7

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         970               85.6134 %
Incorrectly Classified Instances       163               14.3866 %
Kappa statistic                          0.4275
Mean absolute error                      0.1584
Root mean squared error                  0.3467
Relative absolute error                 62.1135 %
Root relative squared error             93.3015 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.543    0.898      0.935    0.916      0.432    0.768     0.919     False
                 0.457    0.065    0.578      0.457    0.511      0.432    0.768     0.472     True
Weighted Avg.    0.856    0.465    0.845      0.856    0.849      0.432    0.768     0.846

=== Confusion Matrix ===

   a   b   <-- classified as
 885  62 |   a = False
 101  85 |   b = True
```

HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

b. Experiment with Pruning. Using the decision-tree-Weka.pptx as a guide re-run you model adjusting each of the parameters below:

i. BinarySplit : True or False

ii. **unpruned**": True or False

iii. **ConfidenceFactor** range

➔ Changing from

BinarySplit = False
Unpruned = True
ConfidenceFactor = 0.25

to

BinarySplit = True
Unpruned = False
ConfidenceFactor = 0.5

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        1000               88.2613 %
Incorrectly Classified Instances       133               11.7387 %
Kappa statistic                          0.5478
Mean absolute error                      0.1437
Root mean squared error                  0.319
Relative absolute error                 56.3328 %
Root relative squared error             85.8666 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.425    0.919      0.943    0.931      0.550    0.750     0.914     False
                 0.575    0.057    0.665      0.575    0.617      0.550    0.750     0.517     True
Weighted Avg.    0.883    0.364    0.877      0.883    0.879      0.550    0.750     0.848

=== Confusion Matrix ===

   a   b   <-- classified as
 893  54 |   a = False
  79 107 |   b = True
```

HW 5: Decision Trees

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/29/2019**

c. Describe what impact, if any, these tuning factors had on model accuracy and tree construction

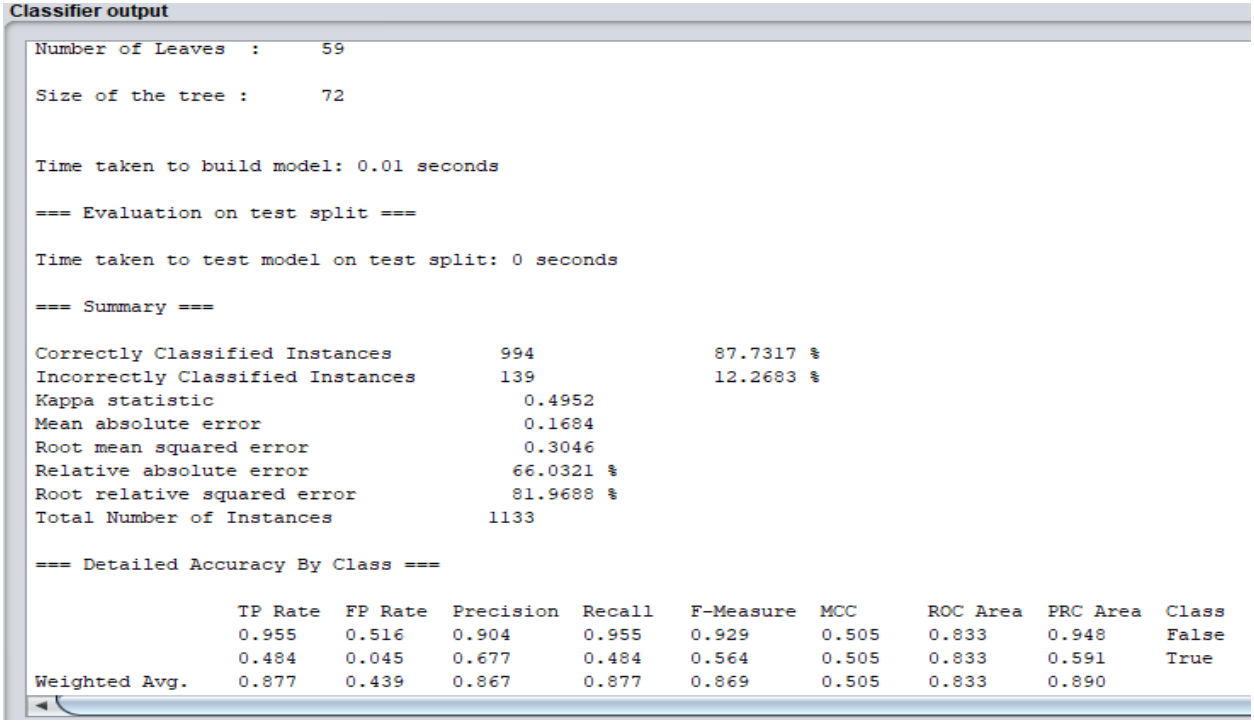
➔ By tuning the above-mentioned factors, the size of the tree and the number of the leaves decreased considerably. Initially, the tree size was 805 and the number of leaves was 743. But after tuning, tree size was 235 and the number of leaves was 118.
Number of "correctly classified instances" also increased from 973 at 85.8782% to 1000 at 88.2613%.

4. Optimal Model

a. Using all of the variables defined above; please define the best model accuracy you can achieve.

➔ By setting the BinarySplit = False, Unpruned = False and the ConfidenceFactor = 0.15, I am getting the number of "Correctly Classified Instances" to be 994 at 87.73172%. The size of the tree has decreased to 72 and Number of Leaves has also decreased to 59.

b. Provide screen shot for optimal model



Classifier output

```
Number of Leaves  :     59

Size of the tree :      72


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         994               87.7317 %
Incorrectly Classified Instances       139               12.2683 %
Kappa statistic                          0.4952
Mean absolute error                      0.1684
Root mean squared error                  0.3046
Relative absolute error                 66.0321 %
Root relative squared error             81.9688 %
Total Number of Instances             1133

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.955     0.516     0.904      0.955     0.929      0.505     0.833     0.948     False
                 0.484     0.045     0.677      0.484     0.564      0.505     0.833     0.591     True
Weighted Avg.    0.877     0.439     0.867      0.877     0.869      0.505     0.833     0.890
```