

Predictive Analytics and Data Mining
Peer-graded Assignment: Module 2 Peer Review Assignment
Aya Anisa Dwinidasari – January 30th, 2021

1. First, I input the data into the rattle with default partition and “yesno” as the target as shown on figure below.

The screenshot shows the Rattle Version 5.4.0 interface. The title bar indicates the file is 'R Data Miner - [Rattle (Pfb6qsiNTsuweqrjS7L2w_e730dd22de6c42059bb8417445c21cba_spam (2).csv)]'. The menu bar includes Project, Tools, Settings, and Help. The toolbar contains icons for Execute, New, Open, Save, Export, Stop, and Quit. The main window has tabs for Data, Explore, Test, Transform, Cluster, Associate, Model, Evaluate, and Log. The 'Model' tab is active, showing the following settings:

- Type: ☒ Tree ☐ Forest ☐ Boost ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ All
- Target: yesno Algorithm: ☒ Traditional ☐ Conditional Model Builder: rpart
- Min Split: 20 Max Depth: 30 Priors: ☐ Include Missing
- Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

Summary of the Decision Tree model for Classification (built using 'rpart'):

```
n= 3220
node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 3220 1275 n (0.6040373 0.3959627)
 2) dollar< 0.0555 2416 564 n (0.7665563 0.2334437)
   4) bang< 0.086 1663 166 n (0.9001804 0.0998196) *
   5) bang>=0.086 753 355 y (0.4714475 0.5285525)
      10) crl.tot< 85.5 383 117 n (0.6945170 0.3054830)
          20) bang< 0.825 304 68 n (0.7763158 0.2236842) *
              21) bang>=0.825 79 30 y (0.3797468 0.6202532) *
          11) crl.tot>=85.5 370 89 y (0.2405405 0.7594595) *
      3) dollar>=0.0555 804 93 y (0.1156716 0.8843284) *
```

Classification tree:

```
rpart(formula = yesno ~ ., data = crs$dataset[crs$train, c(crs$input,
  crs$target)], method = "class", model = TRUE, parms = list(split = "information"),
  control = rpart.control(usesurrogate = 0, maxsurrogate = 0))
```

Variables actually used in tree construction:

```
[1] bang    crl.tot dollar
```

The Decision Tree model has been built. Time taken: 0.10 secs

2. Evaluate Step.

After input the data, I evaluate the error by using VALIDATION as shown on the figure below.

The screenshot shows the R Data Miner application window. The title bar indicates the file path: R Data Miner - [Rattle (Pfb6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417445c21cba_spam (2).csv)]. The menu bar includes Project, Tools, Settings, and Help. The toolbar contains icons for Execute, New, Open, Save, Export, Stop, and Quit. The main menu bar has tabs for Data, Explore, Test, Transform, Cluster, Associate, Model, Evaluate, and Log. The Evaluate tab is active, showing various configuration options.

Type: ☒ Error Matrix ☐ Risk ☐ Cost Curve ☐ Hand ☐ Lift ☐ ROC ☐ Precision ☐ Sensitivity ☐ Pr v Ob ☐ Score

Model: ☒ Tree ☐ Boost ☐ Forest ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ KMeans ☐ HClust

Data: ☐ Training ☒ Validation ☐ Testing ☐ Full ☐ Enter ☐ CSV File ☐ (None) ☐ R Dataset

Risk Variable: Report: ☒ Class ☐ Probability Include: ☐ Identifiers ☒ All

Error matrix for the Decision Tree model on Pfb6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

Predicted			
Actual	n	y	Error
n	382	46	10.7
y	52	210	19.8

Error matrix for the Decision Tree model on Pfb6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

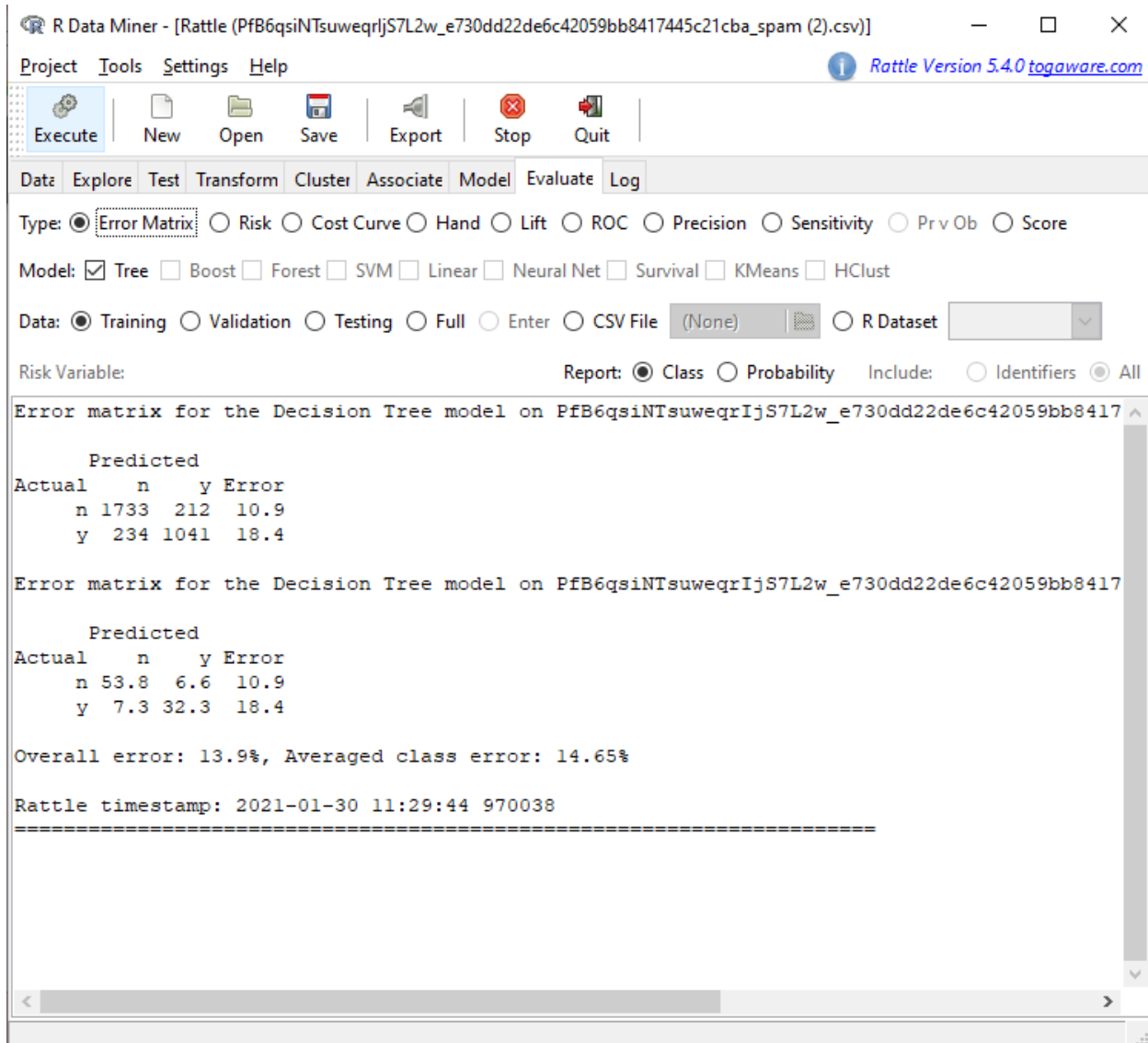
Predicted			
Actual	n	y	Error
n	55.4	6.7	10.7
y	7.5	30.4	19.8

Overall error: 14.2%, Averaged class error: 15.25%

Rattle timestamp: 2021-01-30 11:31:34 970038

Generated confusion matrix.

After I did a validation data, the next step is comparing the error result with the TRAINING mode as shown on the figure below. The VALIDATION and TRAINING mode would give different error result.



3. Error Comparison

In this section we get two different errors from those two methods (Training and Validation)

- **VALIDATION:**
Overall error: 14.2 %, Averaged class error: 15.25%
- **TRAINING:**
Overall error: 13.9 %, Averaged class error: 14.65%

As the matter of fact, we define that the **error rate** from **VALIDATION** is **higher** than the TRAINING. The higher the error rate indicates that there are overfitting or too complex model. Overfitting is caused by many variance and noise in the model, so we should reduce the overfitting to get less error rate.

There are some ways to reduce the overfitting:

- Increase MIN SPLIT, to control the node to be less split.
- Increase MIN BUCKET, to control the tree from expanding too bigger.
- Increase COMPLEXITY PARAMETER (CP), to control the parameter of tree too be less growing.
- Decrease MAX DEPTH, to control the tree not too be too depth/big.
- Lets try changing the CP number.

So I did some iteration in those four parameters to reduce the overfitting, with the model parameter as shown on the figure below.

The screenshot shows the RStudio Model Builder interface. The 'Model' tab is selected. The configuration is as follows:

- Type: ☒ Tree
- Target: yesno
- Algorithm: ☒ Traditional
- Model Builder: rpart
- Min Split: 25
- Max Depth: 10
- Priors: (empty)
- Include Missing: ☐
- Min Bucket: 12
- Complexity: 0.0005
- Loss Matrix: (empty)
- Buttons: Rules, Draw

Below the configuration, the 'Summary of the Decision Tree model for Classification (built using 'rpart'):' is displayed. The summary shows the tree structure with nodes and their associated statistics.

```

n= 3220

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 3220 1275 n (0.60403727 0.39596273)
2) dollar< 0.0555 2416 564 n (0.76655629 0.23344371)
4) bang< 0.086 1663 166 n (0.90018040 0.09981960)
8) crl.tot< 15.5 410 12 n (0.97073171 0.02926829) *
9) crl.tot>=15.5 1253 154 n (0.87709497 0.12290503)
18) money< 0.01 1228 141 n (0.88517915 0.11482085) *
19) money>=0.01 25 12 y (0.48000000 0.52000000)
38) crl.tot< 170.5 13 6 n (0.53846154 0.46153846) *
39) crl.tot>=170.5 12 5 y (0.41666667 0.58333333) *
5) bang>=0.086 753 355 y (0.47144754 0.52855246)
10) crl.tot< 85.5 383 117 n (0.69451697 0.30548303)
20) bang< 0.825 304 68 n (0.77631579 0.22368421)
40) crl.tot< 51.5 192 28 n (0.85416667 0.14583333) *
41) crl.tot>=51.5 112 40 n (0.64285714 0.35714286)
82) bang< 0.171 28 2 n (0.92857143 0.07142857) *
83) bang>=0.171 84 38 n (0.54761905 0.45238095)
166) bang< 0.4435 60 22 n (0.63333333 0.36666667)
332) crl.tot< 50.5 27 10 n (0.77777778 0.22222222) *

```

Then I did Validation evaluation from the new model.

Execute | New | Open | Save | Export | Stop | Quit

Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Type: ☒ Error Matrix ☐ Risk ☐ Cost Curve ☐ Hand ☐ Lift ☐ ROC ☐ Precision ☐ Sensitivity ☐ Prv Ob ☐ Score

Model: ☒ Tree ☐ Boost ☐ Forest ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ KMeans ☐ HClust

Data: ☐ Training ☒ Validation ☐ Testing ☐ Full ☐ Enter ☐ CSV File (None) ☐ R Dataset

Risk Variable: Report: ☒ Class ☐ Probability Include: ☐ Identifiers ☒ A

Error matrix for the Decision Tree model on PfB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

		Predicted	
Actual	n	y	Error
n	404	24	5.6
y	55	207	21.0

Error matrix for the Decision Tree model on PfB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

		Predicted	
Actual	n	y	Error
n	58.6	3.5	5.6
y	8.0	30.0	21.0

Overall error: 11.4%, Averaged class error: 13.3%

Rattle timestamp: 2021-01-30 13:32:16 970038

=====

After that, I did TRAINING Evaluation from the model too.

Execute | New | Open | Save | Export | Stop | Quit

Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Type: ☒ Error Matrix ☐ Risk ☐ Cost Curve ☐ Hand ☐ Lift ☐ ROC ☐ Precision ☐ Sensitivity ☐ Pr v Ob ☐ Score

Model: ☒ Tree ☐ Boost ☐ Forest ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ KMeans ☐ HClust

Data: ☒ Training ☐ Validation ☐ Testing ☐ Full ☐ Enter ☐ CSV File (None) ☐ R Dataset

Risk Variable: Report: ☒ Class ☐ Probability Include: ☐ Identifiers ☒ All

Error matrix for the Decision Tree model on PFB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

		Predicted		
Actual	n	y	Error	
n	1841	104	5.3	
y	246	1029	19.3	

Error matrix for the Decision Tree model on PFB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

		Predicted		
Actual	n	y	Error	
n	57.2	3.2	5.3	
y	7.6	32.0	19.3	

Overall error: 10.8%, Averaged class error: 12.3%

Rattle timestamp: 2021-01-30 13:32:52 970038

From the new model, I got new error rate, there are:

- **VALIDATION:**
Overall error: 11.4 %, Averaged class error: 13.0%
- **TRAINING:**
Overall error: 10.8 %, Averaged class error: 12.3%

From the new model above, we clearly get new smaller error rate than the previous model.

4. Forest

First, I built the forest using default setting from Rattle, as shown on the figure below.

R Data Miner - [Rattle (Pfb6qsiNTsuweqrjS7L2w_e730dd22de6c42059bb8417445c21cba_spam (2).csv)]

Project Tools Settings Help Rattle Version 5.4.0 togaware.com

Execute New Open Save Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☐ Tree ☒ Forest ☐ Boost ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ All

Target: yesno Algorithm: ☒ Traditional ☐ Conditional Model Builder: randomForest

Trees: 500 Sample Size: Importance Rules 1

Variables: 2 ☒ Impute Errors OOB ROC

Summary of the Random Forest Model

Number of observations used to build the model: 3220
Missing value imputation is active.

Call:

```
randomForest(formula = yesno ~ .,
              data = crs$dataset[crs$train, c(crs$input, crs$target)],
              ntree = 500, mtry = 2, importance = TRUE, replace = FALSE, na.action = randomForest::na.omit)

Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 2

OOB estimate of error rate: 11.52%


Confusion matrix:



|   | n    | y    | class.error |
|---|------|------|-------------|
| n | 1847 | 98   | 0.0503856   |
| y | 273  | 1002 | 0.2141176   |



Analysis of the Area Under the Curve (AUC)



Call:



The Random Forest model has been built. Time taken: 36.36 secs


```

VALIDATION EVALUATION

Data	Explore	Test	Transform	Cluster	Associate	Model	Evaluate	Log
------	---------	------	-----------	---------	-----------	-------	----------	-----

Type: ☒ Error Matrix ☐ Risk ☐ Cost Curve ☐ Hand ☐ Lift ☐ ROC ☐ Precision ☐ Sensitivity ☐ Prv Ob ☐ Score

Model: ☒ Tree ☐ Boost ☒ Forest ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ KMeans ☐ HClust

Data: ☐ Training ☒ Validation ☐ Testing ☐ Full ☐ Enter ☐ CSV File ☐ (None) ☐ R Dataset

Risk Variable: Report: ☒ Class ☐ Probability Include: ☐ Identifiers ☒ All

Error matrix for the Decision Tree model on PFB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

		Predicted		
Actual	n	y	Error	
n	55.4	6.7	10.7	
y	7.5	30.4	19.8	

Overall error: 14.2%, Averaged class error: 15.25%

rattle timestamp: 2021-01-30 13:42:39 970038

=====

Error matrix for the Random Forest model on PFB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

		Predicted		
Actual	n	y	Error	
n	406	22	5.1	
y	57	205	21.8	

Error matrix for the Random Forest model on PFB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

		Predicted		
Actual	n	y	Error	
n	58.8	3.2	5.1	
y	8.3	29.7	21.8	

Overall error: 11.5%, Averaged class error: 13.45%

TRAINING EVALUATION

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ Error Matrix ☐ Risk ☐ Cost Curve ☐ Hand ☐ Lift ☐ ROC ☐ Precision ☐ Sensitivity ☐ Prv Ob ☐ Score

Model: ☒ Tree ☐ Boost ☒ Forest ☐ SVM ☐ Linear ☐ Neural Net ☐ Survival ☐ KMeans ☐ HClust

Data: ☒ Training ☐ Validation ☐ Testing ☐ Full ☐ Enter ☐ CSV File ☐ (None) ☐ R Dataset

Risk Variable: Report: ☒ Class ☐ Probability Include: ☐ Identifiers ☒ A

Error matrix for the Decision Tree model on PfB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

	Predicted			
Actual	n	y	Error	
n	53.8	6.6	10.9	
y	7.3	32.3	18.4	

Overall error: 13.9%, Averaged class error: 14.65%

Rattle timestamp: 2021-01-30 13:43:40 970038

=====

Error matrix for the Random Forest model on PfB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

	Predicted			
Actual	n	y	Error	
n	1921	24	1.2	
y	219	1056	17.2	

Error matrix for the Random Forest model on PfB6qsiNTsuweqrIjS7L2w_e730dd22de6c42059bb8417

	Predicted			
Actual	n	y	Error	
n	59.7	0.7	1.2	
y	6.8	32.8	17.2	

Overall error: 7.5%, Averaged class error: 9.2%

As we see from the evaluation, I did training and validate error ratte comparison. It can be seen from the figure above that the **Overall** error from the **Random Forest Model is smaller** than the Decision Tree Model. But the disadvantage of the Random Forest Model is we cannot get a clear interpretation model like shown in the decision tree model.