



Easy Tree-sy, continued

CAS RPM Seminar

San Diego, California

Peggy Brinkmann, FCAS, MAAA

Advantages of Trees

- Easy to Interpret
- Automatic Variable Selection
- Automatic Interactions and Local Effects
- Handles Missing Values
- Handles Outliers
- Handles Monotonic Transformation



Applications of Trees

Enhancing GLMs. Decision trees do not require a lot of pre-processing of predictor variables to handle missing values and non-linear relationships – making them ideal for quickly screening a large number of potential predictor variables. Analyzing the “residuals” from a GLMs with a decision tree can help you identify additional transformations and/or interactions to improve the fit of your model, and as a check to make sure that no “signal” has been missed.

Portfolio diagnostics. A decision tree run with loss ratio as the target variable can help you identify segments with good profitability (to target marketing efforts) and poor profitability (for pricing revisions and/or underwriting action).

Checking/Quality Control. Ever tried to figure out why your complex calculated value (e.g. rerated premium, credit score) doesn't match to another source (e.g. company inforce premium, vendor calculated score)? Use the input variables as the predictors and the difference in values as the target variable, and you'll quickly find the sources of the discrepancies.

Demo

- Commercial lines automobile carrier
- Eight years of loss and premium data
- Want better segment underwriting and pricing
- Have looked at variables in one-way analysis
- Now what?

Data, part 1

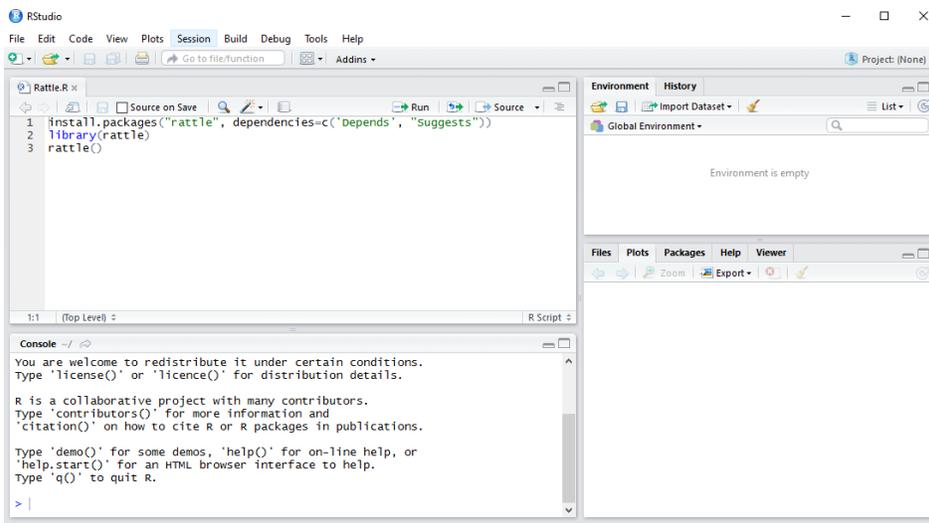
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	reccount	polno	state	effdate	effyr	expdate	cancel	lapse	poltype	vehtype	deluxmulti	color	V96
2	1	FLHP0000001	OH	20140415	2014	20150415	NA	NA	delux	full	No	White	G
3	2	FLHP0000002	OH	20140411	2014	20150411	NA	NA	delux	full	No	White	G
4	3	FLHP0000006	OH	20140411	2014	20150411	NA	NA	delux	full	No	White	G
5	4	FLHP0000009	OH	20140407	2014	20150407	NA	NA	delux	stand	No	White	H
6	5	FLHP0000010	OH	20140411	2014	20150411	NA	NA	delux	full	No	White	G
7	6	FLHP0000011	OH	20140411	2014	20150411	NA	NA	delux	full	No	White	G
8	7	FLHP0000012	OH	20140428	2014	20150428	NA	NA	delux	econ	No	White	G
9	8	FLHP0000017	OH	20140421	2014	20150421	NA	NA	delux	full	No	Blue	H
10	9	FLHP0000018	OH	20140425	2014	20150425	NA	NA	delux	full	No	White	H
11	10	FLHP0000019	OH	20140414	2014	20150414	NA	NA	delux	full	No	White	G
12	11	FLHP0000020	OH	20140407	2014	20150407	NA	NA	delux	econ	No	White	H
13	12	FLHP0000022	OH	20140411	2014	20150411	NA	NA	delux	full	No	White	G
14	13	FLHP0000023	OH	20140417	2014	20150417	NA	NA	delux	full	No	White	G
15	14	FLHP0000025	OH	20140411	2014	20150411	NA	NA	delux	econ	No	White	G
16	15	FLHP0000027	OH	20140408	2014	20150408	NA	NA	delux	stand	No	White	H
17	16	FLHP0000028	OH	20140408	2014	20150408	NA	NA	delux	stand	No	White	G
18	17	FLHP0000029	OH	20140408	2014	20150408	NA	NA	delux	stand	No	Blue	G

Data, part 2

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
limit	prem	incloss	claimct	exposure	modelyea	latepay	mileage_g	countynar	size_num	deductibl	power_nu	policyage
100	573	0	0	1	1996	NA	4	F	1500		1	0
100	1246	0	0	1	1978	NA	3	G	1500		1	0
100	991	0	0	1	1986	NA	2	H	1500		1	0
25	739	0	0	1	2004	NA	5	B	1500		1	0
100	732	0	0	1	1998	NA	4	F	2000		1	0
25	742	0	0	1	2000	NA	4	F	1500		1	0
100	1448	0	0	1	1990	NA	5	B	2000		1	0
100	2646	0	0	1	1985	NA	4	A	2000		1	0
500	1138	0	0	1	2006	NA	3	D	4000		1	0
100	1773	0	0	1	1988	NA	4	D	2000		1	0
500	895	0	0	1	2001	NA	6	F	3000		1	0
100	1063	0	0	1	2005	NA		E	2000		1	0
100	1219	0	0	1	1960	NA	3	B	2000		1	0
100	1537	0	0	1	1983	NA	4	F	1500		1	0
500	1585	0	0	1	1990	NA	3	B	2500		1	0
100	717	0	0	1	1987	NA	3	B	2000		1	0
100	2585	0	0	1	1974	NA	5	A	2000		1	0

Screenshots

7



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains three lines of R code:

```
1 install.packages("rattle", dependencies=c('depends', "suggests"))
2 library(rattle)
3 rattle()
```
- Environment:** Shows "Global Environment" and "Environment is empty".
- Console:** Displays the R startup message:

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

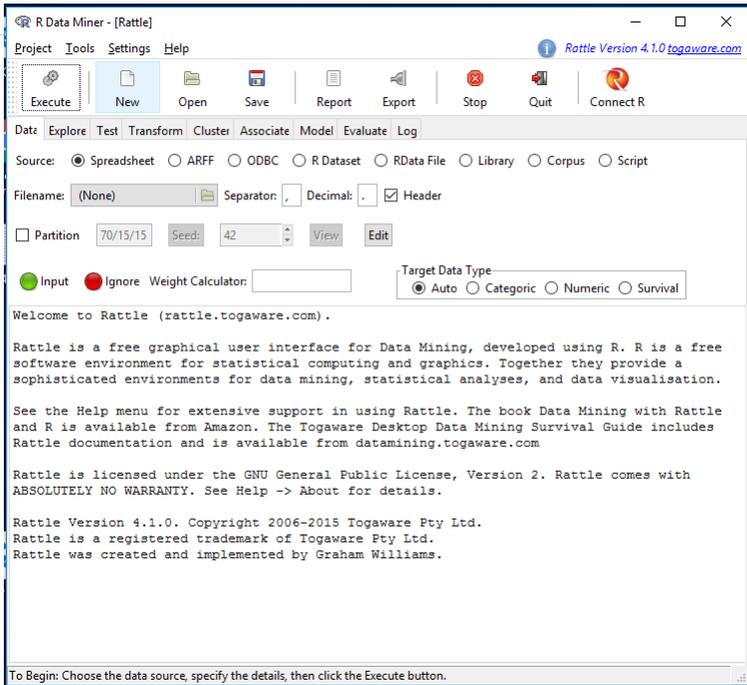
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

This is all the R code
you need to write to
run Rattle!

Click Run to submit

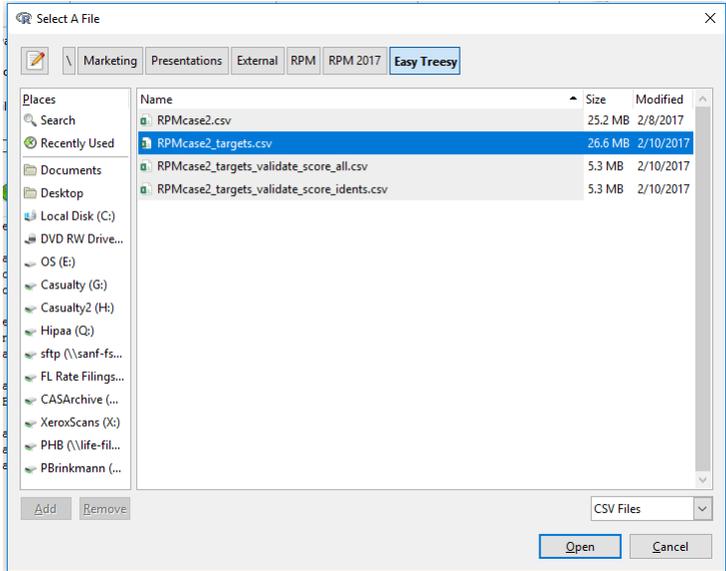


The screenshot shows the Rattle software interface. The 'Data' menu is open, displaying options for data source selection: Spreadsheet (selected), ARFF, ODBC, R Dataset, RData File, Library, Corpus, and Script. Below this, there are fields for 'Filename: (None)', 'Separator:', 'Decimal:', and a checked 'Header' box. There are also 'Partition' and 'Seed' settings, and a 'Target Data Type' section with 'Auto' selected. The main window contains a welcome message and instructions: 'To Begin: Choose the data source, specify the details, then click the Execute button.'

Select your data format, filename, delimiter (if applicable), and if it has a header with variable names

Then click "Execute"

Milliman



The screenshot shows a 'Select A File' dialog box. The current directory is 'Marketing'. A list of files is shown with columns for Name, Size, and Modified. The file 'RPMcase2_targets.csv' is selected.

Name	Size	Modified
RPMcase2.csv	25.2 MB	2/9/2017
RPMcase2_targets.csv	26.6 MB	2/10/2017
RPMcase2_targets_validate_score_all.csv	5.3 MB	2/10/2017
RPMcase2_targets_validate_score_idents.csv	5.3 MB	2/10/2017

Milliman

R Data Miner - [Rattle (RPMcase2_targets.csv)]

Project Tools Settings Help

Rattle Version 4.1.0 togaware.com

Execute New Open Save Report Export Stop Quit Connect R

Date Explore Test Transform Cluster Associate Model Evaluate Log

Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script

Filename: RPMcase2_target... Separator: Decimal: Header

Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator: Target Data Type: Auto Categorical Numeric Survival

No. Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1 reccount	Ident	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 238531
2 polno	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 102607
3 state	Categorical	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
4 effdate	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2731
5 effyr	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 8
6 expdate	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2731
7 cancel	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2 Missing: 33739
8 lapse	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2 Missing: 33739
9 poltype	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
10 vehetype	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 4
11 deluxmulti	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 3
12 color	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 3
13 V96	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 4 Missing: 15233
14 limit	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Uniaue: 4

Roles noted. 238531 observations and 26 input variables. The target is state. Categorical 2. Classification models enabled.

Rattle will guess how to use each variable; here reset the target to loss ratio, specify the weight (prem), and ignore variables like frequency, exposure, and cancellations



R Data Miner - [Rattle (RPMcase2_targets.csv)]

Project Tools Settings Help

Rattle Version 4.1.0 togaware.com

Execute New Open Save Report Export Stop Quit Connect R

Date Explore Test Transform Cluster Associate Model Evaluate Log

Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script

Filename: RPMcase2_target... Separator: Decimal: Header

Partition 50/50/50 Seed: 42 View Edit

Input Ignore Weight Calculator: Target Data Type: Auto Categorical Numeric Survival

No. Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
16 inccloss	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 4157
17 claimct	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 8
18 exposure	Constant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 1
19 modelyear	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 49
20 latepay	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 5 Missing: 33739
21 mileage_group	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 7 Missing: 110
22 countyname	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 8
23 size_numeric	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 11
24 deductible	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 5 Missing: 5949
25 power_numeric	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 4 Missing: 27790
26 policyage	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 8 Missing: 27790
27 freq	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 8
28 pureprem	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 4157
29 lossratio	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 5032

Roles noted. 238531 observations and 16 input variables. The target is lossratio. Numeric. Regression models enabled.

After making selections, hit Execute to refresh the data – see note at the bottom of the window



Basic statistics for each variable created from Summary - Describe

```

R Data Miner - [Rattle (RPMcase2_targets.csv)]
Project Tools Settings Help
Rattle Version 4.1.0 togaware.com
Execute New Open Save Report Export Stop Quit Connect R
Date: Explore Test Transform Cluster Associate Model Evaluate Log
Type:  Summary  Distributions  Correlation  Principal Components  Interactive
 Summary  Describe  Basics  Kurtosis  Skewness  Show Missing  Cross Tab
Below is a description of the dataset.
The data is limited to the training dataset.
crs$dataset[crs$sample, c(crs$input, crs$risk, crs$target)]
17 Variables 119265 Observations
-----
state
  n missing distinct
119265 0 2
Value AZ OH
Frequency 74013 45252
Proportion 0.621 0.379
-----
effyr
  n missing distinct Info Mean Gmd
119265 0 8 0.967 2012 1.898
Value 2007 2008 2009 2010 2011 2012 2013 2014
Frequency 408 2164 9392 15861 17853 22409 24451 26727
Proportion 0.003 0.018 0.079 0.133 0.150 0.188 0.205 0.224
-----
poltype
  n missing distinct
119265 0 2
Find: _____ Find Next
Data summary generated.
  
```



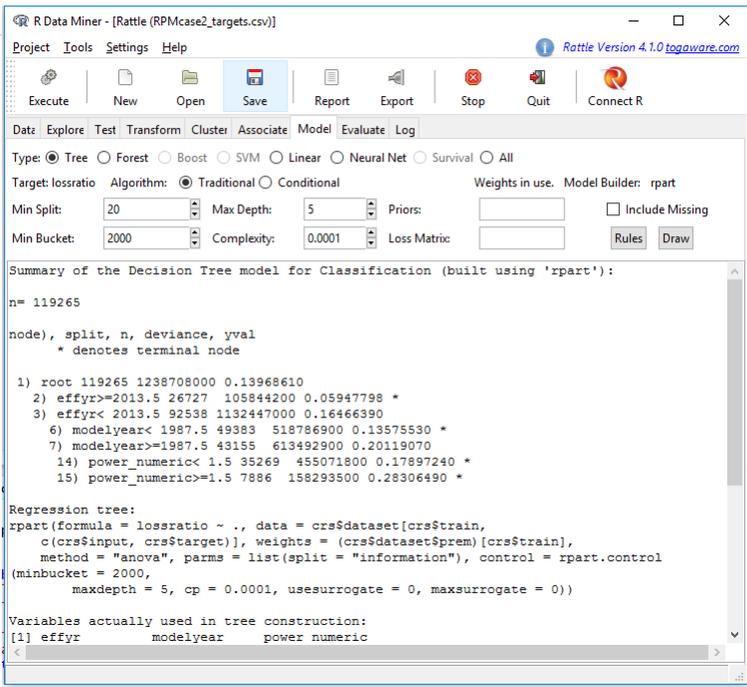
Rattle can fit other types of models besides trees, as well as k-means clusters and association analysis

Note the default parameters for trees – not appropriate for loss ratio analysis!

```

R Data Miner - [Rattle (RPMcase2_targets.csv)]
Project Tools Settings Help
Rattle Version 4.1.0 togaware.com
Execute New Open Save Report Export Stop Quit Connect R
Date: Explore Test Transform Cluster Associate Model Evaluate Log
Type:  Tree  Forest  Boost  SVM  Linear  Neural Net  Survival  All
Target: lossratio Algorithm:  Traditional  Conditional Weights in use: Model Builder: rpart
Min Split: 20 Max Depth: 30 Priors: _____  Include Missing
Min Bucket: 7 Complexity: 0.0100 Loss Matrix: _____
Decision Tree Model
A decision tree model is one of the most common data mining models. It is popular because the resulting model is easy to understand. The algorithms use a recursive partitioning approach.
The traditional algorithm is implemented in the rpart package. It is comparable to CART and ID3/C4.
The conditional tree algorithm is implemented in the party package. It builds trees in a conditional inference framework.
Note that the ensemble approaches (boosting and random forests) tend to produce models that exhibit less bias and variance than a single decision tree.
  
```





The screenshot shows the R Data Miner interface with the following configuration:

- Type: Tree
- Algorithm: Traditional
- Model Builder: rpart
- Min Split: 20, Max Depth: 5, Complexity: 0.0001
- Min Bucket: 2000

The summary of the decision tree model for Classification (built using 'rpart') is as follows:

```

n= 119265
node), split, n, deviance, yval
* denotes terminal node
1) root 119265 1238708000 0.13968610
2) effyr>=2013.5 26727 105844200 0.05947798 *
3) effyr< 2013.5 92538 1132447000 0.16466390
6) modelyear< 1987.5 49383 518786900 0.13575530 *
7) modelyear>=1987.5 43155 613492900 0.20119070
14) power_numeric< 1.5 35269 455071800 0.17897240 *
15) power_numeric>=1.5 7886 158293500 0.28306490 *

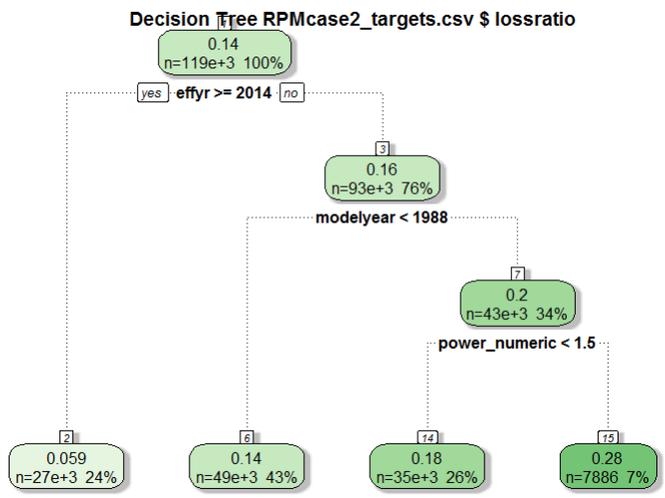
Regression tree:
rpart(formula = lossratio ~ ., data = crs$dataset[crs$train,
c(crs$input, crs$target)], weights = (crs$dataset$prem)[crs$train],
method = "anova", parms = list(split = "information"), control = rpart.control
(minbucket = 2000,
maxdepth = 5, cp = 0.0001, usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] effyr modelyear power numeric
    
```

Reset parameters to require at least 1K policies per node, and limit tree to 5 levels of splits

For regression tree, complexity parameter is the minimum change in R squared to make a split





Decision Tree RPMcase2_targets.csv \$ lossratio

```

graph TD
    Node1["0.14  
n=119e+3 100%"]
    Node2["0.059  
n=27e+3 24%"]
    Node3["0.16  
n=93e+3 76%"]
    Node4["0.14  
n=49e+3 43%"]
    Node5["0.2  
n=43e+3 34%"]
    Node6["0.18  
n=35e+3 26%"]
    Node7["0.28  
n=7886 7%"]

    Node1 -- "yes effyr >= 2014" --> Node2
    Node1 -- "no" --> Node3
    Node3 -- "modelyear < 1988" --> Node4
    Node3 -- "no" --> Node5
    Node5 -- "power_numeric < 1.5" --> Node6
    Node5 -- "no" --> Node7
    
```

Rattle 2017-Mar-09 14:43:33 peggy.brinkmann

Output from the Draw function is sent to R Studio plot window



Evaluate tab has variety of measures that can be applied to training, validation, or full dataset

For regression tree, the most useful is Score, which applies the tree nodes to a dataset

Milliman

	A	B	C	D	AD
1	recount	polno	state	effdate	rpart
2	1	FLHP0000001	OH	20140415	0.059478
3	3	FLHP0000006	OH	20140411	0.059478
4	4	FLHP0000009	OH	20140407	0.059478
5	5	FLHP0000010	OH	20140411	0.059478
6	8	FLHP0000017	OH	20140421	0.059478
7	9	FLHP0000018	OH	20140425	0.059478
8	10	FLHP0000019	OH	20140414	0.059478
9	11	FLHP0000020	OH	20140407	0.059478
10	13	FLHP0000023	OH	20140417	0.059478
11	14	FLHP0000025	OH	20140411	0.059478
12	15	FLHP0000027	OH	20140408	0.059478
13	21	FLHP0000035	OH	20140409	0.059478
14	22	FLHP0000036	OH	20140409	0.059478
15	24	FLHP0000039	OH	20140428	0.059478
16	25	FLHP0000041	OH	20140415	0.059478
17	28	FLHP0000045	OH	20140415	0.059478
18	31	FLHP0000051	OH	20140414	0.059478
19	34	FLHP0000057	OH	20140411	0.059478

The column "rpart" shows the expected loss ratio for the node based on the training data

Milliman

	A	B	C	D	E
1					
2					
3	Row Labels	Sum of exposure	Sum of prem	Sum of incloss	Sum of Loss ratio
4	0.059477976	26509	49133517	3088809	0.06286562
5	0.135755252	49354	88307588	12283788	0.139102293
6	0.178972424	35439	55420814	10251548	0.184976496
7	0.283064949	7964	15219797	3560380	0.233930847
8	Grand Total	119266	208081716	29184525	0.140255115
9					
10					

Summarize the data by node in Excel to see how the tree performs on new data



```

R Data Miner - [Rattle (RPMcase2_targets.csv)]
Project Tools Settings Help
Execute New Open Save Report Export Stop Quit Connect R
Data Explore Test Transform Cluster Associate Model Evaluate Log
 Export Comments  Rename Internal Variables: From crs$ to MY
# Rattle is Copyright (c) 2006-2015 Togaware Pty Ltd.
#-----
# Rattle timestamp: 2017-03-07 08:47:34 x86_64-w64-mingw32
# Rattle version 4.1.0 user 'peggy.brinkmann'
# This log file captures all Rattle interactions as R commands.
Export this log to a file using the Export button or the Tools
# menu to save a log of all your activity. This facilitates repeatability. For example, ex
# to a file called 'myrf01.R' will allow you to type in the R Console
# the command source('myrf01.R') and so repeat all actions automatically.
# Generally, you will want to edit the file to suit your needs. You can also directly
# edit this current log in place to record additional information before exporting.
# Saving and loading projects also retains this log.
# We begin by loading the required libraries.
library(rattle) # To access the weather dataset and utility commands.
library(magrittr) # For the %>% and %<>% operators.
# This log generally records the process of building a model. However, with very
# little effort the log can be used to score a new dataset. The logical variable
# 'building' is used to toggle between generating transformations, as when building
# a model, and simply using the transformations, as when scoring a dataset.
building <- TRUE

```

We can get the R code to perform the analysis from the Log, like recording a macro in Excel

This can be a way to get a R program started, and you can modify and customize the code in R Studio





Thank You

peggy.brinkmann@milliman.com
415-394-3726