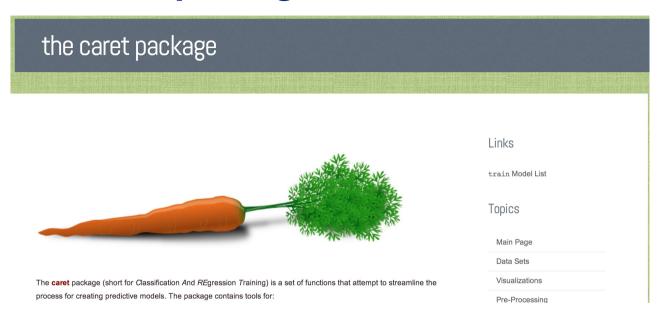


The caret package

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

The caret R package



http://caret.r-forge.r-project.org/

Caret functionality

- · Some preprocessing (cleaning)
 - preProcess
- Data splitting
 - createDataPartition
 - createResample
 - createTimeSlices
- Training/testing functions
 - train
 - predict
- · Model comparison
 - confusionMatrix

Machine learning algorithms in R

- · Linear discriminant analysis
- · Regression
- Naive Bayes
- · Support vector machines
- · Classification and regression trees
- · Random forests
- Boosting
- · etc.

Why caret?

obj Class	Package	predict Function Syntax
lda	MASS	<pre>predict(obj) (no options needed)</pre>
${ t glm}$	stats	<pre>predict(obj, type = "response")</pre>
gbm	gbm	<pre>predict(obj, type = "response", n.trees)</pre>
mda	mda	<pre>predict(obj, type = "posterior")</pre>
rpart	rpart	<pre>predict(obj, type = "prob")</pre>
Weka	RWeka	<pre>predict(obj, type = "probability")</pre>
LogitBoost	caTools	<pre>predict(obj, type = "raw", nIter)</pre>

http://www.edii.uclm.es/~useR-2013/Tutorials/kuhn/user_caret_2up.pdf

SPAM Example: Data splitting

```
[1] 3451 58
```

SPAM Example: Fit a model

```
set.seed(32343)
modelFit <- train(type ~.,data=training, method="glm")
modelFit</pre>
```

```
Generalized Linear Model
3451 samples
 57 predictors
  2 classes: 'nonspam', 'spam'
No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 3451, 3451, 3451, 3451, 3451, ...
Resampling results
 Accuracy Kappa Accuracy SD Kappa SD
 0.9 0.8 0.02
                      0.04
                                                                                       7/11
```

SPAM Example: Final model

```
modelFit <- train(type ~.,data=training, method="glm")
modelFit$finalModel</pre>
```

```
Call:
       NUTLL
Coefficients:
                                                    address
                                  make
                                                                             all
                                                                                               num3d
      (Intercept)
        -1.78e\pm00
                             -7.76e-01
                                                  -1.39e-01
                                                                        3.68e-02
                                                                                            1.94e + 00
                                                                        internet
                                                                                               order
               our
                                  over
                                                     remove
         7.61e-01
                              6.66e-01
                                                   2.34e \pm 00
                                                                       5.94e-01
                                                                                            4.10e-01
              mail
                               receive
                                                       will
                                                                          people
                                                                                              report
         4.08e-02
                              2.71e-01
                                                  -1.08e-01
                                                                      -2.28e-01
                                                                                           -1.14e-01
        addresses
                                                   business
                                                                           email
                                  free
                                                                                                  you
         2.16e+00
                                                                                            6.91e-02
                              8.78e-01
                                                  6.49e-01
                                                                       1.38e-01
           credit
                                  your
                                                       font
                                                                         num000
                                                                                               money
         8.00e-01
                              2.17e-01
                                                   2.17e-01
                                                                       2.04e+00
                                                                                            1.95e+00
                hp
                                   hpl
                                                     george
                                                                         num650
                                                                                                  lab
                                                                                           -1.89e+00 <sub>8/11</sub>
        -1.82e+00
                             -9.17e-01
                                                  -7.50e\pm00
                                                                       3.33e-01
              labs
                                telnet
                                                     num857
                                                                            data
                                                                                              num415
```

SPAM Example: Prediction

```
predictions <- predict(modelFit,newdata=testing)
predictions</pre>
```

- 1													
	[1]	spam	spam	spam	nonspam	nonspam	nonspam	spam	spam	spam	spam	spam	
	[12]	spam	nonspam	spam	spam	spam							
	[23]	nonspam	spam	nonspam	nonspam	spam	spam	spam	spam	spam	spam	spam	
	[34]	spam	spam	spam									
	[45]	spam	spam	spam	spam	nonspam	spam	nonspam	spam	spam	spam	spam	
	[56]	spam	nonspam	nonspam	spam	spam	spam	spam	spam	nonspam	spam	spam	
	[67]	spam	spam	spam									
	[78]	nonspam	nonspam	nonspam	spam	spam	nonspam	spam	nonspam	nonspam	spam	spam	
	[89]	spam	spam	spam	spam	spam	spam	nonspam	spam	spam	spam	spam	
	[100]	spam	spam	spam	nonspam	spam	nonspam	spam	spam	spam	spam	spam	
	[111]	spam	spam	spam	spam	nonspam	spam	spam	spam	spam	spam	spam	
	[122]	spam	nonspam	spam	spam	nonspar	ı						
	[133]	spam	spam	spam									
	[144]	spam	spam	spam	nonspam	spam	spam	spam	spam	spam	spam	spam	
	[155]	nonspam	spam	nonspam	spam	nonspam	spam	spam	spam	spam	spam	spam	
	[166]	spam	spam	spam (9/11								
	[177]	spam	spam	spam									

SPAM Example: Confusion Matrix

confusionMatrix(predictions, testing\$type)

```
Confusion Matrix and Statistics
```

Reference

Prediction nonspam spam

nonspam 665 54 spam 32 399

Accuracy: 0.925

95% CI: (0.908, 0.94)

No Information Rate: 0.606
P-Value [Acc > NIR]: <2e-16

Kappa : 0.842

Mcnemar's Test P-Value: 0.0235

Sensitivity: 0.954 Specificity: 0.881

Pos Pred Value: 0.925

Further information

- · Caret tutorials:
 - http://www.edii.uclm.es/~useR-2013/Tutorials/kuhn/user_caret_2up.pdf
 - http://cran.r-project.org/web/packages/caret/vignettes/caret.pdf
- · A paper introducing the caret package
 - http://www.jstatsoft.org/v28/i05/paper