Caret R Package

http://topepo.github.io/caret/index.html

R Caret Package

- Robust Functions to Manage Classification and Regression Model Workflows

Supports:

- Preprocessing
- Training
- Cross Validation
- Model Comparison and Evaluation

Data Exploration Visualizations

- featurePlot (wrapper for lattice plots)
  - Scatter Matrix
  - Overlayed Density Plots
  - Box Plots
  - Scatter Plots
- x,y,plot=”pairs|ellipse|density|box|scatter”
- layout, autokey

Caret Preprocessing

- dummyVars - convert factor columns to numeric
- nearZeroVar - generates metrics on predictors w/ zero or near zero variance (zeroVar,nzv)

```
nzv <- nearZeroVar(mdrrDescr)
filteredDescr <- mdrrDescr[, -nzv]
dim(filteredDescr)
```

- **cor + findCorrelation**
  - cor - create correlation matrix
  - findCorrelation - find features that are highly correlated to be used to remove duplicate columns

Caret Preprocessing continued

- findLinearCombos - uses QR decomposition to find list of linear combinations in the features
- preProcess - used to center and scale the data and impute missing values - uses **predict.preProcess**

  ```
  preProcess(training, method = c("center", "scale"))
  ```

  - imputation can use k-nearest neighbor or bagging (greater computation cost)

Caret - Training a Model

- Create a trainControl object (optional) and call train method

```
fitControl <- trainControl(## 10-fold CV
                           method = "repeatedcv",
                           number = 10,
                           ## repeated ten times
                           repeats = 10)

gbmFit1 <- train(Class ~ ., data = training,
                 method = "gbm",
                 trControl = fitControl,
                 ## This last option is actually one
                 ## for gbm() that passes through
                 verbose = FALSE)
```

Caret - trainControl

- method - resampling method for cross validation
- number - number of folds
- repeats - number of times folds performed
- PCAThresh, ICAThresh, and k
- allowParallel - used with doMC - more on that later
- many others

Caret - Methods - sampling

rf - random forest

lmt - logistic model trees

lasso - lasso

pls - partial least squares

lm - linear regression

nb - naive bayes

svm - support vector machines (multiple kernel functions available

Caret - Train

method - the algorithm you want to use (e.g., “rf”)

trControl - trainControl object

verbose

tuneGrid - grid with columns = fitting function’s arguments

preProc - preProcessing parameter - same as running before but integrated with the iterations

tuneLength - number of levels for each tuning parameters that should be generated by train method

metric - choice for model evaluation metric (e.g, ROC)

Caret - Predict

Generates predictions for each sub-model in the passed in trained model object

- object - model
- newData - new data you want to predict values for (e.g., testing data)
- type - raw (regression) or prob (classification)

Caret - Model Performance

sensitivity
specificity
posPredValue
negPredValue
postResample
confusionMatrix
mnLogLoss
multiClassSummary
twoClassSim
lift
calibration

doMC - Adds Multi-core Support

For use on servers with multiple cpu cores

```
install.packages(“doMC”)
library(doMC)
registerDoMC(cores=19)
train(class~.,data=thedata, method=”LMT”, allowParallel = TRUE)
```