Data Mining in R

- SEMINAR WINTER SEMESTER 2015/2016 -

Overview of the caretEnsemble package

Submitted by:

Juan David Correa

CaretEnsemble

- Why such package?
- The caret package
- The caretEnsemble package
- How does it work?
- Application
- Summary

Why such package ?

There are so many models in R for Predictive Modeling (aka machine learning/patter recognition) → Hence, **heterogeneity** in syntax !!!

obj **Class**	**Package**	predict **Function Syntax**
lda	MASS	`predict(obj)` (no options needed)
glm	stats	`predict(obj, type = "response")`
gbm	gbm	`predict(obj, type = "response", n.trees)`
mda	mda	`predict(obj, type = "posterior")`
rpart	rpart	`predict(obj, type = "prob")`
Weka	RWeka	`predict(obj, type = "probability")`
LogitBoost	caTools	`predict(obj, type = "raw", nIter)`

Generating class probabilities using different packages

Source: Kuhn (2014), useR! 2014

Classification **A**nd **RE**gression **T**raining, the caret package:

Max Kuhn (2005-2007)

Uniform interface for model training and/or prediction

Standardize common tasks (tuning parameters, variable importance, data splitting etc.)

Increase computational efficiency (parallel processing)

Among other functions, additional info at: caret.r-forge.r-project.org

Source: Kuhn (2014), Applied Predictive Modeling in R, useR! 2014

caretEsemble

Zachary A. Mayer, Jared E. Knowles

Making "easy" for the user to combine models to produce a meta-model with superior fit than the sub-models

caretList: creates lists of the models (caret) from the training data

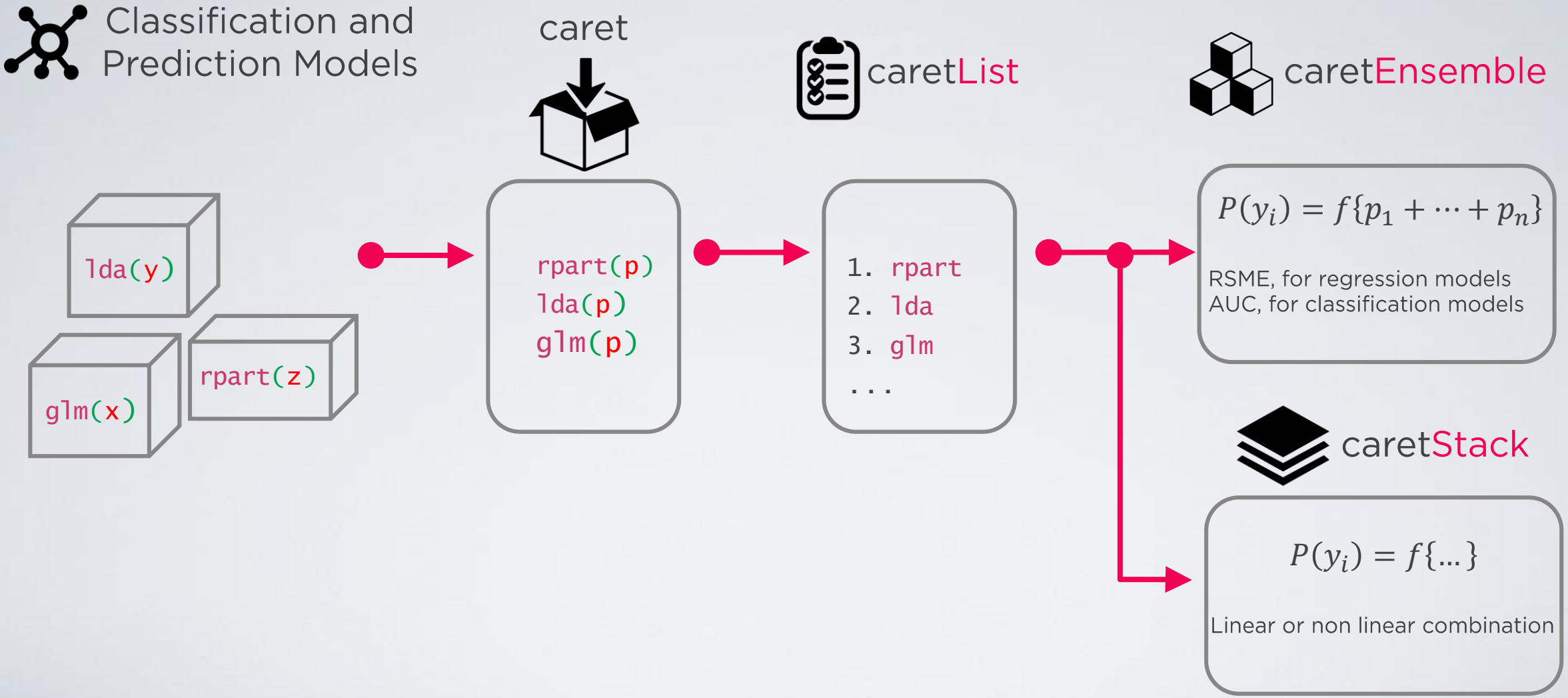
caretEnsemble: ensembles models from caretList via **weights** (greedy optimization)

caretStack: uses a caret model to merge the outputs from several components via **stacking**

Additional features to interact with the models (e.g. predict, summary, plot, etc.)

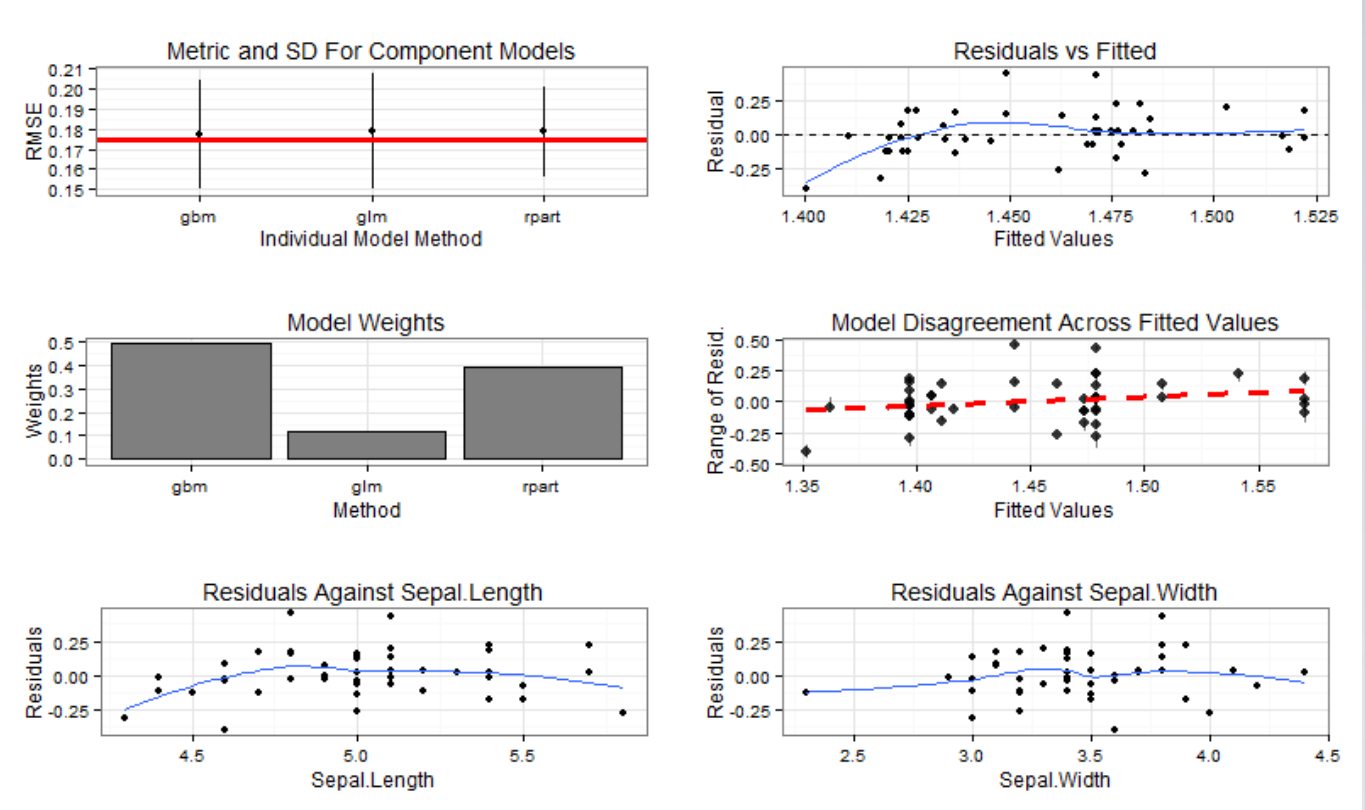
Source: Knowles (2015), Mayer (2013)

How it works...



Source: Knowles (2015), Mayer (2013)

CaretEnsemble provides additional tools to analyze models




```
The following models were ensembled: glm, rpart, gbm
They were weighted:
0.12 0.39 0.49
The resulting RMSE is: 0.1745
The fit for each individual model on the RMSE is:
 method    metric   metricSD
    glm 0.1790451 0.02851506
  rpart 0.1784955 0.02200974
    gbm 0.1770780 0.02687245
```

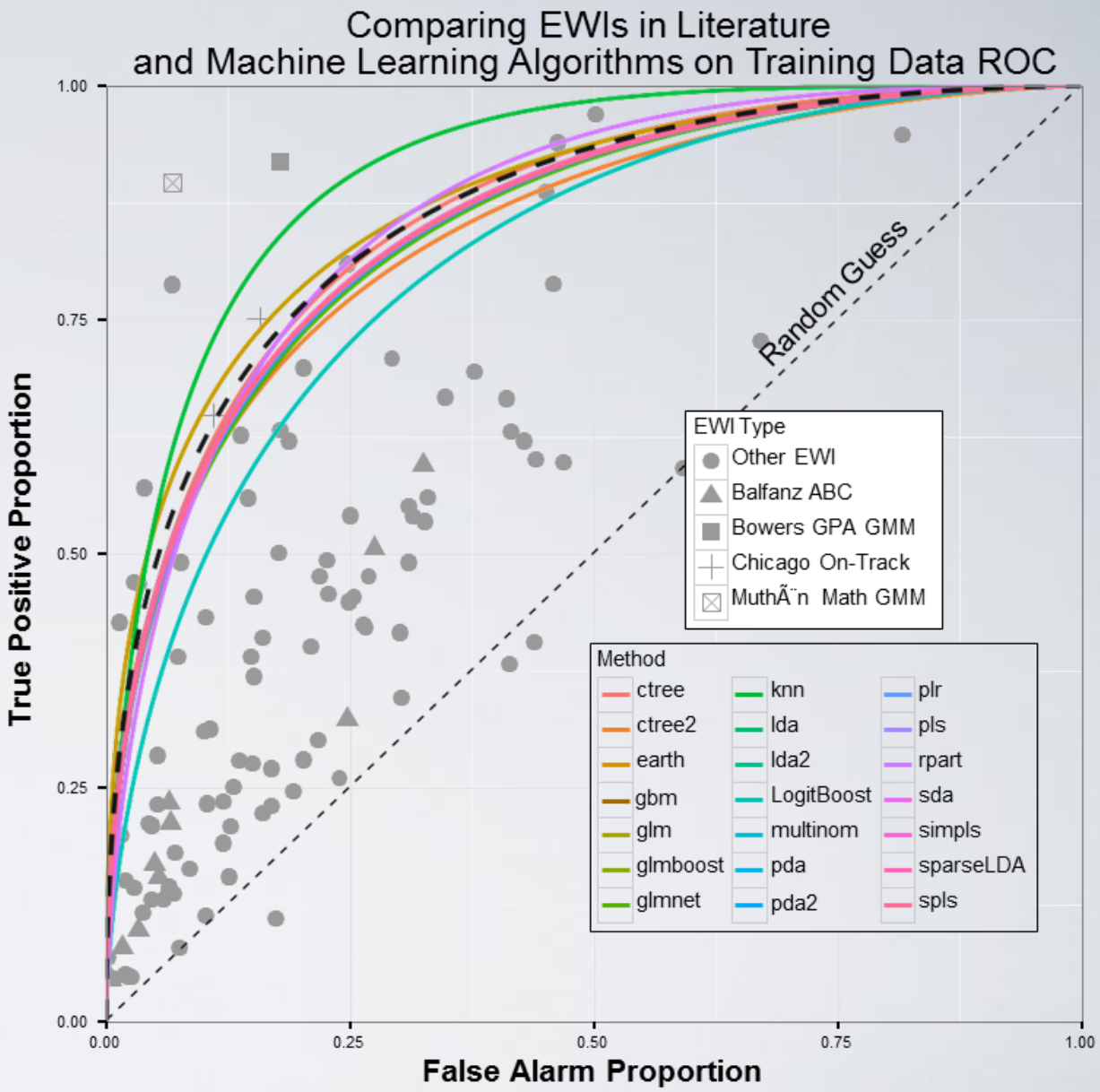
Grid of diagnostic plot plus summary output for the ensemble of 3 models

Source: Output from R (Iris dataset)

Application: predicting dropout risk for students

- Knowles (2015) implements **caretEnsemble** to provide additional predictive power in the Wisconsin Dropout Early Warning System (DEWS)

- DEWS uses the receiver-operating characteristic (ROC) metric to identify the best possible set of statistical models for making predictions about individual students



ROCs for Machine Learning Algorithms Implemented on Training Data

Source: Knowles (2015)

Summary

Caret

streamlines the process of predictive modelling (aka machine learning)

CaretEnsemble

provides a handy set of commands to combine the predictions of multiple models

Ensembles

can be as simple as weighted averages or as complex as a secundary models

Accuracy

can be increased and overfit

can be hedged with ensembles models