caret Package: Things to Note

Vinh "MBALearnsToCode" Luong

`caret` is a popular R package that wraps around 200+ Machine Learning algorithms and data processing procedures.

This note concerns a number of important things to know when working with `caret`.

Let's first load some libraries and import the *Boston Housing* data set for illustration:

```r
library(caret)
library(data.table)
library(gbm)

boston_housing <- fread(
  'https://raw.githubusercontent.com/ChicagoBoothML/DATA___BostonHousing/master/BostonHousing.csv')

# sort data table by 'lstat'
setkey(boston_housing, lstat)
```

Always use `x=model.matrix(...), y=...`, *NOT* `formula=..., data=...`

There are 2 conventions for specifying models for training in R:

1. The convenient `formula=..., data=...`; and
2. The explicit `x=..., y=...`.

Convention #1 is generally more convenient to write. However, its use is **STRONGLY DISCOURAGED** when working with large data sets because it invokes repeated calls to a `model.frame` function to expand the covariates in the formula to a proper model structure, resulting in slow performance. The explicit convention #2 is better because all the data have been prepared properly before being passed into the training procedure.

More importantly, `caret` being an inteface for high-performance Machine Learning often calls the underlying algorithms (e.g. GBM) using the high-performance convention #2 (`x=..., y=...`). If users use convention #1 (`formula=..., data=...`), this mismatch can result in non-obvious, buggy behaviors.

When using the preferred `x=..., y=...`, always explicitly expand the X matrix using **`model.matrix`** before passing it into the `train` and `predict` functions.

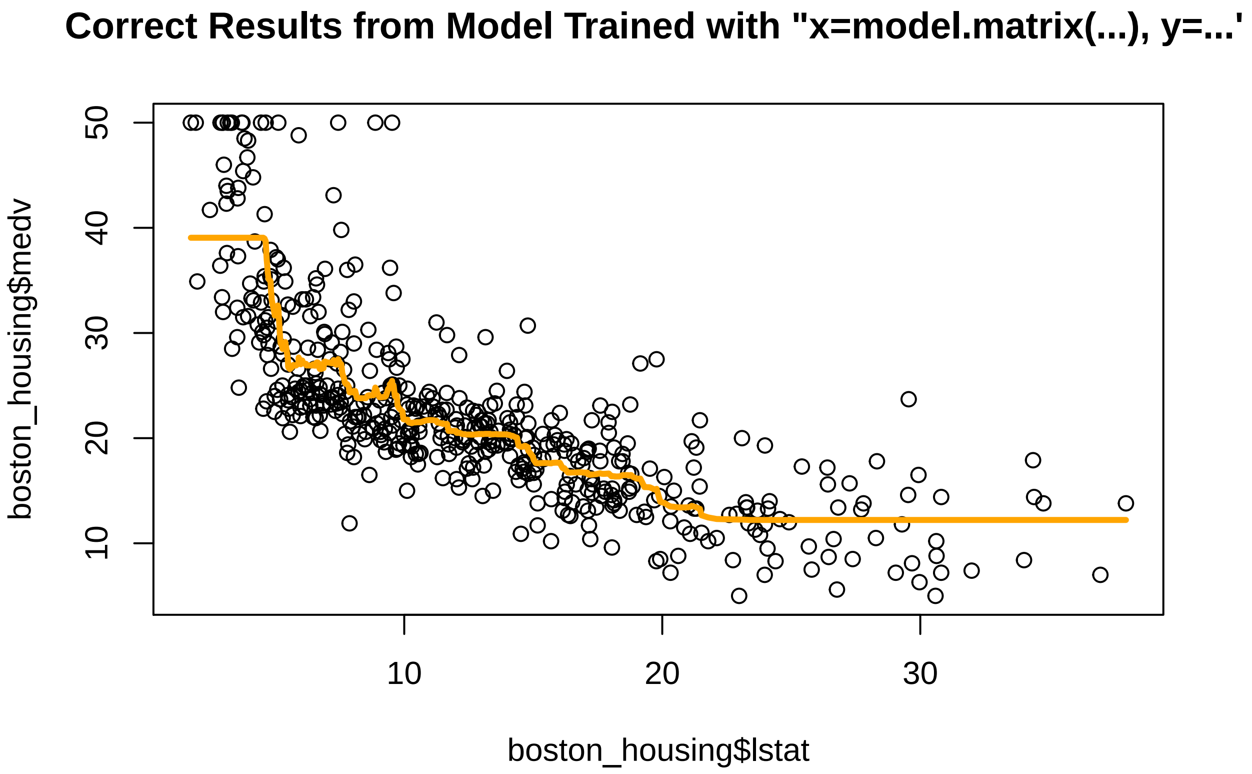
Below is code illustrating correct, recommended usage:

```r
model_matrix_formula <- ~ -1 + lstat   # want X with lstat and NO INTERCEPT

boost_model <- train(
  x=model.matrix(model_matrix_formula, data=boston_housing),   # explicitly-expanded X
  y=boston_housing$medv,                                        # explicit y vector
  method='gbm',
  verbose=FALSE,
  trControl=trainControl(
    method='repeatedcv',   # Repeated Cross-Validation
    number=5,              # 5 Folds
    repeats=3,             # 3 Repeats
  ),
  tuneGrid=expand.grid(
    n.trees=1000,
    interaction.depth=2,
    n.minobsinnode=30,
    shrinkage=.01
  ))
```

```
boost_pred <- predict(
  boost_model,
  newdata=model.matrix(
    model_matrix_formula,
    data=boston_housing))   # at Prediction time, also explicitly expand X

plot(x=boston_housing$lstat, y=boston_housing$medv)
lines(x=boston_housing$lstat, y=boost_pred, col='orange', lwd=3)
title('Correct Results from Model Trained with "x=model.matrix(...), y=..."')
```


Correct Results from Model Trained with "x=model.matrix(...), y=...'
boston_housing$medv
boston_housing$lstat