Caret / Recursive Partitioning

Ryan Guzalowski

5/16/17

Exercise 1: caret/logistic regression (5 points)

Rebuild your logistic regression model from the previous week, this time using the `caret` package.

- Calculate the training or apparent performance of the model.
- Calculate an unbiased measure of performance
- Create a ROC Curve for your model

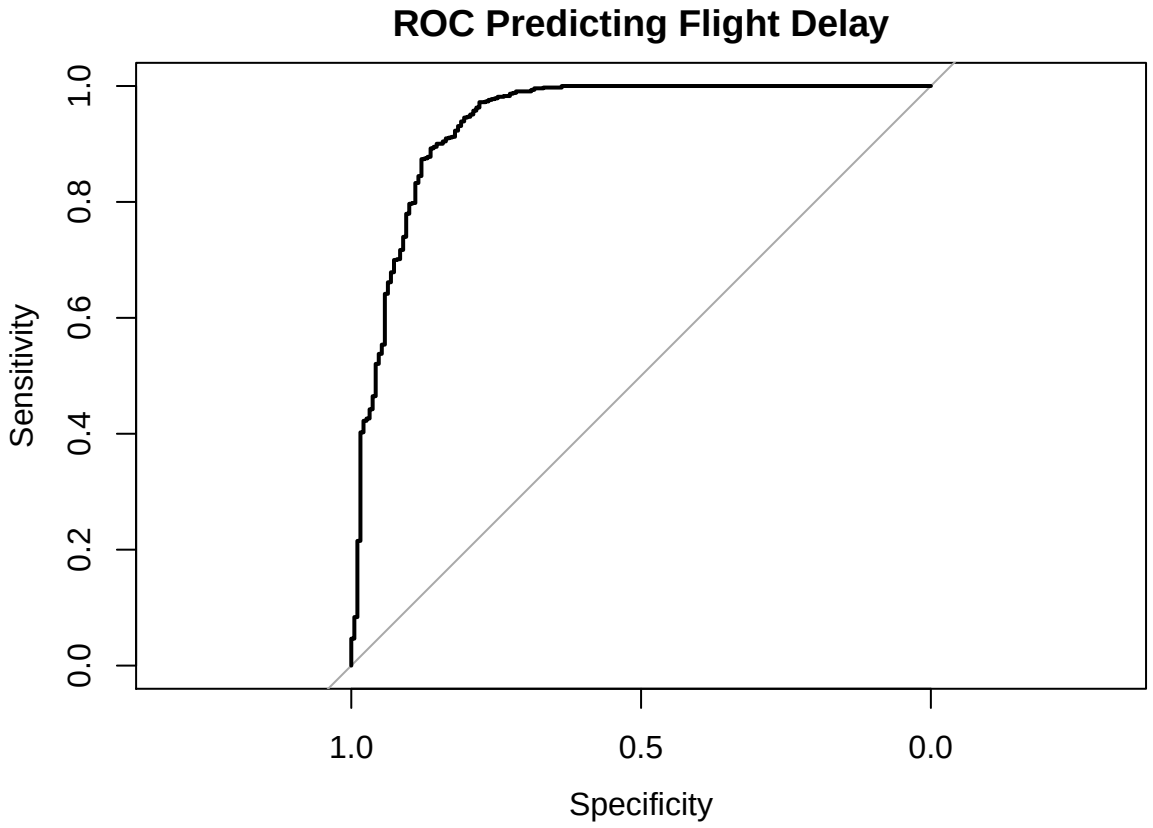
Show all work.

```
YX <- readRDS( "YX.rds")
predictors <- c("dep_delay", "month", "air_time", "distance", "carrier", "lat.dest", "lon.dest")
response <- c("late")
yx <- subset(YX, select = c(response, predictors)) %>% sample_n(1000) %>% na.omit()
yx$carrier <- yx$carrier %>% as.factor()
yx$late <- yx$late %>% as.character()
yx$late[yx$late=="TRUE"] <- "late"
yx$late[yx$late=="FALSE"] <- "notLate"
yx$late <- yx$late %>% as.factor()
form <- late ~ dep_delay + month + air_time + distance + carrier + lat.dest + lon.dest

ctrlLD <- trainControl(method = "cv", number = 10,
                       savePredictions = TRUE,
                       classProbs = TRUE)

fitLD <- train(form, data = yx,
               method = "glm", family = "binomial",
               trControl = ctrlLD,
               na.action = na.omit)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
y <- fitLD$pred$obs
yhat <- fitLD$pred$late
rocCurve <- roc(y, yhat)

plot(rocCurve, main = "ROC Predicting Flight Delay")
```



Exercise 2: caret/rpart (5 points)

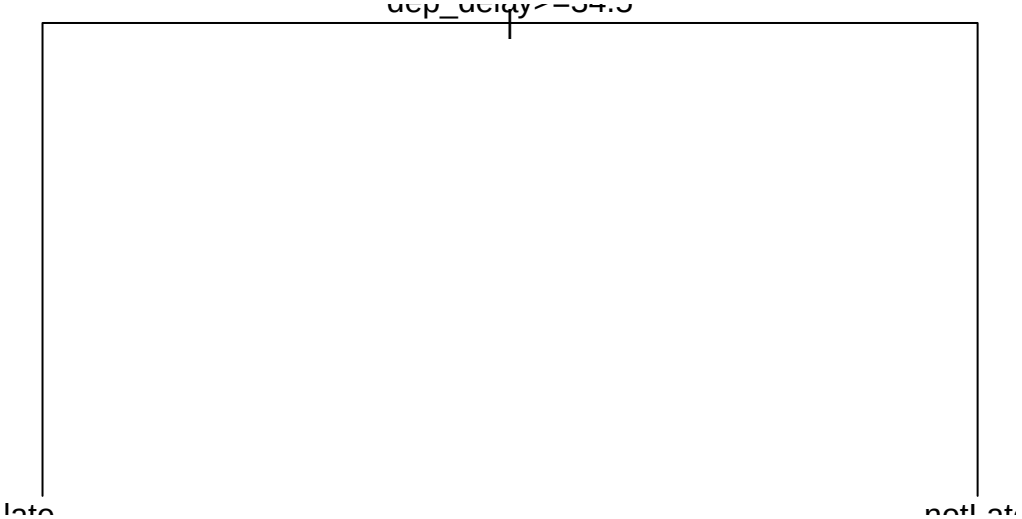
Using the `caret` and `rpart` packages, create a **classification** model for flight delays using your NYC FLight data. Your solution should include:

- The use of `caret` and `rpart` to train a model.
- An articulation of the the problem your are
- An naive model
- An unbiased calculation of the performance metric
- A plot of your model – (the actual tree; there are several ways to do this)
- A discussion of your model

Show and describe all work

Articulation of the Problem: - dlsjfklds

```r
# Use of 'caret' and 'rpart' to train a model:
fitCART <- train(form, data = yx,
                 method = "rpart")


# Plot of the model:
plot(fitCART$finalModel)
text(fitCART$finalModel)
```



Questions:

- Discuss the difference between the models and why you would use one model over the other?
- How might you produce an ROC type curve for the *rpart* model?