SmartFly: Prepare Data For Modeling

Cindy Lamm

January 12, 2015

Load preprocessed data from the previous step "Exploratory Data Analysis"

```
rm(list=ls())   #clear memory
load("../01_exploratory_data_analysis/trainDataTyped.Rdata")
```

Split the train data based on simple bootstrap resampling into a series of train and test sets

```
library(caret)
set.seed(998)
#use simple bootstrap resampling to split data into a series of train and test set
inTraining <- createDataPartition(trainDataTyped$is_delayed, p = .5, list = FALSE)
training <- trainDataTyped[ inTraining,]
testing  <- trainDataTyped[-inTraining,]
```

Create custom function to specifiy the type of resampling and a grid for tuning parameters[1]

```
cctrl4 <- trainControl(method = "cv", number = 3, classProbs = TRUE)
eGrid <- expand.grid(.alpha = seq(.05, 1, length = 15), .lambda = c((1:5)/10))
```

Save R environment

```
save.image(file="prepared_data.Rdata")
```

[1] both taken from `https://github.com/topepo/caret/blob/master/RegressionTests/Code/glmnet.R`