**## 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*. I created two models and compare them each other. P value is 0.3691. My second model didn’t improve anything. I used the first model I created.*

- Calculate an unbiased measure of performance*. Misclassification Rate is 0.9586813 (with delay variable)*

- Create a ROC Curve for your model. *See below.*

|  |
| --- |
| > library(rpart)  > library(caret)  **# Training**  > data1$arr\_delay22 <- ifelse(data1$arr\_delay > 22,1,0)  > data1$arr\_delay221<- as.factor(data1$arr\_delay22)  > sample1<-data1[sample(nrow(data1), 10000), ]  > Train <- createDataPartition(sample1$arr\_delay22 , p=0.7, list=FALSE)  > training <- sample1[ Train, ]  > testing <- sample1[ -Train, ]  > mod\_fit <- train( as.factor(arr\_delay22) ~ season+dep\_delay +  + flight+ origin+ dep\_delay+ air\_time + distance + seats, data=training, method="glm", family="binomial")  >  > summary(mod\_fit)  Call:  NULL  Deviance Residuals:  Min 1Q Median 3Q Max  -2.4188 -0.2766 -0.1580 -0.0731 3.4354  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -5.564e+00 3.004e-01 -18.520 < 2e-16 \*\*\*  season2 3.957e-01 1.531e-01 2.585 0.00974 \*\*  season3 1.168e+00 1.758e-01 6.643 3.07e-11 \*\*\*  season4 3.973e-02 1.675e-01 0.237 0.81252  season5 7.436e-01 2.360e-01 3.151 0.00163 \*\*  dep\_delay 1.258e-01 3.784e-03 33.250 < 2e-16 \*\*\*  flight -4.368e-05 4.513e-05 -0.968 0.33304  originJFK 6.369e-02 1.350e-01 0.472 0.63697  originLGA 9.423e-02 1.435e-01 0.657 0.51131  air\_time 9.942e-02 5.003e-03 19.872 < 2e-16 \*\*\*  distance -1.298e-02 6.660e-04 -19.487 < 2e-16 \*\*\*  seats 9.261e-04 1.029e-03 0.900 0.36808  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 6795.5 on 6999 degrees of freedom  Residual deviance: 2394.8 on 6988 degrees of freedom  AIC: 2418.8  Number of Fisher Scoring iterations: 7  > exp(coef(mod\_fit$finalModel))  (Intercept) season2 season3 season4 season5 dep\_delay flight originJFK  0.003834646 1.485453233 3.216105230 1.040531488 2.103502812 1.134082754 0.999956318 1.065764962  originLGA air\_time distance seats  1.098808068 1.104533389 0.987105224 1.000926550  > testp<-predict(mod\_fit,testing,type='prob')[,2]  > pred<-prediction(testp,testing$arr\_delay22)  > perf <-performance(pred,"tpr","fpr")  >  > #performance  > auc.perf = performance(pred, measure = "auc")  > auc.perf@y.values  [[1]]  [1] 0.9586813 |
|  |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | |  | | --- | | **> #Evaluation the performance**  > mod\_fit\_one <- glm(arr\_delay22 ~ season+dep\_delay +  + flight+ origin+ dep\_delay+ air\_time + distance + seats, data=training, family="binomial")  > mod\_fit\_two <- glm(arr\_delay22 ~ season+dep\_delay +  + flight+ origin+ dep\_delay+ air\_time + distance , data=training, family="binomial")  >  > anova(mod\_fit\_one, mod\_fit\_two, test ="Chisq")  Analysis of Deviance Table  Model 1: arr\_delay22 ~ season + dep\_delay + flight + origin + dep\_delay +  air\_time + distance + seats  Model 2: arr\_delay22 ~ season + dep\_delay + flight + origin + dep\_delay +  air\_time + distance  Resid. Df Resid. Dev Df Deviance Pr(>Chi)  1 6988 2394.8  2 6989 2395.6 -1 -0.80667 0.3691  >  > library(lmtest)  > lrtest(mod\_fit\_one, mod\_fit\_two)  Likelihood ratio test  Model 1: arr\_delay22 ~ season + dep\_delay + flight + origin + dep\_delay +  air\_time + distance + seats  Model 2: arr\_delay22 ~ season + dep\_delay + flight + origin + dep\_delay +  air\_time + distance  #Df LogLik Df Chisq Pr(>Chisq)  1 12 -1197.4  2 11 -1197.8 -1 0.8067 0.3691  **#ROC Curve**  library(ROCR)  library(pROC)  plot(perf,col="green",lwd=2,main="ROC Curve for Logistic ")  abline(a=0,b=1,lwd=2,lty=2,col="gray")    **## 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  **#Creating a Tree**  training$arr\_delay222 <- factor(training$arr\_delay22,levels=c(1,0),  labels=c("YES","NO"))  library(party)  tree<-ctree(arr\_delay222 ~ season+dep\_delay +  flight+ dep\_delay+ air\_time + distance,data=training, control = ctree\_control(mincriterion = 0.90,maxdepth=10))  plot(tree) | |  | | |  | | --- | |  | |   treepred <- as.data.frame(do.call("rbind", treeresponse(tree, newdata = testing)))  testing$tscore<-treepred[,1]  treepred<-prediction(testing$tscore,testing$arr\_delay22)  treeperf <- performance(treepred,"tpr","fpr")  library(ROCR)  library(pROC)  plot(treeperf,col="green",lwd=2,main="ROC Curve for TREE ")  abline(a=0,b=1,lwd=2,lty=2,col="gray")  plot(treeperf,col='red',lwd=2, lty=1,main='ROC-Logistic vs Tree');  plot(perf, col='blue',lwd=2, add=TRUE,lty=2);  legend(0.6,0.6,c('Tree','Logistic'),col=c('red','blue'),lwd=2)  abline(a=0,b=1,lwd=2,lty=4,col="gray") | |

*I used Logistic and Tree to compare the performance. Logistic had better performance than the tree.*