Celijah_ML3

1. Divide the data into 60% training and 40% validation

Response:

```r
library(pROC)
library("gmodels")
library(caret)
library(ISLR)
#install.packages("e1071") #install first
library(e1071)
library(naivebayes)
library(dplyr)
library(ggplot2)
library(psych)
a3 <- read.csv("FlightDelays.csv")
summary(a3)
#Convert week and time variables to factors
a3$time <- as.factor(a3$DEP_TIME)
a3$day <- as.factor(a3$DAY_WEEK)
colnames(a3)
#Select the 5 variables plus Flight.Status
newdata<-a3[,c(1,2,8,13,14,15)]
str(newdata)
```

```r
```{r}

#1.  Divide the data into 60% training and 40% validation
set.seed(123)
Index_Train<-createDataPartition(newdata$Flight.Status, p=0.6, list=FALSE)
Train <-newdata[Index_Train,]
Train
validate  <-newdata[-Index_Train,]
validate
```

data.frame 1321 x 6

data.frame 880 x 6

	CRS_DEP_TIME <int>	CARRIER <fctr>	ORIGIN <fctr>	Flight.Status <fctr>	time <fctr>	day <fctr>
3	1245	DH	IAD	ontime	1245	4
5	1039	DH	IAD	ontime	1035	4
7	1240	DH	IAD	ontime	1243	4
13	930	DL	DCA	ontime	932	4
15	1430	DL	DCA	ontime	1429	4
21	900	MQ	DCA	ontime	853	4
22	1300	MQ	DCA	ontime	1254	4
23	1400	MQ	DCA	ontime	1356	4
25	1900	MQ	DCA	ontime	1853	4
27	900	US	DCA	ontime	858	4

1-10 of 880 rows   Previous 1 2 3 4 5 6 ... 88 Next
```

data.frame 1321 x 6

data.frame 880 x 6

	CRS_DEP_TIME <int>	CARRIER <fctr>	ORIGIN <fctr>	Flight.Status <fctr>	time <fctr>	day <fctr>
1	1455	OH	BWI	ontime	1455	4
2	1640	DH	DCA	ontime	1640	4
4	1715	DH	IAD	ontime	1709	4
6	840	DH	IAD	ontime	839	4
8	1645	DH	IAD	ontime	1644	4
9	1715	DH	IAD	ontime	1710	4
10	2120	DH	IAD	ontime	2129	4
11	2120	DH	IAD	ontime	2114	4
12	1455	DL	DCA	ontime	1458	4
14	1230	DL	DCA	ontime	1228	4

1-10 of 1,321 rows Previous 1 2 3 4 5 6 ... 100 Next

2. Run the Naive Bayes model to predict whether the flight is delayed or not. Use only categorical variables for the predictor variables. Note that Week and Time variables need to recoded as factors

Response:

~~~~
```{r}
#2. Run the Naive Bayes model to Predict whether flight delayed or not. Use only categorical variables for the predictor variables. Note that Week and Time variables remains as factors
#First remove Non Categorical Variables
New_model <- subset(a3,select = -c(CRS_DEP_TIME,DEP_TIME, DISTANCE, FL_DATE, FL_NUM, Weather))
New_model$DAY_WEEK <- as.factor(New_model$DAY_WEEK)
New_model$DAY_OF_MONTH <- as.factor(New_model$DAY_OF_MONTH)
New_model <- transform(New_model, Flight.Status = ifelse(New_model$Flight.Status == "delayed", 1, 0))
New_model
~~~~

~~~~
```{r}
#Q2 Continues: Defining the Naive Bayes Model

set.seed(123)
nb_model <-naiveBayes(Flight.Status~.,data = Train, usekernel = T)
Predicted_validate_labels <-predict(nb_model,validate, type = "raw")
Predicted_validate_labels <- transform(Predicted_validate_labels, Prediction = ifelse(Predicted_validate_labels[,2] > .5, 1, 0))
Predicted_validate_labels
```
~~~~

delayed <dbl>	ontime <dbl>	Prediction <dbl>
0.155806142	0.84419386	1
0.506574170	0.49342583	0
0.084015320	0.91598468	1
0.056629120	0.94337088	1
0.036673116	0.96332688	1
0.063761208	0.93623879	1
0.072968184	0.92703182	1
0.031888507	0.96811149	1
0.266115219	0.73388478	1
0.003896570	0.99610343	1

1–10 of 880 rows

3. Output both a counts table and a proportion table outlining how many and what proportion of flights were delayed and on-time at each of the three airports.

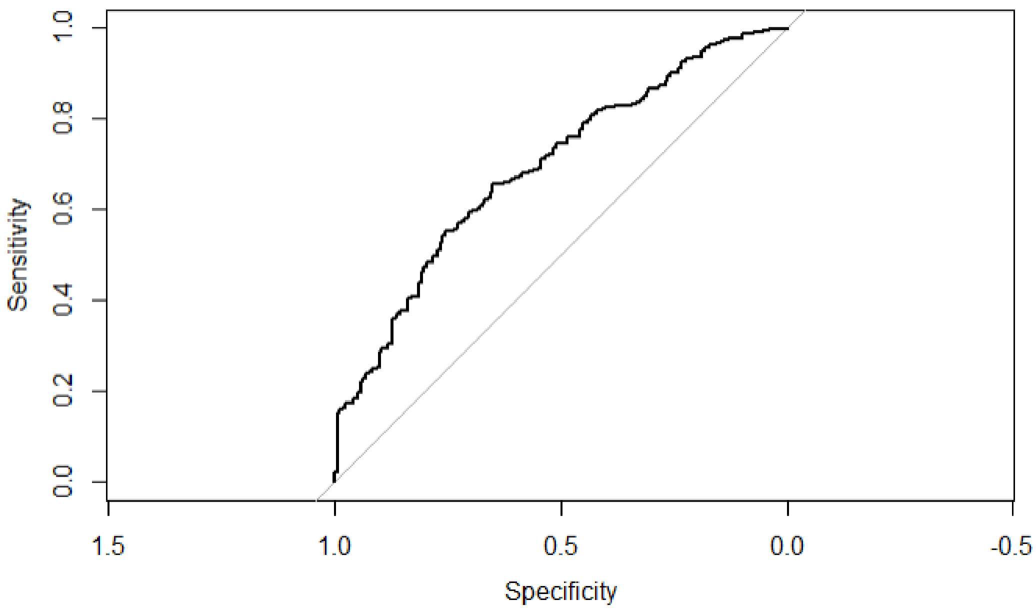
Response:

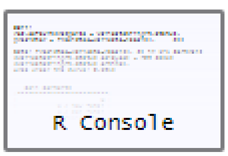
```
```{r}
#3. Output both a counts table and a proportion table outlining how many and what proportion of flights were delayed and on-time at each of the three airports.

#Creating the Counts and Proporation Table

Table <- a3 %>%
    group_by( DEST ) %>%
    summarise( count = n(), proportion = n()/length(a3$DEST) )

Table

```
```

DEST <fctr>	count <int>	proportion <dbl>
EWR	665	0.3021354
JFK	386	0.1753748
LGA	1150	0.5224898

3 rows

4. Output the confusion matrix and ROC for the validation data

```
```{r}
#4.Output the confusion matrix and ROC for the validation data

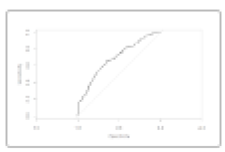
roc(validate$Flight.Status, Predicted_validate_labels[,2])

plot.roc(validate$Flight.Status,Predicted_validate_labels[,2])

CrossTable(x=validate$Flight.Status,y=Predicted_validate_labels[,3], prop.chisq = FALSE)
```
```



```
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  880


                       | Predicted_validate_labels[, 3]
validate$Flight.Status |           0 |           1 | Row Total |
-----------------------|-------------|-------------|-----------|
               delayed |          35 |         136 |       171 |
                       |       0.205 |       0.795 |     0.194 |
                       |       0.438 |       0.170 |           |
                       |       0.040 |       0.155 |           |
-----------------------|-------------|-------------|-----------|
                ontime |          45 |         664 |       709 |
                       |       0.063 |       0.937 |     0.806 |
                       |       0.562 |       0.830 |           |
                       |       0.051 |       0.755 |           |
-----------------------|-------------|-------------|-----------|
          Column Total |          80 |         800 |       880 |
                       |       0.091 |       0.909 |           |
-----------------------|-------------|-------------|-----------|
```