Author: Atmin Sheth Andrew Sen

Date 09-23-2022

Data

The note book is acquired on kragle dataset called Airlines showing the airline and their flights being delay(1) or not (0)

The predictor include: * id: int (unique number) * Airline: factor of char * Flight: int (unique number) * AirpotFrom: char * AirportTo: char * DayofWeek: int(constant) * Time: int * Delay: boolean Delay(1) not delay(0) targeted predictor: * Airline * Flight * Delay

overview of Logistic regression Logistic Regression allows to classify the data into n- parts to determine the section the factors belong and predict upon that for this dataset is a classification of if the airlines will have delay or not , flight is to see a corelatrion wiht airline

strengths of logistic The benefit of using logistic reggresuib model let's you separate the classes and see a distinct differences, it is easy computation where the result is in probability bases on the predictor and and relation of probability or mean to other predictors. The output is easy to comprehend.

weakness of logistic there is a chance of too close of probability so not able to classify due to being clustered. Not able to sometimes make a model line.

Strength of naive bayes Naive bayes is easy to implement and interpret. It works well with small data. it also works well with high dimension

weakness of naive bayes Naive bayes tend to outperform in larger set.there are many guesses that are not sometimes made in a train dataset.there is a limitation if the predictors are not independent.

the evaluation The benefit of using this was see the probability of a airline being delay or not. **logistic vs naiv bais** for this data there is more information given through logistic regression gml and in naive bayes. In naive bayes it shows the mean of delay with each factor but does not give a good prediction. with logistic regression it seem a better accuracy is beter and naive bayes but the accuracy sits the best in the roc function.

reflection of classification matrix The matrices shows the pobability of the each predictor being delay or not. i the predictors we have, the flights predictor was least usefu as all the factors were unique so cannot give any estimation of future garuntee. where we can evaluate on the airlines for a better prediction. So the linear graph for tpr an fpr is accounted for both predictor and a mix evaluation of delay or not.

##Reading csv file want find the difference of delay and not delay the airlines that are tend to be delay

```
flights <- read.csv("Data/Airlines.csv", header = TRUE)
str(flights)
```

```
## 'data.frame':    539383 obs. of  9 variables:
##  $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Airline    : chr  "CO" "US" "AA" "AA" ...
##  $ Flight     : int  269 1558 2400 2466 108 1094 1768 2722 2606 2538 ...
##  $ AirportFrom: chr  "SFO" "PHX" "LAX" "SFO" ...
##  $ AirportTo  : chr  "IAH" "CLT" "DFW" "DFW" ...
##  $ DayOfWeek  : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Time       : int  15 15 20 20 30 30 30 30 35 40 ...
##  $ Length     : int  205 222 165 195 202 181 220 228 216 200 ...
##  $ Delay      : int  1 1 1 1 0 1 0 0 1 1 ...
```

###data cleean getting the data we want to work with that be airlines,flight and delay

```
flights <- flights[,c(2,3,9)]
flights$Airline <- as.numeric(factor(flights$Airline))
head(flights)
```

	Airline <dbl>	Flight <int>	Delay <int>
1	5	269	1
2	15	1558	1
3	2	2400	1
4	2	2466	1
5	3	108	0
6	5	1094	1

6 rows

##Data Exploration spliting the data to training and testinng in a ration of 80/20(.4) train/test

```
set.seed(3)
i <- sample(1:nrow(flights), .4*nrow(flights),replace=FALSE)
train <- flights[-i,]
test <- flights[i,]
```

```
summary(train)
```

```
##     Airline          Flight         Delay
##  Min.   : 1.00   Min.   :   1   Min.   :0.0000
##  1st Qu.: 6.00   1st Qu.: 710   1st Qu.:0.0000
##  Median :11.00   Median :1809   Median :0.0000
##  Mean   :10.34   Mean   :2428   Mean   :0.4456
##  3rd Qu.:16.00   3rd Qu.:3742   3rd Qu.:1.0000
##  Max.   :18.00   Max.   :7814   Max.   :1.0000
```

seeing the head of the train set

```
head(train)
```

	Airline <dbl>	Flight <int>	Delay <int>
2	15	1558	1
3	2	2400	1
4	2	2466	1
5	3	108	0
7	6	1768	0
9	6	2606	1

6 rows

the end of the train data set

```
tail(train)
```

	Airline <dbl>	Flight <int>	Delay <int>
539376	9	58	0
539377	4	717	1
539379	5	178	0
539380	9	398	0
539381	9	609	0
539383	15	1442	1

6 rows

Seeing how many airplines we are working with in training

dimention

```
dim(train)
```

```
## [1] 323630      3
```

```
table(train$Airline)
```

```
##
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 12379 27319  6930 10775 12630 36582 16696  3894 12531  3345 22091  7569 30178
##    14    15    16    17    18
## 16609 20732 56511 18681  8178
```

Seeing the sum of delay flight how many of the files in the train dataset was delaye

```
sum(train$Delay)
```

```
## [1] 144198
```

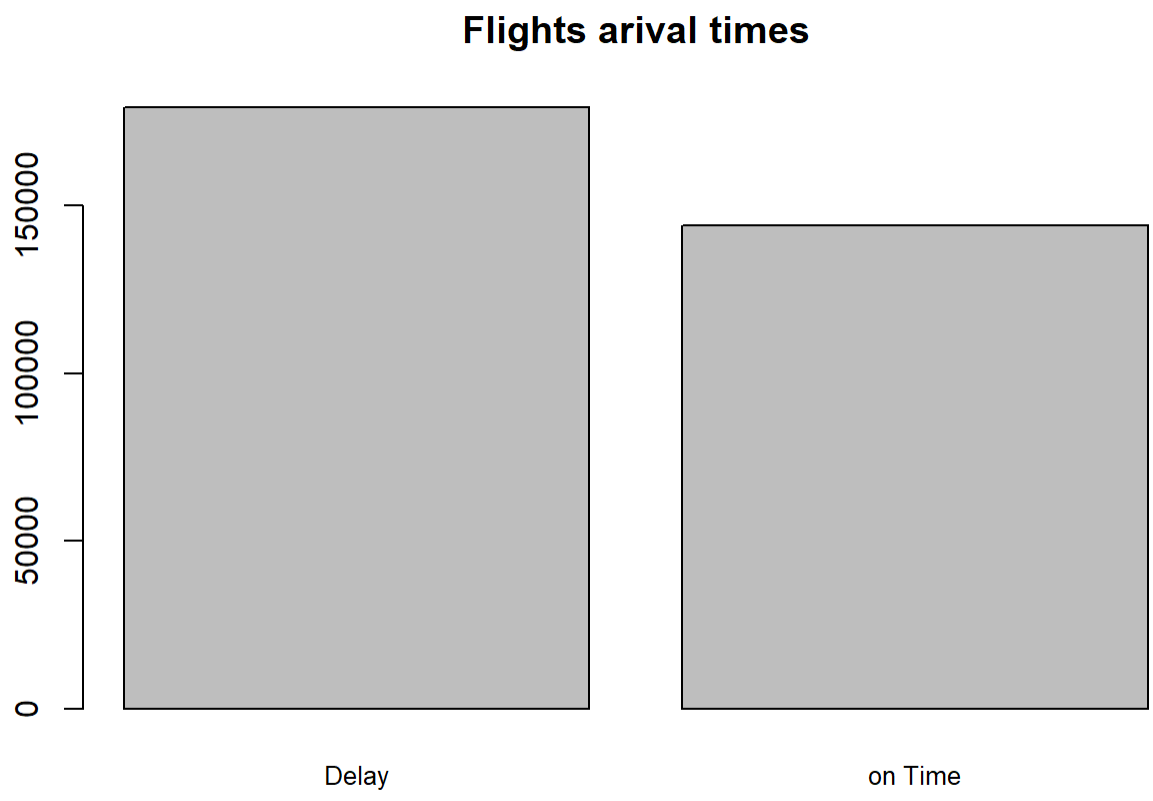
where are the flight most going to

```
dim(train )
```

```
## [1] 323630      3
```

bar graph to show the delay vs non delay in all of the airline in train data set

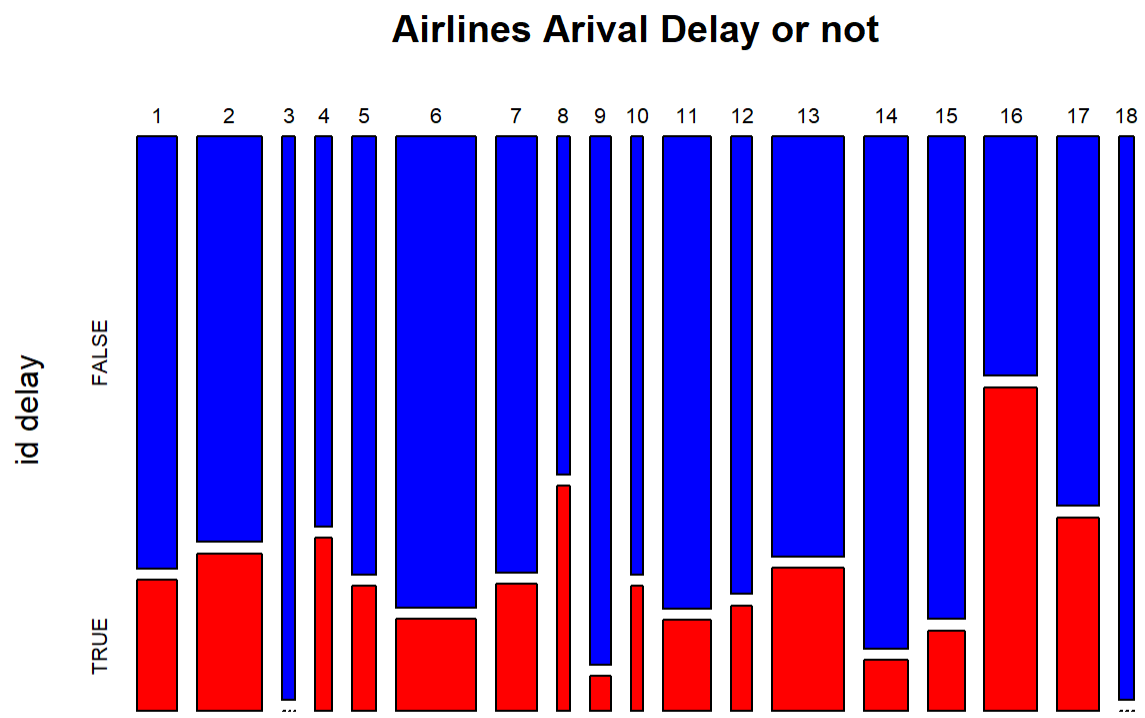
```
count <- table(train$Delay)
bp<- barplot(count,main="Flights arival times", names.arg=c("Delay", "on Time"),cex.names=.8)
```



There are more delay in overall flights then reaching in time

let's see a airlines delay in a subset of about 100 in train

```
sub<-train[1:500,]
T<-table(sub$Airline,(sub$Delay==1))
plot(T, ylab="id delay", col= c("blue","red"), main="Airlines Arival Delay or not")
```



in the subset it can see that there are airlines, w

##logistic regression predictor

```
glm1 <- glm(Delay~. , family=binomial, data=train)
summary(glm1)
```

```
##
## Call:
## glm(formula = Delay ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2044  -1.0722  -0.9952   1.2534   1.4306
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.759e-01  8.582e-03   -43.8   <2e-16 ***
## Airline      2.730e-02  6.741e-04    40.5   <2e-16 ***
## Flight      -5.213e-05  1.732e-06   -30.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 444803  on 323629  degrees of freedom
## Residual deviance: 442452  on 323627  degrees of freedom
## AIC: 442458
##
## Number of Fisher Scoring iterations: 4
```

It took 4 itteration to get the results There is a drop from the Null deviance to the Residual deviance indicating a good prediction for all two predict or for airline and flight there is a good p value of less the 2e^-16 The z value coming to be 40.5 for airline and -30.1 for flight

##naive basis

```
library(e1071)
nb1<- naiveBayes(Delay~.,data=train)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##         0         1
## 0.5544356 0.4455644
##
## Conditional probabilities:
##    Airline
## Y       [,1]     [,2]
##   0 10.02369 5.257896
##   1 10.73411 5.332890
##
##    Flight
## Y       [,1]     [,2]
##   0 2513.890 2101.811
##   1 2320.616 2020.345
```

here we can see a mean of airlines and flights the likelihood of airline and flights being delay

#model testing

```
probs <- predict(glm1, newdata = test, type="response")
preds <- ifelse(probs==0,0,1)
acc1<- mean(preds==test$Delay)
print(paste("glm1 accuracy =",acc1))
```

```
## [1] "glm1 accuracy = 0.445259162097398"
```

```
table(preds,test$Delay)
```

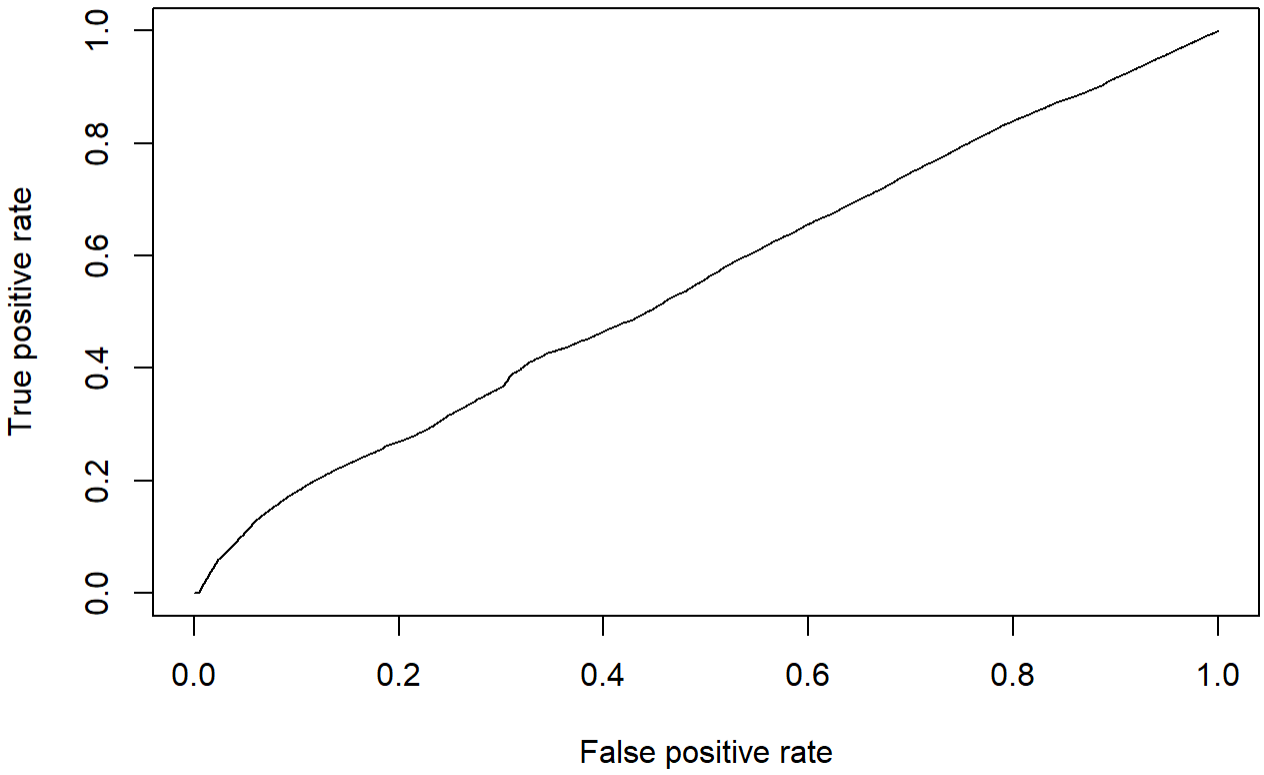
```
##
## preds      0      1
##     1 119687  96066
```

From this models of test data it can be seen that the rate of being late is low the matrix shows the true positive to the reference of being delay. there is a higher flase positive in the data showign a less likely of being delay the mean shows the rate being .58

#addition add

```
library(ROCR)
p<-predict(glm1, newdata = test,test="response")
pr<- prediction(p,test$Delay)
prf <- performance(pr,measure="tpr", x.measure="fpr")

plot((prf))
```



There is a relative growth in the rate or true positive and false positive

```
auc <- performance(pr, measure="auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.5519707
```

the auc is .55,

```
print("prediction of naive bais")
```

```
## [1] "prediction of naive bais"
```

```
predN <- predict(nb1,newdata = test, type="class")
table(predN,test$Delay)
```

```
##
## predN      0      1
##     0 107857  78115
##     1  11830  17951
```

```
print("mean of delays in the naive bais ")
```

```
## [1] "mean of delays in the naive bais "
```

```
mean(predN==test$Delay)
```

```
## [1] 0.5831112
```

the matrix shows the likelihood of prediction being delay . keeping delay as a reference you can sees mean probability of being delay or not. There is a better accuracy in tbhe naive bais compare to logistic regression of .58