Flight Delay

The data set provided contains flight details for 5,819,811 flights for the different months of 2014. The features include categorical variables such as Month, Carrier, Origin Airport Id and Destination Airport Id. Ther are ordinal variables such as Day of month, Day of week and cancelled. There also are numeric variables such as Departure delay, carrier delay, weather delay and distance.

require(dplyr)

require(gbm)

load(file="2014flights.Rdata")

# Feature Engineering

We wish to build a model that would predict the delay time of a flight. Hence, we need remove all the cancelled flights as there cannot be any record of a delay linked to those flights. The carrier delay and weather delay have missing values for most of the observations so it s better to drop those features. Our response is the arrival delay as that would be the final delay for the flight. The arrival delay has a few missing values. There could be two possibilities for this. The first being that there was no delay and hence these were left blank, or the delay was not recored at all. The second case seems to be more likely as the flights that had no delay had a value of 0 and the flights that do not have the arrival delay recorded are also mssing the actual elapsed time and the air time. Hence I removed these observations as imputing zeroes here may have resulted in a weaker model.

df <- df %>% filter(CANCELLED==0)  
#df<-df[,-c(4,6,8,11,12,13,20,24,25,26,27,28,29)]  
df<-df[,-c(4,6,8,9,1,12,13,15,16,20,24,28,29)]  
df<-na.omit(df)  
format(object.size(df),units='MiB')

## [1] "477.5 MiB"

Y<- df$ARR\_DELAY  
X<- df[-13]

# Training and Test set

I split the data into training and testing set for the purpose of getting the estimate of the risk. I split the data randomly in the ration of 70:30.

smp\_size <- floor(0.7 \* nrow(df))  
  
set.seed(123)  
train\_ind <- sample(seq\_len(nrow(df)), size = smp\_size)  
  
X\_train <- X[train\_ind, ]  
X\_test <- X[-train\_ind, ]  
  
Y\_train <- Y[train\_ind]  
Y\_test <- Y[-train\_ind]

# Modelling

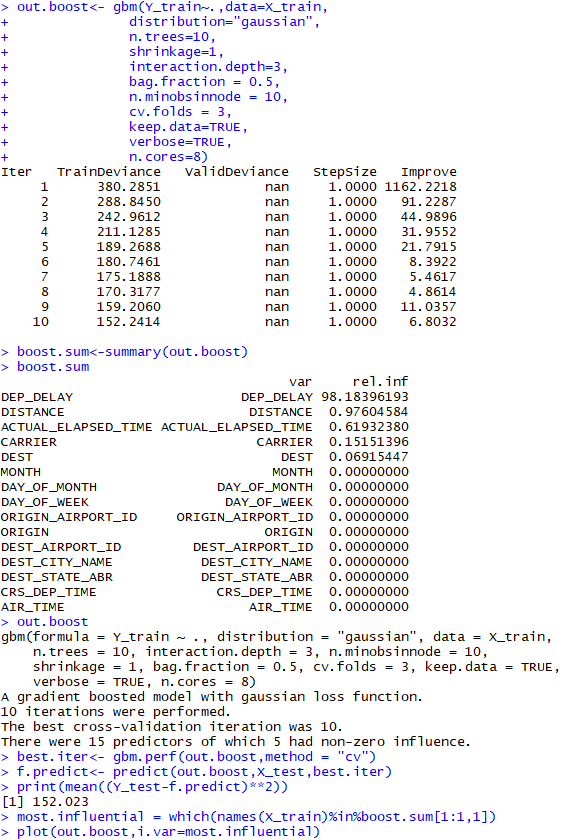
I chose the boosting method to generate a predictive model for the follwing reasons. 1] The model can handle more categories than any other modelling technique. 2] I can control the number of iterations easily. 3] The learning rate can be changed to create a model that learns slowly 4] It is easier to utilize all 8 cores of the machine to process in parallel.

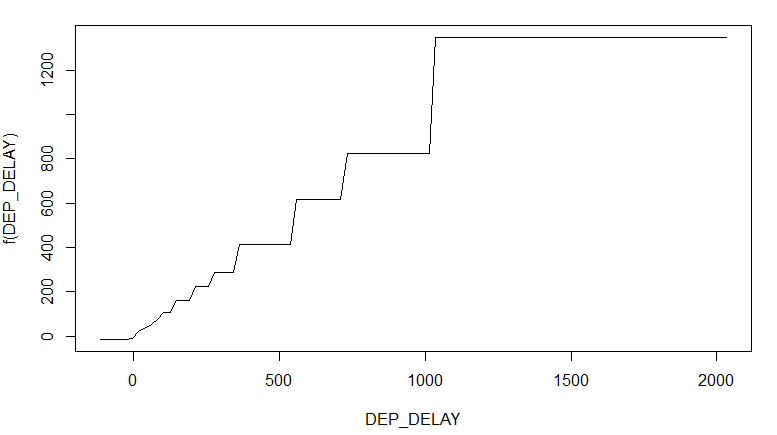
An ideal model would be one which would learn slowly over a long perod of time. In other words, a model which has a very low learning rate and a very high number of iterations. This is not possible as we are limited by the computing power available.

# Model 1

I started of with a very basic model. This model performs only 10 iterations with a very high learning rate of 1. I used the gaussian distribution in order to get the mean squared error. The model performs a 3 fold cross validation and utilizes all 8 cores of the machine.

At the end of 10 iterations, the model has a training error of 152.023 which is very high but considering the small amount of iterations, is not very surprising. The model provides the departure delay as the most influential feature. It is highly correlated with the arrival delay, which makes sense as the late a flight leaves its place of origin, the more the chances are that it will arrive late at its destination. The other influential variables are the distance, actual time elapsed, carrier and the destination. These are the features that influence the delay time of a flight. The model performs only as well on the test set as it did on the training set with a risk estimate of 152.023

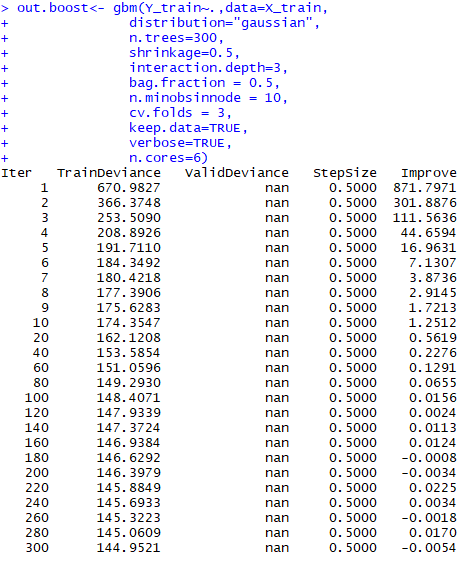


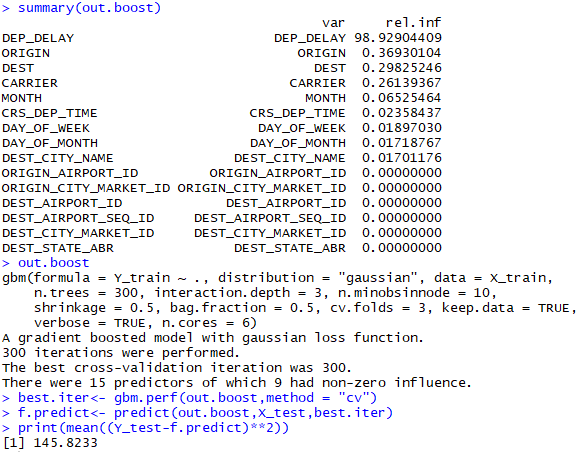


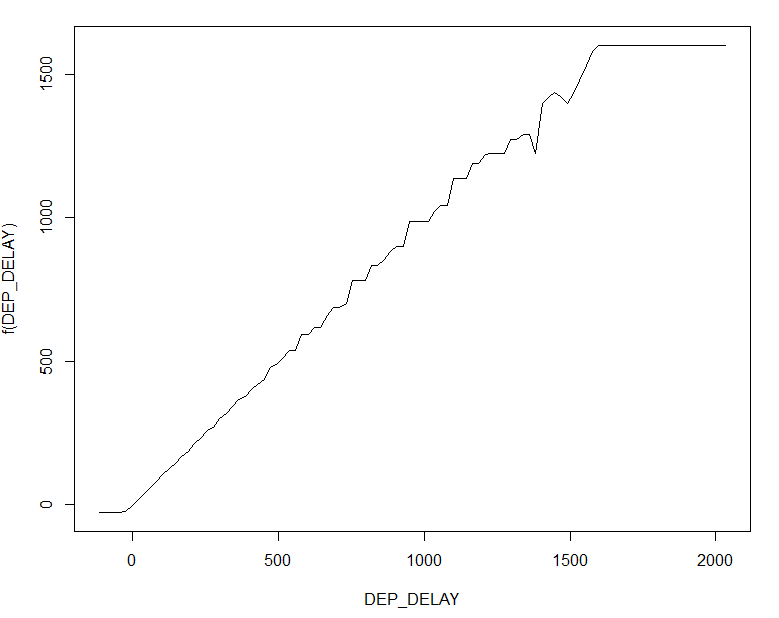
# Model 2

The second model I generated performed 300 iterations with a learning rate of 0.5. As expected, the model learns slower than the previous one but the higher number of iterations succesfully form a better model than before.

The training error for this model is 144.9521. The most influential feature is the departure delay once again. The other influencial feature include origin, destination, carrier and month. This model performs better on the test set as well with a estimated risk of 145.8233







# Model 3

This model is close to the kind of model we would look for through the boosting process. It performs 1000 iterations and learns extremely slow with a learning rate of only 0.01

The training error for this model is 108.2119. The most influential feature is the departure delay. Other important features include actual time elapsed, distance, carrier, origin and destination. The estimated risk via this model is 108.4275.

> out.boost<- gbm(Y\_train~.,data=X\_train,

+ distribution="gaussian",

+ n.trees=1000,

+ shrinkage=0.01,

+ interaction.depth=3,

+ bag.fraction = 0.5,

+ n.minobsinnode = 10,

+ cv.folds = 3,

+ keep.data=TRUE,

+ verbose=TRUE,

+ n.cores=8)

Iter TrainDeviance ValidDeviance StepSize Improve

900 113.8680 nan 0.0100 0.0519

920 112.5226 nan 0.0100 0.0921

940 111.4374 nan 0.0100 0.0756

960 110.4612 nan 0.0100 0.0558

980 109.3279 nan 0.0100 0.0505

1000 108.2119 nan 0.0100 0.0838

> boost.sum<-summary(out.boost)

> boost.sum

var rel.inf

DEP\_DELAY DEP\_DELAY 96.45393413

ACTUAL\_ELAPSED\_TIME ACTUAL\_ELAPSED\_TIME 1.94958108

DISTANCE DISTANCE 1.29600429

CARRIER CARRIER 0.17870909

ORIGIN ORIGIN 0.08555602

DEST DEST 0.03621540

MONTH MONTH 0.00000000

DAY\_OF\_MONTH DAY\_OF\_MONTH 0.00000000

DAY\_OF\_WEEK DAY\_OF\_WEEK 0.00000000

ORIGIN\_AIRPORT\_ID ORIGIN\_AIRPORT\_ID 0.00000000

DEST\_AIRPORT\_ID DEST\_AIRPORT\_ID 0.00000000

DEST\_CITY\_NAME DEST\_CITY\_NAME 0.00000000

DEST\_STATE\_ABR DEST\_STATE\_ABR 0.00000000

CRS\_DEP\_TIME CRS\_DEP\_TIME 0.00000000

AIR\_TIME AIR\_TIME 0.00000000

> best.iter<- gbm.perf(out.boost,method = "cv")

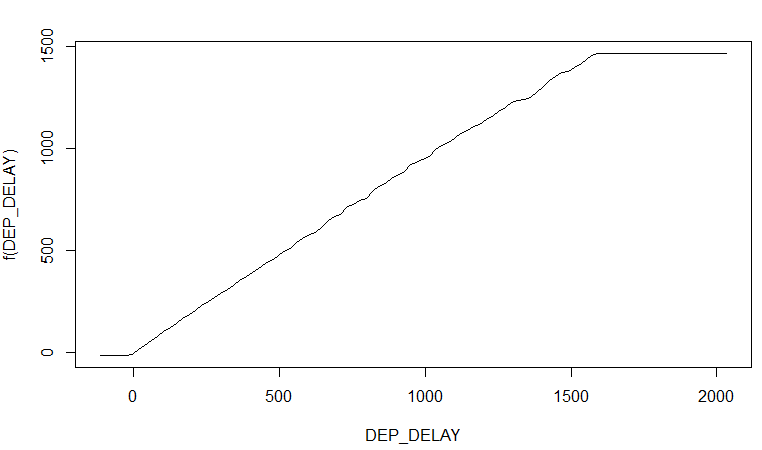
> f.predict<- predict(out.boost,X\_test,best.iter)

> print(mean((Y\_test-f.predict)\*\*2))

[1] 108.4275

> most.influential = which(names(X\_train)%in%boost.sum[1:1,1])

> plot(out.boost,i.var=most.influential)



So from the three models, we can say that departure delay, actual time elapsed and distance are the three most significant features in predicting the delay time of the flight.