Statistics 652: Statistical Learning and Prediction

Prediction of Departure Delay of Flights departing from NYC

Chidambaram Allada

APPENDIX Software Details : R-Version- 3.6.1 LIst Of All Packages Required: 1)library(tidyverse) 2)library(nycflights13) 3)library(lubridate) 4)library(corrplot) 5)library(corrgram) 6)library(gam) 7)library(gbm) 8)library(xgboost) 9)library(tree) 10)library(randomForest) 11)library(dplyr)

Time to Knit : 1min 40 seconds

## Dataset

Dataset being analyzed is data from the nycflights13 package

library(tidyverse)  
library(nycflights13)  
library(lubridate)  
library(dplyr)  
library(corrplot)  
library(corrgram)  
library(randomForest)  
library(gam)  
library(gbm)  
library(xgboost)  
library(tree)  
  
#help(flights)  
#help(weather)  
#help(airports)  
#help(planes)  
fltrain <- read\_csv("C:/Users/Chidu/Documents/fltrain.csv.gz")  
fl\_test <- read\_csv("C:/Users/Chidu/Documents/fltest.csv.gz")  
fltrain

## # A tibble: 200,000 x 43  
## year.x month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2013 11 7 600 600 0 826 825  
## 2 2013 10 30 1252 1250 2 1356 1400  
## 3 2013 12 18 1723 1715 8 2008 2020  
## 4 2013 11 20 2029 2030 -1 2141 2205  
## 5 2013 10 21 1620 1625 -5 1818 1831  
## 6 2013 11 7 852 900 -8 1139 1157  
## 7 2013 9 29 1519 1529 -10 1639 1649  
## 8 2013 12 21 1526 1530 -4 1654 1710  
## 9 2013 11 7 1650 1650 0 1910 1906  
## 10 2013 3 31 1652 1700 -8 1810 1821  
## # ... with 199,990 more rows, and 35 more variables: arr\_delay <dbl>,  
## # carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,  
## # air\_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time\_hour <dttm>,  
## # temp <dbl>, dewp <dbl>, humid <dbl>, wind\_dir <dbl>, wind\_speed <dbl>,  
## # wind\_gust <dbl>, precip <dbl>, pressure <dbl>, visib <dbl>, name <chr>,  
## # lat <dbl>, lon <dbl>, alt <dbl>, tz <dbl>, dst <chr>, tzone <chr>,  
## # year.y <dbl>, type <chr>, manufacturer <chr>, model <chr>, engines <dbl>,  
## # seats <dbl>, speed <dbl>, engine <chr>

dim(fltrain)

## [1] 200000 43

Dimension of Training Dataset

## Handling of Missing data

fl <- fltrain  
for(i in 1:ncol(fl)) {  
 if(typeof(fl[[i]]) == "character") {  
 fl[[i]] <- factor(fl[[i]])  
 }  
}

counting the missing values in each variable.

missing\_value\_features <- function(x) { sum(is.na(x)) }  
sapply(fl,missing\_value\_features)

## year.x month day dep\_time sched\_dep\_time   
## 0 0 0 4898 0   
## dep\_delay arr\_time sched\_arr\_time arr\_delay carrier   
## 4898 5169 0 5584 0   
## flight tailnum origin dest air\_time   
## 0 1492 0 0 5584   
## distance hour minute time\_hour temp   
## 0 0 0 0 948   
## dewp humid wind\_dir wind\_speed wind\_gust   
## 948 948 5862 982 152260   
## precip pressure visib name lat   
## 937 23092 937 4484 4484   
## lon alt tz dst tzone   
## 4484 4484 4484 4484 4484   
## year.y type manufacturer model engines   
## 34298 31163 31163 31163 31163   
## seats speed engine   
## 31163 199415 31163

Discarding the variables that have more than 5% missing values which is around 10000.

fl <- fl%>% select(-year.y,-type,-manufacturer,-model,-engines,-seats, -speed, -engine,-wind\_gust,-pressure)  
summary(fl)

## year.x month day dep\_time sched\_dep\_time  
## Min. :2013 Min. : 1.000 Min. : 1.0 Min. : 1 Min. : 106   
## 1st Qu.:2013 1st Qu.: 4.000 1st Qu.: 8.0 1st Qu.: 907 1st Qu.: 905   
## Median :2013 Median : 7.000 Median :16.0 Median :1401 Median :1359   
## Mean :2013 Mean : 6.553 Mean :15.7 Mean :1349 Mean :1344   
## 3rd Qu.:2013 3rd Qu.:10.000 3rd Qu.:23.0 3rd Qu.:1745 3rd Qu.:1729   
## Max. :2013 Max. :12.000 Max. :31.0 Max. :2400 Max. :2359   
## NA's :4898   
## dep\_delay arr\_time sched\_arr\_time arr\_delay   
## Min. : -43.0 Min. : 1 Min. : 1 Min. : -79.000   
## 1st Qu.: -5.0 1st Qu.:1104 1st Qu.:1124 1st Qu.: -17.000   
## Median : -2.0 Median :1535 Median :1557 Median : -5.000   
## Mean : 12.7 Mean :1502 Mean :1537 Mean : 6.969   
## 3rd Qu.: 11.0 3rd Qu.:1941 3rd Qu.:1945 3rd Qu.: 14.000   
## Max. :1301.0 Max. :2400 Max. :2359 Max. :1272.000   
## NA's :4898 NA's :5169 NA's :5584   
## carrier flight tailnum origin dest   
## UA :34734 Min. : 1 N725MQ : 350 EWR:71658 ATL : 10319   
## B6 :32355 1st Qu.: 561 N723MQ : 300 JFK:65951 ORD : 10186   
## EV :32217 Median :1499 N722MQ : 294 LGA:62391 LAX : 9472   
## DL :28731 Mean :1975 N711MQ : 290 BOS : 9217   
## AA :19415 3rd Qu.:3470 N713MQ : 260 MCO : 8425   
## MQ :15608 Max. :8500 (Other):197014 CLT : 8319   
## (Other):36940 NA's : 1492 (Other):144062   
## air\_time distance hour minute   
## Min. : 20.0 Min. : 17 Min. : 1.00 Min. : 0.00   
## 1st Qu.: 82.0 1st Qu.: 502 1st Qu.: 9.00 1st Qu.: 8.00   
## Median :129.0 Median : 872 Median :13.00 Median :29.00   
## Mean :150.5 Mean :1038 Mean :13.18 Mean :26.22   
## 3rd Qu.:191.0 3rd Qu.:1389 3rd Qu.:17.00 3rd Qu.:44.00   
## Max. :695.0 Max. :4983 Max. :23.00 Max. :59.00   
## NA's :5584   
## time\_hour temp dewp   
## Min. :2013-01-01 10:00:00 Min. : 10.94 Min. :-9.94   
## 1st Qu.:2013-04-04 20:00:00 1st Qu.: 42.08 1st Qu.:26.06   
## Median :2013-07-03 15:00:00 Median : 57.20 Median :42.80   
## Mean :2013-07-03 12:05:05 Mean : 56.98 Mean :41.62   
## 3rd Qu.:2013-10-01 12:00:00 3rd Qu.: 71.96 3rd Qu.:57.92   
## Max. :2014-01-01 04:00:00 Max. :100.04 Max. :78.08   
## NA's :948 NA's :948   
## humid wind\_dir wind\_speed precip   
## Min. : 12.74 Min. : 0.0 Min. : 0.000 Min. :0.0000   
## 1st Qu.: 43.99 1st Qu.:130.0 1st Qu.: 6.905 1st Qu.:0.0000   
## Median : 57.69 Median :220.0 Median :10.357 Median :0.0000   
## Mean : 59.57 Mean :201.5 Mean :11.107 Mean :0.0045   
## 3rd Qu.: 75.33 3rd Qu.:290.0 3rd Qu.:14.960 3rd Qu.:0.0000   
## Max. :100.00 Max. :360.0 Max. :42.579 Max. :1.2100   
## NA's :948 NA's :5862 NA's :982 NA's :937   
## visib name lat   
## Min. : 0.000 Hartsfield Jackson Atlanta Intl : 10319 Min. :21.32   
## 1st Qu.:10.000 Chicago Ohare Intl : 10186 1st Qu.:32.90   
## Median :10.000 Los Angeles Intl : 9472 Median :36.10   
## Mean : 9.252 General Edward Lawrence Logan Intl: 9217 Mean :36.02   
## 3rd Qu.:10.000 Orlando Intl : 8425 3rd Qu.:41.41   
## Max. :10.000 (Other) :147897 Max. :61.17   
## NA's :937 NA's : 4484 NA's :4484   
## lon alt tz dst   
## Min. :-157.92 Min. : 3.0 Min. :-10.000 A :192358   
## 1st Qu.: -95.28 1st Qu.: 26.0 1st Qu.: -6.000 N : 3158   
## Median : -83.35 Median : 433.0 Median : -5.000 NA's: 4484   
## Mean : -89.44 Mean : 582.5 Mean : -5.748   
## 3rd Qu.: -80.15 3rd Qu.: 748.0 3rd Qu.: -5.000   
## Max. : -68.83 Max. :6602.0 Max. : -5.000   
## NA's :4484 NA's :4484 NA's :4484   
## tzone   
## America/New\_York :114518   
## America/Chicago : 44400   
## America/Los\_Angeles: 27368   
## America/Denver : 6069   
## America/Phoenix : 2759   
## (Other) : 402   
## NA's : 4484

dim(fl)

## [1] 200000 33

na.omit used to omit the missing values

fl <- na.omit(fl)  
dim(fl)

## [1] 184316 33

summary(fl)

## year.x month day dep\_time sched\_dep\_time  
## Min. :2013 Min. : 1.000 Min. : 1.00 Min. : 1 Min. : 500   
## 1st Qu.:2013 1st Qu.: 4.000 1st Qu.: 8.00 1st Qu.: 910 1st Qu.: 905   
## Median :2013 Median : 7.000 Median :16.00 Median :1408 Median :1359   
## Mean :2013 Mean : 6.553 Mean :15.67 Mean :1353 Mean :1342   
## 3rd Qu.:2013 3rd Qu.:10.000 3rd Qu.:23.00 3rd Qu.:1747 3rd Qu.:1729   
## Max. :2013 Max. :12.000 Max. :31.00 Max. :2400 Max. :2345   
##   
## dep\_delay arr\_time sched\_arr\_time arr\_delay   
## Min. : -43.00 Min. : 1 Min. : 1 Min. : -79.000   
## 1st Qu.: -5.00 1st Qu.:1106 1st Qu.:1123 1st Qu.: -17.000   
## Median : -2.00 Median :1545 Median :1602 Median : -5.000   
## Mean : 12.67 Mean :1511 Mean :1544 Mean : 7.014   
## 3rd Qu.: 11.00 3rd Qu.:1946 3rd Qu.:1950 3rd Qu.: 14.000   
## Max. :1301.00 Max. :2400 Max. :2359 Max. :1272.000   
##   
## carrier flight tailnum origin dest   
## UA :32252 Min. : 1 N725MQ : 322 EWR:65512 ATL : 9726   
## B6 :29282 1st Qu.: 544 N723MQ : 271 JFK:60327 ORD : 9443   
## EV :29137 Median :1499 N711MQ : 268 LGA:58477 LAX : 9185   
## DL :26998 Mean :1966 N722MQ : 268 BOS : 8674   
## AA :17742 3rd Qu.:3448 N351JB : 247 MCO : 8131   
## MQ :14382 Max. :8500 N258JB : 244 CLT : 7822   
## (Other):34523 (Other):182696 (Other):131335   
## air\_time distance hour minute   
## Min. : 20.0 Min. : 80 Min. : 5.00 Min. : 0.00   
## 1st Qu.: 81.0 1st Qu.: 502 1st Qu.: 9.00 1st Qu.: 8.00   
## Median :127.0 Median : 866 Median :13.00 Median :29.00   
## Mean :149.5 Mean :1035 Mean :13.15 Mean :26.06   
## 3rd Qu.:184.0 3rd Qu.:1372 3rd Qu.:17.00 3rd Qu.:43.00   
## Max. :695.0 Max. :4983 Max. :23.00 Max. :59.00   
##   
## time\_hour temp dewp   
## Min. :2013-01-01 10:00:00 Min. : 10.94 Min. :-9.94   
## 1st Qu.:2013-04-04 13:00:00 1st Qu.: 42.08 1st Qu.:26.06   
## Median :2013-07-03 17:00:00 Median : 57.02 Median :42.08   
## Mean :2013-07-03 11:34:16 Mean : 56.86 Mean :41.33   
## 3rd Qu.:2013-10-02 10:00:00 3rd Qu.: 71.96 3rd Qu.:57.20   
## Max. :2013-12-30 23:00:00 Max. :100.04 Max. :78.08   
##   
## humid wind\_dir wind\_speed precip   
## Min. : 13.00 Min. : 0.0 Min. : 0.000 Min. :0.000000   
## 1st Qu.: 43.71 1st Qu.:130.0 1st Qu.: 6.905 1st Qu.:0.000000   
## Median : 57.14 Median :220.0 Median :10.357 Median :0.000000   
## Mean : 59.12 Mean :201.9 Mean :11.202 Mean :0.004059   
## 3rd Qu.: 74.29 3rd Qu.:290.0 3rd Qu.:14.960 3rd Qu.:0.000000   
## Max. :100.00 Max. :360.0 Max. :42.579 Max. :1.210000   
##   
## visib name lat   
## Min. : 0.000 Hartsfield Jackson Atlanta Intl : 9726 Min. :21.32   
## 1st Qu.:10.000 Chicago Ohare Intl : 9443 1st Qu.:32.90   
## Median :10.000 Los Angeles Intl : 9185 Median :36.08   
## Mean : 9.294 General Edward Lawrence Logan Intl: 8674 Mean :35.97   
## 3rd Qu.:10.000 Orlando Intl : 8131 3rd Qu.:41.41   
## Max. :10.000 Charlotte Douglas Intl : 7822 Max. :61.17   
## (Other) :131335   
## lon alt tz dst   
## Min. :-157.92 Min. : 3.0 Min. :-10.000 A:181313   
## 1st Qu.: -95.34 1st Qu.: 26.0 1st Qu.: -6.000 N: 3003   
## Median : -83.35 Median : 433.0 Median : -5.000   
## Mean : -89.58 Mean : 582.3 Mean : -5.757   
## 3rd Qu.: -80.15 3rd Qu.: 748.0 3rd Qu.: -5.000   
## Max. : -68.83 Max. :6602.0 Max. : -5.000   
##   
## tzone   
## America/Anchorage : 3   
## America/Chicago : 41444   
## America/Denver : 5819   
## America/Los\_Angeles: 26461   
## America/New\_York :107586   
## America/Phoenix : 2629   
## Pacific/Honolulu : 374

## Summaries of the response variable dep\_delay

The departure delays variable is highly right-skewed that is showed.

range(fl$dep\_delay)

## [1] -43 1301

fivenum(fl$dep\_delay)

## [1] -43 -5 -2 11 1301

quantile(fl$dep\_delay,probs = c(0,0.01,0.05,0.1,0.25,.5,.75,.90,.95,.99,1))

## 0% 1% 5% 10% 25% 50% 75% 90% 95% 99% 100%   
## -43 -12 -9 -7 -5 -2 11 49 88 193 1301

mean(fl$dep\_delay >= 60) # about 15,000 or 8% of flights

## [1] 0.08210356

Top 10 delays.

fl%>% arrange(desc(dep\_delay)) %>% head(10)

## # A tibble: 10 x 33  
## year.x month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2013 1 9 641 900 1301 1242 1530  
## 2 2013 9 20 1139 1845 1014 1457 2210  
## 3 2013 3 17 2321 810 911 135 1020  
## 4 2013 7 22 2257 759 898 121 1026  
## 5 2013 12 5 756 1700 896 1058 2020  
## 6 2013 5 19 713 1700 853 1007 1955  
## 7 2013 2 10 2243 830 853 100 1106  
## 8 2013 12 19 734 1725 849 1046 2039  
## 9 2013 12 17 705 1700 845 1026 2020  
## 10 2013 12 14 830 1845 825 1210 2154  
## # ... with 25 more variables: arr\_delay <dbl>, carrier <fct>, flight <dbl>,  
## # tailnum <fct>, origin <fct>, dest <fct>, air\_time <dbl>, distance <dbl>,  
## # hour <dbl>, minute <dbl>, time\_hour <dttm>, temp <dbl>, dewp <dbl>,  
## # humid <dbl>, wind\_dir <dbl>, wind\_speed <dbl>, precip <dbl>, visib <dbl>,  
## # name <fct>, lat <dbl>, lon <dbl>, alt <dbl>, tz <dbl>, dst <fct>,  
## # tzone <fct>

Summaries of departure delay by NYC airport:

Q3 <- function(x) { quantile(x,probs=.75) }  
fl %>% group\_by(origin) %>%   
 summarize(n=n(),med\_d = median(dep\_delay),Q3\_d = Q3(dep\_delay), max\_d = max(dep\_delay)) %>%   
 arrange(desc(Q3\_d)) %>% head(10)

## # A tibble: 3 x 5  
## origin n med\_d Q3\_d max\_d  
## <fct> <int> <dbl> <dbl> <dbl>  
## 1 EWR 65512 -1 15 896  
## 2 JFK 60327 -1 10 1301  
## 3 LGA 58477 -3 7 911

Summaries of departure delay by airline (carrier).

fl %>% group\_by(carrier) %>%   
 summarize(n=n(),med\_d = median(dep\_delay),Q3\_d = Q3(dep\_delay), max\_d = max(dep\_delay)) %>%   
 arrange(desc(Q3\_d)) %>% head(10)

## # A tibble: 10 x 5  
## carrier n med\_d Q3\_d max\_d  
## <fct> <int> <dbl> <dbl> <dbl>  
## 1 EV 29137 -1 25 536  
## 2 WN 6897 1 18 471  
## 3 F9 388 0 17.2 853  
## 4 9E 10179 -2 16 430  
## 5 FL 1832 1 16 602  
## 6 YV 312 -3 13 387  
## 7 B6 29282 -1 12 502  
## 8 UA 32252 0 11 483  
## 9 MQ 14382 -3 9 486  
## 10 VX 2991 0 7 653

fl %>% group\_by(origin,carrier) %>%   
 summarize(n=n(),med\_d = median(dep\_delay),Q3\_d = Q3(dep\_delay), max\_d = max(dep\_delay)) %>%   
 arrange(desc(Q3\_d)) %>% head(10)

## # A tibble: 10 x 6  
## # Groups: origin [3]  
## origin carrier n med\_d Q3\_d max\_d  
## <fct> <fct> <int> <dbl> <dbl> <dbl>  
## 1 EWR OO 3 4 67.5 131  
## 2 EWR EV 23565 -1 26 443  
## 3 LGA EV 4769 -2 22 473  
## 4 JFK 9E 8126 -1 20 430  
## 5 JFK EV 803 -2 19 536  
## 6 EWR WN 3487 2 18 440  
## 7 LGA WN 3410 1 18 471  
## 8 LGA F9 388 0 17.2 853  
## 9 EWR MQ 1156 -2 17 381  
## 10 LGA FL 1832 1 16 602

fl %>% group\_by(dest,carrier) %>%   
 summarize(n=n(),med\_d = median(dep\_delay),Q3\_d = Q3(dep\_delay), max\_d = max(dep\_delay)) %>%   
 arrange(desc(Q3\_d)) %>% head(10)

## # A tibble: 10 x 6  
## # Groups: dest [10]  
## dest carrier n med\_d Q3\_d max\_d  
## <fct> <fct> <int> <dbl> <dbl> <dbl>  
## 1 STL UA 2 77.5 116. 155  
## 2 DTW OO 2 61 96 131  
## 3 TYS EV 183 8 68.5 285  
## 4 PBI EV 3 50 67.5 85  
## 5 ORD OO 1 67 67 67  
## 6 RDU UA 1 60 60 60  
## 7 TUL EV 185 3 53 251  
## 8 OKC EV 184 8.5 51.5 207  
## 9 BHM EV 175 3 50 325  
## 10 CAE EV 57 10 48 163

Summaries of departure delay by date:

fl %>% group\_by(month,day) %>%   
 summarize(n=n(),med\_d = mean(dep\_delay),max\_d = max(dep\_delay)) %>%   
 arrange(desc(med\_d)) %>% head(10) # what happened on march 8?

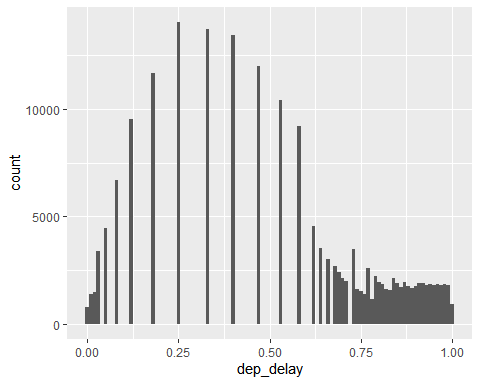
## # A tibble: 10 x 5  
## # Groups: month [7]  
## month day n med\_d max\_d  
## <dbl> <dbl> <int> <dbl> <dbl>  
## 1 3 8 461 79.5 470  
## 2 7 1 505 58.1 363  
## 3 7 10 471 56.6 576  
## 4 9 2 438 53.7 696  
## 5 12 5 458 52.2 896  
## 6 5 23 453 51.5 410  
## 7 4 19 511 50.4 812  
## 8 9 12 444 50.4 602  
## 9 6 13 469 50.3 388  
## 10 7 22 476 49.9 898

Summaries of departure delay by precipitation:

fl %>% mutate(haveprecip = factor(precip>0)) %>% group\_by(haveprecip) %>%   
 summarize(n=n(),med\_d = median(dep\_delay),Q3\_d = Q3(dep\_delay), max\_d = max(dep\_delay)) %>%   
 arrange(desc(med\_d)) %>% head(10)

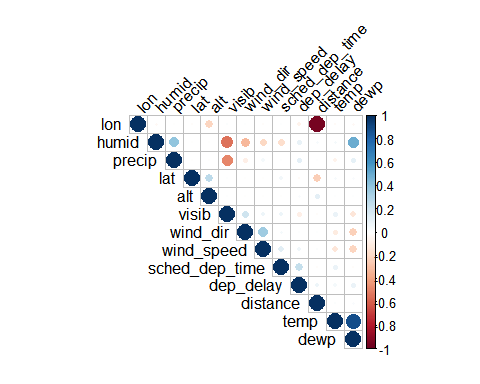
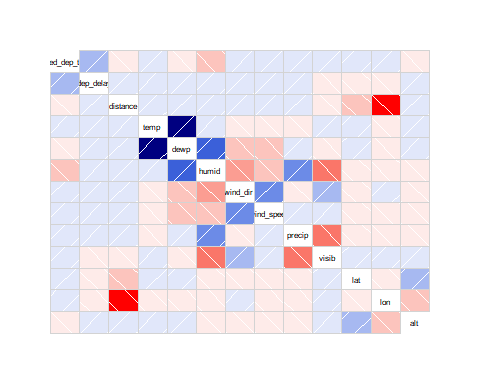
## # A tibble: 2 x 5  
## haveprecip n med\_d Q3\_d max\_d  
## <fct> <int> <dbl> <dbl> <dbl>  
## 1 TRUE 11804 5 41 853  
## 2 FALSE 172512 -2 9 1301

den <- nrow(fl)+1  
fl <- fl %>% mutate(dep\_delay = rank(dep\_delay)/den)  
ggplot(fl,aes(x=dep\_delay)) + geom\_histogram(binwidth=.01)



# Data Wrangling Part

fl <- fl %>%   
 mutate(dep\_date = make\_date(year.x,month,day)) %>%   
 select(-year.x,-month,-day,-dep\_time,-arr\_time,-arr\_delay,  
 -sched\_arr\_time,-tailnum,-flight,-name,-air\_time,  
 -hour,-minute,-time\_hour,-tz,-dst,-dest) %>%  
 mutate(precip = as.numeric(precip>0))  
corrplot(corrgram(fl),type = "upper", order = "hclust",tl.col = "black", tl.srt = 45)



## Relationship between dep\_delay and quantitative predictors

ggplot(fl,aes(x=dep\_date,y=dep\_delay)) + geom\_point(alpha=.01) + geom\_smooth()  
  
ggplot(fl,aes(x=sched\_dep\_time,y=dep\_delay)) + geom\_point(alpha=0.01) + geom\_smooth()  
  
ggplot(fl,aes(x=distance,y=dep\_delay)) + geom\_point(alpha=0.01) + geom\_smooth()  
ggplot(fl,aes(x=log(distance),y=dep\_delay)) + geom\_point(alpha=0.01) + geom\_smooth()  
  
  
ggplot(fl,aes(x=temp,y=dep\_delay)) + geom\_point(alpha=0.01) + geom\_smooth()  
  
ggplot(fl,aes(x=dewp,y=dep\_delay)) + geom\_point(alpha=0.01) + geom\_smooth()

fl <- mutate(fl,logdistance = log(distance)) %>% select(-distance)  
fl <- mutate(fl,logalt = log(alt)) %>% select(-alt)

## Splitting the Training set into train and validation set for tuning of the parameters

set.seed(123)  
tzone1 <- as.factor(fl$tzone)  
fl <- mutate(fl,tzone=tzone1)  
tr\_size <- ceiling(2\*nrow(fl)/3)  
train <- sample(1:nrow(fl),size=tr\_size)  
fl\_tr <- fl[train,]  
fl\_te <- fl[-train,]  
  
var\_dd <- var(fl\_te$dep\_delay)  
var\_dd

## [1] 0.08311941

## Transforming the Test data set as per the trained model

set.seed(9)  
  
for(j in 1:ncol(fl\_test)) {  
 if(typeof(fl\_test[[j]]) == "character") {  
 fl\_test[[j]] <- factor(fl\_test[[j]])  
 }  
}  
  
fl\_test <- fl\_test%>% select(-year.y,-type,-manufacturer,-model,-engines,-seats, -speed, -engine,-wind\_gust,-pressure)  
fl\_test <- na.omit(fl\_test)  
dim(fl\_test)

## [1] 126034 33

den\_test <- nrow(fl\_test)+1  
fl\_test <- fl\_test %>% mutate(dep\_delay = rank(dep\_delay)/den\_test)  
  
fl\_test <- fl\_test %>%   
 mutate(dep\_date = make\_date(year.x,month,day)) %>%   
 select(-year.x,-month,-day,-dep\_time,-arr\_time,-arr\_delay,  
 -sched\_arr\_time,-tailnum,-flight,-name,-air\_time,  
 -hour,-minute,-time\_hour,-tz,-dst,-dest) %>%  
 mutate(precip = as.numeric(precip>0))  
  
fl\_test <- mutate(fl\_test,logdistance = log(distance)) %>% select(-distance)  
fl\_test <- mutate(fl\_test,logalt = log(alt)) %>% select(-alt)  
  
tzone2 <- as.factor(fl\_test$tzone)  
carrier\_test <-as.factor(fl\_test$carrier)  
origin\_test <- as.factor(fl\_test$origin)  
fl\_test <- mutate(fl\_test,carrier=carrier\_test,origin=origin\_test,tzone=tzone2)  
  
dep\_date\_numeric\_t <- as.numeric(fl\_test$dep\_date)  
dep\_date\_numeric\_t <- dep\_date\_numeric\_t - mean(dep\_date\_numeric\_t)  
fl\_test <- mutate(fl\_test,dep\_date = dep\_date\_numeric\_t)

set.seed(67)  
form <- formula(dep\_delay ~ s(dep\_date) + s(sched\_dep\_time) + carrier + origin + tzone + s(logdistance) +  
 s(temp) + s(dewp) + s(humid) + s(wind\_dir) + s(wind\_speed) + precip + s(visib))  
gam\_fit <- gam(form, data=fl\_tr,family=gaussian)   
summary(gam\_fit)  
plot(gam\_fit,se=TRUE)  
gam\_pred <- predict(gam\_fit,newdata=fl\_te)  
mse\_gam <- mean((fl\_te$dep\_delay-gam\_pred)^2)  
  
mse\_gam  
abs(mse\_gam - var\_dd)/var\_dd

dep\_date\_numeric <- as.numeric(fl\_tr$dep\_date)  
dep\_date\_numeric <- dep\_date\_numeric - mean(dep\_date\_numeric)  
fl\_tr\_tem <- mutate(fl\_tr,dep\_date = dep\_date\_numeric)  
  
dep\_date\_numeric <- as.numeric(fl\_te$dep\_date)  
dep\_date\_numeric <- dep\_date\_numeric - mean(dep\_date\_numeric)  
fl\_te\_tem <- mutate(fl\_te,dep\_date = dep\_date\_numeric)

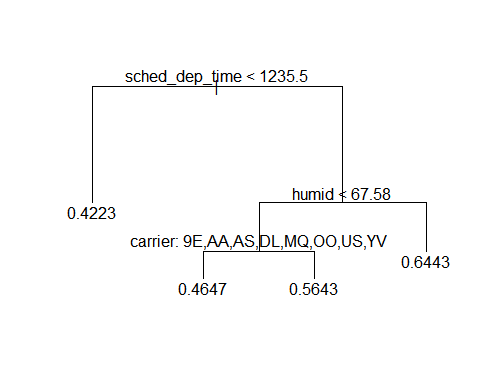
set.seed(142)  
  
gbm\_fit <-gbm(dep\_delay ~ .,data=fl\_tr\_tem,distribution="gaussian",  
 n.trees = 2000, shrinkage = 0.2, interaction.depth = 3)  
summary(gbm\_fit)  
  
gbm\_pred <- predict(gbm\_fit,newdata=fl\_te\_tem,n.trees = 2000)  
gbm\_pred\_test <- predict(gbm\_fit,newdata=fl\_test,n.trees = 2000)  
mse\_gbm <- mean((fl\_te$dep\_delay-gbm\_pred)^2)  
mse\_gbm\_test <- mean((fl\_test$dep\_delay-gbm\_pred\_test)^2)  
mse\_gbm  
mse\_gbm\_test  
  
abs(mse\_gbm - var\_dd)/var\_dd

Decision trees for regression

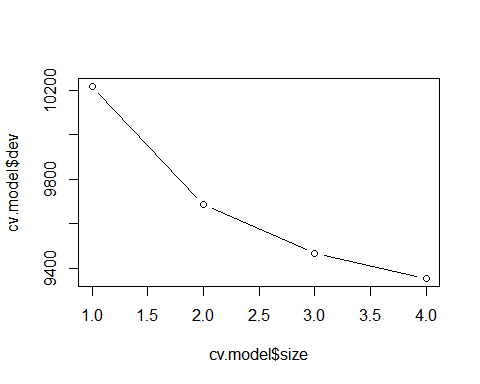
set.seed(123)  
  
tree.model <- tree(dep\_delay~., fl\_tr\_tem)  
summary(tree.model)

##   
## Regression tree:  
## tree(formula = dep\_delay ~ ., data = fl\_tr\_tem)  
## Variables actually used in tree construction:  
## [1] "sched\_dep\_time" "humid" "carrier"   
## Number of terminal nodes: 4   
## Residual mean deviance: 0.07595 = 9332 / 122900   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.644300 -0.239700 0.006016 0.000000 0.228100 0.577700

plot(tree.model)  
text(tree.model,pretty=0)



cv.model = cv.tree(tree.model)  
plot(cv.model$size,cv.model$dev, type = "b")



prune.model= prune.tree(tree.model, best=4)  
yhat=predict(prune.model, newdata=fl\_te\_tem)  
yhat\_test = predict(prune.model, newdata=fl\_test)  
mean((yhat-fl\_te\_tem$dep\_delay)^2)

## [1] 0.07628095

mean((yhat\_test-fl\_test$dep\_delay)^2)

## [1] 0.07586896

Using Random Forest method for regression

set.seed(432)  
fl\_rf <- fl\_tr\_tem[1:35000,]  
bag.flights <- randomForest(fl\_rf$dep\_delay~., data=fl\_rf,ntree=100, importance=TRUE,na.action=na.omit)  
bag.flights  
yhat.bag <- predict(bag.flights,newdata= fl\_te\_tem)  
yhat.bag\_test <- predict(bag.flights, newdata=fl\_test)  
mean((yhat.bag-fl\_te\_tem$dep\_delay)^2)  
mean((yhat.bag\_test -fl\_test$dep\_delay)^2)

set.seed(134)  
data.xg <- xgb.DMatrix(data=data.matrix(fl\_tr\_tem[-2]), label=fl\_tr\_tem$dep\_delay)  
bst\_dmatrix <- xgboost(data=data.xg, max.depth=10,eta = 0.3, nrounds=100, eval\_metric ="rmse")

## [1] train-rmse:0.272287   
## [2] train-rmse:0.262661   
## [3] train-rmse:0.256055   
## [4] train-rmse:0.251486   
## [5] train-rmse:0.246958   
## [6] train-rmse:0.244102   
## [7] train-rmse:0.241053   
## [8] train-rmse:0.238290   
## [9] train-rmse:0.236164   
## [10] train-rmse:0.234420   
## [11] train-rmse:0.232542   
## [12] train-rmse:0.231670   
## [13] train-rmse:0.230701   
## [14] train-rmse:0.229577   
## [15] train-rmse:0.227457   
## [16] train-rmse:0.226650   
## [17] train-rmse:0.225021   
## [18] train-rmse:0.223991   
## [19] train-rmse:0.223346   
## [20] train-rmse:0.222551   
## [21] train-rmse:0.221195   
## [22] train-rmse:0.220478   
## [23] train-rmse:0.220037   
## [24] train-rmse:0.218570   
## [25] train-rmse:0.217824   
## [26] train-rmse:0.216946   
## [27] train-rmse:0.216574   
## [28] train-rmse:0.215742   
## [29] train-rmse:0.215258   
## [30] train-rmse:0.214199   
## [31] train-rmse:0.213031   
## [32] train-rmse:0.212376   
## [33] train-rmse:0.211394   
## [34] train-rmse:0.210862   
## [35] train-rmse:0.210556   
## [36] train-rmse:0.209610   
## [37] train-rmse:0.208961   
## [38] train-rmse:0.207760   
## [39] train-rmse:0.206760   
## [40] train-rmse:0.205982   
## [41] train-rmse:0.205357   
## [42] train-rmse:0.204501   
## [43] train-rmse:0.203894   
## [44] train-rmse:0.202993   
## [45] train-rmse:0.202228   
## [46] train-rmse:0.201912   
## [47] train-rmse:0.201732   
## [48] train-rmse:0.201252   
## [49] train-rmse:0.200998   
## [50] train-rmse:0.200289   
## [51] train-rmse:0.199479   
## [52] train-rmse:0.199096   
## [53] train-rmse:0.198797   
## [54] train-rmse:0.198283   
## [55] train-rmse:0.197604   
## [56] train-rmse:0.196841   
## [57] train-rmse:0.196028   
## [58] train-rmse:0.195582   
## [59] train-rmse:0.194972   
## [60] train-rmse:0.194614   
## [61] train-rmse:0.193754   
## [62] train-rmse:0.193029   
## [63] train-rmse:0.192260   
## [64] train-rmse:0.191463   
## [65] train-rmse:0.190951   
## [66] train-rmse:0.190069   
## [67] train-rmse:0.189479   
## [68] train-rmse:0.189206   
## [69] train-rmse:0.188774   
## [70] train-rmse:0.188042   
## [71] train-rmse:0.187462   
## [72] train-rmse:0.186601   
## [73] train-rmse:0.186187   
## [74] train-rmse:0.185811   
## [75] train-rmse:0.185496   
## [76] train-rmse:0.184813   
## [77] train-rmse:0.184535   
## [78] train-rmse:0.184119   
## [79] train-rmse:0.183726   
## [80] train-rmse:0.183387   
## [81] train-rmse:0.182908   
## [82] train-rmse:0.181861   
## [83] train-rmse:0.181406   
## [84] train-rmse:0.180916   
## [85] train-rmse:0.180203   
## [86] train-rmse:0.179564   
## [87] train-rmse:0.179230   
## [88] train-rmse:0.178760   
## [89] train-rmse:0.178625   
## [90] train-rmse:0.178368   
## [91] train-rmse:0.178248   
## [92] train-rmse:0.177726   
## [93] train-rmse:0.177124   
## [94] train-rmse:0.176468   
## [95] train-rmse:0.175937   
## [96] train-rmse:0.175059   
## [97] train-rmse:0.174326   
## [98] train-rmse:0.173833   
## [99] train-rmse:0.173589   
## [100] train-rmse:0.173253

bst\_dmatrix

## ##### xgb.Booster  
## raw: 3.8 Mb   
## call:  
## xgb.train(params = params, data = dtrain, nrounds = nrounds,   
## watchlist = watchlist, verbose = verbose, print\_every\_n = print\_every\_n,   
## early\_stopping\_rounds = early\_stopping\_rounds, maximize = maximize,   
## save\_period = save\_period, save\_name = save\_name, xgb\_model = xgb\_model,   
## callbacks = callbacks, max.depth = 10, eta = 0.3, eval\_metric = "rmse")  
## params (as set within xgb.train):  
## max\_depth = "10", eta = "0.3", eval\_metric = "rmse", silent = "1"  
## xgb.attributes:  
## niter  
## callbacks:  
## cb.print.evaluation(period = print\_every\_n)  
## cb.evaluation.log()  
## # of features: 16   
## niter: 100  
## nfeatures : 16   
## evaluation\_log:  
## iter train\_rmse  
## 1 0.272287  
## 2 0.262661  
## ---   
## 99 0.173589  
## 100 0.173253

data.val.xg <- xgb.DMatrix(data=data.matrix(fl\_te\_tem[-2]), label=fl\_te\_tem$dep\_delay)  
data.test.xg <- xgb.DMatrix(data=data.matrix(fl\_test[-2]), label=fl\_test$dep\_delay)  
pred <- predict(bst\_dmatrix,data.val.xg)  
pred\_test <- predict(bst\_dmatrix,data.test.xg)  
xgb\_val\_mse <- mean((pred-fl\_te\_tem$dep\_delay)^2)  
xgb\_val\_mse

## [1] 0.06584833

xgb\_te\_mse <- mean((pred\_test-fl\_test$dep\_delay)^2)  
xgb\_te\_mse

## [1] 0.06541933