***DMPM Assignment 4***

Name: Rushikesh Jyoti

Division: A

Roll no: 27

SRN: 201901139

Question

* Read the dataset **"FlightDelays.csv"** that is provided to you.
* Build a suitable logistic regression model using R and predict the status of the flights in your test data set.
* Draw the ROCR and Lift curves of your model and comment on the effectiveness of your model.
* Develop some metrics to determine the accuracy of your classification model (Error % using confusion matrix)

***Code***

library(dplyr)

library(caret)

library(reshape2)

library(pROC)

library(corrplot)

library(caTools)

flight = read.csv("FlightDelays.csv")

head(flight)

summary(flight)

str(flight)

table(flight$delay)

flight = flight %>% mutate(delay = ifelse(delay == "ontime",0,1))

# We need to convert categorical data to numeric data aka encoding the data

encode\_category = function(x, order = unique(x)) {

as.numeric(factor(x, levels = order, exclude = NULL))

}

flight[["tailnu"]] = encode\_category(flight[["tailnu"]])

flight[["dest"]] = encode\_category(flight[["dest"]])

flight[["origin"]] = encode\_category(flight[["origin"]])

flight[["carrier"]] = encode\_category(flight[["carrier"]])

# We don't really need date

flight = select(flight, -date)

head(flight)

# Split the data for training and testing sets

set.seed(101)

sample = sample.split(flight$delay, SplitRatio = .70)

train = subset(flight, sample == TRUE)

test = subset(flight, sample == FALSE)

head(test)

# Plot the correlation heat map

corrplot(cor(train), tl.col="black")

# Build the model

logreg = glm(delay ~ ., family = binomial(link = 'logit'),

data = train)

summary(logreg)

# Predict the values using the model

prob = logreg %>% predict(test\_new, type = "response")

test$prob = prob

threshold = 0.3

# If prediction is less than threshold then put 0 otherwise 1

test = test %>% mutate(predicted = ifelse(prob < threshold,0,1))

head(test\_new)

# The confusion matrix

mat = table(test$delay, test$predicted)

mat

# Metrics to check efficiency of model

accuracy = (mat[1] + mat[4]) / (sum(mat))

error\_rate = 1 - accuracy

precision = mat[1] / (mat[1] + mat[3])

recall = mat[1] / (mat[1] + mat[2])

cat("Accuracy: ", accuracy \* 100,

"%\nError Rate:", error\_rate \* 100,

"%\nPrecision: ",precision \* 100,

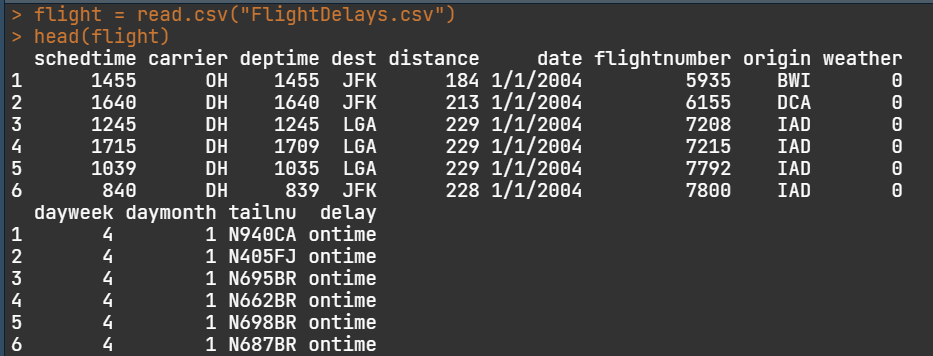
"%\nRecall:",recall \* 100,"%")

# ROCR curve

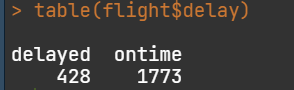
roc = roc(test$delay ~ prob, plot = TRUE, print.auc = TRUE)

***Output***

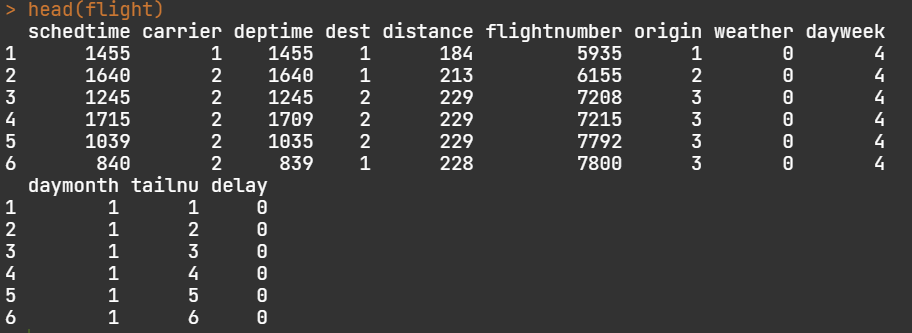
Taking a look at the dataset



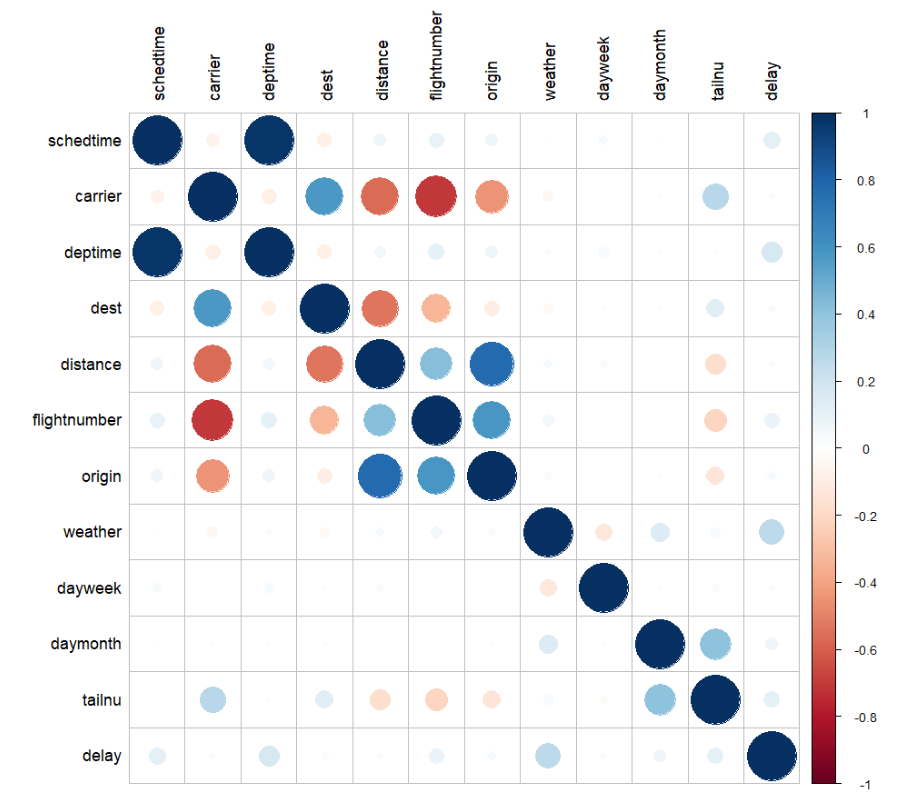
We need to predict if flights are delayed or not, so let’s look at delays in the dataset



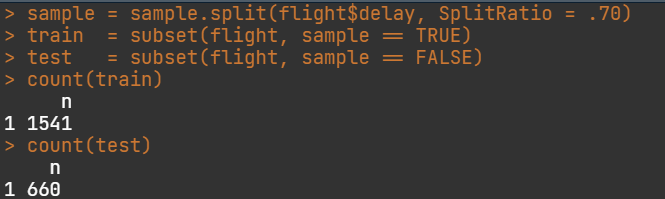
After encoding the categorical variables,



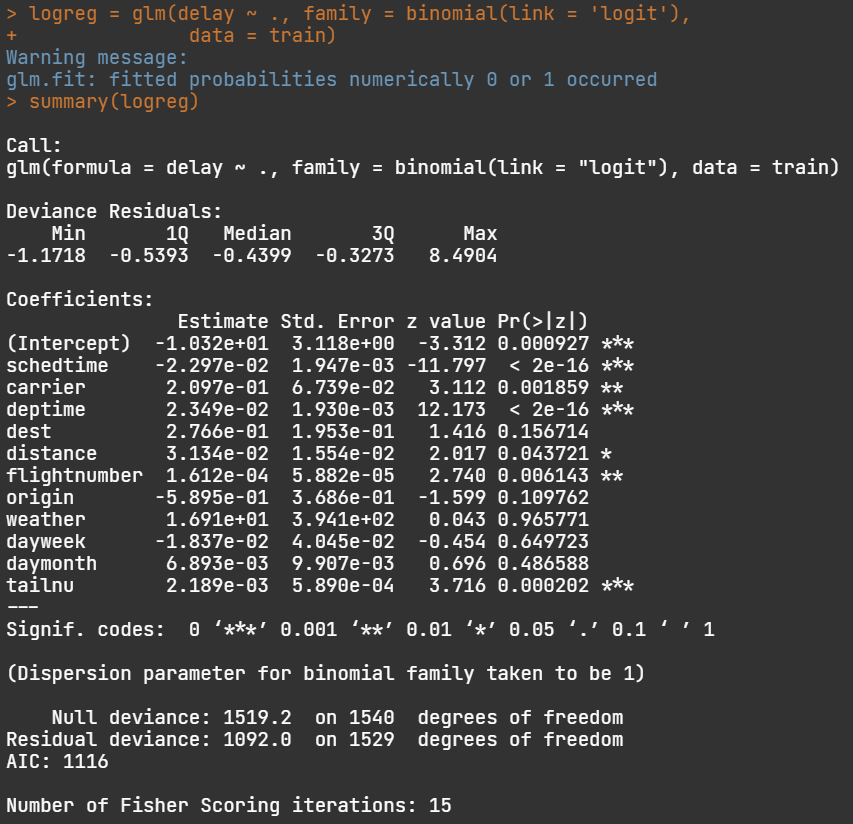
Taking a look at correlation between variables



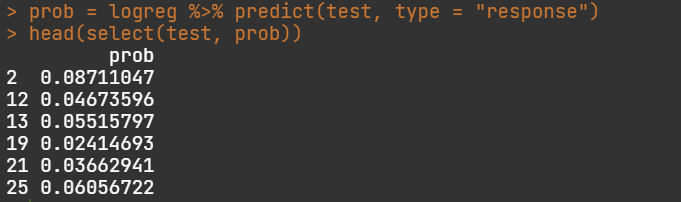
Split the dataset into 70-30



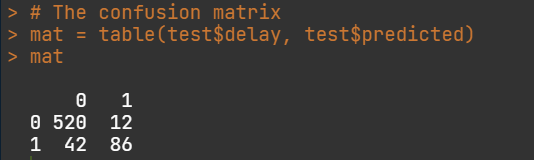
Building the model



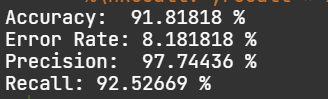
Predicting the values



The confusion matrix



Metrics of the model



Sidenote: woho! Those are some good numbers!

ROC Curve

