DMPM Assignment 4

Name: Saniya S. Inamdar

SRN: 201900913

Roll no. : 17

R CODE:

```
library(dplyr)
library(caret)
library(reshape2)
library(pROC)
library(corrplot)
library(caTools)

flight <- read.csv("FlightDelays.csv")
head(flight)
summary(flight)
summary(flight$tailnu)
str(flight)



flight %>%
  count(delay)

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

summary(flight)

encode_ordinal <- function(x, order = unique(x)) {
  x <- as.numeric(factor(x, levels = order, exclude = NULL))
  x
}

flight[["tailnu"]] <- encode_ordinal(flight[["tailnu"]])
flight[["dest"]] <- encode_ordinal(flight[["dest"]])
flight[["origin"]] <- encode_ordinal(flight[["origin"]])
flight[["carrier"]] <- encode_ordinal(flight[["carrier"]])
head(flight)

flight=within(flight, rm(date))
head(flight)

set.seed(101)
sample = sample.split(flight$delay, SplitRatio = .60)
train = subset(flight, sample == TRUE)
test  = subset(flight, sample == FALSE)
```

```
test_new = within(test, rm(delay))

head(test)
corrplot(cor(train), method="pie",shade.col=NA, tl.col="black",
tl.srt=45)



logreg <- glm(delay ~ ., family = binomial(link = 'logit'), data =
train)
summary(logreg)

prob <- logreg %>% predict(test_new, type = "response")

test_new$prob = prob

test_new <- test_new %>%
  mutate(predicted = ifelse(prob<0.3,0,1))

head(test_new)

table(test$delay, test_new$predicted)

accuracy = (672+106)/(672+106+37+65)

error_rate = 1- accuracy

precision = 672/(672+37)

recall = 672/(672+65)

cat("Accuracy: ",accuracy*100,"%\nError Rate:
",error_rate*100,"%\nPrecision: ",precision*100,"%\nRecall:
",recall*100,"%")

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

OUTPUT:

```
There aren't any missing values.
This is how the data looks like.
Need to encode categorical variables.
```

```
> flight <- read.csv("FlightDelays.csv")
> head(flight)
  schedtime carrier deptime dest distance     date flightnumber origin weather dayweek daymonth tailnu  delay
1      1455      OH    1455  JFK      184 1/1/2004         5935    BWI       0       4        1 N940CA ontime
2      1640      DH    1640  JFK      213 1/1/2004         6155    DCA       0       4        1 N405FJ ontime
3      1245      DH    1245  LGA      229 1/1/2004         7208    IAD       0       4        1 N695BR ontime
4      1715      DH    1709  LGA      229 1/1/2004         7215    IAD       0       4        1 N662BR ontime
5      1039      DH    1035  LGA      229 1/1/2004         7792    IAD       0       4        1 N698BR ontime
6       840      DH     839  JFK      228 1/1/2004         7800    IAD       0       4        1 N687BR ontime
> summary(flight)
   schedtime       carrier             deptime          dest              distance          date             flightnumber
 Min.   : 600   Length:2201        Min.   :  10   Length:2201        Min.   :169.0   Length:2201        Min.   : 746
 1st Qu.:1000   Class :character   1st Qu.:1004   Class :character   1st Qu.:213.0   Class :character   1st Qu.:2156
 Median :1455   Mode  :character   Median :1450   Mode  :character   Median :214.0   Mode  :character   Median :2385
 Mean   :1372                      Mean   :1369                      Mean   :211.9                      Mean   :3815
 3rd Qu.:1710                      3rd Qu.:1709                      3rd Qu.:214.0                      3rd Qu.:6155
 Max.   :2130                      Max.   :2330                      Max.   :229.0                      Max.   :7924
    origin             weather          dayweek         daymonth          tailnu             delay
 Length:2201        Min.   :0.00000   Min.   :1.000   Min.   : 1.00   Length:2201        Length:2201
 Class :character   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 8.00   Class :character   Class :character
 Mode  :character   Median :0.00000   Median :4.000   Median :16.00   Mode  :character   Mode  :character
                    Mean   :0.01454   Mean   :3.905   Mean   :16.02
                    3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:23.00
                    Max.   :1.00000   Max.   :7.000   Max.   :31.00
>
```

Dataset is a bit imbalanced:

```
> flight %>%
+   count(delay)
    delay    n
1 delayed  428
2  ontime 1773
>
```

After Ordinal encoding on all categorical variables:

```
> head(flight)
  schedtime carrier deptime dest distance     date flightnumber origin weather dayweek daymonth tailnu delay
1      1455       1    1455    1      184 1/1/2004         5935      1       0       4        1      1     0
2      1640       2    1640    1      213 1/1/2004         6155      2       0       4        1      2     0
3      1245       2    1245    2      229 1/1/2004         7208      3       0       4        1      3     0
4      1715       2    1709    2      229 1/1/2004         7215      3       0       4        1      4     0
5      1039       2    1035    2      229 1/1/2004         7792      3       0       4        1      5     0
6       840       2     839    1      228 1/1/2004         7800      3       0       4        1      6     0
>
```

Removed the "date" column;

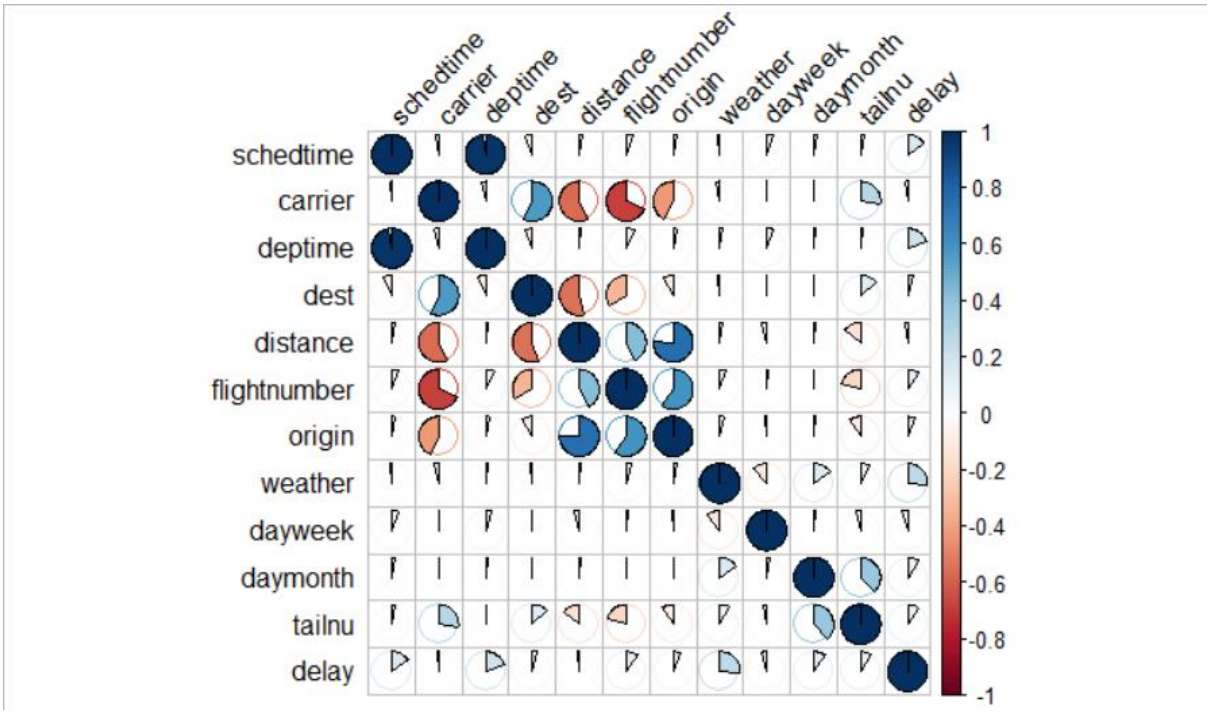
```
> flight=within(flight, rm(date))
> head(flight)
  schedtime carrier deptime dest distance flightnumber origin weather dayweek daymonth tailnu delay
1      1455       1    1455    1      184         5935      1       0       4        1      1     0
2      1640       2    1640    1      213         6155      2       0       4        1      2     0
3      1245       2    1245    2      229         7208      3       0       4        1      3     0
4      1715       2    1709    2      229         7215      3       0       4        1      4     0
5      1039       2    1035    2      229         7792      3       0       4        1      5     0
6       840       2     839    1      228         7800      3       0       4        1      6     0
>
```

Created The tes and train dataset. Did a 60-40 split of the dataset randomly.

```
> set.seed(101)
> sample = sample.split(flight$delay, SplitRatio = .60)
> train = subset(flight, sample == TRUE)
> test  = subset(flight, sample == FALSE)
> test_new = within(test, rm(delay))
> head(test)
   schedtime carrier deptime dest distance flightnumber origin weather dayweek daymonth tailnu delay
3       1245       2    1245    2      229         7208      3       0       4        1      3     0
4       1715       2    1709    2      229         7215      3       0       4        1      4     0
9       1715       2    1710    1      228         7812      3       0       4        1      9     0
11      2120       2    2114    2      229         7924      3       0       4        1     11     0
12      1455       3    1458    1      213          746      2       0       4        1     12     0
13       930       3     932    2      214         1746      2       0       4        1     13     0
>
```

Removed the "delay" column from the test dataset and saved it in test_new dataframe

Correlation Plot:



Trained Logistic regression model on the training dataset:

```
> logreg <- glm(delay ~ ., family = binomial(link = 'logit'), data = train)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(logreg)

Call:
glm(formula = delay ~ ., family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1062  -0.5531  -0.4395  -0.3210   8.4904

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.038e+01  3.441e+00  -3.017  0.00256 **
schedtime    -2.076e-02  2.063e-03 -10.066  < 2e-16 ***
carrier       2.181e-01  7.172e-02   3.041  0.00236 **
deptime       2.165e-02  2.049e-03  10.569  < 2e-16 ***
dest          3.044e-01  2.116e-01   1.438  0.15043
distance      2.852e-02  1.690e-02   1.688  0.09142 .
flightnumber  1.602e-04  6.384e-05   2.509  0.01212 *
origin       -4.963e-01  3.888e-01  -1.277  0.20169
weather       1.665e+01  4.349e+02   0.038  0.96945
dayweek      -5.223e-02  4.268e-02  -1.224  0.22100
daymonth      1.632e-02  1.043e-02   1.565  0.11763
tailnu        1.420e-03  6.159e-04   2.305  0.02116 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1301.85  on 1320  degrees of freedom
Residual deviance:  968.83  on 1309  degrees of freedom
AIC: 992.83

Number of Fisher Scoring iterations: 15
```

Calculating the probabilities by fitting the model to the test data.
Calculating the predictions using a threshold of 0.3 .

```
> prob <- logreg %>% predict(test_new, type = "response")
> test_new$prob = prob
> test_new <- test_new %>%
+   mutate(predicted = ifelse(prob<0.3,0,1))
> head(test_new)
   schedtime carrier deptime dest distance flightnumber origin weather dayweek daymonth
3       1245       2    1245    2      229          7208      3       0       4        1
4       1715       2    1709    2      229          7215      3       0       4        1
9       1715       2    1710    1      228          7812      3       0       4        1
11      2120       2    2114    2      229          7924      3       0       4        1
12      1455       3    1458    1      213           746      2       0       4        1
13       930       3     932    2      214          1746      2       0       4        1
   tailnu       prob predicted
3       3 0.09815537         0
4       4 0.12708537         0
9       9 0.10569568         0
11     11 0.19107544         0
12     12 0.04589189         0
13     13 0.04613564         0
>
```

Confusion Matrix:

```
     0   1
0  672  37
1   65 106
```

Metrics evaluation from the confusion matrix:

```
Accuracy:  88.40909 %
Error Rate:  11.59091 %
Precision:  94.78138 %
Recall:  91.18046 %
>
```

ROC Curve: