Final Project

2023-05-26

R Markdown

```
#THIS FILE IS MLR ONLY
FixedAirlineData<-read.csv("CleanAccurate2019.csv")
x<-FixedAirlineData[c('DEP_DELAY','DEP_TIME','TAXI_OUT','TAXI_IN','ARR_TIME',
'AIR_TIME','DISTANCE','CARRIER_DELAY',
                      'WEATHER_DELAY','NAS_DELAY','SECURITY_DELAY',
                      'LATE_AIRCRAFT_DELAY','ARR_DELAY','ProperDepartureTimes
6',
                      'ProperArrivalTimes6')]
y<-FixedAirlineData[c('ARR_DELAY','ProperArrivalTimes6')]

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

# Splitting the data into training and test sets
set.seed(122515)  # Equivalent to random_state in Python

train_indices <- createDataPartition(y, p = 0.7, list = FALSE)  # 70% for tra
ining

## Warning in createDataPartition(y, p = 0.7, list = FALSE): Some classes hav
e no
## records ( ) and these will be ignored

## Warning in createDataPartition(y, p = 0.7, list = FALSE): Some classes hav
e a
## single record ( ) and these will be selected for the sample

xtrain <- x[train_indices, ]
ytrain <- y[train_indices]

xtest <- x[-train_indices, ]
ytest <- y[-train_indices]

# Printing the shapes of the training and test sets
cat("xtrain shape:", dim(xtrain), "\n")

## xtrain shape: 2 15

cat("xtest shape:", dim(xtest), "\n")

## xtest shape: 1027640 15
```

```
cat("ytrain shape:", length(ytrain), "\n")

## ytrain shape: 2

cat("ytest shape:", length(ytest), "\n")

## ytest shape: 2055282

head(train_indices,10)

##      Resample1
## [1,]         1
## [2,]         2

head(ytest,10)

##  [1] " -16" " -14" " -25" " -19" "   9" "   3" " -22" " -14" "  -7" " -32"

head(ytrain,10)

## [1] "  -1" " -36"

head(xtrain,10)

##   DEP_DELAY DEP_TIME TAXI_OUT TAXI_IN ARR_TIME AIR_TIME DISTANCE CARRIER_D
ELAY
## 1         1      601       22       8      722       51      300
0
## 2        -5     1359       15       4     1633       75      596
0
##   WEATHER_DELAY NAS_DELAY SECURITY_DELAY LATE_AIRCRAFT_DELAY ARR_DELAY
## 1             0         0              0                   0        -1
## 2             0         0              0                   0       -36
##   ProperDepartureTimes6 ProperArrivalTimes6
## 1                 06:00               07:23
## 2                 14:04               17:09

head(xtest,10)

##    DEP_DELAY DEP_TIME TAXI_OUT TAXI_IN ARR_TIME AIR_TIME DISTANCE CARRIER_
DELAY
## 3         -5     1215       18       6     1329       50      229
0
## 4         -6     1521       14       7     1625       43      223
0
## 5        -15     1847       18       5     1940       90      579
0
## 6         -7      853       25       5      953       90      574
0
## 7         -5     1553       33      14     1832       52      341
0
## 8         -4     1551       31       8     1824      114      585
0
```

```
## 9         -8     1037       17       4     1239      101      833
0
## 10         0     1245       15       2     1318       76      533
0
## 11        -5     1410       22       5     1700       83      533
0
## 12        -3      557       10       6      737       84      528
0
##    WEATHER_DELAY NAS_DELAY SECURITY_DELAY LATE_AIRCRAFT_DELAY ARR_DELAY
## 3              0         0              0                   0       -16
## 4              0         0              0                   0       -14
## 5              0         0              0                   0       -25
## 6              0         0              0                   0       -19
## 7              0         0              0                   0         9
## 8              0         0              0                   0         3
## 9              0         0              0                   0       -22
## 10             0         0              0                   0       -14
## 11             0         0              0                   0        -7
## 12             0         0              0                   0       -32
##    ProperDepartureTimes6 ProperArrivalTimes6
## 3                  12:20               13:45
## 4                  15:27               16:39
## 5                  19:02               20:05
## 6                  09:00               10:12
## 7                  15:58               18:23
## 8                  15:55               18:21
## 9                  10:45               13:01
## 10                 12:45               13:32
## 11                 14:15               17:07
## 12                 06:00               08:09
```

```r
# Import the required library
library(caret)

# Create a linear regression model object
lm_model <- lm(ytrain ~ ., data = xtrain)

# Print the model summary
summary(lm_model)
```

```
##
## Call:
## lm(formula = ytrain ~ ., data = xtrain)
##
## Residuals:
## ALL 2 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (14 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -6.833        NaN     NaN      NaN
```

```
## DEP_DELAY                       5.833         NaN      NaN       NaN
## DEP_TIME                           NA          NA       NA        NA
## TAXI_OUT                           NA          NA       NA        NA
## TAXI_IN                            NA          NA       NA        NA
## ARR_TIME                           NA          NA       NA        NA
## AIR_TIME                           NA          NA       NA        NA
## DISTANCE                           NA          NA       NA        NA
## CARRIER_DELAY                      NA          NA       NA        NA
## WEATHER_DELAY                      NA          NA       NA        NA
## NAS_DELAY                          NA          NA       NA        NA
## SECURITY_DELAY                     NA          NA       NA        NA
## LATE_AIRCRAFT_DELAY                NA          NA       NA        NA
## ARR_DELAY                          NA          NA       NA        NA
## ProperDepartureTimes614:04         NA          NA       NA        NA
## ProperArrivalTimes617:09           NA          NA       NA        NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared: 1
## F-statistic:   1.4901e+05 on 1 and 0 DF,  p-value: 2.2e16

intercept <- coef(lm_model)[1]
coefficients <- coef(lm_model)

# Print the intercept term
cat("Intercept:", intercept, "\n")

## Intercept: -6.833333

print(coefficients)

##                 (Intercept)                   DEP_DELAY
##                   -6.833333                    5.833333
##                    DEP_TIME                    TAXI_OUT
##                          NA                          NA
##                     TAXI_IN                    ARR_TIME
##                          NA                          NA
##                    AIR_TIME                    DISTANCE
##                          NA                          NA
##               CARRIER_DELAY               WEATHER_DELAY
##                          NA                          NA
##                   NAS_DELAY              SECURITY_DELAY
##                          NA                          NA
##         LATE_AIRCRAFT_DELAY                   ARR_DELAY
##                          NA                          NA
## ProperDepartureTimes614:04   ProperArrivalTimes617:09
##                          NA                          NA

xtrain$ProperDepartureTimes6 <- as.character(xtrain$ProperDepartureTimes6)
xtest$ProperDepartureTimes6 <- factor(xtest$ProperDepartureTimes6, levels = l
evels(xtrain$ProperDepartureTimes6))
xtrain$ProperArrivalTimes6 <- as.character(xtrain$ProperArrivalTimes6)
```

```
xtest$ProperArrivalTimes6 <- factor(xtest$ProperArrivalTimes6, levels = level
s(xtrain$ProperArrivalTimes6))
b <- predict(lm_model, newdata = xtest)

## Warning in predict.lm(lm_model, newdata = xtest): prediction from a
## rank-deficient fit may be misleading

head(b,10)

##          3          4          5          6          7          8
9
## -36.000000 -41.833333 -94.333333 -47.666667 -36.000000 -30.166667 -53.5000
00
##         10         11         12
##  -6.833333 -36.000000 -24.333333

missing_values <- is.na(ytest) | is.na(b)

## Warning in is.na(ytest) | is.na(b): longer object length is not a multiple
of
## shorter object length

ytest <- as.numeric(ytest[!missing_values])

## Warning: NAs introduced by coercion

b <- as.numeric(b[!missing_values])
rss <- sum((ytest - b)^2)  # Residual sum of squares
tss <- sum((ytest - mean(ytest))^2)  # Total sum of squares
r_squared <-1-(100-(rss / tss))/100

# Print the R-squared value
print(r_squared)

## [1] 0.2173564

# Calculate MSE
mse <- mean((ytest - b)^2)

# Calculate RMSE
rmse <- sqrt(mse)

# Print the RMSE value
print(rmse)

## [1] 246.0989

FixedAirlineData$ProperDepartureTimes6 <- as.character(FixedAirlineData$Prope
rDepartureTimes6)
FixedAirlineData$ProperDepartureTimes6 <- factor(FixedAirlineData$ProperDepar
tureTimes6, levels = levels(FixedAirlineData$ProperDepartureTimes6))
FixedAirlineData$ProperArrivalTimes6 <- as.character(FixedAirlineData$ProperA
```

```
rrivalTimes6)
FixedAirlineData$ProperArrivalTimes6 <- factor(FixedAirlineData$ProperArrival
Times6, levels = levels(FixedAirlineData$ProperArrivalTimes6))
FixedAirlineData$ARR_DELAY <- as.numeric(as.character(FixedAirlineData$ARR_DE
LAY))
a = predict(lm_model, newdata=FixedAirlineData)

## Warning in predict.lm(lm_model, newdata = FixedAirlineData): prediction fr
om a
## rank-deficient fit may be misleading

head(a,10)

##          1          2          3          4          5          6
7
##  -1.000000 -36.000000 -36.000000 -41.833333 -94.333333 -47.666667 -36.0000
00
##          8          9         10
## -30.166667 -53.500000  -6.833333

mlr_Airline=FixedAirlineData
mlr_Airline['MLR_Prediction']=a
head(mlr_Airline['MLR_Prediction'],10)

##    MLR_Prediction
## 1       -1.000000
## 2      -36.000000
## 3      -36.000000
## 4      -41.833333
## 5      -94.333333
## 6      -47.666667
## 7      -36.000000
## 8      -30.166667
## 9      -53.500000
## 10      -6.833333

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

Airline_Results <- mlr_Airline %>%
  filter(MLR_Prediction == a) %>%
  select(OP_UNIQUE_CARRIER, ORIGIN, DEST, MLR_Prediction) %>%
```

```
  arrange(OP_UNIQUE_CARRIER)

head(Airline_Results,10)

##    OP_UNIQUE_CARRIER ORIGIN DEST MLR_Prediction
## 1                 9E    GNV  ATL      -1.000000
## 2                 9E    MSP  CVG     -36.000000
## 3                 9E    DTW  CVG     -36.000000
## 4                 9E    TLH  ATL     -41.833333
## 5                 9E    ATL  FSM     -94.333333
## 6                 9E    DAY  MSP     -47.666667
## 7                 9E    JAN  ATL     -36.000000
## 8                 9E    LGA  CVG     -30.166667
## 9                 9E    JAX  LGA     -53.500000
## 10                9E    ATL  BMI      -6.833333

positive_valuesMLR <- Airline_Results$MLR_Prediction[Airline_Results$MLR_Pred
iction >= 0]
negative_valuesMLR <- Airline_Results$MLR_Prediction[Airline_Results$MLR_Pred
iction < 0]

length(positive_valuesMLR)

## [1] 320233

length(negative_valuesMLR)

## [1] 676152

percentnegatvieMLR<-length(negative_valuesMLR)/(length(negative_valuesMLR)+le
ngth(positive_valuesMLR))
print(percentnegatvieMLR)

## [1] 0.6786052

print(1-percentnegatvieMLR)

## [1] 0.3213948
```