Flight\_Delay\_Prediction

JPOLANKI

2024-05-11

## Motivation of the Project

##Alternative Hypothesis:

The alternative hypothesis for the project is to address the critical issue of predicting flight delays exceeding 120 minutes using historical data on carriers, airports, and other relevant factors.

Flight delays have significant consequences for both airlines and passengers. Airlines face increased operational costs, compromised scheduling, and diminished customer loyalty due to delays, while passengers experience inconvenience, potential missed connections, and reduced overall satisfaction with their travel experience. Accurate delay predictions enable proactive resource management for airlines and informed decision-making for travelers, contributing to operational efficiency, customer satisfaction, and advancement in the air travel industry.

## Methodology

The project focuses on utilizing historical flight data to predict flight delays, employing R packages like tidyverse, CARET, and ggplot2 for comprehensive data analysis and visualization. Through exploratory data analysis and preprocessing, anomalies are identified and data is prepared for modeling. Predictive models, including logistic regression, decision trees, and random forests, are then developed and evaluated for accuracy. The insights derived from these models are presented through an intuitive dashboard created using Shiny, aiming to improve decision-making in the aviation sector.

##Setting up a knitr File setting up knitr file for the project and not including the Knitr file code into the end knitted document.

## Installing R Packages

install.packages(“readr”) install.packages(“ggplot2”) install.packages(“dplyr”) install.packages(“tidyr”) install.packages(“Caret”) install.packages(“stargazer”) install.packages(“randomForestExplainer”) install.packages(“randomForest”) install.packages(“rpart”) install.packages(“rpart.plot”) install.packages(“pscl”) install.packages(“fmsb”)

library(readr)  
library(ggplot2)  
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

library(tidyr)  
library(caret)

## Loading required package: lattice

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library(randomForest)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(randomForestExplainer)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(rpart)  
library(rpart.plot)  
library(fmsb)

## Registered S3 methods overwritten by 'fmsb':  
## method from  
## print.roc pROC  
## plot.roc pROC

## Importing the DATA into rmd file

This Data set includes Historical flight data such as carrier Information which airport and type of delays and this was sourced from Bureau of Transportation statistics.

Flight\_Data <- read.csv("G:/My Drive/CIS 663 R programming/ATC\_R\_Project/DATA/air\_traffic\_main\_Data.csv")

##Over view of the DATA After the Data was imported, to get the overview of the DATA like Descriptive statistics like Min, MAX, AVG of the columns and basic structure of the Data along the with Null Values.

glimpse(Flight\_Data)

## Rows: 3,351  
## Columns: 21  
## $ year <int> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 20…  
## $ month <int> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12…  
## $ carrier <chr> "9E", "9E", "9E", "9E", "9E", "9E", "9E", "9E", "9…  
## $ carrier\_name <chr> "Endeavor Air Inc.", "Endeavor Air Inc.", "Endeavo…  
## $ airport <chr> "ABE", "ABY", "AEX", "AGS", "ALB", "ATL", "ATW", "…  
## $ airport\_name <chr> "Allentown/Bethlehem/Easton, PA: Lehigh Valley Int…  
## $ arr\_flights <int> 44, 90, 88, 184, 76, 5985, 142, 147, 84, 150, 123,…  
## $ arr\_del15 <int> 3, 1, 8, 9, 11, 445, 14, 10, 14, 19, 9, 7, 5, 1, 3…  
## $ carrier\_ct <dbl> 1.63, 0.96, 5.75, 4.17, 4.78, 142.89, 5.36, 6.04, …  
## $ weather\_ct <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 11.96, 0.00, 1.00, 0…  
## $ nas\_ct <dbl> 0.12, 0.04, 1.60, 1.83, 5.22, 161.37, 7.70, 1.00, …  
## $ security\_ct <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ late\_aircraft\_ct <dbl> 1.25, 0.00, 0.65, 3.00, 1.00, 127.79, 0.94, 1.96, …  
## $ arr\_cancelled <int> 0, 0, 0, 0, 1, 5, 1, 0, 1, 3, 0, 2, 0, 0, 0, 0, 0,…  
## $ arr\_diverted <int> 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1,…  
## $ arr\_delay <int> 89, 23, 338, 508, 692, 30756, 436, 1070, 2006, 846…  
## $ carrier\_delay <int> 56, 22, 265, 192, 398, 16390, 162, 838, 1164, 423,…  
## $ weather\_delay <int> 0, 0, 0, 0, 0, 1509, 0, 141, 619, 0, 0, 0, 0, 0, 8…  
## $ nas\_delay <int> 3, 1, 45, 92, 178, 5060, 182, 24, 223, 389, 26, 93…  
## $ security\_delay <int> 0, 0, 0, 0, 0, 16, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
## $ late\_aircraft\_delay <int> 30, 0, 28, 224, 116, 7781, 92, 67, 0, 34, 19, 172,…

summary(Flight\_Data)

## year month carrier carrier\_name   
## Min. :2019 Min. :12 Length:3351 Length:3351   
## 1st Qu.:2019 1st Qu.:12 Class :character Class :character   
## Median :2019 Median :12 Mode :character Mode :character   
## Mean :2019 Mean :12   
## 3rd Qu.:2020 3rd Qu.:12   
## Max. :2020 Max. :12   
##   
## airport airport\_name arr\_flights arr\_del15   
## Length:3351 Length:3351 Min. : 1.0 Min. : 0   
## Class :character Class :character 1st Qu.: 35.0 1st Qu.: 5   
## Mode :character Mode :character Median : 83.0 Median : 12   
## Mean : 298.3 Mean : 51   
## 3rd Qu.: 194.5 3rd Qu.: 33   
## Max. :19713.0 Max. :2289   
## NA's :8 NA's :8   
## carrier\_ct weather\_ct nas\_ct security\_ct   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. : 0.0000   
## 1st Qu.: 1.49 1st Qu.: 0.000 1st Qu.: 0.82 1st Qu.: 0.0000   
## Median : 4.75 Median : 0.060 Median : 2.98 Median : 0.0000   
## Mean : 16.07 Mean : 1.443 Mean : 16.18 Mean : 0.1373   
## 3rd Qu.: 12.26 3rd Qu.: 1.010 3rd Qu.: 8.87 3rd Qu.: 0.0000   
## Max. :697.00 Max. :89.420 Max. :1039.54 Max. :17.3100   
## NA's :8 NA's :8 NA's :8 NA's :8   
## late\_aircraft\_ct arr\_cancelled arr\_diverted arr\_delay   
## Min. : 0.00 Min. : 0.000 Min. : 0.0000 Min. : 0   
## 1st Qu.: 0.90 1st Qu.: 0.000 1st Qu.: 0.0000 1st Qu.: 230   
## Median : 3.28 Median : 0.000 Median : 0.0000 Median : 746   
## Mean : 17.17 Mean : 2.885 Mean : 0.5758 Mean : 3334   
## 3rd Qu.: 10.24 3rd Qu.: 2.000 3rd Qu.: 0.0000 3rd Qu.: 2096   
## Max. :819.66 Max. :224.000 Max. :42.0000 Max. :160383   
## NA's :8 NA's :8 NA's :8 NA's :8   
## carrier\_delay weather\_delay nas\_delay security\_delay   
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 68.5 1st Qu.: 0.0 1st Qu.: 21.5 1st Qu.: 0.000   
## Median : 272.0 Median : 3.0 Median : 106.0 Median : 0.000   
## Mean : 1144.8 Mean : 177.6 Mean : 749.6 Mean : 5.401   
## 3rd Qu.: 830.5 3rd Qu.: 82.0 3rd Qu.: 362.0 3rd Qu.: 0.000   
## Max. :55215.0 Max. :14219.0 Max. :82064.0 Max. :553.000   
## NA's :8 NA's :8 NA's :8 NA's :8   
## late\_aircraft\_delay  
## Min. : 0   
## 1st Qu.: 31   
## Median : 205   
## Mean : 1257   
## 3rd Qu.: 724   
## Max. :75179   
## NA's :8

## Column\_Names

Changed the column names to more readable format, which makes easy to remember and work with the column for the further operations.

colnames(Flight\_Data) <- c("Year",  
 "Month", "Carrier",  
 "Carrier\_Name", "Airport",   
 "Airport\_Name",  
 "Arrival\_Flights",  
"Arrival\_Delay\_Flights",  
"Carrier\_Delayed\_Flights",   
"Weather\_Delayed\_Flights",  
"NAS\_Delayed\_Flights",  
 "Security\_Delayed\_Flights",  
 "Late\_Aircraft\_Delayed\_Flights",  
 "Arrivals\_Cancelled",  
 "Arrivals\_Diverted",  
 "Total\_Arrival\_Delay",  
 "Carrier\_Delay\_Time",  
 "Weather\_Delay\_Time",   
 "NAS\_Delay\_Time",   
 "Security\_Delay\_Time",  
 "Late\_Aircraft\_Delay\_Time")  
  
names(Flight\_Data)

## [1] "Year" "Month"   
## [3] "Carrier" "Carrier\_Name"   
## [5] "Airport" "Airport\_Name"   
## [7] "Arrival\_Flights" "Arrival\_Delay\_Flights"   
## [9] "Carrier\_Delayed\_Flights" "Weather\_Delayed\_Flights"   
## [11] "NAS\_Delayed\_Flights" "Security\_Delayed\_Flights"   
## [13] "Late\_Aircraft\_Delayed\_Flights" "Arrivals\_Cancelled"   
## [15] "Arrivals\_Diverted" "Total\_Arrival\_Delay"   
## [17] "Carrier\_Delay\_Time" "Weather\_Delay\_Time"   
## [19] "NAS\_Delay\_Time" "Security\_Delay\_Time"   
## [21] "Late\_Aircraft\_Delay\_Time"

## Filtering the DATA

From the code above summary statistics shows the outliers of the data in few delays. Filtering the data to take out the outliers and capped the data to 720 minutes (12 Hours) in the Total\_Arrival\_Delay column(sum of all the delays).

Delays <- Flight\_Data %>%  
 filter(Total\_Arrival\_Delay <= 720)  
  
dim(Delays)

## [1] 1643 21

## DATA Reshaping for the Exploratory Data Analysis

Reshaped the Data by gather function which forms a new column i.e., Type\_Of\_Delay consists of all the delay types that exists as each column in the data and their respective values in the Delay\_Minutes column.

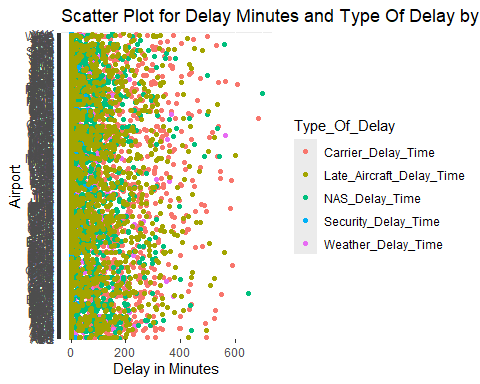
Delays\_Reshaped <- Delays %>%  
 gather(Type\_Of\_Delay,Delay\_Minutes,Carrier\_Delay\_Time,Weather\_Delay\_Time,NAS\_Delay\_Time,  
 Security\_Delay\_Time,Late\_Aircraft\_Delay\_Time)

## DATA PLOTTING for Exploratory Data Analysis

##Plotting Data by Airport

Plotting the distribution of delay minutes for each airport based on delay type enables a visual understanding of delay patterns.

ggplot(Delays\_Reshaped, aes(x = Airport, y = Delay\_Minutes, color = Type\_Of\_Delay)) +  
 geom\_point() +  
 coord\_flip() +  
 labs(x = "Airport", y = "Delay in Minutes", title = "Scatter Plot for Delay Minutes and Type Of Delay by Airport")

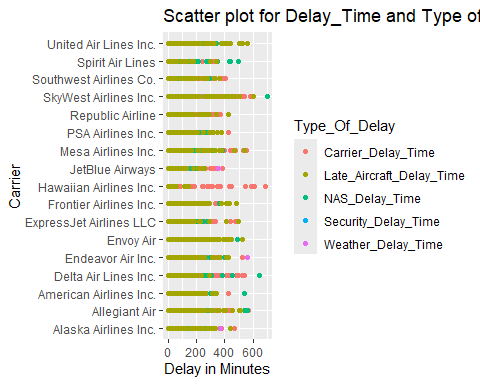


From the above scatter\_plot it is visible that Security\_Delay is mostly around the 0 Minutes and rest of the values are near to 0 with no significant outliers.Even though, NAS\_Delay’s (“National Airspace System”) are near to 0 but have have notable No.of values away from 0 and Carrier\_Delay and Weather\_Delay are mostly scattered away from Zero and have extreme outliers in the column when plotted for each Airport.

## Plotting Data by each CARRIER.

Plotting the distribution of delay minutes for each Carrier based on delay type enables a visual understanding of delay patterns

ggplot(Delays\_Reshaped, aes(x = Carrier\_Name, y = Delay\_Minutes, color = Type\_Of\_Delay)) +  
 geom\_point() +  
 coord\_flip() +  
 labs(x = "Carrier", y = "Delay in Minutes", title = "Scatter plot for Delay\_Time and Type of Delay for Carrier")



##Exploratory DATA Analysis The Exploratory Data Analysis revealed key insights into the factors contributing to flight delays, such as carrier performance, weather conditions and NAS delays. In this of distribution of Flight Delays a scatterplot was created and its analysis demonstrated that significant number of flights experience minimum number of flight delays indicating efficient operations across many flights. This Scatter plot on delays by carriers indicated variations in delay times across different carriers. This insight could help airline to benchmark performance and identify areas for operational improvements. The analysis on impact of weather and NAS Delays have a noticeable impact on flight punctuality.

## Adding the column

adding a new column called “Delay\_120” to the data set. If a flight’s total arrival delay is more than 120 minutes, it gets marked as 1 in this column. Otherwise, it’s marked as 0. This helps easily identify longer delays for analysis.

# Create the new column indicating total arrival delay >= 120 minutes  
Delays$Delay\_120 <- ifelse(Delays$Total\_Arrival\_Delay >= 120, 1, 0)  
  
# Converting a numeric variable into factor variable to perform regression analysis and for building a predictive model.  
Delays$Delay\_120 <- as.factor(Delays$Delay\_120)  
  
is.factor(Delays$Delay\_120)

## [1] TRUE

##DATA Preparation After ensuring the variable as factor, preparing the data for CARET’s Train/test split.

set.seed(123) #for Reproducibility  
index <- createDataPartition(Delays$Total\_Arrival\_Delay, p=0.8, list = FALSE)  
  
Trainingset <- Delays[index, ]  
Testingset <- Delays[-index, ]  
  
#setting up training control for cross-validation  
trainControl <- trainControl(method = "cv", number = 10)

##Regression Models

library(randomForest)  
  
  
# Logistic Regression  
logistic\_model <- glm(Delay\_120 ~ Carrier\_Delay\_Time + NAS\_Delay\_Time   
 + Weather\_Delay\_Time + Late\_Aircraft\_Delay\_Time,  
 data = Trainingset, family = binomial)  
  
# Decision Tree  
tree\_model <- rpart(Delay\_120 ~ Carrier\_Delay\_Time + NAS\_Delay\_Time +  
 Weather\_Delay\_Time + Late\_Aircraft\_Delay\_Time,  
 data = Trainingset, method = "class")  
  
# Random Forest  
forest\_model <- randomForest(Delay\_120 ~ Carrier\_Delay\_Time + NAS\_Delay\_Time +  
 Weather\_Delay\_Time + Late\_Aircraft\_Delay\_Time,  
 data = Trainingset)

##Testing Testing the accuracy of the Models.

# Predictions on Testing Set  
logistic\_pred <- predict(logistic\_model, newdata = Testingset, type = "response")  
  
tree\_pred <- predict(tree\_model, newdata = Testingset, type = "class")  
  
  
forest\_pred <- predict(forest\_model, newdata = Testingset)

##Accuracy

# Logistic\_Regression\_Accuracy  
logistic\_accuracy <- mean(ifelse(logistic\_pred > 0.5, 1, 0) == Testingset$Delay\_120)  
  
cat("Logistic Regression Accuracy:", logistic\_accuracy, "\n")

## Logistic Regression Accuracy: 0.9938838

#Decision\_Tree\_Accuracy  
tree\_accuracy <- mean(tree\_pred == Testingset$Delay\_120)  
  
cat("Decision Tree Accuracy:", tree\_accuracy, "\n")

## Decision Tree Accuracy: 0.9541284

#Random\_Forest\_Accuracy  
forest\_accuracy <- mean(forest\_pred == Testingset$Delay\_120)  
cat("Random Forest Accuracy:", forest\_accuracy, "\n")

## Random Forest Accuracy: 0.9663609

Logistic Regression model has an accuracy of 99.38% the decision tree model has an accuracy of 95.41% and the random forest model has an accuracy of 0%. Among these, the Logistic regression has Highest accuracy in predicting flight delays.

# Creating a table for the Co-efficients and Standard errors in the Logistic regression.  
  
library(stargazer)  
  
stargazer(logistic\_model, type = "text", title = "Logistic")

##   
## Logistic  
## ====================================================  
## Dependent variable:   
## ---------------------------  
## Delay\_120   
## ----------------------------------------------------  
## Carrier\_Delay\_Time 0.249\*\*\*   
## (0.047)   
##   
## NAS\_Delay\_Time 0.276\*\*\*   
## (0.054)   
##   
## Weather\_Delay\_Time 0.252\*\*\*   
## (0.060)   
##   
## Late\_Aircraft\_Delay\_Time 0.255\*\*\*   
## (0.048)   
##   
## Constant -30.548\*\*\*   
## (5.769)   
##   
## ----------------------------------------------------  
## Observations 1,316   
## Log Likelihood -25.775   
## Akaike Inf. Crit. 61.550   
## ====================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

library(pscl)

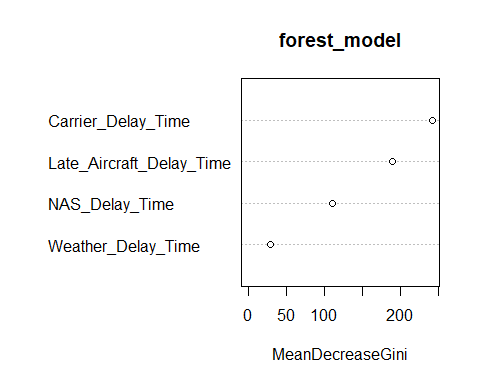
## Classes and Methods for R originally developed in the  
## Political Science Computational Laboratory  
## Department of Political Science  
## Stanford University (2002-2015),  
## by and under the direction of Simon Jackman.  
## hurdle and zeroinfl functions by Achim Zeileis.

pseudo\_r\_squared <- pR2(logistic\_model)

## fitting null model for pseudo-r2

stargazer(pseudo\_r\_squared,type ="text", title = "R2")

##   
## R2  
## ===============================================  
## llh llhNull G2 McFadden r2ML r2CU   
## -----------------------------------------------  
## -25.775 -824.114 1,596.677 0.969 0.703 0.984  
## -----------------------------------------------



## Analysis Report:

## Logistic Regression Results:

The logistic regression analysis revealed insights into the relationship between delay durations exceeding 120 minutes and various delay types. Coefficients in the table indicate the impact of each delay type on the likelihood of experiencing a long delay. For instance, a positive coefficient such as 0.249 for Carrier\_Delay\_Time suggests that as carrier delays increase by one unit, the likelihood of a long delay increases by 0.249 units. Conversely, negative coefficients like -30.548 for the constant term indicate a decrease in the likelihood of long delays. Smaller coefficients signify a lesser impact of the delay type on the likelihood of long delays.

# Pseudo 𝑅2 Analysis:

Comparison of the logistic regression model with a null model provided insights into the model’s predictive performance. Pseudo-𝑅2 measures, such as McFadden’s 𝑅2 (0.969), Maximum Likelihood 𝑅2 (0.703), and Cox & Snell 𝑅2 (0.984), assessed the proportion of variance explained by the model compared to the null model. Higher pseudo-R2 values indicate a better fit of the model to the data.

## Conclusion:

The logistic regression analysis revealed significant associations between delay types and the likelihood of long delays. For example, Carrier\_Delay\_Time had a coefficient of 0.249, indicating its positive impact on the likelihood of long delays. The pseudo-𝑅2 measures demonstrated the model’s effectiveness in explaining the variance in delay duration. These findings offer valuable insights for operational improvements aimed at minimizing long delays.

## References

Loura, J. (2014). Human factors and workload in air traffic control operations-a review of literature. International Journal of Management and Social Sciences Research, 3(3), 1-5. Hilburn, B. (2004). Cognitive complexity in air traffic control: A literature review. EEC note, 4(04), 1-80.

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. Transportation research part C: Emerging technologies, 44, 231-241.

Pamplona, D. A., & Pinto Alves, C. J. (2019). Mitigating Air Delay: An analysis of the Collaborative Trajectory Options Program. Engineering, Technology & Applied Science Research, 9(3).

Xing, J., & Manning, C. A. (2005). Complexity and automation display of air traffic control: Literature review and analysis.

Kontogiannis, T., & Malakis, S. (2009). A proactive approach to human error detection and identification in aviation and air traffic control. Safety Science, 47(5), 693-706.

Osetrov, E. (2023, October 10). Airline delays. Kaggle. <https://www.kaggle.com/datasets/eugeniyosetrov/airline-delays?resource=download>.

Controller workload (Arad, 1964; Grossberg, 1989; Redding, 1992; Athènes, S., Chaboud et al., 2000) A-Z index. A-Z Index | Bureau of Transportation Statistics. (n.d.). <https://www.bts.gov/A-Z-Index> Team, T. H. (n.d.). The Epidemiologist R Handbook. 42 Dashboards with R Markdown. <https://epirhandbook.com/en/dashboards-with-r-markdown.html>