ORD Flight Delays Methodology

Aden Ramirez

February 24, 2023

1 Research Question 1

Can flight delays be predicted?

For this question we want a good overall look at stages of a flight, taxi, departure, landing, taxi, and arrival. The goal is to categorize if a flight will be delayed, this is useful for airlines and people for expectations and planning around it. However there should also be an attempt at predicting the length of the delay to make the information more meaningful. This question will be solved with three different methods.

1.1 Method 1

The first method for this question that will be applied is logisitic regression. This method will have two possible categorical outputs, being the flight is delayed or the flight is not delayed (Using >15m FAA definition). Logisitic regression is great for two class problems, and will fit well here.

1.2 Method 2

The second method for this research question is also a random forest decision tree, but using more categorical outputs. These being our delay groups that are used in the dataset. They are not delayed, 15 minutes - 30 minutes, and builds all 15 minute increments up to 180 minuts for a total of 12 possible groups. The power of the decision tress will allow to go beyond saying a flight is delayed or not, and offer a range following the ranges the DOT provides.

2 Research Question 2

What role do airports play in flight delays?

For this question focus is put closer to the airport and the delays themselves. To determine an airports fault there is a couple categories such as security and late aircraft that can be considered airport delays, but also adding a new type with the taxi time at each airport. Getting this information I can predict the categories of delays.

2.1 Method 1

The first method for this question that will be applied is logisitic regression. This method will have two possible categorical outputs, being the airport is the main delay or other.

2.2 Method 2

The second method being applied for this research question is a random forest decision tree. This method will output multiple categories corresponding to the type fo delay possible.

3 Research Question 3

Can airport congestion be predicted based on flight delays?

This final question puts a focus on the operations around a flight, specifically late aircraft, taxi times, and departure delays. The answers for this question should be if the airport is congested (or busy) during that flight. If most of the flights are during congested times the airport can generally be considered congested.

3.1 Method 1

The first method being applied for this research question is logistic regression. This method will have two possible categorical outputs, being the airport is congested with many flight delays or not congested.

3.2 Method 2

The second method being applied for this research question is a random forest decision tree. This method will have two possible categorical outputs, being the airport is congested with many flight delays or not congested.

4 Validation

With the methods being used, there will need to be many validation methods used. For logisitc regression there will need to be simple classification measures such as accuracy and precision that can be determined testing the model on a subset of the data. With this info sensitivity and specificity can also be calculated, and all reported in a table before the next method. An ROC curve will also be used as a nice visual for these values calculated. For decision trees there will be variable inmportance measures and predictions verfied on a subset of the data. Both will start with 75% of the data being used for training and the last 25% to be used in validation.

5 Data

For the models in this research question we will provide:
Airline
Origin
Departure Block
Departure Delay
Destination
Arrival Block
Arrival Delay (Time in minutes)
Arrival Delay (True or False)
Distance
Carrier Delay
Late Aircraft Delay
Air System Delay
Security Delay
Weather Delay
Taxi Time