Individual Learner

US Domestic Flight Delays

Ian Gledhill

12/06/2019

Introduction

Overview

This project will look at the delayed flight data recorded in the US for the 3 months from January to March 2018 and use data analysis to create a model to predict whether a flight would be delayed from its scheduled departure time. The data is provided by the US Department of Transport and covers 18 carriers operating from 20 airports.

Dataset

Source

The data is provided by:

https://www.transtats.bts.gov

BUREAU OF TRANSPORTATION STATISTICS

U.S. Department of Transportation

1200 New Jersey Avenue, SE

Washington, DC 20590

855-368-4200

The page to download the data used in this project can be found here:

https://www.transtats.bts.gov/Fields.asp?Table_ID=236

Sample Size for Analysis

The DoT record more than 500,000 domestic flights per month. The period covered by this project contains 100,000 data points. Each data point has 35 measurements.

The original data has 1,702,836 observations. For the purposes of expediency a sample of 1e+05 has been randomly selected from the original data. In addition only the top 20 airports based on the number of departures over the data period have been analzed. These can be seen in Appendix 1.

Shape of the Data

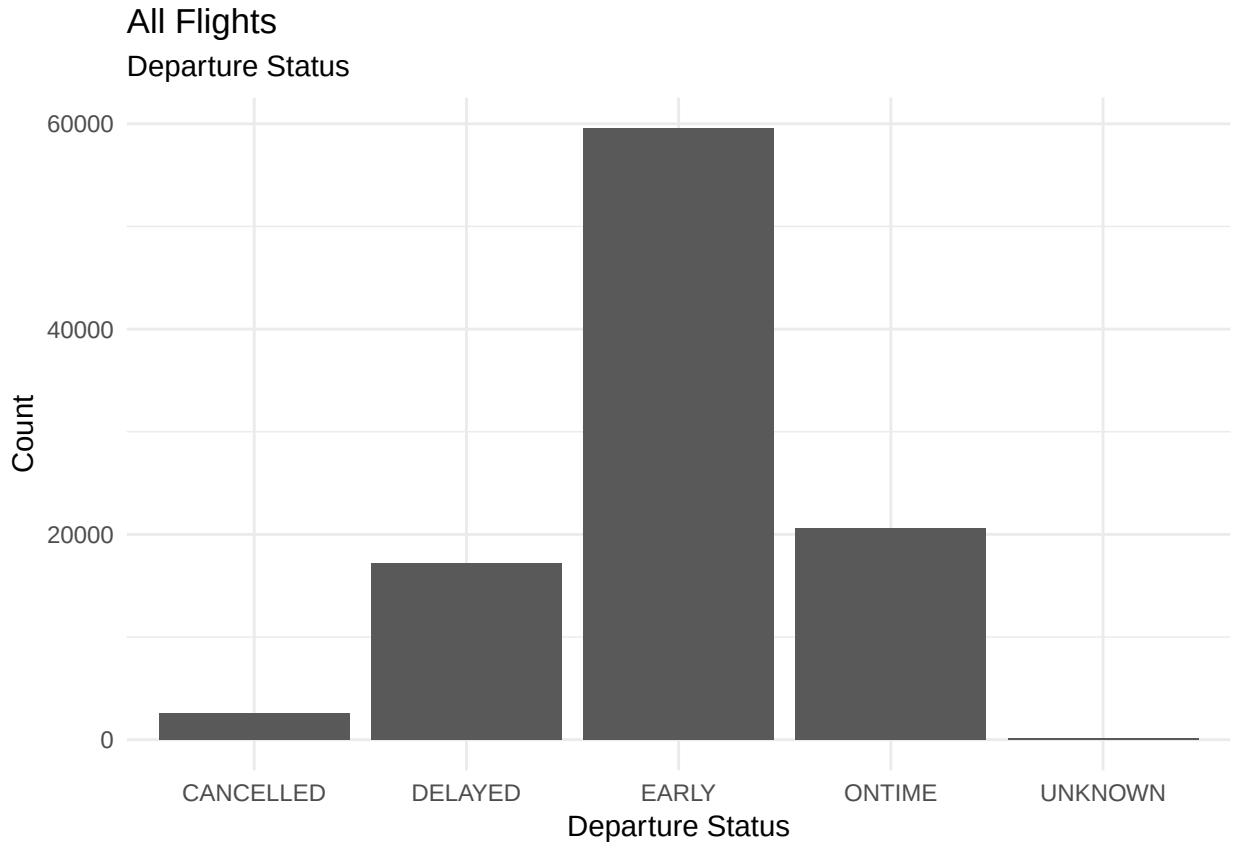
The data that the DoT make available falls into the 11 groupings shown below (the number of measurements in each group are shown in brackets). Not all the groupings are relevant to this project, which is only concerned with delayed departures; for example, the arrival performance at the destination airport is not relevant. The groupings that are considered useful are shown in bold.

- **Time Period** (6 catagorical)
- **Airline** (5 catagorial)
- **Origin** (9 catagorical)
- Destination (9 catagorical)
- **Departure Performance** (3 catagorical, 6 continuous)

- Arrival Performance (3 catagorical, 6 continuous)
- **Cancellations and Diversions** (3 catagorical)
- **Flight Summaries** (3 catagorical, 3 continuous)
- **Cause of Delay** (5 catagorical)
- **Gate Returns** (3 continuous)
- Diverted Airport Information (4 x sub-groups reprenting each diversion with 4 catagorical and 3 continuous)

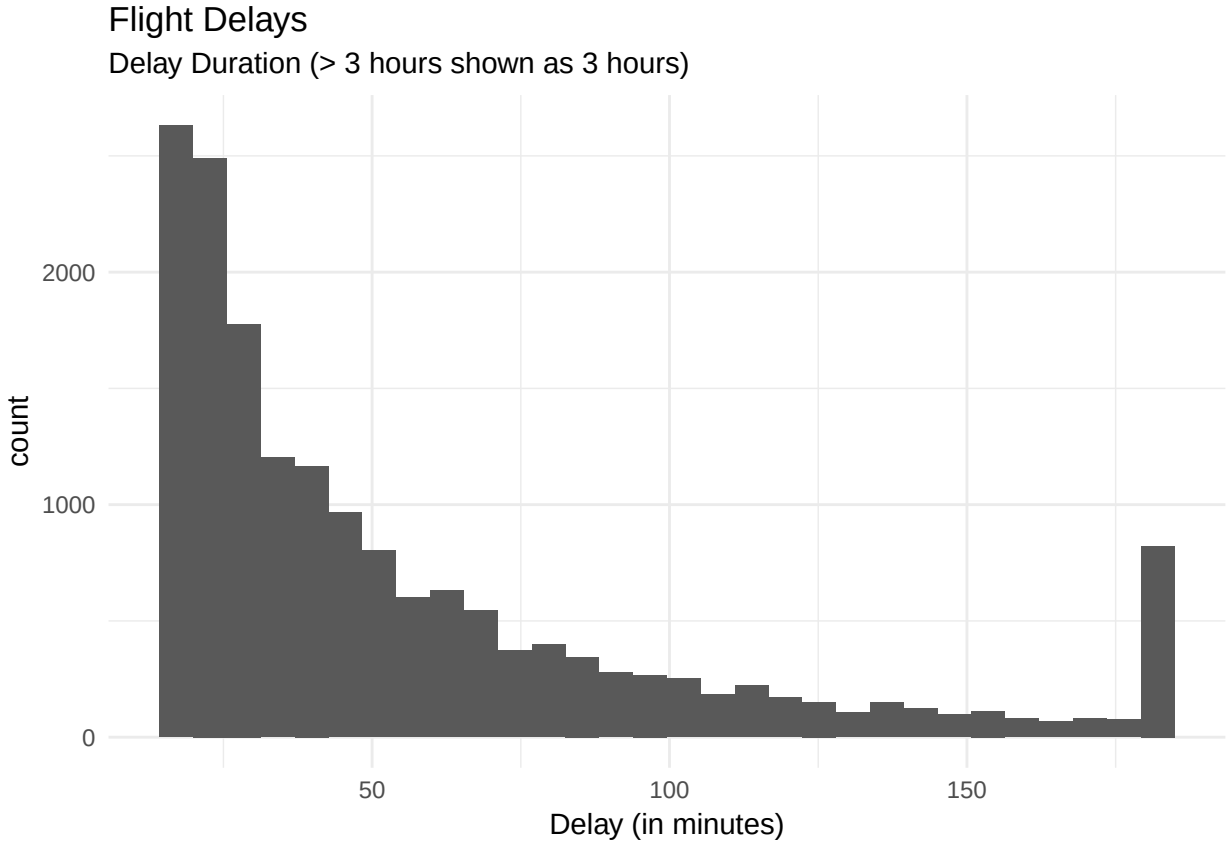
Overview of the Data

There are 100,000 deomestic departures from US airports in the period covered by the data. The break down of these flights by departure status is show below:

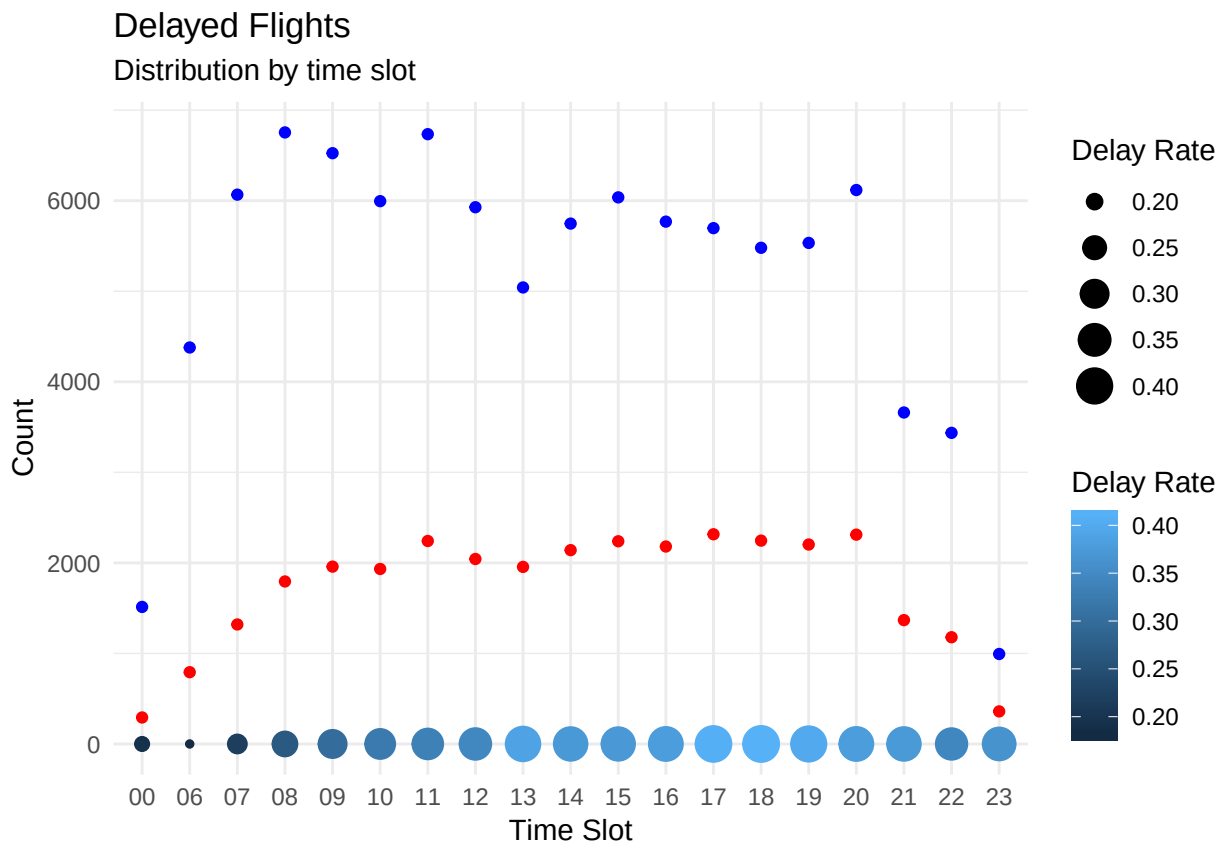


- 2.55% of flights were cancelled
- 59.6% of flights departed ahead of their scheduled time
- 20.6% of flights departed at their scheduled time
- 17.2% of flights departed after their scheduled time

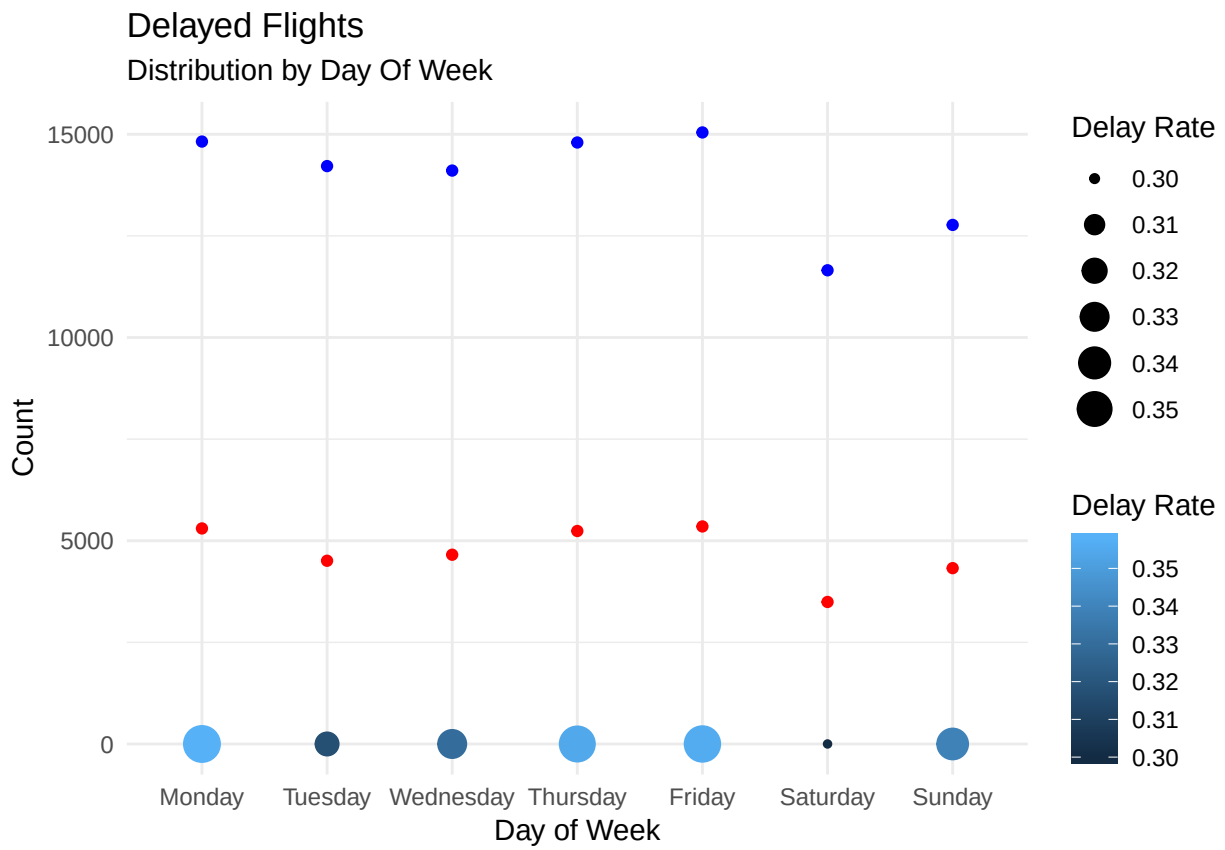
Note that the 0.106% percentage of flights (equal to 106 flights) were not cancelled but have no depature time recorded. This is assumed to be errant data and is excluded from further analysis.



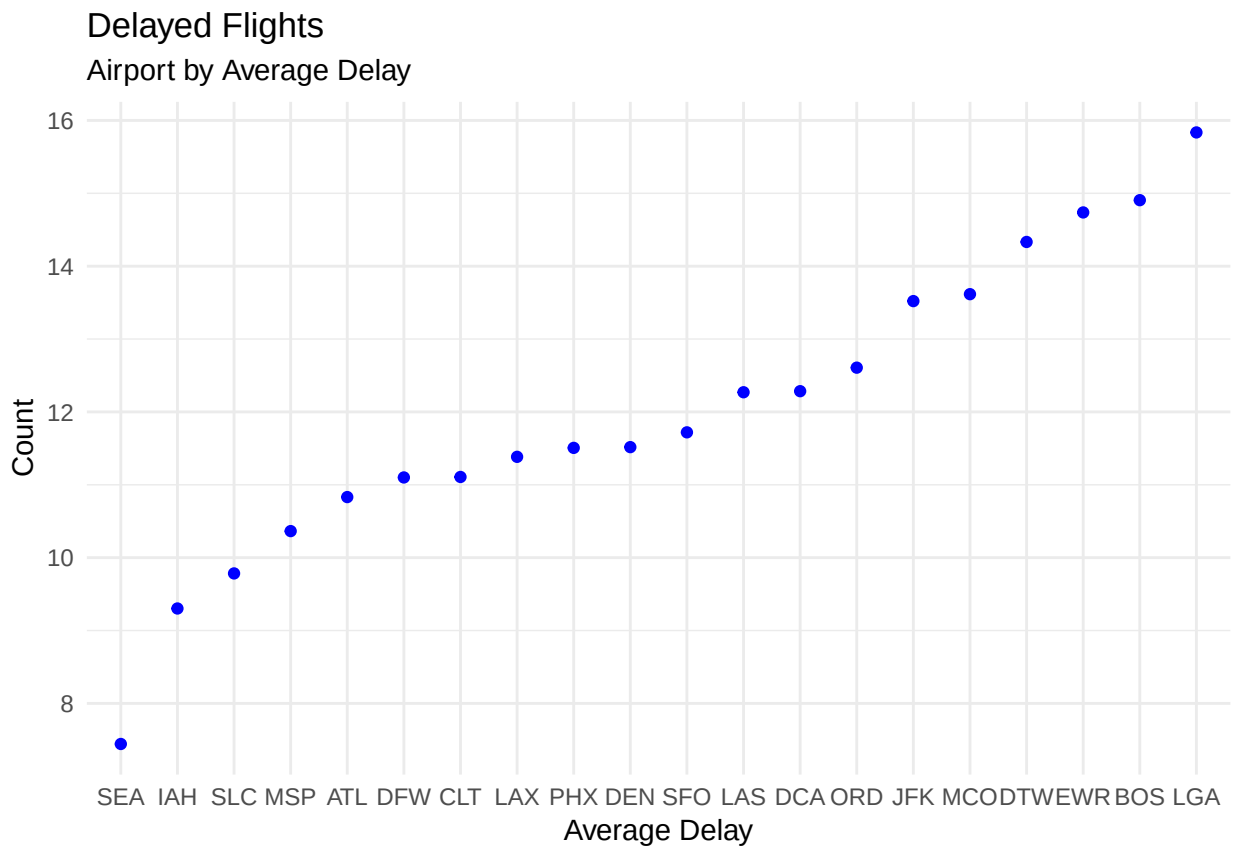
There are a small number of outliers which have exceptionally long delay times. These have been included in the 180 minutes bin to make the histogram more readable. We can see that most delays are within 30 minutes.



This shows that the delay rate changes significantly depending on the time slot peels from early afternoon to early afternoon.



This shows that although the number of departures peeks mid-week and dips at the weekend the delay rate does not vary significantly by day.



Some airports have considerably longer average delay times. The full list of airport codes can be seen here https://www.transtats.bts.gov/FieldInfo.asp?Field_Desc=Origin%20Airport&Field_Type=Char&Lookup_Table=L_AIRPORT&Table_ID=236&SYS_Table_Name=T_ONTIME_REPORTING&Sys_Field_Name=ORIGIN

Training and Test Data Partitions

Goal

The goal of this project is to create a prediction model that can estimates the probably deviation from the scheduled departure time for any domestic flight from any US airport covered by the DoT data.

Approach

Data Cleansing

Cancelled Flights

The aircraft will either leave early, leave on time or be delayed. The expected departure time is given by *CRS_DEP_TIME* and the number of minutes relative to this time of the actual departure is given by *DEP_DELAY*. The number of flights cancelled is 2,549. As these planes don't have a departure time the *DEP_DELAY* for the flight will not be available and will therefore be given a value of N/A in the data.

Modelling Techniques

The data set is quite large and has many features. This makes it computationally expensive. To aleviate this Principal Component Analysis (PCA) will be used to reduce the number of features.

CART (*classification and regression trees*) and Random Forest have been adopted to build a predictive model.

After training the models:

- the best CP (*complexity parameter*) for CART was found to be 0.2
- the number of nodes used for Random Forests was 200

Additionally, random forest is used as a comparison.

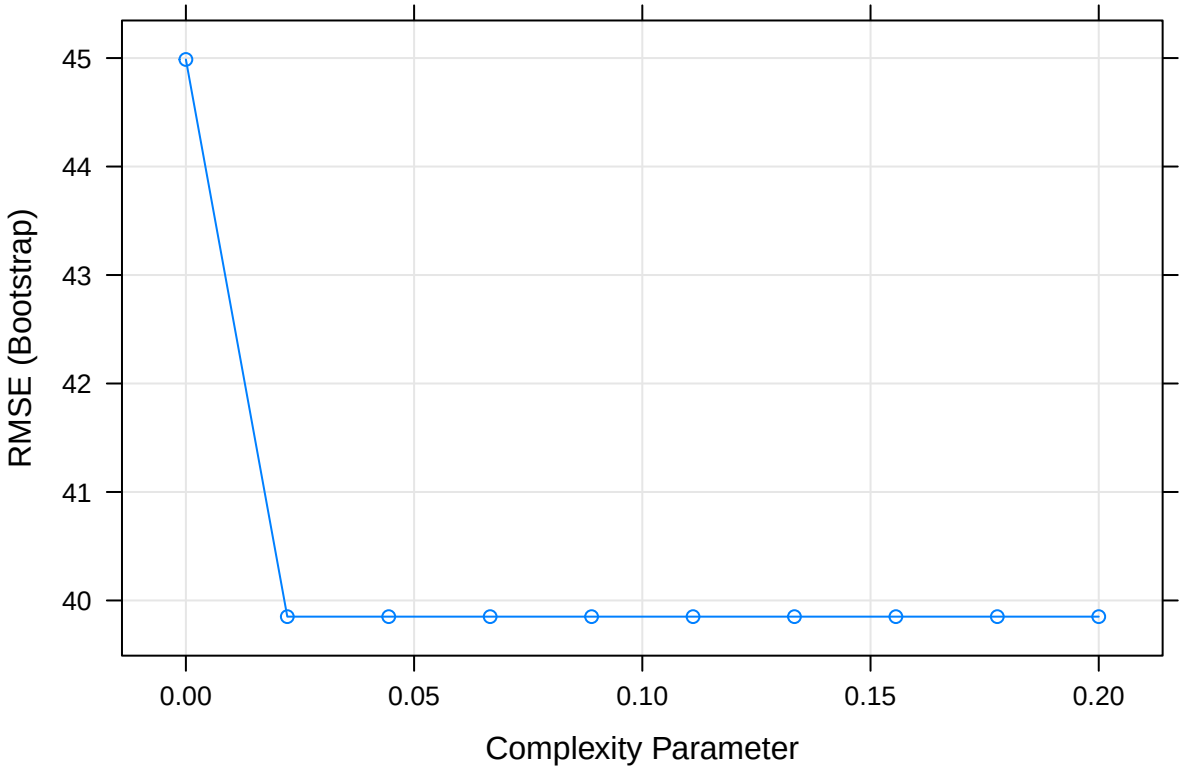
Table 1: Feature Importance

	%IncMSE	IncNodePurity
x.DEP_TIME_BLK	19.349625	3904132
x.DAY_OF_WEEK	8.988356	1979285
x.OP_UNIQUE_CARRIER	14.832756	3139002
x.ORIGIN	11.786985	3435697

Results

CART

The RMSE using CART was 40.024714.




```
## CART
##
## 77928 samples
##     4 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 77928, 77928, 77928, 77928, 77928, 77928, ...
## Resampling results across tuning parameters:
##
##   cp          RMSE      Rsquared     MAE
##   0.00000000  44.98696  0.002436071  20.07290
##   0.02222222  39.85086          NaN  19.00364
##   0.04444444  39.85086          NaN  19.00364
##   0.06666667  39.85086          NaN  19.00364
##   0.08888889  39.85086          NaN  19.00364
##   0.11111111  39.85086          NaN  19.00364
##   0.13333333  39.85086          NaN  19.00364
##   0.15555556  39.85086          NaN  19.00364
##   0.17777778  39.85086          NaN  19.00364
```

```
##   0.20000000  39.85086          NaN  19.00364
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.2.
```

Random Forests

The RMSE using Random Forests was 39.7692484.

```
##
## Call:
##  randomForest(formula = y ~ ., data = train_df, ntree = 200, keep.forest = TRUE,      importance = TR
##                Type of random forest: regression
##                      Number of trees: 200
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 1541.215
##                     % Var explained: 1.76
```

Conclusion

There is some predictive capability but the residual error is quite large which makes the accuracy not very high. Using a larger data set may help this.

Appendix 1: Top Airports

Only data from the following airports has been analyzed. These are the 20 busiest airports.

Table 2: Top Airports

ORIGIN	n
ATL	92454
ORD	75763
DFW	66085
CLT	54864
DEN	53590
LAX	52460
PHX	43769
IAH	41770
SFO	41565
LGA	41036
LAS	38761
DTW	37557
MSP	36527
MCO	35300
EWR	34504
BOS	33730
DCA	32113
SEA	31434
JFK	30111
SLC	27605

Bibliography and References