Big Data Analytics: Final Project Presentation

Presented by: Niki, Katia and Andrew

June 9th, 2018

The Problem

- Flight delays and cancellations can be expensive for businesses and ruinous for vacation travellers.
- Stakeholders are hungry for better departure and arrival information for US flights.
- The goal of our research, analysis and modelling is to provide travellers and other stakeholders with a tool that can be used to better predict on-time performance and likelihood of cancellation.

Our Data

- The dataset includes nearly six million lines of information from 2015 flights, and is sourced from the US Department of Transportation [1].
- Info on departures, arrivals, departure and arrival delays, as well as the frequency of flight cancellations and categorizes the reasons why.
- Data was very imbalanced and had to be compiled from multiple files and sources.
- Contained many “NA” entries that had to be removed and reconciled

[1]https://www.kaggle.com/usdot/flight-delays

Methods Used

- We ran several predictive models for several outcomes with the same set of predictors: Scheduled Departure, Scheduled Arrival, Airline, Month and Day of week of travel.

- The outcomes that required classification were: cancellations, any delay and delay > 30 minutes (given that a delay is likely). We also used regression models to predict the extent of any delay.

Methods Used

- The classification algorithms carried out in R were Random Forests and Support Vector Machines (cancellations only).
- In python, the models were Random Forests, Naïve Bayes, Decision Trees and K-Neighbors.
- In R, the model accuracy was assessed using 20% sample of the data and was based on confusion matrix statistics and the ROC/AUC.
- In python we used the confusion matrix statistics, recall and precision and also implemented 5- or 10-fold cross validation.

Technical Challenge

```
dff <- df[ sample( 1:nrow(df), size=100000 ) , ]

select <- sample( 1:nrow(dff), size=nrow(dff)*0.8 )
training <- dff[ select, ]
test <- dff[ -select, ]
fit <- randomForest( CANCELLED ~ MONTH + DAY_OF_WEEK + AIRLINE
predicted <- predict( fit, test )
confusionMatrix( predicted, test$CANCELLED )

## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 19734   266
##          1     0     0
##
##                Accuracy : 0.9867
##                  95% CI : (0.985, 0.9882)
##     No Information Rate : 0.9867
##     P-Value [Acc > NIR] : 0.5163
##
##                   Kappa : 0
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9867
##          Neg Pred Value :    NaN
##              Prevalence : 0.9867
##          Detection Rate : 0.9867
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##

predicted <- predict( fit, test, type = "prob" )
plot( roc(test$CANCELLED,  predicted[,2] ) )
```

We chose 100,000 lines of data to run our model, as there was too much to process.

Our initial model results came back with a +99% accuracy.

Technical Challenge

```
dff <- df[ sample( 1:nrow(df), size=100000 ) , ]

select <- sample( 1:nrow(dff), size=nrow(dff)*0.8 )
training <- dff[ select, ]
test <- dff[ -select, ]
fit <- randomForest( CANCELLED ~ MONTH + DAY_OF_WEEK + AIRLINE
predicted <- predict( fit, test )
confusionMatrix( predicted, test$CANCELLED )

## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 19734   266
##          1     0     0
##
##                Accuracy : 0.9867
##                  95% CI : (0.985, 0.9882)
##     No Information Rate : 0.9867
##     P-Value [Acc > NIR] : 0.5163
##
##                   Kappa : 0
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9867
##          Neg Pred Value :    NaN
##              Prevalence : 0.9867
##          Detection Rate : 0.9867
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##

predicted <- predict( fit, test, type = "prob" )
plot( roc(test$CANCELLED,  predicted[,2] ) )
```

We chose 100,000 lines of data to run our model, as there was too much to process.

Our initial model results came back with a +99% accuracy.

Great!!!

Technical Challenge

```
dff <- df[ sample( 1:nrow(df), size=100000 ) , ]

select <- sample( 1:nrow(dff), size=nrow(dff)*0.8 )
training <- dff[ select, ]
test <- dff[ -select, ]
fit <- randomForest( CANCELLED ~ MONTH + DAY_OF_WEEK + AIRLINE
predicted <- predict( fit, test )
confusionMatrix( predicted, test$CANCELLED )

## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 19734   266
##          1     0     0
##
##                Accuracy : 0.9867
##                  95% CI : (0.985, 0.9882)
##     No Information Rate : 0.9867
##     P-Value [Acc > NIR] : 0.5163
##
##                   Kappa : 0
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9867
##          Neg Pred Value :    NaN
##              Prevalence : 0.9867
##          Detection Rate : 0.9867
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##

predicted <- predict( fit, test, type = "prob" )
plot( roc(test$CANCELLED,  predicted[,2] ) )
```

We chose 100,000 lines of data to run our model, as there was too much to process.

Our initial model results came back with a +99% accuracy.

~~Great!!!~~

Only ~1.5% of all flights were cancelled. A classic example of an imbalanced dataset...

Technical Challenge

```
dff <- df[ sample( 1:nrow(df), size=100000 ) , ]

select <- sample( 1:nrow(dff), size=nrow(dff)*0.8 )
training <- dff[ select, ]
test <- dff[ -select, ]
fit <- randomForest( CANCELLED ~ MONTH + DAY_OF_WEEK + AIRLINE
predicted <- predict( fit, test )
confusionMatrix( predicted, test$CANCELLED )

## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 19734   266
##          1     0     0
##
##                Accuracy : 0.9867
##                  95% CI : (0.985, 0.9882)
##     No Information Rate : 0.9867
##     P-Value [Acc > NIR] : 0.5163
##
##                   Kappa : 0
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9867
##          Neg Pred Value :    NaN
##              Prevalence : 0.9867
##          Detection Rate : 0.9867
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##

predicted <- predict( fit, test, type = "prob" )
plot( roc(test$CANCELLED,  predicted[,2] ) )
```

To overcome this issues we:

- Created a new dataset
- Selected all cancelled flights (~90,000)
- Added the same number of randomly selected non-cancelled flights
- Re-ran our model...

The Results

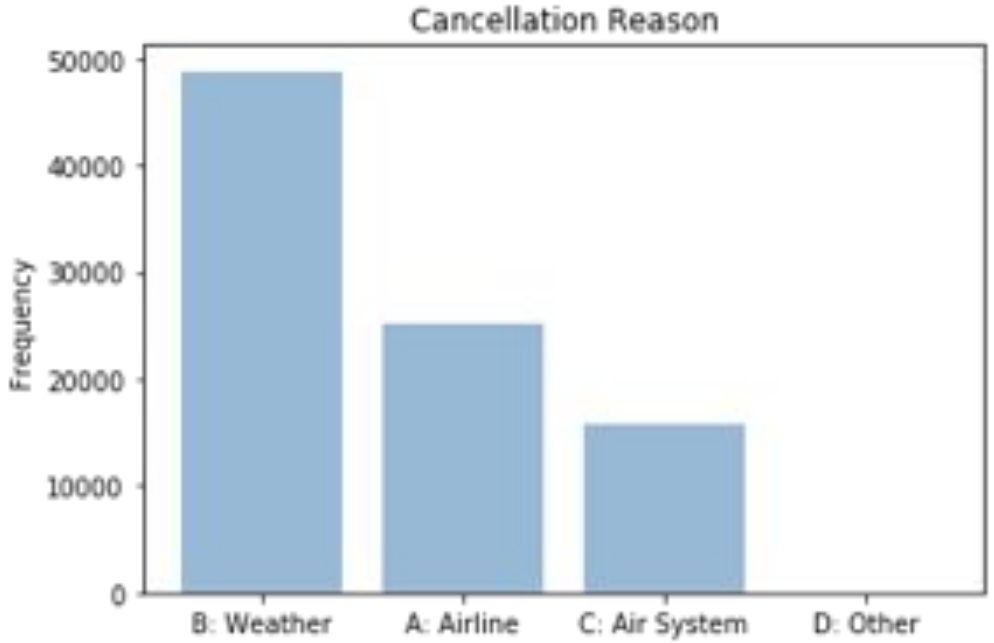
Our approach was to analyze data from millions of flights to create predictive models that use past performance to indicate the likelihood of cancellation and future on-time airline performance.

- Cancellations: The Random Forest classifier outperformed the others with 74% accuracy(R) and precision of 0.72, recall of 0.73 (python). Note that the SVM had accuracy of 0.66 (R).

- Any delay: Again, the Random Forests outperformed the others with precision and recall values of 0.68 and 0.84, respectively.

- Delays over 30 minutes The predictive accuracy was poor with precision and recall of the best model (Naïve Bayes) equal to 0.37 and 0.39 respectively.

The Results


Cancellation Reason
Frequency
50000
40000
30000
20000
10000
0
B: Weather
A: Airline
C: Air System
D: Other

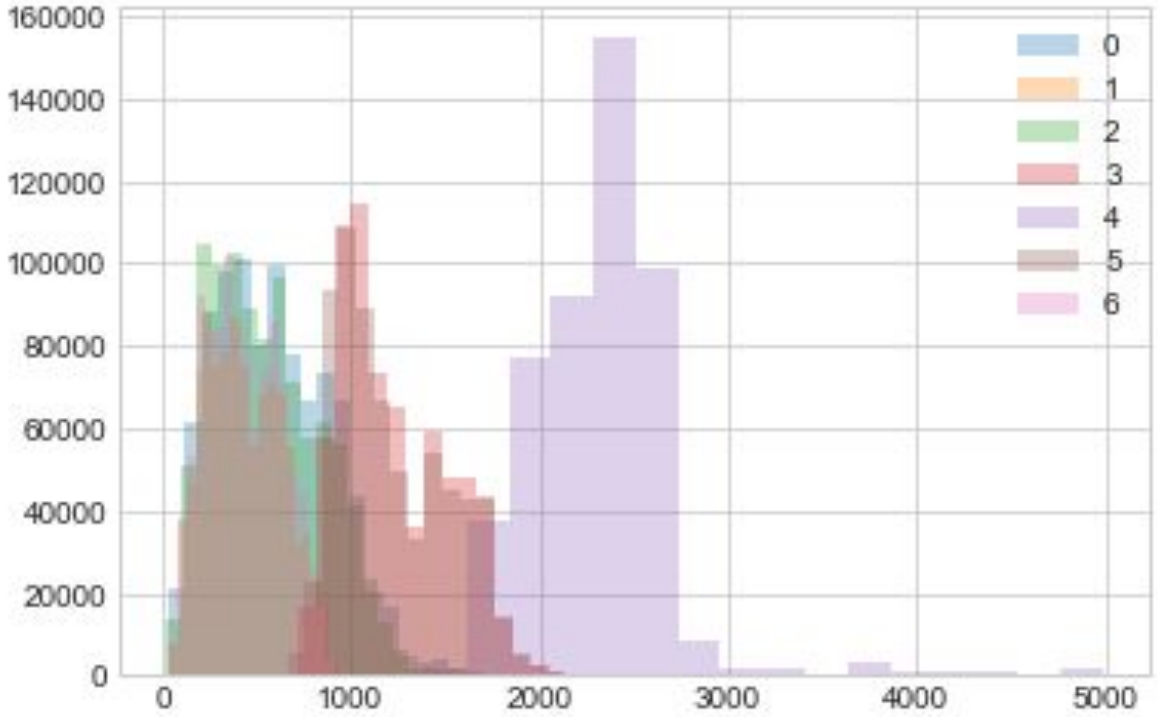
The Results

Kmeans cluster analysis showed 7 distinct clusters, based on "month", "day_of_week", "distance" and "scheduled time".

- Most short flights are clustered at the beginning of the week (Mon to Thurs, clusters 0, 1, 2, and 6) and longer flights closer to weekend (Fri, Sat and Sun, clusters 3 to 5).

- Most of the long flights (cluster 4) are taken in June and mostly July.

The Results


160000
140000
120000
100000
80000
60000
40000
20000
0
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6

The Results

1. **With air traffic congestion and weather events set, are we able to adequately evaluate individual airline performance?**

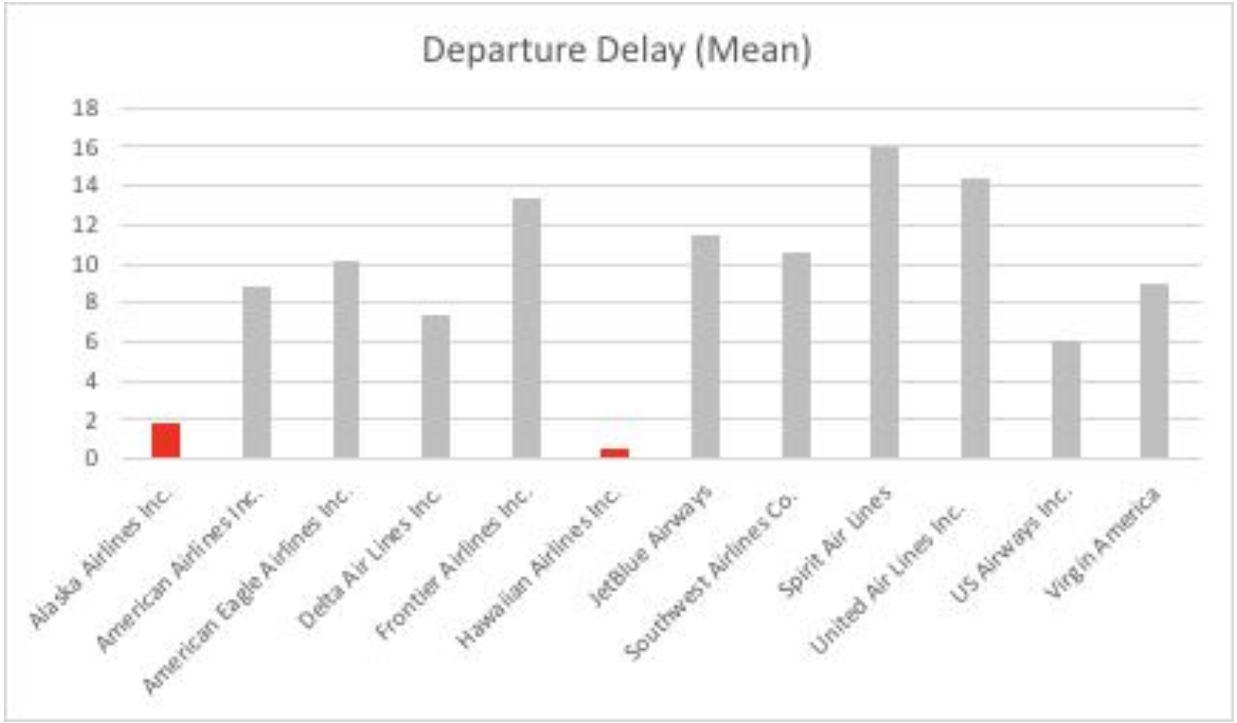
   Yes. Comparing on-time arrivals, departures and the amount of cancelled flights, we're able to evaluate the differences between carriers.

The Results

2. **Is there a significant difference between national carriers?**

   Yes. Our analysis showed a wide gap between performances, with Alaska Airlines having the most on-time arrivals, second best on-time departures and second least amount of cancelled flights in the group.

The Results


Departure Delay (Mean)
18
16
14
12
10
8
6
4
2
0
Alaska Airlines Inc.
American Airlines Inc.
American Eagle Airlines Inc.
Delta Air Lines Inc.
Frontier Airlines Inc.
Hawaiian Airlines Inc.
JetBlue Airways
Southwest Airlines Co.
Spirit Air Lines
United Air Lines Inc.
US Airways Inc.
Virgin America

The Results

3. **What percentage of delays/cancellations are due to airlines versus weather or air traffic congestion? Is it significant?**

   Biggest contributing factors for delays are "late_aircraft delay" (40% of the total delay time) followed by "airline_delay" (32%), "air system_delay" (23%) and "weather_delay (5%).

   There is no significant delay related to security.

   For cancellations, weather is the biggest driver, followed by Airline issues.

The Results

4. **Do flight speeds impact on-time arrivals?**

   Most airlines fly at similar speeds (400-450 miles/h), and it seems that there is not any strong correlation between speed and on-time performance.

The Results

5. **Can we predict which airline will have the greatest chance of an on-time departure and arrival?**

   Based on past performance, we feel confident that future performance can be accurately predicted using our model.

Conslusion

Given our analysis, we feel confident that we'll be able to use the historical flight data to estimate the chance of cancellation and any delay based on the user input of airline, scheduled departure, arrival, flight time, day of week and month of travel. We have less confidence in the predicted chance of long delays (>30mins) and delay time.