Estimated Off-Block Time Prediction

Anubhav Reddy Nallabasannagari

BUAN 6V98.091

Spring 2017

May 8th 2017

Preface

This report contains my findings and learnings from internship with American Airlines from 27th March 2017 to 8th May 2017. I worked as intern with Centre of Excellence (COE) for Machine Learning. The objective of COE is to develop machine learning capabilities at American Airlines and provide internal consulting and guidance to business units in American Airlines. Mr. Venkata Pilla is the Manager at COE and I reported to him during the duration of my internship. He was very supportive and helped me perform to best of my abilities. He can be reached at (817)-931-1712 to understand more about COE and development of machine learning at American Airlines.

The report is broadly divided into:-

- Introduction
- Objective
- Analysis
- Findings
- Learnings

Table of Contents

Introduction

American Airlines is the world's largest airline when measured by fleet size, revenue, passengers flown and destinations served. Its headquarters is located in Fort Worth Texas. Its serves more than 350 destinations in more than 50 countries. American airlines founded Oneworld alliance. The alliance comprises of British Airways, Iberia, Finnair, Cathay Pacific, Nippon Airlines and American Airlines.

American Airlines was started in 1930 with union of more than 80 small airlines. Due to tough market situation and downturn in the airline industry the American Airline's parent company AMR Corporation filed for bankruptcy protection in 2011. In 2013 the company merged with US Airways.

American Airlines Group has a total revenue of $ 40.180 billion, operating income of $ 5.284 billion and a profit of $ 2.676 billion. It currently employs over 118,500 employees.

Problem Statement

Flights do not always depart at the scheduled time to departure. Sometimes flights get delayed and in other situations they are ready to depart early. During delays a flight may miss its take of sequence and may be moved many places down the departure list hence, leading to more delay. When a flight is ready to take of earlier than scheduled time the airlines can benefit from this situation if the early departure is communicated in advance. This will help airline make more effective use of their equipment.

Objective

The objective is to predict time to departure at regular intervals prior to departure. This in turn will help us determine Estimated Off-Block Time (EOBT).

The example below will help to understand the objective.

Flight NBR	Passengers Boarded	SCHD Departure time	Current Time	Time to Departure	Estimated Off-Block Time	Actual Departure time
850	5%	11:21 AM	10:53 AM	25	11:18 AM	11:18 AM
850	25%	11:21 AM	10:57 AM	22	11:19 AM	11:18 AM
850	50%	11:21 AM	11:00 AM	19	11:19 AM	11:18 AM
850	75%	11:21 AM	11:04 AM	16	11:20 AM	11:18 AM
850	95%	11:21 AM	11:07 AM	12	11:19 AM	11:18 AM

Time to departure – Target Variable

Estimated Off-Block Time = Current Time + Time to departure

Actual time to departure – recorded when the blocks are remove

The prediction will occur when 5, 25, 50, 75 and 95 percentile of passengers have boarded the flight.

Current Model

The current model determines time to departure from historical data. Based upon a percentile value the time departure is selected from the trend that is built using the historical data. The percentile value itself is determined based on:

- Fleet Type
- Airport Type (Hub, Gateway, Spoke)
- Bags / Passengers left

Two values are derived for time to departure. One is obtained based on bags left and other obtained based on passengers left to board. The maximum of both value is taken as final value for time to departure.

Current model has following challenges:

- Very few variables are used which prevents the model from accurately identifying underrepresented patterns in the data
- Single airport requires multiple curves to be derived which is not a scalable solution if the number of hubs and variables increase

Machine Learning Model

The machine learning models chosen to predict the time to departure are:

- Decision Tree Regression
- Random Forest Regression

These models were chosen because the preprocessing required for train these models is very less compared to other regression models like Linear Regression, Polynomial Regression and Neural Network Regression.

Data Pre – Preprocessing

Data Description

The dataset comprises of flight details for all flights originating from Dallas Forth Worth and Charlotte airport for American Airlines for the year 2016. The dataset contains flight details for approximately 143,000 flights. Another dataset containing passenger boarding details is obtained and combined with the flight details. The final dataset has 143,000 flights and 55 columns of data.

Reshaping Data

The merged dataset has the flight details repeated for every instance of boarding i.e. same flight repeats itself 10 times. Reshape the entire dataset that all details are available on a flight level rather than transaction level. This makes data easier to manage and read.

Feature Creation and Selection

Apart from the existing features in the dataset new features need to be generated in order to explain the changes in time to departure. Many of the features existing the dataset were a derivation of other features. For example a date-time feature was again represented by a separate date and time feature. Such redundant features need to be removed. After feature selection and feature creation following list of features is obtained.

List of Features	
Existing Features	**Derived Features**
Origin airport code	Passenger left to board
Outbound length of haul	Departure Bank
Inbound length of haul	Weekend or Weekday
Outbound number of passenger	Fleet type
Inbound number of passenger	Time to scheduled departure
Is previous departure airport International	Tight turn indicator
Is arrival airport International	Day of week
Is outbound leg international	Week of month
Is inbound leg international	Month of year

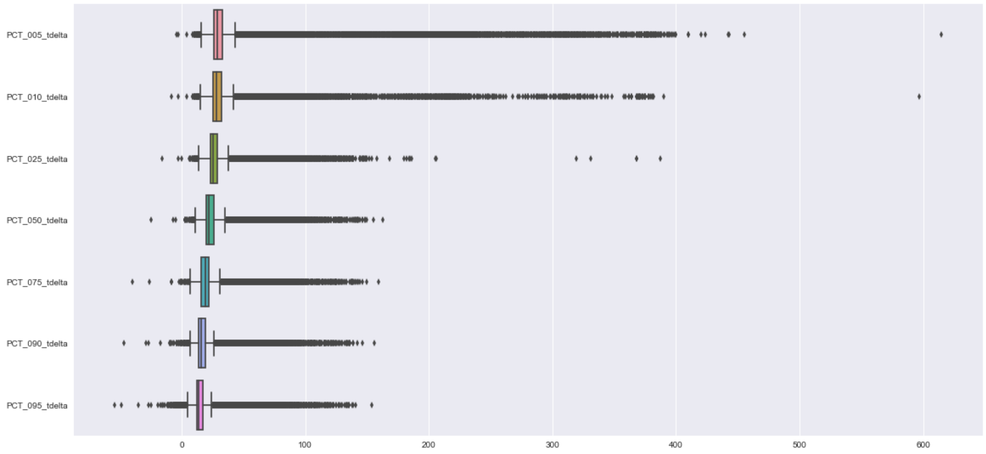
- Passenger left to board = Total number passenger boarding at the airport – Number of passengers boarded
- Departure bank – this information is obtained from another source and combined with the current dataset
- Weekday or Weekend flag – it is based upon scheduled departure date
- Fleet Type – The Fleet code is further grouped into wide-body, narrow body, regional small and regional large
- Time to Scheduled Departure = Time at boarding - Scheduled Departure Time
- Tight Turn Indicator – Minimum On Ground Time > Available On Ground Time
- Day of the week – this based on scheduled departure date
- Week of the month - this based on scheduled departure date
- Month of the Year - this based on scheduled departure date

Target Variable

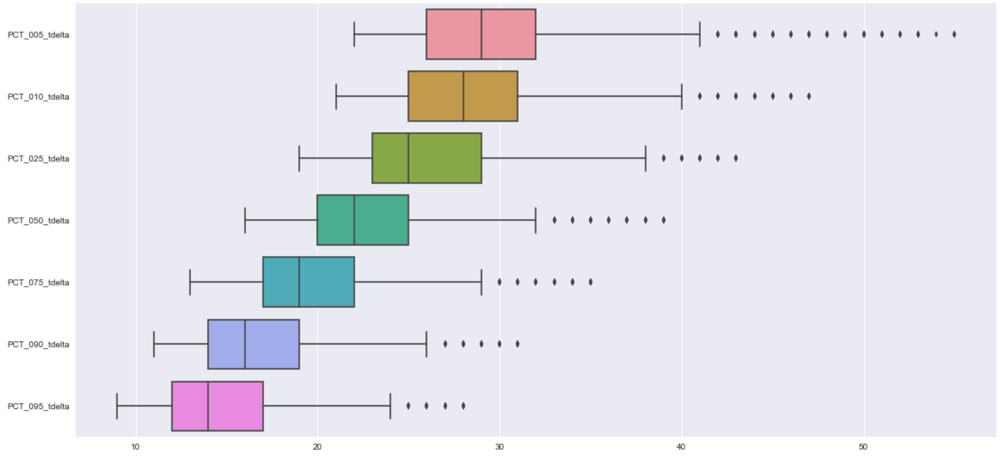
Time to actual departure = Time at boarding – Actual Departure Time

The target variable has very extreme values on negative and positive side. These extreme values need to be removed as they represent significant deviation from normal process or extreme delays which are beyond the scope of the project.

The representation of target variables can be seen below:



We observe that extreme values exist for the all the target variables. We remove the outliers by removing the all data points below 0.05 percentile and 0.95 percentile. Hence we can see the new distribution of the target variables below.



Split Dataset - Train & Test

The dataset needs to split into training data and test data. The training data will be used to help the supervised algorithm learn. The test data is used to estimate the performance of the algorithm. We split the dataset into 70:30 ratio.

- Training Data – 81,431 flights
- Test Data – 34,899 flights

Parameter Selection and Power Tuning

Parameters for Decision Tree Regression (DTR)

- Maximum Depth: 1-50
- Min sample in leaf: 100

Parameters for Random Forest Regression (RFR)

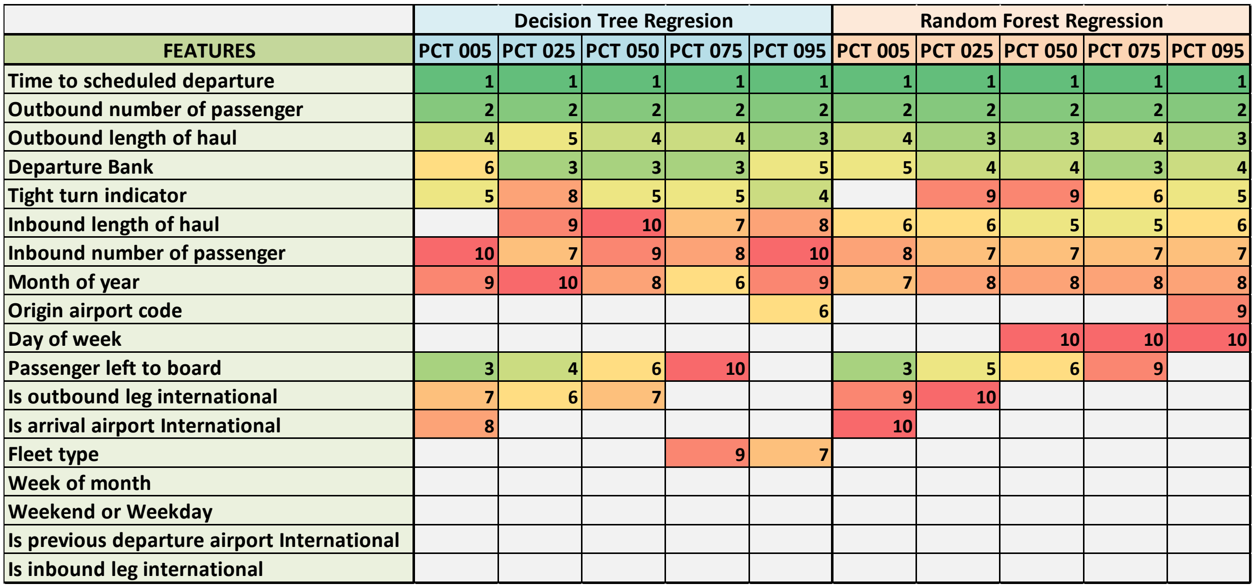
- No. of Trees: 20, 100, 200, 300
- Min No. of samples in leaf: 100

The selection of the best parameters is achieved through a function called “GridSearchcv”. Grid search function searches through each of the given parameters to determine the best parameters. It determines the best parameter by comparing the MSE (mean square value) generated by each parameter selection. The MSE itself is generated through a 3-fold cross validation of training data set hence, eliminating the need for validation dataset and prevents overfitting of the trees.

Analysis

Feature Importance Matrix

Feature importance matrix is generated by extracting the feature importance from all the models in order to understand importance of features at different stages of boarding.



	Decision Tree Regresion					Random Forest Regression				
FEATURES	PCT 005	PCT 025	PCT 050	PCT 075	PCT 095	PCT 005	PCT 025	PCT 050	PCT 075	PCT 095
Time to scheduled departure	1	1	1	1	1	1	1	1	1	1
Outbound number of passenger	2	2	2	2	2	2	2	2	2	2
Outbound length of haul	4	5	4	4	3	4	3	3	4	3
Departure Bank	6	3	3	3	5	5	4	4	3	4
Tight turn indicator	5	8	5	5	4		9	9	6	5
Inbound length of haul		9	10	7	8	6	6	5	5	6
Inbound number of passenger	10	7	9	8	10	8	7	7	7	7
Month of year	9	10	8	6	9	7	8	8	8	8
Origin airport code					6					9
Day of week								10	10	10
Passenger left to board	3	4	6	10		3	5	6	9	
Is outbound leg international	7	6	7			9	10			
Is arrival airport International	8					10				
Fleet type				9	7					
Week of month										
Weekend or Weekday										
Is previous departure airport International										
Is inbound leg international										

We can see that some of the features are not important across all the boarding stages. This helps us understand that the current process is biased as it depends upon only e features.

PCT 005 – Passenger boarding at 5 percentile

PCT 025 – Passenger boarding at 25 percentile

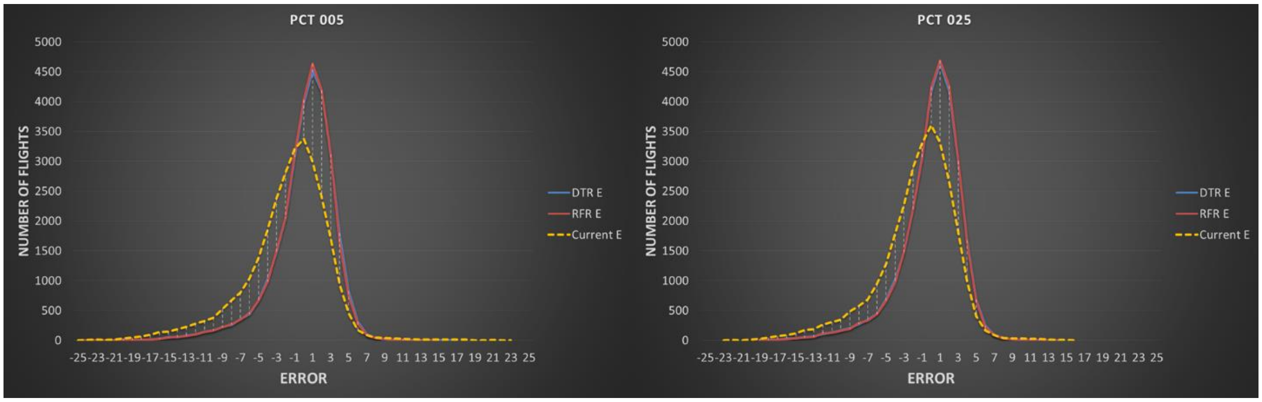
PCT 050 – Passenger boarding at 50 percentile

PCT 075 – Passenger boarding at 75 percentile

PCT 095 – Passenger boarding at 95 percentile

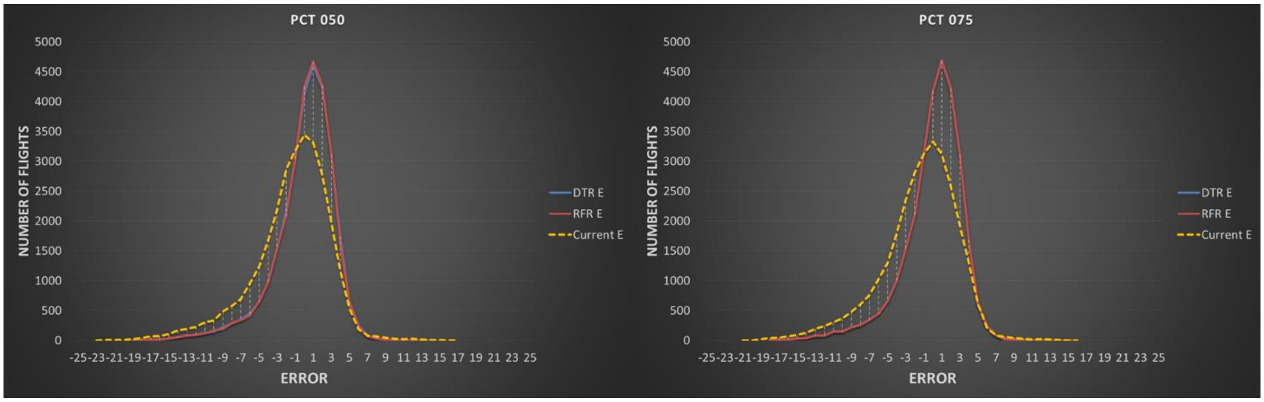
Error Distribution

In order to compare the results we must understand the distribution of errors from our prediction and compare them against the errors of current model.

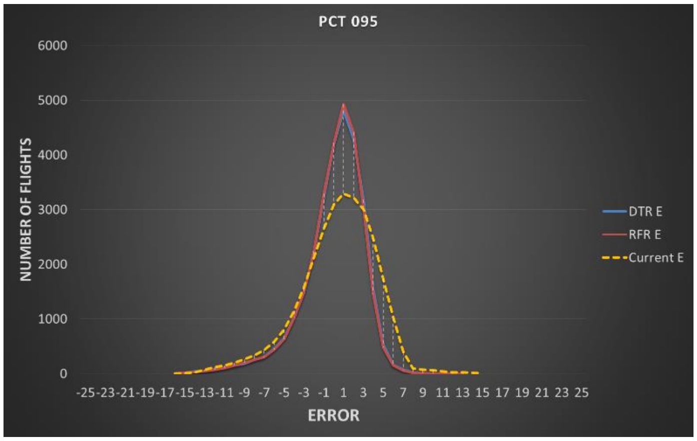


The above charts indicate the error distribution when passenger boarding is at 5 percentile and 25 percentile.

- RFR – Random Forest Regression
- DTR – Decision Tree Regression
- Current model has larger spread of errors
- Random Forest and Decision Tree have identical distribution of errors



The charts represent error at 50 percentile and 75 percentile. The chart below displays the distribution at 95 percentile.



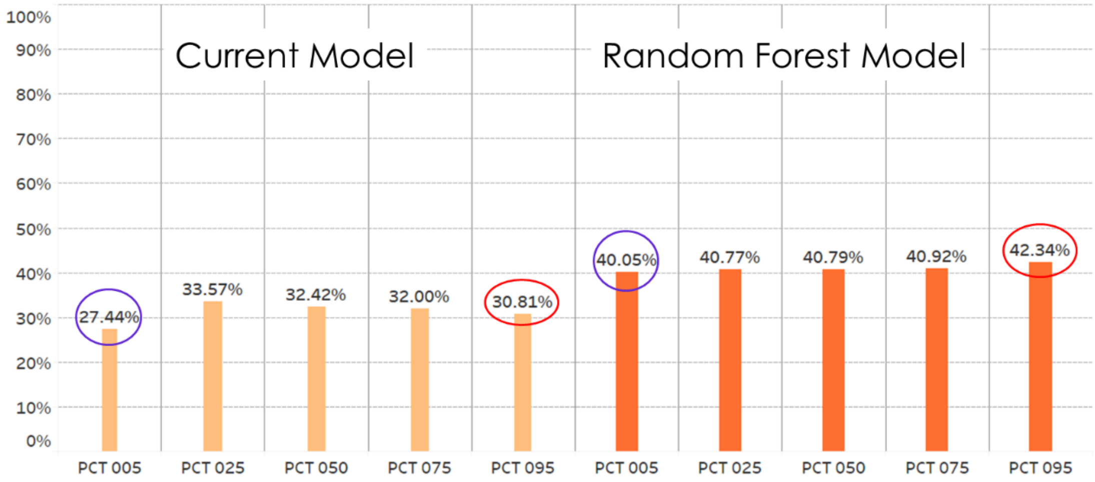
We observe from the above charts that the current model has more negative errors at the start of boarding process and more positive errors at the end of boarding process. This tells us that the current model more aggressive when boarding starts and towards the end make more liberal predictions. At all the stages we observe that the machine learning model outperforms the current model.

We choose Random Forest Model over Decision Tree because it is more robust and less prone to overfitting than the Decision Tree. As we are focused on our prediction accuracy rather than the rules are leading to a certain prediction, Random Forest Model is better choice.

Results

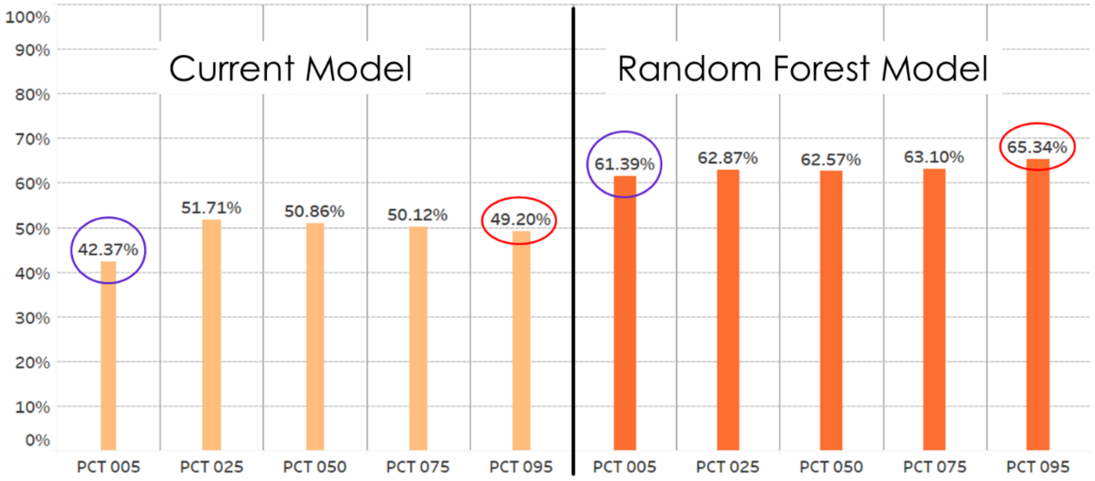
In order to measure performance in business terms we will measure accuracy of the predictions in the interval range of one, two and five minutes.

Accuracy (+/- 1 minute)



- Accuracy from Random Forest is consistent and higher than Current Model
- Random forest is able to predict time to departure within 1 min range for 40-42% of the flights
- Accuracy for current model reduces as we approach end of boarding process

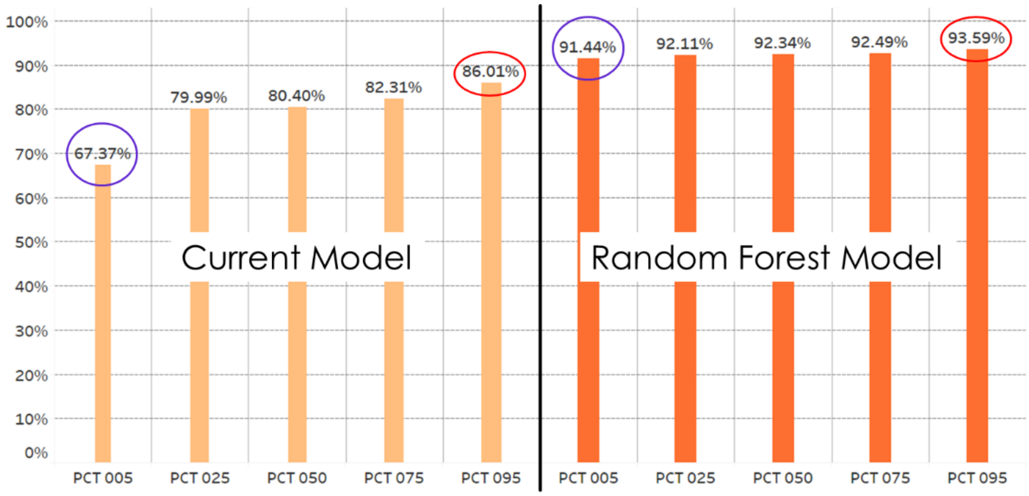
Accuracy (+/- 2 minutes)



- Accuracy from Random Forest is consistent and higher than Current Model

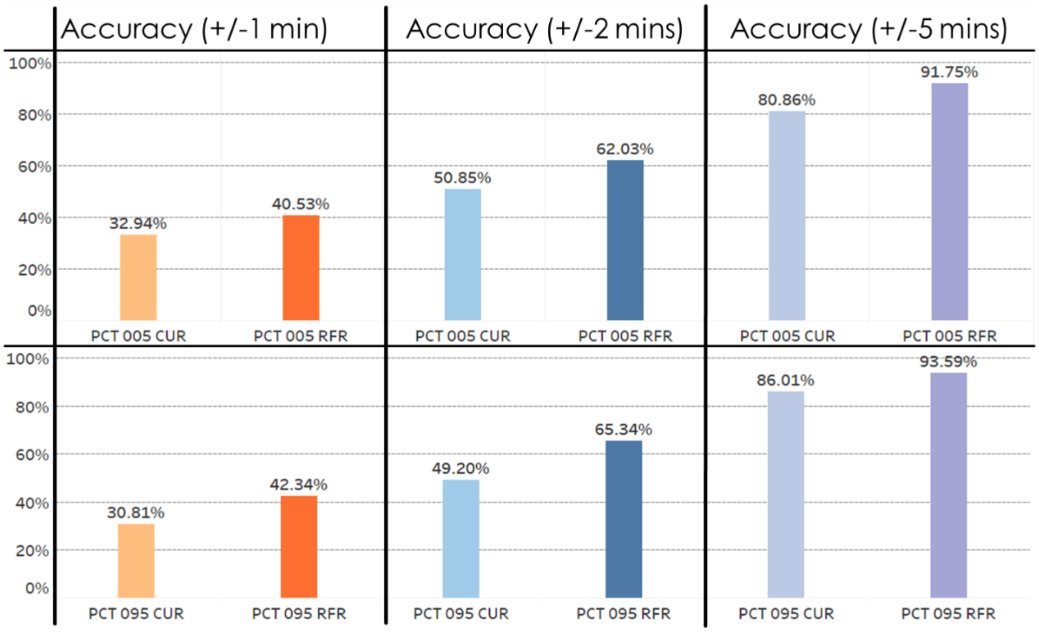
- Random forest is able to predict time to departure within 2 mins range for 60-65% of the flights
- Accuracy for current model reduces as we approach end of boarding process

Accuracy (+/- 5 minutes)



- Accuracy from Random Forest is consistent and higher than Current Model
- Random forest is able to predict time to departure within 5 mins range for 91-93% of the flights

Accuracy Summary



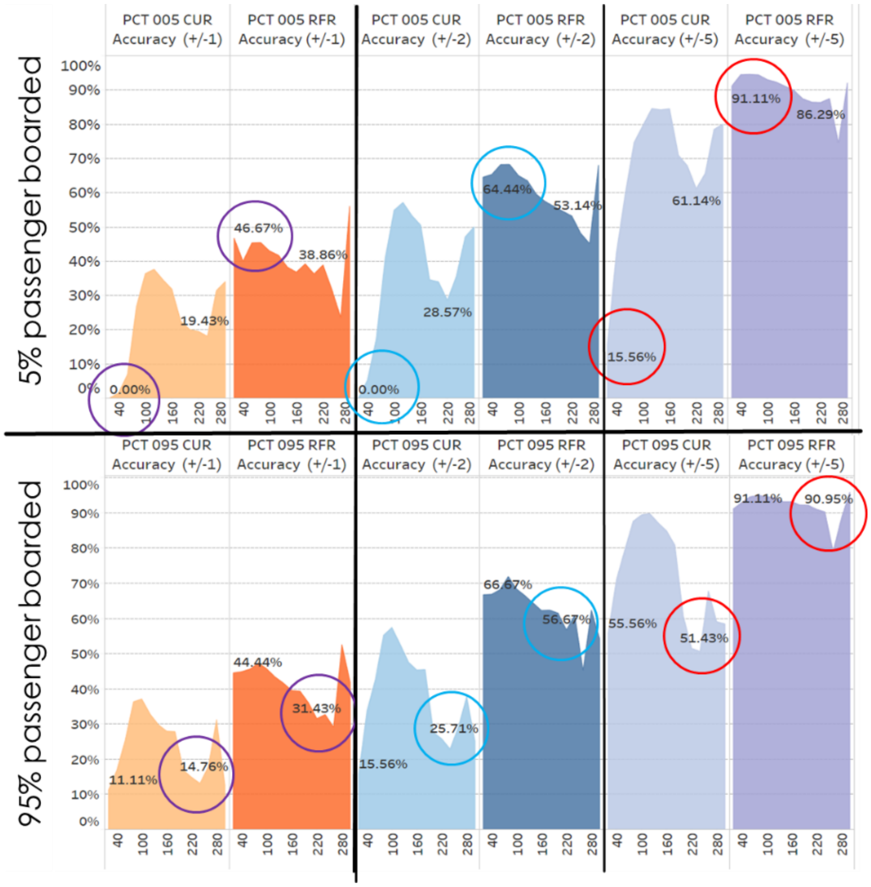
- Random Forest Model has higher accuracy than current model in all accuracy intervals

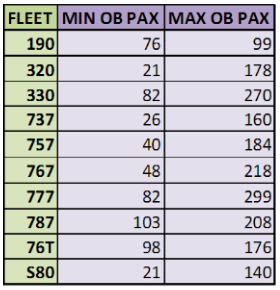
Performance Comparison

We would like to go a step beyond and compare the performance in smaller pockets of data to verify if the machine learning model is able to pick up the pattern represented by a very small subset of data. We will some the important features to identify performance.

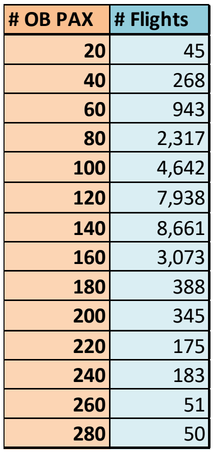
Outbound number of passengers

This features tells us about total number of passengers leaving on a flight.





FLEET	MIN OB PAX	MAX OB PAX
190	76	99
320	21	178
330	82	270
737	26	160
757	40	184
767	48	218
777	82	299
787	103	208
76T	98	176
S80	21	140

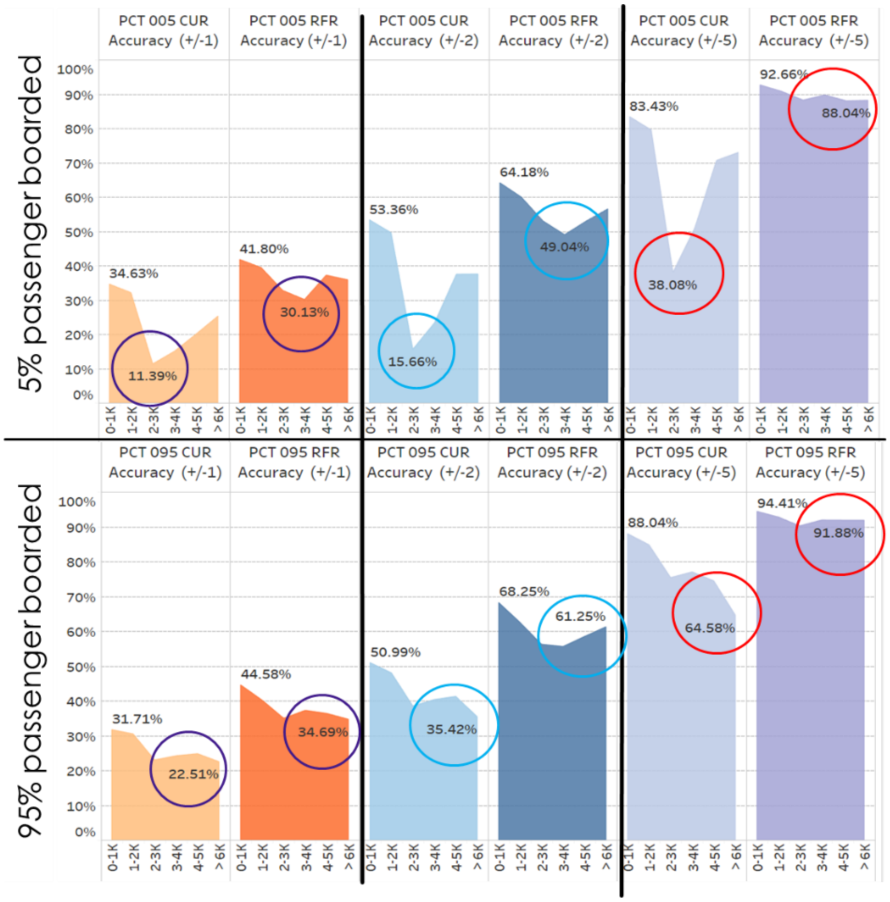


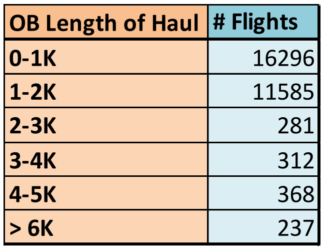
# OB PAX	# Flights
20	45
40	268
60	943
80	2,317
100	4,642
120	7,938
140	8,661
160	3,073
180	388
200	345
220	175
240	183
260	51
280	50

We observe from the highlighted regions that the random forest model outperforms current model significantly. When outbound passengers are less than 60 or between 180 and 260 the performance of random forest extremely good compared to current model.

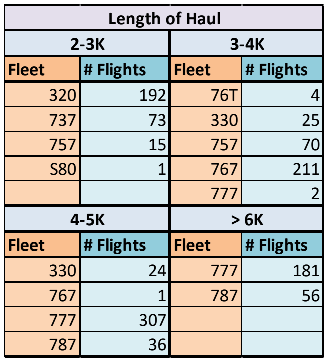
Outbound Length of Haul

This feature tells the length of haul to the destination of the flight.





OB Length of Haul	# Flights
0-1K	16296
1-2K	11585
2-3K	281
3-4K	312
4-5K	368
> 6K	237

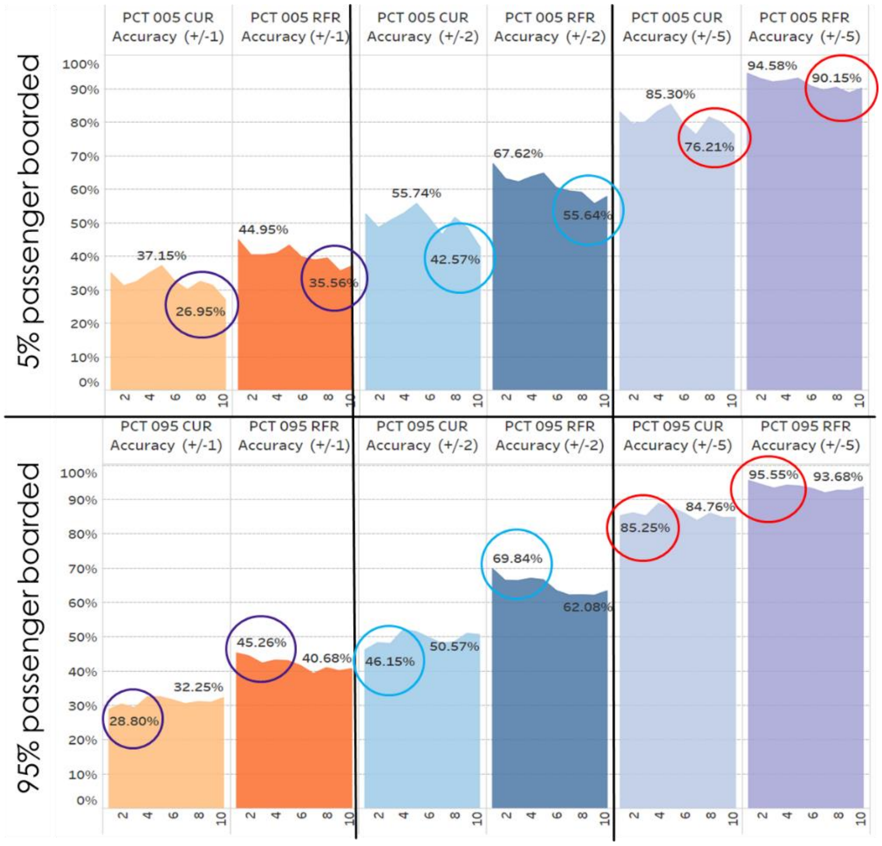


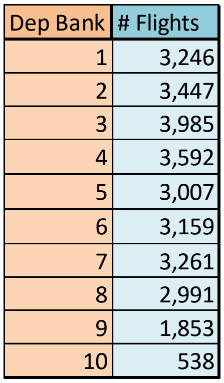
Length of Haul			
2-3K		3-4K	
Fleet	# Flights	Fleet	# Flights
320	192	76T	4
737	73	330	25
757	15	757	70
S80	1	767	211
		777	2
4-5K		> 6K	
Fleet	# Flights	Fleet	# Flights
330	24	777	181
767	1	787	56
777	307		
787	36		

We can observe from the chart above that the current model performs very poorly for the flight with length of haul greater between 2000 – 4000 miles. The tables on the right show us the distribution of flights and type of flights.

Departure Bank

The flights are grouped together into banks based on their departure time. Bank 1 represents the flights that leave early in the morning and Bank 10 represents flights leaving late at night.





Dep Bank	# Flights
1	3,246
2	3,447
3	3,985
4	3,592
5	3,007
6	3,159
7	3,261
8	2,991
9	1,853
10	538

We observe that as we reach towards the end of day the accuracy for both models falls. The Random forest model still outperforms the current model by significant margin.

Conclusion

The key findings of this the project are:

- Random Forest provides higher accuracy than current model
- Random Forest is able to identify patterns in small pockets of data
- Random Forest is accurate for over 90% of fights in +/- 5 mins range
- Random Forest is accurate for over 60% of fights in +/- 2 mins range
- Random Forest is easily scalable to be deployed on complete network

Key Learnings

The learning from this project are:

- Better understanding of terminology used in Airline Industry
- Better understanding of Decision tree and Random Forest Models
- Better understanding of how trees are used for regression and why it is easier to implement than a regular regression in terms of pre-processing of data
- Better understanding of Pandas and Scikit-learn i.e. machine learning package for python
- Date and time manipulation in python; and developing appropriate structure for the given problem to get the desired output in terms of extracting the results and reshaping the data
- Importance of understanding the audience and creating appropriate presentation to effectively communicate the findings
- Learned to develop a structure to approach data analysis and effectively manage time to meet deadlines