Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

Abstract: Our idea is to perform exploratory analysis using big data technologies to analyze aviation data that consists of airlines, flights and airports and combining it with weather data to understand:

- ***What causes delays in flights?***
- ***Build a model that can predict the delay based on such factors.***
- ***Test the model for accuracy.***
- ***Perform text-analysis on user reviews to get a sense of common-topics.***
- ***Combine all of the above to develop a sound recommender system.***

Based on the results of this analysis, we aim to develop a prediction logic that can predict how much delay can be expected on a given path and which carrier would be a better choice when it comes to choosing a flight on the said path.

Introduction

This project will serve as a guideline for frequent fliers to choose which flight they should avoid which has a record of frequent delays or which carrier has a history of huge number of delays. It will also provide predictive analysis based on weather and many other factors like the day of your travel from a source to a destination using a particular airline might or might not cause delay. It also analyzes reviews by users for different airlines and airports which will help owners understand what drives the public opinion and how can they improve their facilities to better serve the passengers. The airline delay analysis module combined with user reviews text analysis module can form a strong recommender system.

Motivation

Providing the travelers the insight on how much delay they can expect during their travel could benefit a traveler to plan their travel diligently. It would also be of great help to airlines to better plan their routes and schedules and provide better service to their customers. This project will provide useful insight to both the flyer as well as the airlines so as to manage their operations efficiently. Making use of the most recent data will further improve the accuracy of results and predictions will be more realistic. Together the system can be scaled to analyze all the airports and all the major airlines.

Data Sources

We used three datasets for our models:

- Flights Data – Origin, Destination, DepDelay, Time, Carrier, Cancellation, etc.
- Weather Data - Time, wind, precipitation, visibility, location, etc.
- User Reviews Data - Carrier, time, review, rating, class, author_details etc

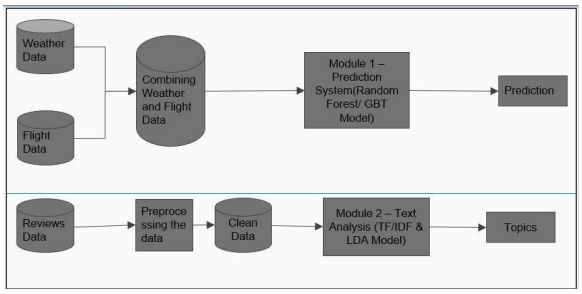
Tech Stack

Apache Spark

- Apache Spark (Scala) - Data Preprocessing for Machine Learning and Delay analysis
- Pyspark - Data preprocessing for Text Analysis
  - ➢ SparkMLlib - Machine Learning, Text Analysis
  - ➢ SparkSQL - Analysis of results

R Programming and Tableau – Visualizations

Architecture Design



Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

Design Goals

• Scalable Design

• Generic System

• Automated Workflow

• Fault Tolerance

• Modularity

Analyzing Delays

For analyzing delays, we took flight datasets of 2 years – 2006 and 2007 and processed it using SPARK SQL. We found some interesting patterns and variables which we used to visualize the data later using R. We used 4 variables for delay analysis:

• Month of year

• Day of Week

• Hour of Day

• Unique Carrier

Analysis: The analysis results are depicted in the result section below.

Visualization

We visualized the data by using R-shiny where we created a web application that generated interactive graphs and charts based on different variables. We used 4 variables to analyze data mentioned above. We also used 3 origin airports for comparison.

Shiny and Plotly

Shiny - Shiny is an R package that makes it easy to build interactive web applications (apps) straight from R. This lesson will get you started building Shiny apps right away.

Shiny apps have two components:

- a user-interface script
- a server script

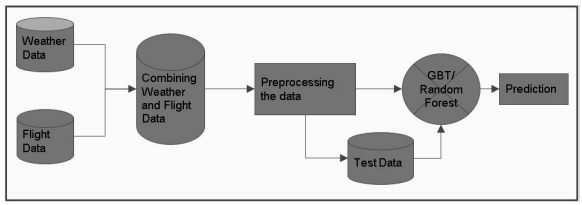
The user-interface (ui) script controls the layout and appearance of your app. It is defined in a source script named ui.R.

The server.R script contains the instructions that your computer needs to build your app.

Plotly – Plotly is graphing library which is used to make interactive graphs ,charts etc.

Module 1

Architecture



Delay Prediction Model

Our design of airline delay prediction consists of combining Flight data and Weather Data and using it to determine airline delay. Flight data consist of attributes like departure time, arrival time, length of flight, origin etc and weather data consist of parameters describing weather like wind speed, precipitation, visibility at the origin airport.

Model

The airline delay prediction model can be described in 4 stages:

- Stage 1: Combining Dataset
- Stage 2: Preprocessing
- Stage 3: Training Machine Learning Models
- Stage 4: Testing trained model

Stage 1: Combining Datasets

This stage comprises of combining weather and flight

Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

dataset. The dataset are combined on the basis of date and time attribute. Since we were considering 3 years worth of data, we took 2006 and 2007 data as the dataset for training and 2008 data as the test dataset.

Stage 2: Preprocessing

Stage 1 is followed by Stage 2 i.e preprocessing stage, after the datasets have been combined they are preprocessed for cleaning the dataset. Since it is a real world data it comprises of many irregularities, so for removing them we used following methods for preprocessing:

- Removal of null attributes:
  - Any row comprising of null values were eliminated from the dataset
- Categorical to numerical conversion.
  - There were attributes that were categorical in nature and since we were training a model for classification we needed to convert those variables into numerical values. This was achieved using String Indexer in scala.

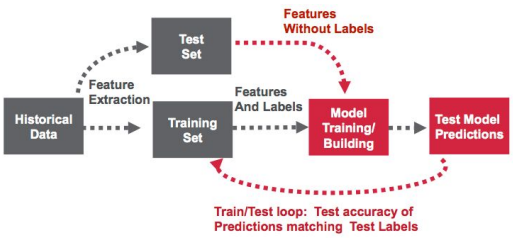
Stage 3: Training

After Preprocessing the data we separate the training and test data and we train the model on training data using cross-validation. We used two models for training purposes:

- Gradient Boosted Trees
- Random Forest

Both models were trained using the preprocessed stage and we are depicting here the best results that were obtained by using different hyper-parameter values.

The Training process is as depicted in diagram below:



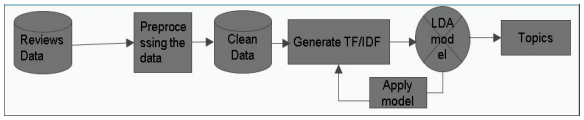
- First, the desired features are extracted from the Historical Data, we are taking into consideration all the parameters of our dataset in our model.
- Second, we divide the dataset into test and training dataset, training dataset consist of combined weather and flight data for the year 2006 and 2007, test data consist of 2008 dataset for flight and weather.
- Third, we train the model in our case Gradient Boosting Tree Model and Random Forest Model using cross validation. The training process keeps on repeating itself for desired number of iterations (hyper-parameter).

Stage 4: Model Prediction Accuracy Determination

The last stage comprises of Testing the model accuracy using test dataset. The results of delay prediction accuracy are determined below. From the Prediction result we can clearly see that as we combine the flight data with more weather data attributes the prediction accuracy of the model keeps on increasing which clearly depicts that weather also plays a big role in delay prediction.

Module 2

Architecture



In this module, the raw data for reviews is first preprocessed by tokenizing each review, removing stop words and converting to lower case. Further, this data is then passed through a TF/IDF model which applies term weights to each word. This model results in a vector which is then used for LDA clustering. The LDA clustering is an iterative step which works on the words repeatedly to generate topics that make the most sense. LDA is a bag of words model, hence the resulting topics are nothing but a bag of words which relate to a common topic, for example seats, legroom, storage etc all talk about cabin space. Based on these topics, interesting analysis can be carried out. If a topic is about mechanical issues, then data can be visualized using that topic to understand which airline faced the most issues due to mechanical problems and so on.

These are the steps used in this process.

Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

Text Analysis

• Pre-process reviews data

• Tokenize, remove stop words, special characters and words of short length

• Generate TF-IDF model

• Apply LDA Clustering

• Combine topics with original dataset

• Analyze interesting topics

Results

- ***Distributed by Month***

year	month	origin	number_of_delayed_flights
2006	1	LAX	3008
2006	1	JFK	1924
2006	1	SFO	2128
2006	10	LAX	3722
2006	10	JFK	2444
2006	10	SFO	2174
2006	11	LAX	3210
2006	11	SFO	2559
2006	11	JFK	2618
2006	12	SFO	3016
2006	12	LAX	4401
2006	12	JFK	3257
2006	2	LAX	3436
2006	2	SFO	2474
2006	2	JFK	1548
2007	2	SFO	2751
2007	2	JFK	2913
2007	2	LAX	3754
2007	3	LAX	3494
2007	3	JFK	3664
2007	3	SFO	1805
2007	4	SFO	2126
2007	4	JFK	3319
2007	4	LAX	3191
2007	5	JFK	2477
2007	5	LAX	2682
2007	5	SFO	2287
2007	6	LAX	3779
2007	6	SFO	2962
2007	6	JFK	3653
2007	7	SFO	3185
2007	7	JFK	3701
2007	7	LAX	4360
2006	2	JFK	1548
2006	3	SFO	3403
2006	3	JFK	1470
2006	3	LAX	4201
2006	4	JFK	1321
2006	4	LAX	3518
2006	4	SFO	2649
2006	5	SFO	1949
2006	5	JFK	1255
2006	5	LAX	3007
2006	6	JFK	2047
2006	6	SFO	2530
2006	6	LAX	3462
2006	7	JFK	2860
2006	7	SFO	2287
2006	7	LAX	3798
2006	8	SFO	2512
2006	8	LAX	3931

Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

2006	8	LAX	3931
2006	8	JFK	2595
2006	9	LAX	2970
2006	9	SFO	1898
2006	9	JFK	2166
2007	1	LAX	3655
2007	1	JFK	2865
2007	1	SFO	2080
2007	10	JFK	2110
2007	10	LAX	3077
2007	10	SFO	2823
2007	11	JFK	1967
2007	11	SFO	2524
2007	11	LAX	3422
2007	12	LAX	5439
2007	12	JFK	2947
2007	12	SFO	3432
2007	2	SFO	2751

- ***Distributed by Hour of Day***

timeOfDay	origin	numberOfDelayedFlights
0	JFK	2
0	LAX	699
0	SFO	369
1	LAX	104
1	SFO	12
2	JFK	1
5	JFK	165
5	LAX	31
5	SFO	3
6	JFK	986
6	LAX	1451
6	SFO	977
7	JFK	1297
7	LAX	2607
7	SFO	1745
8	JFK	4001
8	LAX	3516
8	SFO	2496
9	JFK	3108
9	LAX	3736
9	SFO	2441
10	JFK	1763
10	LAX	5392
10	SFO	4883
11	JFK	1693
11	LAX	5360
11	SFO	4614
12	JFK	2337
12	LAX	6739
12	SFO	6067
13	JFK	2920
13	LAX	5820
13	SFO	5265
14	JFK	1911
14	LAX	5701
14	SFO	4090
15	JFK	3281
15	LAX	5369
15	SFO	4421
16	JFK	7596
16	LAX	6293
16	SFO	3643
17	JFK	6982
17	LAX	5596
17	SFO	2624
18	JFK	4540
18	LAX	7170
18	SFO	4302
19	JFK	6366
19	LAX	4972
19	SFO	2534
18	LAX	7170
18	SFO	4302
19	JFK	6366
19	LAX	4972
19	SFO	2534
20	JFK	6326
20	LAX	5324
20	SFO	3497
21	JFK	3275
21	LAX	4070
21	SFO	2007
22	JFK	1774
22	LAX	4114
22	SFO	3613
23	JFK	277
23	LAX	2381
23	SFO	1131

Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

- ***Distributed by Week***

DayOfWeek	origin	numberOfFlightsDelayed
1	JFK	8626
1	LAX	12248
1	SFO	8860
2	JFK	7137
2	LAX	10122
2	SFO	8248
3	JFK	7869
3	LAX	11025
3	SFO	8267
4	JFK	9114
4	LAX	13444
4	SFO	8708
5	JFK	10745
5	LAX	15459
5	SFO	10141

Took 1 min 43 sec. Last updated by anonymous at May 08 201

2	LAX	10122
2	SFO	8248
3	JFK	7869
3	LAX	11025
3	SFO	8267
4	JFK	9114
4	LAX	13444
4	SFO	8708
5	JFK	10745
5	LAX	15459
5	SFO	10141
6	JFK	8346
6	LAX	10348
6	SFO	7435
7	JFK	8764
7	LAX	13799

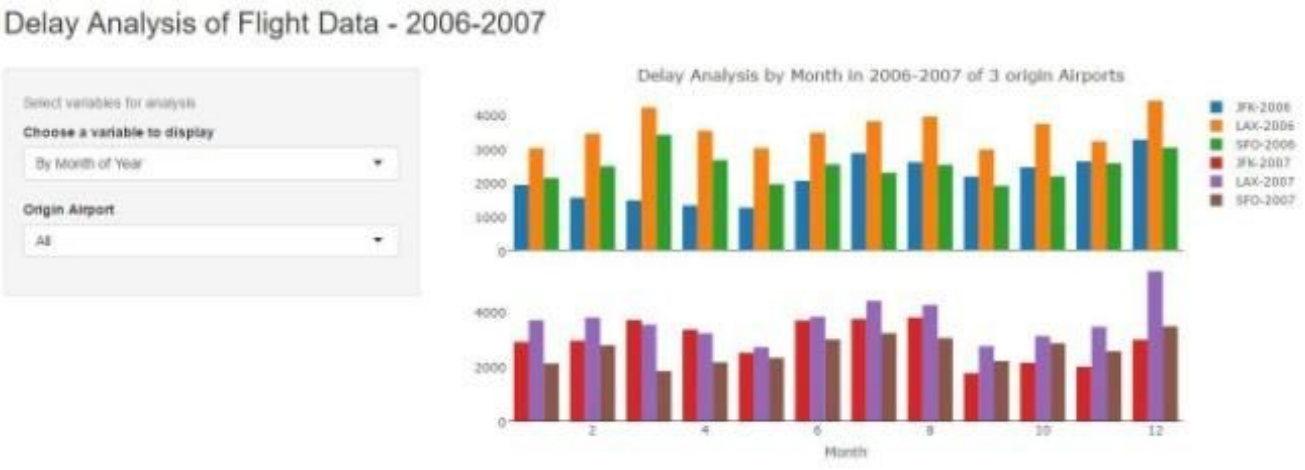
- ***Distributed by Carrier***

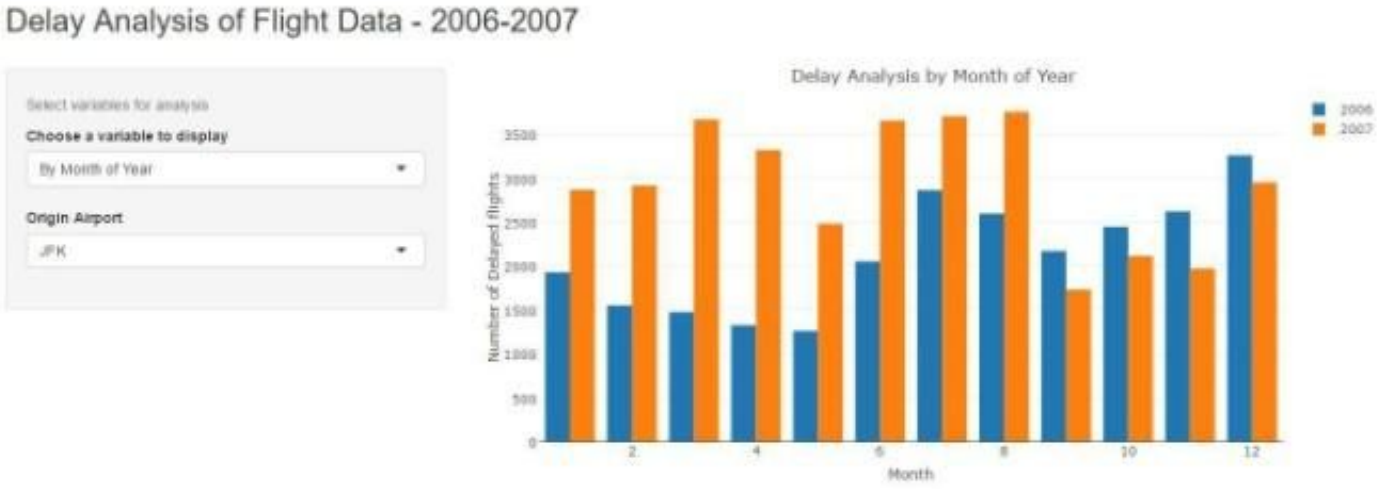
number_of_delayed_flights	origin	UniqueCarrier
992	JFK	NW
98	LAX	TZ
1073	SFO	NW
13606	LAX	AA
3475	LAX	AS
2544	LAX	US
2	LAX	B6
133	SFO	TZ
481	JFK	CO
2950	SFO	AS
22321	SFO	OO
13371	JFK	OH
2624	LAX	CO
635	SFO	MQ
3964	LAX	MQ

3475	LAX	AS
2544	LAX	US
2	LAX	B6
133	SFO	TZ
481	JFK	CO
2950	SFO	AS
22321	SFO	OO
13371	JFK	OH
2624	LAX	CO
635	SFO	MQ
3964	LAX	MQ
157	JFK	XE
92	SFO	YV
363	LAX	FL
184	LAX	HA
12808	LAX	UA

Exploratory Aviation Analysis and Recommendations

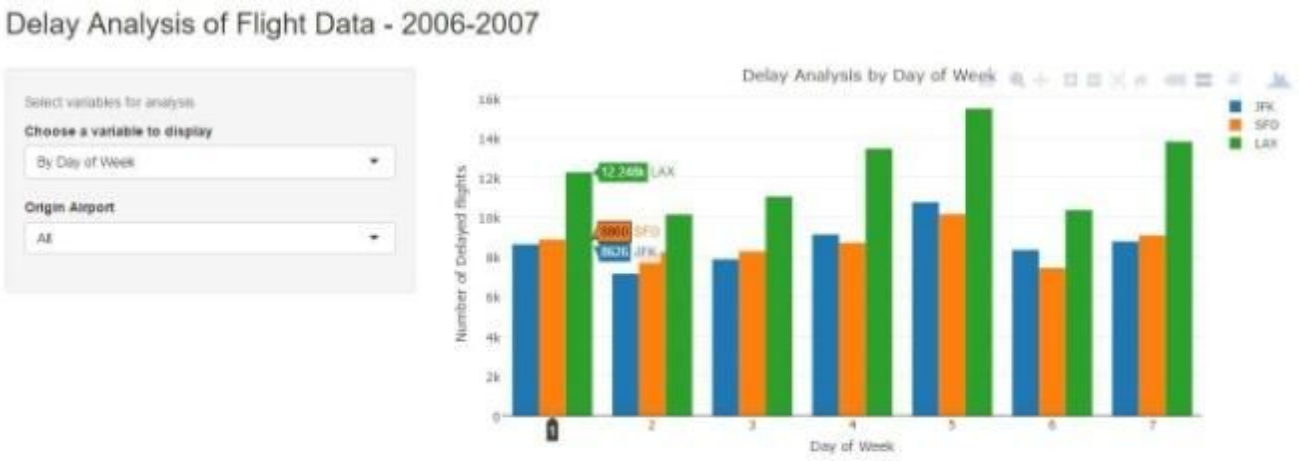
Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

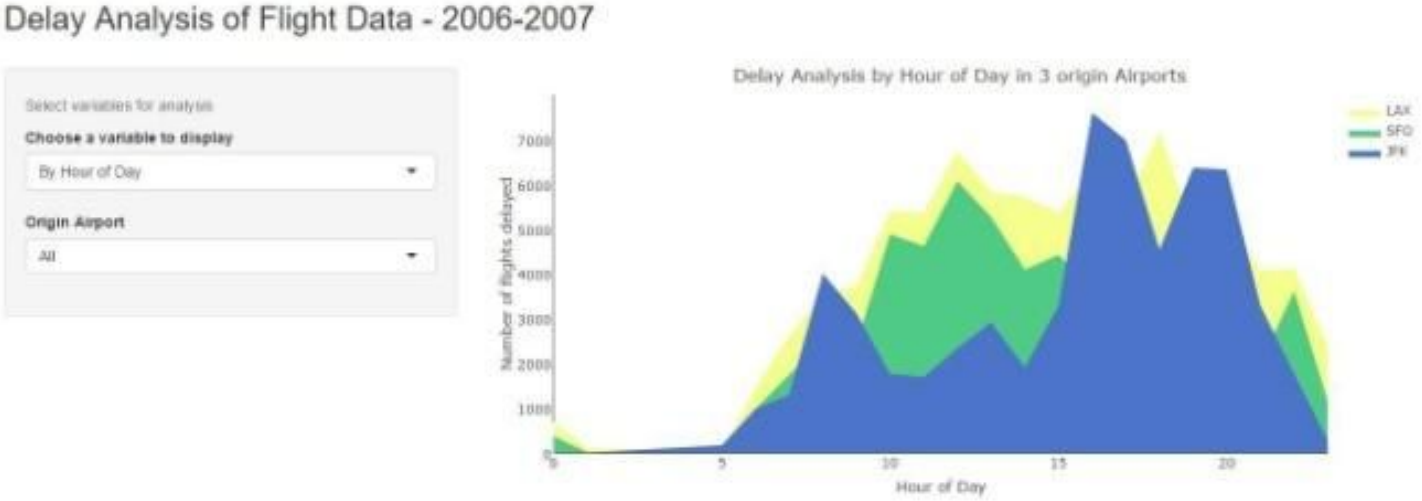




Exploratory Aviation Analysis and Recommendations

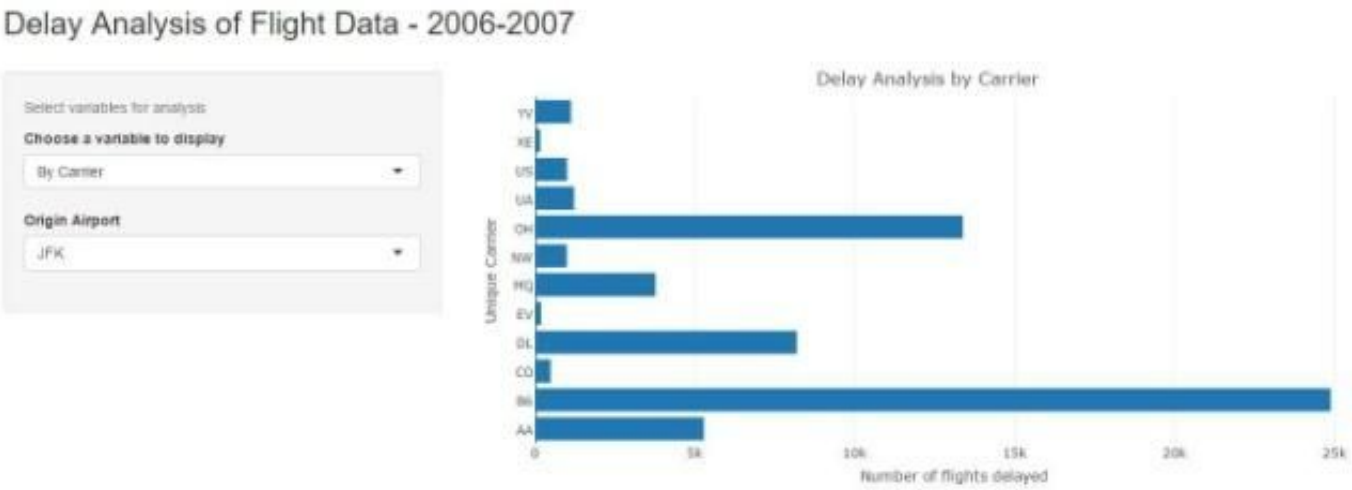
Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)





Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)



Module 1: Results

Model	FlightData	FlightData+WindSpeed	FlightData + WindSpeed + Visibility	FlightData + WindSpeed + Visibility + Precipitation
Gradient Boosted Tree	73.8%	74.02%	74.8%	76.5%
Random Forest Model	71.8%	73.5%	73.93%	74.2%

Module 2: Results

topic	topic_desc
0	[extra, row, main, segments, legroom, exit, cabin, pay, seat, seats]
1	[vegas, las, upgrade, economy, upgraded, seat, online, internal, dec, seats]
2	[oakland, nearly, phoenix, less, united, york, couple, lax, time, flights]
3	[san, phoenix, day, missed, connection, told, desk, connecting, diego, leaving]
4	[son, supervisor, boarding, different, told, seattle, agent, southwest, got, check]
5	[clt, wife, airways, class, miles, lga, club, coach, first, seats]
6	[phl, louisville, philadelphia, washington, laguardia, york, set, airways, rude, houston]
7	[carry, overhead, recline, seat, fit, ice, may, crew, drink, price]
8	[sfo, dtw, southwest, orlando, lhr, delta, trip, flights, seats, storm]
9	[min, delta, representative, check, bus, landing, anyone, terminal, baggage, got]
10	[lady, proceeded, wheelchair, ticket, name, cleveland, columbus, luggage, lost, ewr]
11	[paris, detroit, philly, hrs, united, wifi, flights, hour, family, plane]
12	[denver, mechanical, hours, delayed, airport, connecting, hour, night, next, stuck]
13	[priority, average, quality, guess, amsterdam, poor, nothing, etc, low, flights]
14	[gate, bags, minutes, plane, passes, milwaukee, told, charlotte, sat, boarding]
15	[business, delta, class, deltas, lax, 777, meal, cabin, sydney, usa]

Exploratory Aviation Analysis and Recommendations

Riya Patni(rsp378) Prateek Chourasia(pc1899) Nishchal Nagar(nn1123)

topic	topic_desc
0	[extra, row, main, segments, legroom, exit, cabin, pay, seat, seats]
1	[vegas, las, upgrade, economy, upgraded, seat, online, internal, dec, seats]
2	[oakland, nearly, phoenix, less, united, york, couple, lax, time, flights]
3	[san, phoenix, day, missed, connection, told, desk, connecting, diego, leaving]
4	[son, supervisor, boarding, different, told, seattle, agent, southwest, got, check]
5	[clt, wife, airways, class, miles, lga, club, coach, first, seats]
6	[phl, louisville, philadelphia, washington, laguardia, york, set, airways, rude, houston]
7	[carry, overhead, recline, seat, fit, ice, may, crew, drink, price]
8	[sfo, dtw, southwest, orlando, lhr, delta, trip, flights, seats, storm]
9	[min, delta, representative, check, bus, landing, anyone, terminal, baggage, got]
10	[lady, proceeded, wheelchair, ticket, name, cleveland, columbus, luggage, lost, ewr]
11	[paris, detroit, philly, hrs, united, wifi, flights, hour, family, plane]
12	[denver, mechanical, hours, delayed, airport, connecting, hour, night, next, stuck]
13	[priority, average, quality, guess, amsterdam, poor, nothing, etc, low, flights]
14	[gate, bags, minutes, plane, passes, milwaukee, told, charlotte, sat, boarding]
15	[business, delta, class, deltas, lax, 777, meal, cabin, sydney, usa]

ID	DATE	AIRLINE	REVIEW
10711	6/6/2015	American Airlines	The relatively short commute between Durango CO and Cleveland OH was the worst trip I have experienced. I have never had any trip since the inception of Denver International Airport that has gone without an aircraft maintenance problem. This recent trip held me hostage on a small aircraft and at the airport for 8 hours. I have never seen such an appalling record of mechanical breakdowns as I've seen with aircraft at Denver's airport. Both the departure and return trips were plagued with mechanical problems. The return trip was on board an aircraft meant for short duration flights of less than 1 hour and was used for a trip spanning over 3 hours. When we were supposed to circle the airport for a short period of time we had to divert to Cheyenne WY to refuel. The plane was too small to carry the necessary amount of fuel. As a result I missed my connection and waited 7 hours before taking off to my final destination. On the outbound trip we had to return to DIA due to a mechanical failure of an engine. We had already flown 45 minutes and instead of landing in an airport we were close to we had to fly back 45 minutes wait for another plane and take off in the middle of the night. We landed at 12.45 am. I have flown from different cities (Manhattan KS and Albuquerque NM) to Alaska three times and not once did I experience any problems with aircraft. The main difference was that none of these trips included United Airlines and Denver International Airport.
10888	4/24/2015	United Airlines	Traveled to Denver from Rochester NY on United twice in the past month. Every plane had a mechanical problem resulting in having to re-book multiple flights because of missed connections. In January I sat on the Tarmac for 3 hours in Rochester because someone had forgotten to drain the water out and it took that long to de-ice. On my return flight from Denver I had to be re-booked 3 times due to mechanical issues and sat in the airport for 8 hours. The only reason I finally got out was because a connecting flight from Idaho landed too late and they refused to hold the plane. In my most recent flight out of Rochester in February the altimeter was broken and we waited 6 hours for parts to be flown in from Atlanta. Oh and - I had paid extra for aisle seats in Economy Plus and ended up in the middle seat in the back of the plane. Now I have to contact Customer Service as no one in the Service Center in the Denver Airport is authorized to do so. I am dreading the return flight home. I will be flying on other carriers from now on.
10765	4/22/2015	American Airlines	My husband and I recently returned from a trip using United Airlines. We arrived at our international destination 12 hours later than the original arrival time. Since that was mainly due to weather I can't blame United. On the way home though we had a terrible experience. Our final flight from Houston to Fayetteville was supposed to board at 6.45pm but they kept delaying it around every half hour until finally at midnight they canceled the flight entirely. There were no more flights going out that night and the next day's flights to our destination were completely booked. We were given vouchers for a hotel which we got to at 2.30am and vouchers for food for the next day at the airport. We had to book a flight to Tulsa where our family had to drive a long way to pick us up. This plane was also delayed for an hour for no apparent reason. Upon reaching Tulsa we discovered our luggage had been separated and my bag was sent to our original destination while my husband's stayed back in Houston airport. We had to drive to Fayetteville airport to pick up my bag. The next day we had to drive back to Fayetteville to pick up my husband's bag. Definitely the last time we will ever fly with United.

Conclusion

- Flight delays are caused due to a combination of factors such as origin, destination, carrier, airTime, time of the year etc.
- Weather factors such as precipitation, visibility and wind speed also impact the flight schedules
- GBT model gave more accurate results in delay prediction
- Prediction can be improved by incorporating more factors such as manufacturer, aircraft design, age of aircraft

etc.

- Based on user reviews, airports and airlines can be rated based on different factors such as cleanliness, seat comfort, shopping experience, value for money etc.
- Text analysis helped us find out the most common topics that decide a user review.
- Using a prediction model on these topics, we can predict the user rating analyzing a new review.

Future Work

- Using this model on live streamed flight and weather data.
- Improve accuracy of prediction by including additional parameters about aircraft design.
- Automating the training process on streamed data after fixed intervals, to improve accuracy.
- Scale-up the model to analyze all the airlines from all sources and destinations across United States.