

Big Data Analysis: Airline performance

Indra Reddy Gayam
Jashwanth Gottipati
Karina Aguiar Goncalves
Srikrishna Krishnarao Srinivasan

Dataset - Airline on-time performance



- 2008 data
- 7,009,728 flights
- American Statistical Association http://stat-computing.org/dataexpo/2009/

Databricks - Spark - Pyspark







What is the most common reason for Cancellation?





What are the routes (Origin, Dest) with more cancellations caused by the Carrier?

- Routes on the top 10:
 - HOU-DAL / DAL-HOU
 - SFO-LAX / LAX-SFO
 - LAS-PHX / PHX-LAS
 - LGA-ORD / ORD-LGA
 - OGG-HNLI / HNL-OGG

- Carriers listing the top 10 Routes:
 - WN (Southwest Airlines Co.)
 - UA (United Air Lines Inc.)
 - HA (Hawaiian Airlines Inc.)
 - US Airways Inc. (US)
 - AA (American Airlines Inc.)



What are the routes (Origin, Dest) with more cancellations caused by the Weather?

- Cities listing the top 10 Routes:
 - Aspen, CO (Mountain location)
 - New York
 - Boston
 - Chicago



What are the routes (Origin, Dest) with more cancellations caused by the National Airspace System (NAS)?

- Airports listing the top 10 Routes:
 - La Guardia in New York
 - Chicago O'Hare International in Chicago

Carriers k-Means Cluster



- Features:
 - Average Carrier Delay in minutes
 - Average Arrival Delay in minutes
 - Average Departure Delay in minutes
 - Number of Flights (Volume)
 - Percentage of Cancellation (when caused by the Carrier)

Elbow Analysis => 6 clusters.





#Cluster Cen	Cluster Centers:										
#	[AvgCarrierDelay	AvgArrDelay	AvgDepDelay	NumFlights	PrgtCancellation]						
# CLUSTER 0:	[14.36	7.97	9.63	461,432.00	2.25%]						
# CLUSTER 1:	[20.97	2.32	2.95	79,122.50	0.79%]						
# CLUSTER 2:	[10.28	5.17	10.38	1,201,754.00	1.03%]						
# CLUSTER 3:	[19.05	9.86	11.04	250,221.00	2.15%]						
# CLUSTER 4:	[16.50	9.00	8.92	361,081.00	1.75%]						
# CLUSTER 5:	[16.62	9.60	10.36	586,022.00	2.53%]						

Cluster 2: Southwest*

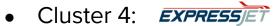
• Cluster 5: **Skyllest**

American Airlines



• Cluster 3: jetBlue Continental Airlines







• Cluster 1:





Analysis of Airlines Dataset to find Insights for Data Driven Decision

To Predict Cancellations



- Descriptive analysis of Cancellation of Flights
 - By Origin, By Airline, By Time period, Cancellation code
- Inferential statistical analysis of Cancellation of Flights
 - Chi-Square test to find variables highly responsible for Cancellation of flight
 - Multi-collinearity test for feature selection for Machine Learning modelling
- Predictive Machine Learning model for Cancellation of flight
 - Support Vector Classifier model to predict Cancellation
 - Model evaluation using Area Under the Curve, Accuracy and F1 Score
- Planned Improvements to the model

Descriptive analysis of Cancellations



Origin_airport Description Description Description American Airlines Inc. Count of Cancellation by Origin Airport and Airline B 54,904 A A Count by Cancellation code B A CancellationCode D A A CancellationCode D A A CancellationCode B A CancellationCode B A CancellationCode B A CancellationCode B A CancellationCode D A CancellationCode B A Co CancellationCode D A Co CancellationCode B A Co CancellationCode B A Co CancellationCode B A Co CancellationCode D A Co CancellationCode B A Co CancellationCode B A Co CancellationCode D A Co CancellationCode D A Co CancellationCode D A Co CancellationCode D A Co Co CancellationCode D A Co Co Co Co Co Co Co Co Co	Cancelled Cour				t of Cancelled	Cancelled		Count of Not Cancelled	
Origin_airport Chicago O'Hare International Dallas-Fort Worth International T,272 William B Hartsfield-Atlanta Intl Count of Cancellation by Airline American Eagle Airlines Inc. Top 3 Count of Cancellation by Airline American Airlines Inc. Top 3 Count of Cancellation by Airline American Airlines Inc. Top 3 Count of Cancellation by Airline B Hartsfield-Atlanta Intl Count of Cancellation by Airlines Inc. Top 3 CancellationCode Count by Cancellation code B 54,904 A 54,330 C 28,188 D 12 Chicago O'Hare International Chicago O'Hare American Airlines Inc. B 54,904 A 54,330 C 28,188 D 12	1			137,43	34	0 6,872,294			
Origin_airport Description American Airlines Inc. Chicago O'Hare International Chicago O'Hare American Airlines Inc. American Eagle Airlines Inc. 4,326 D Chicago O'Hare American Airlines Inc. American Airlines Inc. 2,926	Origin_airportAirportChicago O'Hare International15,050Dallas-Fort Worth International7,272					American Eagle Airlines Inc. American Airlines Inc.		Cancellation by Airline 18,331 17,440	Тор 3
International Chicago O'Hare International Chicago O'Hare International Chicago O'Hare American Eagle Airlines Inc. American Eagle Airlines Inc. 4,326 D 12 Chicago O'Hare American Airlines Inc. 2,926	Origin_airport	by			by Origin Airport and				
Chicago O'Hare International American Eagle Airlines Inc. 4,326 Chicago O'Hare American Airlines Inc. 2,926 D 12									
		American Eagle Airlines Inc.			4,326				
International	Chicago O'Hare International				2,926				

Inferential and Predictive Analytics of Cancellation:



Description						
American Eagle Airlines Inc.	13.34					
American Airlines Inc.	12.69					
Skywest Airlines Inc.	9.05					
Southwest Airlines Co.	9.01					
United Air Lines Inc.	7.67					

The area under the curve is 0.5

The area under the PR curve is 0.01969750732259068

The accuracy of the model is 0.9803024926774093

Below is the confusion matrix:

[[1030493 0 [20706 0]

pearson(features)

1.0 0.06575570036830161 ... (8 total) 0.06575570036830161 1.0 ...

0.0657252726295638 -0.24011421864393012

-0.0026678381935330564 -0.014908711207421689 .. 0.0068635648746472626 7.861196209730063F-4

-0.040400298332104526 -0.06761987574141898 ...

-0.009227404522645217 -0.005782797044295207 ... 0.0014049443489825 -4.897018765511002F-4

pValues: [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0]

degreesOfFreedom: [19, 302, 303, 4, 6, 1440, 11, 30] statistics: [22182.74291141317,30551.853383112266,29447 1565.8787188880342,4863239.444302091,20553.587598560807

Following are displayed here

- % Cancellation by Airline
- 2. Multi-collinearity (Pearson) results
- 3. Chi-Square test results (pValues)
- 4. Predictive model evaluation results (AUC, PR, ...)

PLANNED IMPROVEMENT:

- Use more features in the model to increase complexity
- Use Bagging and Ensemble method to improve generalization of the model

Analysis of Airlines Dataset to find Insights

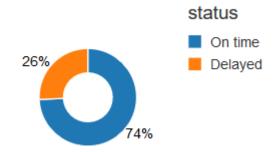


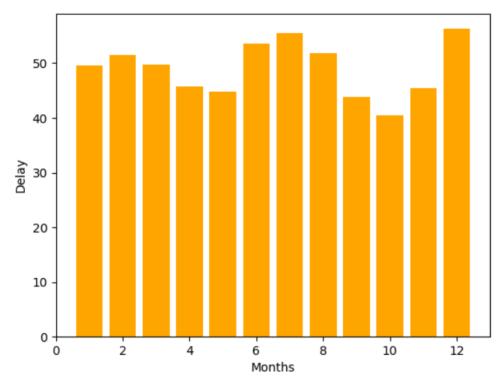
- Descriptive analysis of Delay of Flights
 - By Arrival, More than 10mins, By Month, Number of times per month
- Feature selection for building predictive model on Delay of Flights
 - Multi-collinearity test for feature selection for Machine Learning modelling
- Predictive Machine Learning model for Cancellation of flight
 - Liner Regression model to predict duration of Delay
 - **Decision Tree model**
 - Model evaluation using Area Under the Curve, Accuracy and F1 Score



Delay analysis







Descriptive Analytics – Delay by Carrier and Day



carrier_name	~	day_of_the_week
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)		friday
Pinnacle Airlines Inc.		friday
Aloha Airlines Inc.		sunday
Skywest Airlines Inc.		monday
American Eagle Airlines Inc.		friday
United Air Lines Inc.		friday
Comair Inc.		friday
Expressjet Airlines Inc.		thursday
Eroption Airlings Inc		thursday

Predictive Analytics of Flight Delay



predicted.show()

										OOOTH I LONIDA
eCarrier Dest 0	rigin M	onth Da	y0fWeek Arr	rTime Di	stance De	pDelay Un [.]	iqueCarrier_index Dest_	_index Origin_	_index feature	s prediction
9E ABE	DTW	1	2	2326	424	38	14.0	146.0	9.0 [14.0,146.0,9.0,1	. 33.83511183213857
9E ABE	DTW	1	5	1551	424	23	14.0	146.0	9.0 [14.0,146.0,9.0,1	. 35.05761422079986
9E ALB	DTW	1	4	1234	488	22	14.0	78.0	9.0 [14.0,78.0,9.0,1	. 36.1934204436858
9E ALB	DTW	1	4	1255	488	61	14.0	78.0	9.0 [14.0,78.0,9.0,1	. 36.17978984114072
9E ALB	MSP	1	3	1644	979	11	14.0	78.0	15.0 [14.0,78.0,15.0,1	. 35.39062499754558
9E ALO	MSP	1	3	2325	166	5	14.0	283.0	15.0 [14.0,283.0,15.0,	. 32.090768518797596
9E ALO	MSP	1	4	2331	166	14	14.0	283.0	15.0 [14.0,283.0,15.0,	. 32.326696809966414
9E ATL	HOU	1	1	1302	696	172	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 37.24755996497636
9E ATL	HOU	1	3	1037	696	3	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 37.89921068564678
9E ATL	HOU	1	3	1557	696	5	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 37.56169100357804
9E ATL	HOU	1	4	1645	696	40	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 37.74439503718987
9E ATL	HOU	1	5	1557	696	19	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 38.041336501655735
9E ATL	HOU	1	5	1600	696	17	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 38.01342622025389
9E ATL	HOU	1	6	1350	696	105	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 38.4155180472104
9E ATL	HOU	1	7	1445	696	149	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 38.59367854664053
9E ATL	HOU	1	7	1551	696	3	14.0	0.0	25.0 [14.0,0.0,25.0,1	. 38.52487645760345
9E ATL	IAH	1	1	1752	689	6	14.0	0.0	5.0 [14.0,0.0,5.0,1.0	. 36.18672674010886
9E ATL	IAH	1	2	2140	689	15	14.0	0.0	5.0 [14.0,0.0,5.0,1.0	. 36.17470788021949
9E ATL	IAH	1	4	1803	689	13	14.0	0.0	5.0 [14.0,0.0,5.0,1.0	. 36.87309209533019
9E ATL	IAH	1	4	2127	689	9	14.0	0.0	5.0 [14.0,0.0,5.0,1.0	. 36.662791370348906
	9E ABE 9E ABE 9E ALB 9E ALB 9E ALO 9E ALO 9E ATL 9E ATL	9E ABE DTW 9E ABE DTW 9E ALB DTW 9E ALB DTW 9E ALB MSP 9E ALO MSP 9E ALO MSP 9E ATL HOU 9E ATL HOU	9E ABE DTW 1 9E ABE DTW 1 9E ALB DTW 1 9E ALB DTW 1 9E ALB MSP 1 9E ALO MSP 1 9E ALO MSP 1 9E ATL HOU 1 9E ATL IAH 1 9E ATL IAH 1	9E ABE DTW 1 2 9E ABE DTW 1 5 9E ALB DTW 1 4 9E ALB DTW 1 4 9E ALB MSP 1 3 9E ALO MSP 1 3 9E ATL HOU 1 3 9E ATL HOU 1 3 9E ATL HOU 1 5 9E ATL HOU 1 6 9E ATL HOU 1 7 9E ATL IAH 1 1 4	9E ABE DTW 1 2 2326 9E ABE DTW 1 5 1551 9E ALB DTW 1 4 1234 9E ALB DTW 1 4 1255 9E ALB MSP 1 3 1644 1 3 2325 9E ALO MSP 1 4 2331 2325 9E ALO MSP 1 4 2331 2325 9E ATL HOU 1 1 3 1037 3 1037 9E ATL HOU 1 3 1557 3 1557 9E ATL HOU 1 4 1645 3 1557 9E ATL HOU 1 5 1557 3 1037 9E ATL HOU 1 7 1445 3 1037 9E ATL HOU 1 7 1445 3 1037 9E ATL HOU 1 7 1551 3 1037 9E ATL ATL 1 1 1 1752 3 1037 9E ATL IAH 1 4 1803 9E ATL IAH 1 4 1803	9E ABE DTW 1 2 2326 424 9E ABE DTW 1 5 1551 424 9E ALB DTW 1 4 1234 488 9E ALB DTW 1 4 1255 488 9E ALB MSP 1 3 1644 979 9E ALO MSP 1 3 2325 166 9E ALO MSP 1 4 2331 166 9E ATL HOU 1 1 1302 696 9E ATL HOU 1 3 1037 696 9E ATL HOU 1 3 1557 696 9E ATL HOU 1 4 1645 696 9E ATL HOU 1 5 1557 696 9E ATL HOU 1 5 1557 696 9E ATL HOU 1 6 1350 696 9E ATL HOU 1 7 1445 696 9E ATL HOU 1 7 1551 696 9E ATL HOU 1 7 1551 696 9E ATL IAH 1 1 1752 689 9E ATL IAH 1 2 2140 689 9E ATL IAH 1 4 1803 689	9E ABE DTW 1 2 2326 424 38 9E ABE DTW 1 5 1551 424 23 9E ALB DTW 1 4 1234 488 22 9E ALB DTW 1 4 1255 488 61 9E ALB MSP 1 3 1644 979 11 9E ALO MSP 1 3 2325 166 5 9E ALO MSP 1 4 2331 166 14 9E ATL HOU 1 1 1302 696 172 9E ATL HOU 1 3 1037 696 3 9E ATL HOU 1 3 1557 696 5 9E ATL HOU 1 4 1645 696 40 9E ATL HOU 1 5 1557 696 19 9E ATL HOU 1 6 1350 696 105 9E ATL HOU 1 7 1445 696 149 9E ATL HOU 1 7 1551 696 3 9E ATL IAH 1 1 1752 689 6 9E ATL IAH 1 1 1752 689 6 9E ATL IAH 1 4 1803 689 13	9E ABE DTW 1 2 2326 424 38 14.0 9E ABE DTW 1 5 1551 424 23 14.0 9E ALB DTW 1 4 1234 488 22 14.0 9E ALB DTW 1 4 1255 488 61 14.0 9E ALB MSP 1 3 1644 979 11 14.0 9E ALD MSP 1 3 2325 166 5 14.0 9E ALO MSP 1 4 2331 166 14 14.0 9E ATL HOU 1 1 1302 696 172 14.0 9E ATL HOU 1 3 1037 696 3 14.0 9E ATL HOU 1 3 1557 696 5 14.0 9E ATL HOU 1 5 1557 696 19 14.0 9E ATL HOU 1 5 1557 696 19 14.0 9E ATL HOU 1 5 1557 696 19 14.0 9E ATL HOU 1 5 1557 696 19 14.0 9E ATL HOU 1 7 1445 696 149 14.0 9E ATL HOU 1 7 1445 696 149 14.0 9E ATL HOU 1 7 1551 696 3 14.0 9E ATL HOU 1 7 1551 696 3 14.0 9E ATL HOU 1 7 1551 696 3 14.0 9E ATL HOU 1 7 1551 696 3 14.0 9E ATL IAH 1 1 1 1752 689 6 14.0 9E ATL IAH 1 1 1 1803 689 15 14.0	9E ABE DTW 1 2 2326 424 38 14.0 146.0 9E ABE DTW 1 5 1551 424 23 14.0 78.0 9E ALB DTW 1 4 1234 488 22 14.0 78.0 9E ALB DTW 1 4 1255 488 61 14.0 78.0 9E ALB MSP 1 3 1644 979 11 14.0 78.0 9E ALD MSP 1 3 2325 166 5 14.0 283.0 9E ALO MSP 1 4 2331 166 14 14.0 283.0 9E ATL HOU 1 1 1302 696 172 14.0 0.0 9E ATL HOU 1 3 1037 696 3 14.0 0.0 9E ATL HOU 1 3 1557 696 5 14.0 0.0 9E ATL HOU 1 5 1557 696 19 14.0 0.0 9E ATL HOU 1 5 1557 696 19 14.0 0.0 9E ATL HOU 1 5 1557 696 19 14.0 0.0 9E ATL HOU 1 5 1557 696 19 14.0 0.0 9E ATL HOU 1 5 1600 696 17 14.0 0.0 9E ATL HOU 1 7 1445 696 149 14.0 0.0 9E ATL HOU 1 7 1445 696 149 14.0 0.0 9E ATL HOU 1 7 1551 696 3 14.0 0.0 9E ATL IAH 1 1 1752 689 6 14.0 0.0 9E ATL IAH 1 4 1803 689 15 14.0 0.0	9E ABE DTW 1 5 1551 424 23 14.0 146.0 9.0 [14.0,146.0,9.0,1] 9E ALB DTW 1 4 1234 488 22 14.0 78.0 9.0 [14.0,78.0,9.0,1] 9E ALB DTW 1 4 1255 488 61 14.0 78.0 9.0 [14.0,78.0,9.0,1] 9E ALB MSP 1 3 1644 979 11 14.0 78.0 15.0 [14.0,78.0,15.0,1] 9E ALD MSP 1 3 2325 166 5 14.0 283.0 15.0 [14.0,283.0,15.0,] 9E ALD MSP 1 4 2331 166 14 14.0 283.0 15.0 [14.0,283.0,15.0,] 9E ATL HOU 1 1 1302 696 172 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 3 1637 696 3 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 3 1557 696 5 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 4 1645 696 40 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 5 1557 696 19 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 5 1557 696 19 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 7 1445 696 149 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 7 1445 696 149 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 7 1445 696 3 14.0 0.0 25.0 [14.0,0.0,25.0,1] 9E ATL HOU 1 7 1445 696 3 14.0 0.0 25.0 [14.0,0.0,5.0,1] 9E ATL HOU 1 7 1445 696 3 14.0 0.0 25.0 [14.0,0.0,5.0,1]

only showing top 20 rows

RMSE: 35.042989241330005

MSE: 1228.0110949679706



