BUDT 733 – Data Analysis for Decision Makers

Team Project

What drives flight delays between DC and Honolulu?

TEAM #1

Prashant Bhaip

Andres Garay

Vedat Kaplan

Grigoriy Vinogradov

Ling Wang

December 11, 2007

Executive Summary

The goal of this project was to find the major factors explaining flight delays on route from Washington, DC to Honolulu, Hawaii. The findings of this analysis that are under control of travelers, can then be used to make better decisions on when to travel or avoid traveling, which Washington, DC airport to use and which airline to fly.

We gathered the statistics on airline's on-time performance, as well as detailed weather conditions for major airports, and some other indicators including Dow Jones Industrial stock index and oil price that reflect the overall state of the economy. Since there is no direct flights from Washington DC area to Honolulu, we gathered the airline on-time information from Washington DC area to various hubs and then from those hubs to Honolulu. Several techniques were used to explore and analyze the data, and as a result a different model was fitted for each one of the legs that makes up a trip from Washington to Honolulu.

The results of this analysis demonstrated that weather does play a major role in explaining why flights are delayed. Unfortunately, in most cases accurate weather forecasts are not available at the time when a flight is booked. On the other hand, our analysis uncovered that summer is the worst time to travel to Hawaii in terms of probability of encountering a flight delay. Another interesting finding is that Continental Airlines has the best track record in terms of not experiencing flight delays in comparison to four other examined airlines: American, Delta, Northwest and United. These two later findings can be of great importance for travelers from Washington, DC who want to avoid delayed flights when traveling on holiday to Honolulu.

Finally, we also found that not all the factors have the same effects when they are interacting versus when they are used alone. Some airports and some airlines are more prone than the others to situations such as high volume of passengers, or adverse weather. This fact just reiterates the difficulty of explaining a flight delay in terms of a basic set of causes, as most situations involve a chain of events that incorporates interaction between different factors.

Technical Summary

Data Selection and Processing

As the primary data source, we used airline on-time performance data collected by Bureau of Transportation Statistics (BTS). We combined the dataset we obtained from BTW with weather information provided by National Oceanic and Atmospheric Administration (NOAA). Finally, we added two additional predictors: daily price of oil from Bloomberg and performance of Dow Jones Industrial Index from Yahoo! Finance. These two variables would act as proxies for a general price index of air travel, as well as an overall state of the economy. A snapshot of data is shown in Exhibit A. Our analysis examined flights between all three Washington, DC airports and six major hubs en route to Hawaii: Atlanta, Dallas/Fort Worth, Houston, Minneapolis, Chicago and San Francisco. We selected flights operated by five major US airlines: American, Continental, Delta, Northwest and United (who operates flights for US Airways as well). The final dataset used for this analysis contained 11,822 records and covered flights from January 2005 through September 2007.

Data Exploration

To explore patterns present in the selected data set, we used Spotfire data visualization software and Excel pivot tables. A sample of informative graphs and related observations can be found in Exhibit A. At a high level, the exploratory analysis gave us an insight that time of travel as well as weather-related factors play a major role in flight delays between Washington, DC and Honolulu.

Model Results

To analyze exploratory power of various predictors available in our dataset, we applied three classification models: classification tree, discriminant analysis and logistic regression.

The resultant classification trees and accompanying confusion matrices for flights between Washington, DC and hubs and from hubs to Hawaii are shown in Exhibits B and C respectively. It can be seen from these exhibits, the most significant predictors, as

shown at the top of the two trees, are weather-related factors, such as destination precipitation, origin and destination visibility and origin and destination wind speed. For flights, between the hubs and Honolulu, however, the most important predictor according to classification tree is distance. This seems a little odd, given that none of the exploratory analysis, or the other models did find that distance was a significant predictor.

The results of discriminant analysis can be found in Exhibit D. Similar to the classification tree analysis, it identified weather-related factors as important predictors. In addition, it identified Continental Airlines as the carrier which has significantly better on-time records than the other four airlines in our dataset.

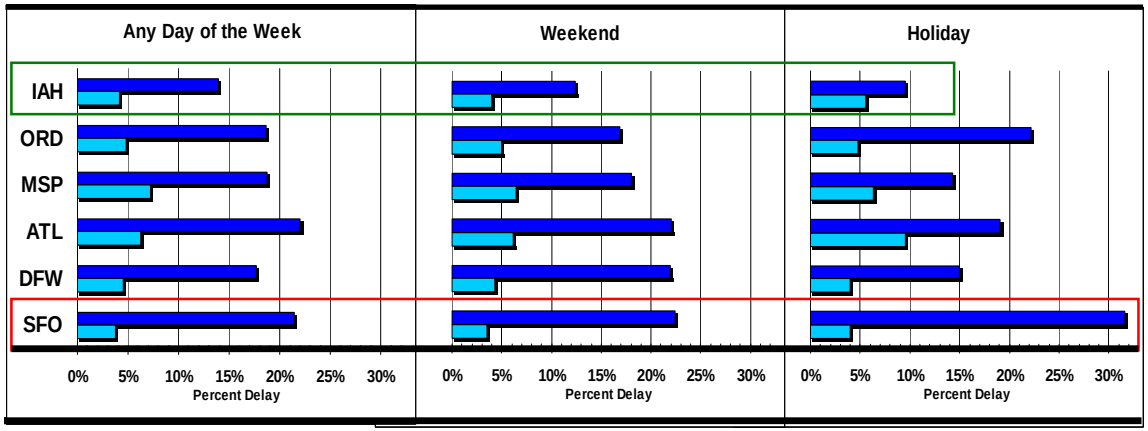
Finally, output of logistic regression is also presented in Exhibit D. We found logistic regression to be the most useful tool for analyzing flight delays based on the selected data set. It produced the lowest overall error rate. As in the case with other two classification methods, logistic regression models identified weather-related factors as significant predictors. These models also supported our finding from discriminant analysis that flights operated by Continental Airlines are least likely to be delayed. In addition, logistic regression identified summer as the most likely season for delays.

Conclusions

- Weather conditions at both origin and destination airports play a major role in explaining flight delays between Washington, DC airports and Honolulu. However, because weather cannot be accurately predicted at the time when travelers book their flights, this finding is not very helpful for future travelers.
- The worst season to travel to Hawaii in terms of flight delays is summer.
- Continental Airlines has the best track record in terms of not experiencing flight delays in comparison to four other examined airlines: American, Delta, Northwest and United.
- In broad terms, flight distance is not an important factor that causes delays. However, the San Francisco airport seems to be an exception to this finding.

Exhibit A – Data, Exploratory Graphs

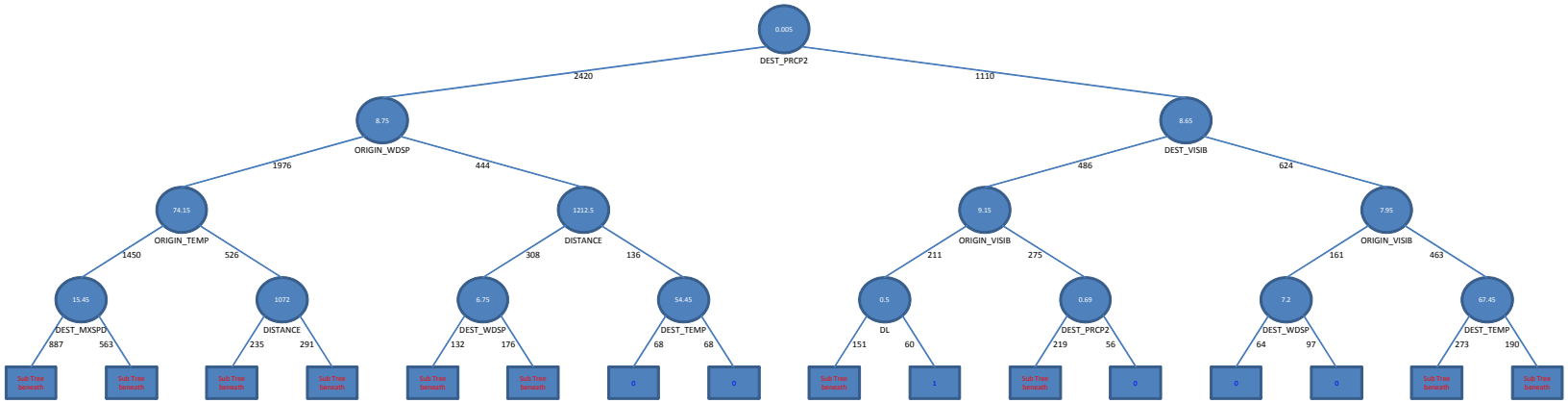
DELAY	Weekend	Holiday	OilPrice	MONTH	CARRIER	DAY_OF_V	ORIGIN	DEST	ORIGIN_TE	DEST_VIS	DJ_Close
No	1	0	76.70	9	UA	7	IAD	SFO	77.7	10	13113.38
No	1	0	76.70	9	UA	7	BWI	ORD	75.5	10	13113.38
Yes	1	0	76.70	9	UA	6	IAD	SFO	81.1	9.5	13113.38
No	1	0	76.70	9	UA	6	BWI	ORD	78.3	10	13113.38
No	0	0	76.70	9	UA	5	IAD	SFO	81.1	10	13113.38
No	0	0	76.70	9	UA	5	BWI	ORD	78.8	8.3	13113.38
No	0	0	76.30	9	UA	4	IAD	SFO	80.5	9.7	13363.35
Yes	0	0	76.30	9	UA	4	BWI	ORD	77.7	8.7	13363.35
No	0	0	75.73	9	UA	3	IAD	SFO	75.9	10	13305.47
No	0	0	75.73	9	UA	3	BWI	ORD	74.7	8.8	13305.47
No	0	0	75.05	9	UA	2	IAD	SFO	70.8	9.7	13448.86
No	0	0	75.05	9	UA	2	BWI	ORD	71.4	9.8	13448.86
No	1	0	81.66	9	UA	7	IAD	SFO	61.9	10	13895.63
No	1	0	81.66	9	UA	7	BWI	ORD	63.3	10	13895.63
No	0	1	74.04	9	UA	1	IAD	SFO	75.1	10	13357.74
Yes	0	1	74.04	9	UA	1	BWI	ORD	72	10	13357.74
No	1	0	81.66	9	UA	6	IAD	SFO	65.4	9.9	13895.63
No	1	0	81.66	9	UA	6	BWI	ORD	65.7	10	13895.63
Yes	0	0	81.66	9	UA	5	IAD	SFO	71	10	13895.63
No	0	0	81.66	9	UA	5	BWI	ORD	72.1	10	13895.63
No	0	0	82.88	9	UA	4	IAD	SFO	77.8	10	13912.94



Observation: Median temperature in case if there was a delay is always higher than median temperature when there was no delay, except for Minneapolis.

Continental has the best on-time record generally speaking.

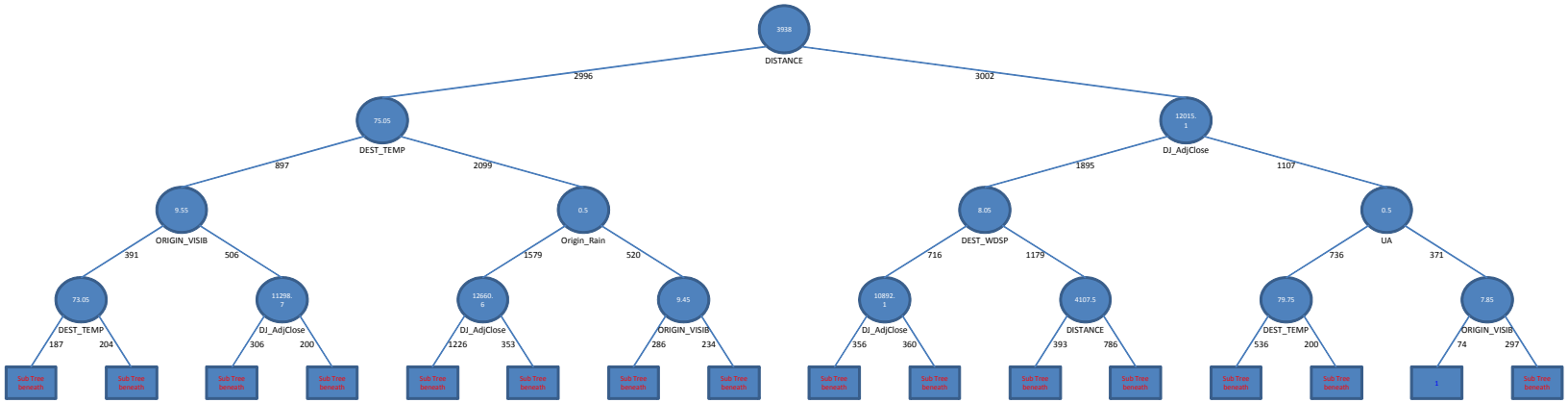
Exhibit B – Classification Tree for Flights between Washington, DC and Selected Hubs



Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	485	182
0	1072	1791

Error Report			
Class	# Cases	# Errors	% Error
1	667	182	27.29
0	2863	1072	37.44
Overall	3530	1254	35.52

Exhibit C – Classification Tree for Flights between Selected Hubs and Honolulu



Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	1645	192
0	2608	1553

Error Report			
Class	# Cases	# Errors	% Error
1	1837	192	10.45
0	4161	2608	62.68
Overall	5998	2800	46.68

Exhibit D – Discriminant Analysis & Logistic regression

Washington, DC to Selected Hubs

Variables	Delay	No Delay	Difference
Constant	-16676.15	-16675.8	-0.32031
Fri_Sun	-10.74369	-10.7999	0.056263
Winter	15.327852	15.274632	0.05321
Spring	104.09368	104.34243	-0.24875
Summer	17.53158	17.215236	0.316350
UA	16.26759	16.318662	-0.051065
DL	31.135717	30.924640	0.211076
AA	28.033428	28.172458	-0.139030
CO	-15.21126	-14.74134	-0.469915
DISTANCE	0.009946	0.009790	0.000155
ORIGIN_TEMP	-1.627411	-1.621275	-0.006136
ORIGIN_DEWP	7.643952	7.648923	-0.00497
ORIGIN_SLP	32.222812	32.2223	0.000480
ORIGIN_VISIB	-4.290945	-4.171614	-0.11933
ORIGIN_WDSP	28.164411	28.116142	0.048269
Origin_Fog	-3.497109	-3.612166	0.115056
Origin_Rain	51.661319	51.716835	-0.055515
Origin_Snow	107.21433	106.91032	0.304008
Origin_Thunder	11.141178	11.078694	0.062483
Origin_Tornado	-66.73384	-68.54547	1.811622
DEST_TEMP	1.579037	1.585049	-0.006011
DEST_DEWP	-3.180112	-3.18889	0.008778
DEST_SLP	0.015332	0.015176	0.000155
DEST_VISIB	13.622522	13.793818	-0.171296
DEST_WDSP	0.79251	0.731281	0.061235
Dest_Fog	0.26841	0.195923	0.072494
Dest_Rain	11.013598	10.708426	0.305171
Dest_Snow	75.689323	75.753402	-0.064079
Dest_Hail	-29.51551	-30.05800	0.54248
Dest_Thunder	-3.435158	-3.794443	0.359285
DJ_AdjClose	0.002438	0.002301	0.000137

Selected Hubs to Honolulu

Variables	Delay	No Delay	Difference
Constant	-1681.757	-1687.81	6.058715
Fri-Sun	6.656792	6.675014	-0.018222
Weekend	-1.884532	-1.805045	-0.079487
Holiday	7.428784	7.335012	0.093772
Winter	94.896202	94.925437	-0.029235
Spring	67.325805	67.347541	-0.021736
Summer	-16.31361	-16.52402	0.210403
UA	17.219245	17.988590	-0.769344
DL	6.803079	7.214820	-0.411741
AA	12.439149	13.216532	-0.777382
CO	16.188005	17.138525	-0.950519
DISTANCE	0.014214	0.014092	0.000122
ORIGIN_TEMP	-0.179785	-0.182327	0.002542
ORIGIN_VISIB	8.621864	8.725420	-0.103556
ORIGIN_WDSP	0.873870	0.85265	0.021218
Origin_Fog	13.84321	13.592679	0.250539
Origin_Rain	10.810284	10.683654	0.126629
Origin_Snow	15.274599	15.094016	0.1805
Origin_Hail	-23.56042	-23.31077	-0.249650
Origin_Thunder	-9.976094	-10.31680	0.340709
DEST_TEMP	17.852846	17.923374	-0.070528
DEST_VISIB	176.68241	176.82189	-0.139480
DEST_WDSP	-5.290013	-5.275408	-0.014605
Dest_Fog	138.67860	138.44540	0.233200
Dest_Rain	36.433460	36.549209	-0.115749
Dest_Thunder	102.84405	102.85617	-0.01212
Dest_Tornado	-48.56667	-48.6370	0.070411
DJ_AdjClose	0.007226	0.007095	0.000131

Washington, DC to Selected Hubs

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.767286	0.458512	0.094243	*
Spring	-0.274237	0.092221	0.00294	0.760151
Summer	0.381736	0.119182	0.00136	1.464825
CO	-0.45578	0.108944	0.000028	0.633949
DISTANCE	0.000128	0.000054	0.018602	1.000128
ORIGIN_TEMP	-0.01239	0.002953	0.000026	0.987678
ORIGIN_VISIB	-0.109990	0.019804	0	0.895842
ORIGIN_WDSP	0.046074	0.011725	0.00008	1.047152
DEST_VISIB	-0.162576	0.025056	0	0.849950
DEST_WDSP	0.053189	0.010504	0.000000	1.054629
Dest_Rain	0.348928	0.087340	0.000064	1.417547
Dest_Thunder	0.35605	0.114137	0.001811	1.427692
DJ_AdjClose	0.000122	0.000031	0.000117	1.000122

Selected Hubs to Washington, DC

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	99.518493	14.110648	0	*
Summer	0.195646	0.08895	0.027855	1.216097
ATL	0.444471	0.079112	0.000000	1.559666
MSP	0.766332	0.081051	0	2.151860
ORIGIN_TEMP	0.016068	0.00570	0.004882	1.016198
ORIGIN_DEWP	-0.017736	0.005940	0.002829	0.982420
ORIGIN_VISIB	-0.152028	0.024924	0	0.858963
ORIGIN_WDSP	0.023464	0.009129	0.010160	1.02374
Origin_Rain	0.208670	0.078001	0.007468	1.232038
Origin_Thunder	0.375603	0.100149	0.000176	1.455870
DEST_TEMP	-0.038676	0.013646	0.004594	0.962061
DEST_DEWP	-0.036387	0.011606	0.001717	0.964266
DEST_SLP	-0.092799	0.013810	0	0.911376