Predicting Flight Delays and Cancellations

Kyle Durfee, Mehrad Moradshahi, Ajit Punj
(kpdurfee, mehrad, apunj)

October 26, 2017

1 Task Definition

1.1 Introduction

Nothing ruins best laid plans more than a canceled or delayed flight. Our project uses data provided by the United States Department of Transportation to predict whether a flight will be canceled, and if not to predict the minute value of any delays. In order to do more than simply identify an inconvenience our project will also provide the user with a list of recommended alternatives should they anticipate issues.

1.2 Initial Development and Evaluation

Our data has been pulled from the United States Department of Transportation, Bureau of Transportation Statistics. We plan on using data from 1990 to 2008 to train and evaluate the effectiveness of our system. Due to the large number of flights in the United States every day we will consider only data in the month of December for this initial stage, as it is of particular interest and is one of the most popular months for travel. We plan on using the first 60% of our data set to train our model, the next 30% as as a validation data set and the most recent 10% to test our system. An example of preliminary data (a subset of 40 flights from one day in 2008) is shown in appendix A. Using the preliminary data in appendix A a user could specify that they are considering a flight and see the output in Appendix B. Note that the flight suggestions in this example are limited to the exact same day and airline given the condensed nature of the preliminary data in appendix A.

2 Infrastructure

To implement our baseline we needed to create two scripts: one to generate "clean" data files, and one to actually calculate our baseline. The cleaning script takes a large set of data as a CSV file, and extracts only the relevant rows of data. Once the clean data script outputs "clean" CSV files, we use the new CSV output files as inputs to the baseline script to predict canceled and delayed flights. We have multiple methods in our baseline script: one with random predictions and getting the error between the predictions and actual values, and one with the domain specific prediction estimated that flights closer to Christmas will have a higher chance of cancellation or delay.

3 Approach

Our project will focus on the ever present and frustrating issue of unexpected flight delays and cancellations. A user will be able to input information about an upcoming flight (starting airport, destination airport, airline, date, time etc) and our system will provide them with a predicted delay (in minutes or a cancellation). To ease the pain and attempt to solve a problem as opposed to simply identifying one, the system will also group planned flights and propose similar alternatives so travelers are able to modify their plans accordingly. To evaluate the success of our system, we will look at the accuracy of the cancellation predictions (as a binary prediction) and the accuracy of the minute to minute delay predictions (difference from actual). This information is part of our data set and so both are known values. As there is no concrete metric for flight similarity we will use a human oracle to determine the effectiveness of the systems proposals for alternate flights based on factors like start and end location, flight duration, and overall difference in arrival time. Our research was unable to unearth a useful oracle as other methods either rely on different data sets (weather, news feeds etc) or are unable to do much better than random guessing. For this reason we consider our Oracle to be 70% accuracy in binary delay prediction and with delay values within 20 minutes.

3.1 Baseline and Oracle

The baseline that we implemented was to consider delays and cancellations as random events (as travelers are essentially forced to do), taking probabilities and delay times from the same source from which we get our data for the year of 2008 . While not ideal, attempting to give certain features weights resulted in worse results than our purely random approach. Doing this gives us a lower bound, if our model is able to determine any meaningful relationships between inputs and outputs it should be able to outperform this random implementation. In this blind, baseline model cancellations are assumed to have a chance of 2.18 percent and the delays were modeled as a 21.7 percent chance with an average delay of 57 minutes. Using this baseline we calculated the accuracy of our cancellation predictor to be 0 percent and the average minute error of our delay predictor to be twenty seven minutes on our cleaned data for the year of 2008 and averaging the results over 1000 iterations. We attempted to predict weighting flights being closer to Christmas to be more likely to be delayed/canceled but actually got worse results, cancellations 0 percent accurate and an increased average delay estimation error of 40.22 minutes was reported. Our baseline recommendations simply chose the three flights with the closest departure time from the same origin airport as the user input to the same destination airport and offered by the same airline. This is a reasonable baseline as it makes a very simple correlation between a minimal number of features. Our human Oracle noted that the recommendations were not optimal as they could be delayed or canceled themselves, and the arrival/departure times as well flight duration varied significantly which is a clear inconvenience from a travel planning standpoint.

3.2 Challenges

Here is a list of challenges we might face during this project and their possible solutions:

- Finding useful features
  - Using CNN to set weights
- Clustering similar flights for prediction
  - K means based on a customized loss function
  - loss focuses on time differences
- Finding hidden factors outside our data set which might impact our model
  - Adding other data features such as delay in arrival flights, total number of people at the airport at the time of flight, or any recent plane accidents
- Treating canceled flights
  - Options: consider them as flights with infinite (high) delay time, nix them, or give a cancellation probability
- Accommodating all the data and making processing more efficient
  - Focus on certain days, airlines, and routes
- Classifying data based on airline or other parameters
  - Different airlines tend to have different delay patterns so grouping them all together might increase the error. Thus, we need to cluster airlines or even departure cities.
- Choosing the best flight among the suggested "similar" flights
  - Different metrics can be used: the closest flight to the time user specified, the flight with zero delay or the lowest chance of being delayed, or high class flight in the "same" airline category

4 Literature Review

There are many previous studies which tries to predict flight delays and cancellation using different set of data such as origin, destination, time, distance, weather, and etc.

In some papers, weather data was been used as the main feature. While in [1] they used a large portion of data (about 130 million of flights) to calculate the results, in another paper [3] they focused on one airport and used the data from the past 2 years to do the prediction. In a similar work [2] , seven major carriers, and fifty highest-traffic airports were under focus. They used Binary Classification, and Probability Estimation to estimate the delay and concluded the latter is a better approach. However, again, there is no guarantee that for a larger subset of data, the same rule applies. In a more recent work [5] , some new features were being incorporated in the model as well as new evaluation techniques. Security and delay of previous flights are among those features.Less error and higher recall has been achieved. It has also been shown [4] that neural networks tend to work better than other methods in this problem since they are able to detect and capture the complex features affecting the end results.

Our goal is to be able to predict with high confidence whether a flight is going to be canceled or delayed, and suggest a number of alternative flights which the user can choose from. First, we limit our scope to certain popular routes and days, and then grow our data set as we see fit.

References

[1] William Castillo Dieterich Lawson. Predicting flight delays. *CS229 Final Report:*, 2012.

[2] Brett Naul. Airline departure delay prediction. *CS229 Final Report*, 2008.

[3] Rafael Guerrero Raj Bandyopadhyay. Predicting airline delays. *CS229 Final Report:*, 2012.

[4] Banavar Sridhar Yao Wang Richard Jehlen, Alexander Klein. Modeling flight delays and cancellations in the us. *Eighth USA/Europe Air Traffic Management Research and Development Seminar (ATM2009)*, 2009.

[5] Jonathan Leaf Romain Sauvestre, Louis Duperier. Modeling flight delays. *CS229 Final Report:*, 2016.

Appendix A: Small Subset of Raw Data

DepTime	CRSDepTime	ArrTime	CRSArrTime	TailNum	ArrDelay	DepDelay	Origin	Dest
636	635	921	945	N454WN	-24	1	ISP	FLL
734	730	958	1020	N712SW	-22	4	ISP	LAS
2107	1945	2334	2230	N798SW	64	82	ISP	MCO
1008	1005	1234	1255	N736SA	-21	3	ISP	MCO
712	710	953	1000	N795SW	-7	2	ISP	MCO
1312	1300	1546	1550	N247WN	-4	12	ISP	MCO
1449	1430	1715	1720	N707SA	-5	19	ISP	MCO
1110	1040	1136	1110	N479WN	26	30	JAX	BNA
1535	1535	1603	1610	N255WN	-7	0	JAX	BNA
1919	1915	1942	1950	N215WN	-8	4	JAX	BNA
1053	1055	1245	1240	N264LV	5	-2	JAX	BWI
1433	1440	1623	1625	N714CB	-2	-7	JAX	BWI
2015	2010	2158	2155	N436WN	3	5	JAX	BWI
2139	2130	2244	2240	N726SW	4	9	JAX	FLL
1500	1500	1602	1615	N399WN	-13	0	JAX	FLL
850	850	1000	1000	N387SW	0	0	JAX	FLL
646	645	752	755	N405WN	-3	1	JAX	FLL
1221	1220	1328	1330	N685SW	-2	1	JAX	FLL
1738	1730	1841	1840	N467WN	1	8	JAX	FLL
1813	1735	1936	1905	N643SW	31	38	JAX	HOU
802	750	1001	955	N263WN	6	12	JAX	IND
1820	1825	1946	1955	N363SW	-9	-5	JAX	ORF
821	820	953	945	N257WN	8	1	JAX	ORF
1734	1650	1941	1905	N521SW	36	44	JAX	PHL
712	700	926	915	N663SW	11	12	JAX	PHL
1318	1310	1410	1400	N376SW	10	8	JAX	TPA
958	900	1052	950	N791SW	62	58	JAX	TPA
1859	1850	1950	1945	N392SW	5	9	JAX	TPA
1538	1445	1753	1710	N799SW	43	53	LAS	ABQ
933	935	1151	1200	N607SW	-9	-2	LAS	ABQ
2248	2125	102	2345	N618WN	77	83	LAS	ABQ
1327	1230	1550	1500	N682SW	50	57	LAS	ABQ
624	625	846	850	N456WN	-4	-1	LAS	ABQ
1614	1600	1833	1825	N509SW	8	14	LAS	ABQ
1917	1915	2136	2140	N293	-4	2	LAS	ABQ
1832	1655	148	30	N473WN	78	97	LAS	ALB
1229	1155	1633	1555	N351SW	38	34	LAS	AMA
1256	1240	1724	1720	N238WN	4	16	LAS	AUS
2118	2015	144	50	N499WN	54	63	LAS	AUS
905	850	1334	1330	N309SW	4	15	LAS	AUS
1739	1640	114	25	N245WN	49	59	LAS	BDL
906	905	1426	1430	N467WN	-4	1	LAS	BHM
816	815	1339	1340	N256WN	-1	1	LAS	BNA
1325	1240	1841	1810	N275WN	31	45	LAS	BNA
1506	1440	2030	2010	N271WN	20	26	LAS	BNA
2039	1930	155	55	N434WN	60	69	LAS	BNA
924	920	1209	1210	N312SW	-1	4	LAS	BOI
1611	1535	1849	1825	N619SW	24	36	LAS	BOI
1824	1715	117	25	N290WN	52	69	LAS	BUF
826	825	930	925	N493WN	5	1	LAS	BUR
2118	2015	2224	2115	N383SW	69	63	LAS	BUR
1818	1740	1916	1840	N608SW	36	38	LAS	BUR
650	650	748	750	N777QC	-2	0	LAS	BUR
2146	2055	2250	2155	N626SW	55	51	LAS	BUR
2241	1910	2340	2010	N369SW	210	211	LAS	BUR
1409	1355	1513	1500	N396SW	13	14	LAS	BUR
1100	1050	1157	1155	N293	2	10	LAS	BUR
1306	1250	1406	1355	N509SW	11	16	LAS	BUR
1726	1630	1832	1740	N409WN	52	56	LAS	BUR

Appendix B: Example User Input and Output

User Input					
Date	Origin	Destination	Carrier	Scheduled Departure Time	Schedule Arrival Time
12/23/08	ABQ	DAL	WN	14:30	17:20

Alternate Flight Suggestions						
Date	Origin	Destination	Carrier	Scheduled Departure Time	Schedule Arrival Time	Expected Delay
12/23/08	ABQ	DAL	WN	16:45	19:23	0
12/23/08	ABQ	DAL	WN	19:55	22:42	0
12/23/08	ABQ	DAL	WN	07:00	09:40	0

Expected to be Canceled?: No
Expected Delay: 36 minutes