FLYDEAL: Investigating Flight Data

Yuyang Bao (ybao1), Kieran Barry (kb25), Albie Brown (ajb7), Taylor DeRosa (tderosa)

INTRODUCTION

This project was inspired by a group of people called “Hobbyists,” who invented the art of airfare hacking in an online community called FlyerTalk. Hobbyists often talk about what they have termed the “mistake fare,” which is how they would describe a flight ticket that has been somehow mispriced by an airline. We started making FlyDeal with the intention of identifying and predicting these so-called mistake fares. Collecting data for this was nearly impossible. Almost every API that included flight price, charged a premium for this valuable piece of information. We were limited to three data sources:

1. Expedia’s Flight Search, an open API built during an internal hackathon
2. Skypicker, a scrapeable flight search engine
3. The Department of Transportation, which releases data about flight delay duration

After collecting our data, we soon realized that mistake fares were fairly random and for the most part unpredictable. We decided to pivot, refining our mission to instead help people learn how to fly. We decided to use this data to identify strategies that would help people book the cheapest flights with the least delays. To accomplish this, we planned to use classifiers on the various other parameters in our dataset to model the algorithm used by airlines to set price in the first place. Through data visualizations, we hoped to identify trends in flight prices that would demonstrate how much each feature would affect prices and delays, and in what way. By applying our new knowledge of data science, we could identify the optimal strategies for choosing flights and begin learning how to fly.

HYPOTHESES

The three features that would most affect flight prices would be:

- Route/Distance/Duration
- Departure Time
- Airline

The three features that would most affect flight delays would be:

- Departure Time
- Date
- Route

DATA

Pricing datasets:
Expedia Flight Search API & SkyPicker API
~160,000 flights
10 major US airports
May 2016 - Feb 2017

Delay dataset:
Department of Transportation
~16 million flights
All US domestic flights
Feb 2013 - Jan 2016

METHODOLOGY

In order to aggregate the delay and pricing datasets, we needed flight matching schema. Initially, we only combined delay and pricing data for flights that were exact matches -- flights with the same route, departure time, and airline. However, even after considering over 16 million past flights, most of the flights that we had pricing data for did not match with any past flights. After trying a number of different matching schemes to combat this issue, we decided to consider any two flights that had the same route and left within a half hour of each other to be a match. Because we were analyzing larger delay patterns (seasonal, weather-related, etc.), it was only necessary to look at route and departure time when cleaning and parsing data. This schema resulted in a significant increase in the amount of integrated data, allowing us to consider average and maximum delay times as features in our machine learning algorithms. We also used this same flight matching schema to aggregate delay information for our delay visualization, as we were still considering grand trends in delay data, and not considering discrete changes in a single day.

CHALLENGES

A majority of the challenges we faced came from collecting and aggregating our datasets. During our initial research, we had a lot of trouble finding pricing datasets. Because there are many commercialized versions of our project (KAYAK, Expedia, etc.) and airlines aim to keep their pricing algorithms private, most pricing datasets costed money or had query limits. We looked into many datasets, including Google's QPX API, KAYAK's API, and OpenFlight's database. Each of these did not fit our needs, and we settled on using Expedia's Flight Search API and SkyPicker's API. These datasets were perfect for our needs, but it took us a bit too long to find them.

Once we had collected all of our data, we realized that the process of integration posed more difficult than we initially believed. If we loaded in more than a third of the delay data, we would fill the disc space on a department machine. In retrospect, we realized that we could have remedied this by requesting more disc space or using larger machines, but realized it a bit too late. Instead, we broke our data integration into two distinct steps. In the first step, we would parse through a portion of the data, clean and integrate it, and output the result to a new csv file. Upon repeating this step, the newly integrated data would be added to the same csv files. Then, in the second step, we parsed through our semi-cleaned data files, and combined them into a final dataset. By doing so, we were able to consider a massive amount of data.

PRICE VISUALIZATION

We thought it would be interesting to visualize how price is affected by our four major features in the dataset: time of day, flight duration, distance traveled and day of week. Various online sources contend that these parameters can be optimized to yield cheapest price. We wanted to see how these claims stood up with our dataset. Day of Week vs. Price clearly illustrates that the best days to fly are Tuesday and Thursday. The worst days to fly are Saturday-Monday, although there are more flights that occur on these days. Percentage On-Time vs. Price shows that you get what you pay for. More expensive flights are less likely to be delayed, but perhaps this is because they are operated by smaller airlines or depart from smaller airports, both of which have limited resources to shuffle things around to ensure an on-time departure. Departure Time vs. Price is fairly noisy, but the general trend seems to suggest that early morning and late night flights are the cheapest, with a slight price increase in the middle of the day. Finally, Duration vs. Price confirms our hypothesis that longer flights mean higher airfares.

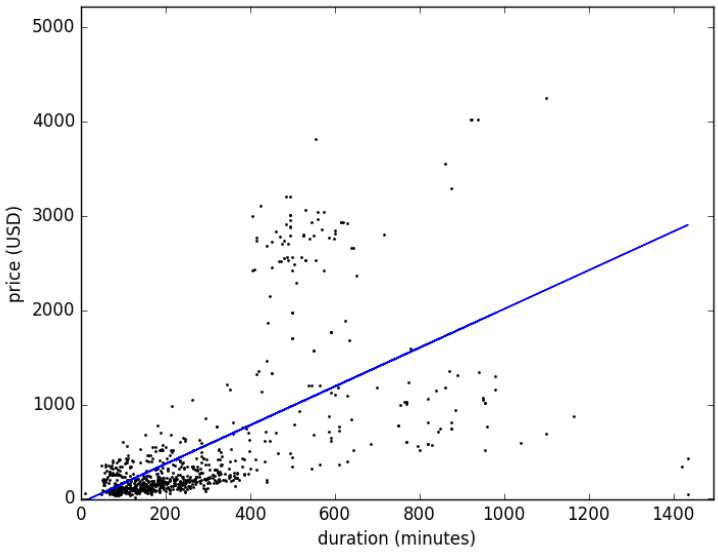
DELAY VISUALIZATION

We wanted to analyze how geography influenced delay patterns for domestic flights in the US. By using a map, we were able to visualize delays for individual routes. Originally, we wanted to create a visualization that featured not only route, but also other dimensions such as airline, airplane model, and weather at origin and destination. Adding these features was a bit too costly performance-wise, but would definitely be a next step for this visualization. Additionally, displaying only route data allowed us to identify cycles and patterns in the data. For example, when viewing flights out of Boston Logan (BOS) over a three year time span, there are clear peaks in delays during winter and early summer. Additionally, there are peaks in delay times during hurricane season in Miami (MIA).

REGRESSION

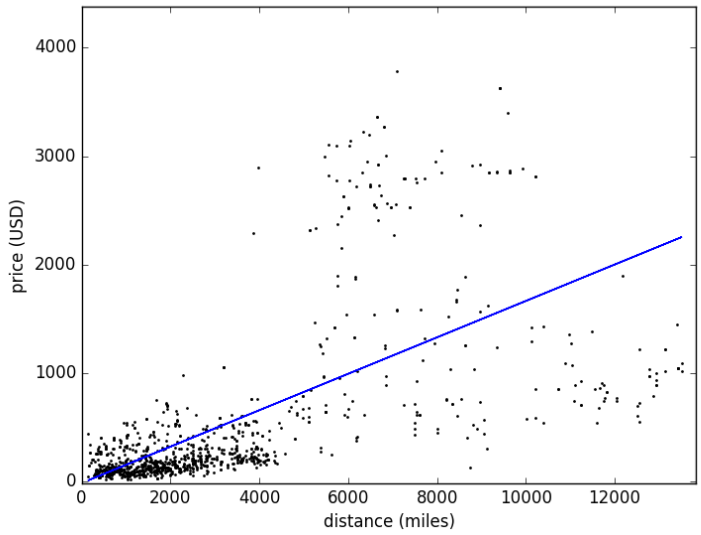
By running a multiple linear regression analysis on the flight dataset, which includes 160,000 flights departing from 10 major US airports in the next 8 months. We thought that ticket price differs at different time of the year, so we first ran linear regression analysis on the week of the year and ticket price. To our surprise we found that flight price was very weakly related to the specific time interval of the year, with a correlation coefficient nearly equal to 0. (The week score corresponds to the “hotness” of the week, which is the price average during that week. The reason for such weak correlation is mostly because what actually affect the ticket price the most are not regular tickets but those outlier tickets, which means the normal tickets across the year is roughly the same.

We then ran linear regression analyses on two single features (distance and duration) and found that there was a moderate correlation between both features and price; both returned a R^2 score of around 0.5. We also found that looking at the changes in flight price by week was the most reasonable according to various online forum discussions and resources.



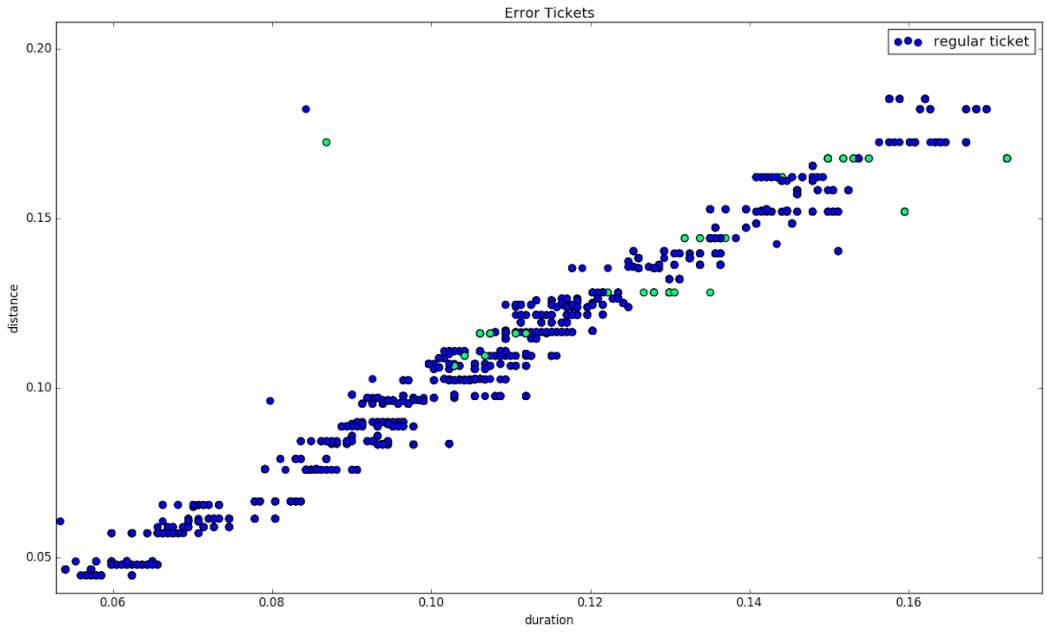
Price v duration regression.

Price v distance regression.



Bonus: We first used the linear regression model we fit to predict the expected prices of the flights. Then we defined an error ticket as a ticket priced less than 30% of the expected price. According to this criterion, we categorized all the flights as either a normal ticket or an error ticket. Next we fit a logistic regression model with the categorized data to predict whether or not a flight ticket is an error ticket based on the distance, duration and the time of the year. The logistic regression model achieved around 93% accuracy.

We then plot the error tickets on a distance and duration graph, and found out that the existence of error tickets actually does not have a correlation with distance and duration. (Green dots being the error tickets)



CONCLUSIONS

Through our classifications and visualizations, we were able to identify a number of trends to help people fly smarter. To optimize for your flight preferences, we make the following recommendations. For the cheapest flights, fly early in the morning or late at night on Tuesday or Thursday and avoid Saturday-Monday, when possible. To decrease the likelihood and duration of a potential delay, we recommend flying in the spring and fall, avoiding the East Coast in the winter, and trying not to fly too much in the month of December, when delays are longest. Remember, you get what you pay for, so a more expensive ticket with a big carrier is much more likely to lead to an on-time flight. We have already begun using this new information in our daily lives as we book flights home at the semester's end.