Avoiding Flight Delays with Supervised Ranking

Dan Garant John Lalor Adam Nelson

Abstract

In this work, we examine the task of selecting flights to minimize the risk of encountering a flight delay. Taking the perspective of a passenger attempting to book a flight one day in advance or thirty days in advance, we gather and explore a number of weather, airport, and air carrier characteristics that can be used to assess the probability of experiencing a delay. These features vary in availability and quality across the thirty-day and one-day horizons, yielding two distinct machine learning tasks. We evaluate a number of supervised learning models for each task, and present results in the form of rankings, consistent with a strategy that could be employed by a flight-booking service.

1 Introduction

In recent years, the U.S. Bureau of Transportation Statistics (BTS) has enabled easy access to a wealth of information on on-time flight performance of domestic air carriers[1]. Given the availability of flight characteristics, the task of determining whether a flight will take off and arrive on time is well-suited for analysis with machine learning. Air transportation has become a central aspect of the U.S. culture for both business and recreation, with the number of domestic flights offerings growing steadily for decades. Accompanying the growth in air travel, numerous online services including Google Flights and Expedia provide passengers the ability to select from a variety of available air carriers, flight dates, and origins when booking a flight. However, these services provide only high-level historic information about patterns in delays for specific scheduled flights, if at all.

Combining historical characteristics of air carriers and airports with weather forecasts and weather trends (or *normals*) would enable flight booking sites to provide more reliable and targeted information about delay risk. Our objective is to use supervised learning models to identify the flights with a low probability of delay. Our analysis will look at flight delay patterns according to two distinct problems: predicting a flight delay for flights leaving one day after booking, and predicting flight delays for flights leaving in thirty days. In section 2, we give an overview of the BTS dataset, previous studies or airline delays using this dataset, and supplemental weather data sources we employ. In section 4, we examine patterns in delays that can be exploited to construct a probabilistic model of delays. Section 5 formally details our method of evaluation, and section 6 explores the results obtained by six supervised learning models of various levels of sophistication.

2 Data Sources

2.1 Bureau of Transportation Dataset

The flight data used in this project was first studied through a visualization challenge issued by the 2009 American Statistical Association Joint Statistical Meetings Data Expo [Wickham, 2011b]. The challenge was to create a graphical summary of flight delays using flight statistics from October 1987 to April 2008. The challenge submissions for this explored a variety of aspects of the data, a number of which are mentioned below.

The data for the challenge was supplied by the U.S. Bureau of Transportation Statistics (BTS), and contains flight statistics for domestic flights. For the

[1] http://www.transtats.bts.gov/

purposes of our project, we focused on flight statistics from the year 2014, as these flights were more recent than the original challenge dataset and still provide a large set of data from which we could obtain conclusions. We obtained the data directly from the BTS website.

The BTS data includes a large number of data fields. Below we enumerate the fields that we included in our analysis, along with a description of each:

- Origin Airport and Departure Airports: Origin and departure airports for each flight in the dataset, encoded as a discrete characteristic with one-hot encoding. 250 airports were represented in this dataset.
- Month, Day of Month, and Day of Week: Included since traffic patterns vary over the course of the year—specific periods could be tied to flight delays. Month and day of week were encoded as discrete characteristics.
- Carrier: Discrete characteristic representing the carrier of the flight. 14 distinct carriers were represented in the data.
- Scheduled Departure and Arrival Time: Scheduled arrival and departure times for each flight, as provided by the Computer Reservation System (CRS) used by travel agents to allow agents to book flights and issue tickets.
- Distance: Distance from the origin to the destination
- Airport volume in and out: For each flight, we calculated the number of other flights scheduled to arrive and depart within the same hour that the reference flight was scheduled to depart.

2.2 Supplemental Data

In addition to the data provided by the BTS, we augmented the feature set for flights with weather forecasts and patterns at each airports to determine the effect that weather might have on flight delays. For our two tasks, this involved obtaining two new sets of features: detailed weather information for the one-day task, and high level weather normals for the one-month task.

When redicting flight delays for tomorrow's flights, detailed weather information will be available for typically very accurate. One month in advance, however, the weather information is less likely to be accurate, so we instead looked at ten-year normals for the flight dates, in order to get an idea of weather one month in advance. Two mirror our two task scenarios, we only included the detailed flight information in our one-day feature set, and only included the high-level weather normals in our thirty-day task feature set.

For the one-day task, a wealth of relatively accurate and detailed weather information is available. We gathered historical weather data for a about 50% of 2014 flights represented in the BTS data using a popular weather API[2]. This API provided 12 features:

- Precipitation intensity, measured in inches per hour
- Recent precipitation intensity, a derived feature, which we computed as the mean of precipitation intensity values for the previous 3 hours
- Dew point
- Visibility in miles
- Temperature
- Apparent temperature, representing a "feels like" temperature
- Pressure in millibars
- Precipitation probability, when actual precipitation data is not available but a forecast is
- Humidity
- Wind speed in miles per hour
- Wind bearing, ranging from 0 to 359 with 0 indicating that the wind is blowing true north
- Weather summary, a concise summary of the weather which we represented as a discrete characteristic

For the one-month task, quality weather forecasts are not available. Historical weather data would be inappropriate to include, so instead we collected ten-year climate normals for the flight dates from the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI) [NCEI]. For our task, we looked at the temperature normals collected over a period of years, from 1981 to 2010. Specifically, we used the

[2] `https://developer.forecast.io/`

high, low, and average temperature for both origin and destination airports.

3 Related Work on BTS Dataset

As part of the 2009 challenge, there were a number of submissions that used the dataset for interesting visualization tasks. We will briefly discuss the top four submissions for that task here. Wicklin and Allison [2009] used SAS Software to provide a number of visualizations of the data, including departure delay heatmaps by month for each of the years included in the data and historical comparisons of flight delays, showing that the number of delays varied widely between years. Hofmann *et al.* [2011] found that Newark airport was the worst offender when it came to flight delays, which one author can back up empirically. Wickham [2011a] looked at two specific airports, Oakland and San Francisco, and compared number of flights between the two airports, looking at trends in flight numbers in the two airports. Dey *et al.* [2009] modeled the data as a graph in order to predict delays for any origin-destination pair.

In addition to work performed as part of the 2009 challenge, there has been additional work using the BTS data. Rupp *et al.* [2006] investigated whether competition for specific routes influences airline performance and found that *busy* routes (that is, routes with many carriers) tend to have more delays. Stone [2015] found that small community airports suffering from departure delays had a large effect on those passengers, causing missed connecting flights and incurring large re-booking costs.

4 Exploratory Analysis

To inform our selection and tuning of supervised learning algorithms we performed a number of exploratory analyses characterizing the nature of dependence between specific features and the response (departure delay in minutes). First, we investigated characteristics of organizations involved in fulfilling the flight. Using a linear model, we estimated the conditional mean and variance of the departure delays (in minutes) given an air carrier. We find significant differences between carriers, shown in Figure 1. Southwest Airlines is the biggest offender, with an expected delay duration of more than 15 minutes. At the other end of the spectrum, Hawaiian Airlines has a negative expectation, indicating that most flights leave before scheduled.

To gain a high-level understanding of the relationship between the weather features and delays, we fit a generalized additive model [Hastie and Tibshirani, 1990] of delay duration (in minutes) to a linear combination of cubic spline transformations of various weather features. Naturally, there are associations between each of the weather characteristics. For instance, humidity will necessarily be high when it is raining. To isolate the *independent* effects of each of the features on the departure delay, we predicted the conditional expectation of delay D with respect to an individual feature X_i, holding other features at their mean value:

$$E[D|X_1 = \bar{X_1}, X_2 = \bar{X_2}, \ldots, X_i = x_i, \ldots, X_n = \bar{X_n}]. \quad (1)$$

Then, sweeping across various settings for x_i and various feature X_i yields a sensible explanatory model of delay with respect to weather characteristics. When stratifying by airport and evaluating (1) with X_i =wind bearing, we find a strong non-linear relationship illustrated in Figure 2. This dependence is likely the result of differing runway orientation at various airports. When wind is blowing *across* a runway, take-off and landing is more challenging, and some pilots may opt to defer until winds are more favorable. Each airport may have a different runway orientation, so an accurate model must be capable of capturing interactions between a discrete characteristic (airport) and a real-valued characteristic (wind direction). We also find intuitive dependencies between precipitation intensity, shown in Figure 3a, and wind speed, shown in Figure 3b.

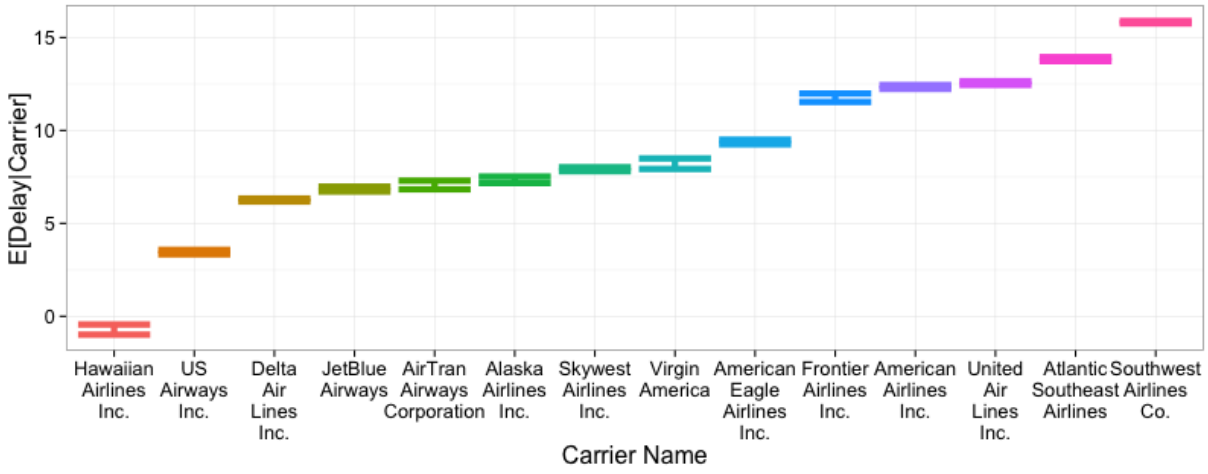
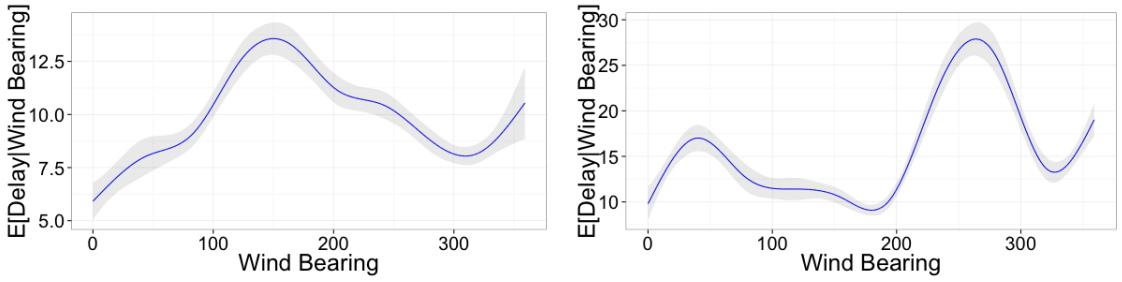


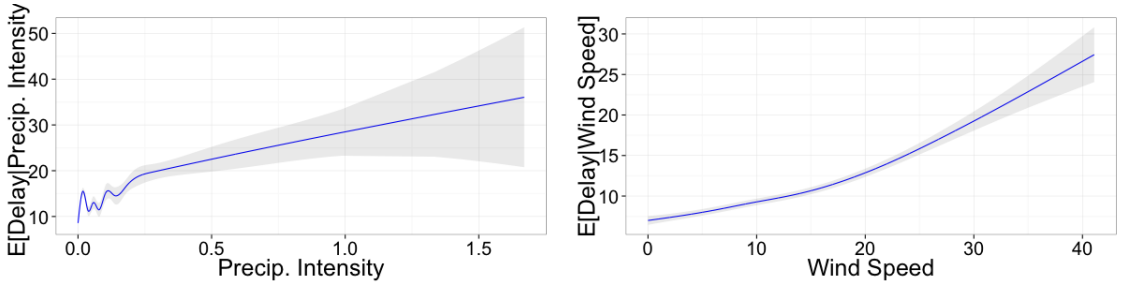
Figure 1: Performance of Major Domestic Air Carriers



(a) Hartsfield-Jackson Atlanta

(b) Dallas/Fort Worth

Figure 2: Relationship between wind bearing and expected delays (in minutes) at various airports. One standard error in the conditional expectation is shown in gray.



(a) Relationship between precipitation and delay

(b) Relationship between wind speed and delay

5 Evaluation Methodology

We evaluated according to two related but distinct tasks: delay classification and risk ranking. In both cases, the training data consists of set of 3,500 flights randomly sampled from all 2014 flights for which weather data was available. Each flight is annotated with a binary delay class, "delay" if the flight left 1 or more minute after the scheduled departure time, or "no delay" otherwise. For the delay classification task, learning algorithms are evaluated based on how well they can predict this binary class label. From a distinct pool of 1,500 test flights, we compute the classification accuracy as the proportion of correctly labeled instances.

The risk ranking task is more closely related to the intended passenger-centric use case. For this task, the test set consists of 750 *pairs* of flights, where each pair has the same origin, destination, and flight date. For each flight in the pair, the delay probability, $P(D = \text{delay}|x_1, x_2, x_3, \ldots, x_n)$, is estimated rather than the delay class. The flight with the smallest delay probability is marked as "favorable". Then, the ranking accuracy is computed as the proportion of pairs for which (1) the pair has differing delay status, and (2) the favorable flight experienced a shorter delay.

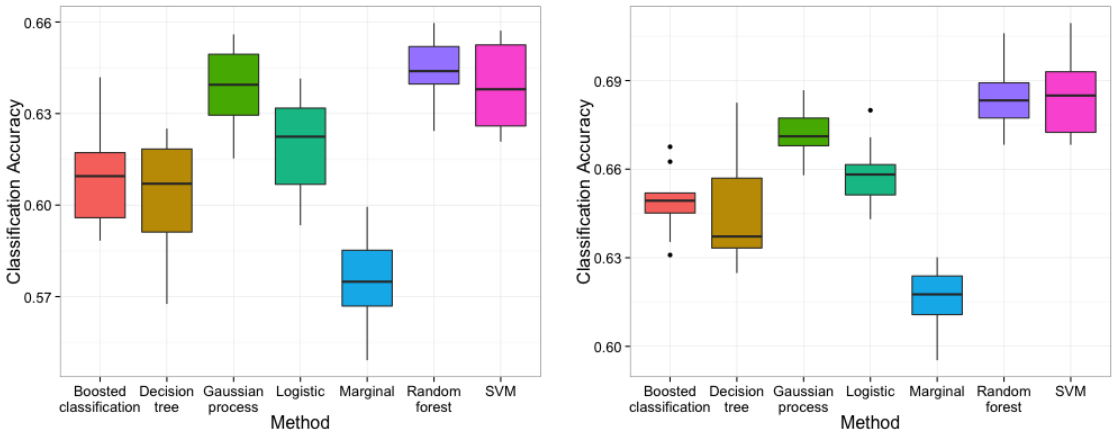
We considered machine learning algorithms from three categories:

- Simple/baseline methods: These included methods that predict from marginal and uniformly random class probabilities. The marginal baseline represents a "null" model for the delay classification task. It always predicts the most common class. The uniformly random baseline represents a null model for the ranking task, as it randomly marks one of the flights as favorable. We also considered classical logistic regression.
- Tree-based methods: This category includes decision trees, random forests, and boosted classification trees. We tuned the complexity penalty of decision trees with linear search. For random forests, we set set the number of trees with cross-validation, and for boosted classification trees, we set the number of boosting iterations with cross-validation.
- Kernel-based methods: In this group, we used support vector machines and Gaussian processes. In each case, we used a radial basis function kernel with bandwidth hyper-parameter set with linear search.

6 Results

Figure 4 shows the performance of classification results for both the one-day and thirty-day tasks. Several properties are apparent. First, all of the methods are able to generally out-perform the marginal baseline. In both cases, support vector machines and random forests are most effective. With the exception of random forests, tree-based methods perform rather poorly, perhaps owing to the relatively high cardinality of discrete features such as airport and carrier. Interestingly, performance in the one-day task is comparable to that in the one-month task. This suggests that weather information is relatively uninformative in predicting delays. Upon inspecting the distribution of weather characteristics during the hour that all 2014 flights were scheduled to depart, we found that 91% of instances are categorized as clear, partly cloudy, mostly cloudy, or overcast. Even while the ratio of clear to rainy days in the U.S. is approximately 3:1, the number of *hours* without precipitation to those with precipitation is approximately 20:1. Thus, while there is a strong relationship between various forms of inclement weather, as shown in section 4, opportunities to exploit this information are fairly limited.

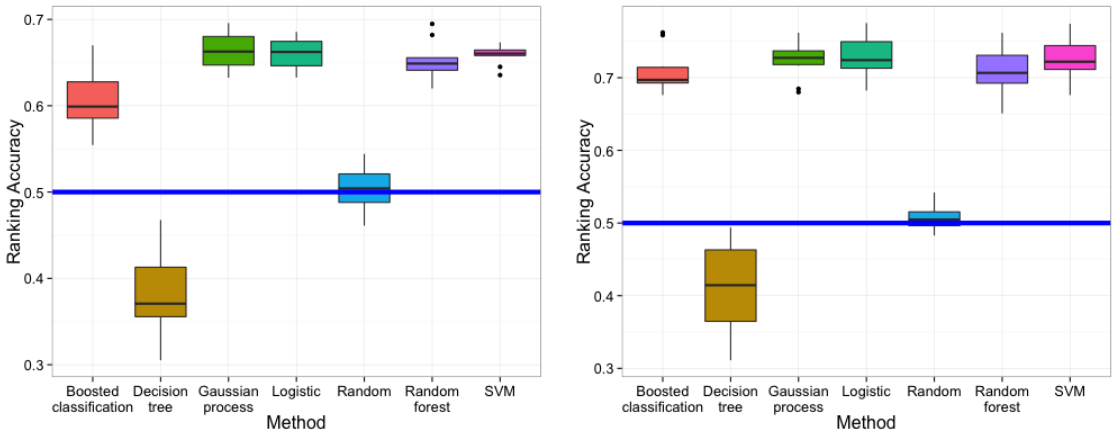
Figure 5 shows the results for the ranking task. In this case, the performance of most methods is fairly comparable, typically 10-20% more accurate than random. Again, the additional weather information available in the one-day horizon task is not an important factor in achieving a reasonable level of performance, suggesting that the thirty-day horizon task is equally as feasible as the one-day horizon task.



(a) One-day horizon

(b) Thirty-day horizon

Figure 4: Classification Accuracy



(a) One-day horizon

(b) Thirty-day horizon

Figure 5: Ranking Accuracy

7 Conclusion

In this project we attempted to answer the following question: Can a machine learning model be used to identify flights with the lowest probability of delay? Our motivation was from the point of view of an airline passenger that seeks to book a flight with the lowest chance of delay. We considered different flight-booking scenarios: booking a flight tomorrow and booking a flight in a month's time. For these two tasks we combined flight data with relevant weather data, and found that a number of machine learning methods were able to outperform random baselines. Further, we found that accuracy in both the ranking and classification tasks were comparable for the one-day and one-month horizons, suggesting that quality predictions are possible long before the flight is scheduled to depart, and without detailed weather information.

A key lesson we learned through this work is the importance of structuring an evaluation around the intended use-case. Our initial evaluation revealed mediocre classification performance in this task. However, since our ultimate objective is to provide flight recommendations, we explored re-framing the task as a ranking problem. Given this context, we use conditional densities rather than class labels, allowing us to assign a clear ranking to any pair of flights, even if they have the same estimated class label.

Future work could include looking at more complex patterns in the flight data. This could include expanding our airport volume feature into numerous volume-oriented features for specific adjacent airports. Since flights typically operate on regular schedules, it may be possible to include historic delay characteristics for a specific flight in question. It would also be interesting to expand this task to international travel, in which multi-stop flights are common and passengers have a potentially wide array of flight choices.

References

Tanujit Dey, David Phillips, and Patrick Steele. Minimizing flight delay. Technical report, Tech. rep., Dept. of Math, College of William & Mary, contact authors for manuscript, 2009.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

Heike Hofmann, Dianne Cook, Chris Kielion, Barret Schloerke, Jon Hobbs, Adam Loy, Lawrence Mosley, David Rockoff, Yuanyuan Huang, Danielle Wrolstad, et al. Delayed, canceled, on time, boarding flying in the usa. *Journal of Computational and Graphical Statistics*, 20(2):287–290, 2011.

NCEI. Climate data online, ncei web services. `http://http://www.ncdc.noaa.gov/cdo-web/webservices/v2`.

Nicholas Rupp, Doug Owens, L Plumly, et al. Does competition influence airline on-time performance. *Advances in airline economics*, 1:251–272, 2006.

Matthew J Stone. Investigating the effect of flight delays and cancellations on travel from small communities. 2015.

Charlotte Wickham. A tale of two airports: Exploring flight traffic at sfo and oak. *Journal of Computational and Graphical Statistics*, 20(2):291–293, 2011.

Hadley Wickham. Asa 2009 data expo. *Journal of Computational and Graphical Statistics*, 20(2), 2011.

Rick Wicklin and Robert Allison. Congestion in the sky: Visualising domestic airline traffic with sas. *ASA Statistical Computing and Graphics Data Expo*, 2009:14, 2009.