Avoiding Flight Delays with Supervised Ranking

Dan Garant John Lalor Adam Nelson

Abstract

With the amount of data available on flights in the United States, one interesting task is to use this data in order to make predictions about flight performance. However, being able to find the correct features to use from the variety of available features is important when trying to build a model of flight delays. In this project, we look at predicting flight delays from the point of view of the passengers on the flight. We look at the task of predicting flight delays from two points of view: a customer looking for a flight that departs in one month, and a customer looking for a flight that departs tomorrow. The two tasks offer different features in terms of available weather data. We evaluate a number of models for each task, and present results in the form of rankings, where the flight with the lowest chance of having a delay is presented as the top result.

1 Introduction

Being able to determine whether a flight will take off and arrive on time is a question that is well-suited for analysis with machine learning. Air transportation is something that many people deal with regularly, and a huge amount of data is recorded by the U.S. Government for all of the flights that operate in the U.S. Combine that with the data available about weather patterns and weather history and you have a lot of information about the conditions surrounding air travel. For this project, we will explore this data, and attempt to answer the following question: Can we use machine learning to avoid flight delays when traveling by plane? In this report, we will give an overview of the dataset, and briefly describe some of the work done before using it as a resource, we will look at the data from a high level and look at trends and patterns that arise, and we will apply a number of supervised learning methods to the data to attempt to rank flights based on their likelihood of having a delay. The objective is to use the models to identify the flights that are least likely to be delayed. Our analysis will look at flight delay patterns according to two distinct problems: predicting a flight delay for flights leaving tomorrow, and predicting flight delays for flights leaving in one month (e.g. in thirty days).

2 BTS Dataset

2.1 Dataset

The original idea for the project came from the 2009 American Statistical Association Joint Statistical Meetings Data Expo [cite]. The challenge was to create a graphical summary of the data, which consisted of flight statistics from October 1987 to April 2008. The submissions for this explored a variety of aspects of the data. (Need to elaborate here).

The data for the expo came from the United States Department of Transportation (DOT). The DOT Bureau of Transportation Statistics (BTS) track flight statistics for domestic flights. For the purposes of our project, we focused on flight statistics from the year 2014, as these flights were more recent than the original expo 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 Airport IDs: Unique IDs used to identify the origin and departure airports for each flight in the dataset.
- Departure Date: Date of flight.
- Month, Day of Month, and Day of Week: Additional date fields at varying levels of detail. These were included in an attempt to determine if there were specific days of the week or other specific periods that could be tied to flight delays.
- Carrier ID: Unique ID for the airline of the flight.
- Scheduled Departure and Arrival Time: Scheduled arrival and departure times for each flight, as provided by the Computer Reservation System (CRS)
- Distance: Distance of the flight

2.2 Related Work

As part of the 2009 task, there were a number of submissions that used the dataset in interesting ways. We will briefly discuss the top four submissions for that task here. [4] 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. [2] found that Newark airport was the worst offender when it came to flight delays, which one author can back up empirically. [3] 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. [1] modeled the data as a graph in order to predict delays for any origin-destination pair.

2.3 Supplemental Data

In addition to the data provided by the BTS, we wanted to augment the feature set for flights by looking at weather patterns for each of the 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. The intuition here was that if you are looking to predict flight delays for tomorrow's flights, a lot of detailed weather information will be available for tomorrow that is most likely 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.

3 Exploratory Analysis

4 Evaluation Methodology

5 Results

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.