Avoiding Flight Delays with Supervised Ranking

Dan Garant John Lalor Adam Nelson

Abstract

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.

2.2 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