Predicting Flight Delays Using Local Weather Data

Samir Menon [samir2] and Neil Movva [nmovva]

CS229 Final Project
2016-12-13

Introduction

The total cost of flight delays in the US is $34bn annually. This cost is shouldered by both the airline industry ($8.3bn) and even more so, passengers ($16.7bn) [1]. In 2008, weather accounted for 45% of all delay minutes [2].

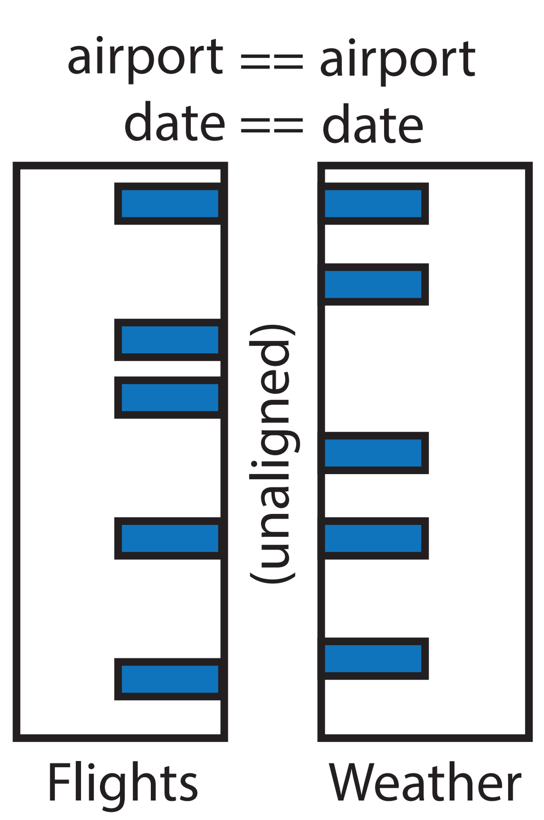
We want to predict flight delays using localized weather data. Since modern meteorological predictions are quite accurate, a good model could offer passengers and carriers more time adjust plans, bringing significant cost savings for everyone.

Data

- Bureau of Transport Statistics provides data on all domestic flights in 2008 (7M)
- Filtered to top 11 airports (2.3M), and we think larger airports follow more consistent protocol
- NOAA has hourly meteorological data at these airports for all of 2008.
- Dropped flights that were cancelled / diverted (<1% of flights in set)
- Cut out columns that were known after takeoff (e.g. ArrivalTime)

- Large, optimized SQL join ties each flight to weather at its departure.

- To force an equi-join, we simplified the data by only including one weather sample per station per hour.



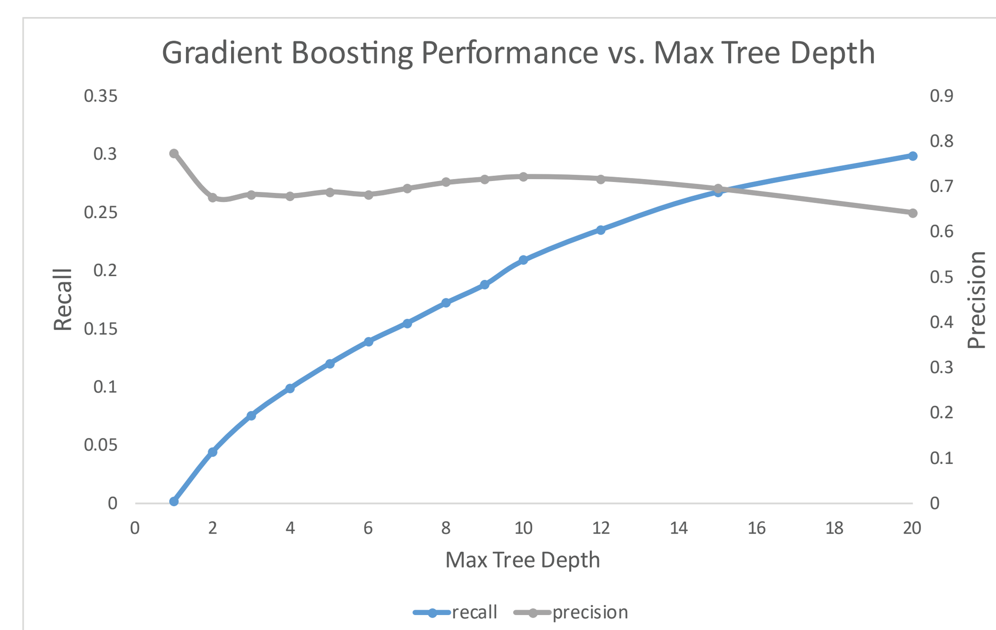
[1] US Bureau of Transportation Statistics, Office of Research and Technology, "Understanding the Reporting of Causes of Flight Delays and Cancellations"
[2] RITA says that extreme weather accounts for 4.1%, and NAS accounts for 24.8% (note that "53.1 percent of NAS delays were due to weather"), so we get 17.2% of total flights delayed.
[3] Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., ... & Britto, R. (2010, January). Total delay impact study. NEXTOR Research Symposium, DC.
[4] UC Berkeley, Hansen, M., "Flight delays cost $32.9 billion, passengers foot half the bill"

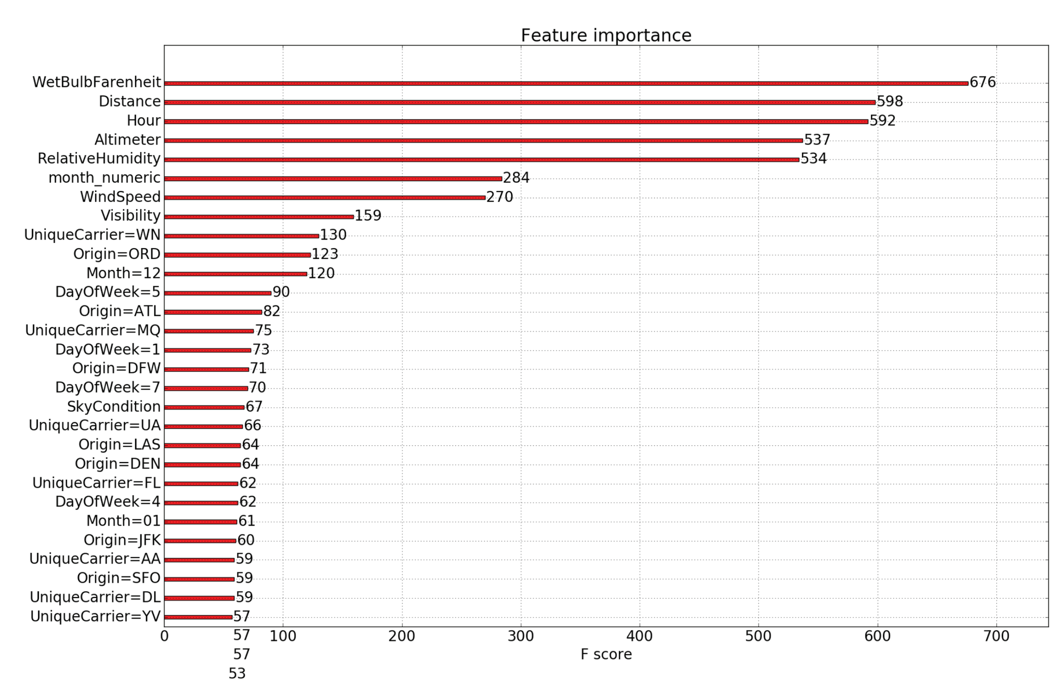
Results

We tried a variety of models, but found that gradient boosting (xgb) performed best. Looking at the decision trees produced, we find a nice semantic mapping to the actual process by which flights are delayed - airlines make a series of choices, some of which are regulated by the FAA for safety reasons, and then ultimately make a delay decision.

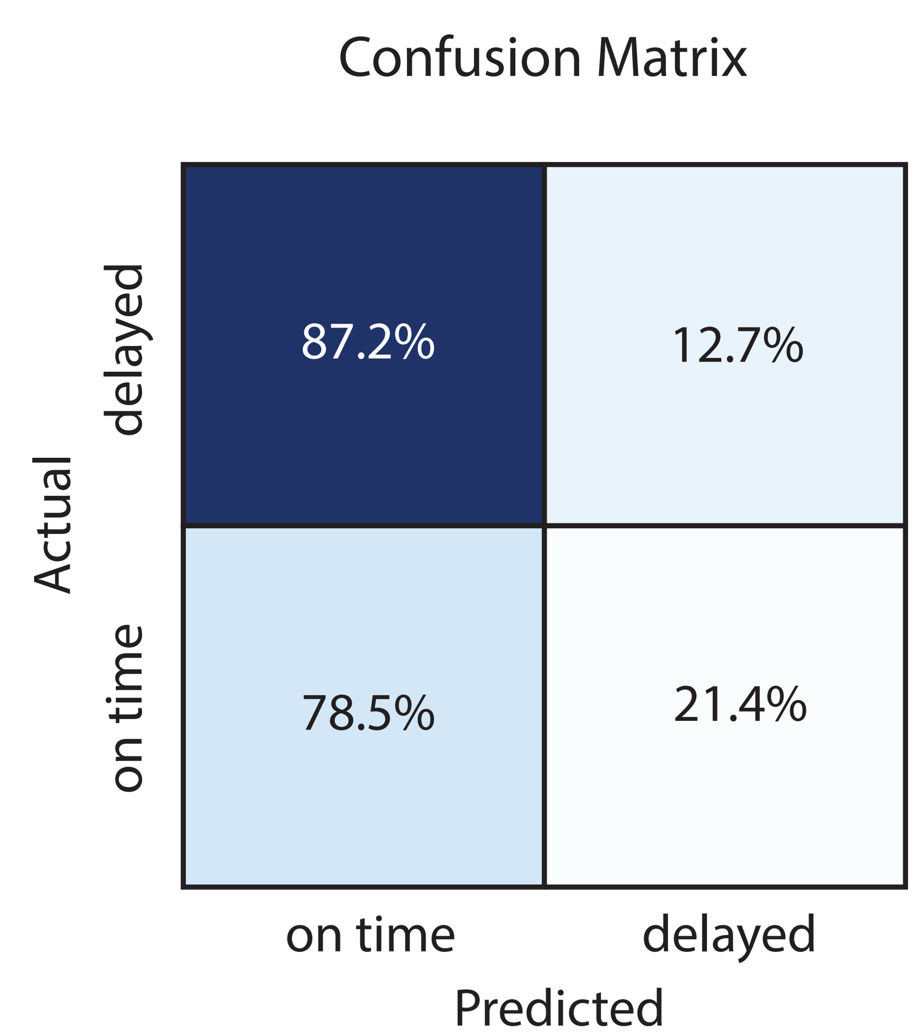
When tuning our gradient boosting algorithm (through XGBoost), we find that increasing the max depth of our decision trees improves performance significantly without incurring too much computational cost. Intuitively, this makes sense given the high dimensionality of our data.

numerical	Visibility Temperature Relative Humidity WindSpeed HourlyPrecip Altimeter Distance Hour Month
categorical	SkyCondition DayOfWeek WeatherType Carrier Origin Month
binary label	DepDelay





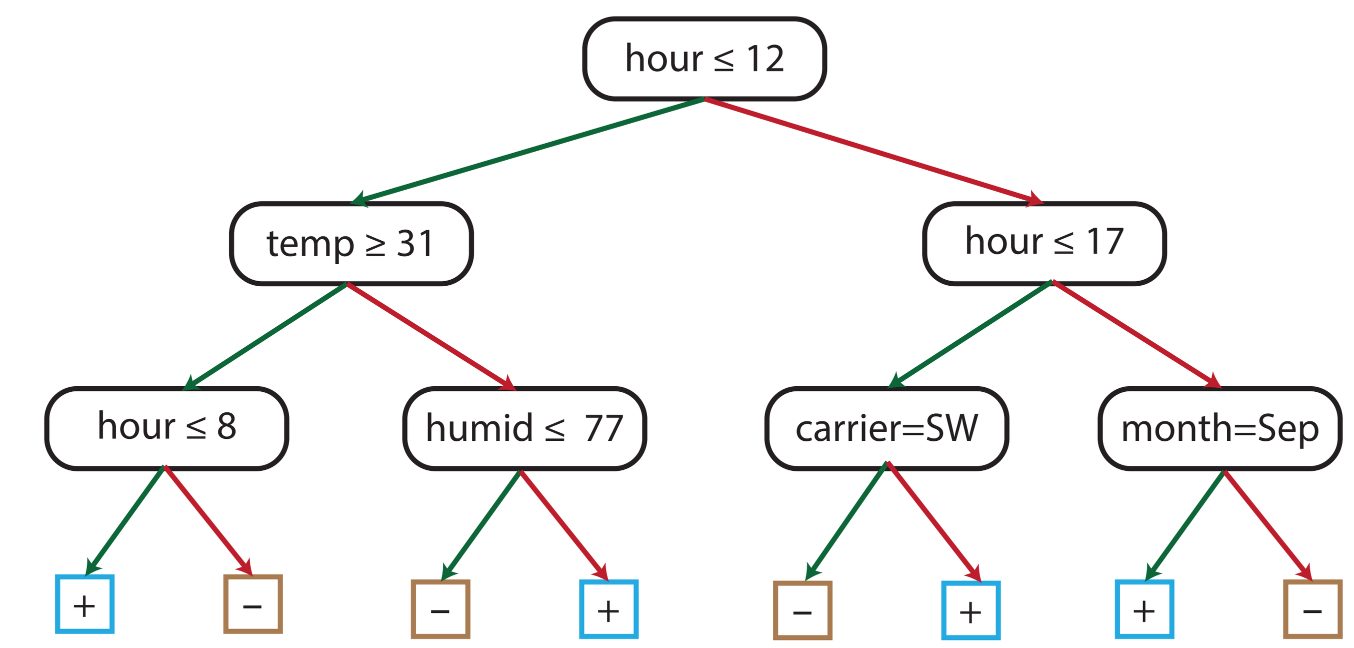
Feature importance measures the relative impact of the features we considered. We find that the model's most powerful features generally follow intuition, giving us confidence that the model has learned a reasonable decision process.



Discussion

Our model's recall is 15.5%, and precision is 70%. The recall may seem rather poor, but note that the majority of flight delays are caused by non-weather events - e.g maintenance, crew, etc. Our model cannot be expected to account for most of these, as it lacks the related feature data. Based on data from BTS, we know that ~17% of flight delays are caused by weather, so our 15.5% recall is relatively performant in context.

Unfortunately, the BTS reporting system does not identify non-extreme weather as the cause of a flight delay - this may be improved soon.



Looking at a sample decision tree, we notice a decision boundary on airtemp < 31F, which is very near the freezing point of water. Interestingly, the model has learned this physical relationship, which we can understand since precipitation / humidity is much more dangerous below the freezing point.

Future Work

We used entirely our own computers to preprocess, train, and test. With more time, we'd set up a cluster and run larger models on data from 1978-2015. Our hardware limited our options when joining the weather and flight data, and made higher resolution analysis infeasible. With more compute, we'd especially to experiment with some exponential decay models on weather prior to the flight's departure, as we had originally intended.