Forecasting Flight Delay and the Impact of Weather

Wessam Kaddoura wkaddoura3
Eric Park epark313
Alex Demory ademory3
Leah Schattenfield lss30
Andrew McAlister amcalister7
Sarah Paustian sfarmer36

Summary

Most online flight delay and forecasting tools are limited in scope or overly simplistic, especially when it comes to accounting for weather changes. We have developed a predictive model along with an intuitive dashboard for users to understand the weather factors causing flight delays at specific airports. Such a tool can also provide the groundwork for further analysis into individual flight data.

Current Industry Practices

Past attempts at synthesizing flight delay and weather data have typically focused on how a single factor affects arrival/departure times of specific flights, which limits the scope of analysis and can lead to inaccurate predictions. Additionally, the few products available for a retail audience present an oversimplified view, leaving much to be desired in terms of probability details.

Our Approach

Our model predicts the **delay rate for 10 of the busiest US airports, up to 15 days in advance**, an improvement on the typical 3-day lead that other services provide. To provide this insight, we trained 4 separate machine learning models on historical flight and weather data, focusing on 8 weather and 3 flight-related features. With this information, one can then calculate a “daily delay rate” for each airport as the percentage of total flights classified as delayed. This approach allows us to build a **user-friendly airport-level view of delay probability and differs from past “flight-centric” methods.**

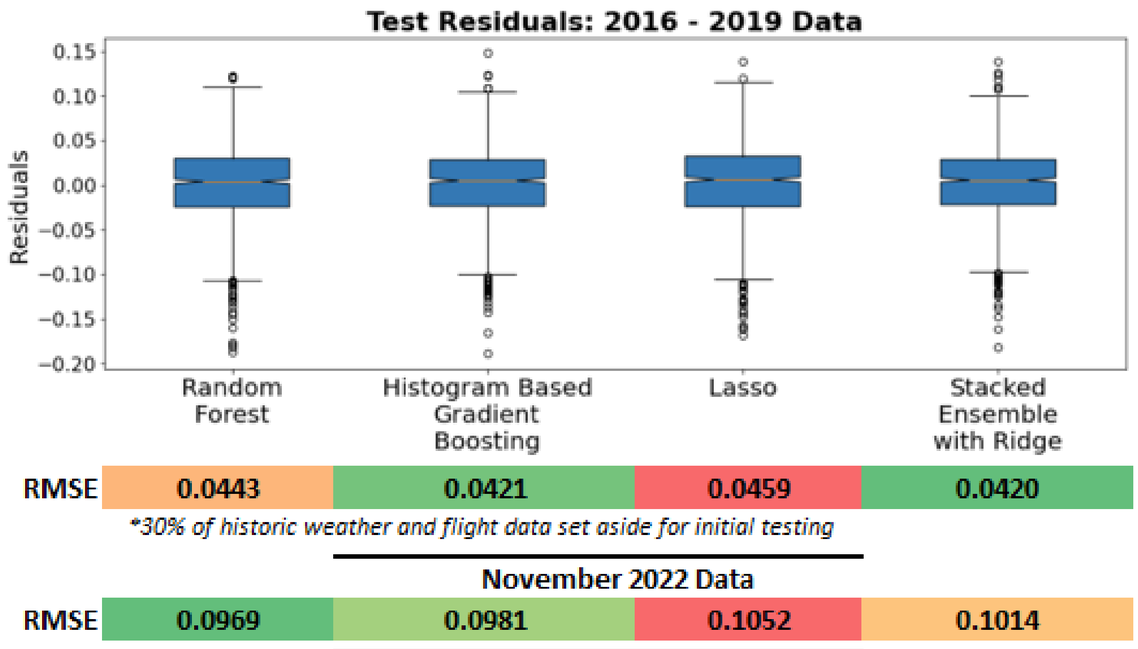
Our dashboard hosts **multiple views of the forecasted delay probability and allows users to customize weather inputs**, if desired. This is a significant improvement versus existing product offerings, which are more generalized and have limited information around the weather factors.

Data

	Historic Weather	Forecast Weather	Historic Flight
Source	Iowa Environmental Mesonet (IEM)	Weatherbit.io	Bureau of Transportation Statistics
Size on Disc	24.7 MB	N/A	880 MB
# of Records	349,937	16 per API call	8,420,999
Time Range	2016 – 2019	Current + 15 days	2016 – 2019
Rate	Hourly	Daily	Not Temporal (flight level)
Retrieval Method	File Download (from web)	API call	File Download via R anyflights package

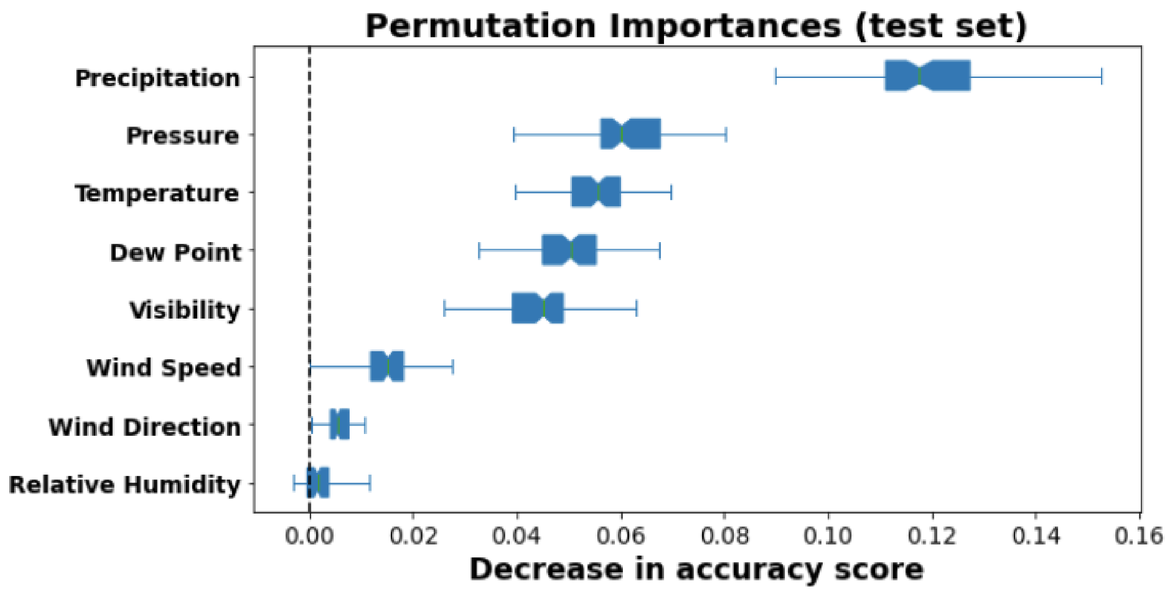
All data sources include records for the following 10 airports: ATL, LAX, ORD, DFW, DEN, JFK, SFO, LAS, SEA, CLT

Experiment #1



Prior to model selection we conducted further testing on daily forecast data through the month of November to assess the impact of forecast error on prediction accuracy. The **Histogram Based Gradient Boosting Model** is selected for use in the final product due to strong performance on both historic and current data.

Experiment #2



***Based on 100 permutations calculated on the test set using the tuned Histogram Based Gradient Boosting Model**

Permutation feature importances calculated for the eight weather features included in the model indicate that **precipitation** is the strongest predictor of flight delays.