Flight delay prediction and exploration in the United States

Team 127 - Aviators:
Aaron Joseph Mathew, Ayush Subedi, Harshini Prabakaran, Hira Laghari, Saumya Shekhar

Summary

Flight delays can have a vast economic and emotional impact, due to which we have devised several ***predictive models,***, including **Random Forests** and **CatBoost** algorithms to find out whether a certain flight has a high chance of delay, based on factors such as airline, location, route and days of the year. Our models have accurately predicted the prevalence of delays by up to 75%. Using these models, we have created a tool which passengers can use to determine whether a flight they are interested in has a chance of delay.

Causes & Impact of Delays

Flight delays can be caused by a variety of factors, with ground and air congestion being one of the main reasons, as well as adverse weather and logistical factors (ie security delays, cargo delays, baggage delays, etc). The economic impact of flight delays is very large, with airlines losing USD 29 billion to delays in the United States in a single year. Airlines with a high tendency of delays also have negative sentiments associated with them by passengers.

Delays Dataset

The dataset we used was collected by the Bureau of Transportation Statistics. It consists of more than **200 million rows** and **109 columns**, spanning the years 1988 to 2021. We used data from 2014 to 2021, which consists of about **50 million rows**, and is 6GB in size. About 20% of the records have delays. Some of the columns we used to predict delays include IATA Airline Codes, Origin & Destination Airport IDs, number of minutes of delay caused by various factors (ie security, weather, late carrier, etc) and the number of minutes of arrival and departure delays.

Modeling & Feature Selection

We analyzed several papers that modeled the prediction of flight delays, and observed the best performing models to be using Random Forests models. We also tested several boosting models, including CatBoost, a model that provides gradient boosting for categorical features. In order to improve accuracy with our models, we also chose to use ANOVA techniques and LASSO Regression to select the best features, narrowing it down to the following: *Month, Day of Month, Day of Week, IATA Airline Codes, Origin Airport ID, Destination Airport ID, Airtime / Distance*

Models Evaluation

Some of the best performing models we tested include Random Forests, Boosting Algorithms (CatBoost, XGBoost, etc) and Neural Nets. The bagging and boosting models were developed and run locally using Pandas and SciKit-Learn, whereas the Neural Net model was developed on BigQuery ML. The data was prepared by first feature-engineering a binary predictor column Delay (delay >15 minutes), limiting the data to only use selected features (with some variation with CatBoost & Neural Net models), one-hot encoding to deal with categorical data, and then using SMOTE to deal with class imbalances. The accuracies of some of the models tested are reported below:

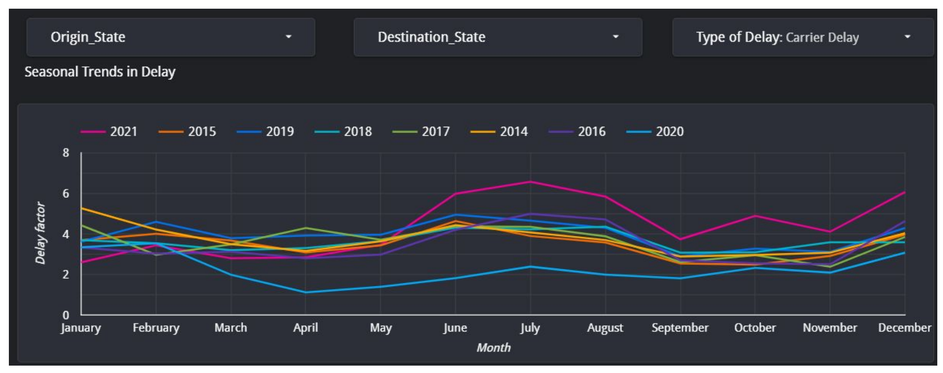
Model	Accuracy
Random Forests	72.06%
CatBoost	75.89%
Deep Neural Network	82.49%
Wide & Deep NN	73.80%

Despite having a lower accuracy, the Random Forests model was used for the prediction model due to it being a faster model and its ease of implementation for the prediction tool. Though the accuracy of our models is slightly lower, we are not using any delay factors to predict delay, whereas most research papers are using one or several delay factors in their prediction models.

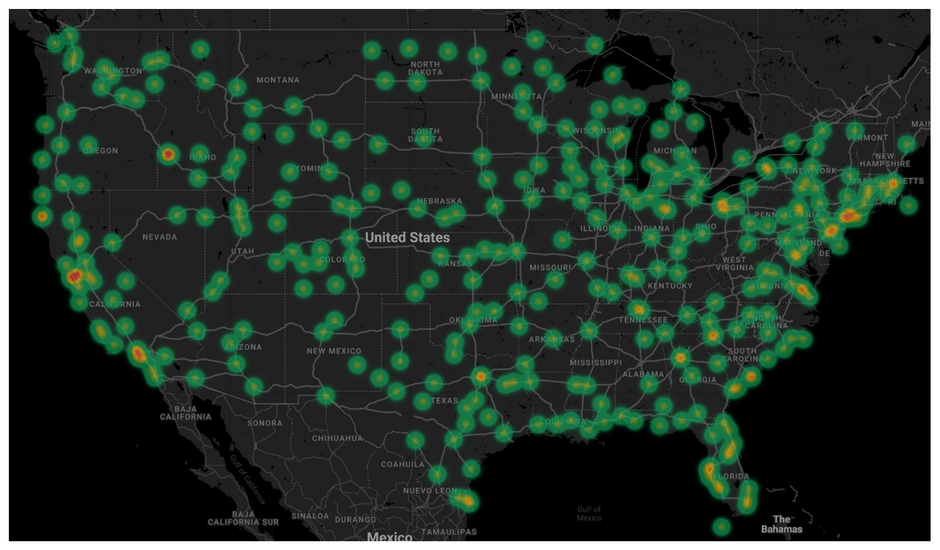
Dashboarding & Visualizations

The visualizations and interactive dashboards are created on Looker Studio. There are 5 dashboards with a variety of visualizations which cover:

1. Geo-Spatial Analysis
2. Trends in Delays
3. COVID Delay Trends
4. Air Carrier Performance
5. Time Plot Chart



The above figure shows seasonal trends in Carrier Delay by month, across 8 years. Using the dropdowns, the types of delays as well as origin and destination states can be varied.



The above figure shows a heat map of delays across the entire United States. The North-East and South-West states have the highest prevalence of delays.

Confusion Matrix

Below is the confusion matrix for the Random Forests prediction model, with an accuracy of 72.06%, run on 1 million randomly selected rows.

Actual \ Predicted	Not Delayed	Delayed
Not Delayed	130843	34332
Delayed	21540	13285

User-Friendly Interface

The final product is a simple form which users as potential passengers can fill to predict whether a certain flight has a possibility of delay, thereby allowing them to plan future trips more efficiently. Passengers can enter information such as the flight date and time, the airline they are flying with as well as the origin and destination airports, which will be used as the predictor variables for the Random Forests model, using which the model will predict the possibility of delay.