Midterm Project Progress Report

On-Time Flight Performance Prediction

IDS 561 | CRN 45604 | Spring 2022

Team

Anupriya Rastogi (arasto6@uic.edu)

Giovanni Alvin Prasetya (gprase2@uic.edu)

Rajaram Ramesh (rrames8@uic.edu)

Problem Setting

Being able to predict flight delays ahead of departure time can potentially reduce the last minute change of schedules and help airlines make the necessary preparations to ensure they are able to stick to the set schedules. To put our datasets into perspective, we are focusing on Airline on-time performance, airline passengers, airline employees, and relevant other factors to gather insights on what could possibly be the key to flight delays and cancellations.

Data Description

All datasets listed below are for the period 2019-2021.

Airline On-Time Performance:	On-time data for flights operated by US domestic carriers. Data includes on-time arrival and departure data for non-stop domestic flights by month and year, by carrier and by origin and destination airport. Data also includes scheduled and actual departure and arrival times, canceled and diverted flights, taxi-out and taxi-in times, causes of delay and cancellation, air time, and travel distance.
Passengers:	Data includes number of passengers that traveled domestically by month and year
Airline Employment Data:	Data includes the number of full-time and part-time employees by airlines, month and year.
COVID Data (CDC): (2020-2021)	Data includes COVID-19 cases, deaths, testing volume and vaccine rollout, nationally and state-wise, by day, month, and year.

Techniques

Based on datasets, we would like to use gradient boost to get the correlation of delays and specific airlines (which airlines have delayed more and what's the direction) then we could make a prediction based cases and vaccine data from CDC focusing on airline recovery post-pandemic kind of analysis. Second, we would like to link airline staffing to cancellations and see if there is any correlation in that area. This was based on an article that had come out last year when there were many flight cancellations due to staffing shortages. For the data visualization part, we would like to create a flight delay/cancellation ranking by airline, as a recommendation to future travelers.

Results

Airline On-Time Performance Dataset
Count of Values for each year

2019	2020	2021
7,422,037	4,688,354	5,995,396

Passengers Dataset
Total for each year

2019	2020	2021
811,471,767	335,045,847	611,908,076

Airline Employment Dataset
Total for each year

2019	2020	2021
8,884,326	8,494,869	8,569,916

Role of Team Members

Member	Role
Anupriya Rastogi	Data Processing, Identifying key variables
Giovanni Alvin Prasetya	Forecasting and Modeling
Rajaram Ramesh	Handling dataset, Visualizations

Progress so far

So far, we have downloaded the key dataset, the Airline On-Time Performance. For the Airline On-Time Performance dataset alone, we collected a total of around 6GB of data. We have merged the monthly files into yearly files to make it easier for use in python. We have also identified and downloaded additional datasets to complement the primary dataset that we believe would help us in generating valuable insights in the airline industry. We are still working on data processing and finalizing models for our project.