# Final 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

Flight delays are something that every traveler has most likely experienced at least once at some point in their journeys. Being able to predict flight delays ahead of departure and arrival time can potentially reduce the last-minute change of schedules. It can also help airlines make the necessary preparations in advance to ensure that they make the journey for their passengers as smooth as possible. The period we focused on for our project is from January 2019 to December 2021, which includes the two key year period of the covid-19 pandemic. To put our datasets into perspective, we are focusing on airline on-time performance, airline passengers, airline employees, and covid-19 cases to gather insights on what could possibly be the key to flight delays and cancellations. In this project, alongside predicting delays, we also try to look at the impact covid had on the airline industry in the US.

## Data Description

All datasets listed below are for the period 2019-2021, unless stated otherwise.

| Dataset | Description | Source |
|---|---|---|
| Airline On-Time Performance (Primary Dataset) | 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. This dataset includes data for 18 passenger airlines including United Airlines, Southwest Airlines, American Airlines and Delta Airlines. | Bureau of Transportation Statistics |
| Passengers | Data includes number of passengers that traveled domestically by month and year | Bureau of Transportation Statistics |
| Airline Employment Data | Data includes the number of full-time and part-time employees by airlines, month and year. | Bureau of Transportation Statistics |
| 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. | Centers for Disease Control and Prevention (CDC) |

**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

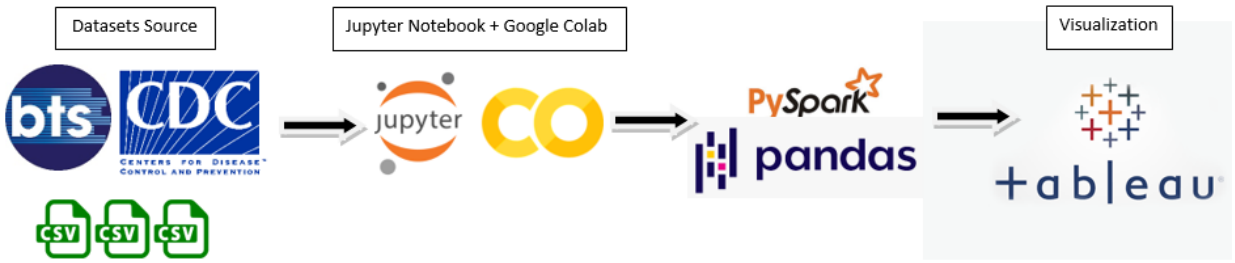
| 2019 | 2020 | 2021 |
|---|---|---|
| 811,471,767 | 335,045,847 | 611,908,076 |

**Airline Employment Dataset**
Total for each year

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

# Techniques

## Workflow

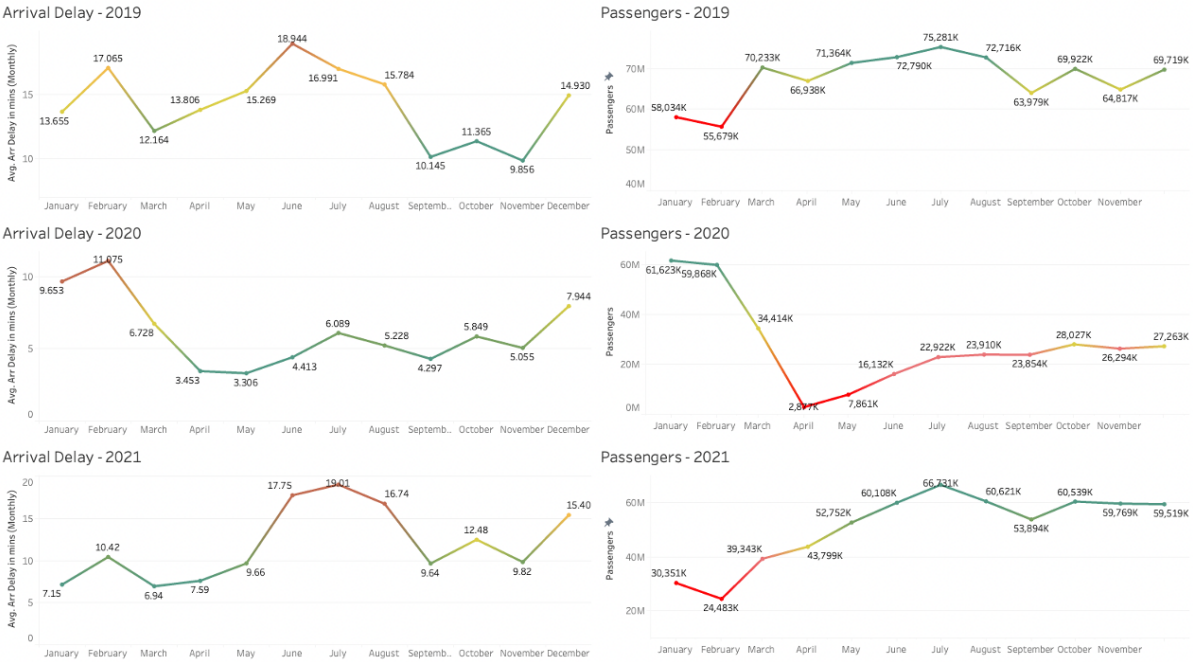



The chart above describes our project workflow. As stated earlier, we download our datasets from the Bureau of Transportation Statistics (BTS) and the Center of Disease Control and Prevention (CDC). We then used Jupyter Notebook and Google Colaboratory for data preprocessing and modeling purposes, respectively. For the preprocessing part, we used pandas to handle our primary dataset, and for the modeling part, we used pyspark to handle the datasets. For data visualization, we used Tableau to be able to generate insights from our datasets.

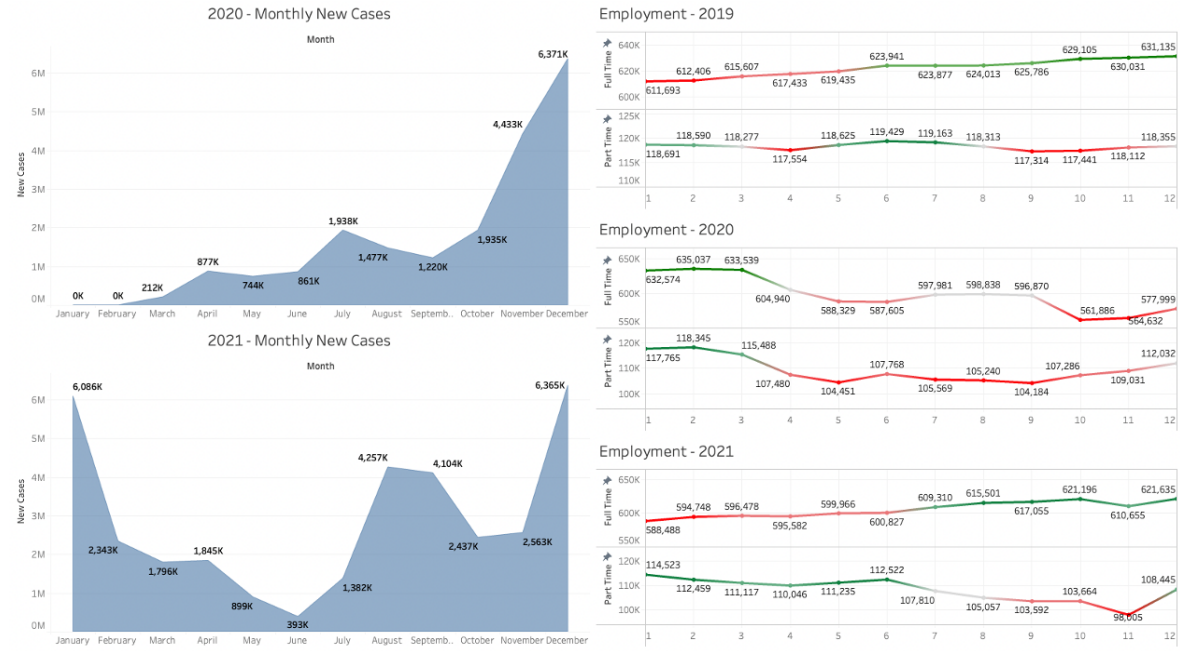
**Exploratory Data Analysis and Visualization using Tableau**

The other datasets that we used in addition to our primary dataset in this project really helped us understand the flight delays and cancellations for our particular time period. In addition to the usual delay factors, covid certainly had a huge impact on the airline industry. The visualizations below show arrival delay and passenger traffic for the respective periods. Passenger traffic and

average arrival delay for April 2020 both reduced together as the traveling came to a near halt at that time. The visualizations here show that passenger traffic is directly proportional to average arrival delay.




The visualizations below show monthly covid cases in the United States and airline employment numbers, both full-time and part-time, for the respective periods. As shown, employment in the airline industry took a direct hit in 2020 during the initial impact of covid-19 and has only recently started to recover in late 2021.



## Machine Learning Models

1. **Random Forest**

   Random forest is a classification algorithm consisting of many decisions trees. It uses bagging and features randomness to try creating an uncorrelated forest whose prediction is more accurate than individual trees. Each individual tree in a random forest shows a prediction and the class with the most output becomes the model used for prediction. We used a random forest model because it could outperform the low correlation key. To put it simply, the trees protect each other from their individual errors. While many trees could come out with an error, other trees would be right, so the result will be in the correct direction.

   For this project, we selected 'YEAR', 'MONTH', 'DAY_OF_MONTH', 'DAY_OF_WEEK', 'CRS_DEP_TIME', 'CRS_ARR_TIME', 'FLIGHTS', 'DISTANCE', 'DIVERTED' as the tree to make the prediction. We utilized “Cancelled” that stands for cancelled flight, “Dep_delay15” that stands for Departure Delay Indicator with condition of 15 minutes or More (1=Yes), and “Arr_delay15” that stands for Arrival Delay Indicator with condition of 15 Minutes or More (1=Yes) as the estimators of the correlation. We consider the delay and the canceled outcome as the uncorrelated models because the majority of the data shown on-time flight.

   Basically the results will show the accuracy of the prediction and how much the error population. The prediction represents the on-time flight population of the data while test error shows the “canceled” and “delayed” flight.

| | Cancelled Flight | | | Arrival Delay | | | Departure Delay | | |
|---|---|---|---|---|---|---|---|---|---|
| Year | 2019 | 2020 | 2021 | 2019 | 2020 | 2021 | 2019 | 2020 | 2021 |
| Accuracy | 0.98 | 0.94 | 0.98 | 0.81 | 0.90 | 0.83 | 0.81 | 0.91 | 0.83 |
| Test-error | 0.02 | 0.06 | 0.02 | 0.19 | 0.10 | 0.17 | 0.19 | 0.09 | 0.17 |

   After we performed the random forest model for each year (2019-2021), we have seen that during 2020 the arrival delay and departure delay is at their best, while the total amount of flights is the shortest compared to other years. It shows that during covid outbreak in 2020, the average on-time performance of US airline industries has performed better due to less busy schedules. It goes similarly to the “canceled flight” in 2020 which shows that the number of canceled flights are less than other years.

2. **Linear Regression**

Given that we wanted to produce a continuous result, we applied a linear regression model (delay in minutes). We also selected it because we expected that there would be a linear or more straightforward connection between the characteristics and labels, in which case a linear regression model would be more appropriate to use. To construct a features column, we developed a vector assembler. We utilized columns like "DEP_DELAY" and "TAXI_OUT" as we thought that they would be more directly related to estimating "ARR_DELAY", given whether the flight was delayed at departure or not. We used one-hot encoding to create seven new columns for the flight day of the week, which helps with converting the existing categorical data and make it useful for use in Linear Regression. For linear regression, we tried with and without regularization and didn't notice any difference in the two methods.

| | Linear Regression w/o Regularization | Linear Regression w Regularization |
|---|---|---|
| R2 | 0.938 | 0.938 |
| RMSE | 11.729 | 11.729 |

3. **XGBoost**

In addition to Random Forest and Linear Regression, we also decided to use XGBoost. XGBoost has become a very popular algorithm for machine learning problems in recent years because of its accuracy and efficiency. It also can be applied in situations where the problem is either classification or regression in nature. Therefore, we wanted to test this augmented model for this dataset to see if we could develop an even more accurate model. For XGBoost, we used the Scikit Learn packages as it best supports the algorithm. We used a smaller dataset of about one million rows to help speed up the tuning process as there were many hyperparameter combinations we wanted to test.

| | XGBoost |
|---|---|
| R2 | 0.947 |
| RMSE | 12.004 |

## Conclusion and Discussion

Given that our dataset is quite a complex one with numerous variables and millions of data, there are certainly many approaches when it comes to being able to accurately predict flight delay at departure time and arrival time. While we limited to using three models in our project, there are other models that may be better suited in this process such as Support Vector Machine (SVM) and many more. The topic of flight delay is quite an interesting field to study and understand the complexities tied to air travel and how it is managed effectively at a national and international level.

## Role of Team Members

| Member | Role |
|---|---|
| Anupriya Rastogi | Data Processing, Identifying key variables |
| Giovanni Alvin Prasetya | ML Modeling |
| Rajaram Ramesh | Handling dataset, Visualizations |