Datathon Summer 2023 Report

Juliette Garcia[1], Zachary Gin[2], Krish Mawalkar[3], Sonica Prakash[4]

July 30, 2023

Yale University[1], University of Pennsylvania[2], Carnegie Mellon University[3], California Institute of Technology[4]

Contents

1 Executive Summary

The United States airline industry has historically been recognized for its success in transporting millions of passengers. Specifically, the Bureau of Transportation reports that, in 2022, U.S. airlines carried 853 million passengers.

However, U.S. airlines face challenges that are dragging their performance, resulting in delays for passengers. The purpose of this report is to predict flight delays based on datasets including flight traffic, airport, and weather information. The broader significance of this report is to help policy makers further understand what factors most affect airline delays, allowing for future actions to be implemented to decrease domestic airline delays.

The primary aim of this report is to formulate predictive models that effectively anticipate airline performance metrics, with a particular focus on arrival delays, by leveraging various traffic-specific metrics. The central research question our team sought to address revolved around the feasibility of accurately predicting the occurrence of delayed airline arrivals.

We utilized a Random Forest Regressor Model and Decision Tree Regressor Model as our main models. We used two Random Forest Regressor models which yielded an R-squared of 0.922 and an R-squared of 0.155. The model which produced the lower R-squared had its variable set altered which changed the outcome of the model. The Decision Tree Regressor yielded a R-squared of 0.846. The difference in our R-squareds demonstrate that there are specific variables that can help predict flight delay. Consequently, These R-squared values demonstrate that our model can reasonably identify flight arrival delays.

The results indicate that there is a significant relationship between flight traffic, airport, and airline data and arrival delays. Our findings have important impacts for the airline industry and could allow airline companies to better communicate with their passengers about arrival times to ensure a smoother customer experience.

2 Technical Exposition

2.1 Data Collection and Preprocessing

We used multiple data files provided to us by Correlation One and Citadel. These datasets enabled us to have insight into airline delays and performance.

2.1.1 Data Collection

The objective of this report is to develop models that predict airline performance metrics, specifically arrival delays, using other traffic-specific metrics. Thus, we wanted to investigate any existing correlation between factors such as date, origin airport, arrival locations, and the initial scheduled departure time. In our models, we utilized the following datasets provided to us:

- **flight traffic**: The flight traffic dataset by the Bureau of Transportation, provides us with over 1 million data points on delays for domestic flights in 2017.
- **airports**: the airports dataset (US Department of Transportation) contains 322 data entries on domestic airport names and location (city, state, latitude, and longitude).
- **airlines**: the airlines dataset contains the list of airline id codes and names. There are 61 airlines listed.

We use the arrival delays as our response variable. From this same dataset, we also used many of the other variables within it, such as airport locations, taxi times, and initial schedules departure times as explanatory variables.

2.1.2 Data Preprocessing

To be able to conduct our data analysis more effectively, we had to do some data preprocessing. This meant making a new dataset that included only the values that we deemed useful within the context of our question. We made modifications to the flight traffic dataset to be able to use it more effectively. This dataset had many separate columns for each different type of delay, including weather delays, air system delays, airline delays, security delays, etc. However, all of these columns had a high percentage of NULL values. To improve the efficiency of our model, we created a new dataframe that only included the non-NULL values of these columns. Next, we manually calculated delays in departure, arrival, and elapsed time in the air through convertion to date time objects.

In the airlines dataset, we also did some minor preprocessing work by encoding all airline ID variables to numbers so that they could be used in our modeling work. Then, when it came to the airports dataset, no preprocessing was needed to conduct our exploratory data analysis. We used the airport data as it was provided.

2.2 Exploratory Data Analysis

2.2.1 Univariate Exploratory Data Analysis

In order to better understand the data at hand, we began by conducting univariate exploratory data analysis. Our main goal when we first started was to understand common airline and airport trends so that we could be more conscious of how airline and airport popularity could potentially skew our final results.

For instance, when we graphed the most common origin airports (as shown in Figure 1) we realized that there were six airports ATL (Atlanta), ORD (Chicago O'Hare), LAX (Los Angeles), SFO (San Francisco), DEN (Denver), and DFW (Dallas/Fort Worth) that were more likely to be the origin airport of a flight than any other in 2017. The reason for this trend is likely due to the fact that all six airports are either common airline hubs or are located in densely populated regions of the United States. We found similar skewed popularity trends when it came to airlines as only three airlines, AA (American Airlines), B6 (Jet Blue), and DL (Delta Air Lines), being responsible for more than half of the US commercial air traffic in 2017.

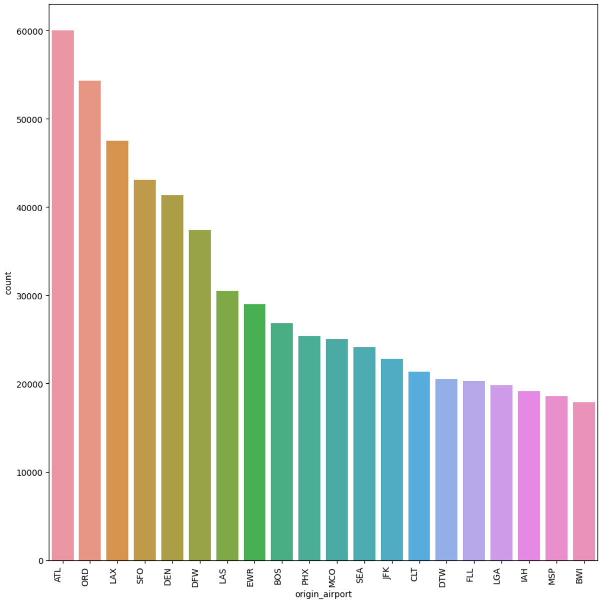


Figure 1: Most Common Origin Airports in 2017

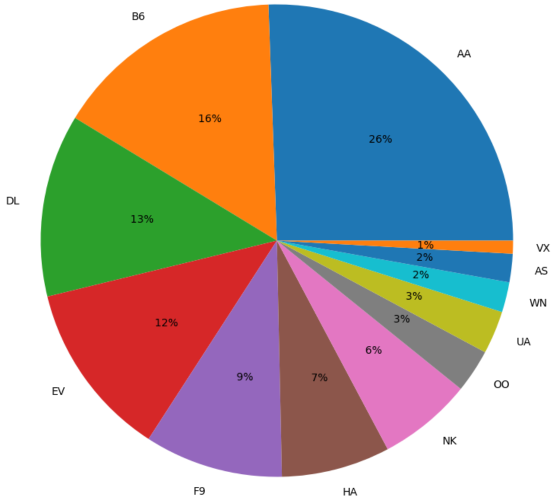


Figure 2: Most Flown Airlines in 2017

2.2.2 Bivariate Exploratory Data Analysis

To further understand airline delays in the U.S. airline industry, we conducted bivariate EDA on various major airline companies. We observe that for some airlines, such as VX (Virgin America), AS (Alaskan Airlines), and WN (Southwest Airlines), there is a noticeably smaller gap in arrival delay as compared to airlines like AA (American Airlines) and OO (SkyWest).

We also conducted bivariate EDA on taxi time by airline, as seen in Figure 4. Taxi time is the amount of time an airplane is on the ground. Taxi-in time refers to the amount of time an airplane is on the ground during landing time and time blocked while parking brakes are on. Taxi-out time is the amount of time between initial movement from a parking position to take off. We see that AA, DL, EV, OO, and UA have total taxi time at around 25 minutes. In our modeling, we want to explore whether taxi time, both taxi-out and taxi-in time, is a significant factor for arrival delay.

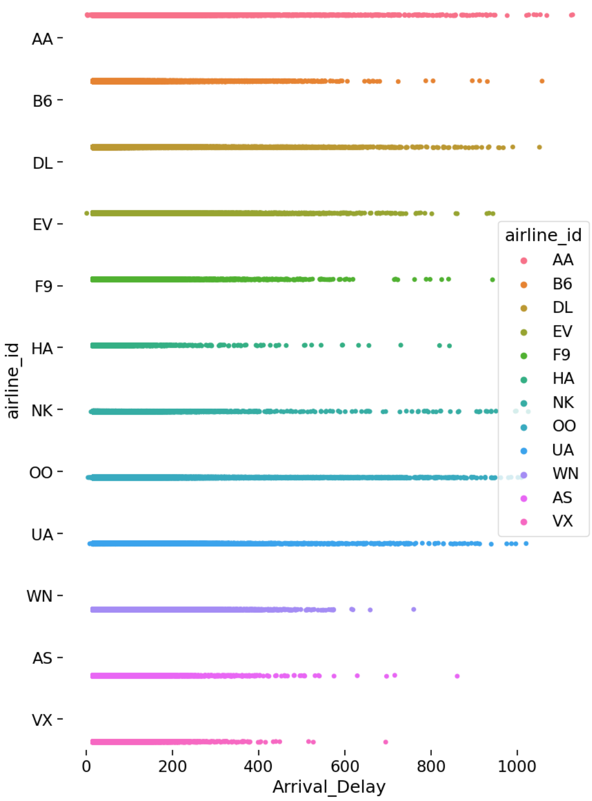


Figure 3: Arrival Delay by Airline Company

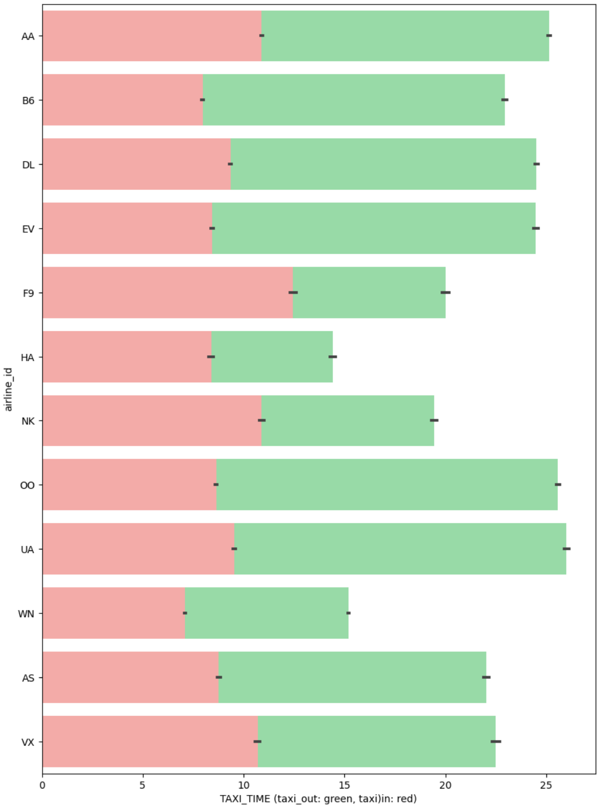


Figure 4: Taxi Time by Airline

2.3 Assumptions

We are conscious of the assumptions we made in our modeling. We assumed that airline, airport, and delay trends in 2017 would not have been different from trends in any other year. In other words, we assumed that 2017 would not be an outlier when it came to airline and airport trends. We would not have made this assumption, for instance, if our data was instead from 2020, when there was a major global event that had significant impacts on airlines and airport trends. However, 2017 did not have any global/national/regional events that would have significantly and abnormally affected the airline industry.

2.4 Developing the Model

As mentioned before, our goal is to develop models that can predict the length of the delay of arrival (in minutes) using various variables as found in the flight traffic dataset. We proceed through two iterations of modeling.

We focus on using Random Forest Regressor Models and Decision Tree Regressor Models as the main implementation for our goal. We chose the Random Forest Regressor due to its ability to predict continuous values such as arrival delays by averaging over multiple decision tree built for the data, such that it has the potential to achieve better fit to larger datasets that comparable techniques such as linear regression and logistic regression. The averaging technique of Random Forest helps eliminate overfitting issues. It is less sensitive to noise in the data and is able to capture non linear relationships between variables which makes it preferable.

We also use Decision Tree Regressors which are easier to interpret if needed for further research because it contains less of "black box" in the modeling than the Random Forest Regressor does.

2.5 Modeling

We first started by building a Random Forest Regressor and Decision Tree regression for predicting Arrival Delays using features that included 'wheels off' and 'wheels on' columns in the data. The complete list of features is listed in the feature importance analysis below.

The Random Forest Regressor achieved an R-squared value of 0.84 and the Decision Tree regression achieved an R-squared value of 0.92. These means that these models explain at least 84 percent and 92 percent of the variation in the arrival delay durations, respectively. All of the predictor variables were standarized using the Standard Scaler methodology in Python sklearn which works by removing the mean and scaling each feature to a unit variance.

We then were interested in examining feature importance of the different predictor variables on arrival delay times to guide further analysis on how we can improve airline efficiency with respect to on-time arrivals.

We used permutation feature importance to determine the importance of various features in the predictive power for both the Random Forest and Decision Tree Regressors. Permutation feature importance reflects the decrease in model score when a single feature has its data shuffled randomly in a column. For both models, we can see that Wheels On time and Departure delays were very significant predictors in arrival delay time. Airline id has a very small importance and all other variables do not seem to impact model performance at all.

However, we noticed that Wheels On is very fundamentally correlated with arrival delay time because these events happen within the same time frame. Thus, in the second round of modelling, we again used a Decision Tree Regressor and Random Forest Regressor on a different set of features. We removed Wheels On/Off time as features in the analysis and added Elapsed Delay as a feature into the model. This allows us to examine other variables that might impact arrival delay on a less obvious level than Wheels On/Off time. (Elapsed Delay was calculated using the difference in the Scheduled and Actual Elapsed Time in the air).

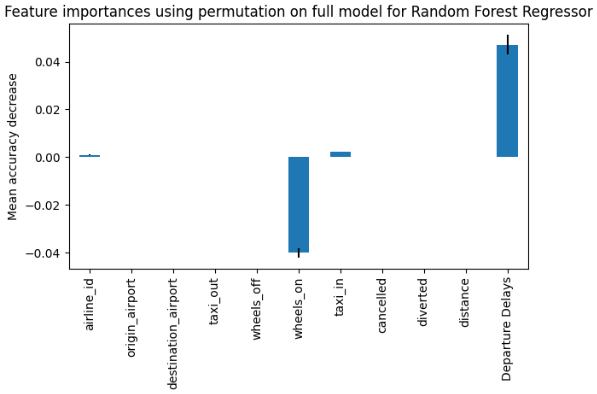


Figure 5: Permutation Feature Performance on Random Forest Regressor

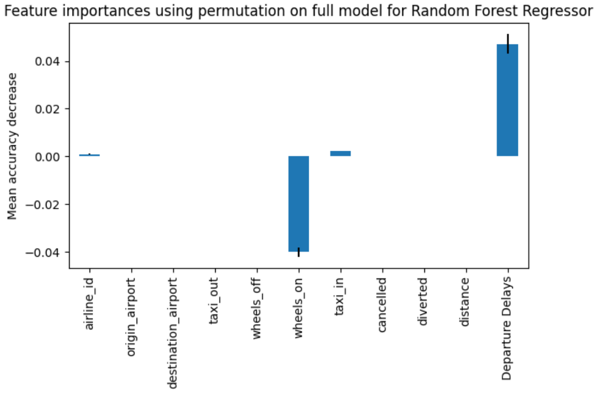


Figure 6: Permutation Feature Performance on Decision Tree Regressor

The Decision Tree Regressor this time only had an R-squared value of -0.7. The Random Forest Regressor performed slightly better with an R-squared value of .15. Even though this does not indicate very high predictive power with either model, we still found it to be interesting to again analyze permutation feature importance. This time, without the Wheels Off and On column, we find that other variables carry more predictive power such as airline id, taxi in/out, elapsed time. Departure delay remains to be a significant predictive factor in both models.

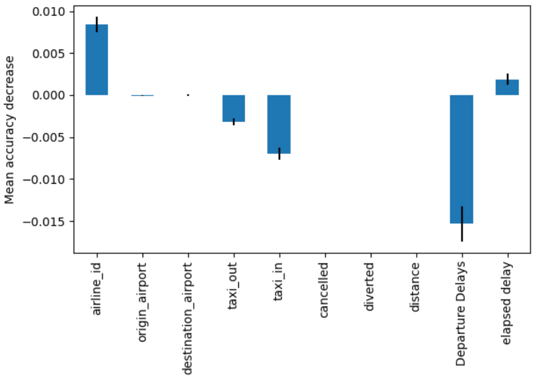


Figure 7: Permutation Feature Importance for Second Random Forest Model

2.6 Results and Discussion

2.6.1 Results

In our research, we employed three primary models to predict flight delays: a Decision Tree Regressor and two distinct Random Forest Regressors, which utilized different sets of variables. The Decision Tree Regressor provided valuable insights into the relationships between arrival delays and the target variables, enabling us to interpret the impact of individual factors on flight delays. On the other hand, the two Random Forest Regressors leveraged the power of ensemble learning to deliver more robust predictions, taking advantage of feature selection variations to optimize performance. The versatility of these models allowed us to explore different combinations of variables and methodologies, facilitating a comprehensive analysis of the factors influencing flight delays in the US airline industry.

Models			
Model Variables	Decision Tree Regressor	Random Forest Regressor (Figure 5)	Random Forest Regressor (Figure 6)
Mean Absolute Error	21.693	19.225	125.177
Mean Squared Error	15292.406	7719.273	83805.403
Root Mean Squared Error	123.662	87.859	289.491
R^2	0.84587	0.92219	0.15534

2.6.2 Discussion

The Random Forest Regressor exhibited a high R-squared value of 0.922, indicating a strong explanatory power in predicting flight arrival delays. Similarly, the Decision Tree Regressor achieved a respectable R-squared value of 0.846, highlighting its effectiveness in capturing the underlying relationships between variables and the target outcome. However, it is important to note that the R-squared value of 0.155 for our last model, which incorporated different variables from the second model, suggests a comparatively weaker predictive performance. Nevertheless, the substantial R-squared values of the initial two models underscore the model's ability to reasonably identify flight arrival delays and reveal a statistically significant relationship between flight traffic, airport, and airline data in influencing arrival delays. These compelling findings hold the potential to enhance passenger communication and operational decision-making within airline companies, offering opportunities for more efficient and customer-centric services.

3 Future Research

Possible future research questions pertaining to airline performance and predicting flight delays could be:

- **Economic Impacts of Flight Delays on Airlines**

  One critical aspect that merits further investigation is the economic impact of flight delays on airlines. Understanding the financial ramifications of delays is crucial for airlines to devise effective strategies and mitigate revenue losses. Future research could delve into quantifying the direct and indirect costs incurred by airlines due to delayed flights. this analysis might encompass factors such as decreased ticket sales, compensation paid to affected passengers, operational inefficiencies, and the potential loss of customer loyalty. Furthermore, examining how airlines manage these costs and the long-term implications on their financial health would be valuable in guiding industry stakeholders

- **Customer Loyalty and Perception amid Lengthy Flight Delays**

  A key research question that remains unexplored pertains to customer loyalty and perception towards airlines experiencing high volumes of lengthy flight delays. Investigating whether passengers maintain their loyalty to specific airlines despite recurrent delays could shed light on the role of brand reputation and customer service in shaping consumer behavior. Furthermore, understanding the thresholds beyond which customer satisfaction declines significantly, leading to a shift in airline preference, can inform airlines on setting acceptable service standards and managing customer expectations during disruptions. This research may involve surveys, customer feedback analysis, and sentiment analysis of social media data to gain insights into passengers' sentiments.

- **Developing Predictive Models that Account for Traveling Popularity**

  An area of research that holds considerable promise is the development of predictive models that can account for traveling popularity and its influence on flight delays. Understanding how varying demand levels impact delays can help airlines optimize their scheduling and resource allocation. Researchers can explore the integration of factors such as seasonality, special events, holidays, and peak travel periods into predictive models. Utilizing machine learning techniques and historical data, such models could offer airlines real-time predictions on the likelihood of delays based on the anticipated popularity of specific routes and destinations. This research can lead to more informed decision-making for airlines, resulting in improved operational efficiency and enhanced customer satisfaction.

- **Dynamic Rescheduling Strategies for Delay Management**

  Exploring dynamic rescheduling strategies to optimize flight operations in the face of delays is another compelling research avenue. Investigating how airlines can efficiently manage and rearrange flight schedules to recover from delays and minimize cascading effects is of paramount importance. Such research could focus on algorithmic approaches that take into account various constraints, including crew availability, aircraft utilization, and regulatory limitations. Implementing effective rescheduling mechanisms can lead to significant improvements in operational efficiency and passenger satisfaction, making this a relevant and impactful research area.

In conclusion, the field of airline performance and predicting flight delays offers numerous avenues for future research. Exploring these questions in a systematic and data-driven manner can contribute to the overall improvement of airline operations, enhance customer experiences, and inform policymakers on measures to reduce domestic airline delays effectively. As advancements in data collection, analytics, and technology continue to evolve, researchers have unprecedented opportunities to gain deeper insights into this dynamic industry and contribute to its sustainable growth and success.

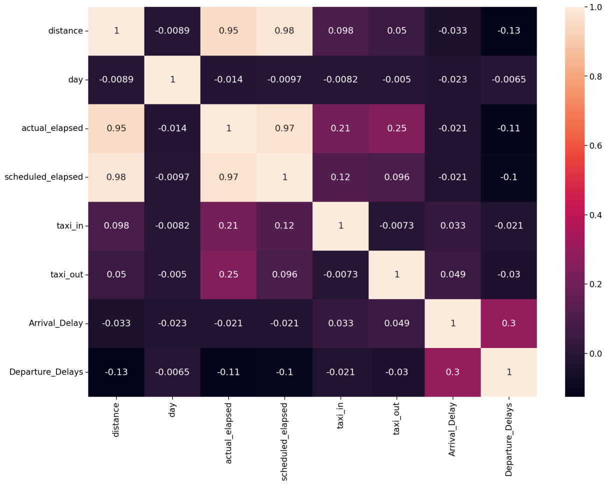
4 Bibliography

O'Hare, Maureen, 2023, https://www.cnn.com/2023/07/22/travel/travel-news-us-air-employment-flight-delays/index.html.

U.S. Department of Transportation's Bureau of Transportation Statistics, 2023.

U.S. Department of Transportation's Bureau of Transportation Statistics, 2015. Kaggle, https://www.kaggle.com/c
delays.

5 Appendix



	distance	day	actual_elapsed	scheduled_elapsed	taxi_in	taxi_out	Arrival_Delay	Departure_Delays
distance	1	-0.0089	0.95	0.98	0.098	0.05	-0.033	-0.13
day	-0.0089	1	-0.014	-0.0097	-0.0082	-0.005	-0.023	-0.0065
actual_elapsed	0.95	-0.014	1	0.97	0.21	0.25	-0.021	-0.11
scheduled_elapsed	0.98	-0.0097	0.97	1	0.12	0.096	-0.021	-0.1
taxi_in	0.098	-0.0082	0.21	0.12	1	-0.0073	0.033	-0.021
taxi_out	0.05	-0.005	0.25	0.096	-0.0073	1	0.049	-0.03
Arrival_Delay	-0.033	-0.023	-0.021	-0.021	0.033	0.049	1	0.3
Departure_Delays	-0.13	-0.0065	-0.11	-0.1	-0.021	-0.03	0.3	1

Figure 8: Correlation Matrix for Flight Traffic Data