Flight Delay Prediction

Group Members: Neha Karna (nak28), Maille Sherry (mm136), Prince Ahmed (pa99), Sophia Maupin (sam199)

I. Introduction

It's estimated that in a given year, one quarter of all U.S. flights are delayed and in the past year, 120,000 flights were completely canceled ("Airport Travel Stats to Know Before You Go", 2015). Flight delays and cancellations inconvenience millions of travelers. Furthermore, delays have a substantial economic impact. Both passengers and airlines experience high costs when a flight runs off schedule. Passengers are often left stranded and financially responsible for meals and accommodations. Delays and cancellations can also be costly for airlines who are still responsible for paying crew members and issuing refunds while losing revenue. For instance, one winter storm in 2015 resulted in thousands of missed flights and is estimated to have cost airlines and passengers a combined $200 million ("Canceled flights cost airlines, fliers millions", 2015) Our analysis targets understanding the most important factors impacting flight delay length as well as projecting flight delays and cancellations.

Our primary research question is: among all North American flights in 2018, what factors are most predictive in determining flight delay length? We also have a sub-question: is there a significant difference between total delay time during the holiday season versus less busy travel times? We selected 2018 because it was pre-pandemic and contained a more diverse representation of airlines than 2019.

Beyond providing insights into an area of economic impact, this topic is relevant among the scientific community. Belcastro et al. (2016) explore the impact of flight information and weather conditions on flight delays and attempt to generate a predictor of arrival delay. Similarly, Abdelghany et al. (2012) investigate flight delays during irregular operation conditions and implement a model to project delay times.

II. Data Sources

For this project we utilized two datasets. The first dataset we found is the "Flight Status Prediction" dataset from Kaggle (linked here). It contains flight information by airline for the 2018-2022 time period. There are five CSV files each corresponding to a year. Each file has around 5 million entries and there are 61 variables for each flight. These variables include flight date, airline, destination, and cancellation status. We used this dataset to originally explore and finetune our research question. More numerical data relating to delays was needed so the original source of the Kaggle dataset was explored. Custom datasets were made from the On-Time : Marketing Carrier On-Time Performance database made by the Bureau of Transportation Statistics. This custom data also has flight date, origin and destination, but it more importantly has detailed data recording delays and causes for these delays. It has delay time in minutes attributed to carrier delays, weather delays, national security delays, security delays, and late aircraft delays. The data from the database was pretty clean. The only way it was processed was by removing NaN values, by either dropping them or replacing them with zeros.

III. Modules

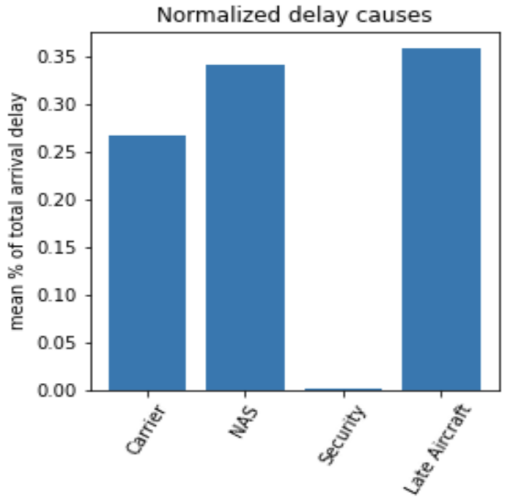
We used modules 4, 5, 8, and 9 in our project. We used Module 4 in the data cleaning stage. Both datasets we found needed some wrangling as they both contained NaN values. We needed to understand why the NaN's occurred in order to appropriately address them to make our data usable. For our sub-question question, we wanted to understand the difference in mean delay time between different time periods. Therefore, we relied on Module 5 to construct confidence intervals to ensure that the 2values we find are statistically significant. To construct

our confidence intervals we used stats.ttest_ind because we had two independent datasets, one for flight delays in September and one for flight delays in December. We additionally used Module 8 to create informative visuals. Originally, scatter plots and bar charts were used to help us understand the data for our data investigation. However, we also used Module 8 in data analysis to create a bar showing the distribution of the causes of the delays in order to better understand what factors were most contributory to delays. Module 9 was also helpful in our data investigation since we created multiple prediction models, utilizing regression, to determine the explanatory value of different factors. We first created Logistic and linear regression models as well as a KNN model. Ultimately, we utilized a Linear Regression to try and create a model to predict whether or not a flight would be delayed to which factors were most predictive.

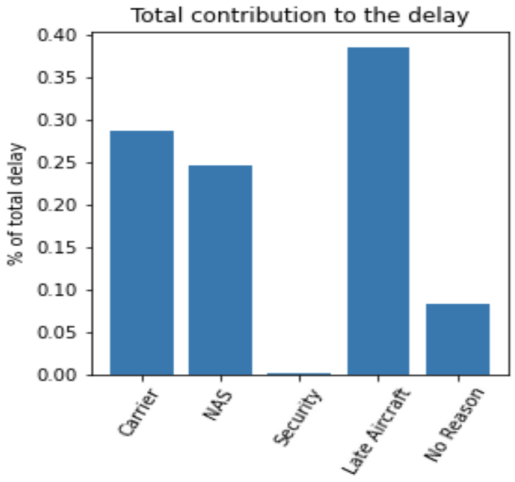
IV. Results and Methods

Primary Question: Predictive Factors

Our overarching research question involved finding the most predictive factors in determining flight delay length. To address this, we began by creating visualizations to find the most contributive factors to flight delay length. The first visualization we created is titled “Normalized Delay Causes” (see below)



For each flight that was delayed, we normalized by dividing each cause by the total arrival delay time. . Then this was plotted as a barchart with the causes on the x-axis. Our next visualization is titled “Total contribution to the delay” (see below).



For this visualization, we calculated the sum of all the delays for each cause as well as the sum of total arrival delay. This allowed us to

show how frequently different causes were involved in flights that were delayed but did not show how much each cause contributed to overall arrival delay length. We then created a bar chart with % of total delay on the y-axis and causes on the x-axis. We also created visualizations exploring the day of the week and different airlines. However, as our research question evolved we found these factors to be less explanatory and significant and ultimately decided not to expand upon them. All our code, and visualizations can be found at https://github.com/prince-ahmed/216_final_project.

Taking these two visualizations together, our results showed that late aircraft and carrier were the most frequent contributing factors to flight delay. Late aircraft contributed about 38% of flight delays and carriers contributed about 27% of flight delays. However, we also showed that late aircraft also contributes the most to delay time. Of the total arrival delay time 35% was attributable to late aircraft, 34% was attributable to National Air System (NAS), and 27% was attributable to Carrier. One finding that we felt was significant was that NAS had a large difference in the two visualizations. This indicates that NAS is not frequently a factor involved in flight delay but that when they are involved they contribute significantly to the length of the delay.

Building off the visualizations we attempted a few different methods to build a predictive model using the different causes of delay. We first built a Linear Regression using the second dataset. We built different linear models for each of the causes to see which one was most accurate in predicting delay length. Our training data for each model included the time of delay attributable to each cause over the month of December 2018. We chose to build a Linear Regression because we felt it was the simplest model and we felt like it would be a good starting point for our analysis.

We then calculated MSE and r^2 for each of the different models to evaluate how accurate each model was. These results have been compiled and are attached below.

Model trained on:	**MSE** ($1000\ minutes^2$)	R^2
Carrier	4.3	0.34
Weather	5.9	0.10
NAS	6.0	0.09
Security	6.5	0.0009
Late Aircraft	4.7	0.28
Baseline	6.5	0

We found that the model trained on Carrier had an MSE of 4.3 which was the smallest. This indicated that of the causes, Carrier had the most predictive power in determining length of delay. This finding is unexpected given our visualizations. Late Aircraft was the biggest contributor to overall delay in the visualizations. Therefore, we expected the model trained on Late Aircraft data to be most accurate. That model came second with an MSE of 4.7. Consequently, we have considered our results inconclusive as our results from the model contradicted our results from the visualizations. However, we believe that the visualizations are more accurate because the data used to create those had been normalized while the data for the models had not.

Due to the unexpected results of the Linear Regression, we also attempted to build a Logistic Regression Classifier and a KNN Model. We tried to predict whether a flight was going

to be delayed for more than 25 minutes as the threshold between the two classes. We used a DummyClassifier that always returned the most common class to test the models. The models performed no better than their respective baselines, even when the class was changed to 0 minutes, and we therefore decided to not pursue them further. They have been excluded from the results.

Sub Question: Holiday Effect

Our research sub-question was whether or not there was a significant difference between total delay time during the holiday season versus less busy travel times. To address this, we utilized hypothesis testing. The null hypothesis was that the arrival delay comes from the same distribution for both September and December. Our alternative hypothesis was that the arrival delay comes from different distributions for the two months. The hypothesis testing allowed us to see if the data collected was significantly different between the two months. This would indicate that one month may have had a different number of delays, total delay length or some other significant difference. We formulated our null hypothesis because we predicted there to be a difference in delay length due to the holiday season in December. We predicted that due to more people flying in December there would be noticeably more delays.

To test this we used stats.ttest_ind because we had two different and complete datasets. Our test reported a p-value of $1.2 \cdot 10^{-32}$, which is less than 0.01 so we rejected our null hypothesis. We concluded that there is a significant difference in delay time between the holiday season and non-holiday months. However, it is important to mention that the p-value we observed was extraordinarily small which we have determined may be potentially due to our large dataset. It is possible that our data has a small effect size and our results overstate the difference. For these periods we also manually calculated the mean delay time to compare.

We also performed another t-test with the same hypothesis but utilizing more data. This time we expanded our holiday season to include December and January. Our non-holiday season included September and October. The p-value for this t-test was $1.6 \cdot 10^{-214}$ so we again rejected our null and concluded that there was a significant difference between the two periods. Similar to our first test, our p-value was incredibly small, which we contributed to the same factors as the first t-test.

V. Limitations and Future Work

There are several important limitations of our research. First, the generalizability of our findings is limited given we only focused on U.S. flights that occurred in 2018. International tourism is on the rise and therefore it may be important to include flights outside the U.S. in future work. Furthermore, the pandemic has transformed the airline industry, making 2018 a potentially less relevant year to explore. Our dataset included 28 distinct airlines, and while sufficient, this is still only a select number of airlines that could be expanded. When attempting to answer our sub-question, we assumed that flights during the month of December were a good representation for the holiday season. We also used September as a proxy for the non-busy season. This is potentially problematic given the holiday season spans a wider range of time and November could be a viable candidate. Similarly, it could be argued that September is not the best choice for the non-busy benchmark. We also faced constraints stemming from our dataset's large size. With over 5 million observations, our dataset was difficult to handle given hardware limitations and insufficient memory storage in the virtual container kernel. This forced us to be more selective in how we screened our data. Additionally, our p-value may not be as meaningful given our huge sample size.

Future work would include a more representative dataset, with deals from outside the U.S. and ones that occurred post-pandemic. For future work, to generate more comprehensive results, we would need computing devices with more computational power to handle datasets

with millions of observations. A different way of testing our hypothesis would need to be explored, since we had conflicting results. Different computational models or hypothesis tests could be used. A way to further our investigation would be to find data specifying more nuanced causes of the flight delay, which would allow delays of a particular cause to be analyzed. For instance, carrier delay could be broken down into more specific causes like baggage, crew problems, and aircraft damage. This would likely generate more meaningful results that could be better employed by industry experts and airline companies to make improvements. The cause for differences in overall delay time between holiday and non-holiday months could be explored more. Only the presence of a difference was confirmed. A second area of interest is comparing flight delays pre- and post-pandemic to see if reduced airline demand impacted delays. If so, it might also be interesting to see if delays have returned to their pre-pandemic levels.

VI. Conclusion

Our two methods, regression and visualization, generated slightly contradictory results, and therefore, we are unsure which cause of delay is most impactful. The regressions revealed that carrier delays had the lowest error, which would indicate they were the best predictor of overall delay time. Based on this method, late aircraft delay was the second best predictor. In contrast, Figures 1 and 2 show that among all delay types included, late aircraft delay is the most common and impactful. Figure 2 shows that carrier delays happen more often, but NAS delays were more impactful when they occurred. We think the visualizations are more accurate, since we are able to look at normalized delay rather than comparing numbers that might be in different magnitudes.

Based on the p-values from our t-tests, we found that there is a significant difference in the overall delay times between holiday and non-holiday seasons. The p-value was essentially 0 so we are confident that there is a difference between the delay time in the combined time period that includes September and January when compared to that of November and December. The only potential problem relates to the previously mentioned limitation that the large sample size may make our p-values less meaningful. We did not, however, explore reasons for the difference. This result should not be generalized, since flying during a holiday month does not guarantee a delayed flight. We can only conclude that on average, the delays experienced during the holiday season were different than in non-holidays months. The hypothesis test is not directional, but by comparing the means, we can see that the average delay in holiday months is longer.