I had a data covering five years of flight information to analyze. The duration included COVID-19.

The goal of my project was “to predict if, before a flight takes off, it will suffer some type of disruption (including cancellation, diversion, & delay).” Of all the options I used the medium sized data set to include more data than the small yet still be able to process it in an appropriate amount of time. I only used data that was provided before take-off since my model is supposed to be trained on prediction. For example, I used scheduled duration for the flight, the origin airport, the airline, year, month, day of week, and what time it was supposed to take off, etc.

I first split the data into training and test sets. We do this to reduce bias and evaluate whatever best model we create on new - unseen - data. After that I cleaned the data and standardized it so that it could be used. Essentially this allows every piece of data to be called into in a manner that is understood by any graph/model trying to be fit to it.

I then visualized the proportion of disruptions in the training and test datasets to ensure they were relatively similar. This is done in case for some reason - by random chance - one of the sets is heavily biased. I also visualized a graph showing disruptions before, during, and after COVID-19 just to witness and convey the impact it had on disruptions in the data.

After this I created a pipeline. I did this because it makes all “transformations in advance” and allows me to apply the same criteria to each data set (train or test). It allows me to standardize values and encode categorical variables if needed. It makes the numerical data scaled to be comparable and the categorical data encoded so that it can be used by the model. These are then combined so that their combined effect can be understood by the model.

After creating this pipeline, I used it to fit three different types of classification machine learning models: Logistic regression, decision tree, and random forest models. Logistic regression is good to use in binary tasks (disruption or not) and even more so because the dataset was imbalanced: a lot more non-disruptions than disruptions (as expected). I used the decision trees method because it is a non-linear model to fit the data, and I used random forests because it is a more comprehensive version of decision trees because it has a higher potential of catching overarching patterns in large datasets while also having a higher chance of not overfitting the data.

I used a ROC curve to display the true positive rate vs false positive rates for each model. The random forest and logistic regression were close (.59 and .61 respectively) while the decision tree lagged behind (.52).

I chose to fine-tune the random forest because it had almost the same recall as the logistic model while having far greater accuracy. The random forest was accurate about 78% of the time and it was good at predicting non-disruptions compared to disruptions, but it was still better than the other two models. I did this by utilizing a hyper parameter grid search. Basically what this does is explore a ton of values for different combinations of parameters to see if they are accurate in effect. This helps the final model perform well across many different datasets vs the one it was trained on. This process was extremely time consuming because of the amount of features I incorporated and the size of the dataset, however this also increased the helpfulness of the model because it was trained on more data and had more possibility of finding significant effects from the features used. The fine tuned model performed slightly better than the original fitted model. The conclusion of this analysis is having a model that can predict whether a flight will be disrupted with a greater chance than not having the model. If refined and utilized on a large scale this could be used to save airline company’s a lot of money by making their logistical models more efficient or it could be utilized by a company to try and disrupt/increase the knowledge a customer has when booking their traveling, hopefully giving the every day traveler a leg up by knowing with a higher degree of certainty if their flight will be disrupted.