Cody Schwarz

Springboard Capstone 2 Final Paper

For my second capstone project I want to attempt to predict the likelihood of a flight being delayed. There are many variables which can affect a flight being delayed or canceled(weather, which airline it is, location, etc.) and we will build a machine learning model using these factors to predict the likelihood of flight delays and cancellations.

Online travel agencies such as Expedia, Hotwire, TripAdvisor, SkyScanner, etc. would be able to significantly increase the quality of their product with this. Even though it targets a narrow market, of flights that get delayed or canceled, it is extremely valuable for the consequences and troubles it produces. These companies would love to have a tool like this they could use to show their customers.

The scenario: Imagine your sister is having a big wedding out of town and you are flying there. An hour before take off, as you are in the airport, it is announced that the flight is delayed or even worse, canceled. Without the foresight to plan ahead and have another option you may be completely out of luck. Imagine, however, if when you booked the flight, you were provided with a warning. ("Due to inclimate weather, the chances of this flight being delayed is 8%").

I decided to use data from the Bureau of Transportation Statistics, and I will use it to predict the likelihood of a flight being delayed. I decided to take the data from the month of April 2022 to conduct my research, as any more months would have been too much data to analyze. One month proved to be very adequate and rigorous. The data was extremely well put together, because it comes from the Bureau of Transportation Statistics, and there was very little work needed to clean it. Besides realigning some columns and removing unnecessary values I was ready to go to work.

Immediately it was interesting to just look at the data. The average distance people traveled was seven-hundred and ninety miles, which was extremely surprising to me as I expected this number to remain near the five-hundreds. This indicates to me that there are a lot more flights traveling further distances and those flights at the very top (furthest distance being just over five-thousand miles) skew the data slightly. As we can see that our 50% for Distance is only six-hundred and fifty miles, while still more than expected, is significantly less than that of our mean.

After creating a graph displaying the frequency of distances traveled it is clear to see that we were correct. While the vast majority of trips range from three-hundred to a thousand, many people refuse to fly short distances. And while there exist flights of extreme distances our data will be skewed in that direction. This type of distance information can be crucial to come back to and recognize patterns with flight delays.

After running our models we obtain accuracy scores of 0.677 for our Logistic Regression model, a score of 0.683 for our Naive Bayes model, a score of 0.718 for our Random Forest Model, and lastly a score of 0.675 for our KNN model. Obviously, these are not ideal accuracy ratings from our models. While the data is very complex and predicting flight delays is difficult, we will have to look towards Hyperparameter Tuning to increase the accuracy of the model and provide more accurate predictions.

From the HyperParameter Tuning of our most accurate model, Random Forest Model, we actually see a slight decrease in accuracy to 0.716. This is somewhat concerning, but given the nature of the project it is very difficult to increase this score. There are many factors that play a role in flight delays or cancellations, but the biggest of which might be the weather. The weather is an extremely difficult thing to predict and thus causes randomness in the model. This makes increasing our accuracy score very difficult.

In conclusion, I really enjoyed this project. It was extremely challenging and much more rigorous than my first capstone. Even though I could not increase the accuracy with the Hyperparameter Tuning I still feel there is much more work to be done. Our model ended up returning a F1 score of 0.81 and while that is nothing to write home about, it does indicate potential for our model to grow. Knowing you have less false negatives and false positives is always a nice place to start.