Cody Schwarz

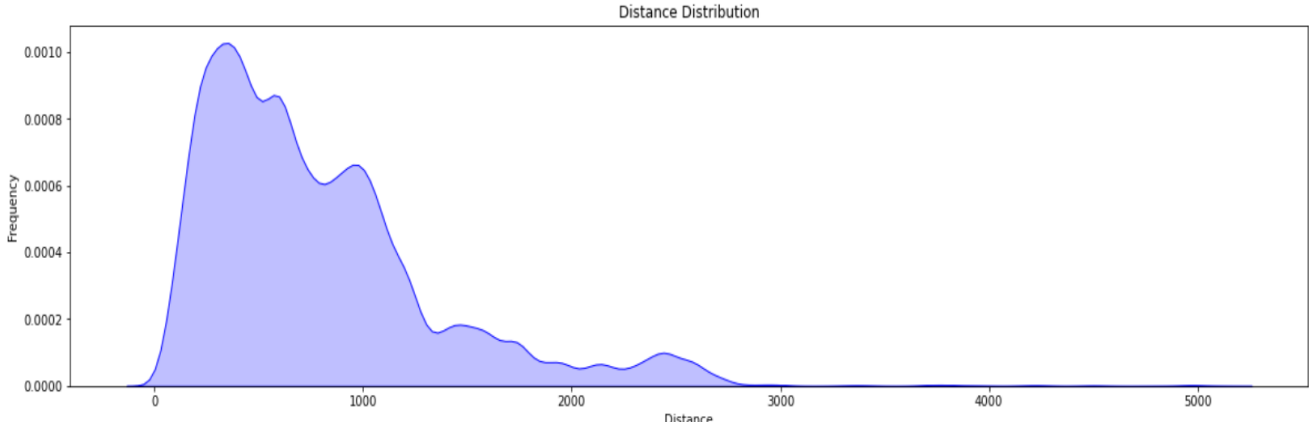
Capstone 2 Milestone Report

Before beginning with the data I made some initial guesses just to see if what I thought was the truth. Upon initial look at the data, however, it was clear to see I was pretty wrong. I suspected the average flight to only be four-hundred miles and it turned out to be almost double that. As we can see from the data:

	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	DEP_TIME	DEP_DEL15	DISTANCE
count	123697.0	123697.000000	123697.000000	123697.000000	123697.000000	123697.000000
mean	4.0	14.594024	4.034795	1453.154814	0.664382	791.301082
std	0.0	8.109583	1.884684	493.498396	0.472208	571.129985
min	4.0	1.000000	1.000000	1.000000	0.000000	31.000000
25%	4.0	7.000000	3.000000	1058.000000	0.000000	370.000000
50%	4.0	14.000000	4.000000	1505.000000	1.000000	646.000000
75%	4.0	21.000000	5.000000	1845.000000	1.000000	1032.000000
max	4.0	30.000000	7.000000	2400.000000	1.000000	5095.000000

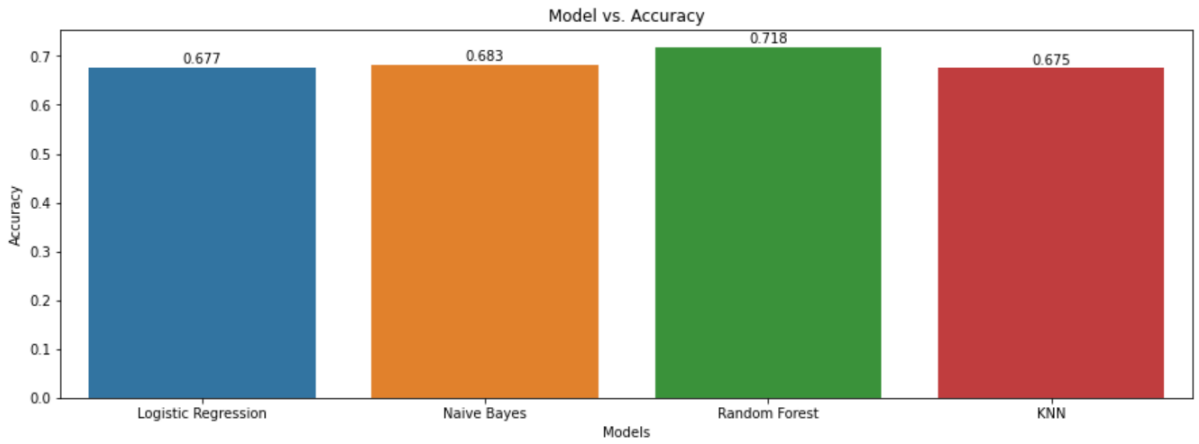
There is a clear disparity in distance. While the other categories from this image have little value because they are days of the month, (14 = 14/30) and days of week (4 = Thursday), the distance statistics are extremely relevant. As we can see even though the average distance is seven-hundred and ninety miles, the 50% value is only six-hundred and forty. This discrepancy is a result of people not flying on short flights. As we can see from the following graph of Distance

vs Frequency:



People very infrequently take flights of distances shorter than two-hundred miles. However, there is the highest point of our graph at slightly less than five-hundred miles where roughly .1% of all flights are the same value there. Moving away from the right side of our graph we can see a very narrow purple line extend all the way down to over five-thousand miles. Therefore, because people do not fly when they can drive short distances, but also people fly extremely far distances because they have to, our data will be skewed to the right. While not immediately important, these kinds of trends and patterns can be critical in building our models.

I will build four different models (Logistic Regression, Naive Bayes, Random Forest, KNN). Using these models I will take the highest accuracy rating and work towards using HyperParameter Tuning to increase our overall accuracy:



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 which will be my next step in completing this capstone.