Capstone 2 Final Presentation

Cody Schwarz

Table of Contents

1) Introduction
2) The Data
3) The Data
4) Refining the Data
5) Distance Distribution
6) Day of Week
7) Time of Day
8) Unique Destination Location
9) Model vs Accuracy
10) HyperParameter Tuning
11) Conclusion

Introduction

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. Without the foresight to plan ahead and have another option you may be completely out of luck.

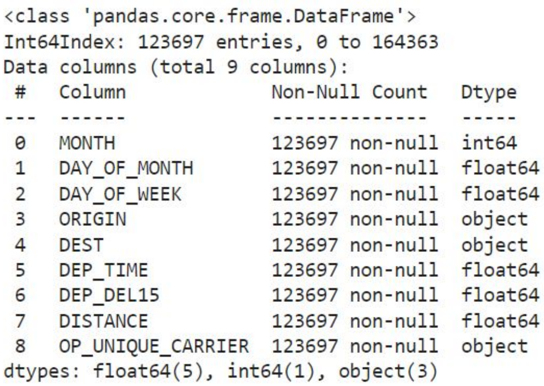
Imagine, however, if when you booked the flight, you were provided with a warning. ("Due to inclimate weather, the chances of this flight being delayed are 8%").

Now, one will be able to plan ahead and prepare for a delay or cancellation because it is at a greater likelihood.
There are many variables which can affect a flight being cancelled (weather, which airline it is, location, etc.)
and we will build a machine learning model using these factors to predict when a flight will be cancelled.

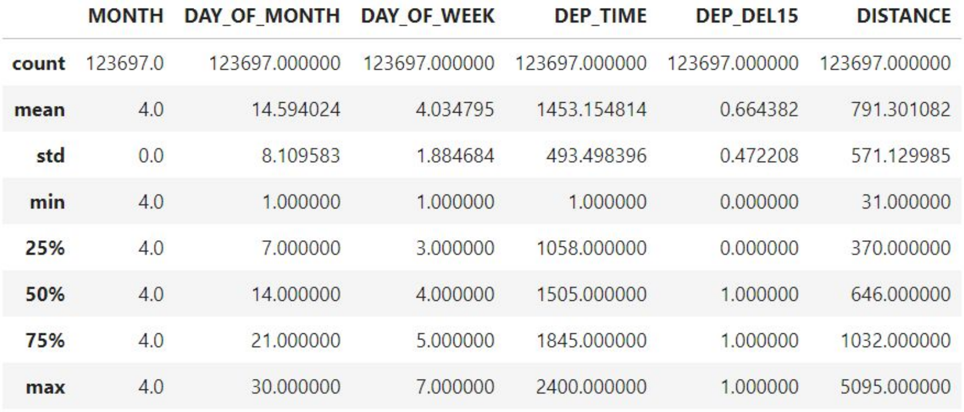
The Data

I decided to use data straight from the Bureau of Transportation Statistics because it is extremely good data and very clean. 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.

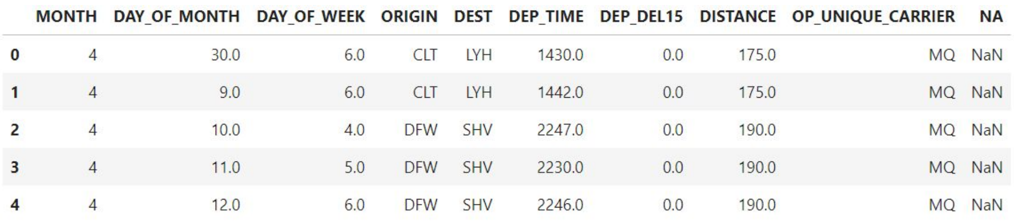
The Data



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 123697 entries, 0 to 164363
Data columns (total 9 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   MONTH              123697 non-null  int64
 1   DAY_OF_MONTH       123697 non-null  float64
 2   DAY_OF_WEEK        123697 non-null  float64
 3   ORIGIN             123697 non-null  object
 4   DEST               123697 non-null  object
 5   DEP_TIME           123697 non-null  float64
 6   DEP_DEL15          123697 non-null  float64
 7   DISTANCE           123697 non-null  float64
 8   OP_UNIQUE_CARRIER  123697 non-null  object
dtypes: float64(5), int64(1), object(3)
```



	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



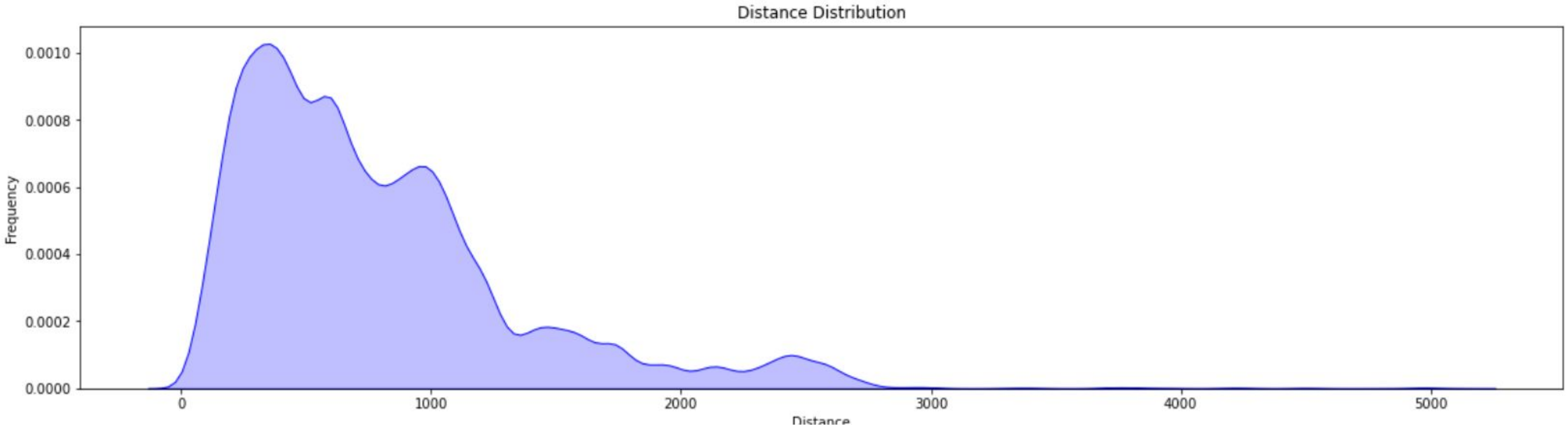
	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	ORIGIN	DEST	DEP_TIME	DEP_DEL15	DISTANCE	OP_UNIQUE_CARRIER	NA
0	4	30.0	6.0	CLT	LYH	1430.0	0.0	175.0	MQ	NaN
1	4	9.0	6.0	CLT	LYH	1442.0	0.0	175.0	MQ	NaN
2	4	10.0	4.0	DFW	SHV	2247.0	0.0	190.0	MQ	NaN
3	4	11.0	5.0	DFW	SHV	2230.0	0.0	190.0	MQ	NaN
4	4	12.0	6.0	DFW	SHV	2246.0	0.0	190.0	MQ	NaN

Refining the Data

There was not much cleaning needed for this data set as it came directly from the Bureau of Transportation Statistics and was very clean. The only thing needed was to fix up some columns so that they were uniform across the board for my prediction. As well as removing some unnecessary values.

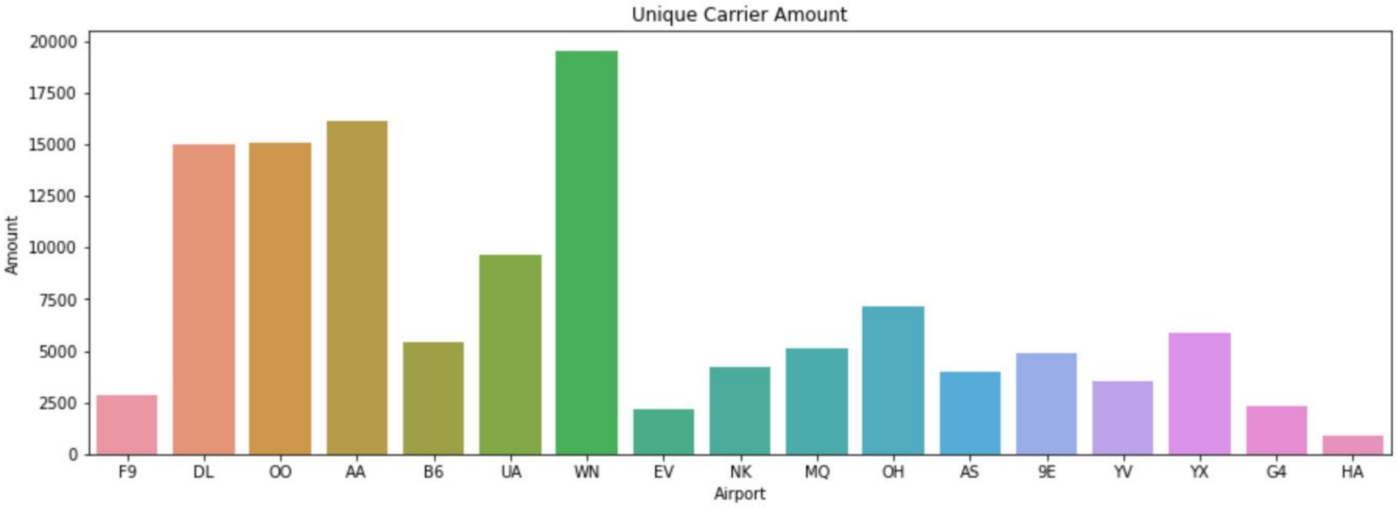
Distance Distribution

This is the frequency of flights based on distance travelled on that flight. As we can see the average flight is under one thousand miles, with over .1% of flights being roughly 500 miles. Most people do not fly for little miles, they will just drive instead. Also, most people won’t take long flights. For many reasons, but as can be seen after one thousand miles there is a steep drop off. While there exist flights ranging up to 5000 miles.



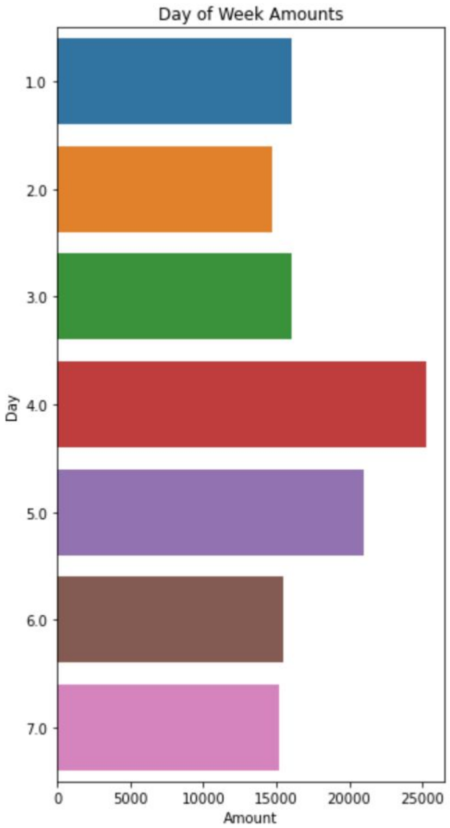
Unique Carrier Amount

This is just the amount of unique carrier trips for every airport. While nothing to touch on immediately, it can prove very useful in identifying patterns with predicting flight delays.



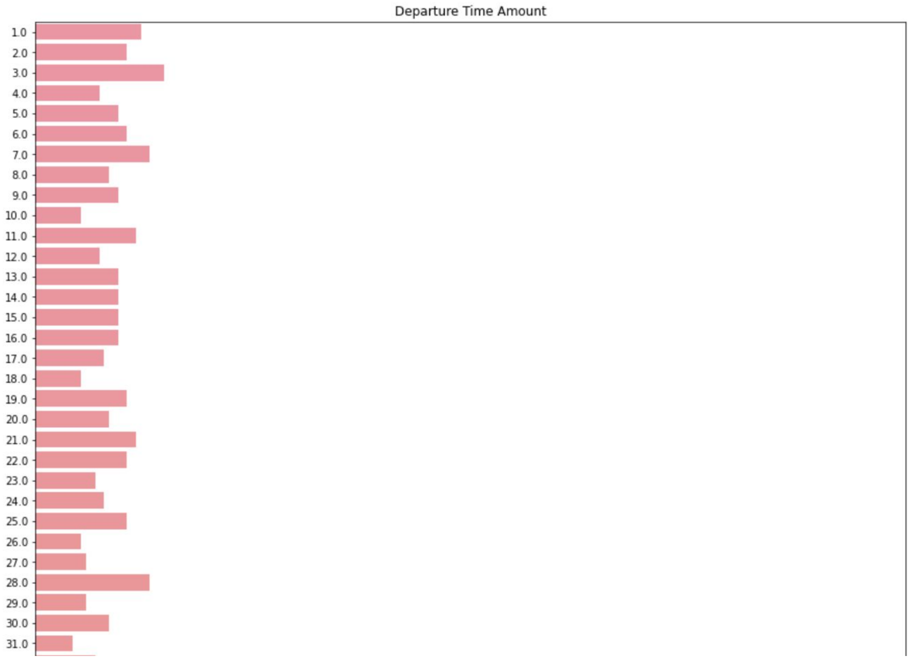
Day of Week

This is the day of the week in which flights leave a certain location. These graphs will be very useful in identifying patterns later on. As we can see, Thursday and Friday are the most frequently traveled days, as Tuesday is the least.



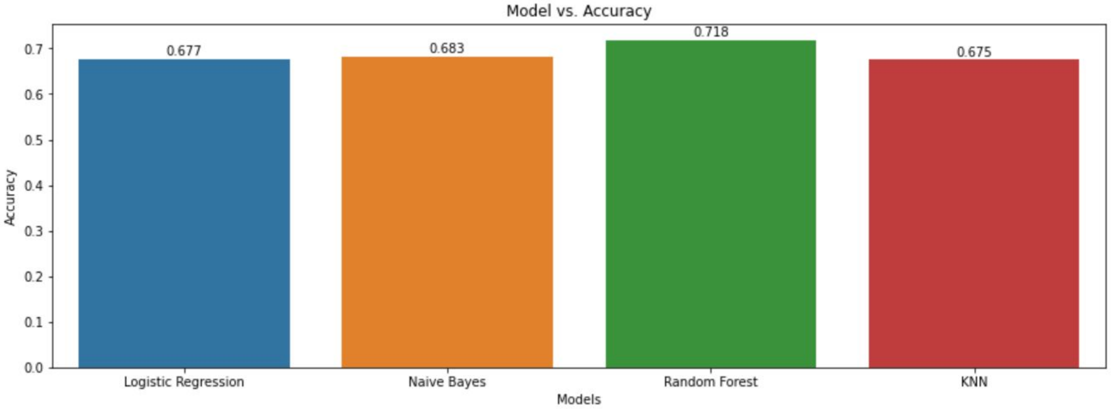
Time of Day

Same as Day of the Week, just when these flights end up taking off. Using these two charts in tandem may provide clues as to what are the likely factors in flight delays. This chart details down to every minute of every day to give us a complete look at the data.



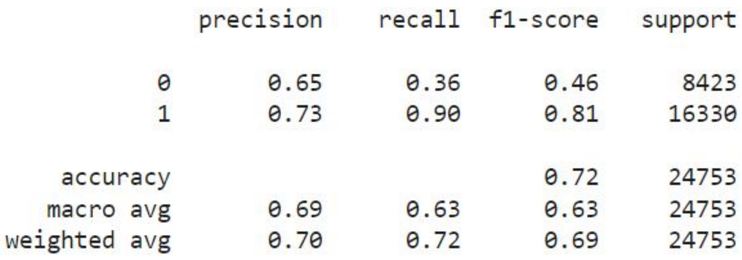
Model vs Accuracy

After running our models we have roughly similar accuracies for all the models. Obviously the Random Forest is the highest accuracy of almost 72 percent, while the KNN and Logistic Regression models provided the lowest accuracy near 67.5 percent. None of these, however, are adequate accuracy rating and we will hope to improve them with our Hyperparameter Tuning.



HyperParameter Tuning

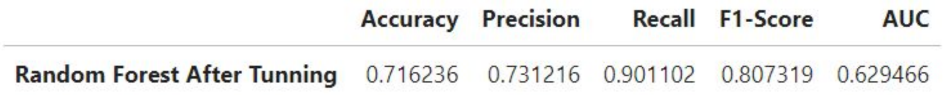
As we can see from our HyperParameter Tuning, our Accuracy did not improve. With further optimizations I do believe it could improve, but as of now it is very difficult to increase the accuracy. However, with a F1-Score of over .8 there is clearly potential with this model. It is performing enough to limit false positives and false negatives which is always good.



```
              precision    recall  f1-score   support

           0       0.65      0.36      0.46      8423
           1       0.73      0.90      0.81     16330

    accuracy                           0.72     24753
   macro avg       0.69      0.63      0.63     24753
weighted avg       0.70      0.72      0.69     24753
```



	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest After Tunning	0.716236	0.731216	0.901102	0.807319	0.629466

Conclusion

This is a difficult project without a doubt. The sheer size of the data is scary enough, but overall I am happy with the progress made. I did not expect to receive a high accuracy score on any of the models considering the difficult task at hand, so to have some success is clearly good progress. There are just a lot of factors that can determine flight delays or cancellations and one of the biggest factors of that is the weather. Weather is extremely difficult to predict and thus leads to some inaccuracies in the predictions. I would love to continue working with this data and the hyperparameter tuning to increase the accuracy score.