Predicting Airline Delays (USA Domestic Flights)

1. Abstract

This project was to identify factors that could allow it to predict airline delays. Two different models were used: Logistic Regression and Random Forest Classifier, with the Random Forest Classifier proving to be better with this dataset. The project manged to identify several factors to help avoid delays. These factors are: (1) Flying in the day, (2) Flying in the 1st half of the month, (3) Flying in the middle of the week, (4) Avoiding Frontier Airlines (F9), Southwest Airlines (WN), American Eagle Airlines (MQ), JetBlue Airways (B6).

In the predictive modelling phase, the two models yielded very different results. In the Logistic Regression model, the confusion matrix, area under the curve and classification report produced a low accuracy score (64%) for the model. This would be barely better than an educated guess. However, when the Random Forest Classifier was used, results improved to an average 91% for the confusion matrix, area under curve, and classification report.

2. Problem Statement and Project Purpose

Onion Travel Agency (OTA) is a new player to the US market and wants to carve a name for themselves by ensuring customers encounter minimal delays when booking flights with them. As such, OTA wants to be able to predict flight delays to make the most convenient bookings for their customers. As the end-of-year holiday season is key, OTA's focus is on the months of November and December. The project aims to be able to identify the factors most related to delays and help customers avoid potential delays to their flight.

3. Dataset Details

Data used was taken from the public Bureau of Transportation website[1]. Data was for the months of November and December, from years 2014 to 2017. In its original format, the chosen datasets had a total of 31 features and 3,716,693 rows. The features and their descriptions can be found on their website[2].

With the data available, the project focused on delays at departure as the target as airlines have been known to "inflate" their arrival times in order to avoid payouts for delays[3]. Only delays of more than 15 mins are considered as flight delays of less than 15 mins are not reported in OPSNET[4].

4. Data Usability for the Project

The first step for the project was to look into the usability of the data. This comes in a few steps. (1) Are there variations to the data in the features? (2) Are there too many null values in the data?

---

[1] Data available from Bureau of Transportation, https://www.transtats.bts.gov/Tables.asp?DB_ID=120
[2] Feature description for the data used can be found at: https://www.transtats.bts.gov/Fields.asp
[3] "Airlines extending flight times to avoid payouts for delays, report claims", The Telegraph, Hugh Morris and Oliver Smith, 28 August 2018. https://www.telegraph.co.uk/travel/travel-truths/Are-airlines-exaggerating-flight-times-so-theyre-never-late/
[4] Types of delay, Federal Aviation Administration (USA) http://aspmhelp.faa.gov/index.php/Types_of_Delay

1) Are there enough variations to the data in the features?

Figure 1 shows the number of unique values in each of the features. There seems to be a variety of unique values in the features which should allow analysis for the project.

YEAR	4	CRS_ARR_TIME	1438
MONTH	2	ARR_TIME	1440
DAY_OF_MONTH	31	ARR_DELAY	1327
DAY_OF_WEEK	7	CANCELLED	2
UNIQUE_CARRIER	15	CANCELLATION_CODE	4
TAIL_NUM	5814	DIVERTED	2
FL_NUM	7078	CRS_ELAPSED_TIME	562
ORIGIN	326	ACTUAL_ELAPSED_TIME	717
ORIGIN_CITY_NAME	322	AIR_TIME	690
DEST	325	DISTANCE	1432
DEST_CITY_NAME	321	CARRIER_DELAY	1122
CRS_DEP_TIME	1326	WEATHER_DELAY	694
DEP_TIME	1440	NAS_DELAY	576
DEP_DELAY	1299	SECURITY_DELAY	182
TAXI_OUT	174	LATE_AIRCRAFT_DELAY	683
TAXI_IN	244	DELAYED	2
		dtype: int64	

Figure 1: Number of unique values in the features

2) Are there too many null values in the data?

In figure 2 below, there are some null values in important features such as departure time (DEP_TIME). These rows are deleted as they will not be useful to our analysis.

YEAR	0	CRS_ARR_TIME	0
MONTH	0	ARR_TIME	40871
DAY_OF_MONTH	0	ARR_DELAY	46962
DAY_OF_WEEK	0	CANCELLED	0
UNIQUE_CARRIER	0	CANCELLATION_CODE	3677479
TAIL_NUM	4516	DIVERTED	0
FL_NUM	0	CRS_ELAPSED_TIME	1
ORIGIN	0	ACTUAL_ELAPSED_TIME	46962
ORIGIN_CITY_NAME	0	AIR_TIME	46962
DEST	0	DISTANCE	0
DEST_CITY_NAME	0	CARRIER_DELAY	3055466
CRS_DEP_TIME	0	WEATHER_DELAY	3055466
DEP_TIME	37516	NAS_DELAY	3055466
DEP_DELAY	37520	SECURITY_DELAY	3055466
TAXI_OUT	38688	LATE_AIRCRAFT_DELAY	3055466
TAXI_IN	40871	DELAYED	0
		dtype: int64	

Figure 2: Null values in the data

After deletion of rows with null values, the project is left with 98.7% of the available data, accounting for 3,716,693 rows.

5. Exploratory Data Analysis (EDA)

In this phase of the analysis, the project commences analysis on on-time performance in the identified months. A new column was introduced as the target. This is to be the “DELAYED” column, where departures delayed for more than 15 minutes would be a “1”, and others, “0”.

Once the target column is added, the analysis then focused on (1) Does the day of the week or month contribute to delays? (2) On-time performance of the airlines, (3) Does origin affect delays?, (5) Does flight time/distance contribute to delays? (6) Is departure time a factor? (7) Correlation of features.

1) Closure and Merger of Airlines

Before starting EDA, it was found out that one of the airlines in the data had closed down, and another two had merged. These were first amended to stay current with today's carrier status. This change reduced the number of rows to 3,662,868, which is still more than sufficient.

2) Day of Week Analysis

As the week goes by, the percentage of delays seem to drop steadily, until Sunday, where it increases again.

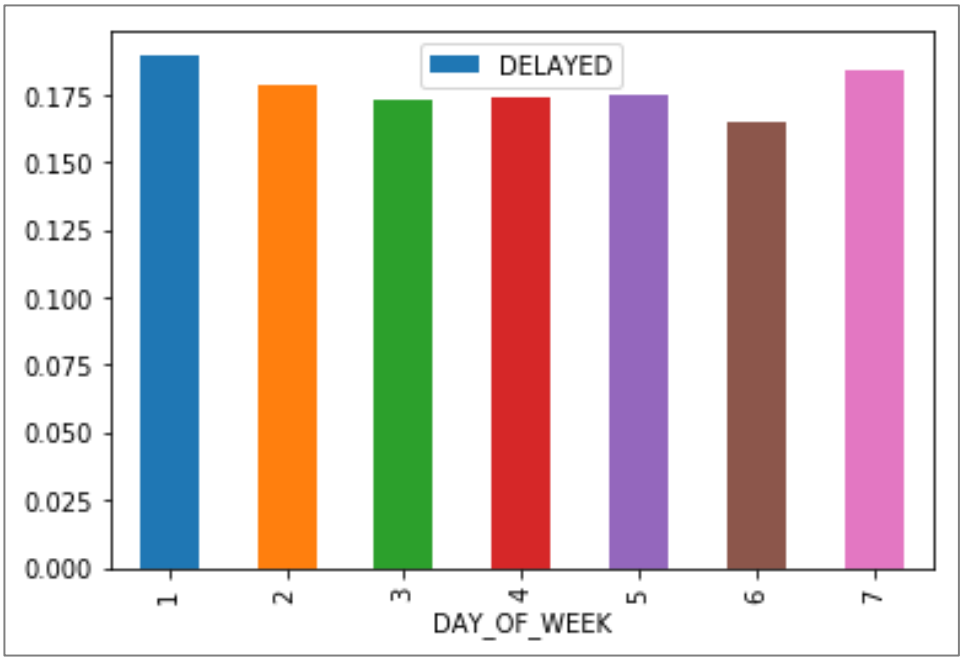

Figure 3: Percentage of delays for each day of the week.

However, this could possibly be due to the number of flights on each day? The number of flights on each day was measured. It was found that Sunday was not even the busiest day of the week. In fact, the busiest day was Wednesday, the day with the second lowest percentage of delayed departure flights.

3) Day of Month Analysis

In the second half of the month, there is an increase in the amount of delays. Figure 4 tells us that the beginning of the month would be a better time fly, with only around 13% of flights being delayed before the 15th of the month. After the 15th, there was more than 15% of flights delayed almost every day (except for the 24th and 25th).

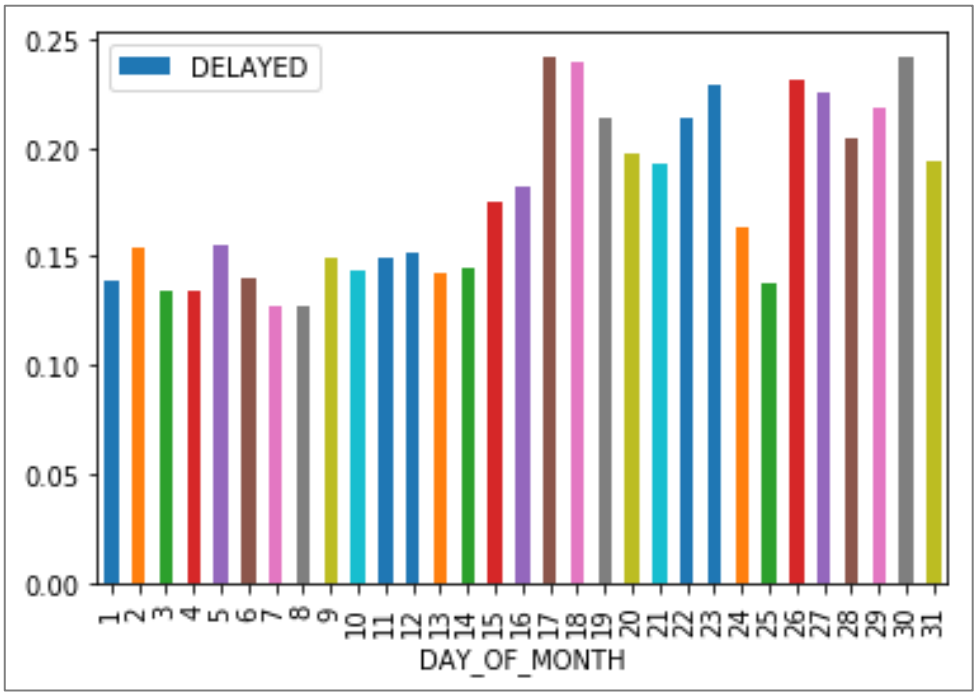

Figure 4: Percentage of delays for each day of the month (Nov & Dec).

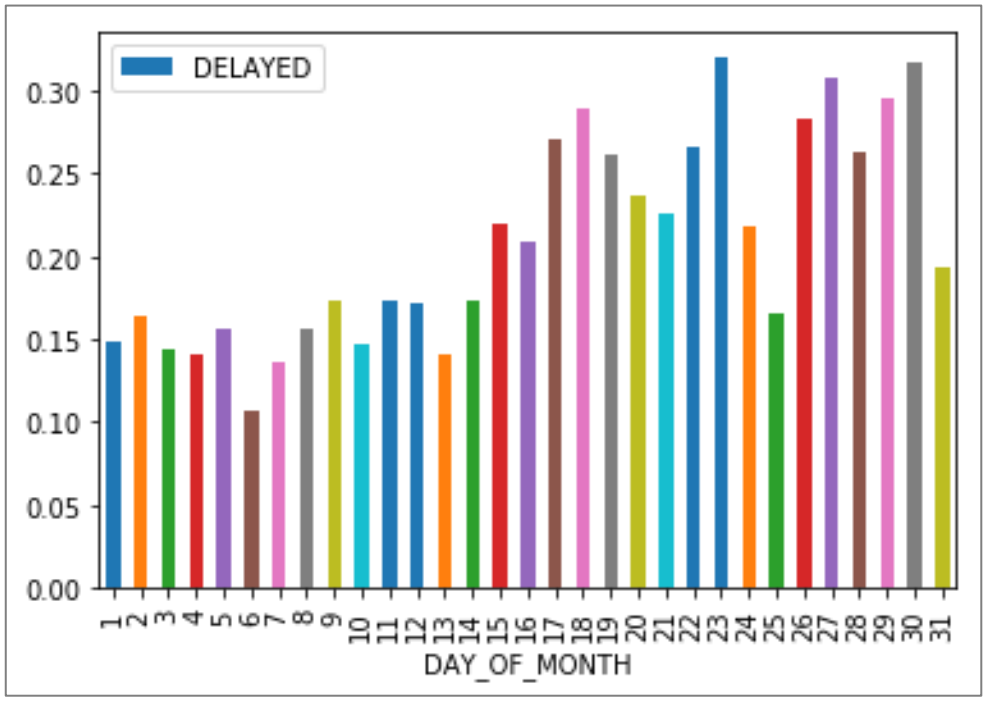

Figure 5: Percentage of delays for each day of the month (Dec only).

This spike was probably caused by the Christmas season as in December, there was a general increase in delays from 15th to 23rd December followed by a drop on Christmas eve and Christmas day. This trend is also seen for New Year's Day.

4) <u>On-Time Performance of Airlines</u>

The following airlines had the highest percentage of flight delays: Frontier Airlines (F9), Southwest Airlines (WN), American Eagle Airlines (MQ), JetBlue Airways

(B6). However, while American Airlines (AA) has a reasonable delay rate, they were responsible for the longest delays (Fig 6).

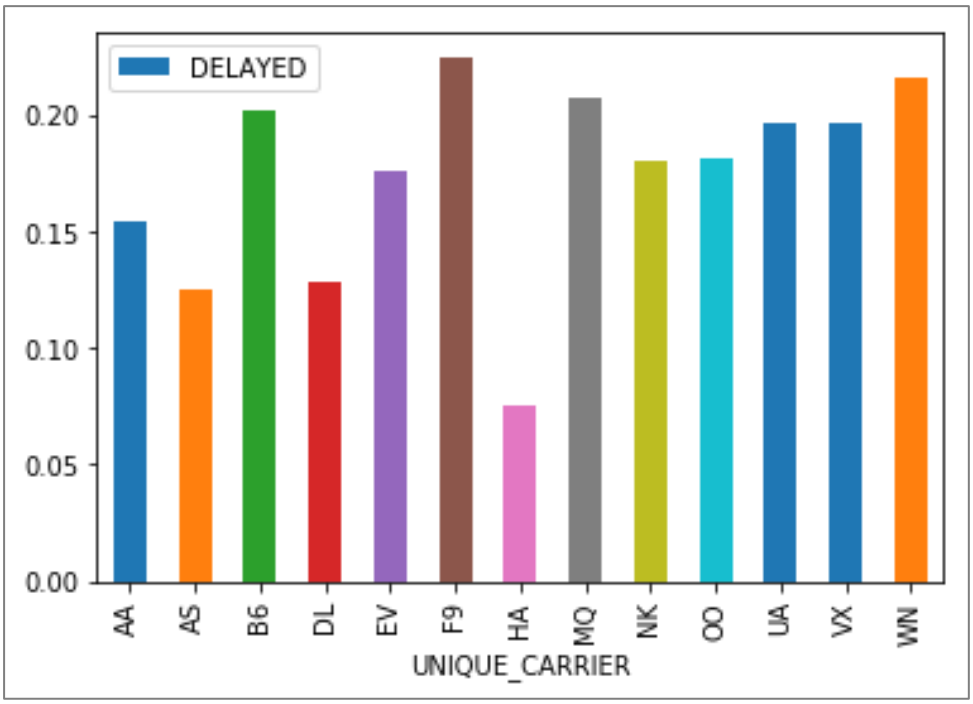

Figure 6: Percentage of delays for each carrier.

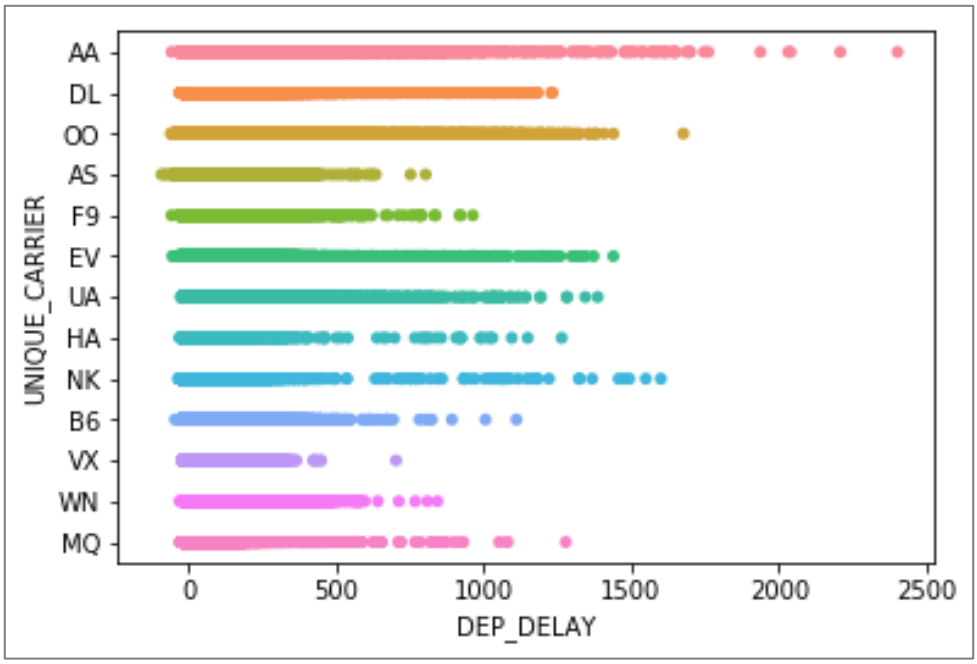

Figure 7: Delay minutes for each carrier.

5) Origin

Airports in Chicago, Atlanta, Denver, Los Angeles and Dallas had the highest number of delays. With the exception of Atlanta, the other airports had 20% of their flights delayed.

ORIGIN	DELAYED	% DELAYED
ORD (O'Hare International Airport)	36,450	20.9%
ATL (Hartsfield-Jackson Atlanta International)	35,072	15.0%
DEN (Denver International Airport)	30,145	21.5%
LAX (Los Angeles International Airport)	29,711	21.7%
DFW (Dallas/Fort Worth International Airport)	29,309	20.6%

Figure 8: Origin with the highest amount of delays.

6) Departure Time

There is a high possibility (>50%) of a delay after midnight, to around 4 am (0400 hours). To avoid delays, flights in the morning (5am to 1pm) would be preferred as less than 20% of these flights are delayed.

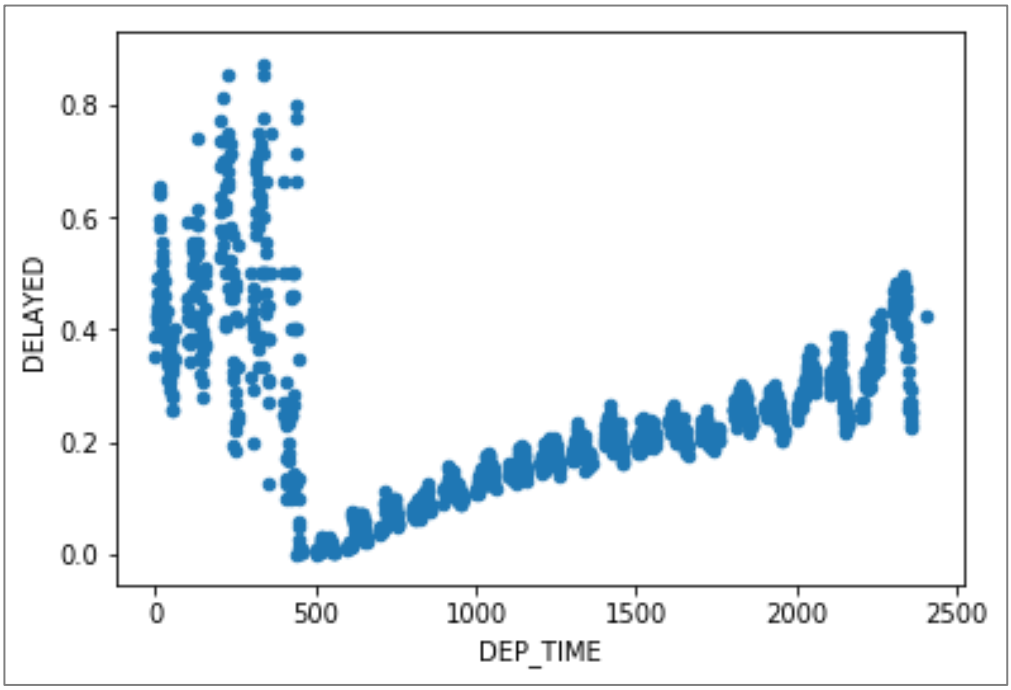

Figure 9: Percentage of flights delayed by time.

7) Flight Distance

Longer flights do not correlate to a higher percentage of delays.

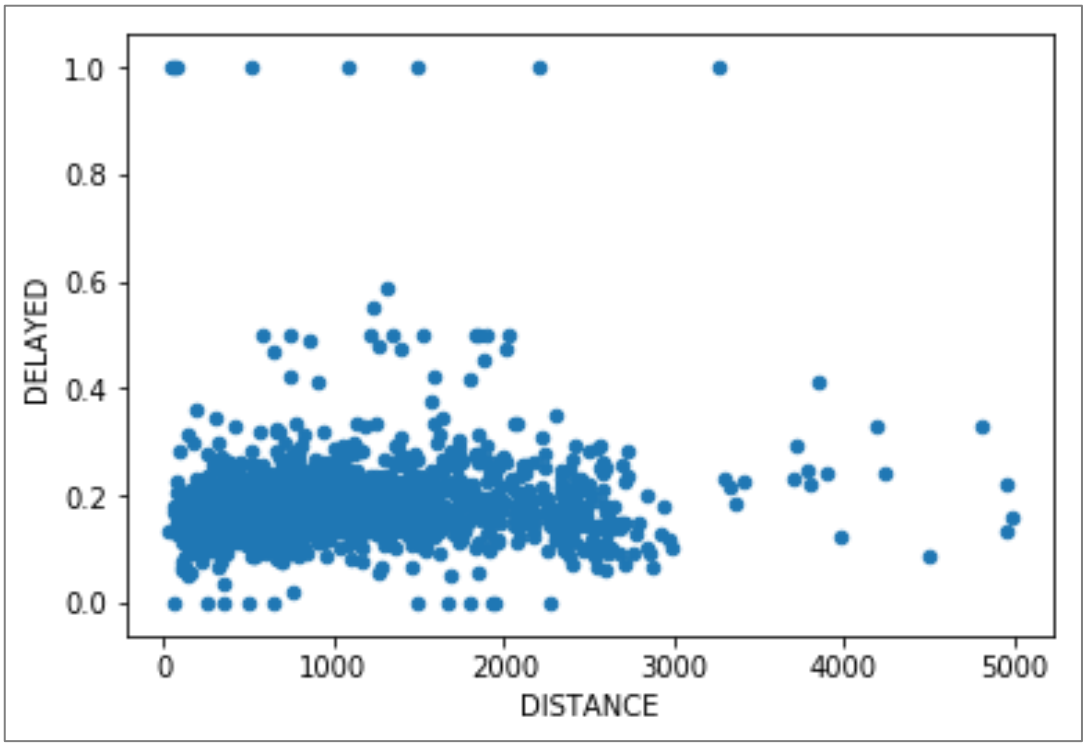

Figure 10: Percentage of flights delayed by distance.

6. Feature Selection

A correlation was carried out to identify collinearity among features. The first round of correlation found the following features to be highly correlated. Some features were removed to prevent biasness.

- CRS_DEP_TIME, DEP_TIME
- CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME, AIR_TIME, DISTANCE
- CRS_ARR_TIME, ARR_TIME

As the project aims to analyze departure delays, the following features are also removed:

- CANCELLED, CANCELLATION_CODE, DIVERTED
- FL_NUM, TAXI_IN, TAXI_OUT

After removing features that are not needed, or are highly correlated, the correlation matrix is shown below:

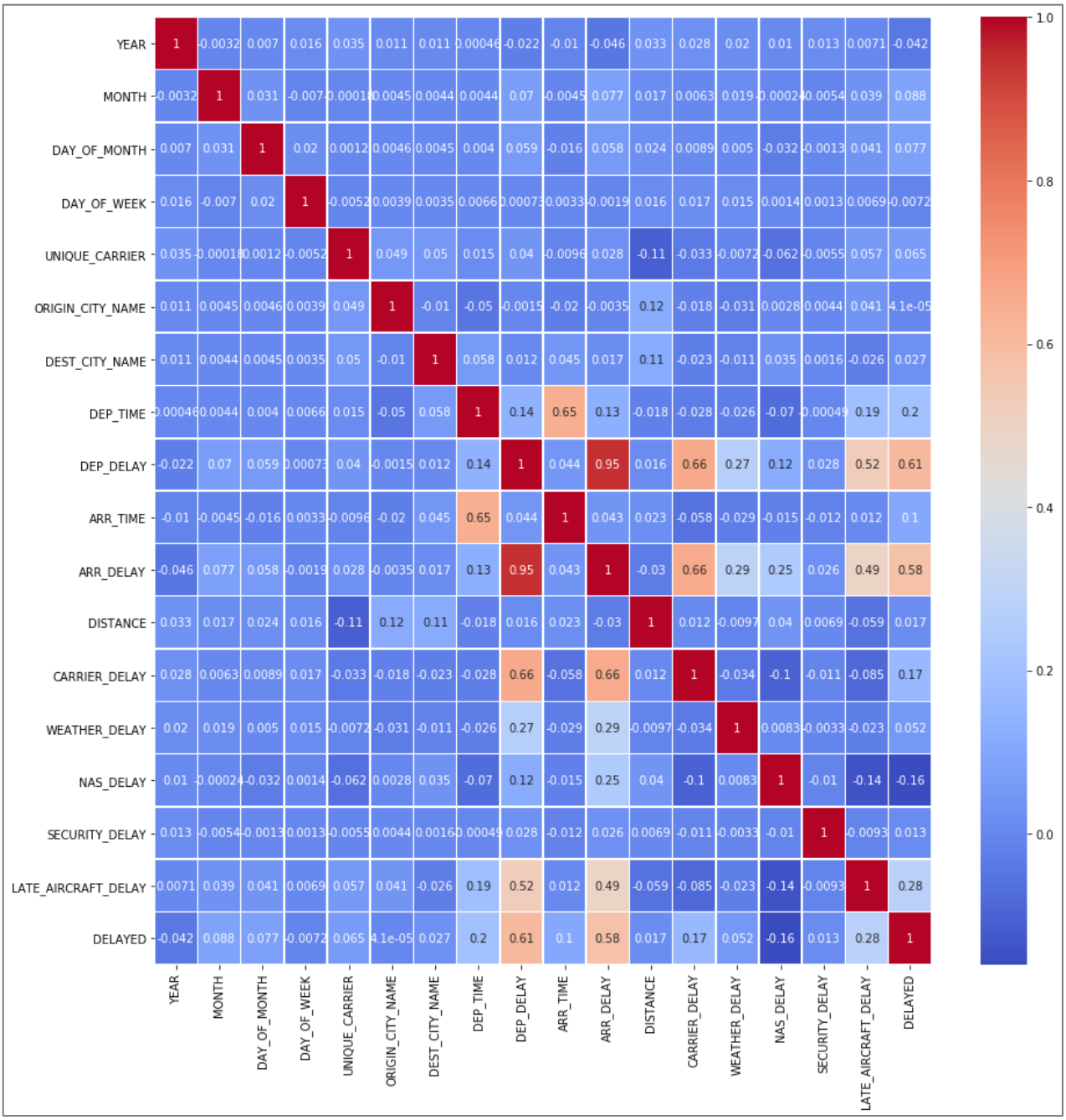

Figure 11: Correlation matrix

From the matrix, the following could be deduced:

- An arrival delay could cause a delay
- A departure delay could end in a delayed arrival

Other than this, there is no strong correlation between the target "DELAYED" and other features. This could cause issues when attempting a predictive model in the next section.

Before proceeding to model a predictive analysis, categorical features will be encoded to allow for predictive modelling to be carried out. The following categories are encoded:

- UNIQUE_CARRIER
- ORIGIN_CITY_NAME
- DEST_CITY_NAME

7. Predictive Modelling

Using the target column, we find that there is a large class imbalance between delayed flights, and non-delayed flights. Only 17.5% of flights were delayed. With such a highly imbalanced dataset, there was a need to apply SMOTETomek to even out the imbalance.

Model 1: Logistic Regression

1) **Logistic Regression** was used as the target variable is dichotomous. In this instance, a "1" or "0". The following steps were carried out:

1) As mentioned, SMOTETomek had to be carried out to address the issue with the high class imbalance.

```
Original y dataset shape Counter({0: 3013882, 1: 648986})
Resampled y dataset shape Counter({0: 2954679, 1: 2954679})
```

Figure 12: Applying SMOTETomek to address class imbalance

After applying SMOTETomek to the dataset, there was a slight under-sampling for the "0" class (3,013,882 to 2,954,679), and a large over-sampling for the "1" (delayed) class (648,986 to 2,954,679).

2) Scaling of features using "StandardScaler()".

3) Selecting a test size for the train_test_split. An arbitrary test size of 50% was chosen. The assumption was that with the large dataset size, there would still be a reasonable training data size for the logistic regression modelling.

4) Applying of Logistic Regression model. The model below was used to perform the logistic regression on the dataset.

```
LogisticRegression(C=20, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

Figure 13: Logistic Regression model applied to dataset

5) Checking of model accuracy. Several methods of checking for model accuracy were used.
   - Confusion Matrix.

	predicted		
actual		Negative	Positive
	Negative	TN: 931,380	FP: 546874
	Positive	FN: 523779	TP: 952,646

Figure 14: Results from the confusion matrix

The results from the confusion matrix showed that only 64% of the predictions were correct. Even worse, 18% of the predictions were false negatives. This

means that the model has predicted that the flight would not be delayed, as opposed to the actual flight which was delayed.

- Area under the curve.

The graph below shows the results from the "auc_roc" score. The model was only given a 0.64 which was a consistent figured when compared to the confusion matrix. With a maximum (ideal) score of 1, this result is barely better than a coin toss.

Given more time and familiarity to python and sklearn, I might have been able to identify a method to adjust the threshold of the curve. That could have given me a slightly better accuracy to the model and reduce the amount of false negatives.

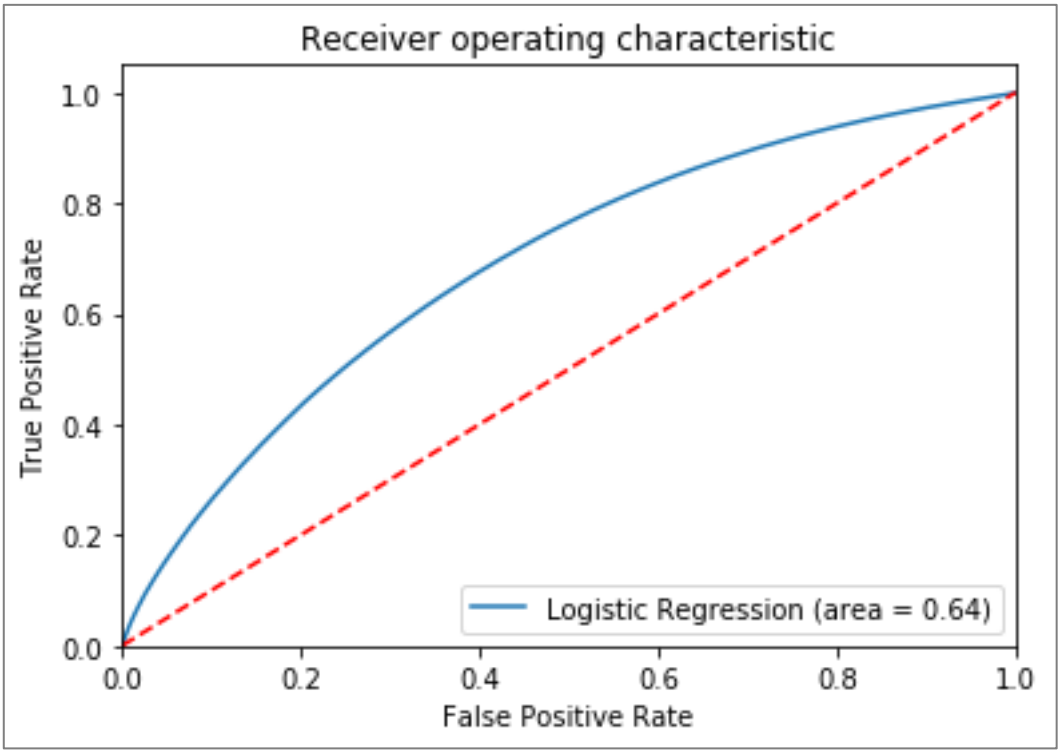

Figure 15: Results from the auc_roc score

- Classification Report
  The classification report gives a precision score of 64%, which means that the model has predicted the correct result 64% of the time. Incidentally, the amount of true positive (recall score) is also given as 64%.

	precision	recall	f1-score	support
0	0.64	0.63	0.64	1478254
1	0.64	0.65	0.64	1476425
avg / total	0.64	0.64	0.64	2954679

Figure 16: Results from the classification report

Model 2: Random Forest Classifier

1) **Random Forest Classifier** was the second method to be used. It is used as it was the capability to be used for both classification and regression problems. The following steps were carried out:

2) As mentioned, SMOTETomek had to be carried out to address the issue with the high class imbalance.

```
Original y dataset shape Counter({0: 3013882, 1: 648986})
Resampled y dataset shape Counter({0: 2954679, 1: 2954679})
```

Figure 17: Applying SMOTETomek to address class imbalance

After applying SMOTETomek to the dataset, there was a slight under-sampling for the "0" class (3,013,882 to 2,954,679), and a large over-sampling for the "1" (delayed) class (648,986 to 2,954,679).

3) Scaling of features using "StandardScaler()".

4) Selecting a test size for the train_test_split. An arbitrary test size of 50% was chosen. The assumption was that with the large dataset size, there would still be a reasonable training data size for the modelling to be accurate.

5) Applying of Random Forest Classifier model. The model below was used to perform the modelling on the dataset.

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

Figure 18: Random Forest Classifier model applied to dataset

6) Checking of model accuracy. Several methods of checking for model accuracy were used.
   - Accuracy Score

   Two accuracy scores were obtained. One for training, and another for testing. The training accuracy measured the y_train and the prediction from training data. Testing accuracy was measured by y_test and prediction for test data.

   Training Accuracy achieved: 0.99 (which is expected)
   Testing Accuracy achieved: 0.91

   - Confusion Matrix

   To confirm the results, we proceed to the confusion matrix. The matrix produced looks to be accuracy, with the largest figures in the diagonal (True Negative and True Positive).

	predicted		
actual		Negative	Positive
	Negative	TN: 1,426,146	FP: 52,108
	Positive	FN: 221,386	TP: 1,255,039

Figure 19: Results from the confusion matrix

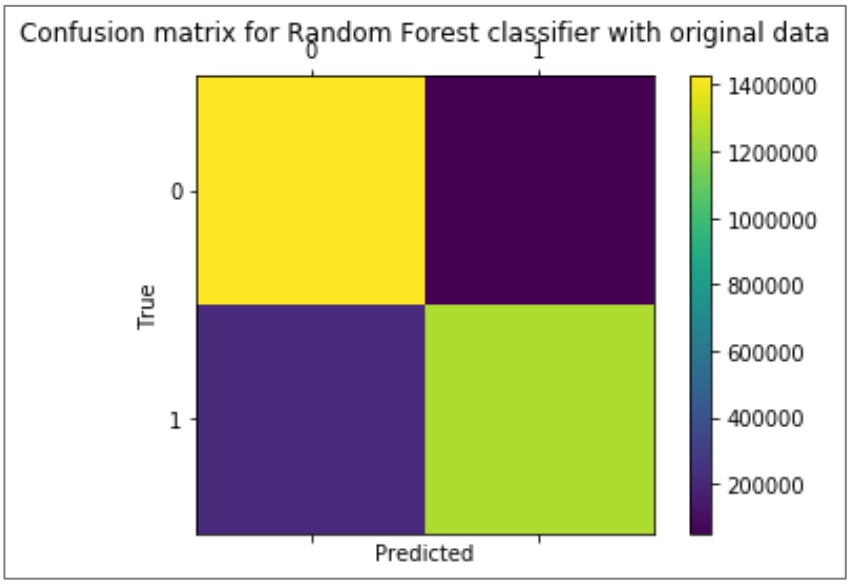

Figure 20: Results from the confusion matrix

The results from the confusion matrix showed that only 64% of the predictions were correct. Even worse, 18% of the predictions were false negatives. This means that the model has predicted that the flight would not be delayed, as opposed to the actual flight which was delayed.

We attempt to calculate accuracy, precision and recall from the figures obtained from the confusion matrix.

Accuracy = Total Correct Predictions/Total Predictions
= 2,681,185 / 2,954,679 = 90%
Precision = True Positive / (True Positive + False Positive)
= 1,255,039 / 1,476,425 = 85%
Recall = True Positive / (True Positive + False Negative)
= 1,255,039 / 1,476,425 = 85%

- Area under the curve.

The graph below shows the results from the "auc_roc" score. The model was only given a 0.64 which was a consistent figured when compared to the confusion matrix. With a maximum (ideal) score of 1, this result is barely better than a coin toss.

Given more time and familiarity to python and sklearn, I might have been able to identify a method to adjust the threshold of the curve. That could have

given me a slightly better accuracy to the model, and reduce the amount of false negatives.

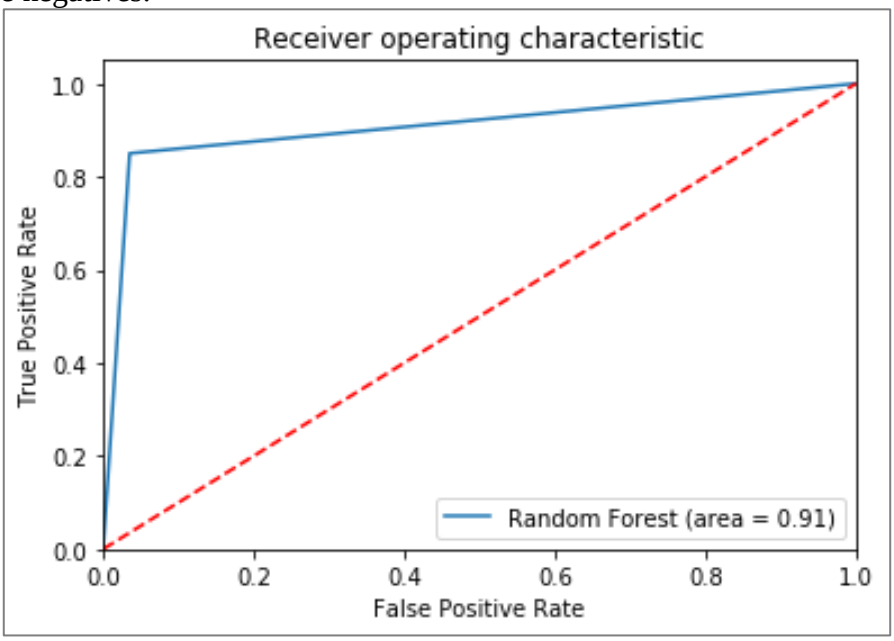

Figure 20: Results from the auc_roc score

- Classification Report
  The classification report gives a precision score of 64%, which means that the model has predicted the correct result 64% of the time. Incidentally, the amount of true positive (recall score) is also given as 64%.

	precision	recall	f1-score	support
0	0.87	0.96	0.91	1478254
1	0.96	0.85	0.90	1476425
avg / total	0.91	0.91	0.91	2954679

Figure 21: Results from the classification report

8. Results

The two predictive models gave very different results. In this particular dataset, the Random Forest Classifier provided a much better model as shown above.

We have also identified several ways that we can avoid delays:

1) Take a day flight
2) Avoid flights on Sundays to Tuesdays
3) Fly in the first half of the month
4) Avoid the following airlines: Frontier Airlines (F9), Southwest Airlines (WN), American Eagle Airlines (MQ), JetBlue Airways (B6).
5) For good measure, avoid American Airlines as well. While a delay is not as common, the delays could be worse.

9. Next Steps/Improvements

1) Dataset Improvement.

For future projects, I should be more aware of the dataset to be used. In this dataset, while there were enough unique values for the features, a predictive analysis dataset should have a better correlation with the intended target.
As seen in the correlation matrix earlier in the report, there were no features that were strongly correlated. This might have affected the accuracy of the predictive model.

2) Learn how to adjust and fine tune Logistic Regression and ROC threshold to provide a better model.

Due to a lack of time, I was unable to find a method to adjust the threshold for the auc_roc method. If I was able to do so, I might have been able to achieve a better accuracy in the model.

3) Making the Model Useful

As there is a model that seems to work (Random Tree Classifier), the next logical step would be to enable the model to provide you with a date/flight with less probability of a delay.