Flight Cancellation Prediction

Jyothi Chandrakanth

Eknath Chunduri

Pavan Kumar

College of Professional Studies, Northeastern University

ALY6110: Data Management and Big Data

Professor Janos Mako

December 19, 2021

Introduction

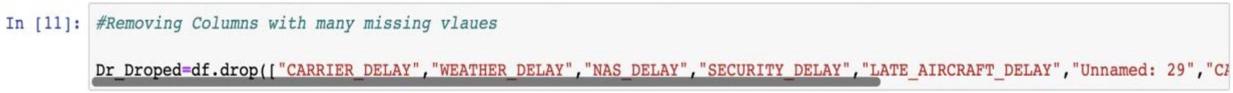
Flights spend less time on the road when commuting. Flight cancellations and delays are becoming a serious concern since they result in resource inefficiency, increased expenditures, and disruptions in passengers travel arrangements, all of which lead to consumer dissatisfaction. According to Lukacs, the US government lost 33.0 billion dollars in 2019 due to airline cancellations and delays (2019). Cancelled flights and delays are caused by a variety of circumstances including weather, security, and technical faults with the aircraft. If airlines do not provide quality service, that has an impact on their image in the aviation business. To tackle this, we'll use the "flights.csv" dataset to discover what variables influence flight cancellations. Technology is radically altering how organizations communicate with their consumers, make business decisions, and create workflows. Airlines are transforming their activities from pre-flight through post-flight, including booking, extra legroom, luggage, boarding, and transportation services, with the use of data. Big data analytics' ultimate benefits include quicker reactions to current and future market needs, enhanced planning and closely aligned decision making, and crystal-clear grasp and monitoring among all major performance drivers relevant to the aviation business. Unplanned maintenance accounts for over 30% of total time delay; airlines suffer large expenditures owing to delays and cancellations, which include maintenance costs and compensation for travelers detained at airports.

In Data Analysis, Machine Learning, and Data Science, cluster analysis has become one of the most essential methodologies. Clustering is a technique that divides a collection of items with similar characteristics into groupings called clusters. Since the dataset which we have is imbalanced in nature we will be utilizing Random Forest and Decision Tree, classification algorithms to predict the factors affecting the cancellation of flights.

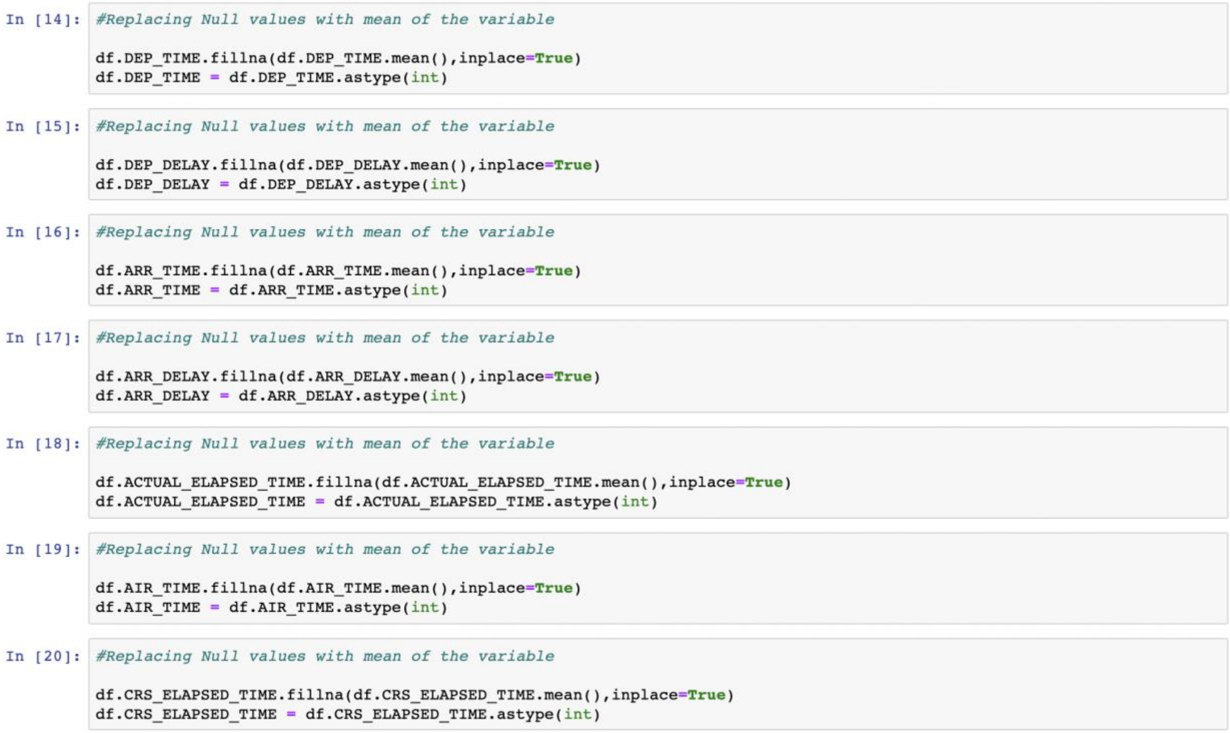
Exploratory Data Analysis

1) Data Cleaning

During the data cleaning procedure, we identified all variables with missing values and the percentage associated with them. The variables with more than 25% missing values were eliminated, while the ones with fewer missing values were imputed with the mean of the variable.



```
In [11]: #Removing Columns with many missing vlaues

         Dr_Droped=df.drop(["CARRIER_DELAY","WEATHER_DELAY","NAS_DELAY","SECURITY_DELAY","LATE_AIRCRAFT_DELAY","Unnamed: 29","CA
```



```
In [14]: #Replacing Null values with mean of the variable

         df.DEP_TIME.fillna(df.DEP_TIME.mean(),inplace=True)
         df.DEP_TIME = df.DEP_TIME.astype(int)

In [15]: #Replacing Null values with mean of the variable

         df.DEP_DELAY.fillna(df.DEP_DELAY.mean(),inplace=True)
         df.DEP_DELAY = df.DEP_DELAY.astype(int)

In [16]: #Replacing Null values with mean of the variable

         df.ARR_TIME.fillna(df.ARR_TIME.mean(),inplace=True)
         df.ARR_TIME = df.ARR_TIME.astype(int)

In [17]: #Replacing Null values with mean of the variable

         df.ARR_DELAY.fillna(df.ARR_DELAY.mean(),inplace=True)
         df.ARR_DELAY = df.ARR_DELAY.astype(int)

In [18]: #Replacing Null values with mean of the variable

         df.ACTUAL_ELAPSED_TIME.fillna(df.ACTUAL_ELAPSED_TIME.mean(),inplace=True)
         df.ACTUAL_ELAPSED_TIME = df.ACTUAL_ELAPSED_TIME.astype(int)

In [19]: #Replacing Null values with mean of the variable

         df.AIR_TIME.fillna(df.AIR_TIME.mean(),inplace=True)
         df.AIR_TIME = df.AIR_TIME.astype(int)

In [20]: #Replacing Null values with mean of the variable

         df.CRS_ELAPSED_TIME.fillna(df.CRS_ELAPSED_TIME.mean(),inplace=True)
         df.CRS_ELAPSED_TIME = df.CRS_ELAPSED_TIME.astype(int)
```

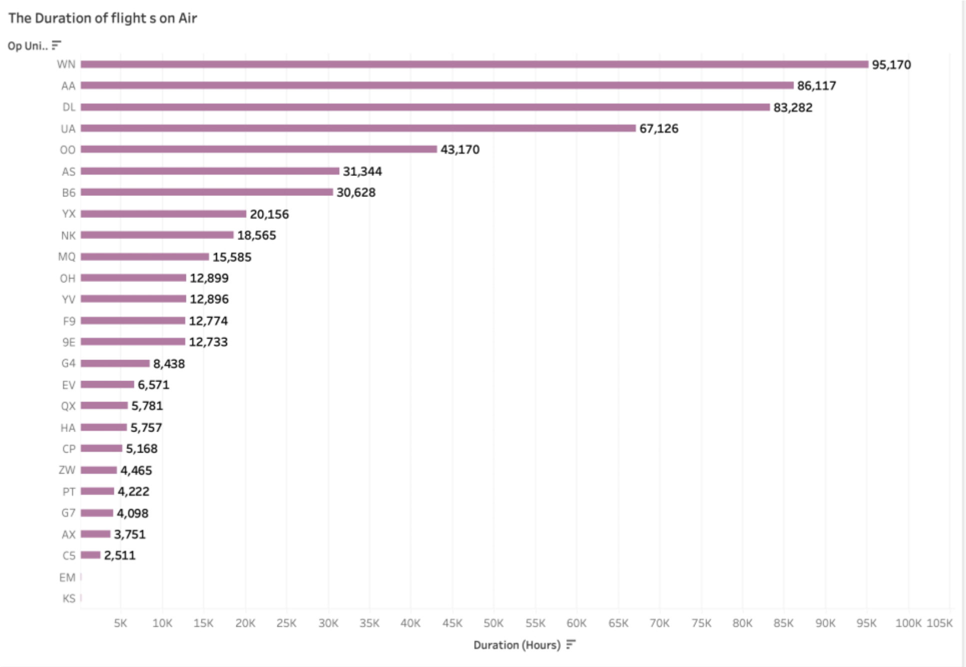
We are trying to answer the below questions to achieve the goal of the assignment.

1) The trends of cancellation based on day of the week and the month
2) Delays Caused due to various factors

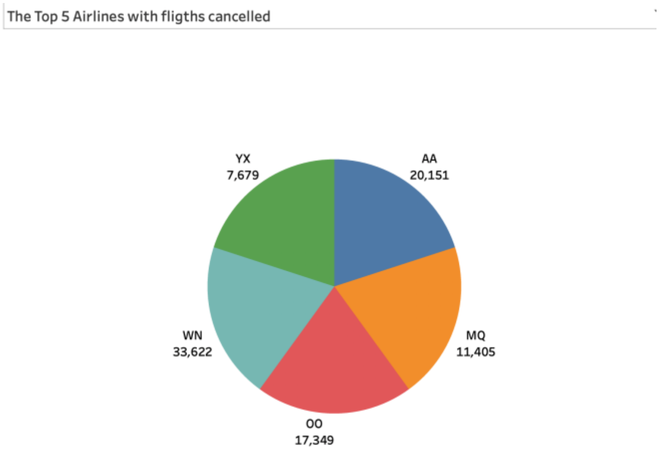
To perform Exploratory Data Analysis, we have extensively used Tableau and produced the below graphs to address the objective of the project.

We performed initial analysis that shows the delays and the cancelation trends

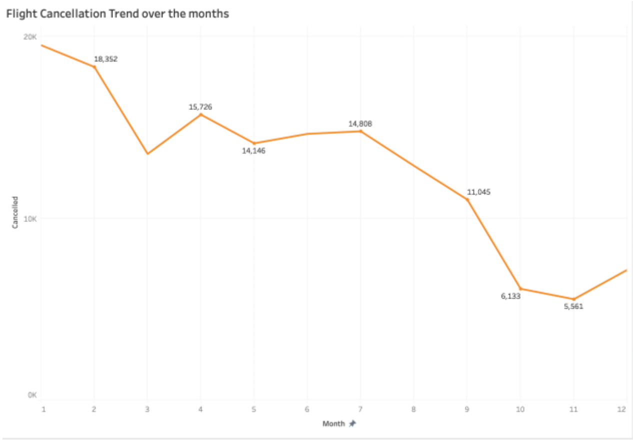
Determining the airtime of each flight



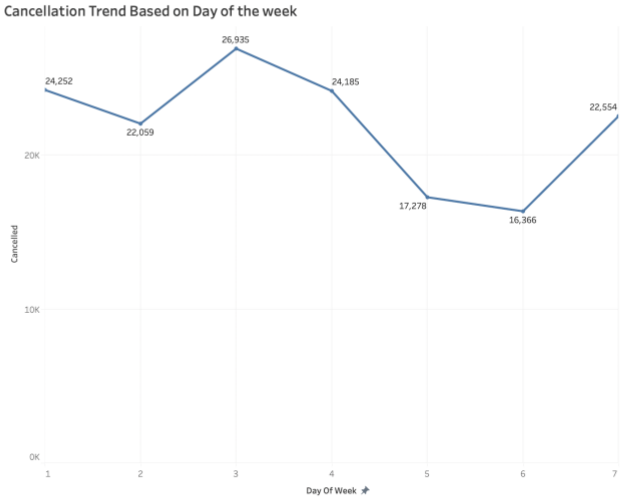
In the above graph, we have tried to identify the airtime for each flight to and converted them into hours for better understanding. Southwest (WN) has the highest airtime followed by American Airlines (AA) and Delta Airlines (DL). The least amount of airtime was found for Empire Airlines (EM) and Peninsula Airways (KS).



The above graph shows the top five flights with highest number of cancellation flights Southwest (WN) and American Airlines (AA) have the highest airtime and the highest cancellation of flights. Although Envoy Air (MQ) has flown for lesser duration when compared the Southwest and American Airlines.

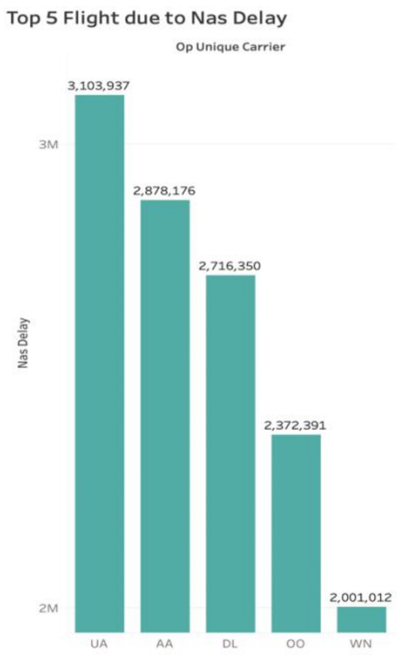


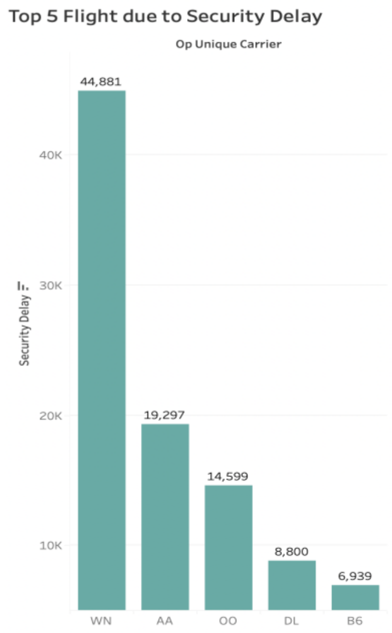
The above graph shows the cancellation trend for each month, The cancellation is high at the beginning of the year. January, February and April have the greatest number of cancellation, October and November months have the least cancellation.



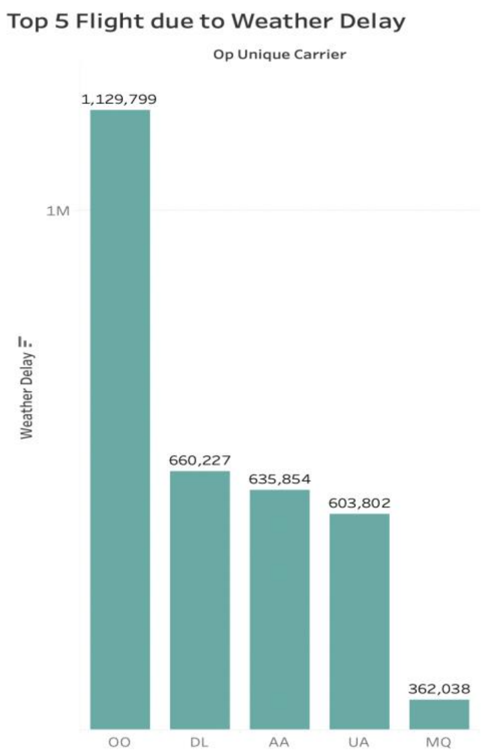
The above graph shows the cancellation trend based on day of the week. Tuesday and Sunday have the maximum number of cancellation while Friday and Thursday have the minimum cancellation.

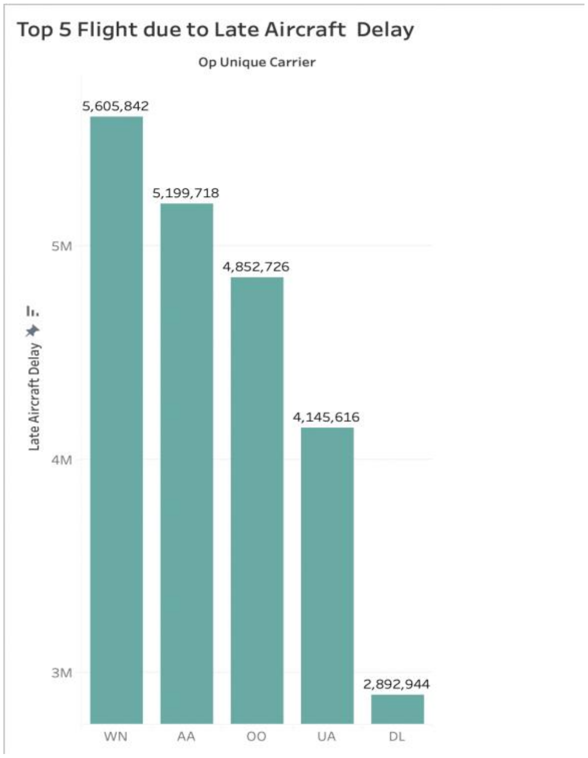
Delays that are under the control of the National Airspace System (NAS), such as non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, and so on, are shown in the graph below. United Airlines is the carrier with the most NAS delay, followed by American Airlines.





Security delays caused by evacuation of a terminal or concourse, re-boarding of aircraft due to a security breach, malfunctioning screening equipment, and/or long waits at screening areas are depicted in the graph above. Southwest Airlines is the airline with the longest security delays, followed by American Airlines.





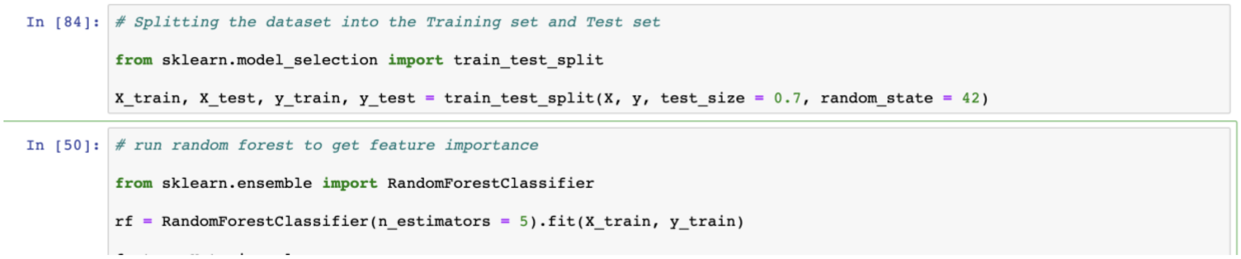
The above graphs show the delay due to various weather conditions. The arrival delay at an airport caused due to the late arrival of the same aircraft at a former airport is depicted in the graph below. Southwest Airlines is the airline with the latest arrivals, followed by American Airlines.

Data Modeling

Because the data we have is imbalanced, we decided to utilize Random Forest and Decision Tree to model the data and find the factors that are significant in causing the flight cancellations in the dataset. "CANCELLED" was used as a predictor variable, while all other factors were used as outcome variables. We excluded the variables that are associated with delays because our goal is to forecast flight cancellations, not delays.

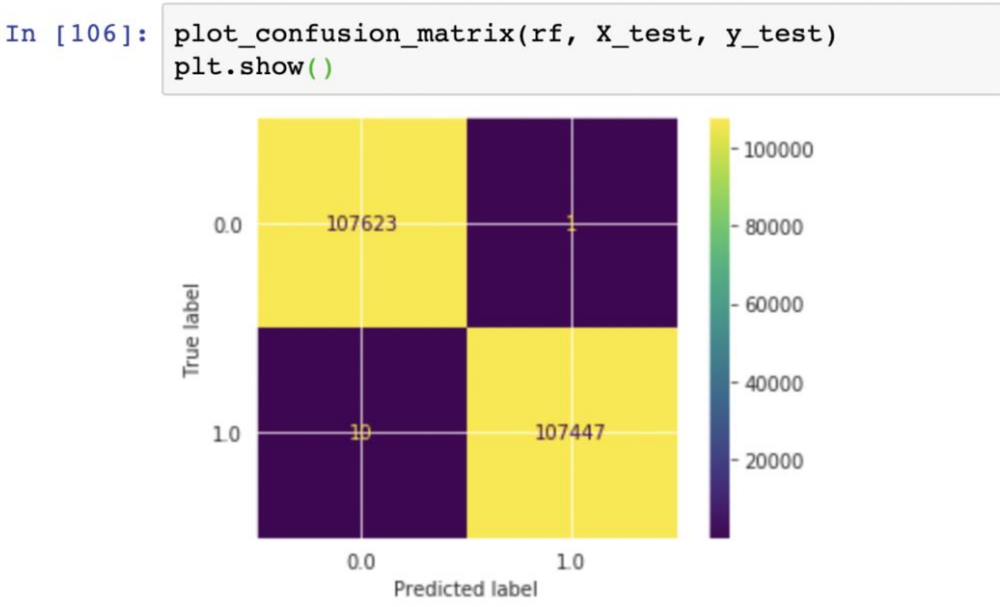
1) Random Forest

We have split the dataset into train and test with a ratio of 70:30 and fit the model and then transformed all the variables to numerical values for faster computation.



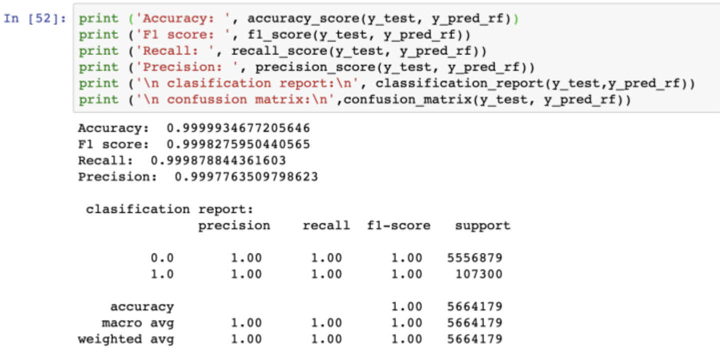
```
In [84]: # Splitting the dataset into the Training set and Test set

         from sklearn.model_selection import train_test_split

         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.7, random_state = 42)

In [50]: # run random forest to get feature importance

         from sklearn.ensemble import RandomForestClassifier

         rf = RandomForestClassifier(n_estimators = 5).fit(X_train, y_train)
```

The confusion Matrix obtained for the Random Forest Model is shown below



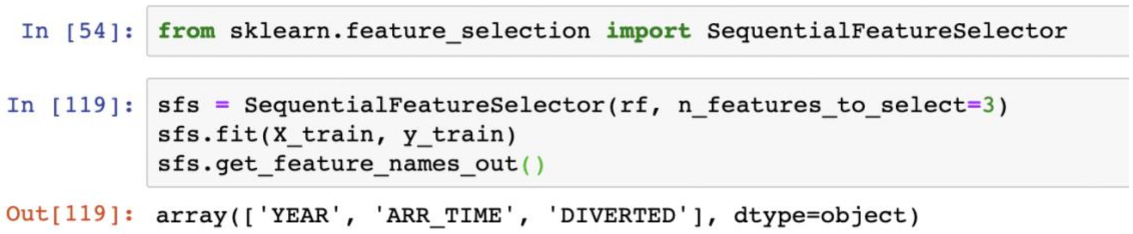
```
In [106]: plot_confusion_matrix(rf, X_test, y_test)
          plt.show()
```

The Accuracy, Recall, Precision and F1 score for the Model is shown below



```
In [52]: print ('Accuracy: ', accuracy_score(y_test, y_pred_rf))
         print ('F1 score: ', f1_score(y_test, y_pred_rf))
         print ('Recall: ', recall_score(y_test, y_pred_rf))
         print ('Precision: ', precision_score(y_test, y_pred_rf))
         print ('\n clasification report:\n', classification_report(y_test,y_pred_rf))
         print ('\n confussion matrix:\n',confusion_matrix(y_test, y_pred_rf))

Accuracy:  0.9999934677205646
F1 score:  0.9998275950440565
Recall:  0.999878844361603
Precision:  0.9997763509798623

 clasification report:
               precision    recall  f1-score   support

         0.0       1.00      1.00      1.00   5556879
         1.0       1.00      1.00      1.00    107300

    accuracy                           1.00   5664179
   macro avg       1.00      1.00      1.00   5664179
weighted avg       1.00      1.00      1.00   5664179
```

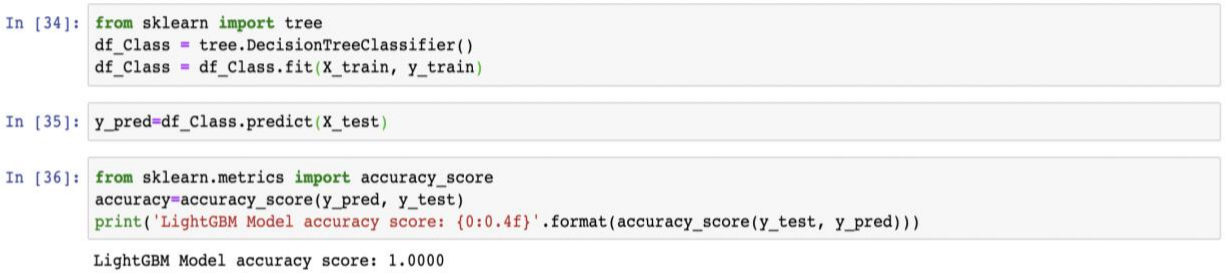
The above picture shows the top three variables that must be considered to minimize the cancelation of flights.



```
In [54]: from sklearn.feature_selection import SequentialFeatureSelector

In [119]: sfs = SequentialFeatureSelector(rf, n_features_to_select=3)
          sfs.fit(X_train, y_train)
          sfs.get_feature_names_out()

Out[119]: array(['YEAR', 'ARR_TIME', 'DIVERTED'], dtype=object)
```

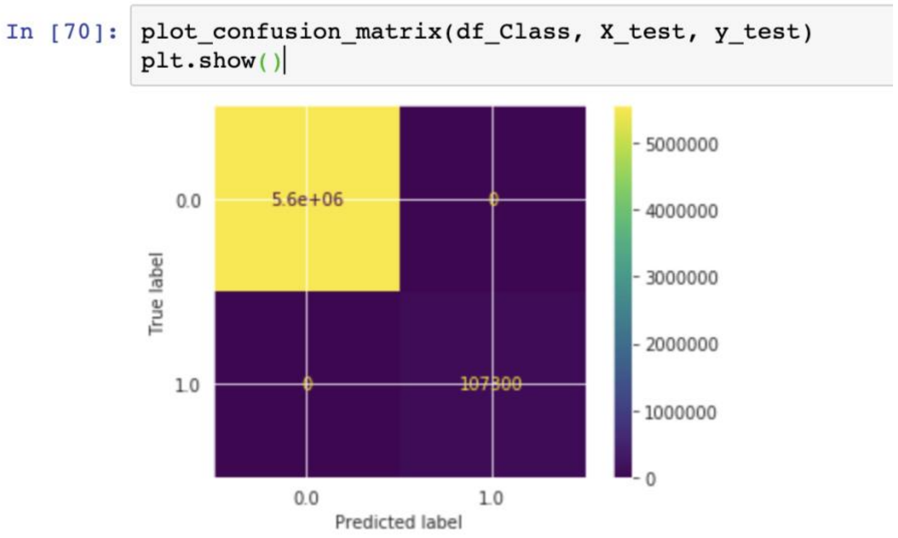
2) Decision Tree

We have fit the model using Decision Tree Classifier



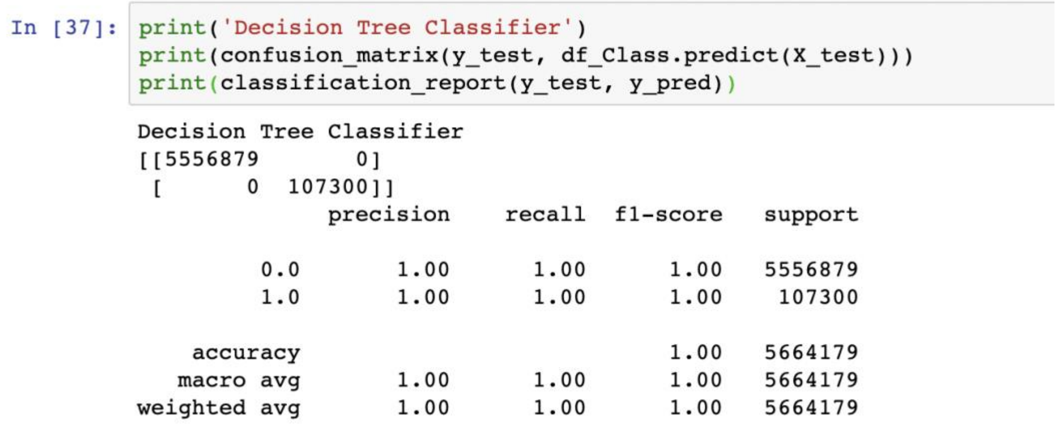
```
In [34]: from sklearn import tree
         df_Class = tree.DecisionTreeClassifier()
         df_Class = df_Class.fit(X_train, y_train)

In [35]: y_pred=df_Class.predict(X_test)

In [36]: from sklearn.metrics import accuracy_score
         accuracy=accuracy_score(y_pred, y_test)
         print('LightGBM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_test, y_pred)))

         LightGBM Model accuracy score: 1.0000
```

The confusion matrix obtained for the decision tree model



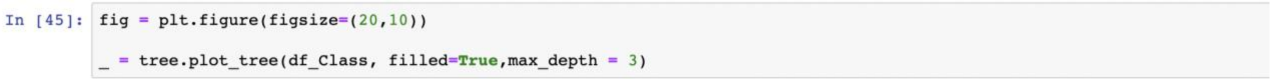
```
In [70]: plot_confusion_matrix(df_Class, X_test, y_test)
         plt.show()
```

The Accuracy, Recall, Precision and F1 score for the Model is shown below

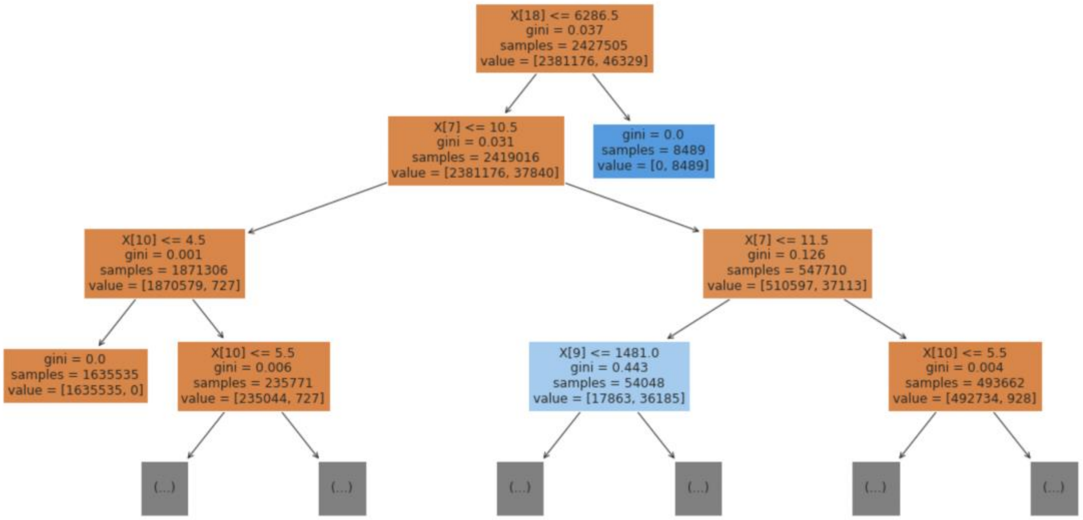


```
In [37]: print('Decision Tree Classifier')
         print(confusion_matrix(y_test, df_Class.predict(X_test)))
         print(classification_report(y_test, y_pred))

         Decision Tree Classifier
         [[5556879       0]
          [      0  107300]]
                       precision    recall  f1-score   support

                  0.0       1.00      1.00      1.00   5556879
                  1.0       1.00      1.00      1.00    107300

             accuracy                           1.00   5664179
            macro avg       1.00      1.00      1.00   5664179
         weighted avg       1.00      1.00      1.00   5664179
```

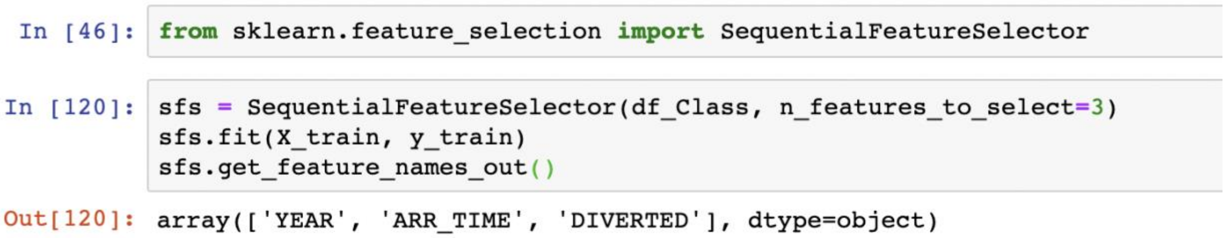
Decision Tree Plot



```
In [45]: fig = plt.figure(figsize=(20,10))

         _ = tree.plot_tree(df_Class, filled=True,max_depth = 3)
```



The top three variables that must be considered to minimize the cancelation of flights.



```
In [46]: from sklearn.feature_selection import SequentialFeatureSelector

In [120]: sfs = SequentialFeatureSelector(df_Class, n_features_to_select=3)
          sfs.fit(X_train, y_train)
          sfs.get_feature_names_out()

Out[120]: array(['YEAR', 'ARR_TIME', 'DIVERTED'], dtype=object)
```

Model Comparison

Based on the above confusion matrix, we may conclude that the decision tree outperformed the random forest. Despite the fact that both models have high accuracy, the decision tree model was able to forecast False Positive and False Negative those accounts equal to 0, eliminating Type I and Type II errors. Hence Decision Tree model can be used to forecast the cancellation of flights.

Recommendation

The most significant factors that contribute to flight cancellations, as indicated in the above feature selection plot obtained by running the Decision Tree, are *Arrival Time* and *Diverted*. To reduce passenger flight cancellations, the aviation industry must emphasize on these factors in order to boost revenue and improve customer satisfaction.

Conclusion

The analysis on the '*flights'* dataset was performed using python and Tableau. The exploratory data analysis using Tableau gave us insights on the patterns of the cancellations, we were able to identify the airlines that flew most frequently, the airlines that has the maximum number of delays and determined top five airlines with the maximum amount of cancellation. Python was used for cleansing the data, for data imputation of data and to identify missing values. Furthermore, we leveraged python for modeling. We used Random Forest and Decision Tree to predict the factors that affected the cancellation of the flights. A

References

Yanying, Y., Mo, H., & Haifeng, L. (2019, December 31). *A classification prediction analysis of flight cancellation based on Spark*. Procedia Computer Science. Retrieved December 13, 2021, from https://www.sciencedirect.com/science/article/pii/S1877050919320241.

(PDF) Flight Delay Prediction based on Deep Learning and ... (n.d.). Retrieved December 13, 2021, from https://www.researchgate.net/publication/346539839_Flight_delay_prediction_based_on_deep_learning_and_Levenberg-Marquart_algorithm.

Secondary navigation. Aviation Data & Statistics. (2021, April 12). Retrieved December 13, 2021, from https://www.faa.gov/data_research/aviation_data_statistics/.

Pathak, P. P. (2021, July 18). *Decision trees and random forests-explained with python implementation.* Medium. Retrieved December 19, 2021, from https://towardsdatascience.com/decision-trees-and-random-forests-explained-with-python-implementation-e5ede021a000