Predicting Flight Delays at Arrival using Machine Learning

Introduction

Flight delays are a known inconvenience for the millions of daily passengers across the US and an economic burden for the airlines that face them (Ball). Delays disrupt tourism, business, connecting travel plans, and more, all of which cause economic loss and increase the stress of travel. According to the Bureau of Transportation Statistics, 22% of flights, over 1 in every 5 flights, in 2024 were not able to arrive on time.

This study will focus on RDU, where 24% of flights in 2024 were delayed, to determine which information about a flight and its journey from the origin to destination airport can help best predict what delay a flight will face.

Personal motivation

As an avid traveler, I know how disruptive flight delays are to travel plans, whether causing travelers to miss their next transportation, forcing them to stay at an airport or hotel without the proper luggage, or making them miss much anticipated plans. While this study will not be able to limit the number of delays travelers face, it will be able to provide a prediction of the delay a flight might face, allowing travelers to be better prepared and increase flexibility in their plans.

Dataset description

The dataset used for this study is taken from the Bureau of Transportation Statistics' "On Time : Reporting Carrier On-Time Performance". This data source has monthly updates from 1987, with the most recent update in December 2024. It carries multiple data fields categorized by Time Period, Airline, Origin, Destination, Departure Performance, Arrival Performance, and more. The relevant data fields are described below:

Field	Description
DayofMonth	Day of the Month
DayofWeek	Day of the Week (1=Monday, 2=Tuesday, etc.)
Reporting_Airline	Airline carrying the flight (identified by a Unique Carrier Code)
Origin	Origin Airport
DestStateName	Destination State
CRSDepTime	Central Reservation System Time of departure
CRSArrTime	Central Reservation System Time of arrival
ArrDel15	Arrival delayed by 15 minutes or more (1=Yes)
ArrivalDelayGroups	15-minute intervals of arrival delays (from <-15 to >180, each assigned an integer value)
CRSElapsedTime	Central Reservation System elapsed time
ActualElapsedTime	Actual elapsed time
CarrierDelay	Delay caused by the airline, in minutes
WeatherDelay	Delay caused by weather, in minutes
NASDelay	Delay caused by National Aviation System, in minutes
SecurityDelay	Delay caused by security, in minutes
LateAircraftDelay	Delay caused by a late aircraft, in minutes

From the Arrival Performance category, ArrivalDelayGroups will be the target for this study, and all other fields, excluding ArrDel15, are potential factors to predict this target.

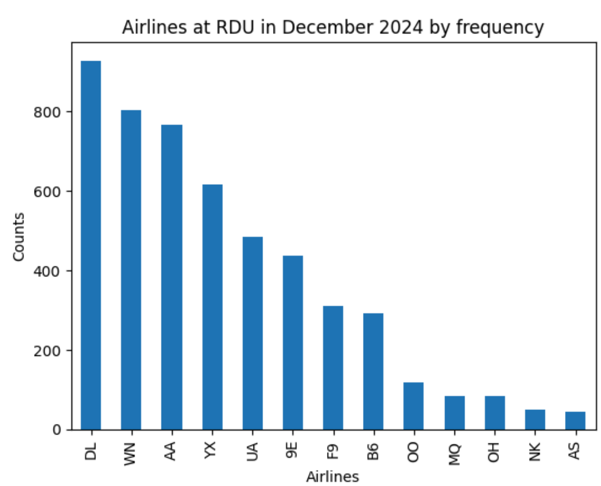
To focus on the most recent data concerning RDU, the downloaded dataset only contains information of travel in North Carolina from December 2024.

Data processing

Processing the data involved removing flights that did not depart from RDU and mapping the String features into integers.

Although the data was downloaded to only include flights in North Carolina, it still contained flights departing from other airports across the state. To remove these, values in the "Origin" column that were not "RDU" were set to NaNs using np.nan, and the rows containing these NaNs were removed using .dropna(inplace=True).

The two features of type String being used were Reporting_Airline, the carrier, and DestStateName, the destination state. Using the example of Reporting_Airline, to give each a meaningful assigned number, the airlines were ordered by their flight frequency from RDU in December 2024, as visualized in the bar graph below.



Using the sorted list, each airline was assigned a number by their position in the list, and that number was mapped to the Report_Airline column. The same was done to convert from String to Int in the DestStateName column.

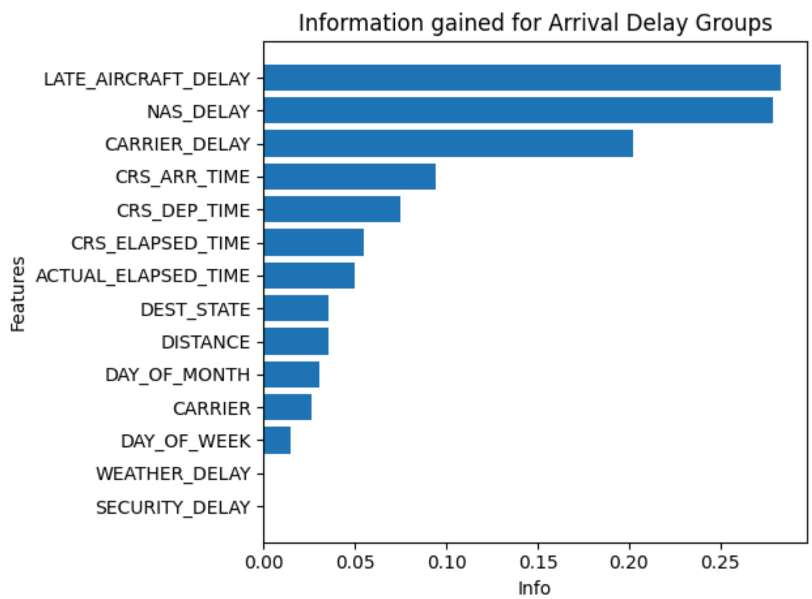
Feature selection

With the columns processed as necessary, feature selection was ready to be used to determine which columns were the most useful for the decision tree. Specifically, this was done using Mutual Info Classification, known to be useful for decision trees as it determines the features that provide the most information with simple calculations.

The amount of information provided by a feature is determined by how much entropy it removes. Using D as a dataset, F as a feature, H representing calculating entropy, the formula is:

$$\text{Information Gain}\ (D,F) = H(D) - H(D|F)$$

After importing mutual_info_classif from sklearn.feature_selection, I calculated each feature's information gain, sorted the features from highest to lowest, and created the bar graph below.



I chose the four features with the highest information gain, Late Aircraft Delay, National Air System Delay, Carrier Delay, all of which relate to causes for delay, and the Computer Reservation System Arrival time. The information gained for each feature is shown in the table below:

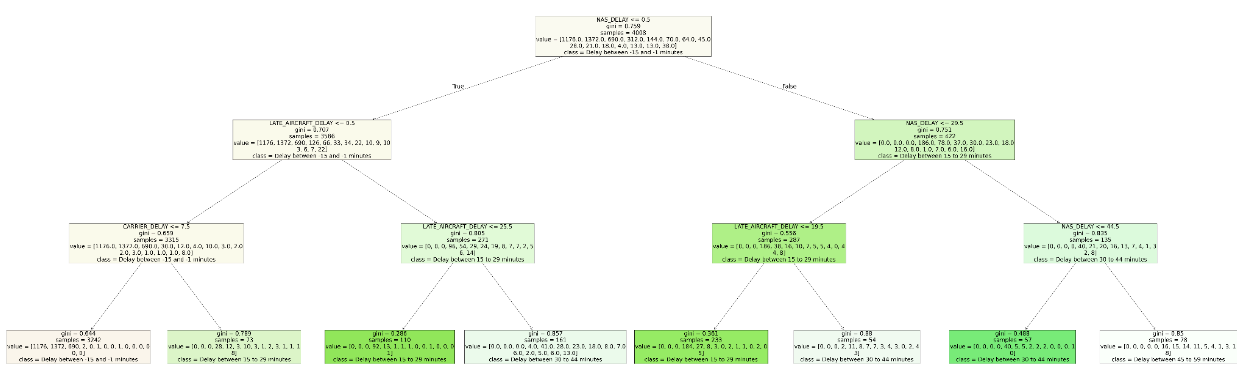
Feature	Information gained
LATE_AIRCRAFT_DELAY	0.2832129738173359
NAS_DELAY	0.2788206101140638
CARRIER_DELAY	0.2024028075647406
CRS_ARR_TIME	0.09415336819209674

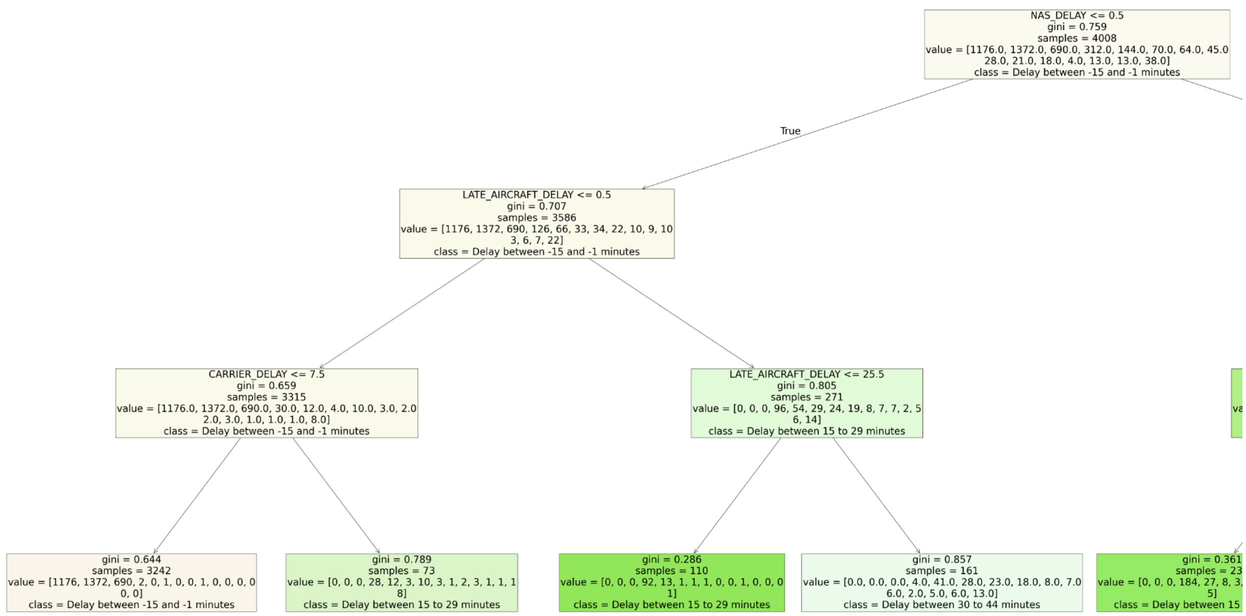
Decision Tree

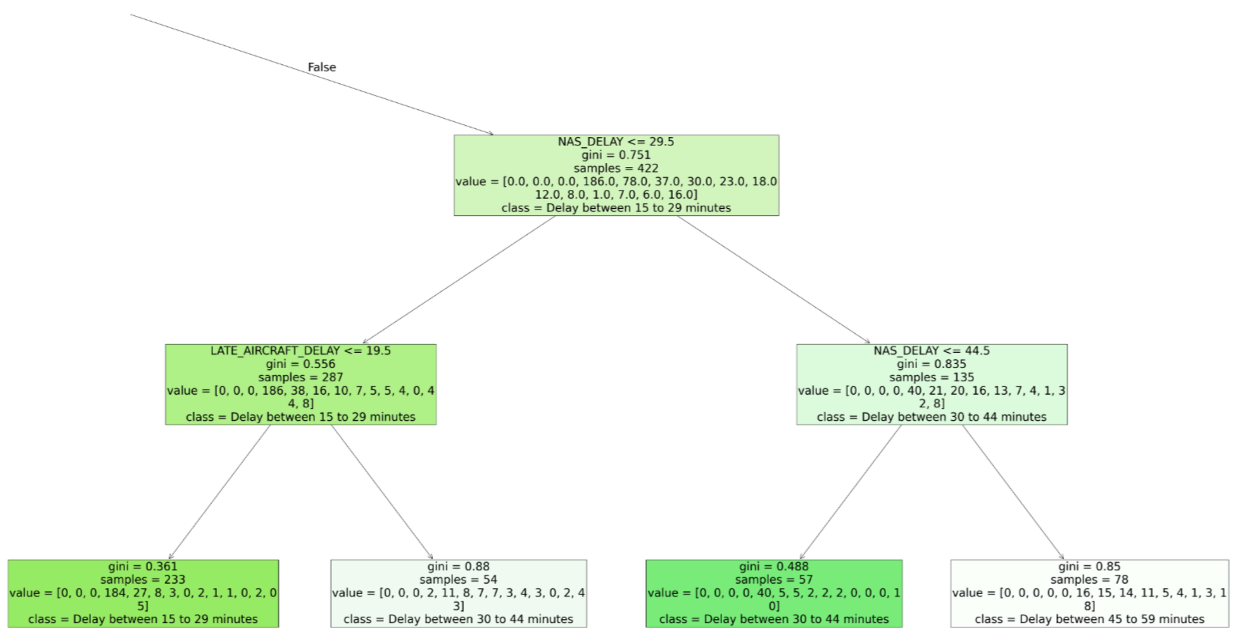
With these four features, I created a decision tree to determine which values of each feature best determine if the flight will be delayed. After assigning the features to an X value and Arrival Delay Groups to a y value, I used train_test_split to create the X_train, X_test, y_train, and y_test variables. Then, with a DecisionTreeClassifier with a max depth of 3, to reduce overfitting, I fit the X_train and y_train variables to the tree. To plot the tree, I set feautre_names to X.columns, and class_names to the description of each Arrival Delay Group, provided in a lookup table by the Bureau of Transportation copied below.

Code	Description	Code	Description
-2	Delay < -15 minutes	6	Delay between 90 to 104 minutes
-1	Delay between -15 and -1 minutes	7	Delay between 105 to 119 minutes
0	Delay between 0 and 14 minutes	8	Delay between 120 to 134 minutes
1	Delay between 15 to 29 minutes	9	Delay between 135 to 149 minutes
2	Delay between 30 to 44 minutes	10	Delay between 150 to 164 minutes
3	Delay between 45 to 59 minutes	11	Delay between 165 to 179 minutes
4	Delay between 60 to 74 minutes	12	Delay >= 180 minutes
5	Delay between 75 to 89 minutes		

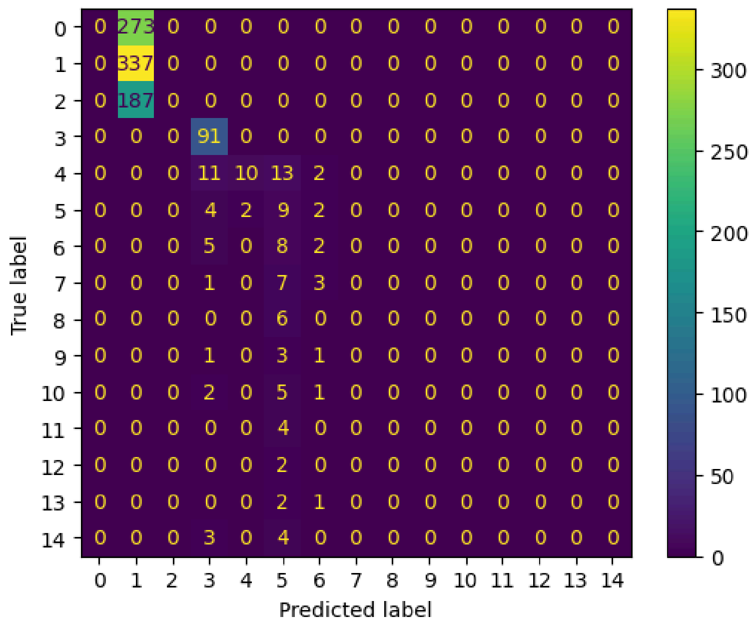
With this, I created the decision tree pasted below and with zoomed-in photos on the next page:







Next, I determined the predicted y values for X_test and compared them to the y_test values to determine that my Decision Tree had an Accuracy of 0.4530938123752495. Finally, I created the Confusion Matrix below to visualize the correct and false positive and negative predictions:

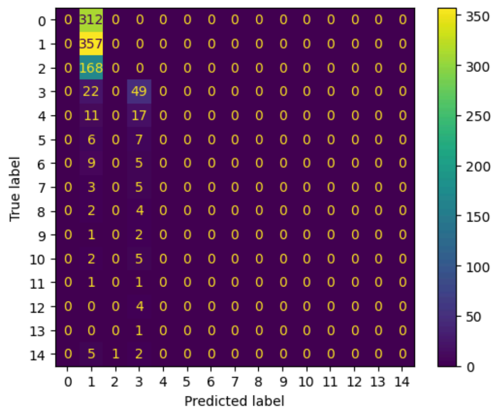
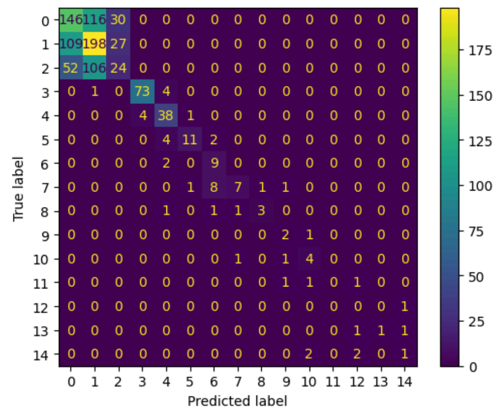


The axes are the order of each Arrival Delay Group in the table above. As can be seen, the decision tree does best predicting delays for flights that have little delay or are ahead of schedule.

Boosting

To improve the accuracy of my decision tree, I used the AdaBoost and Random Forest boosters. As one of the features, CRS_ARR_TIME, had a low information gain score, I chose AdaBoost as it combines weaker learners as well as reduces overfitting. I also used Random Forest as, by creating multiple trees, it is able to make the best predictions and it is known to be a strong boosting algorithm.

Using a similar procedure of creating each model with X_train and y_train and testing by making predictions for y based on X_test that were compared to y_test, I determined the accuracy and created a Confusion Matrix for each boosting method.

<table>
<tr><td>Boosting method</td><td>AdaBoost</td><td>Random Forest</td></tr>
<tr><td>Accuracy</td><td>0.405189620758483</td><td>0.5159680638722555</td></tr>
<tr><td>Confusion Matrix</td><td>
</td><td>
</td></tr>
</table>

As can be seen, both boosters perform similarly to the Decision Tree in terms of which categories they are best able to predict, but, ultimately, AdaBoost had an accuracy less than that of the Decision Tree while Random Forest was able to surpass it.

Conclusion

The predictions provided by the Decision Tree and improved by Random Forest have an impressive accuracy given that they are categorized into 15 delay intervals. Following the decisions made at each level of the Decision Tree and knowing the cause for their flight delay can help travelers determine how much delay to expect and to then be better prepared.

This model can be improved by using data from more time periods and be useful to more travelers by expanding to airports outside of RDU.

Citations

Ball, Michael, et al. "Total delay impact study : a comprehensive assessment of the costs and impacts of flight delay in the United States." *US Transportation Collection*. 1 October 2010.

Link to the code

Flight Delays