Flight Delay Classifier

Kellen Rice / Maria Alvarez / Michelle Cheung / Jennie Kim

Motivation

According to the Federal Aviation Administration, they handle over 16,405,000 flights yearly, that is over 2.9 million airline passengers daily. With this kind of sheer volume, it is imperative to have efficient and transparent operations. However recently we've seen a spike in the number of flights delayed, specifically domestic flights with over 20% of flights being delayed in the first 5 months of 2022 according to the Bureau of Transportation Statistics. This alarming statistic shows the necessity of clear and reliable communication around flight delays. Our goal is to provide clarity on both sides of the ecosystem, from airlines to passengers, the ability to better predict flight delays is essential for more efficient travel.

On the passenger side, the benefits are numerous. The inconvenience caused from a flight delay should not be underestimated, it can cause additional financial difficulties, lost time, and missed opportunities. With a better understanding of flight delays, a passenger can be better educated and therefore make better decisions when booking their travel plans and getting to the airport, this makes for a more efficient flow of people all around.

From an airline perspective, consistent unexpected delays and lack of communication can have severe consequences. From customer complaints, to increased costs, and an overall lower satisfaction, airlines want all they can do to improve their ability to predict delays. Some benefits on the airline side include better communication with passengers about the timing and probability of delays, reduced costs from overstaffing flights or airports dealing with these delays, and a more economical, productive, and effective system where the flow of passengers is normalized. The more prepared the airline can be, the better strategies they can create for dealing with these situations. Our project aims to solve the problem of poor communication and prediction for delays in order to implement the benefits stated above.

Data

The data that we used for analysis is from November 2019 to December 2020, about flights leaving the JFK airport. The data consists of 28820 individual flight information collected during the one year period. We were able to obtain this open data from Kaggle, which states that the data was scraped from an academic paper under review by IEEE transportation. The 23 attributes for each flight are month, day of month, day of week, carrier code (represents carrier company), tail number (air flight number), destination, departure delay, scheduled journey time, distance, scheduled departure, actual departure, scheduled arrival time, temperature, dew point, humidity, wind direction, wind speed, wind gust, pressure, condition of the climate, number of flights scheduled for departure, number of flights scheduled for arrival, and taxi-out time.

In order to process the data, we had to change non-integer variables to an integer variable. The dew point attribute was given as an object, so we first converted it into integers. Categorical variables that had very little to no influence on the prediction of flight delays were dropped, such as tail num. Then, we went on and used one-hot encoding to convert categorical attributes to integer columns. Since the ultimate prediction we want to make is delay, we turned the variable Departure Delay ("DEP_DELAY") into a binary variable, where 1 means that there was a delay and 0 there was not. If the departure delay (time difference between scheduled departure and actual departure in minutes) was longer than 0 minutes, we considered that as a delay. After making such modifications, we split the data set into training and testing sets (test size = 30%). The target variable is column "DEP_DELAY".

Analytics models

For our project, we created a binary classification model to predict the feature "DEP_DELAY," 0 if there is No Delay and 1 if there is. We trained and deployed the following algorithms: Baseline Model, Cross-validated Classification Tree (CART), Linear Discriminant Analysis (LDA), Logistic Regression, Vanilla Bagging (Random Forest model where *max_features* = the total number of features). The Baseline, Classification Tree, Linear Discriminant Analysis, and Logistic Regression models are all base classifiers whereas the Random Forest model is an ensemble method.

The baseline model was created by predicting the majority class in the train set on the test set, which was No Delay ("DEP_DELAY" = 0). We used a list comprehension to create an array of zeros to always predict No Delay for each value of the test set. Through utilizing a confusion matrix, it was identified that the baseline model yielded 6,300 true negatives and 2,346 false negatives. Using the accuracy_score function from the sklearn.metrics module, the Baseline Test Accuracy yielded a score of 0.72866.

The Logistic Regression model was created using LogisticRegression from the sklearn.linear_model module. A Logistic Regression model with all the features was first created. Then, a VIF (Variance Inflation Factor) Test was run on the features to identify multicollinearity. Next, features that had a high VIF (VIF >= 10) were removed and the logistic regression model was fit on this newly cleaned data set. The final Logistic Regression model yielded a test set accuracy of 0.7339.

The CART model was created using the DecisionTreeClassifier from the sklearn.tree module. The optimal ccp_alpha value for the tree was identified using cross-validation from GridSearchCV from the sklearn.model_selection module. This cross-validation process fitted 10 folds for each of the 201 ccp_alpha candidates (201 values between 0 and 0.1 inclusive). The other parameters for this model were: {'min_samples_leaf': [5], 'min_samples_split': [20],

'max_depth': [30], 'random_state': [88]}. The Cross-Validated CART model yielded a test set accuracy of 0.9669.

The Linear Discriminant Analysis (LDA) model was created using LinearDiscrimnantAnalysis from the sklearn.discriminant_analysis module. The LDA model yielded a test set accuracy of 0.7270.

Lastly, the Vanilla Bagging model was created using RandomForestClassifier from the sklearn.tree module and setting the parameter *max_features* equal to the total number of features (*max_features* = len(X_train.columns)). The Vanilla Bagging model yielded a test set accuracy of 0.9840.

	Model	Accuracy
4	Random Forest	0.984039
1	CV CART	0.966921
3	Logistic Regression	0.733865
0	Baseline Model	0.728661
2	LDA	0.727041

(*Fig. 1, Model Performance on Test Set*)

As seen in *Fig. 1*, Random Forest (Vanilla Bagging) and CV CART performed the best on the test set. Followed by much lower accuracies in our logistic regression, baseline and LDA models. This lower accuracy could be due to a number of reasons, one possible one being our classes are not very well separable with the current features in the data set. Another could be data skewing heavily towards the non-delayed option. However, overall we have a wide variety of models that provide us with interesting insight into the dataset. With our highest accuracy model being Random Forest, we performed bootstrapping to determine a 95% confidence interval. This gabes us an accuracy in an extremely tight window of [0.98155 , 0.98658], further proving that our model is extremely accurate and would be beneficial to airlines and customers across the board, with little thought or drawback to false positives, or false negatives.

We are confident about our results but would extend our analysis by using further cross-validation techniques to increase our confidence in our results. Additional analysis can also include applying different ensemble methods (i.e. Boosting, Stacking) and further tuning hyperparameters on the models.

Impact

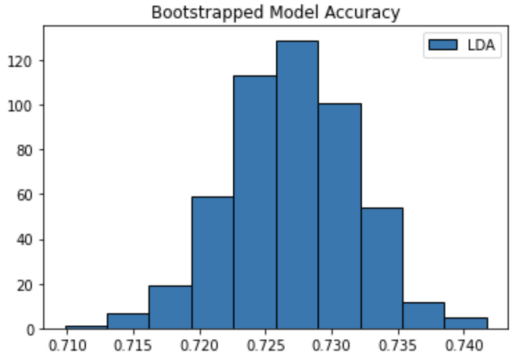
The potential impact our model can have on both airlines and passengers is enormous. From the airline's perspective, communicating with passengers about a flight delay early on has huge advantages. Sure everyone hates flight delays, but like many other obstacles in life they are bound to happen. Thus if an airline can communicate early on with their customers about a flight delay, they can offer them accommodations on another flight or the passenger themselves can plan accordingly. The effects that word of mouth can have on a brand's image are immense, thus if an airline can avoid passengers spreading the word about how they missed an important meeting, dinner, or whichever important event they had due a flight delay which was informed to them last minute, that is huge. Likewise, passengers would appreciate knowing ahead of time if their flight will be delayed; particularly if they've already arrived at the airport.

Our model also has cost savings opportunities for both parties. From the customer's side there is a money saving opportunity for the transportation they may be forced to take to the airport and back if the delay is longer than they are willing to wait at an airport for. Food savings is another factor, since a longer stay at the airport means passengers will need to pay for pricey airport food. Amongst this other unaccounted expenses could include extra clothes, hotel stay, and car rental fees. This of course does not consider non-quantitative losses such as time. On the airline's side, they are saving in expedited personnel or equipment that would be needed urgently due to delay, possible baggage claims for passengers with connecting flights, and the brand image.

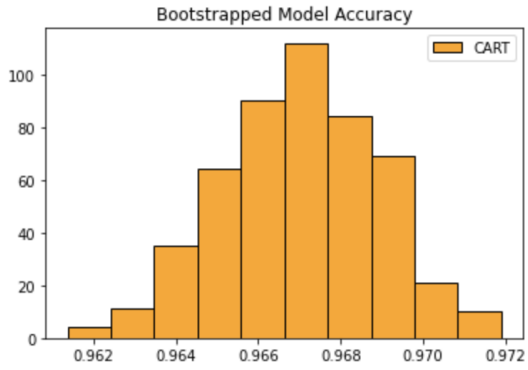
To further improve our analysis we can definitely expand our scope to other airports and airlines. With the given dataset our analysis was limited to JFK and the airlines that operate at that airport. However, to get a better understanding of flight delay causes and possibly improve our model our scope must expand. Although two of the models got an accuracy of over 90%, when exploring other airports, we may find that the features chosen for this model may not produce an accurate model. Therefore, other features and data should also be considered in the expansion of the scope.

Appendix

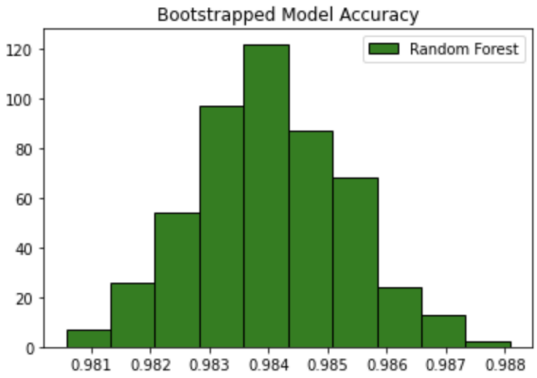
- Data
  - Kansal, Deepankur. “Flight Take off Data - JFK Airport.” *Kaggle*, 11 June 2021, https://www.kaggle.com/datasets/deepankurk/flight-take-off-data-jfk-airport.
- Additional Figures



(*Fig. 2, Bootstrapped LDA Accuracy*)



(*Fig. 3, Bootstrapped CART Accuracy*)



(*Fig. 4, Bootstrapped Random Forest Accuracy*)

- Code
  - https://drive.google.com/drive/folders/1j43at07jlYM9GODV7asYVJ61g6VkuePn?usp=sharing
- References

  - "Air Traffic by the Numbers." *Air Traffic By The Numbers | Federal Aviation Administration*, https://www.faa.gov/air_traffic/by_the_numbers.
  - Adams, Kurt. "2022 Has Brought More Air Travel Delays and Cancellations - and Nearly Double the Risk of Having a Bag Mishandled." *ValuePenguin*, ValuePenguin, 22 Aug. 2022, https://www.valuepenguin.com/travel/delays-cancellations-bags-study.
  - "FLIGHT DELAY PREDICTION." *ScholarWorks*, https://scholarworks.calstate.edu/.
  - "Data Profile: Airline On-Time Performance Data." *BTS*, https://transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&DB_URL=f7owrp6_VQ&f7owrp6_Qr5p=cn55r0tr4%FDg4n8ry&Z1qr_VQF=D.
  - "Airline Service Quality Performance 234 (on-Time Performance Data)." *Airline Service Quality Performance 234 (On-Time Performance Data) | Bureau of Transportation Statistics*, https://www.bts.gov/browse-statistical-products-and-data/bts-publications/airline-service-quality-performance-234-time.