08.04.2021

Michael Bollig, Emily Brantner, Sarah Iranpour, Abhi Sharma

UC Berkeley MIDS

Current airlines have too many delays and not enough lead time on notifications when delays inevitably happen. MASE aims to shake up the industry by being the low-delay airline. We’re accomplishing this by investing heavily in our data science team to create predictive models to know when and why delays occur so we can stop them at the source.

# Overview

Flight delays cost consumers over 18 billion dollars annually, and cost airlines another 10 billion. Flights are only becoming more delayed (with the effects of covid being the exception). Between 2016 and 2017, the number of on-time flights decreased by 8.5% (Thiagarajan et al. 2021). Delayed and canceled flights cause a loss of trust between the consumer and the airline, which we at MASE hope to avoid. The goal of our project is to use weather and airline data to predict whether flights will be delayed- we define this as taking off fifteen minutes or more beyond the scheduled time or canceling the flight altogether. Past attempts to predict delays based on weather and flight data have been largely successful, with at least one deep learning model achieving 96.2% accuracy (Yazdi et al., 2020). Another study looked at airline data without weather data, and found that between random forest, logistic regression, KNN, decision trees, and Naive Bayes, their random forest models had the best outcomes by over 4%, managing to get 66% accuracy, with logistic regression following behind it with 62% accuracy (Huo et al., 2021). A quick google search results in large number of papers, theses, dissertations, and kaggle contests, all surrounding how best to predict flight delays, indicating that we've chosen a very important topic to focus on.

To become the premier "low delay airline", the specific question we're answering is:

**Two hours before the planned departure time, can we use weather and airline data to predict with over 80% confidence whether a flight will be canceled or delayed by more than 15 minutes?**

To accomplish this, we use four datasets, three which were provided (airlines, weather and stations), and one which we sourced from OpenAirlines. These four datasets were joined together to create one massive dataset with all flight and weather features on US domestic flights from 2015-2019. From there, we created additional features from the data such as percent and mean delays across different aggregations, holiday indicators, prior delays, delay potential, and others.

Because we’re working with time series data, we can't just split the data without thought, since time may have a factor in whether flights are delayed. Instead, we used cross validation to split into our train and test sets. We have significantly more on-time flights than delayed flights, which makes our data set unbalanced. In at least one previous study, upsampling the delayed data was shown to overfit (Yazdi et al., 2020) so we undersampled our on-time flights to balance the two classes. Previous After preparing and transforming the data, we selected four models to predict delays: logistic regression, support vector machine, gradient boost tree, and random forest. We chose to use F1 scores as our main evaluation as it's a decent blend between both precision and recall.When a flight is delayed without warning (low recall), passengers will be unhappy and less likely to book with the airline again, and when a flight is predicted as a delay, but ends up being on time (low precision), passengers may miss their flights and revenue may be lost from making alternative arrangements. We expected our random forest to give us the best results based on previous studies (Huo et al., 2021), but our logistic regression model had the best results with a maximum F1 score of 0.904, and a mean of 0.898. With these scores, we feel confident that we’ll be able to reduce and better predict delays for our flights.

# Data

## Initial Datasets

For this predictive modeling task, three primary datasets have been provided: a flights table, a weather station table, and a weather table. In order to get a baseline understanding of these datasets, the first quarter data from 2015 will be loaded for all three tables. The flights table has additionally been preprocessed to only contain flights originating from the two busiest US airports: Chicago (ORD) and Atlanta (ATL). While we did do a lot of our initial work with the dataset using the three months of data, rather than the full data, we wrote our work into functions so that we could reuse them easily with the full data.

## Cleansing

### Airport Data

The airlines dataset is robust, with some features that were redundant, could be combined or needed to be added. We changed the names of the variables for ease of use, and our more technical changes are detailed below.

**Feature removal**

* additional names for airline carriers, airports, cities, and states removed, leaving OP\_UNIQUE\_CARRIER, ORIGIN/DEST\_AIRPORT\_ID, ORIGIN/DEST, ORIGIN/DEST\_STATE\_ABR, ORIGIN/DEST\_CITY\_NAME
* removed information about dept, 2nd, 3rd, 4th, and 5th diverted flights. Because we're only looking at whether a flight is or is not 15 minutes delayed, and not HOW delayed it is, any diversion is going to knock a flight more than 15 minutes off track, and the 4th and 5th diverted flights are all NULL in the 3mo dataset.
* removed aditional gate departure time columns as that departure time is built into the departure delay.

**Feature combinations**

* to make up for removing the information about the specific diverted flights, we added the diverted flight arrival delays into the arrival delay. Diverted delays previously had their own column and now they're combined. We know what the delay is from based on the diverted column showing 1 or 0.
* cancellations were added as 1s into the outcome variable dep\_is\_delayed

**Null Values**

* the NULL values are generally where canceled flights do not have a departure time, or flights that aren't diverted don't have a flight diversion time (and other cases along those lines). For now, we're leaving them as NULL, although we may have to change them to non-nulls depending on the model we use.

**Duplicate values**

* duplicate values did not exist in the three month dataset, but did exist in the full dataset, and were removed.

**Additional datasets and timezones**

* because the original dataset uses local timezone and does not have the same airport code type as our weather data, we joined the open airlines dataset which has both timezones and the ICAO airport codes. As a bonus, we also have altitude information about the airport in this dataset
* because we wouldn't be able to see the future, our "time\_at\_prediction\_utc" feature is the time at which we'd be predicting the delay (2 hours before the planned departure).

Station Data

Weather Data

## Joining

# EDA

# Feature Engineering

##### Previous Flight Delay (prev\_fl\_del)

Prior flight performance can have an impact on flight departure times. We tracked flights by their tail numbers and arrival times in order to determine if the previous flight was delayed. If the previous flight was delayed, we set the indicator to 1 and if it was not delayed, we set the indicator to 0.

##### 

##### Potential for Delay (poten\_for\_del)

Previous flight arrival times can also have an impact on flight departure times. After landing, the plane needs to be refueled, cleaned, and maintenanced. The cabin crew and pilots may need to be changed. The more time in between the flight’s arrival time and next departure time, the less likely the flight departure will be delayed. We calculated the time in between flights by tracking the tail number and actual arrival time and created an indicator where flights with more than 2 hours in between flights were indicated with a 1 and less than 2 hours were indicated with a 0. Flights that were canceled, diverted, or did not have a previous flight were null and were indicated with a -1.

##### 

##### Holiday (holiday)

Airports typically see the most traffic during the holiday seasons. We captured this information by setting flights that depart on a US holiday to a “holiday” category. We also set the two days prior and after a holiday to “holiday\_adjacent” category since many people travel to a location before the actual holiday, spend time with their family or friends, and then fly back home after the holiday.

##### Average Delay 2-4 Hours Prior to Planned Departure

At times, there may be certain issues, such as security, weather, maintenance, etc., that can affect flight performance at the airport or carrier level. We created a few indicator variables to capture the average delay minutes 2-4 hours prior to planned departure times. If the average delay 2-4 hours prior is less than 15 minutes, it is assigned a 0, and if it is greater than 15 minutes, it is assigned a 1. There are null values if there are no flights in the 2-4 hour measurement window. We assign nulls a -1.

Origin Airport Average Delay 2-4 Hours Prior (oa\_del\_ind) - this feature is created based on calculating the average delay minutes 2-4 before planned departure at the origin airport.

Destination Airport Average Delay 2-4 Hours Prior (da\_del\_ind) - For this feature, we are determining average arrival delay at the departure 2-4 hours prior to the departure time at the origin airport. The concept is similar to the previous feature. If there is an issue that is affecting the destination airport, then flights to that airport may be delayed.

Carrier Average Delay 2-4 Hours Prior by Origin Airport (carrier\_del\_ind) - If a specific airline is low on maintenance or cleaning staff on a certain day, it may impact departure times. We created this feature by calculating the average delay minutes by carrier at the origin airport.

Five Delay Categories Average Delay 2-4 Hours Prior by Origin Airport (security\_window\_del\_ind, nas\_window\_del\_ind, carrier\_window\_del\_ind, weather\_window\_del\_ind, late\_ac\_window\_del\_ind) - There were five features, security, NAS, carrier, weather, and late aircraft delay, that provided the total delay minutes for the specific category. We created five features to capture the average delay minutes of each category by origin airport

##### 

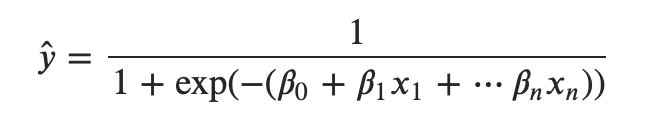
##### Aggregate Percent and Mean Delay

xxx

# Models

##### Model 1 Baseline: Logistic Regression

For the baseline model, we chose to implement a logistic regression model since it is easy to implement and efficient to train. It can help us determine feature importance and the relationship between features by measuring the coefficient size and direction of association. The features would need to be scaled before applying the model. Logistic Regression uses the Sigmoid Function, where it returns a probability between 0 and 1. If the probability is greater than 0.5, then the predication class will be 1 and if the probability is less than 0.5, the prediction class will be 0.



Although regression models are effective in determining a baseline relationship and are efficient to train, one of the major limitations is that it assumes linearity between the dependent and independent variables. This makes it difficult to determine complex relationships that are typically found in real world situations. Logistic regression also requires low or no multicollinearity between the variables. If a dataset has high dimensions, it can lead to overfitting.

https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

##### 

##### Model 2: Random Forest Classifier

For our second model, we chose to implement the random forest classifier over decision trees due to the problem of overfitting. Random Forest is an ensemble of decision trees, and utilizes the bagging technique. The trees are grown without any influence from other trees in the model and the result is determined by taking a majority vote from all the results of all the trees.

Random Forest has the power to handle higher dimensionality and can include features with lower correlation that may need to be excluded from other models, such as logistic regression. It can handle many variables and identify the most significant ones. Other advantages of random forest is that the features do not need to be encoded or scaled and it has an effective method for estimating null values. It can maintain accuracy even when big portions of data are missing. Based on our research on prior work on this topic, we expected this model to perform the best.

##### Model 3: Gradient Boost Tree Classifier

We chose another tree-based method as our third model, Gradient Boost Tree. It is also a set of decision trees, but differs from random forest in that it utilizes the boosting technique. The trees are grown sequentially where each tree is grown by using information from the previous tree in order to minimize the errors (based on the residuals).

However, Gradient Boost Trees may take longer to train with large datasets and may be more difficult to tune. The number of trees is important since boosting emphasizes small errors and noise. Too many trees can cause overfitting. We expected this model to be the least efficient in terms of training.

##### Model 4: Linear Support Vector Machine

The fourth model we chose was a linear support vector machine (SVM), which creates a hyperplane that separates data into classes. It differs from logistic regression in that it finds the distance between the line and support vectors and uses the hinge loss function. It tries to find an optimal hyperplane rather than focusing on maximizing the probability of the data. It is less prone to overfitting compared to logistic regression, however, may take longer to train. We expected SVM to perform better than logistic regression.

Hinge loss = [0, 1 - yf(x)]

##### Rebalancing

The dataset is imbalanced, with approximately 80% of flights departing on-time. When we run our models on the imbalanced data, accuracy is high, but F1 scores are low. Accuracy is high because the models are only predicting the majority class, however, low F1 scores indicate poor recall and precision. Because of this, we rebalanced the data by undersampling (due to the large amount of data).

# Course Concepts

##### One Hot Encoding

Some algorithms, such as logistic regression and support vector machine, are unable to work with categorical data directly. They need to be converted to numerical form by first assigning each category an integer value. Variables that do not have an order will then be one hot encoded to avoid algorithms treating the numbers as an attribute of significance.

# References

Huo, Jiage & Keung, K.L. & Lee, C. & Ng, Kam K.H. & Li, K.C.. The Prediction of Flight Delay: Big Data-driven Machine Learning Approach. (2021).

10.1109/IEEM45057.2020.9309919.

Thiagarajan B, et al. A machine learning approach for prediction of on-time performance of flights. In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). New York: IEEE. (2017).

Yazdi, M.F., Kamel, S.R., Chabok, S.J.M. et al. Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. J Big Data 7, 106 (2020). https://doi.org/10.1186/s40537-020-00380-z