Andrew Reilly, Jiesen Wang, Damien Klein, Bryce Sheedy

Stat 451 Project Proposal, Group 25

11/19/23

## Data Reading

```
import pandas as pd

df_raw = pd.read_csv("Airlines.csv")

df_raw.head()

df = df_raw.drop(['id','Flight'],axis =1)
```

We will use this dataset to attempt to predict flight delays and answer the questions below from the given variables above. The “id” and “Flight” variables were dropped because they were deemed unimportant in predicting a delay. See below for descriptions of the kept variables.

## Data Description

A brief summary of the raw data is necessary: there are approximately 540,000 observations for seven different possible predictors. These are listed below. The prediction variable is a categorical variable: 1 if the flight was delayed, and 0 if not. Upon further inspection, we noticed that the “AirportFrom” and “AirportTo” variables had around 300 different categories. This prompted us to do some kind of data imputation so as to make it unnecessary to use all ~300 airports. For this we have three possible solutions - either create a separate variable denoting the airport’s geographic region (e.g. airports in the general New England region would be reclassified as “NEN” for New England), create a new numerical variable corresponding to how many times that airport appeared in the dataset (e.g. if an airport has 463 instances in the dataset, the new variable will have a value of 463 at all instances of that airport), or use the airports that cover a decent percentage of the data (i.e. the airports which together cover 50% or 75% of the data, for example). All three solutions have their advantages and drawbacks, which is why we will test all of them in our project and determine which is more effective at predicting a delay.

Andrew Reilly, Jiesen Wang, Damien Klein, Bryce Sheedy
Stat 451 Project Proposal, Group 25
11/19/23

**Variables**

- **`Airline`** - flight airline (categorical)
- **`Flight`** - type of aircraft (categorical)
- **`AirportFrom`** - IATA code for source airport (categorical)
- **`AirportTo`** - IATA code for destination airport (categorical)
- **`DayOfWeek`** - day of week of flight (categorical)
- **`Time`** - departure time measured in minutes from midnight (numerical)
- **`Length`** - duration of the flight in minutes (numerical)
- **`Delay`** - whether the flight is delayed or not (categorical)

**Questions**

1. Is it possible to predict if there's a delay of a flight based on different characteristics like day of week, time of departure, and length of duration of the flight?
2. Which machine learning classification method is the most accurate at classifying if there's a delay for the flight or not?
3. What are some general trends in this dataset? Are there certain airlines, airports, or regions that have higher amounts of delays?

**Methods**

We will try two classification methods: logistic regression and decision tree. The dataset will be split into train and test data. The training data will be used to train the models, and the test data will be used for evaluating the performance. Also, we will do the parameter tuning by using Grid Search to find the best hyperparameters for both models and then using ROC to see which one performs better.