Group 2 Project Proposal

**Contributors:**

Beatriz Buquerin

Jamila Clayton

Gunnar Cukor

Mason Goss

Emma Resmini

Katie Stewart

ABSTRACT:

The client is experiencing issues with flight delays, causing their employees to miss meetings. Our objective is to use flight data to minimize missed meetings due to delayed flights. We plan to use a  logistic regression model to show the probability of an airline’s flights being delayed, as well as the usual length of delay.

BUSINESS PROBLEM:

The client has employees who utilize air travel on a weekly basis to travel to sites. The client has recently experienced issues with flight delays, resulting in missed meetings and customer complaints. Nearly all commercial airlines experience flight delays and cancellations. Additionally, delays and cancellations have negative impacts on the U.S. economy. Inefficiencies within air transportation increase the cost of business for other sectors, which makes the associated businesses less productive[[1]](#footnote-1). Our objective is to minimize missed meetings due to delayed flights. By optimizing flight schedules, our goal is to positively impact corporate synergy and ideation. We also seek to reduce expenditures and increase productivity.

Recently, the airline carriers used by the client have been experiencing delays. The problem is that the client doesn’t know if there will be a delay until their employees are already at the airport. Our project will determine which airline carriers have the lowest frequency of delays. Additionally, we will determine, for each airline, the length of time a delay usually lasts.  Our results can help avoid future delays.

This project needs to be completed as soon as possible before the consulting firm starts losing business. With our chosen data, the model will be run once. Our chosen dataset is static - it has not been updated since 2018. Thus, the model would only be updated either if new data becomes available, or with an entirely new dataset.

Presently, there are no physical constraints. Timing constraints: we will be implementing a comprehensive solution to our client as soon as possible to retain their business, with a final deadline of the week of December 12th. Analytical constraints: we are analyzing all major domestic airlines, although the client may have a preferred airline and airport.

The consulting firm funds travel expenses, the travel office of the consulting firm will be implementing the solution, and the consulting firm staff and clients will be the ones affected by the outcome of the project.

ANALYTIC PROBLEM:

We believe that by using a supervised classification model, we will be able to calculate a certain airline’s probability of delay, as well as the usual length of delay. This will allow the client to make an airline decision based on the modeled output.

We will use a dataset from Kaggle labelled “*Airline Delay and Cancellation Data, 2009 - 2018*”[[2]](#footnote-2). The site lists the data source as the Bureau of Transportation Statistic (BTS) site, although that could not be confirmed. However, certain attributes of the dataset are similar/identical to those found in the data provided by BTS. Therefore, it may be assumed that either BTS has since changed their data collection methods since 2018 and no longer publicly provides that specific data, or the Kaggle dataset is a combination of BTS data and third-party sources.

After comparing multiple datasets, it was confirmed that this dataset will likely provide the most information for the client. The dataset includes information such as date, carrier, arrival delay, departure delay, scheduled arrival/departure, actual arrival/departure, types of delay (carrier, weather, NAS, security, cancelled, diverted), and origin/destination, among others. We will be able to provide the client with a probability of flight delay per airline, but additionally will be able to include information about the length of delay, which will give further context. No airline will be perfectly on time, so knowing the length of delay will give the client the ability to choose, for example, an airline that is consistently 30 minutes late over one that is consistently 60 minutes late.

We will focus on flight data from 2014-2018 (five years of data). Since we want to provide the client with data on carrier delay, specifically, we will only consider the flights in the dataset whose delay type is due to the carrier. Therefore, flight entries with delays caused explicitly by weather, the National Aviation System (NAS), and security will be removed in the data cleaning process. Other dataset attributes that will likely be removed in the cleaning process include flight distance, air time (i.e. how long the flight lasted), time on the tarmac (several attributes), and flight number, as these likely have no effect on delay.

METHODOLOGY:

We will implement a combination of statistical summaries, visual analysis, and regression to solve this problem and present our findings to the client. We intend to execute these deliverables using Python.

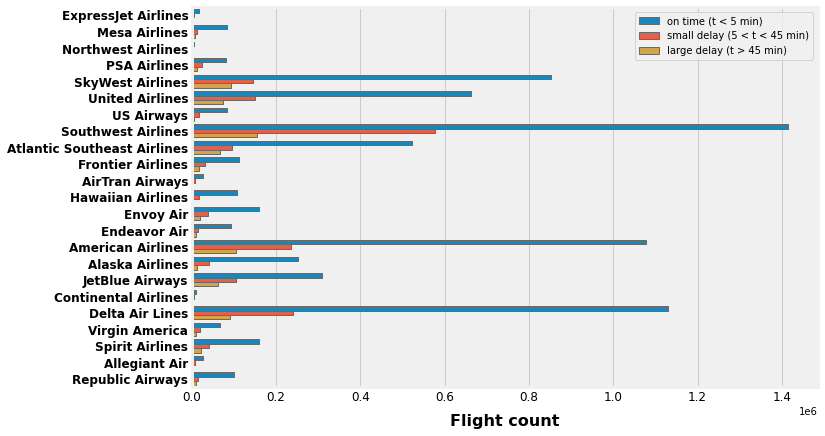
Initial data analysis on the dataset will occur after the data has been cleaned to assess the quality of data and need to perform transformations (as well as any further data cleaning). This can include identifying and appropriately handling extreme values and imputing missing values. Furthermore, we will identify broad characteristics of the data with the use of summary statistics and visualizations, which will further help prepare the data for the model. Data visualizations presented to the client will give an overview of the problem to be solved as well as a visual of any statistics deemed important for viewing.

We will use a logistic regression model, which is a supervised classification model. Our goal is to determine, for each airline, the probability of flight delay. The first model will be a binary logistic regression model with two outcomes: delayed or on-time. Classification will require thresholds to be defined. For example, flights with an arrival delay time of less than 30 minutes could be classified as on-time and flights with delays of more than 30 minutes classified as delayed.

An additional regression model can be made to determine the length of delay. This model will be a multinomial regression model, with the outcome categories as different time ranges (e.g. up to 1 hour, 1-2 hours, more than 2 hours).

The graph below follows how we expect one of our models to display its output: showing how, for each airline carrier, the flights fall into different delay categories:

**Flight Delays by Airline**



TIMELINE:

**Definitions:**

All: Refers to the entire group.

Admin Team: Jamila Clayton, Mason Goss, and Katie Stewart.

Data Team: Beatriz Buquerin, Gunnar Cukor, and Emma Resmini.

|  |  |  |
| --- | --- | --- |
| **Date** | **Task** | **Responsible Party** |
| 1-Sep | First Group Meeting to discuss roles and coding language | All |
| 8-Sep | Second Group Meeting to determine problem and general approach | All |
| 13-Sep | Found the data set | Data Team |
| 17-Sep | Business problem brainstorm and draft begin | Admin Team |
| 22-Sep | Business problem complete | Admin Team |
| 23-Sep | Analytic Problem Begin Draft | Admin Team |
| 27-Sep | Analytic Problem Final Draft Due | Admin Team |
| 28-Sep | Data Wrangling and Cleanup Begin | Data Team |
| 5-Oct | Data Cleanup Complete | Data Team |
| 7-Oct | Begin Model Coding | Data Team |
| 12-Oct | Final Team Review of Proposal | All |
| **12-Oct** | **Turn in Proposal** | - |
| 21-Oct | Complete Model Build, Begin Validation and testing | Data Team |
| 4-Nov | Complete Testing | Data Team |
| 5-Nov | Begin prepping for production and conclusions | All |
| 10-Nov | Meet to begin building final report | All |
| 15-Nov | Final Report Drafted and Reviewed | All |
| 21-Nov | Thanksgiving Recess Begin | - |
| 28-Nov | Thanksgiving Recess End | - |
| 1-Dec | Final Report Re-Reviewed | All |
| 4-Dec | Begin Prepping for presentation | All |
| 8-Dec | Present in Class | All |

1. Berkeley News, “Flight delays cost $32.9 billion, passengers foot half the bill”, <https://news.berkeley.edu/2010/10/18/flight_delays/> [↑](#footnote-ref-1)
2. Kaggle,  “Airline Delay and Cancellation Data, 2009 - 2018”, <https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018> [↑](#footnote-ref-2)