# Project Scoping

## DS4A – Team 16

### Shantel Culver, Tosin Adebote, Ashley Drewry, Carolina Pacheco,

### Melissa Flores, Jiaxuan Ji, Sylvia Nanyangwe

**How to Reliably Predict Airline Delays**

Overview of industry, business, or problem

The airline industry provides air transportation for passengers and cargo by using aircraft. Flight delays and cancellations can significantly affect customer experience and cause financial loss to the business. The total cost of delays from 2016-2019 was over US$23 billion.

Define the specific problem that should be solved

Assist customers in making better decisions when booking flights and help businesses mitigate financial loss due to delays.

* Predict flights that will be delayed allowing passengers to avoid those flights if they chose.
* Allow businesses to do root-cause analysis on flights with habitual delays as well as apply mitigation strategies to avoid negative customer experiences related to delays.

How to accomplish the problem

* Analyze flight delay data: detecting seasonality, airport and connection flights patterns and any correlation.

Why does this problem matter?

Delays have costs airlines over US$23 billion. Additionally, negative customer experiences further erode a company’s reputation and future profit by loss of business.

Potential Audience

* Anyone who is booking flight arrangements (for self or others) would be interested. This information would allow them to make informed decisions regarding flights and/or airports.
* Airline executives could benefit by identifying frequently delayed flights and do deep dives into cause and effects.

First glance at the Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Datasets** | **Sources** | **Description** | **Topics** | **Related Topics** |
| Airline Delay and Cancellation Data, 2009-2018 | <https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018>  Original source is Office of the Secretary, Department of Transportation website, which has data from 1987-now | * US flights * Jan 2009-Dec 2018 * 6.43m for 2009, 6.45m for 2010, 6.07m for 2011, 6.10m for 2012, 6.37m 2013, 5.82m for 2014, 5.82m for 2015, 5.62m for 2016, 5.67m for 2017, 7.21m for 2018 * Has column for departure delay, a column where e 2% of entries have missing data for this variable * Has unique carrier code and the flight number. * Captures arrival and departure delays (with how long the delay was rather than a flag). | Transportation, operations |  |
| January Flight Delay Prediction | <https://www.kaggle.com/divyansh22/flight-delay-prediction>  Original source is Bureau of Transportation Statistics | * US flights * Jan 2019- Jan 2020 * 587k observation for 2019 and 607k for 2020 * Data is very complete * There is a flag columns indicating it was delayed by 15 min or more. * This dataset has the unique carrier code and the flight number. * Captures arrival and departure delays (with how long the delay was rather than a flag). |  |  |
| Feb 2020 US Flight Delay | <https://www.kaggle.com/rowhitswami/feb-2020-us-flight-delay>  Original source is Bureau of Transportation Statistics | * US flights * Feb 2020 * 574k observations * Has a flag for departures that were more than 15 minutes late, but not how long the departure was. * Departure time and the departure delay flag are missing data for 1% of entries * Has unique carrier code but not the actual flight number. * Has 9 columns when the previous two datasets had 28 and 21, respectively. * Captures only departure delays, with a flag rather and not the duration of the delay. |  |  |

Methods and Models

Exploratory Data Analysis (EDA)

* Pre-processing
  + Identify NaN
  + One-hot encoding
* Data exploration
  + Define the label (operationalize the parameters)
  + Skewed dataset (% of delayed flights Single and multi-line charts)
  + Identify significant key, temporal variance --> if there's commonality
  + Cross-validation (test vs. train)
  + Histograms to analyze any relationships between the variables
  + Scatter Plots
  + Box Plots to display the summary of behavioral variables (minimum, quartiles, median, maximum and presence of outliers).
  + Spatial/geographical visualization
* Data analysis
  + Feature importance: which factors contribute the most to delayed flights (PCA?)
  + Heatmaps to visualize correlations between locations with high delays/cancellations
  + Time Series Analysis
  + Model selection, development, and testing

What Libraries Will Be Utilized

The exploratory data analysis is the first crucial step in modeling: provides summary level insights, reveals underlying patterns, missing data, outliers, biased, unbalanced data and relates the available data to the business opportunity and the modeling.

In python, pandas will be used for profiling to ensure an effective data exploration and Matplotlib will be used for creating more detailed visualizations such as scatterplots, box plots, line charts and histograms.

Models

* Supervised ML for regression problems: Neural Networks, Classification models - ie SVM, Logistic Regression or Decision Trees
* Linear (Multi) Regression
* Statistical tests to find significance differences between subgroups

Interface

Below is an example of the interface we will be working towards for version 1.

Graphical user interface

Description automatically generated

1. A timeline which shows the dates that airlines experience the highest overall reported delays and cancellations.
2. A histogram that shows a snapshot of airports that have the highest reported delays and cancellations.
3. A pie graph that breaks down the highest reported airports by showing which airlines represent the largest portion of these delays.
4. A heat map that will show where the highest reported airports are located.

Milestones

**Version 1:** Build a simple dashboard with visualization to show airlines and airports most prone to delays and cancellations. Additionally, show a heat map which plots the airports with highest reported delays and cancellations.

**Version 2:** Build prediction model to show projected delays and cancellations for future use.

**Version 3:** Integrate neural networks to further enhance the prediction model and accuracy.

**Version 4**: Build interactive interface that airlines can use to toggle variables which will predict likelihood of delays.

Timeline

|  |  |  |
| --- | --- | --- |
| **Date** | **Deliverable** | **Responsible** |
| 9/18/21 | Project Team Formation and Introductions | All |
| 9/24/21 | Team Meeting with TA | All |
| 9/25/21 | Submit Project Description Document | Lead |
| 9/29/21 | Team Meeting - Scoping Document Review | All |
| 10/1/21 | Team Meeting with TA & Mentor | All |
| 10/2/21 | Submit Project Scoping Document | Lead |
| 10/3/21 | Begin Exploratory Data Analysis (1 week) | Practitioners |
| 10/3/21 | Begin Working on Report through 2B | Non-Practitioners |
| 10/8/21 | Team Meeting with TA | All |
| 10/9/21 | Submit Report Draft (through 2B) | Lead |
| 10/10/21 | Wrap Up EDA and Begin Coding/Analysis (1 week) | Practitioners |
| 10/10/21 | Continue Working on Report | Non-Practitioners |
| 10/17/21 | Finalize coding and analysis | Practitioners |
| 10/17/21 | Continue Working on Report and Presentations | Non-Practitioners |
| 10/22/21 | Review and Sign off on Final Report | All |
| 10/23/21 | Submit Final Report | Lead |
| 10/26/21 | Submit Final Presentation and Datafolio | Lead |

Concerns

* 2019-2020 data may capture delays caused by a rare event (COVID-19). Using these points may impact the prediction. While the entries sometimes have information on the cause of the delay, this is often not available.
* We may need to define delays carefully. [One analysis](https://www.milantomin.com/2018-u-s-airlines-delay-analysis/) defined a delayed flight as one that arrives late. A flight can have a delayed departure and still arrive on time. If we use the February dataset, that definition will not work for us.
* Do we care about delays of 15 or more minutes or any delays.
* Time constraint to achieve our version 4 milestone.