Wharton Analytics Fellows Fall 2023 Data Challenge

- Submission link
- Dataset
- Data dictionary

Instructions:

In this data challenge, you will be evaluating the on-time performance of passenger airlines and potential causes for flight delays. In this dataset, each instance (row) corresponds to a flight, some of which departed on time and some of which were delayed. Each instance is also accompanied by a variety of its attributes, including the operating airline, departure/arrival airports, delay duration, and weather conditions.

Using this dataset, we want you to show us what insights you can obtain about the flights–delayed, or not. Your task is to create a **1-5 slide presentation** containing your findings and answers to the questions below. This should include any relevant data visualizations, tables, and graphs. **Please also include a file with your code/script for review.** If you are not comfortable coding and plan on conducting your analysis in Excel or another tool, please include that file as well.

Please format your deck as if you are going to present it to an airline client who is relatively familiar with the topic at hand (Director of Flight Scheduling), but not necessarily an expert in data science and machine learning. You will present this deck during your interview, and we will then dive deeper into the technical side of your analysis after your presentation. Clear visuals and a well designed presentation are a big plus.

Note: The timeframe for this challenge has been shortened so that you don't spend too much time on it. You do not need to fully answer all of the questions below - if you feel you're short on time, rather choose a few to answer and deliver well-thought out, polished responses both from a technical and visual perspective. We expect about 4-6 hours of work to be put into this data challenge.

For a data dictionary and a description of the various columns in the dataset, check here.

Data Challenge Questions

You may choose to tackle some or all of these questions. Use whatever tools, programming languages, and models you are comfortable with and have at your disposal. Answering one or two questions thoroughly and deeply is preferable to attempting all of the questions without understanding your model or assumptions.

1: **EDA**: Does the data have any interesting quirks or features? Describe any important preprocessing/wrangling steps you took to clean the data.

2: **Classification**: What are the most predictive factors in determining if a flight is delayed for more than 15 minutes? Brainstorm or build a predictive classification model for the 'DEP_DEL15' column of the dataset. Explain your feature selection process. Feature engineering may be useful for this prediction.

3. **Regression:** Similar to question 2, what are the most predictive factors in determining the duration of a flight delay? Brainstorm or build a predictive regression model for the 'DEP_DELAY_NEW' column of the dataset. Explain how you evaluate the performance of the model, and please be prepared to defend your assumptions around the model performance.

4. **So What?** What are your final business recommendations to the business management team at this airline? What are the major insights and business implications? Incorporate the analysis from the models used in previous questions to recommend steps or action items to better identify flight delays beyond using the model itself. (Note: Once again, this question is open-ended and does not have one right answer. We're looking for creativity in how you might recommend the airline to proceed!)

Once you are finished, please submit your PDF here.

The data challenge is due on **September 17th at 11:59 PM ET**.