[Airline Delay Analysis]

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
  - The Airline industry is volatile in the sense that there are hundreds of thousands of passengers daily across the nation and many natural disasters, incidents, and shutdowns stopping flights or delaying flights all the time. I wanted to create a model that could predict delays or give airline's a chance to predict when abnormal stoppages could occur, so they are better prepared to more easily handle the situations.
- What industry/realm/domain does this apply to?
  - The airline industry
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
  - My motivation is due to the upcoming holiday travel… We have all had flights delayed or cancelled and I want to see if specific routes or time periods affect travel.

Data Understanding

- What data will you collect?
  - Shouldn't need much data collection as Kaggle has many datasets on the topic. Specifically, this dataset or this dataset have peaked my interest thus far (see links in my repo).
- Is there a plan for how to get the data (API request, direct download, etc.)?
  - Direct download should be sufficient.
- Are the features that will be used described clearly?
  - Very, I will be looking into flight data across the industry from a relatively large timespan with features clearly described which should make cleaning a lot faster.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
  - Encoding, to help the model determine the airline easier.
  - Some form of dictionary, key sorting (not sure if this is encoding) to assign values to orgin/destination codes. For example, DFW might be like {dfw:1, pdx:2} etc.
- What are some of the cleaning/pre-processing challenges for this data?
  - Somewhat described above, I also foresee that I will need to do some research to figure out what specific column labels mean (like OP_Carrier).
  - Deciding if certain columns/features are relevant to model performance.

Modeling

- What modeling techniques are most appropriate for your problem?
    - I have not performed any analysis yet. Some techniques that might be appropriate include hyperparameter tuning (as overfitting has come up in past analysis) and using tools such as GridSearchCV, also I may try out pipelines to help as well.
- What is your target variable? (Remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - Target will for sure be flight Delay
- Is this a regression or classification problem?
    - Classification because although we would like to know the delay time it is more important in this sense to know if our flight will be delayed at all since most flights are not going to have a delay. However, I can also see the argument for regression where we would look at how long the flights will be delayed by.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - RMSE

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - Random Forest has given me success in the past and is typically less computationally intense but may also try decision trees or Xgboost as neccesary.