Features

Categorical

1. Month
   1. air travel is seasonal
   2. Visual?
2. Day of the week
   1. travel is cyclical, certain days see more traffic and
   2. Visual?
3. Carrier
   1. every carrier has its own unique logistics or procedures that could lead to different delays
   2. For example, Southwest, they are famous for no assigned seating because they are concerned about delays
   3. Visual?
4. Origin Airport Code
   1. there are differences in airport procedures and quality. Some airports are unpredictable because of weather
   2. LAX is notorious for delays
   3. SFO is notorious for canceling because of fog
5. Destination Airport Code
   1. Same as origin airport code
6. Departure Hour
   1. traffic accumulates throughout the day. If you’re the first flight to leave, you are less likely to be delayed
   2. Visual?

Categorical maybes?

1. Origin State - similar to airport code
2. Origin City - similar to airport code
3. Destination State - similar to airport code
4. Destination City - similar to airport code
5. Origin Wind Type - any type of extreme weather could be a factor for delay
6. Origin Wind Angle - any type of extreme weather could be a factor for delay

Numeric

1. Origin Cloud Angle (-0.06) - any type of extreme weather could be a factor for delay
2. Origin Wind Speed (0.09) - any type of extreme weather could be a factor for delay
   1. Summer storms
3. Origin Visibility Distance (-0.08) - any type of extreme weather could be a factor for delay
   1. SFO fog
4. Origin Temperature (-0.007) - any type of extreme weather could be a factor for delay
   1. Winter?
5. Origin Dew (-0.001) - any type of extreme weather could be a factor for delay
   1. Similar to fog?
6. Flight Distance (0.01)
   1. Longer flights show longer delays
7. Planned Duration (don’t have the correlation yet)
   1. Longer flights show longer delays (pick this one or flight distance?)

Numeric Engineered - pretty self explanatory i think? These tell us the overall avg and % delays from origin, destination, route, and state

1. % Delayed from Origin
2. Mean Delay from Origin
3. % Delayed to Destination
4. Mean Delay to Destination
5. % Delayed for Route
6. Mean Delay for Route
7. % Delayed from State
8. Mean Delay from State
9. % Delayed to State

Categorical Engineered

1. Potential for delay
   1. checks if the prior actual arrival time is > 2 hours prior to planned departure time.
2. Previous Flight Delay
   1. checking if the prior flight was delayed
3. Origin Airport Average Departure Delay
   1. These are sort of like stress variables. It looks at the average delay 2-4hrs prior to planned departure time. The greater the delay, the more likely the next flights will probably be delayed
4. Destination Airport Average Arrival Delay
   1. Same as origin airport average departure delay

Hypothesis 1: these variables are more related to date time and location

* Month, day of the week, carrier, origin\_airport\_code, dest\_airport\_code, dep\_hour, planned duration OR flight distance

Hypothesis 2: taking hypothesis 1 and adding weather variables

* Month, day of the week, carrier, origin\_airport\_code, dest\_airport\_code, dep\_hour, planned\_duration OR flight\_distance + weather

Hypothesis 3: taking hypothesis 2 and added engineered features that have to do with the prior flight

* Month, day of the week, carrier, origin\_airport\_code, dest\_airport\_code, dep\_hour, planned\_duration OR flight\_distance, poten\_for\_del, prev\_fl\_delay, oa\_avg\_del\_ind, da\_avg\_del\_ind

Hypothesis 4: taking hypothesis 3 and adding the engineered features that tells us overall avg and % delay for origin, route, state, destination

* Month, day of the week, carrier, origin\_airport\_code, dest\_airport\_code, dep\_hour, poten\_for\_del, prev\_fl\_delay, oa\_avg\_del\_ind, da\_avg\_del\_ind, "pct\_delayed\_from\_origin", "mean\_delay\_from\_origin", 'pct\_delayed\_to\_dest', 'mean\_delay\_to\_dest', 'pct\_delayed\_for\_route', 'mean\_delay\_for\_route', 'pct\_delayed\_from\_state', 'mean\_delay\_from\_state', 'pct\_delayed\_to\_state',

This week we worked on multiple items. We worked on setting up the model pipelines for 3 additional models, Random Forest, Gradient Boost Tree, and Support Vector Machine. We also began using a subset of the data to test the pipelines, the models, and writing train and test results to storage. We were seeing poor scores overall as a result of unbalanced data so we chose to downsample the data. Another item we implemented was cross validation for time series data where we are training on 6 months of data and testing on 2 months on a 3 month rolling basis from 2015 to 2019. Once the pipelines and models were in good shape, we standardized the metrics we will be measuring across all the models. In terms of features, we engineered a few features that tracked departure delays 2-4 hours prior to planned departure time by airport and carrier. We went through the given features and the engineered features and began to narrow them down and form hypotheses of feature combinations we wanted to test in our models. Finally, we have started putting together our presentation for Tuesday.