*Predicting Upcoming Flight Delays*

**Project Updates**

After receiving feedback from my first delivery, I decided to make a few changes in order to improve my project overall. I have decided to only focus on weather induced delays from my dataset. I decided to do weather induced delays because it seemed appealing and about 18.56% of the delays in the dataset were due to weather related issues. As a result, I plan to find weather related data that will match my 2015 flights dataset. Additionally, I will focus on multiple ML algorithms instead of just one. My goal of doing so is to compare the accuracy of each algorithm to see which one works best for my dataset(s).

**Literature/industry research and outcome**

Many projects started off by constructing a model with a binary classifier which means that they were predicting that either the flight was delayed or not delayed. However, I want to use a multi-class model that will predict the magnitude of the delay. Therefore, I will be skipping the binary classifier model.

**Project differentiation**

As mentioned before, I am only focusing on weather induced delays. Other projects focus on various types of delays such as, aircraft delays, security delays, etc. Additionally, other projects only focus on American Airlines, while my project will analyze fourteen airlines. It also only focuses on five airports, while mine will focus on 322 airports.

**Transformation, Data Cleaning, and Exploratory Data Analysis (EDA)**

I cleaned, transformed, and performed EDA on the flights dataset. Luckily, my dataset was already in pretty good shape beforehand. Because I am only focusing on weather induced delays, I deleted all rows that stated that flights were cancelled. The flight dataset originally had 5,819,079 rows. 89,884 of those rows were cancelled flights, which left me with 5,729,195 rows after I cleaned the dataset. I then also deleted the “reason of cancellation” column because it became irrelevant.

**Next Steps**

During my next steps I plan to find weather data that will match my dataset. I am considering DarkSky API. I also plan to work on more visualizations that will include the weather dataset as well. Lastly, I will be working on Delivery 3 which will include finalizing the data exploration stage and constructing my models.

**Link to presentation:** <https://www.youtube.com/watch?v=hJMMTjQ7ZAw>

**References:**

* <https://www.kaggle.com/usdot/flight-delays#airlines.csv>
* <https://arxiv.org/pdf/1903.06740.pdf>
* <https://engineering.upside.com/applying-predictive-analytics-to-flight-delays-85413ca4939f>