Capstone 2 Project Proposal: Flight Arrival Predictive Model

Flight delays and late arrivals are a significant issue for airline companies, affecting operational efficiency, customer satisfaction, and financial performance. Accurately predicting if a flight will arrive late, and by how long, can enable airlines to proactively manage operations and improve the passenger experience. Knowing that a flight is likely to be late, airlines can establish policies to immediately assuage customer complaints. For example, planning to have extra complimentary meals on flights that are likely to be delayed or arrive late.

Using flight data from the United States's Department of Transportation's Bureau of Transportation Statistics, I will develop a predictive learning model to determine if a flight will be late, and estimate the duration of that delay. I will also look at identifying the key factors that contribute to delays and late arrivals. This will help airlines gain a better idea as to how many flights, and which flights, are likely to be delayed, giving them better data to work off of to develop strategies, policies, and procedures to addressing delays.

The biggest hurdle in determining this will be in the specificity, or lack thereof, of the data. While the dataset includes the length of the flight, and whether it arrived early or late, only some of the flights list a reason for the delay. In addition, the causes for delay are split into general categories, without specific reasons listed. For example, if the cause of delay is listed as "carrier delay," we do not know if this means there was a delay in fueling the plane, cleaning the plane, pilot running late, etc.

To create the model, I will need to first train the model to identify whether a given flight is likely to be delayed. Once the model is able to determine this, I will refine the model to predict an estimated delay duration. Finally, I will analyze the data to see if there are any patterns in causes for delays, and if so, what the most correlated factors are. After cleaning the data, some data wrangling that I will do will include creating a new feature column that separates delayed flights from flights that arrived early or on time. Then I will test different regression models to predict the length of the delay duration and to determine what the key factors are that lead to a delay.

By accurately predicting flight delays and their duration, this project aims to enhance the airline's operational efficiency and customer satisfaction. The insights gained will allow for better resource allocation, proactive issue management, and improved passenger communication, ultimately leading to a more reliable and customer-friendly airline service.