Flight Delay/Cancellation Prediction

Project Proposal

Members: Eugene Choi (ec727)[ORIE 4741], Ryan Mao (rwm275)[ORIE 5741], Corwin Zhang (csz9)[ORIE 4741]

Kaggle Dataset: https://www.kaggle.com/datasets/usdot/flight-delays

Github Link: https://github.com/Corfish123/ORIE4741FinalProject

As we navigate through the intricacies of managing airline operations, it becomes increasingly apparent that we need to enhance our ability to predict and mitigate potential delays/cancellations. Airline delays are an unavoidable phenomenon that is frustrating for both consumers and service providers, justifying the need for accurate predictions.

Predicting airline delays can have significant implications for airlines from a business perspective. By accurately predicting delays, airlines can better manage their resources such as crew schedules, gate availability, and aircraft allocation. This can lead to smoother operations and reduced costs associated with idle resources or last-minute adjustments. Additionally, looking from a consumer standpoint, giving the consumers a proactive notice and providing alternative options will increase customer satisfaction and loyalty. Lastly, predicting delays allows airlines to optimize their revenue management strategies. They can adjust ticket prices based on the likelihood of delays, potentially offering discounts to passengers willing to take flights with a higher probability of delays, while charging premium prices for more reliable flights.

Given this issue and our rationale for solving it, we propose to work on a project where, given data on airline delays/cancellations, we aim to construct an airline delay/cancellation prediction model. We have chosen to use the “2015 Flight Delays and Cancellations” dataset from Kaggle. While we do have more recent data, much of the flight data in recent years has been affected by the COVID pandemic, thus we concluded it would be more insightful to look into pre-pandemic data. We aim to be able to predict two main variables, one being whether we expect a cancellation, and the other being the expected delay time, if any. This dataset comes packaged with 28 different potential explanatory variables. This ranges from general variables, such as the exact date of the flight and the origin/destination airports, to more specific variables such as weather delay information and cancellation reasoning.

Since we are investigating two response variables, we expect to use a variety of techniques for prediction. Predicting airline cancellation can be separated into a binary classification problem, while predicting airline delay time is more of a regression problem. Given our extensive dataset, we expect to be able to find some correlation between some subset of the 28 explanatory variables and the response variables.

Thus, being able to predict airline delays will provide a necessary strategic advantage to today's dynamic aviation landscape. By investing in predictive modeling capabilities, we can enhance operational efficiency, improve customer experience, reduce costs, and uphold safety standards. We hope that our project will be able to provide the necessary insights for growth in this airline market.