Flight Cancellation Prediction
Capstone Project Proposal

Problem statement
Quick Fly (QF), an air cargo company, has experienced higher than industry average claim payments for delayed air cargo shipments caused by flight cancellations in the past year. What can QF do to bring down its claim payments by 30% to industry average by the end of this year?

Context
Quick Fly is a leading air cargo company providing guaranteed cargo delivery for its customers throughout the US. Recently, it was brought to the attention of the leadership that QF has been paying late delivery claim charges much higher than the industry average, mostly for shipments departing from Los Angeles International Airport (LAX). To improve customer satisfaction, keep the cost down and increase the profit, QF wants to identify and avoid those airlines, airports and outbound flight routes from LAX with a higher chance of delays and cancellations.

Criteria for success
- A list of the airlines, airports and flights that have the highest chances of cancellations.
- A model showing how the selection of different airlines, airports, flight routes, departure date/time, etc. will impact the likelihood of flight delays and cancellations.

Scope of solution space
- This project will focus on all flights departing from LAX. Flights departing from other airports will not be considered.
- Only flight cancellations will be investigated as the main cause for late cargo delivery claims. Other elements such as flight delays, diversions, ground transportation, etc. are not in the scope of this project.

Constraints
- Quality and integrity of the raw data
- Potential air freight costs increase as a result of avoiding certain airlines, airports, flights, etc.
- Elements other than flight delays and cancellations could also have contributed to late deliveries and hence claim payments, but are not investigated here.

Stakeholders
- Data Science team
- Operations team
- Airline Interface team
- Financial team

Data sources

Dataset of flight delays and cancellations published by the U.S. Department of Transportation's (DOT) from Kaggle
https://www.kaggle.com/datasets/usdot/flight-delays

Outline of problem solving approach

- Import, organize, define and clean all relevant datasets
- Explore to understand the relationship between data and features
    - What are the major features contributing to flight delays and cancellations?
    - Do certain departure airports/time, arrival airports/time, airlines, flights, etc have a higher chance of experiencing delays and cancellations?
    - Are there any data/elements that do not have identifiable correlationships with flight delays and cancellations and therefore can be ignored in the investigation?
- Pre-processing, standardize and train the dataset
    - Remove outliers and unreasonable elements.
    - If needed, split the dataset into testing and training subsets.
- Select, train and deploy a model to make predictive insights
    - Using the top features to build a model to estimate probabilities of delay/cancellation of a given flight.
    - Identify the top airlines, airports and flights with highest chances of delays/cancellations
- Document the work and share the finding

Capstone Project Deliverables

- A python model
- A final project paper or slides deck
- A GitHub repository containing all the work completed for each step of the project