Capstone Two - Project Proposal

Shirley Li

Problem statement formation

When booking flights for clients in 2023, we want to ensure that each flight has less than a 5% chance of being cancelled or delayed by more than an hour.

Context

Our clients are travelling for business or leisure and would obviously prefer not to have their itineraries disrupted. They're using our services to ensure a streamlined experience (from booking to travel itself); if they have cancelled or significantly delayed flights they will not have a good experience overall.

In addition if their flights are cancelled or significantly delayed there will be a cost to us: we will need to rebook their flights which may cost time and monetary resources.

Criteria for success

Success would be 1) given the data, each flight having less than a 5% chance of cancellation or >1 hour delay; 2) given all flights booked in 2023, no more than 5% of them being cancelled or delayed >1 hour.

Scope of solution space

We'll focus on the info for airline delay and cancellation 2009-2018 (we can pretend that this goes up to the present). We may find that particular airlines or airports are best to avoid, or that particular times of the year are better to avoid if possible.

Constraints

We actually don't have data from 2018-present, and a lot has changed since then! Our model will be hugely outdated. Airlines have massively changed their policies post-covid, some are in the middle of merges, etc.

In addition, the data we have doesn't give any info about ticket prices, which is also important to ensure business profitability. Even if we can minimize cancellations/delays, we don't want to spend an exorbitant amount of money on this one goal.

Stakeholders

Customer Experience team
Clients who don't want flights delayed

Data sources

Our data is on Kaggle: Airline Delay and Cancellation Data, 2009 - 2018 and is sourced from OST. We have info/columns on airline, departing airport, month, (perhaps also year if I need more columns/data,) time of departure, departure delay, time between cabin closed and wheels off, actual elapsed time between cabin doors closed and open, air time, carrier delay, weather delay, air system delay, and security delay.

Project Plan

We'll first do **data wrangling & exploration**: which columns are relevant, what data is available/missing, which flights were cancelled or had a delay of more than an hour (need to create a binary column for this). From here we'll explore which columns seem to be correlated with flight cancellation or arrival delay.

Perhaps we can see if it'd be easy to scrape more data for 2018-22 from the OST website and combine it with the data we already have.

Next we'll do **data preprocessing**, split into train/test sets, train a dummy regressor, and obtain baseline metrics (R squared, MSE, MAE).

Then we'll do **data training** to create a few models: a linear model, incorporate k best parameters and cross validation using GridSearchCV, and a random forest regressor. We'll test these and obtain performance metrics.

Finally we'll deploy our **model** to make predictions about which flights will be cancelled/delayed by >1 hour.

Deliverables

We will present a slide deck and provide a written project report summarizing the findings. Hopefully there will also be an "operationalizeable" piece of code: where the bookings team can enter an airline, airport, date, etc. and see the probability of a cancellation or delay.