**Problem Statement:**

To predict the flight prices given a dataset with information such as 'airline', 'source\_city', 'departure\_time', 'stops', 'arrival\_time', 'destination\_city', 'class', 'duration', 'days\_left', and, 'price'

**Dataset:**

Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 datapoints and 11 features in the cleaned dataset.

Link : <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

Uploaded dataset to github repo and connected it to Colab file

**Data Preprocessing:**

1. The original dataset had two files, one for business class and another for economy, a third fike was also available which was combined from the two and had some basic data cleaning performed on it
2. Dropped the columns that won’t have any impact on the model or prediction such as: ‘flight’ which was the flight code/number.
3. Used ColumnTransformer and,OneHotEncoder to encode categorical variables so that they can be used in the model
4. Used variable scaling wherever required

75% of the data was used for training the model and the remaining 25% was used for testing

**Models and Techniques Used:**

1. LightBGM

Accuracy: 99.98%

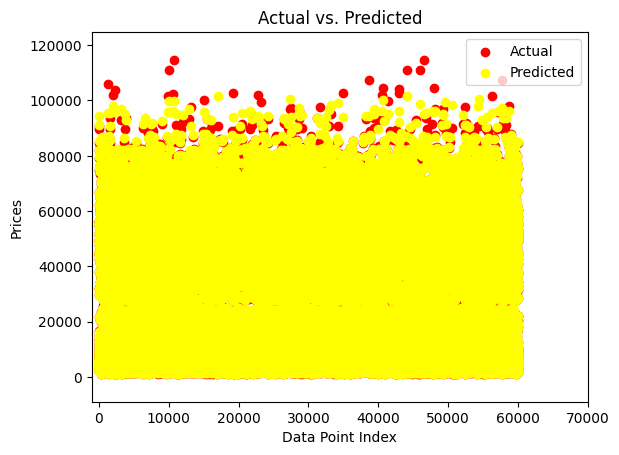
1. XGBoost

Accuracy: 99.97%

These models managed to perform really well and there isn’t much room for improvement

**Final Results:**

LightBGM and XGBoost models achieved exceptional accuracy rates of 99.98% and 99.97% respectively.



Plotting of the predicted vs actual prices in the test set

(high overlap)