**Kaggle Project 8**

**Tabular Classification with Reservation Cancellation dataset**

**Prepared by: Dwaipayan Mukherjee (2211569, AIA-AIML)**

**About the project**:

Customer behaviour and booking possibilities have been radically changed by online hotel reservation channels. Cancellations or no-shows cause a significant number of hotel reservations to be cancelled. Cancellations can be caused by a variety of factors, such as scheduling conflicts, changes in plans, etc. In many cases, this is made easier by the possibility of doing so free or at a low cost, which is beneficial for hotel guests but less desirable and possibly revenue-diminishing for hotels.

The dataset for this competition (both train and test) was generated from a deep learning model trained on the Reservation Cancellation Prediction dataset. Feature distributions are close to, but not the same, as the original.

The target variable is **booking\_status** that indicates if the booking was cancelled or not. The aim is to predict whether a customer cancels a reservation or not. The evaluation metric is Area Under the ROC Curve.

Steps that I have followed for completion of this project:

1. **Imports**:

First, I have done the necessary imports of libraries. Then I have imported the train, test, original and the sample submission datasets as Pandas Dataframes.

1. **Dealing with missing values**:

I have investigated for missing values in the train, test and original datasets and have found none.

1. **Understanding the data**:

I have checked out the summary statistics of the data, the unique value counts of the different columns and the unique values of the categorical columns.

1. **Exploratory data analysis**:

First, I have looked at the distribution of the target variable (i.e., **booking\_status**) for the train and the original data. I found that the train data was well representative of the original data and that in both the data, the distribution of **booking\_status** is a bit unbalanced.

Next, I have looked into the distribution of the categorical variables and then the distribution of the numerical columns. Lastly, I have explored the correlations between the features in the train data.

1. **Data Preparation**:

Here I have performed four distinct steps:

[5.1] I have added the original dataset to the training data. This has allowed me to train my models on a more extensive dataset.

[5.2] I have removed the **id** columns from the train and the test data.

[5.3] I have one-hot encoded all the categorical columns.

[5.4] I have separated out the features and target columns in the train data.

1. **Modelling**:

I have used four models for this classification exercise. I have always used 5-fold stratified cross-validation while training. The evaluation metric has been Area under the ROC curve.

[6.1] **Random Forest Classifier**:

The average AUC score came out to be 0.8989.

[6.2] **XGBoost Classifier**:

The average AUC score came out to be 0.9033.

[6.3] **LightGBM Classifier**:

The average AUC score came out to be 0.9001.

[6.4] **CatBoost Classifier**:

The average AUC score came out to be 0.8921.

1. **Final Prediction**:

Since all the model have performed more or less at par, so I have taken the mean of the predictions from each of the models to get the final prediction on the test data.