Machine Learning Project

Model Training and Testing

You are provided with a training dataset (**train.csv**) which consists of a sample of rides completed by a ride-hailing company. This consists of 17,176 trip entries out of which 1,681 are labeled as having incorrect fares. Your goal is to train a classification model based on the training dataset and test it.You can split the data to test/train as required for model building.

Stages of a Trip

Content of Training Data:

There are 14 fields in the given dataset and you may use as many fields as required to build the model.

tripid	Trip ID
additional_fare	Additional fare charged in rupees
duration	Duration of the trip in seconds from trip-start to trip-end
meter_waiting	Duration in seconds indicating the total time the vehicle was immobile from trip-start to trip-end
meter_waiting_fare	Fare for the time the vehicle was stopped within trip-start to trip-end duration. This may be due to traffic or other reason ex: stop at supermarket
meter_waiting_till_pickup	Time between trip-accept to trip-start
pickup_time	Date and time at pickup
drop_time	Date and time at drop off
pick_lat	Pickup latitude
pick_lon	Pickup longitude
drop_lat	Drop latitude
drop_lon	Drop longitude

fare	Trip fare
label	Label indicating if it is a correct or incorrect trip fare

Evaluation of the Model

You are provided with a testing dataset (**test.csv**) which consists of 8,576 entries. Format of the test.csv is the same as the training dataset with the exception of the label column not provided. Your task is to predict whether the fare is correct or incorrect for each entry in the test.csv.

Submission

For evaluation purposes, we make use of the kaggle platform. You can access the competition using the following url: https://www.kaggle.com/t/43bb8b5bdffd4353b6742b25dd8137e8
Submission to the kaggle should contain the header and be in the following format:
tripid, prediction
189123628,1
189125358,1
189125719,1
.....
See the **sample_submission.csv** file for the format. For each tripId in the test set, you must predict whether the fare is correct (1) or incorrect(0). Valid submission should contain a header followed by 8,576 predictions.

Submission Limitations:

You are allowed to make a maximum of 3 submissions per day.

Evaluation and Evaluation Metrics

Submissions will be evaluated based on their macro F1 score. Public Leaderboard is calculated with approximately 40% of the test data. The final results will be based on the other 60% after the contest deadline. You are allowed to select 2 submissions for final evaluation.

Please note that the data provided in the platform is dummy due to privacy concerns and you should be using the data files(train.csv and test.csv) given in the LMS for training, testing and evaluation.

Code and Version Controlling

In the final evaluation, you should be able to produce the code which resulted in your best score. Therefore you are advised to use proper version control throughout. However, make sure,provided data files are not committed to public repositories.

IMPORTANT

Data Security Policy : Downloading data from the learning management system(LMS), you agree to use reasonable and suitable measures to prevent persons who have not access to this from gaining access to the Competition Data. You agree not to transmit, duplicate, publish, redistribute or otherwise provide or make available the Data to any party not participating in the Competition. You agree to notify immediately upon learning of any possible unauthorized transmission or unauthorized access of the Data and agree to help rectify any unauthorized transmission. You agree that participation in the Competition shall not be construed as having or being granted a license (expressly, by implication, estoppel, or otherwise) under, or any right of ownership in, any of the Data. Further, competition organizers will not be held responsible in case you failed to adhere to Data Security Policy.