IML Challenge 2022

Challenge Guidelines

Second Semester, 2022

Dear students,

As part of the whole IML course experience, where you learn new concepts and algorithms used to tackle different data challenges, we are excited to release the IML Course Challenge. In this semester-long challenge you are given the hands-on opportunity to try out your developing machine learning skills and help you become better machine learning experts.

This challenge is a healthy competition between teams of students. We urge and expect you to practice Gracious Professionalism®: learn and compete like crazy but treat one another with respect and kindness while doing so. Help each other to improve, *even if you are not in the same team*.

So successfully participate in the challenge please read the following instructions carefully.

The Challenge

The challenge begins on *Thursday 25/3/2022* and ends at *16/6/2022*. During this time you will be solving the problem of predicting cancellations of hotel bookings. The training set composed of the feature matrix $\mathbf{X}_{\mathbf{train}}$ and the cancellation labels $\mathbf{y}_{\mathbf{train}}$ (the column named **cancellation_datetime**) can be found on IML.HUJI/challenge with a detailed explanation about the different fields available on Moodle. Please note that this data is messy and requires data wrangling. Use this data to train your model.

While the challenge takes place, a new test data file will be published (via Moodle) every Friday morning at 08:00. You are then welcome to work on your models and improve them. The test set contains only the feature matrix $\mathbf{X}_{\mathbf{test}}$. Your model computes the predicted labels $\hat{\mathbf{y}}_{\mathbf{test}}$ which you then submit for grading. Your goal is to minimize the misclassification loss over the test set. Submissions will be accepted until the following *Thursday at 23:59*.

Teams

To participate in the challenge you must form teams of 3 students only. Please register your team (with a cool team name) through the registration forum available on Moodle. If you are searching for team members please use the designated team members search forum on Moodle. The team's

name will be used for the challenge's leaderboard with the default name being the surnames of the team members.

Submissions

The weekly released test set consists of 500 samples not present in the training data. Used your trained model to predict the labels of these observations. Then, export the results in a `csv` file format with 500 rows (one for each observation) containing the predicted values and a single column named "predicted_values". The file name should be `id1_id2_id3.csv` where id_i corresponds to each of the team members' id number. This file will be automatically tested for prediction accuracy.

Predictions should be submitted via Moodle. Make sure to follow this format. Submissions that do not follow this exact format (e.g., bad filename, columns instead of rows, too many or not enough predictions) **will not be checked**.

In addition, you are also expected to provide your code for examination. You are expected to upload your code (for preprocessing of the data, fitting of the model and predicting values) to your *personal* fork of the git repository. Teams who make it to the leaderboard will have their code examined by the TAs. Failure in the examination may lead to the team's exclusion from the challenge.
A notebook fails if one of the following events happens:

- When running the notebook on that week's test-set the predicted values differ from those submitted. That is, if we take your code, feed in the test-set data and the predicted values do not match the values that you submitted - your submission not be considered for that week and might lead to exclusion from the challenge.
- The code uses libraries/languages that are not included in the course virtual environment.
- If we find signs of code sharing/pilfering between teams.

Grading

Each week the top 10% of teams enter the leaderboard. The leaderboard reflect the performance of the most recent submission only (i.e., each week is independent of the previous weeks).

Grading is computed as a function of the number of (not consecutive) weeks a team has made it into the leaderboard:

- **3 points for the first 3 weeks** and then 1 point per week. That is, if a team is in the weekly leaderboard for **3 weeks** (not necessarily consecutive) they get **3 points**.
- For any additional week that this team is in the leaderboard they get an additional **1 point** to a **total sum of 9 points** .
- Teams that are in the leaderboard for **less than 3** weeks get **no points**.

Please note that the overall bonus depends on meeting the course requirements - if you fail to meet them you won't get the bonus

Support

The course staff will provide **technical support only**. All other issues are not supported by us. You will find a technical support forum in the course Moodle.

Permitted Libraries and Tools

Students may use any library/algorithm covered by the course environment ‘iml.env‘. You may not use any programming language other than Python.

Review

Each week, after the scoring process has terminated, we will publish the following:

- A leaderboard - with three columns: team name, prediction score for current week and the number of weeks in the leaderboard
- The correct labels of that week’s test set - use the score and the labels wisely to improve your model!

The course team would like to pay its thanks to Agoda for providing the data and support.

We wish you good-luck!