The Analytics Edge (Summer 2023): Competition

Dates/Times

- Start of Kaggle competition: 20 July 2023
- End of Kaggle competition: 28 July 2023 (11:59 pm UTC)
- Submission of report: 4 August 2023

Kaggle Competition

Compeition link: https://www.kaggle.com/t/56bfe1d9ff734a1d9fedab24564160b7.

Goal: In this competition, you will be predicting the choice among bundles of safety features in cars. Every observation has four bundles of safety features from which one bundle is selected.

Data: The training data is provided in train.csv (total of 14250 observations) and the test data (total of 7315 observations) is provided in test.csv. The sample submission file is sample_submission.csv. You will need to submit your team's predictions in the format of the file sample_submission.csv.

Maximum Daily Submissions: Every team can make up to two submissions daily. Participants will need to wait until the next UTC day after submitting the maximum number of daily submissions.

Evaluation: Evaluation of the predictions is done using the multi-class loss version of log likelihood. In each observation, exactly one of the four alternatives (bundles) is chosen. For each observation, you must submit a predicted probability for each alternative. The metric that is computed to evaluate performance is the average negative log likelihood of the model defined as follows:

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log(p_{ij})$$

where:

$$n = \text{Number of observations}$$

$$m = \text{Number of alternatives} = 4$$

$$\log = \text{Natural logarithm}$$

$$y_{ij} = 1 \text{ if in observation i, alternative j is chosen and 0 otherwise}$$

$$p_{ij} = \text{Predicted probability in observation i, alternative j is chosen}$$

The benchmark performance is to predict a probability of 1/4 for each alternative for each observation. In the leaderboard, this benchmark shows up as my_sample_submission_1.csv with a score of:

$$\text{LogLoss of benchmark} = 1.38629$$

Lower the LogLoss metric of your model, the better is your performance. The values you submit need not necessarily sum up to 1 in the submission file. If it does not sum up to 1, the values will be automatically rescaled (each is divided by the sum) before evaluation. However I do not personally recommend this approach due to numerical instabilities and would suggest you provide predicted probabilities that sum up to 1 for each observation.

Ranking: It's conceivable that teams could overfit a solution - which might be great for winning a competition, but not valuable for a real-world application. To prevent this, 70 percent of the observations in the test set will be used to evaluate performance on the public leaderboard and 30 percent of the observations in the test set will be used to evaluate performance on the private leaderboard. You will be able to monitor your performance on the public leaderboard. However you will not know which observations are used for the public and private leaderboards respectively. The private leaderboard performance is kept a secret until after the competition deadline. Your final rankings are based on the private leaderboard. At the close of the competition, your best scoring submission on the public leaderboard (you do not need to select which one) will be used to score the performance on the private leaderboard.

Team Name: You are allowed only one Kaggle account per team. Name your account with the name of one team representative followed by the team's number. For example if the team representative is Walter White and team number is 1; name the team walter_white_1. Drop an email to Arjun and me with your team name. Team numbers will be provided by Arjun. Only use this account to make predictions on the Kaggle link.

Codes: You can only use R for the competition (any R package is allowed). You are free to use methods that were not covered in the class. You are free to refer to online resources. However you should only work on the projects within your groups and no help from other individuals is allowed.

Report

- The report is a short document (maximum of 8 pages, font size 12) containing a a high-level description of the approach you used, a short description of the results, and a brief discussion on interpretability and limits of the approach. An executive summary is not needed. The reports must be submitted using the eDimension submission inbox.
- When submitting the report, please also upload a zipped folder containing the code you developed. All files needed to run your code must be available and if more than one files are used, list them in a readme file. Your code must must produce the same results as you uploaded on Kaggle and we should be able to recreate the results.

Grading

- 15 marks - performance on public leaderboard
- 15 marks - performance on private leaderboard
- 10 marks - report