GSEM, University of Geneva
Machine Learning (‘S403011’)
Prof. Sebastian Engelke

Spring Semester 2023
Data Competition
Assistants: Olivier Pasche, Manuel Hentschel

Data Competition: Advertisement Impact Prediction

1 The competition

The competition is hosted on `https://www.kaggle.com/competitions/ml-unige-2023`. In order to join the competition, you will have to:

1. Create an account on `https://www.kaggle.com/` Please use a “Display name” that identifies yourself, for example your last name. When logged in, you can join the competition on

   `https://www.kaggle.com/t/c3039ddd013443d6896699c96fa76102`

   Please do not share this link with anyone else outside of the course.

2. When all team members have joined the competition, you can create a team.

The competition ends on **28 May 2023 at 23:59 Swiss time**.

2 Data set and goal

A company wants to assess the quality of their online advertisement campaign. Online users are the main interest in this campaign. The users see a web banner during their browsing activity. For each user, they want to be able to predict whether they subscribe to the advertised product through the advertisement banner, based on the information they have about them. To subscribe, the user has to click on the banner and then subscribe to the service. The target variable name is `subscription`.

3 Prediction Evaluation

The framework and the evaluation metric of the data competition are explained on the Kaggle website above. There are two different types of rankings on Kaggle:

1. **Public Leaderboard:** Each day you may (not mandatory) submit up to two prediction files. These predictions are directly evaluated on 30% of the test data (always the same, randomly chosen beforehand). This score is shown on the Public Leaderboard and gives you an indication of the accuracy of your prediction. The Public Leaderboard **does not count** for the final evaluation.

2. **Private Leaderboard:** Before the end of the Kaggle competition, you can choose one of your submissions for the final evaluation. After the end of the competition, the predictions of this submission are evaluated on the 70% remaining test data, resulting in the Private Leaderboard. The final scores in the Private Leaderboard determine the accuracy of your predictions and the competition winners. You should choose a final prediction that you believe performs best on new data, which might not necessarily be the best-scoring Public Leaderboard prediction.

4 Rules

1. You can participate in teams of up to 3 students.

2. Predictions should be based only on the training data and information from the Public Leaderboard.

3. You should use the python programming language. You are allowed to use any pre-built modules or packages in python, as long as they are explicit in your code.
4. No cheating of any kind.
5. You have to explain your main prediction approaches in detail in the final notebook as markdowns.

5 Prize

The first 3 teams in the Private Leaderboard receive bonus points for the course (which has 100 points in total).

- 1st place: all team members receive 10 bonus points.
- 2nd place: all team members receive 6 bonus points.
- 3rd place: all team members receive 3 bonus points.

6 Deliverables and grading

You will document your different approaches to solving the data competition and code in an "IPython Notebook", which is due on **28 May 2023 at 23:59 Swiss time**. The main notebook should be in `ipynb`-format, well-structured and well-documented. It should contain your python code corresponding to the Kaggle submission you selected for final evaluation. It should output the same exact csv file uploaded on Kaggle, so make sure that your code is reproducible.
Code that is not relevant in the main notebook can be submitted in additional notebooks and scripts. Make sure that all the code for your main approach, your analysis, and model comparisons (you can refer to the other notebooks's results), are in your main notebook.
The notebooks can be submitted on the `Moodle` course website where assignment modules will be created. Only one member per team has to submit the notebook, but make sure to mark the names of all team members and the Kaggle team name clearly on it. The main notebook should describe in detail the data set and how you approached the problem. A possible structure might be:

1. Introduction: Description of the data set, imports, notebook structure
2. Exploratory data analysis and feature engineering
3. Description of the best predictive model used, comparison of different methods, tuning parameters analysis, model selection approach
4. Best model diagnostics and final Kaggle prediction
5. Conclusion

You can concentrate on the best model, but you should also compare the results (training, CV and test errors) with the other approaches. The notebook should contain plots to illustrate your findings. Grading will be based on scientific correctness, originality and presentation. The analysis and code in the notebook(s) will count for 50% **of your final grade**.