

In this assignment we are going to focus on playing Counter Strike GO (not bad, right?). Well, hold your horses since we are going to use Machine Learning not quite for playing but for deciding who's is going to win a match.

If you are not familiarized with this game: *"CS:GO is a tactical shooter, where two teams (CT and Terrorist) play for a best of 30 rounds, with each round being 1 minute and 55 seconds. There are 5 players on each team (10 in total) and the first team to reach 16 rounds wins the game. At the start, one team plays as CT and the other as Terrorist. After 15 rounds played, the teams swap side. There are 7 different maps a game can be played on. You win a round as Terrorist by either planting the bomb and making sure it explodes, or by eliminating the other team. You win a round as CT by either eliminating the other team, or by disarming the bomb, should it have been planted."* (from the task description).

In particular, we are going to use the scenario proposed by the following Kaggle Dataset:

<https://www.kaggle.com/christianlillelund/csgo-round-winner-classification>

For this assignment, I present you a realistic machine learning scenario: you have a task, a description of what is expected, a dataset to work with and a brief description of the data. So, I will not give you any markdown or initial code. The only pre-processing step that I have done is to split the dataset into training and test so I can have a proper benchmark to evaluate your submission.

You have to make sense of the problem and the data and use your ML toolbox to solve the problem. In addition, I am also expecting from you a final report as a summary of your work, findings and conclusions. This report should be enough to understand your work, results, findings and conclusions even without deep technical knowledge on Machine Learning.

There is no need to say that I will help you in whatever doubt or problem you may have.

Details

- This is a binary-class classification task (the target variable is ***round_winner***)
- I will use accuracy to evaluate your predictions, so I recommend you use the same metric for the training process.
- You can apply any of the algorithms we have explained in class to address the task: Regression, Decision Trees, Naive Bayes, KNN...
- You can use R or Python
- You can use Dataiku for Data Cleaning and Feature Engineering
- This is a group assignment

Recommendations

- Look to the problem description. It contains valuable information on the dataset and the task
- Feature Engineering is important for this dataset. It includes categorical and numerical values, so, you should properly deal with each of them: remove outliers, remove useless features, look for NAs and missing data and impute them, create new features if you consider them useful, reduce the dimensionality if it makes sense.
- You can use whatever library you want. Nevertheless, I would recommend you review caret (for R) and sklearn (for Python) documentation. These libraries provide you with almost anything you need to solve the task.

Submission

You must send me:

- The Markdown, Notebook or code you have implemented to generate your results. This markdown has to provide a comprehensive description of the machine learning pipeline: ideas that you have tested, the problems that you have faced, and you should explain and justify any decision you take.
- Final report in a separate file where you summarize all your findings, the insights you could have extracted from your data, the conclusions of the different steps of your analysis, as well as the final conclusions from your results.
- The CSV with your predictions on the test set.

Evaluation

For grading your work, I will consider the 3 following aspects:

- The Machine Learning pipeline in your markdown: 40%
- The report summarizing your work, findings and conclusions: 40%
- The final accuracy in the competition (40%)