Predictive Modeling with Sports Data

Homework 2 (Mini-Project)

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting relevant code you want to include in your PDF. For Jupyter notebooks, you can save them as LaTeX or PDF before including them. Make sure your answers to each problem are clearly stated in the submitted PDF. The graders should not have to look through your code to find the solution.

Any code (Jupyter Notebooks and other source files) used to compute your answers should also be uploaded to Gradescope, along with a `readme` file telling the graders how to run your code if they need to, and if you are using any special packges. If the problem requires you to train and test a model, both the training and testing code should be submitted. Please indicate in your `readme` file if running your code requires special hardware (like a GPU, or 64GB+ of ram), a compilation step (if you decide to use Cython), or a signficant amount of time (longer than 5 minutes).

All of the questions in this homework will use the data in `soccer18.csv`. You may not use any data other than what is given (e.g., do not go online looking for additional datasets, and do not use the market-based probabilities given in homework 1).

1. (Average Goal Differentials)

   (a) For each game compute the *historical* average goal differentials for each team. That is, the average goal differentials using all games that occur *strictly before* the given game. If a team has played no previous games, we will say its average goal differential is zero. Throughout this part, only use games strictly before 2018 (`Y<18`). For the purpose of brevity, please rename the team "Evian Thonon Gaillard" to "Evian".

      i. Give a table containing the 7 games with the largest absolute disparity between the historical average goal differentials of the home and away teams. Your table should include the division (`Div`), the year (`Y`), the home and away

teams, the average goal differentials, the absolute disparity, and the number of games played by the home and away teams prior to that game. [Hint: You should have two games from Ligue 1, two games from La Liga, and three games from Serie A.]

ii. Repeat the previous part restricted to games where each team had previously played at least 100 games in our dataset (that is, 100 or more).

iii. Almost all games in the solution to part (i) come from the 2014 season (the first season in our dataset), but one comes from the 2017 season. In a few words, explain what is special about it.

(b) Fit a logit model to predict the probability of the home team winning (draws count as non-wins) using only an intercept term. You should fit the model on the data before 2018 (`Y<18`).

i. Report your coefficient value.

ii. Report the Brier score of your out-of-sample predictions on 2018 (`Y=18`). [See `sklearn.metrics.brier_score_loss` for Brier score computation in Python.]

(c) The intercept coefficient from the previous part is negative. Does this imply there is no home field advantage? In other words, if home teams are favored, shouldn't the intercept be positive? [Hint: A trick question.]

(d) Repeat part (b) using the intercept, and the historical average goal differentials from each team as features (three features in total).

2. (Go For It!) In the previous question we built some simple logit models to predict the probability of the home team winning. In this question you will improve those models. You may use any ML techniques you have learned (logit models, random forests, gradient boosting, neural networks, etc.), and you can use any features you come up with. Your goal is to get the best out-of-sample performance on 2018. Remember the following.

- You cannot use any additional data. Only use what is given in `soccer18.csv`. Below we describe some additional features in the dataset not present in the previous homework. These features are all computed after the game is completed, and are **not** known before the game is played.
  - `HS` and `AS` are the shots taken by the home and away teams, respectively.
  - `HST` and `AST` are the shots-on-target taken by the home and away teams, respectively.
  - `home_xG` and `away_xG` are the expected number of goals by the home and away teams, respectively. These are computed after the game is over by feeding shot data from the game into a statistical model.
- All of the data used to predict a specific game in the test data must come from strictly earlier games. More precisely, when predicting the outcome of a game

from the 2018 season, you can use data from earlier seasons, or from strictly earlier games in the 2018 season.

- Do not train/validate your model on the 2018 data. Use the data before 2018 (`Y<18`) to train and validate your model. To validate you can either use cross-validation, or split the training set into two sets, but in both cases keep the validation set after the training set chronologically.
- If we use market implied probabilities (which are not allowed on this assignment), the resulting out-of-sample Brier score is 0.2102. If your model achieves this score or better, you very likely have some type of bug.

In addition to uploading your code for training and testing the model, please include each of the following in your submission (we do not need snippets in your submission).

(a) Your out-of-sample Brier score on 2018.

(b) The type of model you fit (logit, random forest, etc.).

(c) A very brief summary of the features used in your model. Be terse but precise so that the reader can figure out exactly how your features are computed.

(d) A write-up of the process you used to build your model. Include a description of what ideas you had, how you evaluated the quality of those ideas, and how you incorporated them into the model. Your explanation can include interesting ideas that didn't end up working, and how you decided they weren't additive.

Your write-up should be as clear as possible. Good write-ups can receive full marks even if your out-of-sample Brier score isn't stellar.

In addition, please complete the following Google form which asks for some basic information about your solution. It is required, but you will not be graded on it. It will help us prepare materials for upcoming lectures. `https://docs.google.com/forms/d/e/1FAIpQLScAlxd_nxD-LoBSDC7JKK2OpREbhQhSRXaSbWBvpO2edNX4Rg/viewform`