Predictive Modeling with Sports Data

Homework 5

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. Do not share any code.

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.

Make sure your answers to each problem are clearly stated in the submitted PDF. Do not include code in your submitted PDF. Any important figures, results, plots, and tables generated by your code should be extracted and inserted into the PDF in a visually appealing way. Think of the PDF as a presentation you are making based on the results you uncovered in your analysis. Your submitted PDF should be self-contained: 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. If your code needs any special configuration in order to run, please include a `readme`. If the problem requires you to train and test a model, the training, validation, and testing code should all be submitted. If necessary, 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_full.csv` and `soccer18_shots.csv`. You may not use any data other than what is given.

1. (Shot Data)

   (a) There is one game in `soccer18_full.csv` where we have no corresponding shot data in `soccer18_shots.csv`. Investigate and give a potential reason the data is missing. You can exclude this game from our analyses in the remainder of this homework.

   (b) For each player, determine their total xG each season they played. A player is credited with xG if they made a shot or assisted in a shot. Display a table with the top 10 results. The table should have the player's name, the season, and the total xG as columns, and should be sorted in descending order by total xG.

(c) Assuming we are using the previous part as a measure of player ability, state a bias present in our ranking.

(d) Use the data in `soccer18_shots.csv` to add the following new columns to `soccer18_full.csv`:

- `OG_Home`, `OG_Away`: number of own goals (`OwnGoal`) scored that game by the home and away teams, respectively.
- `SP_Home`, `SP_Away`: number of shots that hit the goal posts (`ShotOnPost`) that game by the home and away teams, respectively.
- `HG_Home`, `HG_Away`: number of header goals (`Head` and `Goal`) that game by the home and away teams, respectively.

For each of the new column pairs (3 tables in total), give a table that lists the games in the dataset that achieved the maximum total number of the given shot (home plus away). For example, the first table will list all games with the most total own goals. Each table should have columns for the names of the two teams, the number of those shots by each team, the date, and the league. [Hint: Your tables should have 3 rows, 5 rows, and 1 row, respectively, with maximum values 3, 5, and 5, respectively.]

(e) Repeating the analysis performed in class, we will investigate to what extent own goals are due to chance or skill. More precisely, first construct a variable `ownGoalVar` analogous to `goalVar` from the lecture. It should be a difference (home minus away) of average own goal differentials using all prior games that season, and should incorporate a Bayesian prior of 0 with a weight of 5 games (just as was done in the lecture). As was done in the lecture, using only games where both teams combined have played at least 10 strictly prior games (`homeGames` + `awayGames` > 9), fit the following regressions using seasons 14-17 and report your coefficients.

i. An OLS model for the response logit(`pH`) using `ownGoalVar` as the only covariate other than the intercept.

ii. A OLS model for the response logit(`pH`) using `goalVar` and `ownGoalVar` as the only covariates other than the intercept.

iii. A OLS model for the response logit(`pH`) using `xgVar` and `ownGoalVar` as the only covariates other than the intercept.

Alternatively, you can use `smf.logit` to fit a "quasi-likelihood" model, but use `.fit(cov_type='HC1')` to get the robust standard errors (we may have some notes/video on this at some point).

Finally,

iv. Following the analysis in class, use the fits from the preceding regressions to determine to what extent own goals are due to (bad) luck or (lack of) skill. Briefly describe what each regression above suggests.

(f) Repeat part (e) for shots that hit a goal post.

(g) Repeat part (e) for header goals.

2. (Predicting Over/Unders) This problem is **optional** and should **not** be submitted. We are happy to discuss your findings in office hours.

   (a) The column `pOver` is the market implied probability that the given game will have more than 2.5 goals in total (home plus away). Using data from the 2018 season, compute the Brier score of the market at predicting which games have over 2.5 goals in total.

   (b) Using data from seasons 14-17, fit a model to predict if a game will have over 2.5 goals in total. State which features you used, and your Brier score on the 2018 season.