Predictive Modeling with Sports Data

Homework 3

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 `soccer18m.csv`. You may not use any data other than what is given.

1. (Elo Ratings) Implement the Elo rating system described in the notes (and at this wiki link). Let every game have a weight of $K = 40$, a home field advantage of 100, let $\beta = 400$, and start each team with a rating of 1000. Use the goal weighting formula given in the link above or the notes to determine G.

   (a) Move through every game in the dataset chronologically, and update each teams Elo rating accordingly. Create a table containing the top 3 teams from each division as ranked by Elo ratings at the end of the 2017 season. The row of the table should include the team's league (`Div`), the team's name, and their Elo

rating. The table should be sorted in increasing order by league and, within each league, in decreasing order by Elo ratings.

(b) Briefly describe a situation where it may be a good idea to temporarily use a higher value of K.

(c) Add the difference in Elo ratings (home Elo minus away Elo) as a feature in one of the models you worked on for homework 2. Include the out-of-sample Brier scores on the 2018 season before and after adding Elo. Make sure to use **pre-game** Elo ratings in your added feature, as post-game Elo ratings would leak information about the outcome of the current match.

2. (Market Implied Probabilities) In this dataset, we have the market implied probabilities `pH`, `pD`, `pA`, of a home win, a draw, and an away win, respectively.

(a) Using data from all seasons before 2018 (`Y<18`), find the 7 greatest upsets. That is, the seven games where a team (home or away) won but had the lowest probability of winning according to the market. Output a table where each row has the league, the season, the home team, the away team, `pH`, `pA`, the home goals, and the away goals.

(b) Is the market less accurate at the start of a season? Determine if this is true by computing the Brier score of the market (at predicting a home win) when each team has strictly fewer than 5 games played that season. Compare this against the Brier score of the market on all games. Use games from before the 2018 season (`Y<18`).

(c) Try to incorporate the market implied probabilities into one of your models from homework 2. **Important note: On the test data from the 2018 season, you can only use `pH`, `pD`, `pA` from STRICTLY EARLIER games and not the game being played.** These probabilities are **not** available until pre-game betting has finished, and are thus not available before the match has started. Submit an explanation of how you incorporated the market implied probabilities into your model, and your out-of-sample Brier scores on the 2018 season before and after your changes.