Predictive Modeling with Sports Data

Homework 4

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 `soccer20m.csv`. You can also use the files `hw4.csv` and `hw4_long.csv` that are derived from this data. You may not use any data other than what is given.

1. (DFL/Massey Ratings)

   (a) To compute our team priors we fit a model with formula

   $$\texttt{GD} \sim \texttt{GD\_Prev}$$

   on the 2015-2017 seasons (that is, each `GD` will take values from 2015, 2016, or 2017, and the corresponding `GD_Prev` will take values from 2014, 2015, or 2016, respectively). This fits the seasonal goal differential average for each team against

the average from the preceding season. The resulting model has an intercept that is nearly zero, and a coefficient on `GD_Prev` that is 0.8767. Why isn't it approximately 1?

(b) To convert logit market probability differentials to goal differentials we fit a linear model with formula

$$\texttt{GD\_Home} \sim \texttt{I}(\texttt{logit}(\texttt{pH}) - \texttt{logit}(\texttt{pA})) - 1$$

on the 2014-2017 seasons. Here `GD_Home` is the goal differential for the home team in a given game. The resulting coefficient was 0.4897.

i. Fit an analogous interceptless model for converting shot differentials to goal differentials, and report your coefficient.

ii. Fit an analogous interceptless model for converting expected goal differentials to goal differentials, and report your coefficient.

You can check your understanding of the problem by verifying the coefficient of 0.4897 above.

(c) Extend the DFL/Massey model in `hw4_model.ipynb` to include shot differentials and expected goal differentials. Adhere to the following guidelines.

- For every game there will be four rows in the X and Y matrices (instead of two). The additional two rows in the Y matrix will contain the shot differentials and expected goal differentials (both should be converted to goal scale). The additional two rows in the X matrix have the same values as the other two rows for the corresponding game.
- Leave all existing weights (for weighted least squares) alone and set the weights for the shot and expected goal differential rows to one.
- Use the existing Bayesian priors for home field advantage, and the team priors included in `hw4_prior.csv`.
- Decay the weights of the new rows in the same way the other rows are decayed.

Construct a table of ratings for teams in the EPL at the end of the 2017 season. Your ratings should be informed by all game results from 2017. Your submitted table should have columns for the team name, your end-of-year rating, and the Bayesian prior we used for the team rating at the start of 2017. The rows of the table should be sorted by end-of-year rating in descending order.

(d) Fit a logistic regression model to predict if the home team wins using data from seasons 2015-2017. Your model should have two features: the intercept, and the difference in ratings between the home and away teams. Report your model coefficients, and your Brier score on 2018.

(e) Fit a similar logistic regression model on seasons 2015-2017, but now use `pH` as the response instead of the home team winning indicator. Report your model coefficients, and your Brier score (at predicting home wins) on 2018.

(f) Tune the weights (used in weighted least squares) of the market prices, goals, shots, and expected goals used to generate the ratings. Try to pick reasonable values that lower the Brier score of our model from the previous part on 2018 (don't expect a huge improvement).

i. State what weight changes you made.

ii. Give a brief justification for the weights you selected.

iii. Report the Brier score of your tuned model on 2018.

iv. Report the Brier score of your tuned model on the pre-Covid data from the 2019 season. There is a several month break in the 2019 season data when Covid started. This computation will require you to compute priors for the 2019 seasons – use the same methodology as described below for the 2020 season.

(g) In this final part, we will use the tuned model from the previous part to forecast the 2020 season. To do this you will need to:

- Compute team priors for the 2020 season. Use the same methodology applied in the other seasons (i.e., base it off of the full season average goal differentials from 2019; that is, include both pre-Covid games and games during Covid). To compute a prior for newly added teams, use the average goal differentials for newly added teams in seasons 2015-2017. After computing all the priors for a given season, we then subtract the mean of all of the priors from that division in that season so the resulting priors have mean zero (in each division each season).
- Choose a new prior for home field advantage to account for the impact of Covid. Your chosen value cannot depend on the data from the 2020 season.
- Use the same multipliers for market logits, shots, and expected goals we fit earlier (don't refit them).
- Generate team ratings for the 2020 season.
- It is not required, but you can also consider changing the weight on the home field advantage prior.
- You can check that your team prior calculations are correct by calculating them on the earlier seasons and comparing with our numbers.

i. State what changes you made.

ii. Report your Brier scores on 2020 using our pre-Covid prior on home field advantage.

iii. Report your Brier scores on 2020 using your newly chosen prior on home field advantage.