DS-GA 3001 HW4

Abhishek Dendukuri

March 2021

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?

   > The coefficient is not approximately 1 largely due to regression to the mean. A coefficient of 1 suggests that team performance will be consistent from one season to the next but it will likely not play out the exact same way two years in a row due to factors such as strength of schedule, roster changes, and injury luck.

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

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

   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.

      > Coefficient for Shot Differential: **0.0819**

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

      > Coefficient for Expected Goal Differential: **0.9938**

(c) Extend the DFL/Massey model in hw4 model.ipynb to include shot differentials and expected goal differentials.

	Div	Y	Team	priorGD	R_Final
0	EPL	17	Man City	0.894973	1.452859
1	EPL	17	Liverpool	0.779622	1.002248
2	EPL	17	Tottenham	1.333306	0.940018
3	EPL	17	Chelsea	1.148744	0.794985
4	EPL	17	Man United	0.525850	0.712594
5	EPL	17	Arsenal	0.710411	0.435926
6	EPL	17	Southampton	-0.212395	-0.035730
7	EPL	17	Crystal Palace	-0.350815	-0.061418
8	EPL	17	Leicester	-0.396956	-0.229420
9	EPL	17	Newcastle	-0.549762	-0.249825
10	EPL	17	Everton	0.364359	-0.305655
11	EPL	17	West Ham	-0.443096	-0.356788
12	EPL	17	Watford	-0.696868	-0.373554
13	EPL	17	Bournemouth	-0.327745	-0.385510
14	EPL	17	Burnley	-0.420026	-0.465068
15	EPL	17	Brighton	-0.549762	-0.468206
16	EPL	17	West Brom	-0.235465	-0.527733
17	EPL	17	Stoke	-0.396956	-0.598827
18	EPL	17	Swansea	-0.627657	-0.604645
19	EPL	17	Huddersfield	-0.549762	-0.676251

(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.

Coefficient with `homeWin` as response: **1.0673**
Corresponding intercept: **-0.1823**
Brier Score: **0.21173**

(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.

Coefficient with `pH` as response: **1.0465**
Corresponding intercept: **-0.2217**
Brier Score: **0.21152**

(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.

Market Weight (`mkt_wt`)	`[10, 12, 15]`
Goal Weight (`goal_wt`)	`[1, 1.2, 1.4]`
Shot Weight (`SD_wt`)	`[1, 1.2, 1.4]`
Expected Goal Weight (`xGD_wt`)	`[1, 1.2, 1.4]`

I ended up choosing `mkt_wt = 15` and `goal_wt, SD_wt, xGD_wt = 1.4`.

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

I started with creating different values of weights and iterated through the DFL/Massey ranking function to calculate brier scores based on each combination of weights. I ended up choosing the combination of values that reported the lowest brier score.

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

Brier Score: **0.21149**

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.

Brier Score: **0.22167**

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

i. State what changes you made.

> The circumstances surrounding the 2020 season is likely similar to those of the end of the 2019 season - no fans generally lead to less of a home field advantage. I changed the `hfa_prior` from the value originally given to the mean of `GD_Home` for post covid games in the 2019 season. It may not be a perfect representation since there are limited games, but since the value is less than the original `hfa_prior` it encapsulates the idea that home teams have less of an advantage.

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

> Brier Score: **0.20884**

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

> Brier Score: **0.20834**