Amanda Kuznecov
anr431

Homework 4

DS-GA 3001.015
March 7, 2021

Question 1: DFL/Massey Ratings

Part a: The model used to compute our team priors has a coefficient on GD_Prev that is 0.8767. The coefficient is not approximately equal to 1 because we are performing a regression on the seasonal average against the previous seasonal average. We are not getting approximately 1 because we are regressing to the mean and accounting for noise that affected the previous seasonal average. By regressing to the mean, we are pulling the high outliers down and pushing the low outliers up. Therefore, we still predict a good seasonal average for goal differential but not quite as good as the previous season by using the 0.8767 coefficient as opposed to 1.

Part b.i: Coefficient for converting shot differentials to goal differentials: 0.0819
Part b.ii: Coefficient for converting expected goal differentials to goal differentials: 0.9938

Part c: Table for teams in the EPL at the end of the 2017 season, containing end-of-year ratings and Bayesian priors for team ratings at the start of 2017.

Team	priorGD	Post_Rating
Man City	0.894973	1.452859
Liverpool	0.779622	1.002248
Tottenham	1.333306	0.940018
Chelsea	1.148744	0.794985
Man United	0.525850	0.712594
Arsenal	0.710411	0.435926
Southampton	-0.212395	-0.035730
Crystal Palace	-0.350815	-0.061418
Leicester	-0.396956	-0.229420
Newcastle	-0.549762	-0.249825
Everton	0.364359	-0.305655
West Ham	-0.443096	-0.356788
Watford	-0.696868	-0.373554
Bournemouth	-0.327745	-0.385510
Burnley	-0.420026	-0.465068
Brighton	-0.549762	-0.468206
West Brom	-0.235465	-0.527733
Stoke	-0.396956	-0.598827
Swansea	-0.627657	-0.604645
Huddersfield	-0.549762	-0.676251

Part d: Fitting a logistic regression model to predict if the home team wins using data from seasons 2015-2017, we yield:
Intercept: -0.1823
Coefficient: 1.0673

Evaluating on the 2018 data:
Brier Score: 0.211731

Part e: Fitting a logistic regression model using pH as the response instead of the home team winning indicator, we yield:
Intercept: -0.2217
Coefficient: 1.0465

Evaluating on the 2018 data:
Brier Score: 0.211516

Part f.i: Weights of the market prices, goals, shots, and expected goals used to generate the ratings were tuned with the goal of lowering the Brier score of our model from the previous part on 2018 data.

mkt_wt:40, goal_wt: 3, shot_wt: 3, xg_wt: 3

Part f.ii: A combination of different weight values were used to fine tune the model, resulting in 225 combinations of weights and the Brier scores they yielded on the 2018 data. We expected the market information to be much more important than goal, shot and expected goal weights as there is already a lot of important information baked into the market probabilities. As discussed during lecture, we expect the market to be worth 10-20 times that of the goal differential, so we would expect that the market weights would be 10-20 times larger than goal weights. We would also expect the other weights (shots and expected goals) to be around the same value as the goal weights. Based on previous models built, expected goals would likely have the least effect and therefore we would expect the weights on those to be one of the lowest compared to the others. In reality, the weight placed on goals, shots and expected goals turned out to be exactly the same, indicating their relative importance to the market weight was equal across the board.

Market weights of [10,20,30,40,60], goal weights of [1,2,3,4,5], shot weights of [1,2,3] and expected goal weights of [1,2,3] were all evaluated during the tuning loop.

Part f.iii: Brier score of tuned model on 2018 data: 0.211357

Part f.iv: Brier score of tuned model on 2019 data: 0.221324

Homework 4

Part g.i: The home field advantage prior was originally set to 0.3739. However, during Covid, it is likely that home field advantage is not as prevalent as previous seasons. For this reason, the home field advantage prior was reduced to 1/3 of its original value. As discussed in class, it seems that home field advantage has decreased greatly and likely by about 33%. This is due to many reasons, for example the fact that there are no fans in the stands, and that the opposing team doesn't have as much travel. The prior was therefore adjusted to 0.12463 to reflect the lower influence and prevalence of home field advantage.

Part g.ii:
Brier scores on 2020 data using our pre-Covid prior on home field advantage: 0.207814

Part g.iii:
Brier scores on 2020 data using newly chosen prior on home field advantage: 0.207572