DS-GA 3001 HW3

Abhishek Dendukuri

February 2021

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, 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.

	Div	Team	Elo
0	Bundesliga	Bayern Munich	1350.621424
1	Bundesliga	Schalke 04	1159.177933
2	Bundesliga	Hoffenheim	1142.152808
3	EPL	Man City	1429.659400
4	EPL	Tottenham	1283.911192
5	EPL	Man United	1258.732460
6	La_Liga	Barcelona	1415.462495
7	La_Liga	Real Madrid	1306.832652
8	La_Liga	Ath Madrid	1220.575107
9	Ligue_1	Paris SG	1352.520538
10	Ligue_1	Monaco	1264.722861
11	Ligue_1	Lyon	1240.383883
12	Serie_A	Juventus	1414.044716
13	Serie_A	Napoli	1337.681320
14	Serie_A	Roma	1282.244900

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

A situation where it may be appropriate to use a higher value of K would be in games with higher stakes. Examples include Eurocup or other intercountry/intercontinental tournaments, the Olympics, and the World Cup.

(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.

`Brier Score Original`	0.21503
`Brier Score with Elo`	0.21417

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.

	GameID	Div	Y	HomeTeam	AwayTeam	pH	pA	FTHG	FTAG
0	2677	La_Liga	16	Barcelona	Alaves	0.891147	0.028831	1	2
1	2128	La_Liga	14	Barcelona	Malaga	0.875453	0.040021	0	1
2	1988	La_Liga	14	Barcelona	Celta	0.861781	0.043664	0	1
3	4291	Bundesliga	15	Bayern Munich	Mainz	0.856920	0.044404	1	2
4	3081	La_Liga	17	Real Madrid	Betis	0.876513	0.048646	0	1
5	2641	La_Liga	15	Levante	Ath Madrid	0.052018	0.798875	2	1
6	4008	Bundesliga	14	Bayern Munich	M'gladbach	0.821218	0.054292	0	2

   (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`).

   As shown in the briar scores below, the market actually isn't less accurate at the start of the season. If we change the number of games played to `Y<4`, the brier score improves which on the surface level doesn't seem to make sense and might be due to high variance.

`Brier Score (Whole Season)`	`0.21061`
`Brier Score (less than 5GP)`	`0.21058`
`Brier Score (less than 4GP)`	`0.20839`

   (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.

I incorporated market probabilities similar to the way I incorporated historical goal differential - I melted the original dataframe to get the team names in one column and then took a rolling average of pH, pD, and pA for each game that a team plays.

I subsequently created a new train/test split with pD and pA removed because with rolling averages, some if not most of the probabilities add up to greater than 1, making the model slightly finicky. Ultimately, we care most about home wins, so that's why I found a new brier score with only pH.

Brier Score Original	0.21503
Brier Score with Elo	0.21417
Brier Score with pH, pD, pA:	0.21435
Brier Score with pH only:	0.21329

As we can see, the model with only pH (and Elo also factored in) performed the best.