Amanda Kuznecov
anr431

Homework 2

DS-GA 3001.015
February 17, 2021

Question 1: Average Goal Differentials

Part a.i: Table containing the 7 games with the largest absolute disparity between the historical average goal differentials of each team before the 2018 season. GD is historical average of goal differentials, Abs_Disparity is the absolute disparity between goal differentials of the home and away teams, and GP is number of games played.

	Div	Y	Team_Home	Team_Away	GD_Home	GD_Away	Abs_Disparity	GP_Home	GP_Away
212	Serie_A	14	Sassuolo	Sampdoria	-3.5	1.000000	4.500000	2	2
31	Ligue_1	14	Evian	Paris SG	-3.5	1.000000	4.500000	2	2
5507	Ligue_1	17	Strasbourg	Lille	-4.0	0.078261	4.078261	1	115
145	Serie_A	14	Empoli	Roma	-2.0	2.000000	4.000000	1	1
101	La_Liga	14	Elche	Granada	-3.0	1.000000	4.000000	1	1
210	Serie_A	14	Palermo	Inter	-0.5	3.500000	4.000000	2	2
82	La_Liga	14	Cordoba	Celta	-2.0	2.000000	4.000000	1	1

Part a.ii: Table containing the 7 games with the largest absolute disparity between the historical average goal differentials of each team before the 2018 season. Table only includes games where each team has previously played at least 100 games.

	Div	Y	Team_Home	Team_Away	GD_Home	GD_Away	Abs_Disparity	GP_Home	GP_Away
5055	La_Liga	16	Granada	Barcelona	-0.875000	2.192308	3.067308	104	104
7265	La_Liga	17	Levante	Barcelona	-0.705357	2.140000	2.845357	112	150
5325	La_Liga	16	Granada	Real Madrid	-0.936937	1.900000	2.836937	111	110
6762	La_Liga	17	Las Palmas	Barcelona	-0.623762	2.208633	2.832395	101	139
7151	La_Liga	17	La Coruna	Barcelona	-0.621622	2.142857	2.764479	148	147
4962	La_Liga	16	La Coruna	Barcelona	-0.519608	2.225490	2.745098	102	102
6300	La_Liga	17	Barcelona	La Coruna	2.186047	-0.527132	2.713178	129	129

Part a.iii: It is noticeable that all teams (except Lille) who played in the 7 games from Part 1a.i had only previously played 1 or 2 games. Therefore it makes sense as to why the absolute disparities are so large, because if a team played extremely poorly or extremely well in their first couple games, their average goal differential would be "artificially" high or low since it is only based on a single game. In the case of the 2017 match involving Lille vs. Strasbourg, Lille has already played 115 games in Ligue 1, however, this match happens to be Strasbourg's second match in this division. It seems they had been promoted up from Ligue 2 at the end of the 2016 season. Although Strasbourg has historical information about goal differentials, it is not stored in this dataset since it is missing Ligue 2 data. Since Strasbourg lost by 4 in their first match, their average goal differential is artificially low and biased based on their previous game. Therefore, the comparison between the average goal differential for Lille and Strasbourg is very large.

Part b: Predicting probability of the home team winning using a logit model with only an intercept-term.

Coefficient: -0.1669
Out-of-sample Brier Score: 0.247356

Part c: Once we take the log odds of the coefficient, we can see that the probability of the home team winning is 45.8%, leaving the remaining 54% accounting for the probability of away team winning and probability of a draw occurring. The probability of the away team winning is 29.3%, meaning that the home team wins 1.56 times more than the away team, indicating that there is in fact home field advantage.

Part d: Predicting probability of the home team winning using a logit model with historical average goal differentials from each team and an intercept term.

Coefficient: -0.1791
Out-of-sample Brier Score: 0.217261

Question 2: Go For It!

Part a: Out-of-Sample Brier Scores for model runs

Model	Out-of-Sample Brier Score
1	0.216032
2	0.215917
3	0.215809
4	0.215500
5	0.215359

Part b: Type of Model Fit
All models used in the analysis were logit models.

Part c: A summary of the model components is provided below. All features have 'Home' and 'Away' components except feature 'Fav_at_Home'. Details on each feature are also provided.

Model Description:

Model 1:
Historical Average Goal Differential, Historical Average Expected Goal Differential

Model 2:
Historical Average Goal Differential, Historical Average Expected Goal Differential, Rolling Win Percentage

Model 3:
Historical Average Goal Differential, Historical Average Expected Goal Differential, Historical Average Shots on Target Differential

Model 4:
Historical Average Goal Differential, Historical Average Expected Goal Differential, Historical Average Shots on Target Differential, Winning Streak, Losing Streak

Model 5:
Historical Average Goal Differential, Historical Average Expected Goal Differential, Historical Average Shots on Target Differential, Winning Streak, Losing Streak, Favoured Team at Home

Feature Description: In creating all features, it is noted that the entire dataset was sorted

on Date to ensure chronological order.

Historical Average Goal Differential: Using all games that occurred strictly before any given game, a team's historical average goal differential was calculated by first determining a given team's goal difference against their opponent (ie. Goals For - Goals Against). Next, a cumulative average was calculated on this goal differential, grouped by team. The cumulative average was shifted down 1 row per team as there is no historical information gathered prior to the first game of play. This feature is stored for both home and away teams as: 'GDcumAvg_Home' and 'GDcumAvg_Away'.

Historical Average Expected Goal Differential: Using all games that occurred strictly before any given game, a team's historical average expected goal differential was calculated by first determining a given team's expected goal difference against their opponent (ie. Expected Goals For - Expected Goals Against). Next, a cumulative average was calculated on this expected goal differential, grouped by team. The cumulative average was shifted down 1 row per team as there is no historical information gathered prior to the first game of play. This feature is stored for both home and away teams as: 'xGDiff_cumAvg_Home' and 'xGDiff_cumAvg_Away'.

Rolling Win Percentage: Using all games that occurred strictly before any given game, a team's rolling win percentage was calculated by first determining if the team won a game. For each team, a cumulative sum of the number of games won was recorded by adding up historical game wins and shifting down 1 row per team as there is no historical information gathered prior to the first game of play. Next, the number of games played by any given team was calculated, by grouping by team and taking the cumulative count of games, also shifted down by 1 row per team. Finally, a rolling win percentage was calculated by dividing number of historical wins by number of games played. In cases where a team has yet to play a game, a perfect record was assumed, setting win percentage to 1 to start. This feature is stored for both home and away teams as: 'Winpct_Home' and 'Winpct_Away'.

Historical Average Shots on Target Differential: Using all games that occurred strictly before any given game, a team's historical shots on target differential was calculated by first determining a given team's shots on target difference against their opponent (ie. Shots on Target For - Shots on Target Against). Next, a cumulative average was calculated on this shots on target differential, grouped by team. The cumulative average was shifted down 1 row per team as there is no historical information gathered prior to the first game of play. This feature is stored for both home and away teams as: 'STDiff_cumAvg_Home' and 'STDiff_cumAvg_Away'.

Winning/Losing Streak: Grouped by team, winning and losing streaks were recorded by grouping by wins and losses and performing a cumulative count if the binary 'Win' value

below was the same as the current 'Win' value. The binary 'Win' column was shifted down 1 row per team, and wins and streaks were set to 0 when the team had yet to play a game. A 'Win_streak' column was created by taking the product of the binary 'Win' column and the 'Streak' column (containing the cumulative streak count). A separate binary 'Loss' column was created based on the binary 'Win' column, and a separate 'Loss_streak' column was created by taking the product of the binary 'Loss' column and the 'Streak' column (containing the cumulative streak count). These features are stored as 'Win_streak_Home','Win_streak_Away','Loss_streak_Home','Loss_streak_Away'.

Favoured Team at Home: Based on the team's goal differential at the start of the game, if the home team's goal differential is greater than the away team's goal differential, they are considered to be favoured to win. Since they are at home, it may be even more advantageous. If the home team is favoured to win, this feature is set to the difference between the home team's historical average goal differential and the away team's historical average goal differential. This places extra weight on the advantageous circumstances where the team is already favoured to win and they are at home where they may have home field advantage. If the home team is not favoured to win, the feature is set to 0. This feature is stored as 'Fav_at_Home'.

Part d: Using the previous models in Part 1 as a baseline, it was noted that adding the historical average goal differential feature added predictive value to the model. I decided to start with this last model and build on by adding more features, however if the features added did not add predictive value or if the coefficient signs were in the wrong direction, the features were removed from future models.

To start, I added the historical average of expected goals differential as this is a pretty good indicator of team performance and their quality of play in terms of shots. Although shots may be seen as a good feature, shots do not actually tell us how well a team is playing. For example, a team can be taking lots of shots, but they may not even be good shots. Therefore, if we look at the quality of their shots in terms of the expected goals, we can get a good sense of how a team is playing by taking the historical average of a team's expected goals differential against their opponent. Adding this feature was extremely helpful, even more so than historical average goal differential in terms of Z-scores. Adding these features significantly improved the Brier Score, and therefore they were kept in future models.

Next, I added a rolling win percentage for each team. I thought this would be indicative of the team's overall performance. Although adding this feature did improve the Brier score, it did not appear to add much predictive value as the Z-scores were quite high. Even worse, the coefficients of the win percentages were in the wrong direction. For these reasons, this feature was removed from future models.

Next, I added the historical average shots on target differential. I thought this would add value because this gives us a way to quantify shots that would have gone in had they not been blocked. It is also indicative of how well a team is playing - they are throwing quality shots at the net and not just in any random direction like regular shots. This feature is likely already captured by the historical average of expected goals differential, which is why it did not add as much predictive value as expected. However, adding these features still improved the Brier score, so they were kept for future models.

Next, I added a winning/losing streak feature. I felt that this captured when teams were really hot and had good momentum/high morale. It is more likely that a team who is on a roll will continue to win. Losing streak is also good at capturing teams that are in a slump and likely to continue down that path. Both features were added to the model. It was quite noticeable that winning streak had much more predictive power than losing streak.

Lastly, I added a feature to upweight home teams who were favoured to win (ie. when the home team's historical average goal differential is greater than the away team's). These teams were upweighted by the difference between their historical average goal differential and the away team's. I felt that the weighting could add predictive power because if a team is already favoured to win, just based on their historical average goal differential, they are even more likely to win if they are playing at home with home field advantage. The feature did not add much predictive power, but it appeared to be more significant than historical average goal differential and losing streak features. The Brier score also improved with this new feature added. The model was run with and without features for losing streak and historical average goal differential, however the Brier score worsened after removing those.

Overall, it seemed that the features with the most predictive power were historical average expected goal differential, winning streak, and historical average shots on target differential.

Another idea that was explored but did not work was trying to evaluate quality of a win. For example, if a huge underdog beat a team, this could be upweighted significantly because the team played extremely well against a difficult opponent. On the other hand, the highly favoured opponent could be penalized the same amount. Another example is if a highly favoured team won by a lot, this would be given a low weighting because the event is expected and the team is just playing at their level. Their opponent, the big underdog would not be penalized by much since this is their level of play. I tried creating a few functions for this, however there did not seem to be a good way to quantify "quality".