Amanda Kuznecov
anr431

Homework 5

DS-GA 3001.015
March 20, 2021

Question 1: DFL/Massey Ratings

Part a: The game between Bastia and Lyon on April 16, 2017 has no corresponding shot data. After a quick Google search, it appears that this game was actually never finished! Bastia fans stormed the field and attacked the Lyon players, who then refused to continue play. The game was abandoned at half-time and since the game was never completed, the match statistics were not recorded as part of the shot data because all data stored in there is from completed matches.

Part b:

Player	Season	Total xG
Lionel Messi	14	53.503719
Cristiano Ronaldo	14	52.909403
Luis Suárez	15	49.109693
Neymar	15	46.124972
Cristiano Ronaldo	15	44.195410
Lionel Messi	17	44.046686
Lionel Messi	15	42.973189
Robert Lewandowski	18	42.756660
Lionel Messi	18	41.332334
Lionel Messi	16	40.840305

Part c: It would be biased to measure player ability solely by expected goals because xG will automatically favour players who get more scoring chances/opportunities. Some players have more playing time so it is likely they will have more expected goals. Also we would not be able to compare all players fairly because it is more likely that forwards and midfielders will have scoring chances, whereas defencemen will have very few to no expected goals, and goalies will have none.

Part d: New features were created for each game which recorded the number of own goals, number of shots that hit the post, and number of header goals for the home and away team. For each new feature, games that achieved the maximum total number (home plus away) are provided below.

Own Goal:

Team_Home	Team_Away	OG_Home	OG_Away	Date	Div
Southampton	Sunderland	0.0	3.0	2014-10-18	EPL
QPR	Liverpool	2.0	1.0	2014-10-19	EPL
Empoli	Napoli	1.0	2.0	2015-04-30	Serie_A

Shots that hit the goal post:

Team_Home	Team_Away	SP_Home	SP_Away	Date	Div
Sociedad	Almeria	4.0	1.0	2014-09-21	La_Liga
Getafe	Espanol	4.0	1.0	2016-01-17	La_Liga
West Ham	Liverpool	2.0	3.0	2017-05-14	EPL
Barcelona	La Coruna	5.0	0.0	2017-12-17	La_Liga
Hoffenheim	Mainz	4.0	1.0	2018-12-23	Bundesliga

Header goals:

Team_Home	Team_Away	HG_Home	HG_Away	Date	Div
Cagliari	Fiorentina	3.0	2.0	2016-10-23	Serie_A

Part e.i: An OLS model for the response logit(pH) using ownGoalVar as the only covariate other than the intercept.

Intercept: -0.2238
Coefficient: -2.5428

Part e.ii: A OLS model for the response logit(pH) using goalVar and ownGoalVar as the only covariates other than the intercept.

Intercept: -0.2072
Coefficient (ownGoalVar): 0.3478
Coefficient (goalVar): 0.9648

Part e.iii: A OLS model for the response logit(pH) using xgVar and ownGoalVar as the only covariates other than the intercept.

Intercept: -0.20442
Coefficient (ownGoalVar): -0.1177
Coefficient (xgVar): 1.2263

Part e.iv:

When modeled on its own, the large negative coefficient on ownGoalVar suggests that there is a strong luck component present that will not help us predict the outcome of future games. This makes sense since own goals are often due to mishaps and accidental ball movement.

When ownGoalVar is modeled with goalVar, we expect a blended weight. The coefficient on goalVar is positive, indicating that the skill information in goals will help us predict outcomes of future games. However, since goals are a bit noisy, the coefficient is not as large as it is

reduced by the luck component. The coefficient on ownGoalVar adjusts the weight that goalVar puts on own goals. Since some of the luck and noise is already captured in goalVar, this gets reflected in the coefficient for ownGoalVar. This suggests that both features provide some information to us in order to predict game outcomes, and since the weight is larger on goalVar, it is more valuable, however both features are very noisy.

Finally, when ownGoalVar is modeled with xgVar, we expect a blended weight once again. However, since xgVar is much less noisy, the signal-to-noise ratio is much larger, which explains the larger weighting on xgVar. As expected, ownGoalVar is luck-based which is why a negative weighting is placed on the feature when modeled with xgVar.

Therefore, for a team to have an own goal situation is the result of bad luck.

Part f.i: An OLS model for the response logit(pH) using shotonPostVar as the only covariate other than the intercept.

Intercept: -0.2167
Coefficient: 2.2059

Part f.ii: A OLS model for the response logit(pH) using shotonPostVar and ownGoalVar as the only covariates other than the intercept.

Intercept: -0.2060
Coefficient (shotonPostVar): 0.6177
Coefficient (goalVar): 0.8855

Part f.iii: A OLS model for the response logit(pH) using xgVar and shotonPostVar as the only covariates other than the intercept.

Intercept: -0.2040
Coefficient (shotonPostVar): 0.3255
Coefficient (xgVar): 1.1770

Part f.iv: When modeled on its own, the feature shotonPostVar has a large positive coefficient, indicating that there is skill information baked into the feature that will help us predict the outcomes of future games.

When modeled withgoalVar, this results in a blended weight between the two. The coefficient on shotonPostVar decreases but remains positive and is slightly less than the coefficient on goalVar. This is because the goalVar feature provides more predictive value than shotonPostVar since goalVar represents shots actually going in the net. However, the skill-based components of shotonPostVar are still reflected in the weighting, because even though the shots did not go in the net, there was still a lot of skill involved in setting up and aiming the shot.

Finally, when shotonPostVar was modeled with xgVar, the coefficient on xgVar is much

greater than the coefficient on shotonPostVar. This is because of the predictive power provided by the feature as well as the level of skill baked in. The signal-to-noise ratio is much larger on xgVar, however shotonPostVar still provides information on skill-based components to aid in game outcome prediction.

Therefore, for a team to have a shot on post is the result of skill. Note that shots are on a different scale than goals, but in general they follow this same pattern.

Part g.i: An OLS model for the response logit(pH) using headGoalVar as the only covariate other than the intercept.

Intercept: -0.2182
Coefficient: 1.9639

Part g.ii: A OLS model for the response logit(pH) using goalVar and headGoalVar as the only covariates other than the intercept.

Intercept: -0.2072
Coefficient (headGoalVar): -0.0982
Coefficient (goalVar): 0.9686

Part g.iii: A OLS model for the response logit(pH) using xgVar and headGoalVar as the only covariates other than the intercept.

Intercept: -0.2042
Coefficient (headGoalVar): 0.1637
Coefficient (xgVar): 1.2077

Part g.iv:

Modeled on its own, headGoalVar has a very large positive coefficient, indicating its strong predictive power and skill-based information.

When modeled with goalVar, a blended weight between the two is expected. However, since the coefficient headGoalVar is now negative, it appears that the value of headGoalVar has been significantly reduced, likely because the skill from this feature is already captured in goalVar, and that too much weight was previously placed on the headGoalVar feature. Perhaps, it was due to the amount of noise in the headGoalVar already being captured by goalVar.

When modeled with xgVar, the coefficient on xgVar is much larger than the headGoalVar coefficient because it is less noisy as the signal-to-noise ratio is much larger. It appears that the headGoalVar provides some information on skill but it also appears to be a bit noisy when modeled with goalVar. In this instance, since headGoalVar has a positive coefficient, it seems that the feature is also skill-based.