Home Field Advantage in Various Soccer Leagues

Johnny Weng

Our goal in utilizing data analytics for our final presentation was to identify exploratory findings on how home-field advantage affects different soccer leagues. Our team utilized two different data sets based on in-game soccer statistics for six soccer leagues from the 2010-2016 seasons: Bundesliga, La Liga, Ligue 1, Major League Soccer (MLS), Premier League, and Serie A. We identified some of this data from *Kaggle,* and we created some of the data based on our own findings. We utilized these six leagues because other than MLS, the other five are arguably the top competitive leagues in the world. We felt that adding MLS would be a good outlier to use in the data and hypothesized that our findings for MLS would be different from the others.

For our first data set in which we utilized for clustering, we used the following variables: home wins, home draws, home losses, home goal differential, home/away win differential, and home points. For our second data set that we utilized for logistic regression, forecasting, and classification and ROC curves, we utilized game data with the following variables: home team goals, away team goals, half time home goals, half time away goals, home team wins, home shots, away shots, home shots on target, away shots on target, home fouls, away fouls, home corners, away corners, home yellow cards, away yellow cards, home red cards, and away yellow cards. For the second set of data, we could not identify a trustworthy source for MLS, so it only consisted of the five other leagues.

Since many sports fans and teams are always given the adage that home field (or court, or ice, etc.) advantage exists, our team wanted to exploit it in more detail to identify what factors go into that advantage. We feel that this information could be beneficial not only for professional soccer teams, but teams in any sport as they could utilize the data to figure out how to exploit (or deter) what variables go into winning a game at home. We feel that through cluster analysis, logistic regression, forecasting, and classification and roc curves, that we can provide analysis into factors such as what goes into winning a home game, what leagues have similar and contrasting home field advantages, and what expectations that can be identified from viewing this data.

Prior to utilizing our analytical tools, we wanted to identify a hypothesis on what we may be able to expect through the various use of our tools. To compare and contrast, we calculated home win percentage (total wins, home win percentage (total games, including draws), home points percentage, home goal differential, and home and away win differential. We figured this would give us a good idea of what league had the best home field advantage, but not why; something we would identify more during our logistic regression analysis. Results are posted in Table 1.

Table 1

Various Home Field Advantage Statistics For Soccer Leagues (2010-2016)

Variable	Bundesliga	La Liga	Ligue 1	MLS	Premier League	Serie A
Games	34	38	38	34	38	38
Home Win % (Total Wins)	59.11%	64.88%	62.60%	68.89%	63.21%	64.18%
Home Win % (Total Games)	44.82%	48.47%	43.89%	49.48%	46.30%	45.85%
Home Points %	57.61%	62.13%	60.01%	64.85%	60.58%	60.92%
Home Goal Differential	5.53	9.32	6.80	8.33	7.64	6.84
Home/Away Win Differential	2.39	3.81	3.16	4.38	3.43	3.44

Note: Amount of games are as of the conclusion of the 2016 season(s).

In viewing these statistics without additional analytical insight, it’s easy to see that MLS seems to have the best home-field advantage. However, what this doesn’t tell us is why they have the best home field advantage. This is one of the reasons why we decided to explore some of the reasons why this exists, and to prove that data analytics can provide much more insight than simply viewing sports summary statistics in the newspaper or at various sports sites.

One of our first curiosities was to identify what leagues had similar home field advantages, or in our case, similar home statistics. We hypothesized that each league would probably fall into its’ own cluster since they operate independently of one another. However, after we performed a cluster analysis with an average of each league season with the aforementioned variables provided, we came upon some unique findings in our dendrogram (shown in appendix, due to size). We ended up identifying three unique clusters of 19, 13, and 10 league seasons each. In one cluster, which we call the “Premier Serie A Ligue” cluster, Premier League, Serie A, and Ligue 1 made up 17 of the seasons, and the teams of Paris St. Germain, Juventus, Manchester United, and Manchester City won 12 of the seasons in that cluster. The next cluster, we called the MLS-La Liga connection, as both leagues made up eight of the ten seasons in that cluster. The last cluster, in which we titled Bundesliga domination, consisted of all Bundesliga seasons and a mix and match of some others. While our hypothesis was correct in regards to Bundesliga, we identified through clustering that home statistics can be both similar and different between different leagues. Upon further inspection in identifying the variables that we used, some statistics were very similar within its’ cluster whereas others were not; this occurred throughout each of the three clusters, so we didn’t feel comfortable identifying our clusters based on most home wins, most home draws, etc. Instead, we felt our reasoning for lumping the clusters on similarity between all variables used was the best way to analyze the clusters.

In using logistic regression, we wanted to compare and contrast the significance of the aforementioned variables in our second data set across each league (MLS excluded due to lack of valid data). We created a holdout sample of 20% in each league to test our model and identified that the only variable that was non-significant in each leagues model was away team fouls. Coefficients that were positive in each model and negative in each model are listed in Table 2, whereas each individual regression we ran for the leagues can be found in Tables 3 through 7.

Table 2
Positive and Negative Coefficients in all Soccer League Logistic Regression Models

Positive Coefficients	Negative Coefficients
Home Shots on Target	Home Shots
Home Fouls	Away Shots on Target
Away Corners	Home Corners
Away Yellow Cards	Home Yellow Cards
Away Red Cards	Home Red Cards
Home Fouls	

Note: Not all variables were significant in every logistic regression model.

Table 3
Variables in Bundesliga Logistic Regression Model

Variable	B	S.E.	Sig	Exp (B)	Lower 95% CI	Upper 95% CI	VIF
Home Shots	-.053	.015	.000	.948	.920	.977	2.391
Away Shots	.045	.017	.006	1.046	1.013	1.081	2.246
Home Shots on Target	.463	.031	.000	1.588	1.495	1.687	1.881
Away Shots on Target	-.382	.032	.000	.682	.641	.726	1.848
Home Corners	-.106	.021	.000	.900	.863	.937	1.375
Away Corners	.079	.024	.001	1.082	1.033	1.134	1.327
Home Yellows	-.169	.045	.000	.845	.774	.922	1.060
Away Yellows	.096	.042	.023	1.100	1.013	1.195	1.047
Home Reds	-.970	.217	.000	.379	.248	.580	1.042
Away Reds	.537	.166	.001	1.710	1.236	2.366	1.028
Constant	-.475	.271	.080	.622			

Note: n = 1713 with holdout sample of 429.

Table 4

Variables in Ligue 1 Logistic Regression Model

Variable	B	S.E.	Sig	Exp	Lower 95%	Upper 95%	VIF
				(B)	CI	CI	
Home Shots	-.096	.014	.000	.908	.884	.933	1.996
Away Shots	.056	.016	.000	1.058	1.025	1.091	2.071
Home Shots on Target	.509	.029	.000	1.664	1.571	1.761	1.710
Away Shots on Target	-.416	.031	.000	.660	.621	.701	1.721
Home Corners	-.120	.018	.000	.887	.855	.919	1.240
Away Corners	.095	.021	.000	1.100	1.056	1.145	1.244
Home Yellows	-.169	.043	.000	.845	.777	.919	1.157
Home Reds	-.869	.158	.000	.419	.308	.572	1.060
Away Reds	.618	.126	.000	1.855	1.450	2.373	1.073
Constant	.102	.231	.659	1.108			

Note: n = 2128 with holdout sample of 532.

Table 5

Variables in La Liga Logistic Regression Model

Variable	B	S.E.	Sig	Exp	Lower 95%	Upper 95%	VIF
				(B)	CI	CI	
Home Shots	-.086	.014	.000	.918	.893	.943	2.381
Away Shots	.040	.015	.008	1.041	1.011	1.072	2.321
Home Shots on Target	.500	.028	.000	1.649	1.562	1.742	1.980
Away Shots on Target	-.381	.030	.000	.683	.645	.725	1.951
Home Corners	-.106	.018	.000	.899	.868	.931	1.067
Away Corners	.060	.020	.003	1.061	1.020	1.105	1.030
Home Yellows	-.074	.032	.020	.928	.872	.988	1.333
Home Reds	-.337	.127	.008	.714	.556	.916	1.308
Away Reds	.321	.108	.003	1.378	1.116	1.702	1.053
Constant	.206	.236	.381	1.229			

Note: n = 2128 with holdout sample of 532.

Table 6

Variables in Serie A Logistic Regression Model

Variable	B	S.E.	Sig	Exp	Lower 95%	Upper 95%	VIF
				(B)	CI	CI	
Home Shots	-.049	.019	.008	.952	.917	.987	2.054
Away Shots	.050	.021	.018	1.051	1.009	1.096	1.941
Home Shots on Target	.482	.042	.000	1.620	1.492	1.759	1.666
Away Shots on Target	-.362	.044	.000	.696	.638	.760	1.609
Home Corners	-.146	.027	.000	.865	.820	.911	1.318
Away Corners	.080	.030	.007	1.083	1.022	1.148	1.278
Home Yellows	-.153	.055	.005	.858	.770	.956	1.013
Home Reds	-.626	.215	.004	.535	.351	.815	1.030
Away Reds	.791	.184	.000	2.206	1.538	.3163	1.039

Table 6 Continued

Variable	B	S.E.	Sig	Exp (B)	Lower 95% CI	Upper 95% CI	VIF
Constant	-.263	.356	.459	.768			

Note: n = 2128 with holdout sample of 532.

Table 7
Variables in Premier League Logistic Regression Model

Variable	B	S.E.	Sig	Exp (B)	Lower 95% CI	Upper 95% CI	VIF
Home Shots	-.036	.019	.003	.964	.941	.988	2.324
Home Shots on Target	.253	.018	.000	1.287	1.242	1.334	1.951
Away Shots on Target	-.227	.018	.000	.797	.769	.825	1.275
Home Fouls	.027	.014	.050	1.027	1.000	1.055	1.197
Home Corners	-.077	.016	.000	.926	.898	.956	1.351
Away Corners	.080	.017	.000	1.083	1.047	1.121	1.228
Home Yellows	-.172	.040	.000	.842	.778	.911	1.189
Home Reds	-1.131	.209	.000	.323	.214	.487	1.030
Away Reds	.644	.144	.000	1.905	1.437	2.525	1.030
Constant	-.119	.248	.630	.888			

Note: n = 2128 with holdout sample of 532.

Through our logistic regressions in each league, most coefficients made logical sense; home shots include shots not on target, so a conclusion is that the total amount of shots will result in a negative coefficient. One outlier we identified that only showed up in one model, was a positive coefficient of home fouls in the premier league. This didn’t make logical sense to us as the more fouls a home team would incur, the chances of them winning would go up. Another variable that was tricky to diagnose was the corner kicks; away corner kicks was a positive coefficient whereas home corner kicks were a negative coefficient. Our reasoning for this is typically when a club takes a corner kick, they missed a chance at goal by a deflection of the opposing team, therefore resulting in a missed opportunity to score a goal. One note to mention is that it barely made the significance level of .05. After testing each of our models in each league, we found the models to fit well. For example, the results of the model applied to a Bundesliga fixture between Bayern Munich and Hannover are shown below:

-.475 (Constant) + (-.053*20 home shots) + (.045*20 away shots) + (.463*11 home shots on target) + (-.382*5 away shots on target) + (-.106*11 home corners) + (.079*5 away corners) + (-.169*0 home yellows) + (.096*1 away yellows) + (-.970*0 home reds) + (.537*0 away reds) = 1.873.

The results of the model include a logit of 1.873, odds of 6.51, and a probability of the home team (Bayern Munich) winning of 87.7%. Bayern Munich did indeed win this fixture with a score of 3-1.

The next analytical tool we utilized was classification and ROC curves. When we scored the data for all leagues (excluding MLS due to lack of valid data), we converted the logit output to probabilities. We then ran the outcome and probability at a cutoff of .50. We decided to use a cutoff rate of .50 since we wanted to see how accurate and predictable the home win results are from a 50% probability level. Generally speaking, when the probability is above 50%, we would tend to predict the team with that percentage being the winner. If our accuracy rates for our leagues from the classification is more than 50%, that would suggest that our model had better accuracy than by guessing. By utilizing our training data, we were able to visualize our results in order to create a classification matrix and ROC curve.

The results showed us that our results were surprisingly accurate even at a .50 cutoff level. First and foremost, looking at our total amount of home wins, we can see that they far outweigh both draws and home losses as both of those variables were baked into the binary of 0. Across all five leagues, we can see that the accuracy rate hovered around 68% - 73% across the board. The league with the lowest accuracy rate with true positives and true negatives was the premier league at 68.16%, while the league with the highest accuracy was La Liga at 73.68%. With a 70% accuracy rate at a .50 cutoff, we can say that the model is able to predict home wins accurately, roughly at a rate of 3 out of every 4 games. Figures 2 through 6 show the ROC curve for each league along with the classification confusion matrix.

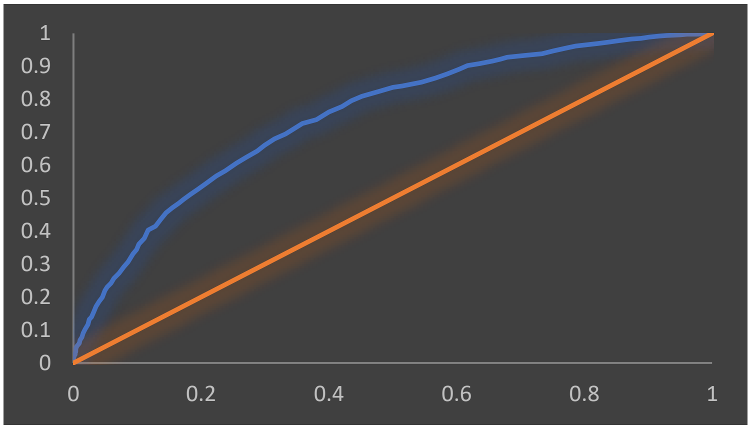

Figure 2. Premier League ROC curve.

Table 8
Classification Confusion Matrix for Premier League

	Predicted	Class	Total
Actual	738	479	1217
Class	368	1075	1443

Note: 686 draws. N = 2660.

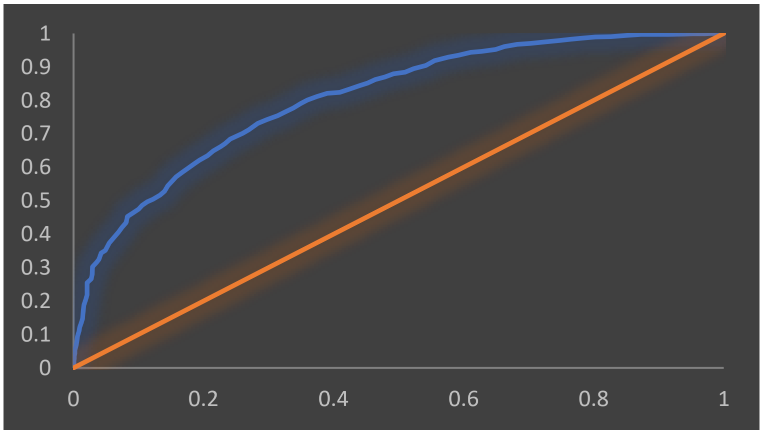

Figure 3. Serie A ROC curve.

Table 9
Classification Confusion Matrix for Serie A

	Predicted	Class	Total
Actual	810	413	1223
Class	326	1111	1437

Note: 711 draws. N = 2660.

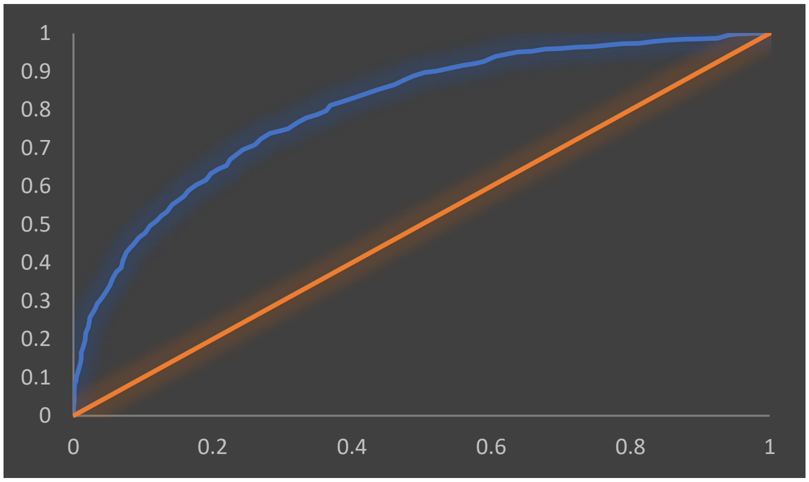

Figure 4. Bundesliga ROC curve.

Table 10
Classification Confusion Matrix for Bundesliga

	Predicted	Class	Total
Actual	628	332	960
Class	260	922	1182

Note: 863 draws. N = 2142.

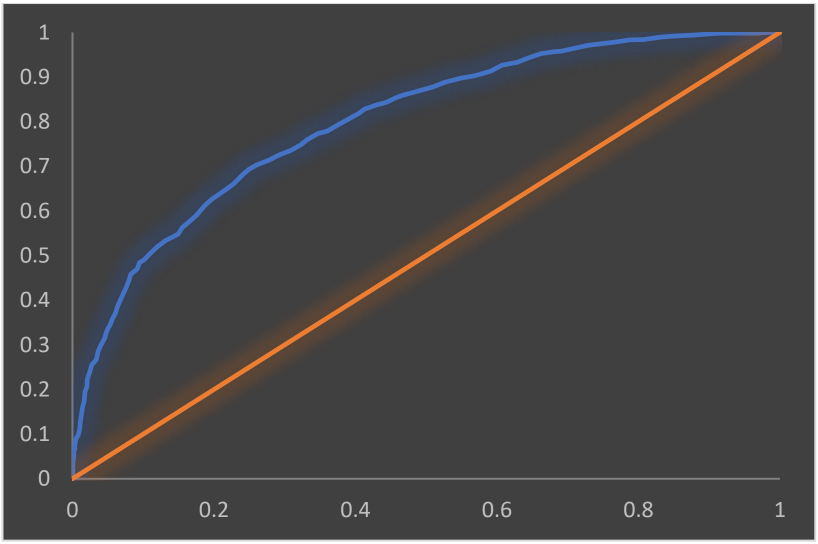

Figure 5. Ligue 1 ROC curve.

Table 11
Classification Confusion Matrix for Ligue 1

	Predicted	Class	Total
Actual	788	406	1194
Class	333	1133	1466

Note: 747 draws. N = 2600.

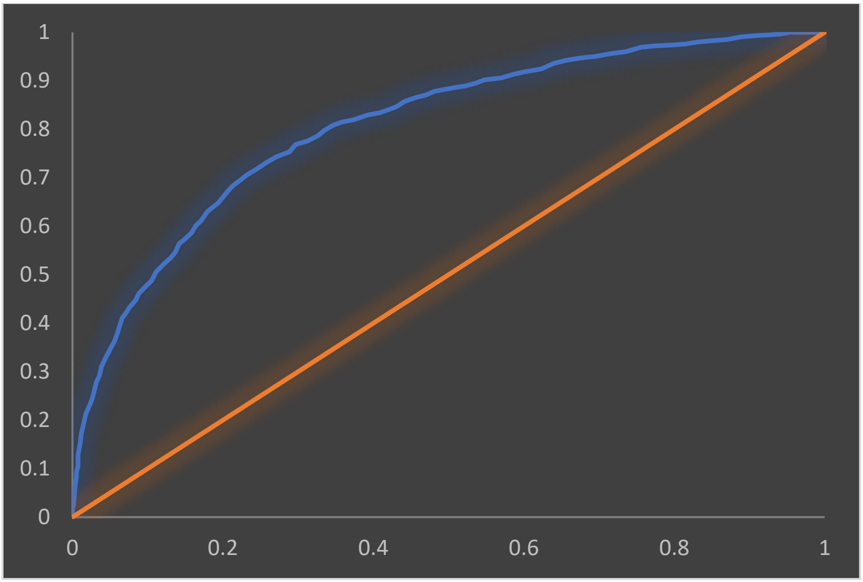

Figure 6. La Liga ROC curve.

Table 12
Classification Confusion Matrix for La Liga

	Predicted	Class	Total
Actual	935	366	1301
Class	334	1025	1359

Note: 621 draws. N = 2600.

The next analytical tool we utilized was a 3-year moving average to forecast home wins for each team. We felt that this was the most efficient way to forecast because using a historical moving 3-year average would be able to account for longer term changes in order to more accurately predict future data as opposed to using a naïve method or prior year data. After plotting the residuals between actual wins and wins forecasted from the years of 2013 through 2016, we identified a fair split between the six leagues we analyzed; The forecasting for Premier League, MLS, and Serie A tended to have higher residuals than Bundesliga, La Liga, and Ligue 1. We figured this was the case as MLS, Serie A, and Premier League tend to have more parity than the other three leagues. This can be explained by the fact that if a league has more parity within it, the level of competition is better within the league. If skill level of all the teams are closer, there is more random spread in terms of wins, and the residual plots would fall further away from the trendline because of the unpredictability of the league. This can definitely be seen in the graph for MLS, as this league has the most competitiveness due to salary cap and a lack of promotion and relegation. On the other hand, teams with less parity will have a better level of predictability within the leagues. In Ligue 1, for example, the teams are very consistent and tend to finish in similar positions year after year across the league. Because the separation between top teams and bottom teams are so prevalent, there is less randomness in the spread. Therefore, as we see from our residuals in Figures 7-12, the plots hug closer to the trendline and we are able to forecast the results better.

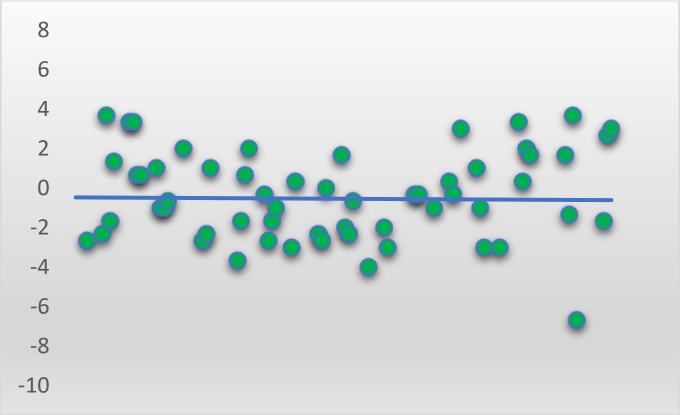

Figure 7. Residual Plot of 3 year moving average forecasting wins for La Liga.

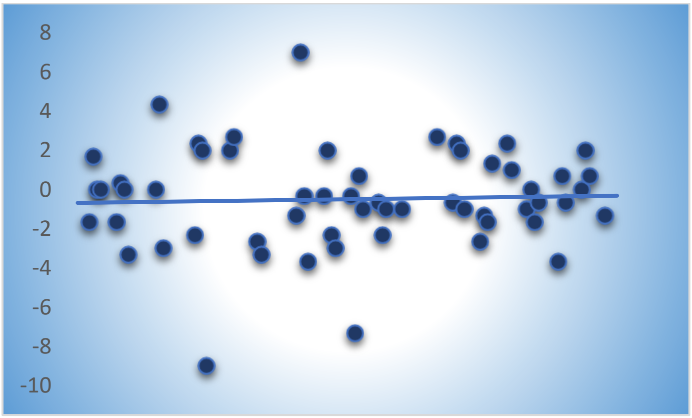

Figure 8. Residual Plot of 3 year moving average forecasting wins for Premier League.

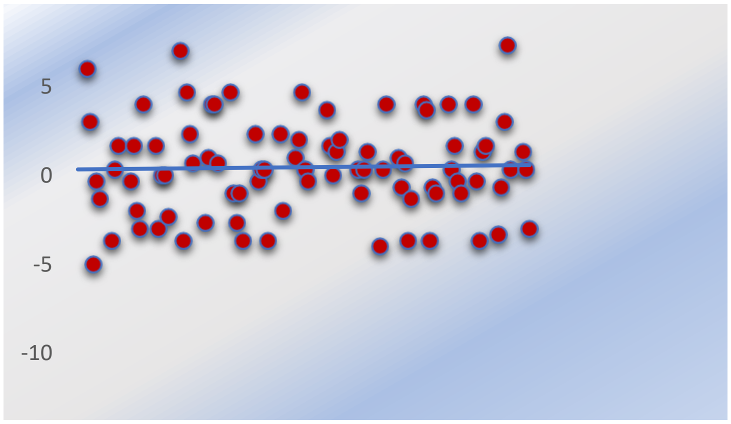

Figure 9. Residual Plot of 3 year moving average forecasting wins for MLS.

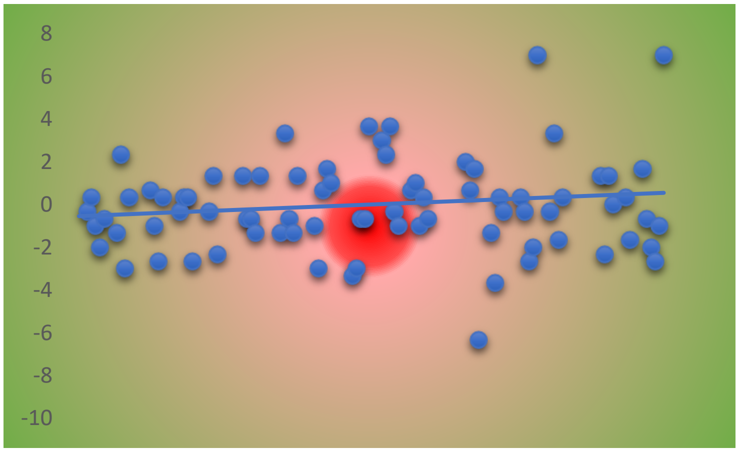

Figure 10. Residual Plot of 3 year moving average forecasting wins for Serie A.

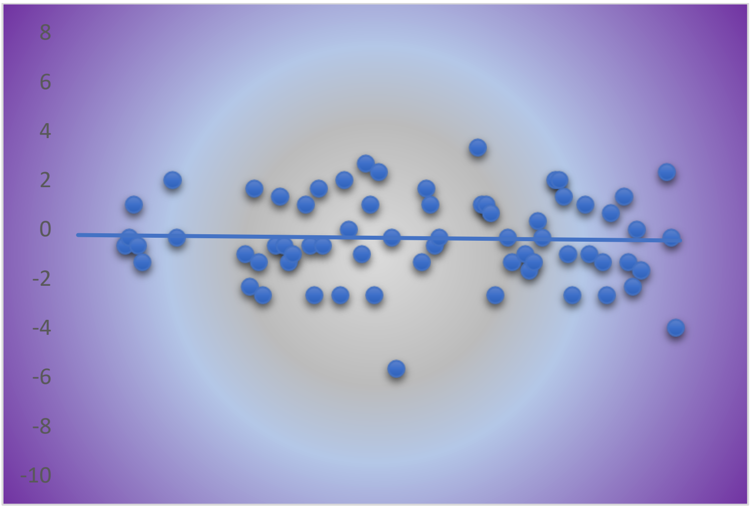

Figure 11. Residual Plot of 3 year moving average forecasting wins for Ligue 1.

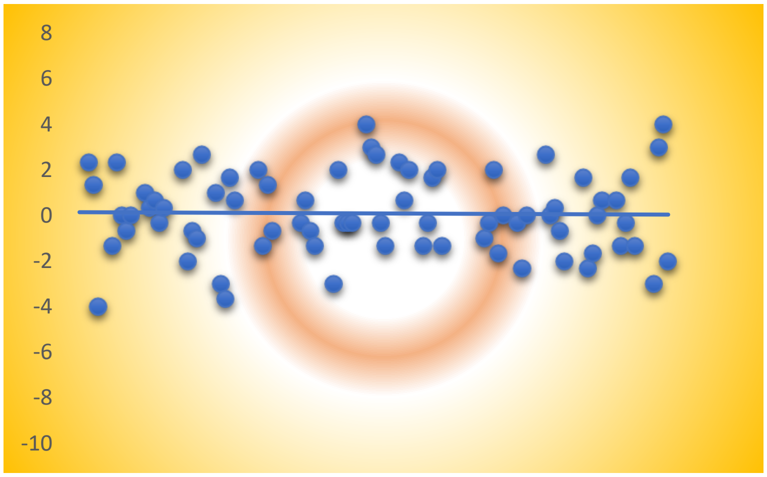

Figure 12. Residual Plot of 3 year moving average forecasting wins for Bundesliga.

After conducting our analysis of our data and using all of the analytical tools available to us, we wanted to reflect and see how our results were affected by the state of world soccer at the time our data was collected. In the 2010-2016 seasons, a lot happened in the soccer world from two World Cups, to new teams dominating their respective leagues, and other surprising results in between. We feel that these events had a huge effect on our results and could potentially affect the results of future analysis were this research conducted again. Some of the more interesting results was the parity of the Italian Serie A. When the data was collected, Serie A had multiple champions due to the fact that their most dominant team, Juventus, had been relegated to the Serie B division. Once they returned they ended up reclaiming their position as the best team in Italy and now boast the same dominance as Bayern Munich does in Germany's Bundesliga. We expect this would have made those results in our clustering and forecasting much more similar to each other instead of the Bundesliga being the sole outlier league. The rise in dominance of Paris-Saint Germain (PSG) in Ligue 1 also created different results for the French league, because before PSG's dominance, the French league crowned five different champions over the past seven seasons, putting it on par with the parity of MLS or the Premier League. We also would have done some things differently were we given the opportunity to do this again. MLS game by game data was not available in the packaged datasets we had found for other leagues and given the rapid growth of MLS, it was difficult to compare them to the other leagues in our regression and forecast. We would have liked to see the comparison side by side to see how truly different MLS is compared to their European counterparts, but from 2010-2016 MLS grew from 16 teams to 20, adding five teams and dissolving one. This would have been difficult to collect data for, particularly for our 3-year moving average forecast, as a number of teams would not have enough data to be counted in the forecast. Regardless, we are proud of our outputs and believe that our regression can be used to accurately score new games and help predict win odds and probability for the home team, while also predicting league results using our 3-year moving average forecast model.

Overall, after performing multiple analyses on different data sets for Bundesliga, La Liga, Ligue 1, MLS, Premier League, and Serie A, we feel like we were able to answer some of the questions we had going into this project. We wanted to know what contributed to home field advantages and come up with conclusions based on data we utilized instead of just making an educated guess. Through clustering, we identified that Bundesliga truly has a home-field advantage of its' own due to all of its' seasons being in the same cluster (in accordance to the variables we used). Logistic regression showed us some unique identifiers in what can contribute to the home team winning, for example, more away team corner kicks. Our ROC curve gave us more insight as to why teams have a higher chance of winning at home as opposed to drawing or losing. Forecasting showed us that parity can make league results unpredictable, as our residual plots showed us for the MLS, Premier League and Serie A. Ultimately, we learned that through using data, we could make better educated guesses as to why there truly is a home-field advantage in soccer.

Appendix

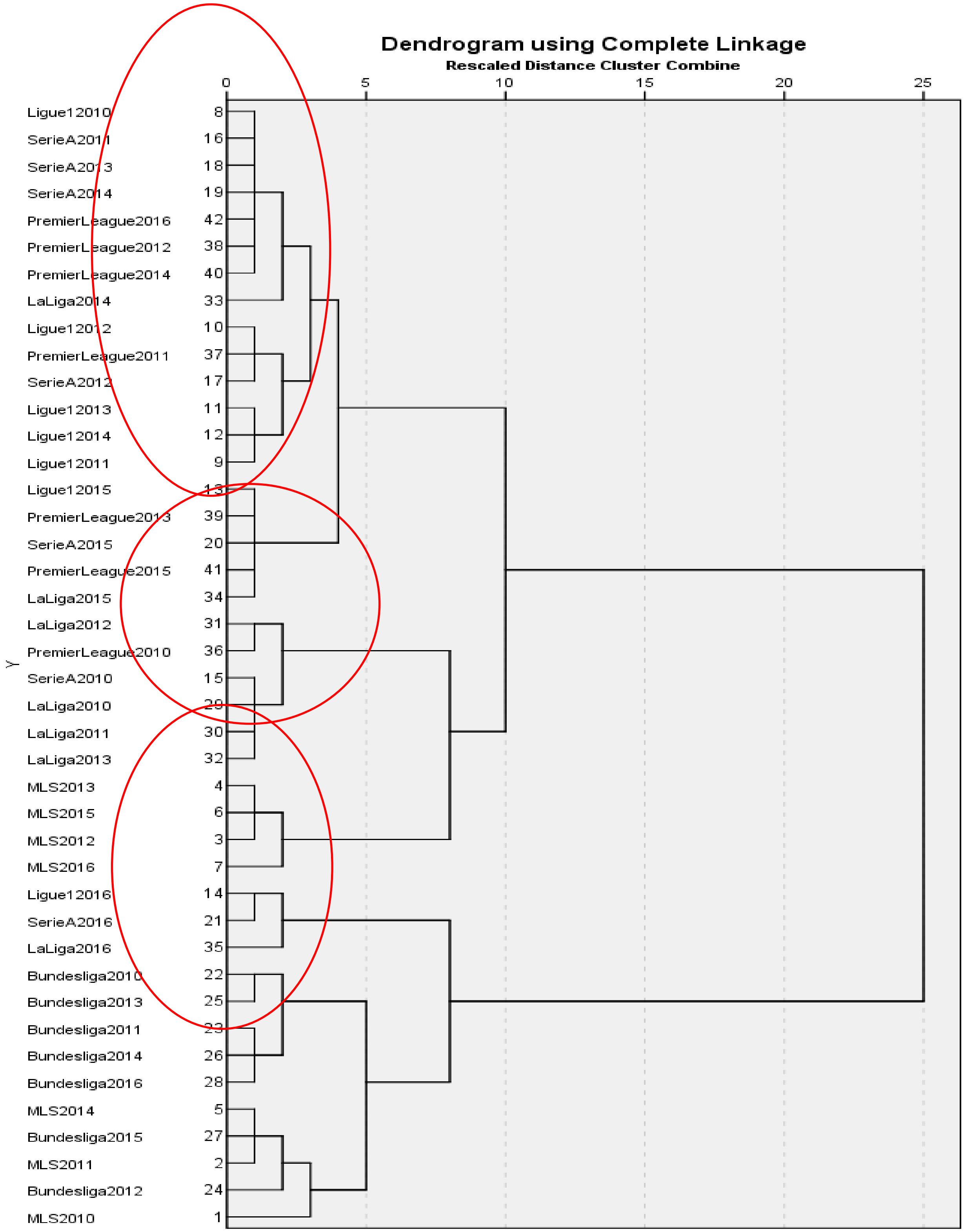
Dendrogram using Complete Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25
Y
Ligue12010 8
SerieA2011 16
SerieA2013 18
SerieA2014 19
PremierLeague2016 42
PremierLeague2012 38
PremierLeague2014 40
LaLiga2014 33
Ligue12012 10
PremierLeague2011 37
SerieA2012 17
Ligue12013 11
Ligue12014 12
Ligue12011 9
Ligue12015 13
PremierLeague2013 39
SerieA2015 20
PremierLeague2015 41
LaLiga2015 34
LaLiga2012 31
PremierLeague2010 36
SerieA2010 15
LaLiga2010 29
LaLiga2011 30
LaLiga2013 32
MLS2013 4
MLS2015 6
MLS2012 3
MLS2016 7
Ligue12016 14
SerieA2016 21
LaLiga2016 35
Bundesliga2010 22
Bundesliga2013 25
Bundesliga2011 23
Bundesliga2014 26
Bundesliga2016 28
MLS2014 5
Bundesliga2015 27
MLS2011 2
Bundesliga2012 24
MLS2010 1

Figure 1. Dendrogram of 42 seasons from 2010-2016 for Bundesliga, La Liga, Ligue 1, Premier League, MLS, and Serie A.