Predicting Serie A Game Outcomes

Biguzzi-Velcich, Stefano[1]

[1] City-University-of-New-York-School-of-Professional-Studies

Abstract

Multiple studies have been conducted on predicting soccer matches, as well as, websites built to provide predictions; however, there is still room to further develop these processes by adding additional data points. The leading strategy to predict matches is to create a projected score for the home team and away team and utilize the projected score as the λ in a Poisson distribution simulation for each team in that game. In this paper I attempt to refine the process of predicting matches by defining a new method to arrive at projected score. Whereas the projected score for websites like FiveThirtyEight uses data that is not free, this method employs only free data that can be scraped online. I then compare the FiveThirtyEight prediction results with my prediction results and determine if any advantage exists to my method of creating λ. The success rates for predicting correct outcomes for matches during the 2021-2022 Serie A season, up to and including match day 35, using the λs from FiveThirtyEight and this paper were 53% and 56% respectively. Within the method for this paper manager experience plays a small role in predicting matches, with most of the weight coming from the Off and Def team ratings. With the continual legalization of sports betting across the country and the dominance of football, basketball, and baseball in the sportsbooks, the ability to develop a process to predict soccer matches using free data and providing the calculations required to arrive at each λ, could increase the market for soccer betting in America.

Keywords: data-science, sports, soccer, football, sports-betting, betting
Word count: 3,565

Correspondence concerning this article should be addressed to Biguzzi-Velcich, Stefano, . E-mail: stefano.biguzzivelcich14@spsmail.cuny.edu

1 Introduction[1]

Sports betting is currently legal in 31 out of 50 states, with 10 more states potentially joining the list in the next year (Rovell, 2021). Just this past year in January 2022 New York was the latest state to legalize sports betting. Doing so added an additional $1.6 billion to the sports betting market. This addition was the driving factor in increasing the total sports betting handle for January by 30% from $7.54 billion to $9.82 billion in the United States (Ordine, 2022). While football, basketball, and baseball dominate the betting world of sports, soccer can still generate a huge amount of profit if selecting the correct outcomes. Some of the possible betting options are the moneyline option, deciding whether the game outcome will be a home win, a tie, or an away win, or the over/under option on total goals scored. The purpose of this project is to attempt to improve on the analytics that are already available to be able to more accurately predict match outcomes and scores.

Predicting goals should be considered a Poisson distribution more than a Negative Binomial when developing a projection model for individual teams. Examining the distribution of goals by all the teams in the league would show a Poisson distribution with variable mean, allowing for a Negative Binomial model to be developed. The problem with this approach is that to predict the outcome of a game each team's goals scoring chances have to be considered independently on the side of the field in which they happen. Assuming the probability to score a goal, p, is constant, and attacks by each team are independent events then a Poisson distribution will fit the data much better (Maher, 1982).

Building on this idea that goals follow a Poisson distribution, (Croucher, 2004), developed a method to predict match outcomes. By taking the home scoring average as the λ_1 and the away scoring average as the λ_2, (Croucher, 2004) created probabilities for the home team and away teams to score 0 to 4 or more goals. Converting these probabilities into matrix form, (Croucher, 2004) was able to assign individual probabilities to each event happening, i.e. the chance a 0-0 game occurs or a 1-0 games occurs. By adding the probabilities at the bottom half of the matrix, (Croucher, 2004) established the home teams winning chances. Adding the probabilities in top half of the matrix, he arrived at the away team winning chances. Finally, by adding the diagonal, he was able to achieve the probability of a tie.

Putting the concept introduced by (Maher, 1982) that goals follow a Poisson distribution and the equations put forth by (Croucher, 2004), (Boice, 2018) developed a more complete prediction model, introducing a new concept to soccer prediction, the Soccer Power Index (SPI) rating. SPI is a measure for the best estimate for a team's overall strength and is used in calculating each teams λ for each future game. (Boice, 2018) refers to this as the projected score 1 for the home team and projected score 2 for the away team. FiveThirtyEight's model simulates each game using the method introduced in (Croucher, 2004) over the course of a season to arrive at season long predictions.

This paper will explore the effects of considering the team manager as an important

[1]The following R libraries were used for this paper: (R Core Team, 2021), (Wickham, François, Henry, & Müller, 2022), (Wickham & Girlich, 2022), (Xie, 2015), (Zhu, 2021), (Aust & Barth, 2020), (Müller & Wickham, 2021), (Kuhn, 2022)

part of the goal scoring process. Each manager has a theory on how a soccer game should be played. A manager could want their team to be more attacking or more defensive, which would affect how many goals are scored and the chances a team has to win a game. Then by calculating a new λ considering the manager in charge for any given game, this paper will add to the predictive methods developed by (Boice, 2018). I will then compare my λ Poisson forecasting method with that of FiveThirtyEight. Each method will create percent chances for games and we can compare which modeling method is the most accurate. The model that forecasts games with a greater level of accuracy can be used to develop a soccer gambling calculator and assist people in selecting game outcomes.

2 Methods

This section contains the process for collecting data, how the data were processed and cleaned, how the data were analyzed, and finally how the new λ was developed. The data includes Italian League Serie A and Serie B games dating from 2015 to present excluding Serie B playoff games. Playoff games are more difficult to predict since they have a much different atmosphere to them. Whereas losing a season game might not automatically disqualify a team from winning the league, losing in a playoff game will eliminate a team from being promoted, giving players extra motivation in playoff games.

In addition to league game data, I also collected all games managed by each team's manager going back from their first game to their most recent relevant game. That is, if a manager was fired halfway through season 2017 and never came back as a manger in Serie A, their historic games would end on their last Serie A game managed for season 2017. Manager data also excludes playoff games or competition games, such as UEFA Champions League, for the same reasoning as excluding Serie B playoff games.

The data was then pre-processed and analyzed using a Poisson regression. The λs were developed by predicting the game scores using the best Poisson regression. These predicted goals, the final λ_1 and λ_2, were then used to simulate each teams chance of scoring x amount of goals within a game. By then applying (Croucher, 2004) method, each game was assigned an outcome percentage[2], which was used to determine the game's final outcome.

2.1 Collection

Data for each Serie A and Serie B games starting from 2016 was collected from the Github repository, (Boice, n.d.). This data is automatically loaded from the Github repository using the *spi_matches.csv* file in the data pre-processing stage. FiveThirtyEight's data only goes back to 2016 for Serie A and 2017 for Serie B, therefore the 2015 Serie A match scores and 2016 Serie B match scores were manually collected using (*Football Transfers, Rumours, Market Values and News*, n.d.) into a file *HistoricData.csv*. The FiveThirtyEight and the older Serie A and B data were combined to create the base scoring data table.

Manager data was collected from (*Football Transfers, Rumours, Market Values and News*, n.d.) and manually copied and pasted into a csv file to build out a database for historical managerial data. Wikipedia was used to determine the manager for each team at

[2]home win, away win, tie

any given time including managerial changes throughout a given season. Each row in the csv file is a game associated to a manger, potentially creating multiple rows per game if the managers for each teams are relevant to the study. Once data was manually updated the data needed to be joined together and each metric to use in creating a model needed to be created.[3]

2.2 Pre-processing[4]

The first data set to be cleaned was the SPI dataset from FiveThirtyEight. A few of the teams needed to be renamed for ease of reading as well as allowing me to more easily match the league team data to the manager data. There are 10 Serie B playoff games each season to determine the three promoted teams for the following Serie A season. I removed these playoff games by assigning a counter to each row, grouped by Serie B and season, and then filtering out any row with a counter greater than the max counter minus ten. Utilizing the method outlined in (Boice, 2018), taking the average of the following three metrics; *adjusted goals*, *Shot-based expected goals*, and *Non-shot expected goals*, I created the composite attacking score and the composite defensive score values for each game.

The two goals for the manager data were to, first, transform the data from a row representing a unique manager game to a row representing one unique game, and, finally, to create the necessary fields for each manager; *off*, *def*, and *exp* columns. The data was collected with one row representing a unique manager and game for ease of collection and it looks something like this:

Table 1
Initial manager data

date_char	team1	team2	g1	g2	manager	manager_team
4/16/2022	Juventus	Bologna	1	1	Sinisa Mihajlovic	Bologna
4/16/2022	Juventus	Bologna	1	1	Massimiliano Allegri	Juventus

The first step was to select the distinct combination of the first five columns of the manager data set; *date*, *team1*, *team2*, *g1*, *g2*. Next I created a team1 id and team2 id by concatenating the date and team names, which were used to merge on the manager name for those particular teams on the given date. This allowed the data to go from Table 1 to

Table 2
Processed manager data

date	team1	team2	g1	g2	manager1	manager2
2022-04-16	Juventus	Bologna	1	1	Massimiliano Allegri	Sinisa Mihajlovic

To create the *off* and *def* columns I took the cumulative mean of all the goals for and all the goals against a specific manager. The *exp* is the share of games that one manager

[3]All the data files I created can be found on my Github
[4]All data pre-processing was done using R-dplyr (Wickham et al., 2022)

managed when facing another manager.[5]. The final manager pre-processing step was to combine the manager data with the game data and create the necessary columns for the analysis.[6].

When a season starts there is not enough data to create the expected offensive and defensive ratings, and the team's actual offensive and defensive ratings. Therefore, we substitute the values for the first game of a new season with the last values from the previous season. Additionally, every season three teams from the second league division, Serie B, are promoted to Serie A. Promoted teams will generally not perform as well in Serie A as they did in Serie B and therefore adjustments need to be accounted for the get more accurate ratings for each team. Taking the each promoted team's Serie B average for season $i-1$ and averaging that with a season adjustment of season $i-1$ performance for promoted team in season $i-1$ gives a better picture of how promoted Serie B teams in season i will perform.

2.3 Analysis - Poisson and Simulation

I will be looking at two methods to create the λs. The first method uses the predicted value from a regression using the significant predictors without transformation. The second method experiments with interacting the terms to create relative predictors, for example Off/Def would be the relative goal variable, and then using that models coefficients to create the λs.

As seen in Figure 1 goals for soccer games look like they follow a normal log distribution.

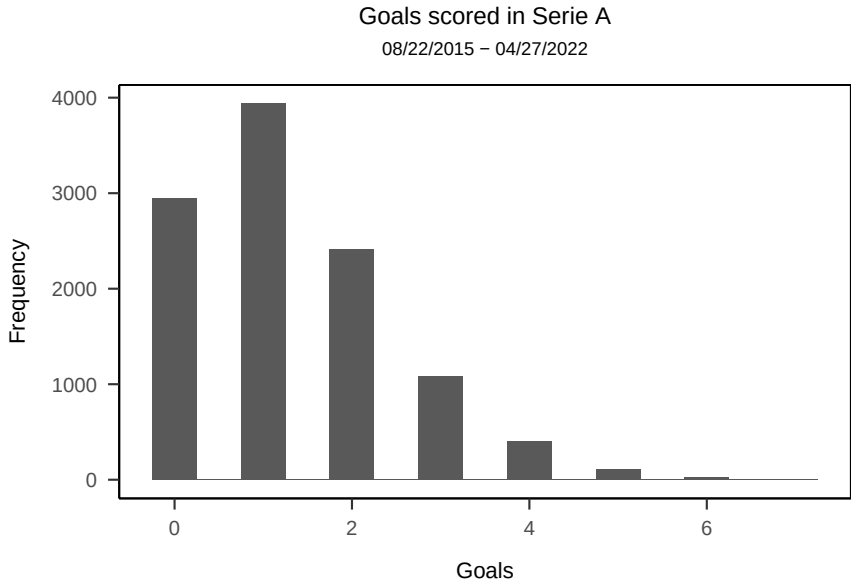

Figure 1

Yet, goals, more accurately, meet the Poisson distribution criteria. A Poisson distri-

[5]If manager1 managed 5 games and manager2 managed 10 games the m1_exp=$\frac{5}{5+10}$ = 0.333 and m2_exp=$\frac{10}{5+10}$ = 0.667

[6]You can get a list of the calculated columns in Appendix A - Calculated Column Logic

bution must follow these three rules: (1) An event can occur any number of times during a time period, (2) Events occur independently, and (3) the rate of occurrence is constant. Goals follow these rule.

First, it is theoretically possible, even though next to impossible, for a team to score an infinite number of goals in a game. Secondly, goals are independent events at the two ends of the field. In each game the attack happens on two separate sides of the field with team1 attacking team2's goal and vice versa. Goals cannot be scored at the exact same time. Furthermore, In each game the attack happens on two separate sides of the field with team1 attacking team2's goal and vice versa. Those events are each a game within itself, rendering them independent of each other, which strengthens the idea that goals follow a Poisson distribution. Finally, the rate of occurrence for Serie A goals is 1.32 goals per team per game.

Additionally, (Croucher, 2004), already showed that a Poisson distribution fits goals much better than other models. Therefore, it does not make sense to try a Negative Binomial or to convert the goals using a log transformation and using a linear regression model. Instead using a Poisson regression model would be the best option to understand the interactions between predictors and goals.

I created two models to determine which variables are significant to predicting goals, which will determine how I create the two λs to predict future games. The first model is a backwards step wise Poisson regression which produces the results in Table 3 as seen below.

Table 3
Model using non relative goals variable

term	estimate	conf.int	statistic	p.value
Intercept	-1.43	[-1.59, -1.27]	-17.79	$< .001$
Off	0.49	[0.44, 0.55]	18.05	$< .001$
Def	0.56	[0.50, 0.62]	18.58	$< .001$
M exp	0.12	[-0.02, 0.25]	1.72	.085
M off	0.09	[0.00, 0.18]	1.90	.058

The second model was also created stepwise where the predictors were transformed to be considered together. For example taking the *Off* variable and dividing it by the *Def* variable.[7] The results of the model are seen below in Table 4.

Table 4
Model using relative goals variables

term	estimate	conf.int	statistic	p.value
Intercept	-0.85	[-0.99, -0.71]	-11.87	$< .001$
Rel spi	1.06	[0.97, 1.15]	23.09	$< .001$
Rel m	0.05	[-0.04, 0.14]	1.09	.274

[7]All relative columns will be in Appendix A - Calculated Column Logic

The first model (Table 3) seems to show that the Off rating of one team versus the Def rating of another team is almost all that's needed to predict goals. All the predictors are significant at the p<0.1 level. Interestingly, *SPI* plays no part in the prediction of goals. Contrary to model one, when looking at the coefficients of each predictor in the second model (Table 4), the *SPI* predictor seems to become the most important predictor for goals. The coefficient for the calculated field *Rel SPI* is almost 21 time larger than all the other coefficients. Further evidence of *SPI's* importance in model two as seen in Table 5 below.

Table 5
Variable importance for each poisson model

predictors	weight
Poisson: No interaction columns	
def	0.4616
off	0.4485
m_off	0.0471
m_exp	0.0428
Poisson: Interaction columns	
rel_spi	0.9547
rel_m	0.0453

How well does each model do at predicting goal? After splitting the data into training (80%) and test (20%) it's possible to get a sense of how well each model predicts goals on a random set of data. By checking the model on the test data, it seems that each model predicts exact goal scored for each team with very similar accuracy as seen in Table 6.

Table 6
Success and Failure for each Poisson model

model	Fail	Success
Poisson: No interaction columns	0.6228	0.3772
Poisson: Interaction columns	0.6440	0.3560

The first model that does not use interaction columns predicts goals about 2% better, not enough to claim a better model alone.

Looking at both the AIC and MAE metric below in Table 7,

Table 7
Comparing no interaction model and interaction model metrics

Model	MSE	MAE	RMSE	AIC
Poisson: No interaction columns	0.6598	0.6377	0.8123	10019.47
Poisson: Interaction columns	0.7026	0.6599	0.8382	10503.45

the model that does not use relative variables preforms better. Given that we are predicting goals, MAE is the most important metric to consider. The MAE are fairly close

for both models, but the model without the interactions preforms slightly better with an MAE of 0.64. Given that both models preformed relatively the same, it's appropriate to test out both models to create each team's λs.

2.3.1 Creating λ_1 and λ_2. Using both models and the `predict()` function will give us the projected score for team1 and team2, which will be considered the λs for each team. The formula for λ_1 using the first model takes uses the Off rating of team 1, the Def rating of team 2, the team 1's manager experience, and team 1's manager Off rating.

$$\lambda_1 = e^{-1.43+0.49\times\text{OffRating}_1+0.56\times\text{DefRating}_2+0.12\times\text{ManagerExperience}_1+0.09\times\text{ManagerOffRating}_1} \quad (1)$$

The formula for calculating λ_1 using the interaction term model uses SPI rating of team 1,SPI rating of team 2,Expected Off rating of team 1, and expected Def rating of team 2, and manager Exp for team 1.

$$\lambda_1 = e^{-0.85+1.06\times\frac{\text{SPI}_1}{\text{SPI}_2}+0.05\times\frac{\text{ManagerOff}_1}{\text{ManagerDef}_2}} \quad (2)$$

For both models the formula for λ_2 uses the same columns as λ_1 but flipping the data from team 1 to team 2.

The two sets of λs created are *proj0.1* and *proj0.2* and *proj1.1* and *proj1.2*. The projected scores starting with *proj0.* use the first model without interaction, while projected scores starting with *proj1.* were created using the second model with interactions. Implementing these formulas you get all the projected score columns as seen in the example below for AC Milan's game on 5/2/2022.

Table 8
Projected scores for the next Milan Serie A game

season	date	team1	team2	proj0.1	proj0.2	proj1.1	proj1.2
2021	2022-05-01	Milan	Fiorentina	1.42	1.08	1.44	1.21

2.3.2 Predicting games and scores. The example matrix in Table 9 shows the score percentages for the Milan vs Fiorentina game on May 1st after simulating the game over 10,000 times using the projected scores from the interaction model.

Table 9
Probability matrix for Milan v Fiorentina 5/1/22 using no interaction model

Milan	Fiorentina					
	0	1	2	3	4	5+
0	8.49	8.63	4.82	1.89	0.47	0.13
1	11.84	12.92	7.10	2.52	0.53	0.21
2	8.11	9.23	4.30	1.80	0.41	0.18
3	3.76	3.85	2.21	0.88	0.23	0.05
4	1.35	1.37	0.89	0.26	0.09	0.01
5+	0.56	0.52	0.30	0.08	0.00	0.01

Adding up the bottom half of the matrix gives Milan's chance of winning the game, 44.33%. While adding the top half of the matrix gives Fiorentina's chance of winning the

game, 28.98%. Finally, adding the diagonal of the matrix gives the chance the game ends in a tie, 26.69%. Even though Milan is projected to win, the best final goal score prediction is a 1-1 draw with a 12.92% chance of happening. To predict total goals or final score on a betting market, the best way to use this matrix is to combine the total outcome percentage with the highest goal scored for that outcome. In the case above that would mean a Milan 1-0 win, with an 11.84% chance of occurring.

Creating predicted outcomes for all Matchday 1 to Matchday 35 games and using the true game outcomes to check how well the each λ predicted winners, the results look slightly better for the λs created using the model without interacting the terms.

Table 10
λ Success Rate

	Percent
No interaction model lambda	
Fail	44%
Succeed	56%
Interaction model lambda	
Fail	47%
Succeed	53%

Given that the no interaction model had the lower MAE, the better ability to predict goals, and created λs with the highest correct win rate, it's makes sense to use the model that does not create interaction terms.

3 Results

In this section, the success rate using the FiveThirtyEight λs and the λs developed in this paper for each game of the 2021-2022 Serie A season was compared to determine the best method to creating λs. To compare my results to those of FiveThirtyEight, I simulated their data using the *proj_score1* and *proj_score2* variables and got the following results:

Table 11
Serie A Matchday 35 FiveThirtyEight Predictions

season	date	team1	team2	proj1	proj2	win1	win2	tie
2021	2022-04-30	Cagliari	Hellas Verona	1.28	1.57	31.90	44.76	25.15
2021	2022-04-30	Napoli	Sassuolo	2.02	1.34	53.72	24.83	21.17
2021	2022-04-30	Sampdoria	Genoa	1.27	0.91	44.69	26.97	27.67
2021	2022-04-30	Spezia	Lazio	1.01	1.89	21.22	58.27	21.69
2021	2022-05-01	Juventus	Venezia	2.13	0.65	71.83	9.67	17.80
2021	2022-05-01	Milan	Fiorentina	1.70	0.82	59.14	17.90	23.54
2021	2022-05-01	Empoli	Torino	1.22	1.40	33.28	41.21	26.39
2021	2022-05-01	Udinese	Inter Milan	0.91	2.17	13.69	66.19	19.52
2021	2022-05-01	Roma	Bologna	1.56	0.95	51.02	22.33	26.00
2021	2022-05-02	Atalanta	Salernitana	2.31	0.90	69.32	12.76	17.43

The *win1*, *win2*, and *tie* column percentages where confirmed with the percentages that were seen on their website for the same dates. Below I compare those results to my predictions using my λ method:

Table 12
Matchday 35 Prediction Differences

season	date	team1	team2	win1_diff	win2_diff	tie_diff
2021	2022-04-30	Cagliari	Hellas Verona	-10.63	11.65	-2.83
2021	2022-04-30	Napoli	Sassuolo	0.95	-2.94	2.27
2021	2022-04-30	Sampdoria	Genoa	-4.82	5.65	-0.16
2021	2022-04-30	Spezia	Lazio	-4.80	4.80	-1.18
2021	2022-05-01	Empoli	Torino	-7.44	5.93	0.63
2021	2022-05-01	Juventus	Venezia	-7.68	3.10	5.28
2021	2022-05-01	Milan	Fiorentina	-14.88	11.42	2.88
2021	2022-05-01	Roma	Bologna	1.45	0.02	-0.82
2021	2022-05-01	Udinese	Inter Milan	1.27	-1.56	0.89
2021	2022-05-02	Atalanta	Salernitana	1.88	-0.75	-0.64

At a quick glance, there are some games that stand out compared to my predictions. Most notably, the game between Milan and Fiorentina on May 1st. FiveThirtyEight predicts Milan to win 59.14% of the time, while my prediction only has them winning 44.33% of the time, a difference of -14.88%.

This could be due to multiple things, but considering that all the win1 percentages are on average -3% points lower, it's possible that FiveThirtyEight's home field advantage adjustment is higher than mine. Furthermore, not knowing the intricacies of how FiveThirtyEight comes up with their projected scores makes it impossible to compare the real reasons for these differences.

Table 13
Outcome comparison

date	team1	team2	True outcome	My outcome	538 outcome
2022-04-30	Cagliari	Hellas Verona	win2	win2	win2
2022-04-30	Napoli	Sassuolo	win1	win1	win1
2022-04-30	Sampdoria	Genoa	win1	win1	win1
2022-04-30	Spezia	Lazio	win2	win2	win2
2022-05-01	Juventus	Venezia	win1	win1	win1
2022-05-01	Milan	Fiorentina	win1	win1	win1
2022-05-01	Empoli	Torino	win2	win2	win2
2022-05-01	Udinese	Inter Milan	win2	win2	win2
2022-05-01	Roma	Bologna	tie	win1	win1
2022-05-02	Atalanta	Salernitana	tie	win1	win1

Considering the results in Table 13 my projected scores and FiveThirtyEight's projected scores predicted the same outcomes for Matchday 35. Both were incorrect when predicting the Roma vs. Bologna game and the Atalanta vs. Salernitana game. Let's take a closer look at those games.

Both my prediction and FiveThirtyEight's prediction were wrong about the Roma vs. Bologna game, meaning that there was no unluckiness to the results. Comparing the final score with the expected goals, both teams deserved to score 1 goal, meaning that a tie is a fair result. On the other hand, for the Atalanta vs. Salernitana game the predictions were much closer than they seem. My win percent of 71.20% and FiveThirtyEight's win percent of 69.32% were fairly accurate. Comparing the final score of 1-1 to the expected final score of 2.52-0.99 shows that Salernitana got lucky and Atalanta should have won the game.

4 Discussion

This paper set out to accomplish what FiveThirtyEight has done and improve on it by considering whether the manager's experience has a role in goal scoring and, therefore, predicting the outcome of a game. The method produced in this paper utilizes manager experience when predicting goals for matches. The importance that the model assigns to manager experience is low, around 4%, even though the p-value was significant at the 0.10 level. The manager statistic that was more important than the manager experience was the manager offense rating at 5%. Manager offensive rating was also significant at the 0.10 level.[8]

Not knowing the mathematical methods behind FiveThirtyEight projected scores creates a barrier in improving their model. The method provided in this paper creates similar outcome percentages for each game, as seen in Table 13. The average difference for home team win percentages of -3.35% and the away team win percentages of 4.23%, suggests that the FiveThirtyEight method to develop a λ includes some sort of adjustment for home field advantage, which this paper's method does not account for.

[8]Reference Table 3

Table 14
Game outcome success failure for the 2021 season

Outcome using lambda method from paper	
Failure	44%
Success	56%
Outcome using lambda from 538	
Failure	47%
Success	53%

Nevertheless, comparing the outcome results for the 2021-2022 season, up to and including Matchday 35, suggests that the model presented in this paper is comparable to that of FiveThirtyEights, with an outcome prediction success rate of 56% and 53% respectively.

In future research it would be interesting to see the effect of starting lineups on outcomes and goals scored. It would have been impossible to collect this data for this project as time and resources were limited. Additionally, it would be interesting to gain a better understanding of the FiveThirtyEight model to be able to compare these results more accurately. Nonetheless, the methods in this paper are a close comparison to FiveThirtyEight and are available to the public for knowledge sharing. With that ability it might be possible to create an even better model than FiveThirtyEight's that can be translated to a betting calculator or betting tool website potentially increasing interest in soccer betting across the United States.

5 References

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from https://github.com/crsh/papaja

Boice, J. (n.d.). Data/soccer-spi at master · fivethirtyeight/data. https://github.com/fivethirtyeight/data.

Boice, J. (2018). How Our Club Soccer Predictions Work. https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/.

Croucher, J. S. (2004). Using Statistics to Predict Scores in English Premier League Soccer. In S. Butenko, J. Gil-Lafuente, & P. M. Pardalos (Eds.), *Economics, Management and Optimization in Sports* (pp. 43–57). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24734-0_4

Football transfers, rumours, market values and news. (n.d.). https://www.transfermarkt.us/.

Kuhn, M. (2022). *Caret: Classification and regression training*. Retrieved from https://CRAN.R-project.org/package=caret

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118. https://doi.org/10.1111/j.1467-9574.1982.tb00782.x

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Ordine, B. (2022). January Falls Just Shy of $10B US Sports Betting Handle. https://usbettingreport.com/sports-betting/january-falls-just-shy-of-10b-u-s-sports-betting-handle-will-march-top-it/.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rovell, D. (2021). Sports Betting Legalization Tracker: Several States Consider Legislation. https://www.actionnetwork.com/news/legal-sports-betting-united-states-projections.

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Girlich, M. (2022). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Zhu, H. (2021). *kableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

5.1 Appendix A - Calculated Column Logic

Note$_1$: All these are the base calculations not specific to team1 or team2.

Note$_2$: f = for and a = against

- **xg__t**: Expected goals- $\frac{\text{ShotBasedXg}_f + \text{NonShotBasedXg}_f + \text{AdjustedScore}_f}{3}$
- **m__exp**: Manager experience- $\frac{\text{ManagerGames}_f}{\text{ManagerGames}_a}$
- **h/a__off**: Home/Away Off Rating- $\frac{\Sigma\text{goals}_f}{\text{games}_f}$
- **h/a__def**: Home/Away Def Rating- $\frac{\Sigma\text{goals}_a}{\text{games}_a}$
- **adj__off/def**: Season adjustments for newly promoted teams- $\frac{\Sigma_i \text{goals}_{f/a}}{\text{games}_i}; i = \text{season}$
- **off**: Off rating- $\frac{\sum_{g=2}^{g=38} \text{Off}_{i-1;38}\text{goals}_f}{\text{games}_f}; i = \text{season}$
- **def**: Def rating- $\frac{\sum_{g=2}^{g=38} \text{Def}_{i-1;38}\text{goals}_a}{\text{games}_f}$
- **xg__off**: Expected Off rating: $\frac{\sum_{g=2}^{g=38} \text{xg\_off}_{i-1;38}\text{xg\_t}_f}{\text{games}_f}; i = \text{season}$
- **xg__def**: Expected Def rating: $\frac{\sum_{g=2}^{g=38} \text{xg\_def}_{i-1;38}\text{xg\_t}_a}{\text{games}_f}; i = \text{season}$
- **rel__spi**: Relative SPI rating- $\frac{\text{SPI}_f}{\text{SPI}_a}$
- **rel__g**: Relative game rating- $\frac{\text{Off}_f}{\text{Def}_a}$
- **rel__xg**: Relative expected game rating: $\frac{\text{xg\_off}_f}{\text{xg\_def}_a}$
- **rel__m**: Relative manager game rating: $\frac{\text{m\_off}_f}{\text{m\_def}_a}$
- **rel__adj**: Relative home/away adjustments: $\frac{\text{h/a\_off}_f}{\text{h/a\_def}_a}$