Can Jürgen Klopp Finish His Liverpool Career with the 2023/2024 EPL Title?*

Analyzing Possession, Goals, Expected Goals, Progressive Passes, and Progressive Carries to Forecast Winning Points in Soccer

Jiwon Choi

April 19, 2024

This study develops a Bayesian regression model to predict English Premier League standings by analyzing how team performance metrics such as possession percentages, goals scored, expected goals, progressive passes, and progressive carries influence league points. The findings indicate that goals scored and expected goals are the most significant predictors of team success, providing a quantitative basis for strategic planning in football. The results demonstrate the potential of using advanced statistical methods to enhance the predictability of sports outcomes, offering valuable insights for team management and sports analysts. By quantifying the impact of specific performance metrics, this research contributes to a deeper understanding of football dynamics and emphasizes the importance of data-driven decision-making in professional sports.

1 Introduction

After Klopp announced that he would leave Liverpool (Liverpool FC 2024), all the soccer fans were spotlighting if his Liverpool would win the league. However, the English Premier League (EPL) is one of the most unpredictable and competitive football leagues globally. This unpredictability, while adding excitement for fans, presents challenges in predictive analytics. In sports analytics, accurately forecasting outcomes such as league standings or match results can significantly benefit team management, betting markets, and strategic fan engagement. However, a notable gap persists in the literature—quantitatively linking team performance metrics to actual league outcomes. This paper addresses this gap by focusing on estimating the impact of specific performance metrics on league points, thus improving the predictability of team standings.

*Code and data are available at: https://github.com/jwonc4602/23-24-EPL-Forecast.

Our study constructs a Bayesian regression model to estimate how key performance indicators—such as possession percentages, goals scored, expected goals, progressive passes, and progressive carries—correlate with league points accumulated over a season. The primary estimand in our analysis is the set of coefficients for these metrics in the regression model, which quantifies their respective contributions to the total points earned by teams in the EPL.

The results demonstrate that certain metrics, notably goals scored and expected goals, have a significant predictive relationship with league standings. This finding is instrumental for teams in prioritizing aspects of gameplay that are statistically proven to correlate with success. Moreover, these insights offer tactical and strategic directions for team management, focusing on measurable performance areas that directly contribute to league success.

The paper is methodically structured to enhance understanding and facilitate further analysis. Section 2 then discusses specifics of the data sources and the variables that were considered important for this study. Section 3 introduces the specifics of our Bayesian regression model. This is followed by the presentation of the results in Section 4, which details the statistical significance and predictive power of the identified performance metrics. We conclude with a discussion of our findings and their implications for team strategy and future research in Section 5. Through this structured approach, the study provides insights that significantly contribute to the field of sports analytics, offering strong methods for predicting football performance and informing strategic decisions.

2 Data

This study utilized two distinct datasets from FBref.com (FBref 2018), retrieved through web scraping with the rvest package (Wickham and Posit Software 2024). The datasets comprise seasonal performance statistics and squad-specific data across various seasons. The analysis focuses on data spanning from the 2017/2018 season to the 2022/2023 season to develop a predictive model for the 2023/2024 season's winner. The regular dataset includes general metrics such as matches played, wins, draws, and losses for each club, while the squad dataset provides detailed metrics including possession rates (Poss) and expected goal rates (xG) for each team. Initially, a decade's worth of data was considered for model development; however, the key variables required for this study were only available starting from the 2017/2018 season, thus limiting the analysis to the last six seasons. The data compilation and analysis were conducted using the R statistical programming language (R Core Team 2023), supplemented by packages such as tidyverse (Wickham et al. 2019), rstanarm (Goodrich et al. 2022), knitr (Xie 2023), here (Müller 2020), ggplot (Wickham 2016) among others, to support the analysis.

Table 1: Sample of Cleaned Regular Season Data

Squad	Rk	W	D	L	Pts	Pts.MP	season
Arsenal	2	26	6	6	84	2.21	22/23
Aston Villa	7	18	7	13	61	1.61	22/23
Bournemouth	15	11	6	21	39	1.03	22/23
Brentford	9	15	14	9	59	1.55	22/23
Brighton	6	18	8	12	62	1.63	22/23

Table 2: Winning Points over Previous Season for Liverpool

season	Wins	Draws	Losses	Pts	Pts.MP	Rk
17/18	21(55.3%)	12(31.6%)	5(13.2%)	75	1.97	4
18/19	30(78.9%)	7(18.4%)	1(2.6%)	97	2.55	2
19/20	32(84.2%)	3(7.9%)	3(7.9%)	99	2.61	1
20/21	20(52.6%)	9(23.7%)	9(23.7%)	69	1.82	3
21/22	28(73.7%)	8(21.1%)	2(5.3%)	92	2.42	2
22/23	19(50.0%)	10(26.3%)	9(23.7%)	67	1.76	5

2.1 Regular Season Data

The dataset for the regular football season provides detailed data across various seasons, including team rankings (Rk), total matches played (MP), wins (W), draws (D), losses (L), home game attendance (Attendance), and the top team scorer. For analysis, the variables selected were 'Squad', 'Rk', 'W', 'D', 'L', 'Pts', and 'Pts.MP'. The 'Pts' denotes the total points accumulated in a season, while 'Pts.MP' calculates the average points earned per match, with all variables specific to each squad. In football leagues, a win awards a team three points, a draw one point, and a loss yields no points. The team with the highest points at the season's end is crowned the champion. Additionally, a 'season' was included to specify the data range from the 2017/2018 season to the 2022/2023 season. (See Table 1)

Figure 1 shows the fluctuating rankings of the three title-contending teams over the past seasons. Manchester City has dominated the EPL, achieving the title in five seasons, except for the 2019/2020 season when Liverpool emerged victorious: their consistent performance and last season's treble highlight Manchester City as a formidable contender (Pollard 2023). Meanwhile, Arsenal, after not winning since the 2003/2004 season (Clifford 2024), has shown significant improvement since the 2020/2021 season, positioning themselves as competitive challengers for the title.

Table 2, Table 3, and Table 4 show each team's performance metrics, including wins, losses, draws, total points, points per match played and rankings. Especially, the 2018/2019 season

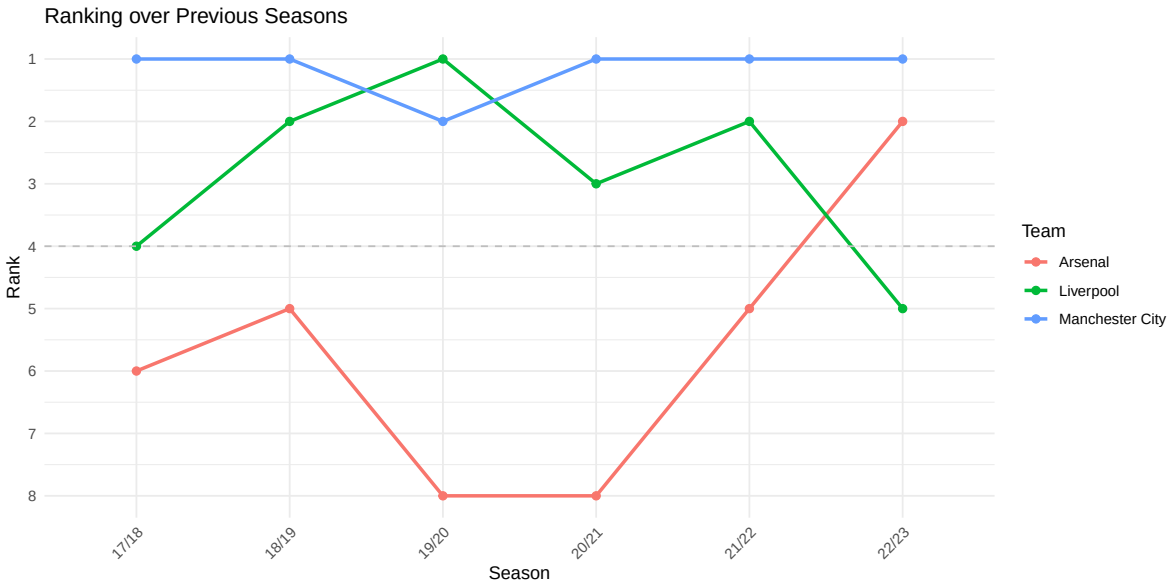

Figure 1: Ranking over Previous Seasons for Top Three Team

Table 3: Winning Points over Previous Season for Manchester City

season	Wins	Draws	Losses	Pts	Pts.MP	Rk
17/18	32(84.2%)	4(10.5%)	2(5.3%)	100	2.63	1
18/19	32(84.2%)	2(5.3%)	4(10.5%)	98	2.58	1
19/20	26(68.4%)	3(7.9%)	9(23.7%)	81	2.13	2
20/21	27(71.1%)	5(13.2%)	6(15.8%)	86	2.26	1
21/22	29(76.3%)	6(15.8%)	3(7.9%)	93	2.45	1
22/23	28(73.7%)	5(13.2%)	5(13.2%)	89	2.34	1

Table 4: Winning Points over Previous Season for Arsenal

season	Wins	Draws	Losses	Pts	Pts.MP	Rk
17/18	19(50.0%)	6(15.8%)	13(34.2%)	63	1.66	6
18/19	21(55.3%)	7(18.4%)	10(26.3%)	70	1.84	5
19/20	14(36.8%)	14(36.8%)	10(26.3%)	56	1.47	8
20/21	18(47.4%)	7(18.4%)	13(34.2%)	61	1.61	8
21/22	22(57.9%)	3(7.9%)	13(34.2%)	69	1.82	5
22/23	26(68.4%)	6(15.8%)	6(15.8%)	84	2.21	2

illustrates this, with Liverpool finishing second with 1 loss and 30 wins despite an impressive record, due to Manchester City's 4 losses and 32 wins. Manchester City clinched the league by just one point, emphasizing that targeting victories, to secure the big winning point of 3, is more strategic than merely avoiding defeats for optimizing point accumulation.

2.2 Squad Standard Stats Data

The Squad Standard Stats Data encompasses a range of metrics assessing team performance, including playing time, possession, and goals. For this analysis, the key metrics selected are 'Possession' (Poss), 'Goals' (Gls), 'Expected Goals' (xG), 'Progressive Carries' (PrgC), and 'Progressive Passes' (PrgP). These variables are important in evaluating the overall quality of a squad and predicting match outcomes.

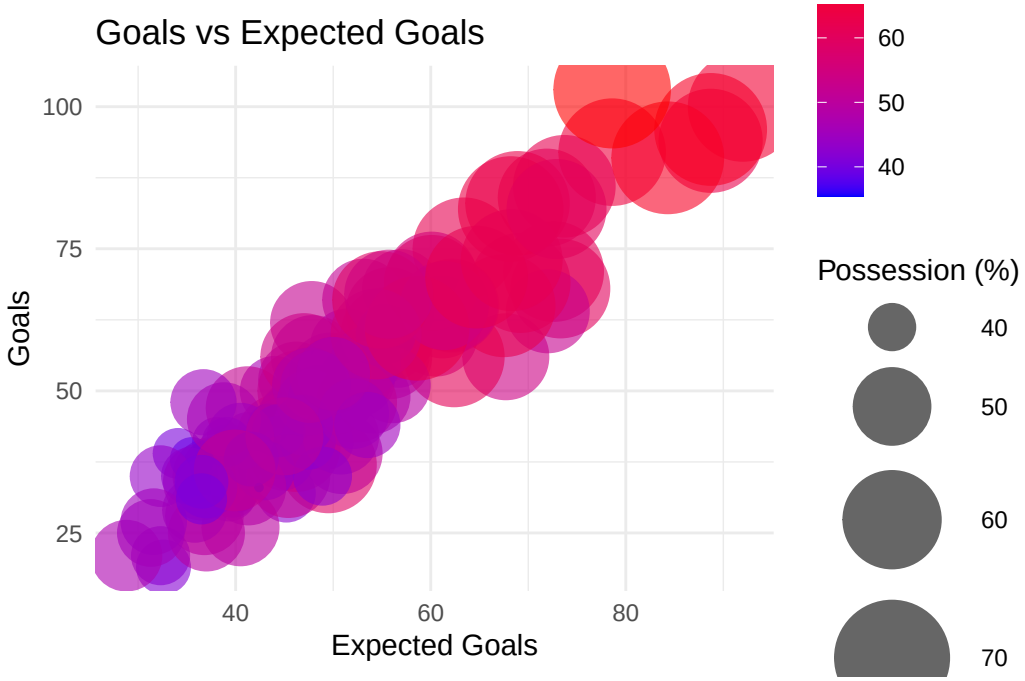

Figure 2: Goals vs Expected Goals with Possession over Previous Seasons

Figure 2 clearly shows that Possession (Poss) continues to be an important metric. Despite a shift in some modern tactics that de-emphasize possession, maintaining control over the ball generally increases a team's opportunities to score. The relationship between 'Goals' (Gls) and 'Expected Goals' (xG) is particularly telling. Expected Goals (xG) quantifies the quality of a scoring chance based on the likelihood of a shot being scored from a similar position and situation in past games, using historical shot data (Whitmore 2023). A discrepancy between xG and actual goals suggests a team's efficiency or lack thereof in capitalizing on scoring opportunities.

'Progressive Carries' (PrgC) measure movements that advance the ball towards the opponent's goal by at least 10 yards from its furthest point in the last six passes, or into the penalty area, but excludes any activity in the defending half of the pitch. 'Progressive Passes' (PrgP) are defined as completed passes that propel the ball forward by at least 10 yards from its furthest point in the last six passes or directly into the penalty area, excluding passes originating from the defending 40% of the pitch. These metrics are critical for assessing how effectively a team progresses the ball into areas where they are more likely to score, reflecting strategic offensive movements. (see Figure 3)

Figure 3: Average Progressive Carries and Progressive Passes over the Previous Seasons

3 Model

For the analysis of English Premier League (EPL) team performance, I have developed a Bayesian linear regression model to predict the points (Pts) for teams based on several predictive metrics. The model-building process began with the dataset `cleaned_data_17_23.csv`, which includes team performance data from the 2017 to 2023 seasons. We specifically focused on five predictors: possession percentage (Poss), goals scored (Gls), expected goals (xG), progressive passes (PrgP), and progressive carries (PrgC). The choice of predictors is driven by their relevance in capturing the team's performance capabilities.

Figure 4 shows that the choice to employ a Gamma distribution for modeling points in our

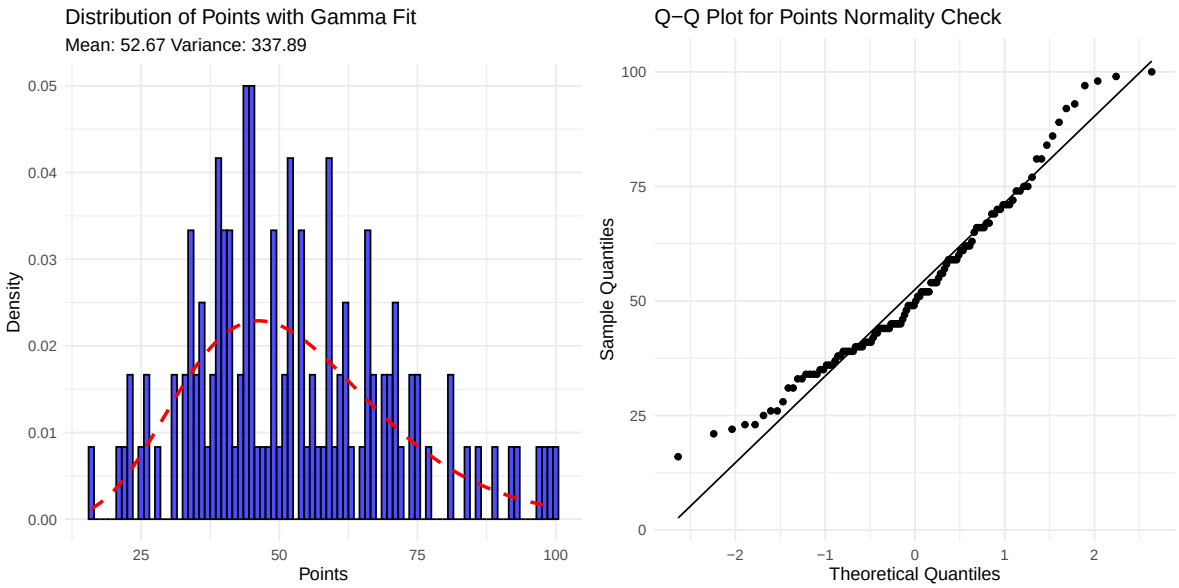

Figure 4: Ranking over Previous Seasons for Top Three Team

Bayesian regression analysis appears well-founded. The histogram of points exhibits a skewed pattern, consistent with the type of distribution seen in continuous data that does not fall below zero. The Gamma fit, as illustrated by the red line, adheres closely to the shape and spread of the observed points data, indicating that it captures the inherent variability effectively. This alignment between the empirical data and the theoretical Gamma distribution curve supports its suitability as a model for the points outcome.

The `stan_glm()` function from the `rstanarm` package allows us to incorporate prior knowledge into our Bayesian models systematically. For both points and wins, the predictors' coefficients are subject to normal priors centered at zero, which reflects a baseline assumption of no effect in the absence of data. However, the flexibility of Bayesian modeling is evident in the employment of a broader prior for the intercept, which accounts for more variability and does not constrain the model to a fixed starting point.

The `stan_glm()` function from the `rstanarm` package (R Core Team 2023) allows us to incorporate prior knowledge into our Bayesian models systematically. The predictors' coefficients are subject to normal priors centered at zero, which reflects a baseline assumption of no effect in the absence of data. However, the flexibility of Bayesian modeling is evident in the employment of a broader prior for the intercept, which accounts for more variability and does not constrain the model to a fixed starting point.

3.1 Model Equation

The regression equation for predicting the variable `Pts` based on the predictors (`Poss`, `Gls`, `xG`, `PrgP`, `PrgC`) with a log link is:

$$\log(\mu) = \beta_0 + \beta_1 x_{\text{Poss}} + \beta_2 x_{\text{Gls}} + \beta_3 x_{\text{xG}} + \beta_4 x_{\text{PrgP}} + \beta_5 x_{\text{PrgC}} \tag{1}$$

Where:

- μ is the expected value of `Pts` based on the gamma distribution assumption.
- β_0 is the intercept.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients for the predictors `Poss`, `Gls`, `xG`, `PrgP`, and `PrgC` respectively.

3.2 Additional Model Details

1. **Gamma Distribution with Log Link**: The response variable `Pts` is modeled with a Gamma distribution, suggesting that it is always positive and the distribution may be skewed. The log link function means that predictions for `Pts` are made on the log scale and must be exponentiated to obtain predictions on the original scale of `Pts`.
2. **Priors**: The regression coefficients $\beta_1, \beta_2, ..., \beta_5$ and the intercept β_0 have normal priors with a mean (`location`) of 0 and a standard deviation (`scale`) of 2.5. The `autoscale` parameter adjusts these scales based on the variability of the predictors, which can help in stabilizing the estimation process.
3. **Model Fitting**: The model is fitted using a Bayesian approach with MCMC sampling, specified to run for 4000 iterations with the first 2000 iterations being the warm-up phase. This warm-up phase helps in tuning the sampling algorithm for better convergence in the subsequent sampling.

3.3 Predicting and Interpreting Coefficients

To predict `Pts` for new observations, use the regression equation to compute $\log(\mu)$, then exponentiate this result to convert it back to the original scale:

$$\widehat{p} = e^{\log(\mu)} \tag{2}$$

This value represents the mean of the Gamma distribution for the predicted `Pts` given the values of the predictors. The coefficients β from the model provide insights into the relationship between each predictor and the log-transformed expected points, indicating how changes in the predictors logarithmically scale the expected points.

Table 5: Predicted Points for the Current Season

Squad	Rk	W	D	L	Pts	Pts.MP	Poss	Gls	xG	PrgC	PrgP	PredictedPts
Manchester City	1	22	7	3	73	2.28	65.9	74	67.1	994	1786	84.62227
Liverpool	3	21	8	3	71	2.22	61.0	68	73.4	773	1736	73.07530
Arsenal	2	22	5	5	71	2.22	59.5	71	62.9	701	1795	72.79558
Newcastle Utd	6	15	5	12	50	1.56	51.9	68	61.4	598	1287	71.60712
Aston Villa	4	19	6	8	63	1.91	54.5	65	55.8	704	1325	69.57045
Chelsea	9	13	8	10	47	1.52	58.8	60	60.7	711	1330	68.03940
Tottenham	5	18	6	8	60	1.88	61.5	60	55.5	814	1758	62.78087
Brighton	10	11	11	10	44	1.38	61.3	47	48.4	731	1577	52.47911
Fulham	12	12	6	15	42	1.27	51.2	48	46.4	588	1325	49.83429
West Ham	8	13	9	11	48	1.45	41.7	51	46.0	493	1092	49.36940
Manchester Utd	7	15	5	12	50	1.56	50.1	47	47.2	621	1260	49.36306
Wolves	11	12	7	13	43	1.34	48.5	43	42.4	596	1012	48.03353
Bournemouth	13	11	9	12	42	1.31	45.1	46	47.0	592	1143	47.10164
Brentford	15	8	8	17	32	0.97	44.2	45	51.4	413	1110	46.05912
Luton Town	18	6	7	20	25	0.76	41.8	43	38.1	590	940	44.94242
Nott'ham Forest	17	7	9	17	26	0.79	40.7	42	42.4	500	973	43.28906
Crystal Palace	14	8	9	15	33	1.03	41.2	37	37.5	408	938	40.02890
Burnley	19	4	8	21	20	0.61	46.7	32	33.5	605	999	39.05474
Everton	16	9	8	15	29	0.91	40.6	32	47.2	424	979	36.84848
Sheffield Utd	20	3	7	22	16	0.50	35.1	26	29.8	272	693	32.61636

Following model estimation, we applied these models to the latest season dataset (`cleaned_data_23_24.csv`) to predict points and wins for the 2023-2024 season. Using the `predict()` function, we calculated predictions for points, subsequently adding these predictions to the dataset to rank the teams. This method enabled us to identify potential league winners based on predicted performance metrics. (see Table 5)

4 Results

Table 6 presents the coefficients derived from our Bayesian regression analysis. Each coefficient estimate is accompanied by error bars, representing the 95% confidence interval for these estimates. The significance of these coefficients can be interpreted directly by their impact on the predicted league points. For example, a positive coefficient for goals scored (Gls) indicates a direct and positive influence on the team's predicted points for the season, reinforcing the importance of offensive strength.

Table 6: Coefficients from the Model for Points

term	estimate	std.error	conf.low	conf.high
(Intercept)	2.7751261	0.2147072	2.4298217	3.1249776
Poss	0.0115327	0.0085027	-0.0022441	0.0252727
Gls	0.0161498	0.0031524	0.0109342	0.0212851
xG	0.0013680	0.0043713	-0.0057928	0.0086352
PrgP	-0.0002544	0.0001844	-0.0005532	0.0000595
PrgC	0.0000583	0.0002528	-0.0003549	0.0004818

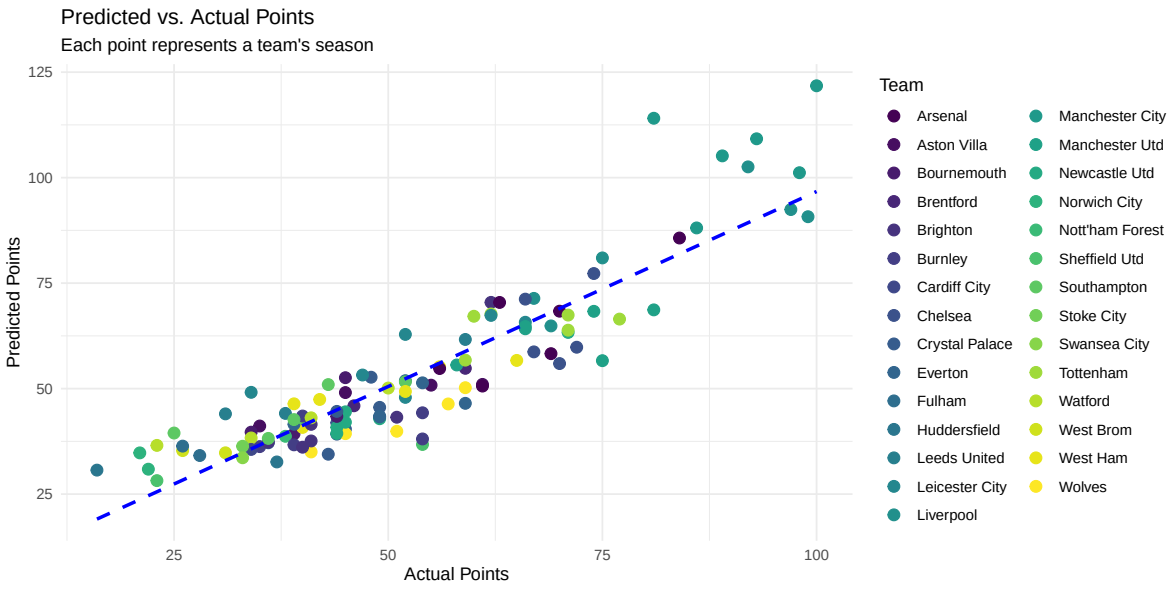

Figure 5: Predicted Points and Actual Points

By integrating these coefficients, we can formulate the predictive equation as follows, adhering to the structure outlined in our model equation 1, where μ represents the expected points, and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients for the intercept, possession, goals, expected goals, progressive passes, and progressive carries, respectively.

Figure 6 presents a analysis of predicted versus actual points for Arsenal, Liverpool, and Manchester City over six Premier League seasons, from 17/18 to 22/23. Across the seasons, the model's predictions align closely with the actual points for Manchester City, indicating a higher predictive accuracy for this team. The red lines, representing the model's predictions, fluctuate in parallel with the actual points denoted by the blue lines, particularly in the 17/18 and 21/22 seasons where the model almost perfectly captures Manchester City's performance.

For Liverpool, the model's predictions demonstrate a consistent underestimation of actual points, especially notable in the 18/19 and 19/20 seasons. This suggests that the model may not fully account for factors that led to Liverpool's actual performance exceeding expectations in those particular seasons. This might indicate that Liverpool might be underestimated for this season and there is still a chance to win Klopp's last season at Liverpool, even though the predicted winner is Mancity.

Arsenal's plot reveals mixed predictive accuracy. In the 20/21 season, the model's predictions were optimistic compared to the actual points. Conversely, for the 21/22 season, the model underestimated the team's performance. Such variances underscore potential areas where the model could be refined to better account for variables influencing Arsenal's season-to-season performance.

Figure 6: Predicted Standings and Actual Standings for Past Seasons

Figure 5 illustrates the relationship between the actual and predicted points for various Premier League teams across different seasons. Each dot represents a team's performance for a season, with the actual points on the x-axis and the predicted points on the y-axis.

A positive correlation is evident: teams with higher actual points tend to have higher predicted points, as shown by the concentration of dots along the dashed trend line. This indicates that, generally, the model's predictions are in sync with the actual outcomes. Teams such as Manchester City, which frequently secure points in the higher range, also show high predicted points, suggesting that the model effectively captures their continued success.

There are instances, however, where the model's predictions deviate from actual points. Some teams with lower actual points than predicted could indicate overestimation by the model or unexpected underperformances by the teams. Conversely, teams with higher actual points than predicted suggest instances where the model may have underestimated the teams' abilities or failed to anticipate successful outcomes.

Clusters of points around the trend line suggest the model's predictions are relatively accurate, while points further from the line indicate greater discrepancies between predicted and actual outcomes. The spread of points along the y-axis for any given value on the x-axis indicates the variability in the model's accuracy for teams with similar actual points.

That said, the model demonstrates a fair degree of predictive accuracy, with a trend line that captures the central tendency of the data. However, some outliers suggest opportunities for refining the model to improve its prediction accuracy for certain teams or under certain conditions.

5 Discussion

5.1 Implications of Findings for Team Strategy

The analysis of performance metrics such as goals scored (Gls), expected goals (xG), and progressive plays highlights their critical importance in predicting team success in the English Premier League. This understanding offers a tangible pathway for teams to enhance their league standings by focusing on these key areas. For instance, improving a team's xG might involve strategic recruitment focusing on players known for high-quality shot creation and finishing, coupled with tactical setups that maximize scoring opportunities. Similarly, enhancing progressive passes and carries could involve training sessions dedicated to improving players' ability to break lines and advance the ball under pressure.

Moreover, these metrics can guide more nuanced strategic decisions during matches. For example, if a team is trailing and needs to improve its chance of scoring, managers could adjust their tactics to prioritize forward movements and riskier, more aggressive plays. This could involve pushing full-backs higher up the pitch, employing more direct attacking midfielders, or switching to formations that overload certain areas of the pitch to create mismatches against

opponents. Thus, the insights from this analysis not only serve to guide long-term strategies but also adapt real-time decisions that capitalize on the dynamics of a game.

5.2 Comparative Analysis of Model Predictions Across Teams

The varying accuracy of model predictions across different teams like Arsenal, Liverpool, and Manchester City invites a deeper exploration into how specific team characteristics may influence predictive outcomes. For example, Manchester City's consistent tactical setup and high-quality squad depth might make their performance easier to predict compared to a team like Arsenal, which has undergone significant tactical shifts and roster changes in recent seasons. This aspect of the analysis can highlight how stability and consistency in playing style, personnel, and management can lead to more predictable performance outcomes.

Further, the discussion could explore how different tactical approaches impact key performance metrics. For instance, teams that focus on a high-pressing style may have higher progressive metrics but could be more vulnerable defensively, which should be accounted for in the model. Similarly, a team that plays a more conservative, counter-attacking style might outperform its expected goals metrics due to the higher quality of chances created on the break. Understanding these nuances could allow for adjustments to the model that better reflect the strategic approaches of individual teams, leading to more accurate predictions and insights into the effectiveness of different football philosophies.

5.3 Limitations and Future Improvements

Despite the model's effectiveness in forecasting team points based on key performance indicators, its limitations must be acknowledged to refine future predictive efforts. One significant limitation is the exclusion of variables such as player injuries, transfers, managerial changes, and even weather conditions, which can all drastically affect game outcomes. Future models could benefit from incorporating these factors, perhaps through dynamic adjustments that take into account the probability of key players missing games or the impact of a new manager's tactics.

Additionally, integrating data from advanced tracking metrics like player work rate, positional heat maps, and psychological factors such as team morale could enhance the model's accuracy. For instance, including the impact of high-profile players returning from injury or new signings integrating into the squad could provide a more accurate reflection of a team's mid-season potential. This approach requires not only more data collection but also advanced analytical techniques that can handle the increased complexity of the model. As predictive analytics in sports continues to evolve, embracing these complexities will be meaningful in developing more accurate models that can anticipate the unpredictable nature of football.

References

Clifford, Sean Markus. 2024. "When Did Arsenal Last Win the Premier League? Record Since Invincibles Season." https://www.sportingnews.com/ca/soccer/news/when-did-arsenal-last-win-premier-league-invincibles/jll8fofzojviikcgtkjzfhzx.

FBref. 2018. "FBref.com: Football Statistics and History." https://fbref.com/en/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Liverpool FC. 2024. "Jürgen Klopp Announces Decision to Step down as Liverpool Manager at End of Season." https://www.liverpoolfc.com/news/jurgen-klopp-announces-decision-step-down-liverpool-manager-end-season.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Pollard, Rob. 2023. "Manchester City Complete Historic Treble with Champions League Success." https://www.mancity.com/news/mens/man-city-champions-league-winners-treble-premier-league-fa-cup-63822026.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Whitmore, Jonny. 2023. "What Is Expected Goals (xG)?" https://theanalyst.com/na/2023/08/what-is-expected-goals-xg/.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jenny Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and PBC Posit Software. 2024. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.