Predicting Premier League Team Performance (2025–2030)

Justin Nam, Moe Jawadi, Steven Al-Sheikh
San Diego State University
CS 549: Final Report

1. Introduction and Research Problem

For the past few years, it's very obvious and clear as day that there has been a shift in player's decision-making skills in football, also known as soccer. Teams are improving through the use of data not for the purpose of observation, but utilizing it as a tool for improvements and a guidance tool in their decision-making regarding their teams. In the Premier League, a league where there is a lot of pressure, competition and expectations. Often leads to financial pressure from the owners and the fans. Using statistics and data to predict how your team is able to perform for future season's is someone to take into account. These stats will not only be able to be viewed and considered by the owners, the players, the staff, but also by the fans to be able to see how each person is performing and their trajectory.

The main focus of this project is to take past English Premier League data to predict each team participating from seasons from the past and predict their performance within the next 5 years. The main goal is to calculate and get a rough estimate of an entire season's performance of a team based on results, points, and performance from previous seasons. The purpose of this is to be ahead in our data collection to obtain our predictions of each team's points by the end of each season from 2025 to 2030. We will obtain this information through the performance of the best and worst teams through machine learning algorithms. We are aiming to use linear regression and apply it to our data collected. This will ultimately help us project what will be the point estimation for each team moving forward. Due to the complexity of this project, we will face challenges with long term goal accuracy due to the risk we face of possible teams being relegated from the premier league based on bad performances and that can risk their point total

There are many factors that can impact a team's performance on and off the field and that will ultimately measure their success due to the implications it may cause and that can be out of the prediction's hands. The perception of this model that can be taken into consideration will ultimately help premier league clubs analyze team trends, performance, and different expectations. This model will create a broader impact on sports analytics and the sports industry by providing machine learning models to help predict the outcome for different real-world scenarios using previous data

The model goal is to keep things very simple, unique, and etiquette. This model does not have super-powers. It's very easy to explain and understand and that's where its strengths come from. The mode's influence is very easy to understand as it's a model based on linear regression that estimates and takes into account the factors of team wins, losses, goals scored, and goals conceded that will ultimately be the deciding factor in these predictions, each of these have a factor in contributing in the predictions of these standings for the premier league. The goal is to give clear, and precise predictions from the upcoming years of 2025 to 2030.

2. Related Work

There have been a number of past studies that have tried using statistics and machine learning to predict football outcomes. Most of the studies focused on predicting the results of individual matches using logistic regression, support vector machines, or other types of models. For example, Tax and Joustra (2015) used public data and machine learning to predict results in the Dutch football league. Their study mostly focused on past stats and team rankings to help them predict future outcomes. In another study, Joseph et al. (2006) applied Bayesian inference to predict match results and showed that using probabilities gave better results than using set rules.

While most models try to predict the outcome of the individual games, our approach focuses on the whole season. Instead of analyzing results on a game-by-game basis, the model aggregates wins, losses, and draws to produce cumulative performance indicators at the end of each season. This strategy supports a more comprehensive analysis of long-term trends across multiple years. Linear regression, despite its relative simplicity, continues to be widely utilized in sports analytics due to its clarity, computational efficiency, and capacity to reveal fundamental linear patterns within the data.

Alongside regression models, other methodologies such as Elo rating systems and Poisson regression are commonly applied in football forecasting. Elo-based systems adapt well to evolving team performance but require careful tuning of match importance parameters. Poisson regression is typically applied to model goal scoring distributions, assuming

independence and fixed scoring rates, which can be limiting for long-term forecasts. Our use of linear regression avoids these assumptions by instead focusing on cumulative outputs, allowing for cleaner trend modeling at the season level.

3. Methodology and Technical Details

We obtained data from the English premier league that covers every season from 1993 to 2024. The data includes a variety of stats for each team and their respective season. These stats include year, team name, the team's final standings, the number of games played that season, the number of wins and losses that season. They also track the most important stats that lead to wins which are, goals scored, goals allowed, the goal difference, and the cumulative points earned. This dataset has given us solid grounds to help analyze how team performance has changed and adapted over time. The data was stored in CSV format before being loaded into the jupyter notebook for processing.

We begin by using the pandas library that helped us clean up and group the appropriate data together as mentioned above. The statistics from the library has given us the knowledge on why teams have done so well in the league. A heatmap was generated (Figure 1) to illustrate the differences in wins and goals across all the different teams. We also implemented plots and bar charts (Figures 2 and 3) to discover and visualize the relationships between performance stats, goals scored and win percentage. These illustrations supported the notion that there is a positive trend between scoring goals and winning games, more than their counterparts. With these positive trends also came negative ones such as conceding goals and overall success.

Our machine learning approach used simple linear regression and we show this by treating each team as an individual entity with a corresponding time series of the total points at the end of the season. These were regressed against the year, with the assumption that performance follows a trend. These trends were either improving, declining, and or stable. This created over 50 team specific models, which each projected the total points for the six future seasons from 2025 and 2030. The teams with incomplete data (new teams or relegated ones) their predictions were still shown but were noted for potentially being less reliable due to fewer samples used for training.

In addition to point-based regression, we also looked at other regressions for the relations between goals scored and win percentage, as well as goals conceded and loss percentage (Figures 4 and 5). These models helped us understand how key performance metrics can influence the outcomes. It's very obvious and clear as day that these models gave us insight into how core performance metrics contribute to a team's success and showed how offensive and defensive stats can be powerful tools in predicting whether a team will perform well or fail.

To make the predictions, the models were used to predict each team's point trend for the future. The results were stored into a dataframe that was sorted by year and team. We used this data to make visualizations that showed stats like the predicted top and bottom teams, team performance categories, and points earned by teams across the league. These were then plotted using matplotlib for easier understanding. These stats would be very useful to the owners, the players, the staff, and also by the fans to be able to see how their team or their rivals could be performing and their trajectory

While linear regression assumes a direct relationship between inputs and outputs, this can miss some other features or complexity in the sport of football. However, this is still useful for pointing out overall trends, helps make basic predictions, and supports early stage testing. In the Premier League, a league where there is a lot of pressure, competition and expectations, utilizing these statistics and data for predicting how their team is able to perform for future seasons is something that should be taken into account, especially when there's immense financial pressure from the owners and the fans.

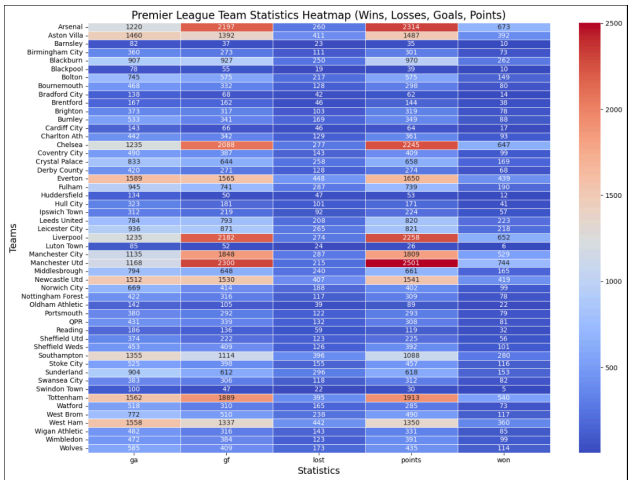

Figure 1. Heatmap of cumulative Premier League statistics (1993–2024) by team, showing total wins, losses, goals scored (gf), goals allowed (ga), and total points.

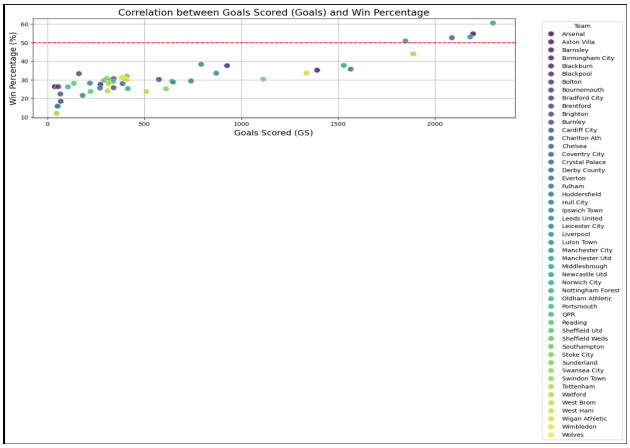

Figure 2. Scatter plot showing correlation between total goals scored and win percentage for all Premier League teams. The red dashed line indicates the 50% win threshold.

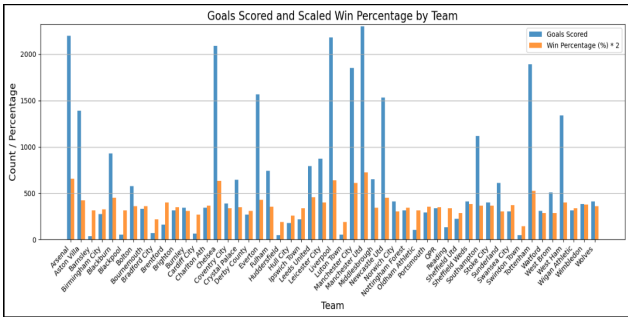

Figure 3. Bar chart comparing total goals scored and scaled win percentage (×2) across all Premier League teams. This highlights the proportional relationship between scoring and team success.

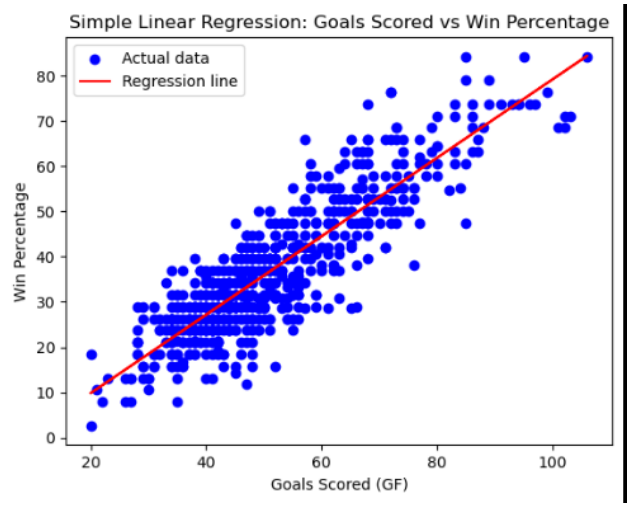

Figure 4. Linear regression model fitted on goals scored versus win percentage. The red line represents the regression fit used to model team success based on offensive output.

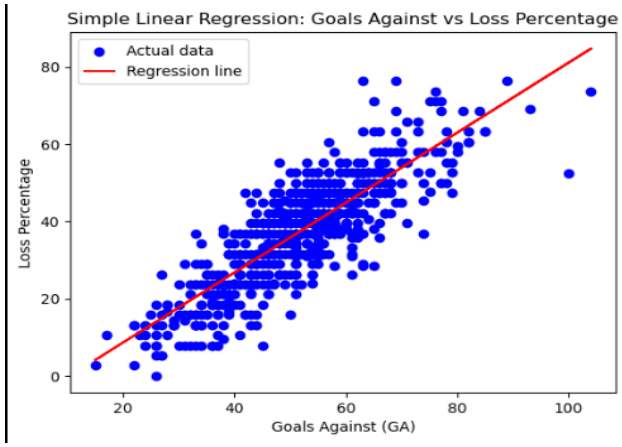

Figure 5. Linear regression model fitted on goals against versus loss percentage. The red line represents the regression fit used to model how conceding goals impacts a team's likelihood of losing.

4. Experimental Results and Analysis

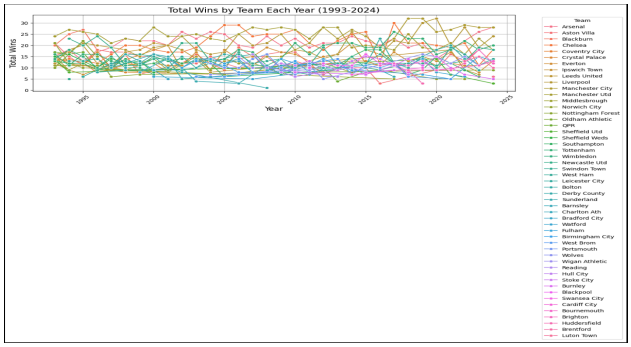

Figure 6. Total wins per team from 1993 to 2024. This historical trend data serves as the foundation for training the linear regression model used to predict future performance.

The model yielded the following insights:

- Top 4 Teams (2025-2030): In Figure, there is a display of the projected dominance and success of clubs like Manchester United, Arsenal, Liverpool, Chelsea and more. Liverpool is projected to achieve the most success out of these 4 clubs. These teams have showcased their consistent success and

play styles that helped them win football games. Their success comes from their growth that comes from their financial strength, player development and stability.

- Bottom 4 Teams (2025-2030): We generated a line plot which is found in Figure 6. In this figure, we showcased a predicted decline in points among many of the underperforming teams that have found themselves at the bottom of the league. Some examples of these clubs are Bradford City, Coventry, Sunderland, and more. Some teams have shown a great decline that their projected points have fallen below zero, which means that they get relegated from the Premier League and go down to The Championship which is the 2nd highest division in English football. This figure also suggests that there is a gap in points between mid-tier teams and lower level teams.
- All Teams: Figure 8 illustrates a full look at the point predictions for the league and it shows that teams are grouped into different levels based on performance. This figure also actually helps us identify how mid-tier teams are performing, it shows whether they are improving in the standing or whether they are declining and might be heading to the lower level group.
- Focused Top 4 View:Figure 9 provides a zoomed-in perspective on just the top 4 clubs, allowing better visual differentiation of their point trajectories. Here, we can see Liverpool pulling ahead slightly, while the other three clubs remain closely matched. This could be used to estimate Champions League qualification probabilities. Additionally, the figure provides a baseline for fan discussions and media projections regarding which top teams are best positioned for future dominance.
- Draw Predictions: Furthermore, to the wins and losses figure, we also modeled the number of draws for each team (Figure 10). Draws showed low differences throughout the league, with almost the majority of teams predicted to end the season with 9-11 draws. The static trend displays that draws do not have a strong correlation to long term team performance trends and more so by in-game differences. This could be studied upon further in the future to see if draws can be used to give teams in this league or any league for that matter, an equal opportunity to win.

Overall, the results have shown us that even a simple linear model can give us insights about how teams are performing over time. The figures produced were easy to interpret and, in most cases, matched what people already expect to see based on the results from past seasons.. These qualities make the model a good starting point for future improvements.

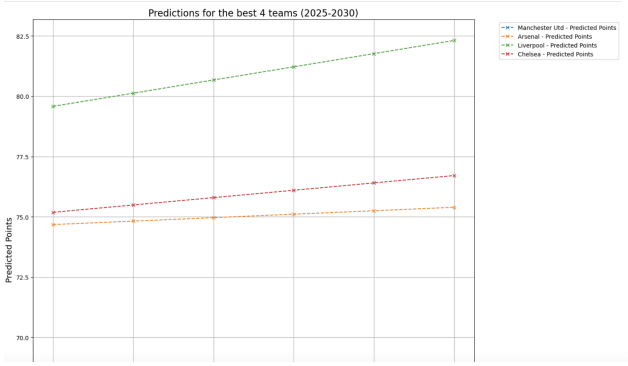

Figure 7. Zoomed-in predictions for the best four Premier League teams (2025–2030), highlighting Liverpool's steady rise and relative consistency from Chelsea, Manchester United, and Arsenal.

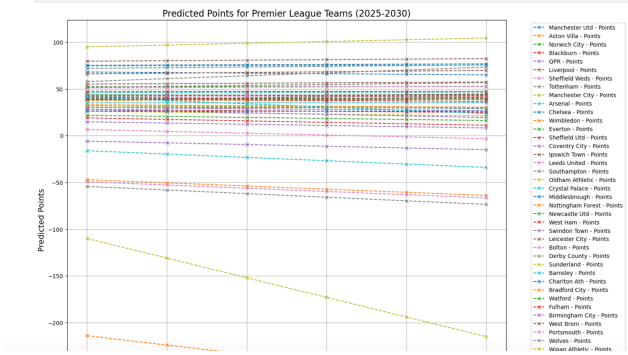

Figure 8. Predicted point trajectories for all Premier League teams from 2025 to 2030. The chart shows stratification into high-, mid-, and low-performing clusters over time.

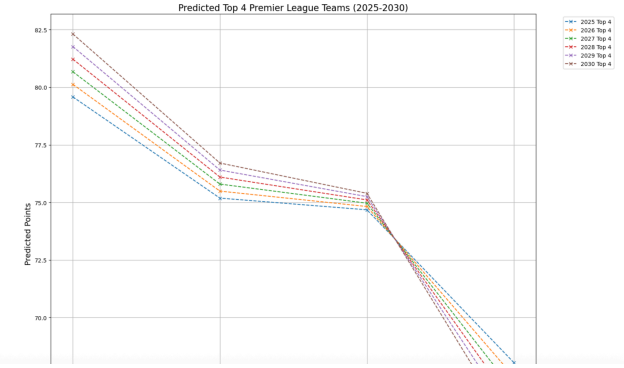

Figure 9. Predicted performance of the top four Premier League teams from 2025 to 2030. Teams such as Liverpool and Manchester United are forecasted to maintain a strong position above the 75-point mark.

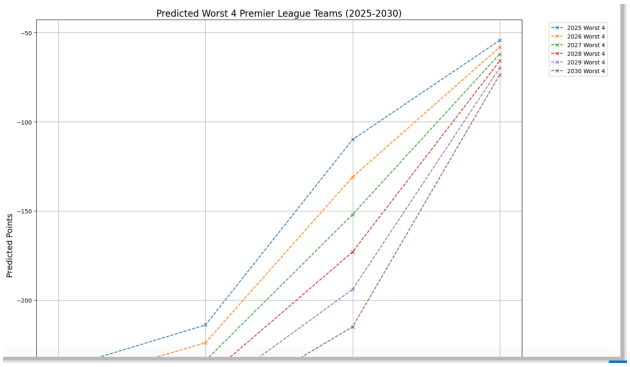

Figure 10. Linear regression predictions for the bottom four Premier League teams by season (2025–2030). The points reflect consistent decline among underperforming teams, with some falling into negative-point territory due to regression trends.

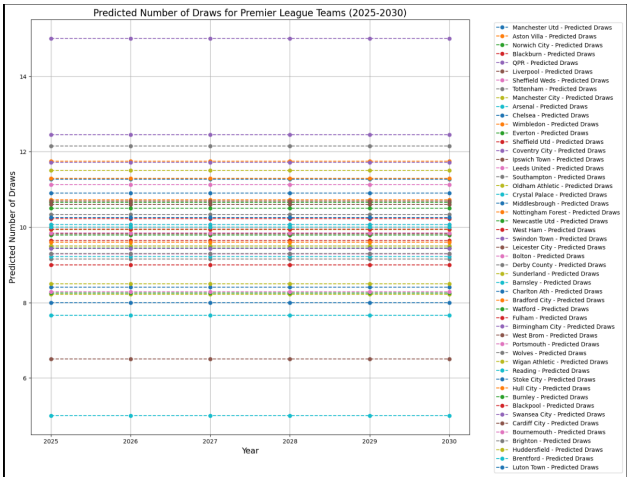

Figure 11. Predicted number of draws for each Premier League team from 2025 to 2030. Most teams are expected to average 9 to 11 draws per season, showing relatively low variance in draw frequency across the league.

5. Conclusion and Future Work

In conclusion, The project's goals of demonstrating the model of linear regression for team standings projection is a big asset and tool to base each Premier League performance in future years. The model has successfully identified the different types of trends of the higher tier team's, and the lower tier team's based on previous results and therefore predicting the future results. Due to relegations in and out of the premier league, certain teams went below the threshold point total and were in the negatives due to bad performances and data analysis of their relegation. We were able to create different tiers of each team based on their record history within the past decade and that ultimately helped notice the different patterns that were being displayed. Pushing updates for this model is not too difficult, in fact it's fairly easy because we simply need to just add the following season's statistics when we have already established prior so it will only be a limited change to introduce new data into our model.

Moving forward, there is a lot of opportunity and room for improvement by implementing different models such as logistic regression, to move away from continuous values and focus on binary classification based on likely outcomes of 0's and 1's. for non-linear purposes that's what we would include moving forward to focus on a different approach for this project as we can provide categories for teams in the top 4 of the premier league race, and predict probability such as will team's finish at a certain standing in the table, the probability that a team will finish with a certain amount of wins and more.

Comparing our current model's predictions to how the 2025 season will conclude will tell us a lot about our model's predictions and how accurate it is. Learning the accuracy, absolute errors, and mean errors are tools that will determine how close we were to the real standings of the 2025 premier league season based on data we have used throughout this project. This will ultimately let us know how effective our model will predict for the season of 2025.

6. Contributions of Each Group Member

Steven Al-Sheikh: Developed the predictive modeling code using Jupyter, trained all team regressors, and generated graphs.

Moe Jawadi: Cleaned the data, conducted exploratory data analysis (EDA), and supported statistical correlation visualizations.

Justin Nam: Authored the final report, compiled graphs and figures, and interpreted results.

References

[1] N. Tax and Y. Joustra, “Predicting the Dutch football competition using public data: A machine learning approach,” *Transactions in Sports Science*, vol. 3, no. 2, 2015.

[2] A. Joseph, N. Fenton, M. Neil, and A. Constantinou, “Predicting football results using Bayesian networks and historical data,” *Knowledge-Based Systems*, vol. 19, no. 7, pp. 544–553, 2006.