How Well Does Non-Penalty xG Difference Explain Premier League Final Position?

Summary

This brief analysis examines the relationship between non-penalty expected goal difference (NPxGD) and final league position in the English Premier League for the previous 11 completed seasons (2014/15 to 2024/25). I find a strong association between NPxGD and final league position using team-season data. Intuitively, it makes perfect sense that teams with the best or highest NPxGD will consistently rank among the highest in the league table, which is, indeed, the conclusion reached in this analysis. By regressing league position on NPxGD, a large share of variation in final position is explained ($R2 = 0.74$), and the estimated relationship is practically unchanged when controlling for season fixed effects. These descriptive results support a key takeaway for performance analysis in soccer: that, over a full season, underlying chance prevention and creation closely track league outcomes.

1. Motivation & Question

Expected goals (xG) has fast become one of the most discussed and utilized statistics to evaluate the quality of chances created or conceded in soccer. Previously, teams used a combination of the number of chances created/conceded, along with raw goals, to analyze and optimize trends. However, this approach was never optimal because these data points were heavily and consistently affected by variance. In comes xG-based metrics, which are much less sensitive to short-run finishing variance and can more comprehensively reflect a team's underlying performance.

In this project, I focus more specifically on non-penalty expected goals because this further reduces the likelihood of variance biasing the results. Penalties are often unpredictable and infrequent actions that cannot be relied upon for goalscoring opportunities, which is why

isolating chance creation from open play and typical attacking situations would add more explanatory power to this analysis.

The core question is straightforward: **How strong is the association between non-penalty expected goal difference (NPxGD) and league position in the English Premier League?** This serves to be more of an applied analysis designed to quantify and visualize how adequately an underlying metric, such as NPxGD, aligns with league position over many seasons.

2. Data

The dataset contains 11 seasons' worth of Premier League team-season observations, spanning 2014/15 to 2024/25. All data were gathered from Understat, then downloaded and compiled into a single dataset, with each season contributing 20 teams.

- Unit of observation: Team-season
- Seasons: 2014/15 to 2024/25
- Total observations: 220

The key variables for the study are final league position (Rank) and non-penalty expected goal difference (NPxGD).

- Final league position (Rank): integer from 1 to 20; 1 representing the champions of that season
- NPxGD is defined as: NPxGD = NPxG - NPxGA, where NPxG represents non-penalty expected goals for, and NPxGA represents non-penalty expected goals against.

Other statistics, such as wins and points, are present in the dataset, but are not controlled for in the main specification because of how they are mechanically linked to league position. Including them would only introduce bad controls and multicollinearity without providing any additional insights about the relationship.

3. Approach

To quantify the relationship between NPxGD and league position, this is the simple linear regression I used:

$$\text{Rank}_{i,t} = \beta 0 + \beta 1 * \text{NPxGD}_{i,t} + \varepsilon_{i,t}$$

i indexes teams and *t* indexes seasons.

I estimate another model with season fixed effects:

$$\text{Rank}_{i,t} = \beta 0 + \beta 1 * \text{NPxGD}_{i,t} + \gamma t + \varepsilon_{i,t}$$

Although the analysis is essentially cross-sectional within each season, the fixed effects are intended to control for any season-wide shifts such as changes in playing style or league-wide scoring output.

4. Results

The baseline regression displays a robust positive relationship between NPxGD and league position with an estimated coefficient of 0.219. This means that a +1 increase in NPxGD for a team over the season is associated with an improvement of 0.22 league positions on average. Furthermore, the estimate is statistically significant at the 1% level.

Adding season fixed effects left our estimate unchanged, which was to be expected because league position is an ordinal outcome within each season, ranging from 1-20, and NPxGD is also fundamentally meaningful within-season. Therefore, they just lead to a shift in intercepts by year, but they do not inherently change the within-season association between NPxGD and league position.

Figure 1 depicts the key takeaway clearly. The relationship between NPxGD and league position is strong and consistent across all of the observed seasons, with an evident upward-sloping pattern. The teams with significantly positive NPxGD are all clustered near the top of the table,

while teams with largely negative NPxGD are clustered in the relegation positions, with no clear outliers present.

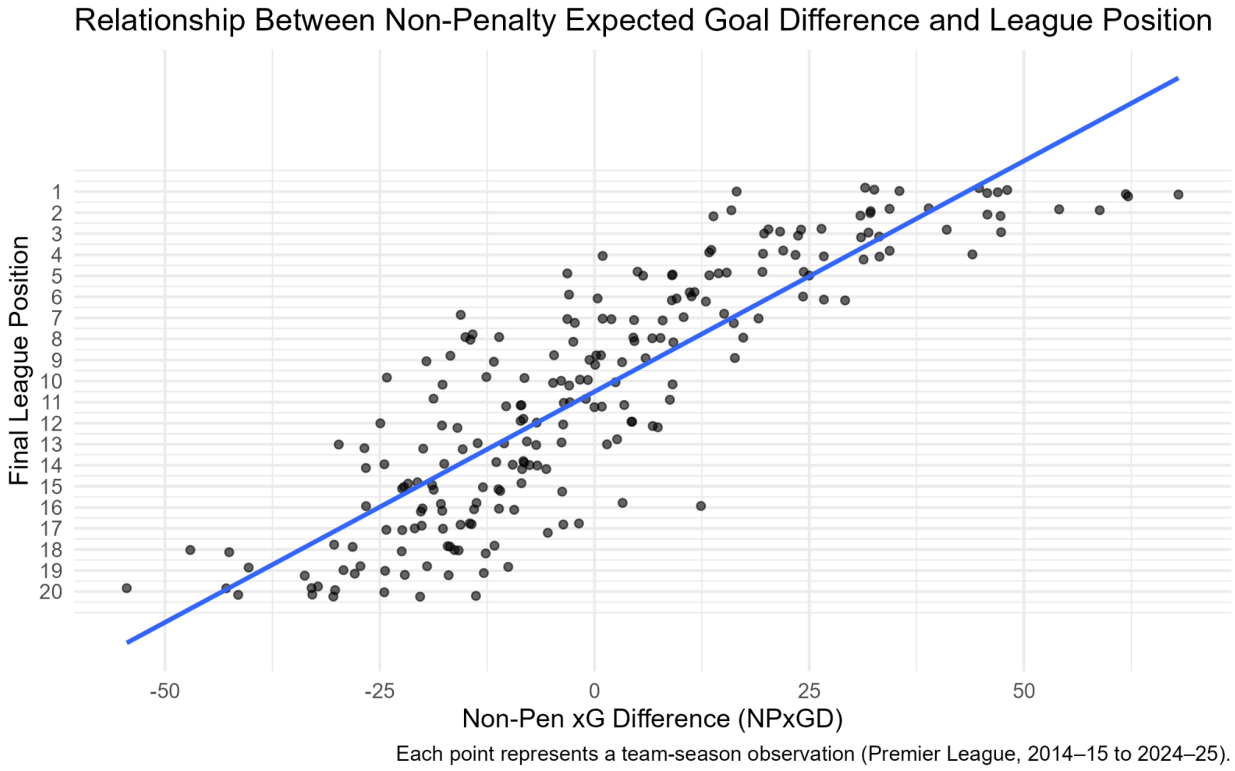

Figure 1

5. Discussion and Conclusion

Ultimately, the findings support any soccer fan's intuition, which is that, essentially, the teams that accumulate the most high-quality chances created, and least high-quality chances conceded, tend to finish near the top of the table. Narrowing down the scope to non-penalty metrics helps keep the analysis centered around a team's genuine ability to create and prevent chances instead of unpredictable and rare actions.

This relationship between NPxGD and league position, no matter how strong, would be considered descriptive rather than causal. Multiple factors, such as squad quality, injuries, and other unobserved factors, can influence both NPxGD and league outcomes, so the analysis does

not identify a causal effect, but rather a very strong correlation. There are other limitations to this study, including the fact that league position is an ordinal variable and that using team-season level data disregards match-level dynamics such as red cards and fixture difficulty. However, this is intended to be a baseline descriptive analysis and not an exhaustive model.

Overall, it is safe to say that the evidence suggests that NPxGD is a thoroughly informative statistic on a team's performance across a full season in the Premier League. The clear visual relationship displayed in Figure 1, along with the robust statistical fit, reinforces our key takeaway: teams that consistently create higher-quality chances than they concede and by a relatively large margin tend to attain a higher final position in the league table.