Predicting Form Maintenance of Mid-Season Overachievers in a Big Five Football League

Gabriel Pacheco
Bradley Department of Electrical
& Computer Engineering
Virginia Tech

1. Background

The "Big Five" football leagues refer to the top professional football markets in Europe: England, Germany, Spain, Italy, and France. [1] Among these, the Premier League in England stands out as the highest revenue generator. Over the past decade, it has produced an astounding €7.4 billion in revenue, far surpassing the Bundesliga in Germany, which follows with €3.9 billion. [2]

In football, it is common to see teams with the most financial power and historical success finishing at the top of their league standings by season's end. However, occasionally, "special teams" emerge, teams that perform exceptionally well during a particular season, defying Fexpectations.

These special teams often begin the season with high confidence and momentum, positioning themselves as contenders for high rankings (top seven positions) by the mid-season break, which typically occurs during the Christmas holidays. This mid-season break serves as a critical juncture: teams either maintain their strong form into the second half or experience a significant drop in performance. The question then arises: "can we predict which special teams will sustain their form?".

An excellent example of this phenomenon is Leicester City during the 2015–2016 Premier League season. Before the start of the season, oddsmakers gave Leicester an almost negligible 0.004% chance of winning the league. [3] Yet, by the mid-season break, they were at the top of the standings, competing strongly against heavyweights like Arsenal and Manchester City. The key question at that time was whether Leicester could sustain their performance or if they would regress to the middle of the table, where they usually stand. Defying all expectations, Leicester City went on to win the Premier League, setting a historic precedent in modern football. [4]

There are two motivations this study of such teams. First off, is economic opportunities. Accurate predictions about special teams sustaining their form could challenge betting odds and create opportunities for significant financial gains. Second off, coaching insights. By identifying key parameters that influence team performance in the latter half of a season, coaches could leverage mathematical models as tools for maintaining team momentum. An example of this is how other data-driven approaches have helped underdogs in sports in the past, such as with the Oakland Athletics back in the 2002 season. [5] By applying mathematical techniques to football, this research aims to contribute to the growing intersection of analytics and sports.

This project will focus on analyzing the Premier League over the past 20 seasons. Using data from these seasons, along with various analytical tools, a model will be developed to predict the probability of teams (among which we can find special teams) maintaining their form. The code for this model is available in the following GitHub repository.

2. Objectives

- Develop a model that analyzes data from individual games played up to the middle of the season and predicts the probability of teams maintaining their form for the remainder of the season. This includes identifying "special teams", which are those whose points exceed the predicted amount, according to betting odds, by a specified margin (e.g., 5 points).
- Create a model that evaluates the predictions developed by mid-season with the actual final standing results, assessing the accuracy of the predictions. This will be done by evaluating both the predictions made using climatology and those made by the model.
- Design the model with scalability in mind, ensuring that it can be easily applied to other leagues with minimal code changes. The model should also allow for the integration of additional parameters in the future, facilitating further development of the project.

3. Model and methodology

The developed model provides a binary prediction on whether a team will maintain its form throughout the rest of the season. At mid-season, it determines the number of points a team should have accumulated by that point, based on a combination of three scores: actual points, performance points, and betting odds points. It then compares this total to the k-nearest neighbors from previous seasons who had a similar number of points. Using historical data, the model calculates a probability based on whether those similar teams maintained their form or not. If the calculated probability exceeds 0.5, the model predicts that the team will maintain its form; otherwise, it predicts a drop in form. Finally, a skill assessment is conducted by comparing the model's predictions with the actual outcomes (derived from the historical end-of-season standings). This evaluation concludes with a skill measurement using ignorance-based scoring. The following section outlines the steps taken to implement this procedure.

3.1. Database

The historic databases from football-data.co.uk [6] were used as the data source for the project. Here, excel sheets with game-by-game data that contained: teams playing, goals, total shots, shots on target, numerous betting odds, corners, fouls, and others were provided.

3.2. Data processing – combination of score

The objective of merging the three types of points: real points, performance points, and betting odds points, was to create a more comprehensive measure of a team's performance at the middle point of the season. While real points reflect the actual results of matches, they do not provide insight into a team's true capabilities or future performance. By incorporating performance points and betting odds points, we can assess whether a team's success is sustainable or simply due to luck, and consider the broader context of their historical performance and future expectations. Merging these three scores provides a more nuanced evaluation of a team's state, beyond what is reflected in the standings table.

- **Real points:** These were determined by verifying the outcome of each match in the current row of the Excel sheet for the given season. Following football rules, 3 points were assigned to the winner, 0 points to the loser, and 1 point to each team if the match ended in a draw. This approach established the "real" number of points the team had earned by the middle of the season.

- **Performance points:** The purpose of these points was to assess whether a team's high position was due to "luck." If a team had few shots on target but scored many goals, it suggested that luck played a significant

role up to that point, and a regression toward their average performance was expected. Conversely, if a team scored highly due to a significant number of shots, it indicated that they were likely capable of maintaining that level of performance. Thus, performance points provided a means of evaluating a team's true capabilities beyond just match outcomes.

After a discussion with Edward Wheatcroft, I was advised that the best approach would involve using the expected number of goals. As explained by him, the number of goals for each team can be modeled using the Poisson distribution, with a conversion rate of 15%, obtained from a football statistics website. [7] The expected goals for each team were calculated as follows:

- Expected home goals = $number\ total\ shots\ home\ team * 0.15$
- Expected away goals = $number\ total\ shots\ away\ team * 0.15$

Based on these Poisson distributions, the difference between the number of goals scored by each team follows a Skellam distribution, assuming independence. Using the Python library "skellam", probabilities for home wins, away wins, and draws were computed. These probabilities were then normalized and used to calculate performance points as follows:

- Points home team = $3(home\ win\ probability) + 1(draw\ probability)$
- Points away team = $3(away\ win\ probability) + 1(draw\ probability)$

- **Betting odds points:** These points were derived from betting odds probabilities, reflecting the market's perception of a team's future performance based on past results. A team with poor betting odds typically indicates a history of poor performance, suggesting they are likely to continue underperforming. These points were included to account for the team's historical trends and how they might influence future results.

  The expected points from betting odds were calculated for both the home team and the away team and added to their respective totals. First, probabilities were derived from the bookmakers' odds as follows:

  - $Home\ win\ probability = \frac{1}{Odds\ home\ win}$
  - $Away\ win\ probability = \frac{1}{Odds\ away\ win}$
  - $Draw\ probability = \frac{1}{Odds\ draw}$

  The expected points were calculated as follows:

  - Points home team = $3(home\ win\ probability) + 1(draw\ probability)$
  - Points away team = $3(away\ win\ probability) + 1(draw\ probability)$

  The combination of these three scores was necessary to determine a more comprehensive measure of a team's state at that point in the season, beyond the simple reflection of the number of points shown in the standings table. The combination of the three scores was calculated using a weighted average, with weights assigned as 0.3 for points, 0.3 for betting odds points, and 0.4 for performance points. Performance points were given the highest weight because they most accurately reflect what actually occurred on the field.

3.3. Data Processing – K-Nearest Neighbors:

The approach described earlier for calculating the combined points up to the middle of the season was applied not only to the season under evaluation, but also to another 20 seasons of data. For each of these seasons, the combined points for all team`s were calculated up to the middle of the season, and then these were compared to their ability to maintain their form up to end of the season (by comparing final standing with mid-season standings).

Therefore, a correlation was established between the number of combined points and the frequency with which teams maintained their form. This allowed for the estimation of probabilities: when a team reached a certain number of points by mid-season, we could compare this to historical data to determine how often teams with a similar number of points had maintained their form in the past.

For example, if a team had accumulated 32 combined points by the middle of the season, the analysis would look at

the *k=5* nearest neighbors in the historical data. These would be teams with combined point totals close to 32. The probability of the current team maintaining its form would then be calculated as:

$$Probability\ of\ maintaining\ form = \frac{\text{Number of teams with 32 points (or close) that maintained form}}{Total\ number\ of\ teams\ with\ 32\ points\ (or\ close)}$$

For instance, if 3 out of 5 teams with approximately 32 points historically maintained their form through the rest of the season, it would give a 60% probability that the current team would also maintain its form. This method provided a way to assess the likelihood of a team sustaining their performance based on historical trends and patterns.

3.4. Processing Data – Climatology

The definition of climatology was considered using the same calculations as before but without combining points from multiple metrics. Instead, only betting odds were used to calculate the points and define the probability of a team maintaining its form. This approach was chosen because betting odds are provided by bookmakers, making them a reliable baseline. Comparing the model's predictions against this baseline allows us to determine whether the model provides any significant advantage or differences compared to what bookmakers can offer.

3.5. Processing Data – Skill

Using the probabilities obtained in the previous section, a threshold was applied to determine whether a team would maintain its form or experience a drop. The threshold was set at 0.5: if the probability exceeded 0.5, the team was predicted to maintain its form. On the other hand, if it was below 0.5, the team was predicted to drop in form. At the end of the season, a comparison between the predicted outcome and the actual result was conducted.

A team was considered not to have dropped in form if it fell by a maximum of two positions in the standings. This adjustment accounts for the possibility that a team can drop slightly in the rankings while still performing well, especially if other teams significantly improve their performance in the second half of the season.

The skill of the prediction was measured using ignorance-based points, which penalize extreme forecasting and overconfidence. This approach utilized the formula provided by Leonard Smith [8], with a pzmin value of 0.0000625 to ensure the smallest possible baseline:

$$\text{IGN points} = 25(1 + \log2(\max(p, pzmin)))$$

This evaluation was performed both for calculations derived from the combination of the three metrics and those based solely on climatology.

3.6. Known Neglecteds

The primary risk associated with this project lies in the significant number of uncertainties inherent in modeling sports performance, many of which cannot be easily quantified. For example, factors like the mental state of players or the manager, though impactful, are not captured in any database but can heavily influence how well a team performs during the remainder of the season. Similarly, unforeseen events such as injuries sustained during the Christmas break can disrupt the momentum of key players, which could be crucial to a team's success.

Additionally, the lack of access to certain valuable data posed limitations on the project and the development of the mathematical model. For instance, information about the "man of the match" awards could help identify standout players contributing significantly to a team's performance. Similarly, possession statistics would provide insight into a team's dominance during games, while data on injuries and injury-prone players could highlight vulnerabilities.

This project focused solely on match performance but did not account for external factors that are equally important. It is essential to acknowledge that these elements were knowingly omitted in the current version of the

project due to data constraints. Future iterations should aim to incorporate these factors for a more comprehensive analysis. The notable areas for improvement include:

⇒ **Mental state of players and managers**: This could influence performance but remains unquantifiable.

⇒ **Possession statistics**: Indicates a team's overall dominance during matches.

⇒ **Injury data**: Tracks how injury-prone key players are and the impact of injuries on performance.

⇒ **Political situations**: Circumstances that may affect a team's focus or resources.

⇒ **"Man of the Match" and "Player of the Month" awards**: Recognizes exceptional player contributions throughout a season.

⇒ **"Manager of the Month" awards**: Reflects a manager's influence on the team's success.

⇒ **Contract incentives**: Rewards for good performance that could motivate players.

⇒ **Participation in other competitions**: Involvement in multiple tournaments can lead to player fatigue.

⇒ **Betting odds**: only the first set of betting odds found on each excel file was taken.

4. Results

To test the model, four seasons were analyzed to evaluate its ability to determine which teams would maintain their form. Table 1 shows the skills scores obtained from the analysis when focused specifically on the special teams of these seasons. On the other hand, table 2 shows the skills scores obtained from the analysis when taking all the teams in the league into consideration. For both cases a comparison between the model and climatology is shown. The seasons selected were 2015-2016, due to its significance as the season when Leicester famously won, as mentioned in the introduction, and the seasons 2021-2022, 2022-2023, and 2023-2024.

From the results, it can be observed that for both overall team predictions and special team cases, the model outperformed climatology in predicting form maintenance. In the 2015-2016 climatology forecasted a "100%" for Man City maintaining its form, which was punished heavily when they very unexpectedly dropped their form. On all other seasons, both approaches had similar results, with positive scores, highlighting good predictability. The same can be noted from the special teams' chart, where the IGN scores for both was positive. Where the overconfidence of my model was rewarded (which will not always be the case). However, it must be heavily noted that neither approach was effective in predicting when a team would drop form, and absolutely all the predictions obtained where above 0.5. Meaning, it was always predicted that teams would maintain form, with only variations on the certainty of it.

Table 1: IGN Score based on predictions for special teams

		Model		Climatology	
Season	Team	Probability	IGN Score	Probability	IGN Score
2015-2016	Leicester	0.79	16.5	0.72	13.15
2015-2016	West Ham	0.72	13.15	0.76	15.1
2015-2016	Crystal Palace	0.72	-20.91	0.72	-20.91
2021-2022	West Ham	0.71	-19.65	0.71	-19.65
2022-2023	Arsenal	0.93	22.38	0.91	21.6
2022-2023	Newcastle	0.91	21.6	0.86	19.56
2023-2024	Aston Villa	0.91	21.6	0.82	17.84
		Total	54.67	Total	46.69

Table 2: IGN Score based on predictions for all teams

Season	Model IGN Score	Climatology IGN Score
2015-2016	138.19	-108.59
2021-2022	199.18	211.63
2022-2023	139.12	103.74
2023-2024	211.71	223.19
Total	688.2	429.97

5. Conclusion

From the results and the IGN scores, it is evident that the model holds value, particularly because it outperformed climatology in 4 out of 5 cases, with higher probabilities of a special team maintaining its form. This demonstrates that the model effectively highlights the limitations of relying solely on bookmakers' odds when predicting whether teams will maintain their form. To make accurate predictions, one must consider not only the team's current standing in the league table but also its performance, as was done in this case using expected goals. This serves as an important foundational step for further developing the project, as it suggests that incorporating additional parameters to calculate expected points from performance metrics, such as possession, number of touches, and dribbles, on a game-by-game basis could yield better results than climatology.

At the same time, it is worth noting that the project is not yet a fully reliable predictive tool, as it consistently outputs positive values. This means the model always predicts that a team will maintain its form, although with varying levels of certainty. This bias is likely because most special teams achieve a high number of points, and when compared to historical top teams who reached similar positions using k-NN, the model tends to produce disproportionately high probabilities.

To address these limitations in the continued development of this project, one promising approach would be to implement an alternative classifier to better evaluate the probabilities of a team maintaining its form. Additionally, it is crucial for the model to progressively integrate more parameters that are significant for understanding a team's progression. Parameters that have been neglected up to this point.

In conclusion, this project provides a valuable stepping stone, showing that bookmakers' odds alone are not the most reliable indicator for betting on whether a team will maintain its form at mid-season. A model that incorporates a broader range of parameters to calculate expected points based on performance metrics would not only provide a more accurate forecasting methodology but also offer insights into the parameters that most influence the maintenance of form. These insights could be obtained by tuning the weights assigned to each parameter in the model, such as the impact of injuries, players' mental states, and other in-match dominance factors.

Finally, considering this conclusion, it is also worth mentioning that the forecasting for Leicester's winning season, predicting a 0.004% chance of success at the beginning of the season, was a poor forecast. This likely resulted from relying heavily on previous bookmakers' odds and league standings, without adequately considering the arrival of new players and a new promising manager. More thorough scouting into the individual characteristics of the new additions to the team could have provided better insight and flagged Leicester as a team to watch during that season.

6. Prediction Ongoing Season (2024-2025)

As a small addition to the project, a prediction of the teams maintaining their form in the current ongoing season which is about to enter the Christmas holidays will be provided by setting a small variation in the code. Here we can observe that the teams highlighted in green are performing better than expected and are predicted to maintain their form.

Position	Team	Prob
1	Liverpool	0.88
2	Chelsea	0.91
3	Arsenal	0.91
4	Man City	0.89
5	Nott'm Forest	0.71
6	Aston Villa	0.71

Figure 1. Prediction table for ongoing season