Sports Analytics for Enhanced Performance Prediction: Premier League

Authors: Akshay Krishnan, Jaya Rishita Pasam

Github: https://github.com/akrishnan96/Sports-Analytics-Performance-Prediction

1. Summary

The Premier League, England's top-tier football league, is renowned globally for its high level of competition and entertainment value. With millions of fans worldwide, the league's influence extends beyond the pitch, especially into the realm of Fantasy Premier League (FPL). In FPL, participants create virtual teams of real-life Premier League players and score points based on those players' actual performances in matches. The ability to accurately predict player performance in the Premier League is crucial for success in FPL. It not only enhances the gaming experience for fantasy league enthusiasts but also provides valuable insights for sports analysts and fans, aiding in understanding player potential and team dynamics.

Given this context, our project is dedicated to forecasting Premier League football player performances for each game week, using fantasy points as a measure of their on-field impact. Our goal is to leverage the increasing interest in sports analytics in football, benefiting fantasy football players, sports analysts, and enthusiasts alike. The project involves assembling a comprehensive dataset, including detailed game week and player statistics from the Premier League, to offer a thorough perspective on player performance.

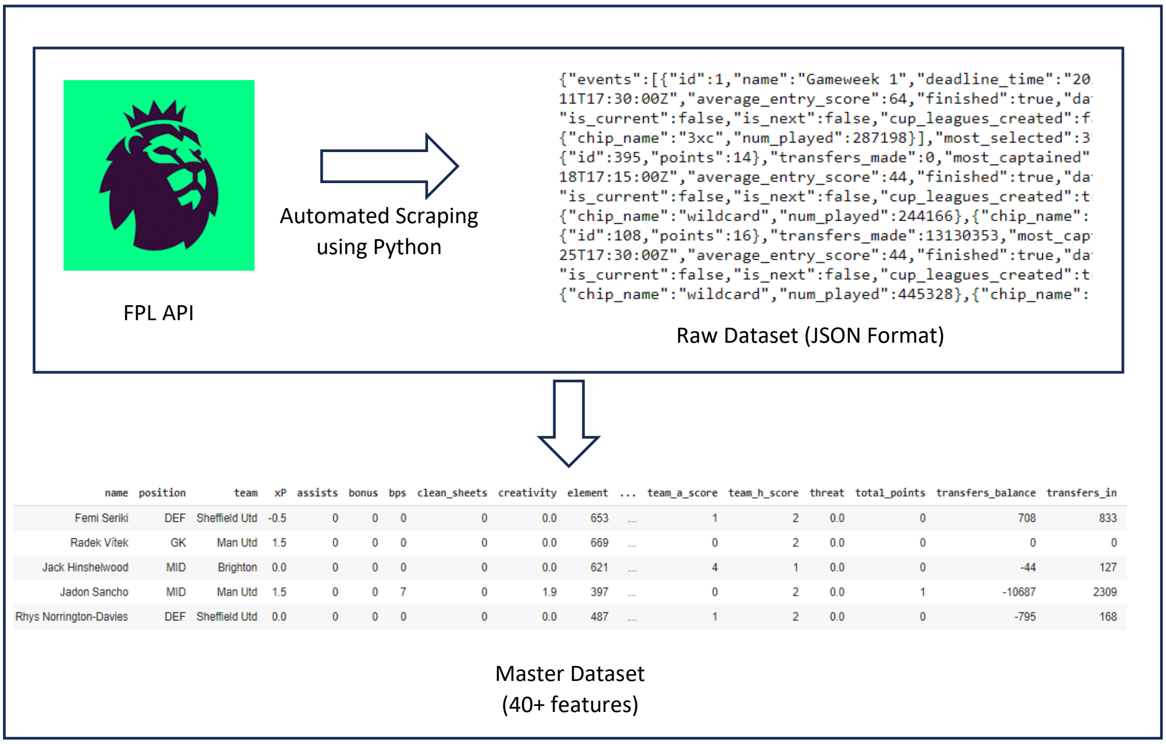
Our methodology encompasses meticulous data extraction, exploratory data analysis (EDA), inventive feature engineering followed by Model Building and Validation. The EDA is designed to reveal hidden patterns and trends within Premier League data, while feature engineering aims to create new metrics that more accurately reflect player and team performance dynamics. To enhance our predictions, we have implemented six different types of machine learning models. These models include Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, and Long Short-Term Memory (LSTM). The performance of each model is assessed using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 score.

Our findings indicate that the Random Forest model with Cross Validation has the highest R2 score, suggesting superior predictive accuracy. Conversely, the Support Vector Machine model shows the lowest MAE, indicating greater precision in its predictions. These results will not only aid in improving player performance predictions in fantasy leagues but also contribute to wider sports analytics discussions, potentially influencing player selection strategies and match analyses in football.

2. Methods

2.1. Data Extraction

The process of data extraction was multifaceted, involving both official APIs and web scraping techniques. For Premier League, we utilized the Fantasy Premier League (FPL) API, which provided extensive data on player and team performance metrics for each game week. This included details such as goals, assists, clean sheets, and more. The integration and harmonization of data from these sources was a critical step. We ensured that data for the different seasons were structured coherently, maintaining uniformity in data formats and measurement units. This process was crucial to facilitate seamless analysis and comparison across different data types, laying the groundwork for robust and comprehensive exploratory data analysis. **Fig 1** below shows the initial methodology of extracting football related information from the Fantasy Premier League API



```
{"events":[{"id":1,"name":"Gameweek 1","deadline_time":"20
11T17:30:00Z","average_entry_score":64,"finished":true,"da
"is_current":false,"is_next":false,"cup_leagues_created":f
{"chip_name":"3xc","num_played":287198}],"most_selected":3
{"id":395,"points":14},"transfers_made":0,"most_captained"
18T17:15:00Z","average_entry_score":44,"finished":true,"da
"is_current":false,"is_next":false,"cup_leagues_created":t
{"chip_name":"wildcard","num_played":244166},{"chip_name":
{"id":108,"points":16},"transfers_made":13130353,"most_cap
25T17:30:00Z","average_entry_score":44,"finished":true,"da
"is_current":false,"is_next":false,"cup_leagues_created":t
{"chip_name":"wildcard","num_played":445328},{"chip_name":
```

name	position	team	xP	assists	bonus	bps	clean_sheets	creativity	element	...	team_a_score	team_h_score	threat	total_points	transfers_balance	transfers_in
Femi Seriki	DEF	Sheffield Utd	-0.5	0	0	0	0	0.0	653	...	1	2	0.0	0	708	833
Radek Vítek	GK	Man Utd	1.5	0	0	0	0	0.0	669	...	0	2	0.0	0	0	0
Jack Hinshelwood	MID	Brighton	0.0	0	0	0	0	0.0	621	...	4	1	0.0	0	-44	127
Jadon Sancho	MID	Man Utd	1.5	0	0	7	0	1.9	397	...	0	2	0.0	1	-10687	2309
Rhys Norrington-Davies	DEF	Sheffield Utd	0.0	0	0	0	0	0.0	487	...	1	2	0.0	0	-795	168

Fig 1: Methodology to extract PL API

2.2. Dataset

The master dataset used in our project comprises a substantial 84,116 rows and 35 columns, offering a detailed view of Premier League football player performances from the 2020-21 season to the 2023-24 season. Each row represents a unique player entry for a particular game week, encapsulating various aspects of their performance. The columns cover a wide range of metrics, including basic player information like name and position, as well as specific game-related statistics such as goals scored, assists, clean sheets, and total points earned. Other notable columns include 'creativity', 'influence', and 'threat', which are advanced metrics used in fantasy football to gauge a player's impact on the game. Additionally, the dataset contains information on the player's team, opponent team, and whether the match was played at home or away. This rich dataset serves as the foundation for our machine learning models, enabling us to predict player performances with greater accuracy.

2.3. Feature Engineering

2.3.1. Team Performance Aggregation

- This aggregation includes calculating the mean **threat**, **points earned** and sum of **total points** for each team per game week and season
- The aggregation is sorted by season, team, and game week, providing a structured view of each team's performance across seasons
- This team-level data can provide insights into overall team performance, effectiveness against specific opponents, and how team dynamics might impact individual player performances

2.3.2. Time Feature Engineering

- The **year, month, day, and day of the week** is extracted from the kick-off time offering a detailed temporal context for each game
- It also categorizes the time of the day into morning, afternoon, evening, or night based on the hour of the kick-off
- This categorization could reveal trends related to performance at different times of the day, potentially influenced by factors like player energy levels or audience presence

2.3.3. Rolling Averages for Player Stats

- Rolling averages are calculated for various player statistics (assists, goals scored, clean sheets, etc.) over window sizes of 1-5 games.
- These averages, along with their lagged versions, are intended to capture the player's form and consistency over time, smoothing out anomalies and highlighting trends in performance
- Such features are vital for predictive models as they incorporate a temporal dimension, reflecting how a player's recent performances might influence their upcoming game

2.3.4. Game Status and Points

- **team_score** and **opponent_score**: These features represent the goals scored by the player's team and the opposing team, respectively. The values are determined by whether the player's team was playing at home or away
- **game_status**: This variable indicates whether a player's team won, drew, or lost a match. It is a categorical feature derived from comparing team_score and opponent_score.
- **points**: Based on game_status, this feature assigns points (3 for a win, 1 for a draw, 0 for a loss). It quantifies the match outcome in terms of league points, offering a direct measure of success

2.3.5. Opponent Team Performance Aggregation

- Metrics like the opponent team's average **threat**, **points** and **total points** would offer a view of the opponent's capabilities and form. Analysing these metrics can help in understanding the level of challenge faced by the player's team in each match
- This aggregation would be particularly useful in modelling to predict player performance, as the strength and form of the opponent team are significant factors in determining a player's likelihood of success in a game

The dataset is enriched with a variety of features that facilitate a more nuanced analysis and enable accurate predictions in sports analytics. The aggregation of team performance metrics, coupled with the opponent team's performance data, adds significant context. This contextual information is pivotal in enhancing the predictive models' accuracy, as it considers not just a player's team form but also the strengths and weaknesses of their opponents.

Moreover, the incorporation of time feature engineering introduces temporal dimensions, allowing for the analysis of trends over time. Rolling averages and game outcome-related variables, such as game status and points, capture ongoing performance trends and the outcomes of matches. Such comprehensive data can inform coaching strategies, aid in player selection, and help in predicting the competitiveness of upcoming matches. By comparing team and opponent metrics, key matchups that could be decisive in the outcome of games are identified, showcasing the dataset's capacity to support sophisticated predictive models in sports analytics.

2.4. Modelling

After feature engineering we had close to 164 features excluding the predictor variable. In this case, our predictor variable would be the points earned by a player per game week so the machine learning model should correctly predict how many points a player might get in a particular game week. We used the following machine learning models as part of the model training process - Linear Regression, Decision Trees, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) neural networks. Each model offers unique strengths: Linear Regression provides a baseline, Decision Trees capture non-linear relationships, Random Forest and Gradient Boosting offer ensemble learning benefits, SVM handles high-dimensional space well, and LSTM is adept at capturing time-series dependencies. The LSTM model is implemented with a specific architecture, including 50 LSTM units and a dropout layer to prevent overfitting. The model is compiled with the Adam optimizer and mean squared error as the loss function, indicative of a focus on minimizing prediction errors. For each model, appropriate data preprocessing steps are taken. For instance, the LSTM model requires reshaping the input data to fit its three-dimensional input structure. Feature scaling is also performed, which is critical for models like SVM and LSTM to function effectively

Cross-validation is a vital step in ensuring the generalizability of the models. It involves splitting the dataset into multiple subsets and training and testing the model on these different subsets. This process helps in assessing the model’s performance across various data samples, reducing the likelihood of overfitting, and ensuring that the model can perform consistently across different sets of data

Model validation is a key aspect of the training process. The LSTM model's performance is evaluated using metrics such as loss, mean squared error (MSE), and mean absolute error (MAE). These metrics provide insights into how well the model is predicting player performances, with a lower MSE and MAE indicating higher accuracy, and ensuring that the model can perform consistently across different sets of data. The results from the LSTM model, along with those from other models, are likely compared and analysed to determine the most effective approach for predicting player performance. This comparative analysis helps in identifying the strengths and weaknesses of each model, guiding the selection of the best model or a combination of models for the final deployment. Overall, the combination of multiple modelling techniques, coupled with rigorous cross-validation and validation processes, exemplifies a comprehensive and robust approach to predictive modelling in sports analytics. The diversity of models ensures a broad exploration of the data's patterns, while validation steps guarantee the models' reliability and applicability to real-world scenarios.

3. Results

3.1. Exploratory Data Analysis

3.1.1. Analysis of Points over Time

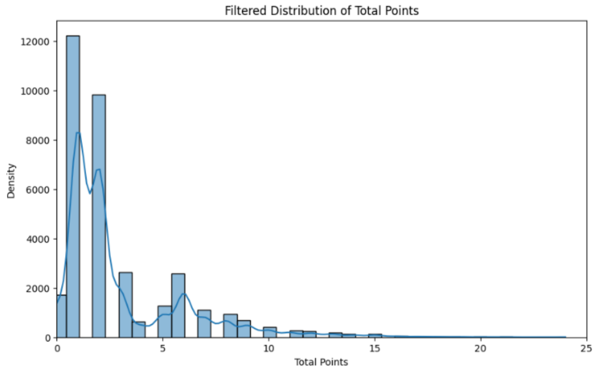


Figure 2a: Distribution of Points over the last 4 seasons

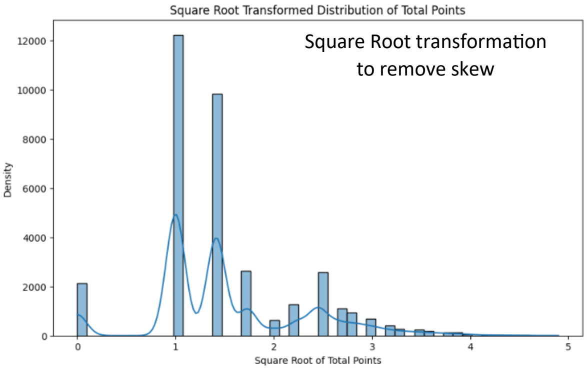


Figure 2b: Square Root Transformed Distribution

Figure 2a represents the distribution of total points for players who have played more than 0 minutes in a game week. The distribution is right-skewed with a high frequency of players scoring between 0 and 5 points. The peak near 0 suggests that many players score very few points, even when they do get playing time. This could be due to various reasons such as playing in a more defensive role, not contributing directly to goals, or getting booked with yellow/red cards. **Figure 2b** shows the square root transformation of points. This has reduced the right skewness of the original data and the distribution appears more symmetrical around the lower values of square root total points, although it still exhibits some skewness. The distribution shows multiple peaks, which could indicate that there are common point totals that players tend to achieve, possibly corresponding to common events in a match like scoring a single goal, assisting, or keeping a clean sheet.

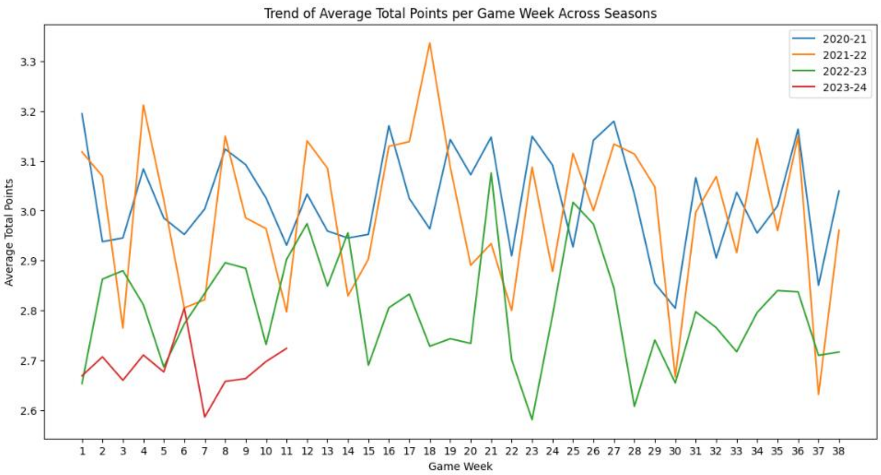


Figure 3: Points earned per player by GW over the last 4 seasons

Figure 3 shows significant week-to-week fluctuations in the average total points for all seasons, indicating variability in player performances. These fluctuations could be the result of various factors, including fixture difficulty, player injuries, tactical changes, or simply the inherent unpredictability of football matches. Each season exhibits its own set of intra-season trends, with peaks and troughs indicating periods of high and low average points. When comparing the seasons, it is evident that the overall pattern of points fluctuation varies from season to season. Some seasons start with high averages that taper off, while others show more stability before experiencing significant changes. This comparative analysis can reveal insights into the evolving nature of the league and team strategies.

3.2. Model Evaluation

The results from the various machine learning models as shown below in Table 1, demonstrate a range of performance, with significant improvements observed in models where cross-validation (CV) was applied. For instance, the Random Forest with CV, Gradient Boosting with CV, and SVM with CV models showed notably better metrics in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 values compared to their counterparts without CV. This improvement underscores the efficacy of cross-validation in enhancing model accuracy and reducing overfitting, ensuring that the model is robust across various data samples.

Additionally, the approach of using only the top 40-50 features for model training played a crucial role in these results. By focusing on the most significant features, the models were able to concentrate on the most impactful variables, leading to more precise and efficient predictions. This feature selection process likely contributed to the improved performance of the models, particularly in the cross-validated versions, by eliminating redundant or less informative variables.

Model	MSE	MAE	R2
Linear Regression	4.3248	1.16507	0.27573
Decision Tree	8.74208	1.44586	-0.46404
Random Forest	4.14154	1.10872	0.30641
Random Forest with CV	**3.95342**	**1.05123**	**0.35095**
Gradient Boosting	4.21589	1.10925	0.29396
Gradient Boosting with CV	**3.98323**	**1.0839**	**0.32233**
Support Vector Machine	4.67762	0.98402	0.21664
SVM with CV	**4.45334**	**0.95123**	**0.25336**
LSTM	4.47294	1.14303	N/A

Table 1: Model Evaluation results

4. Discussion

The results of this project reveal several key insights and implications for predictive modelling in sports analytics. Firstly, the diversity of machine learning models used, including Linear Regression, Decision Trees, Random Forest, Gradient Boosting, SVM, and LSTM, highlights the complexity of predicting player performance in football. Each model brought unique strengths to the table, with the LSTM model being particularly notable for its ability to capture time-series dependencies, a crucial aspect in sports performance analysis. The implementation of these models with specific architectures and preprocessing steps, such as reshaping data for LSTM and feature scaling, was critical in optimizing their performance.

One of the most significant findings was the effectiveness of cross-validation in improving model accuracy. Models that underwent cross-validation, such as Random Forest with CV, Gradient Boosting with CV, and SVM with CV, showed marked improvements in performance metrics (MSE, MAE, R2). This underscores the importance of cross-validation as a technique to ensure the generalizability and robustness of predictive models across different data samples, thereby reducing the likelihood of overfitting.

Moreover, the decision to focus on the top 40-50 features for model training proved to be pivotal. This approach of feature selection ensured that the models concentrated on the most impactful variables, leading to more precise and efficient predictions. The feature selection process, especially in the context of cross-validated models, helped eliminate less informative variables, thereby streamlining the models and enhancing their performance.

Reflecting on this phase of the project, one notable decision was to opt out of incorporating sentiment analysis from Twitter. This decision was made due to the complexities and challenges involved in analysing data at both the season and game week levels, combined with the large volume of data required and constraints imposed by Twitter. In future iterations of the project, exploring alternative methods or sources of data for sentiment analysis could be considered to enrich the predictive models further. Additionally, continuing to refine the feature selection process and exploring different combinations or architectures of models could further enhance the project's outcomes. The learnings from this phase lay a solid foundation for future explorations in sports analytics, particularly in the realm of player performance prediction.

5. Statement of Contributions

1. Project Conceptualization:
    - Akshay Krishnan – Initial conceptualization of the project, outlining the primary objectives and scope
    - Jaya Rishita Pasam – Refining the project goals and establishing key research questions

2. Data Collection:
    - Akshay Krishnan – Data Extraction from the Fantasy Premier League API, Twitter API, Web scraping

3. Exploratory Data Analysis (EDA):

- Jaya Rishita Pasam – EDA focusing on identifying key variables and uncovering patterns in the Premier League data, visualizations and interpreting the findings from the EDA

4. Feature Engineering:
    - Akshay Krishnan – Developing new features from the raw data, focusing on creating metrics that could enhance the predictive model, provided support by evaluating the relevance and impact of these new features

5. Model Development and Testing:
    - Akshay Krishnan – Building and testing various predictive models, selecting appropriate algorithms and validation techniques

6. Data Visualization and Reporting:
    - Jaya Rishita Pasam – Comprehensive data visualizations and reports to communicate the project findings

7. Project Documentation and GitHub Management:
    - Jaya Rishita Pasam – Project documentation and reviews
    - Akshay Krishnan – Maintaining the GitHub repository and codes

8. Final Presentation and Review:
    - Akshay Krishnan and Jaya Rishita Pasam – Preparation of the final presentation of the project, consolidating all findings and insights and future scope

6. References

1) Premier League. (2023). Premier League Data. Retrieved from https://www.premierleague.com/stats

2) Fantasy Premier League. (2023). Premier League Data. Retrieved from https://fantasy.premierleague.com/api/bootstrap-static/

3) R. Pariath, S. Shah, A. Surve and J. Mittal, "Player Performance Prediction in Football Game," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 1148-1153, doi: 10.1109/ICECA.2018.8474750

4) S. Manish., V. Bhagat and R. Pramila, "Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp. 1-5, doi: 10.1109/INCET51464.2021.9456424

5) Stoppels, Eloy. Predicting race results using artificial neural networks. MS thesis. University of Twente, 2017.

6) Accessing Fantasy Premier League data using Python. Retrieved from https://medium.com/analytics-vidhya/getting-started-with-fantasy-premier-league-data-56d3b9be8c32

7. Appendix

Analysis of Points by Position

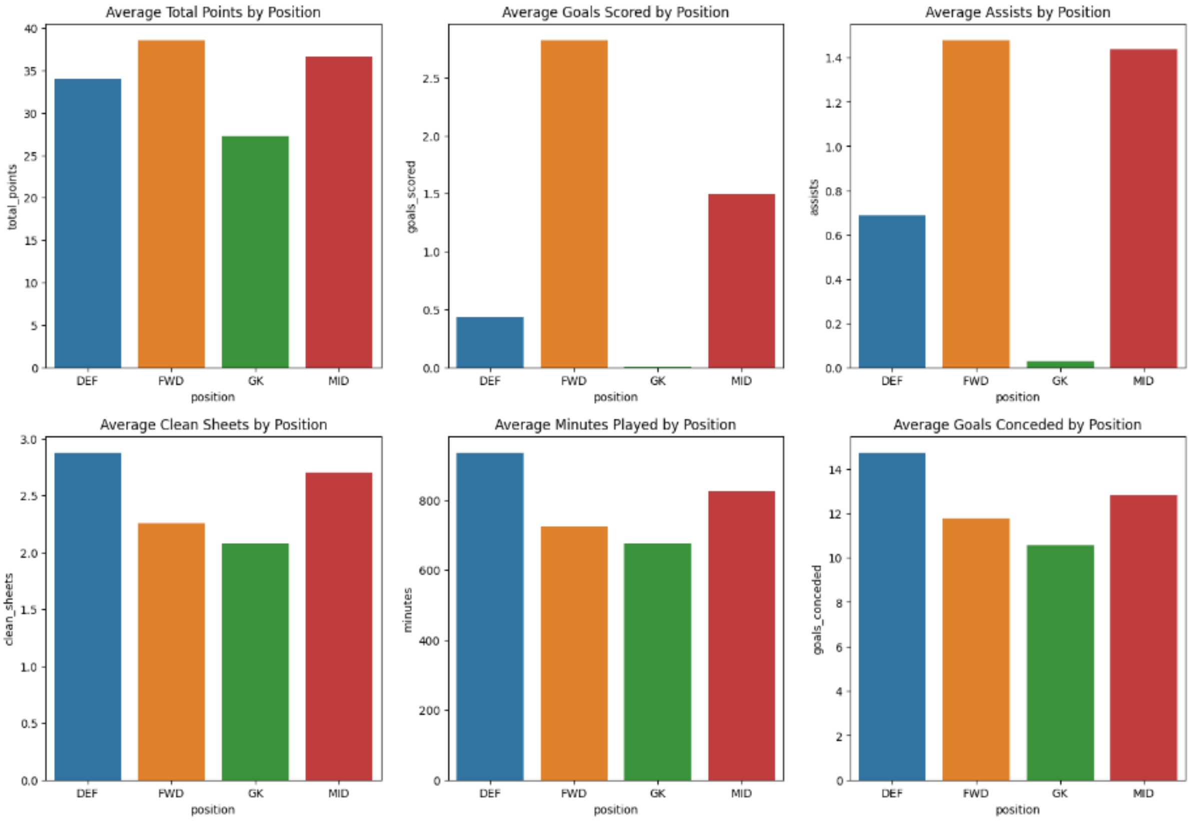


Figure 4: Average performance metrics per position

From **Figure 4**, we see that Midfielders (MID) seem to score the highest average total points, followed closely by forwards (FWD), indicating that these positions contribute significantly to the team's success in ways that are captured by the points system (e.g., goals, assists). Midfielders (MID) provide the most assists, which is typical given their role in creating scoring opportunities. This chart also suggests that midfielders (MID) and defenders (DEF) play the most minutes on average, indicating their crucial role in the team's core structure and gameplay over the duration of matches. From this graph we see that each position has distinct responsibilities and contributions to a team's performance which in turn affects player performance as well.

Analysis of Player Consistency

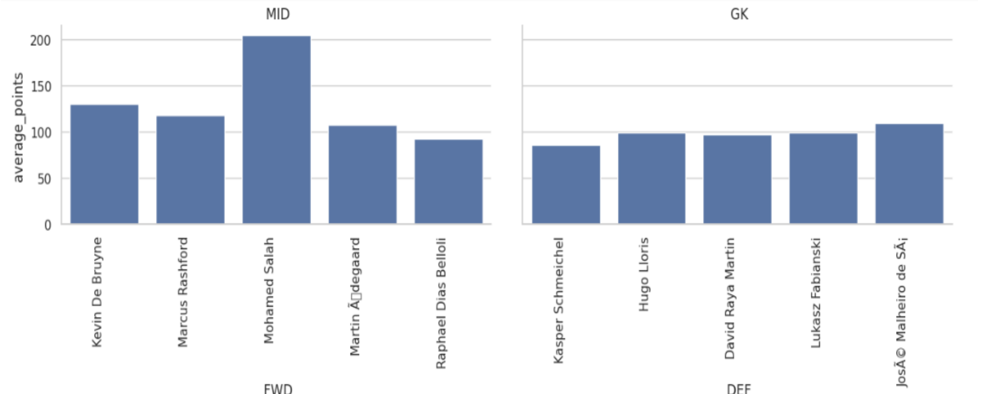


Figure 5: Player-Level Consistency by Position for the latest 4 seasons

From **Figure 5**, we see that high-scoring midfielders like Mohamed Salah are typically involved in both goals and assists, contributing heavily to their points total. We also see that forwards such as Harry Kane have high average points due to their primary role in scoring goals, which usually carries the highest points reward in fantasy leagues. The goalkeepers' average points are often lower than outfield players but can be boosted by clean sheets and saves. Therefore, we can conclude that the value of each position can be assessed in terms of their points contribution. Midfielders and forwards often have higher average points due to their attacking roles. Also, within each position, players can be compared to identify standouts and those who may be underperforming relative to their peers. We also see that if the data is averaged over multiple seasons, players with high average points and low variance would be considered consistent performers.