Fantasy Football Prediction System

Deban Kumar Shahi, Piyush Priy, Sanket Adlak, Priyank Nagarnaik

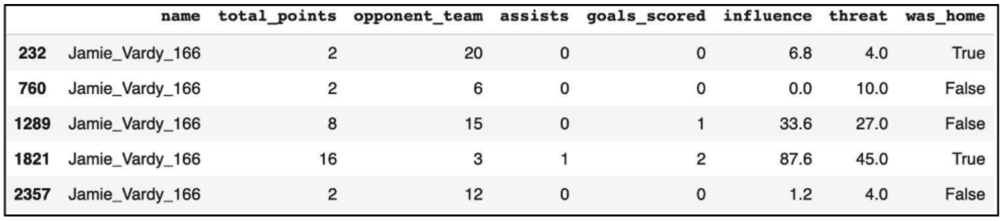
1 Data Sources and Relevance

Data Sourcing

The official FPL API provides access to a wealth of data on individual players, teams, and fixtures. Although many statistics are updated weekly, historical data from previous gameweeks is overwritten, preventing the construction of detailed time series profiles for players. The Varstaav FPL repository addresses this issue by logging FPL API data each week into conveniently formatted files. Data from all available seasons (2019/20 to 2024/25) was used. The primary sources were:

- Merged gameweek files: Containing a game-by-game breakdown of player performance.
- Raw player data files: Providing essential details such as position, team, and FPL price.

Together, these files supply the comprehensive information necessary to build the prediction system. The Pandas library in Python 3 was used to read and process the repository files.



	name	total_points	opponent_team	assists	goals_scored	influence	threat	was_home
232	Jamie_Vardy_166	2	20	0	0	6.8	4.0	True
760	Jamie_Vardy_166	2	6	0	0	0.0	10.0	False
1289	Jamie_Vardy_166	8	15	0	1	33.6	27.0	False
1821	Jamie_Vardy_166	16	3	1	2	87.6	45.0	True
2357	Jamie_Vardy_166	2	12	0	0	1.2	4.0	False

Figure 1: Subset of Statistics Available for Leicester Player Jamie Vardy

Figure 1 illustrates a selection of features available for each gameweek. Metrics such as influence and threat serve as metadata that offer deeper insight into a player's performance—quantifying aspects that traditional statistics might overlook. For instance, while conventional metrics might focus solely on goals scored, the threat metric measures the danger posed to the opposition's goal.

Data Relevance

The Fantasy Premier League dataset offers a comprehensive platform for statistically-driven fantasy football performance prediction. Its granular player-level data, spanning over 20 performance metrics, enables robust modeling of offensive, defensive, and advanced player contributions. The longitudinal structure facilitates temporal analysis, controlling for contextual factors, while the standardized, objective nature of the variables aligns with the goal of minimizing subjective bias in team selection. This dataset provides an ideal foundation for developing purely statistical prediction models that overcome human cognitive biases and preference-based distortions.

2 Data Exploration and Analysis

Data Preparation

In our approach, we split player data by field position (e.g., quarterbacks, wingers, running backs, etc.), so that the models we trained and tested on these segmented datasets can be more accurate due to the distinct play styles inherent to each position.

We experimented with different presentations of previous performance statistics. We used data from n previous games as features to predict performance in a given gameweek. Varying the value of n produced different levels of model accuracy, and through trial and error, we determined an optimal value. The context in which the data is presented is also critical. For example, consider a player (P1) who scores three goals against the worst team in the league versus another player (P2) who scores two goals against the best team. Although raw numbers suggest P1 performed better, we believe P2's achievement against a stronger opponent should be weighted more favorably. Additionally, we addressed outlier statistics; for instance, when measuring goals per minute, a substitute who plays only the last five minutes and scores will show an inflated metric—not due to high scoring frequency, but because of limited playing time.

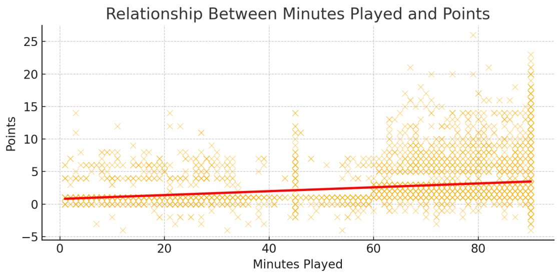


Figure 2: Scatter Plot of Minutes vs Points

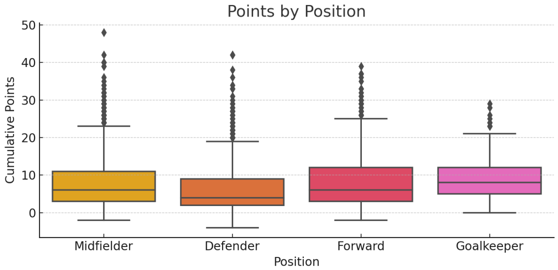


Figure 3: Boxplot comparing cumulative points across different player positions

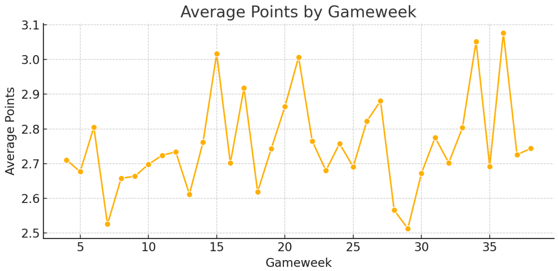


Figure 4: Average points change over different gameweeks

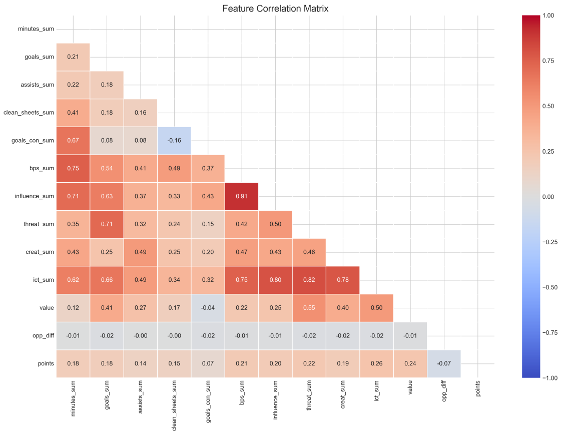


Figure 5: Relationships between features and points

Exploratory Data Analysis

The exploratory data analysis carried out here is essential to understand statistical trends underpinning fantasy football performance and provide an empirical basis for further predictive modeling. Correlation analysis indicated that offensive statistics (goals_sum, assists_sum) and engagement statistics (minutes_sum, influence_sum) had the highest correlations with fantasy points, and the multicollinearity between some of the predictor variables (notably between threat_sum, creat_sum, and ict_sum) made our use of regularized regression methods justifiable. Home bias was statistically significant which justified its use as a predictive factor, whereas difficulty of opponents had a weak negative correlation with performance, which substantiated our difficulty rating model based on league rank. Value efficiency analysis revealed significant variability across positions, with defenders providing better points-per-cost returns than high-priced forwards, which lends empirical evidence to our optimization algorithm's budget split approach. Together, these results support our methodological choices, from feature design to position-specific modeling and team optimization, as well as our primary research question by setting objective statistical links that reduce subjective human choice in fantasy team picks.

3 References

1. *FPL Data source:* `https://fantasy.premierleague.com/api/bootstrap-static/`
2. *Github Repository:* `https://github.com/vaastav/Fantasy-Premier-League`