Optimizing Fantasy Premier League Strategies: A Machine Learning Approach

|  |  |
| --- | --- |
| Naishal Shah | Yashwi Passary |
| Luddy School of Informatics | Luddy School of Informatics |
| Indiana University, Bloomington | Indiana University, Bloomington |
| [shahnai@iu.edu](mailto:shahnai@iu.edu) | [ypassary@iu.edu](mailto:ypassary@iu.edu) |

1. **Abstract**

Fantasy Premier League (FPL) is a globally popular fantasy sports game where participants create virtual teams of English Premier League players, earning points based on real-world player performance. Success in FPL requires strategic decisions, such as player selection, transfers, and captaincy, while considering factors like form, fixture difficulty, and budget constraints. This project leverages machine learning models, including Random Forest, Gradient Bossting and Reinforcement Learning to optimize FPL decision-making. Our approach focuses on three key areas: captaincy selection, dynamic fixture difficulty ratings (FDR), and team optimization strategies. By analyzing historical and real-time data, we offer actionable insights to predict player performance, maximize points, and guide FPL managers in navigating the game's complexities.

1. **Introduction**

Fantasy Premier League (FPL) is the official fantasy football game of the English Premier League, captivating over 15 million players globally. This free-to-play game enables participants to take on the role of virtual football managers, assembling squads of 15 players within a £100m budget. Players earn points based on their real-life performances, with metrics including goals, assists, clean sheets, and penalties, while deductions occur for actions such as yellow cards, missed penalties, or own goals.

The game operates under specific rules: a squad must consist of 2 goalkeepers, 5 defenders, 5 midfielders, and 3 forwards, with no more than 3 players from a single team. Managers select 11 players to start each game week, adhering to positional requirements, and can make one free transfer per week, with additional transfers incurring a 4-point deduction. Managers also designate a captain whose points are doubled, and a vice-captain as a backup. Strategic decisions regarding transfers, captaincy, and chip usage are critical to maximizing weekly points.

Success in FPL relies on analytical decision-making, as the game evolves weekly with dynamic changes such as player injuries, form fluctuations, and fixture difficulty. This study leverages machine learning techniques to provide actionable insights, aiming to simplify complex decisions and empower managers to optimize their performance.

FPL’s ever-changing nature demands constant recalibration, as match schedules, player injuries, and form fluctuations can significantly impact decisions. Unlike traditional sports, FPL intertwines elements of skill, strategy, and chance, creating a unique challenge for managers. Success often hinges on the ability to interpret data and make calculated decisions. This correlation between fantasy games and their real-world counterparts reveals the shared complexities of chance and expertise, making FPL not only a game of entertainment but also a test of analytical prowess. For those without a knack for numbers, the journey can be arduous, as strategic planning and data interpretation are crucial to staying ahead in this fast-evolving competition.

1. **Related works**

The predictive modeling and feature engineering methodologies in this work are inspired by various studies and practices:

1. **Predictive Analytics in Sports:** Research from IEEE (source: https://ieeexplore.ieee.org/document/9909447) demonstrates the use of machine learning models to predict sports outcomes, emphasizing the importance of feature selection and dynamic rating adjustments. This study provided insights into modeling complex relationships in sports data.
2. **Fantasy Premier League Strategy:** An article on Medium (source: https://medium.com/@nedara\_98396/winning-with-data-my-approach-to-fantasy-premier-league) outlines a data-driven strategy for optimizing FPL performance, including captaincy selection and transfer decisions. The approach highlighted the importance of leveraging recent trends and opponent analysis, which parallels our use of rolling averages and streak metrics.
3. **Fixture Difficulty and Team Performance:** Studies from PMC (source: https://pmc.ncbi.nlm.nih.gov/articles/PMC7928501/) and ORiON (source: https://orion.journals.ac.za/pub/article/view/753) explore methods for ranking team performances and assessing fixture difficulties. These works influenced our integration of predicted goal differences and dynamic FDR levels.
4. **General Data Analytics in FPL:** Research from IJRASET (source: https://www.ijraset.com/research-paper/data-analytics-as-a-key-to-fantasy-premier-league) emphasizes the importance of data preparation and feature selection in building effective FPL models. This aligns with our preprocessing steps and feature engineering for predictive accuracy.

These references collectively informed our choice of machine learning models and guided the design of feature engineering strategies, ensuring a robust and adaptive methodology for FPL data analysis.

1. **Problem Statement**

Managing a successful Fantasy Premier League team requires decisions that balance player form, fixture difficulty, and budget constraints. Key challenges include predicting captaincy choices for maximum points, assessing fixture difficulty dynamically as team and player performances evolve, and assembling an optimal team under budget limitations. These complexities demand data-driven strategies to guide decisions and overcome the limitations of intuition or static heuristics. This project aims to address these challenges through machine learning models, enabling better captain selection, dynamic FDR evaluations, and actionable recommendations for team selection.

Fantasy Premier League challenges managers to navigate three critical aspects of team management:

1. **Captaincy Selection**: The captain earns double points, making the choice of player pivotal each game week. Predicting the best-performing captain requires accurate analysis of player form, fixture difficulty, and other performance indicators.
2. **Fixture Difficulty Rating (FDR)**: Evaluating the difficulty of upcoming matches for teams is essential for optimizing transfers and player lineups. This requires dynamic adjustments based on team and player performance trends.
3. **Wildcard Team Strategy**: The wildcard chip allows managers to rebuild their team without point deductions. Using this effectively requires selecting an optimized lineup within budget and team composition constraints.

A football field with different teams

Description automatically generated

**FPL Rules and Scoring System**

FPL’s rules and scoring system introduce unique layers of complexity. Key rules include:

* Managers have a £100m budget to select 15 players: 2 goalkeepers, 5 defenders, 5 midfielders, and 3 forwards.
* A maximum of 3 players can be selected from the same Premier League team.
* Points are awarded based on real-life player performance:

1. **Goalkeeper:** Earns 4 points for a clean sheet, 5 points for each penalty saved, 6 points per goal, 3 points per assist, and 1 point for every 3 saves. Loses 1 point for every 2 goals conceded.
2. **Defender**: Earns 4 points for a clean sheet, 6 points per goal, and 3 points per assist. Loses 1 point for every 2 goals conceded.
3. **Midfielder**: Earns 5 points per goal, 1 point for a clean sheet, and 3 points per assist.
4. **Forward**: Earns 4 points per goal and 3 points per assist.
5. **All Players**: Earn 1 point for playing less than 60 minutes and 2 points for playing 60 minutes or more. Lose 1 point for yellow cards, 3 points for red cards, and 2 points for own goals or missed penalties. Additionally, the top 3 players in each match receive 3-1 bonus points.

* Bonus points are awarded for three standout players in each match.

These rules ensure managers must balance offensive and defensive player contributions while maximizing their budget and adhering to constraints. This study aims to address these complexities through data-driven machine learning models, helping managers optimize key decisions and improve their overall rank in this globally competitive game.

1. **Model**
2. **Random Forest Regressor Model (Captaincy Selection)**

The **Random Forest Regressor** is an ensemble learning method that operates by constructing multiple decision trees during training and combining their predictions to produce a final output. For regression tasks, the model aggregates the predictions by averaging the outputs of all the decision trees, which helps reduce variance and improve generalization. Each tree in the forest is built using a random subset of the training data (bagging) and considers a random subset of features at each split, making the model robust to overfitting and noise in the data.

Random Forest is particularly effective for non-linear and high-dimensional data, as it can capture complex relationships between input features without requiring assumptions about their distribution. Furthermore, the model provides feature importance scores, which allow us to identify the most influential variables in predicting player performance. This interpretability, coupled with its ability to handle missing data and interactions between features, makes Random Forest a suitable choice for FPL captaincy prediction.

**Why Random Forest Approach?**

The **Random Forest Regressor** is an ideal choice for Captaincy Selection problem due to the following reasons:

1. **Handling Non-Linear Relationships**: Player performance in FPL often depends on complex and non-linear factors, such as recent form, fixture difficulty, and positional influence. Random Forest models excel at capturing these relationships without requiring the data to follow linear patterns.
2. **Feature Importance Analysis**: Random Forest naturally provides feature importance scores, helping to identify which variables (e.g., *bps*, *goals\_scored*, *minutes*, *form*) contribute most to predicting player performance. This insight can be used to refine the feature selection process and improve interpretability.
3. **Robustness to Missing Data**: FPL data often contains missing values (e.g., when a player does not feature in a specific gameweek). Random Forest is resilient to incomplete data and can effectively utilize available information for predictions.
4. **Ensemble Learning for Better Predictions**: Random Forest aggregates predictions from multiple decision trees, reducing the risk of overfitting and improving generalization. This ensemble approach ensures that the model delivers reliable predictions across different gameweeks.
5. **Reduced Overfitting**: By averaging the predictions of multiple trees and using randomized subsets of data and features, Random Forest mitigates overfitting, a critical factor when training on historical FPL data with limited observations.
6. **Scalability and Performance**: Random Forest scales well with high-dimensional datasets, making it suitable for FPL data, where multiple features influence player performance.

In this model, the Random Forest Regressor effectively handles the complexity of FPL performance prediction by leveraging historical and current data. Its robustness, accuracy, and ability to generalize across gameweeks make it a powerful tool for recommending optimal captain and vice-captain choices in FPL. This results in a data-driven and reliable strategy for enhancing decision-making for fantasy managers.

1. **XGBoost Model (Fixture Difficulty Rating)**

XGBoost (Extreme Gradient Boosting) is a powerful and efficient implementation of gradient boosting algorithms. It builds sequential decision trees in an additive manner, where each subsequent tree corrects the residual errors of the previous trees. It optimizes the learning process using techniques like gradient descent and regularization to improve both accuracy and generalizability.

Key features of XGBoost include:

1. Gradient Boosting Framework: It combines weak learners (individual trees) iteratively to minimize prediction errors.
2. Regularization: XGBoost applies both L1 (Lasso) and L2 (Ridge) regularization, preventing overfitting to the training data.
3. Tree Pruning: It reduces model complexity and improves computational efficiency by pruning underperforming branches.
4. Parallelization: XGBoost efficiently builds trees in parallel, reducing training time significantly for large datasets.

**Why XGBoost approach?**

XGBoost is particularly well-suited for the Fixture Difficulty Model due to the following reasons:

1. **Complex Relationships in Data:** Football data is inherently complex and non-linear. Team performance, streaks, and opponent difficulty interact in ways that traditional linear models cannot capture. XGBoost handles these non-linear interactions effectively using decision trees.
2. **Feature Engineering Compatibility:** The engineered features, such as rolling averages, streaks, and goal differences, introduce diverse numerical inputs. XGBoost excels at leveraging such structured, tabular data without requiring significant data scaling or transformation.
3. **Regularization for Generalization:** By applying L1 and L2 penalties, XGBoost prevents overfitting. This ensures that the model performs well on unseen data, such as the 2023-24 season test set, and generalizes to new matches accurately.
4. **Robust to Missing Data:** Football datasets often have missing or incomplete information (e.g., certain match results or player stats). XGBoost can natively handle missing values without requiring imputation.
5. **Efficiency and Speed:** XGBoost is optimized for speed and performance with parallelized tree building. This allows the model to train efficiently on multi-season data while delivering superior predictive accuracy.
6. **Feature Importance:** XGBoost provides feature importance metrics, enabling us to identify key factors influencing goal difference predictions. This transparency helps validate the significance of engineered features, such as **recent form** and **streaks**, in fixture difficulty.

The combination of XGBoost’s efficiency, regularization, and ability to model non-linear relationships makes it the best choice for predicting goal differences and dynamically calculating Fixture Difficulty Ratings (FDR). The model integrates rolling features, streak metrics, and opponent data to produce adaptive and reliable FDR values, ensuring that the ratings reflect real-world team performances.

1. **Experiments**
2. **Data Preprocessing and Cleaning:**

**Workflow Steps:**

1. **Data Loading:** Historical FPL data is loaded from CSV files covering multiple seasons, including detailed player statistics.
2. **Data Cleaning:** Missing values are handled, such as replacing null value entries with appropriate default values. Inconsistent or duplicate records are removed.
3. **Feature Engineering:**
   * *home\_advantage*: A binary feature indicating whether a match was played at home or away.
   * form: Computed as the rolling average of a player's points over the last 5 gameweeks, capturing recent performance trends.
   * Aggregated Features: Metrics such as total\_points, minutes\_played, and bps are aggregated for historical and recent gameweeks to analyze performance consistency.
4. **Data Transformation:** Historical and recent gameweek data is merged usingweighted aggregation (e.g., 70% weight for recent performance, 30% for historical data) to emphasize current trends in player form.
5. **Correlation Analysis:** Feature correlation with the target variable (total\_points) is computed, with a heatmap visualization identifying the strongest predictors.
6. **Feature Selection:** Key features like bps, ict\_index, threat, goals\_scored, and form are selected for modeling based on correlation analysis results.
7. **Model-Ready Dataset Preparation:** Data is split into training and testing sets, ensuring time-consistent splits to prevent data leakage (e.g., training on earlier seasons and testing on 2023-24 gameweeks).
8. **Visualization:** Exploratory graphs and heatmaps are generated to evaluate trends, distributions, and relationships between features.
9. **Data Export:** Processed datasets are saved as CSV files for downstream machine learning applications.

This preprocessing pipeline ensures that the data is cleaned, enriched, and formatted correctly, setting the foundation for building robust predictive models. This dataset is then read as input for both the models- Captaincy Selection and Fixture Difficulty Rating.

1. **Random Forest Regressor Model (Captaincy Selection)**

**Workflow Steps:**

1. **Correlation Analysis**: Correlation between features and the target variable (total\_points) is calculated to identify the most relevant predictors. A heatmap is generated to visualize the top 12 features strongly correlated with total\_points, ensuring that the most meaningful variables are retained.
2. **Feature Selection**: Based on the correlation analysis, a subset of features is selected for model training ( Top 12 features). These include bps, influence, bonus, ict\_index, goals\_scored, minutes, clean\_sheets, threat, and form.
3. **Data Preparation**: The prepare\_combined\_features function merges historical and recent gameweek data, applying a weighted approach to balance the influence of recent and older performance (70% weight for recent performance, 30% for historical data). The prepare\_gameweek\_data function organizes data into a suitable format for training and testing the model, ensuring that only relevant features are included.
4. **Model Training and Prediction**: The Random Forest Regressor is trained using historical FPL data (from the 2016-17 to 2022-23 seasons). For each gameweek in the 2023-24 season, the trained model predicts the total points for all players. Based on the predicted points, the top two players are selected as captain and vice-captain, respectively.
5. **Accuracy Calculation**: A custom calculate\_accuracy function is implemented to compare predicted points with actual points. The accuracy is calculated based on the closeness of predicted points to actual points, with additional bonuses awarded if the selected players rank among the top N performers.
6. **Exporting Results**: The predictions, including predicted\_points, actual\_points, and accuracy metrics, are stored in a DataFrame and exported to a CSV file for further analysis.
7. **XGBoost Model (Fixture Difficulty Rating)**

**Workflow Steps:**

1. **Feature Engineering**: Rolling averages and streak-based features are calculated to capture recent team and opponent performance:
   * + **Team Features**: Recent scoring form, conceded goals, home performance, winning streaks, and losing streaks.
     + **Opponent Features**: The same metrics are derived for opponent teams to account for fixture difficulty.

These rolling features are computed over specified windows (e.g., 5-7 games), ensuring that short-term trends are captured effectively.

1. **Model Training and Prediction**: The dataset is split into training (historical seasons) and testing (2023-24 season). The XGBoost Regressor is trained to predict goal difference using team and opponent features. Predictions from the model are stored and used as a key component in determining fixture difficulty.
2. **Dynamic FDR Assignment**: A weighted average combines the predicted goal difference (from the model) with opponent recent form, creating a robust measure of fixture difficulty. The weighted measure is categorized into FDR levels ranging from 1 (easy) to 5 (difficult). Historical FDR, based on win percentage and team ranking, is also factored in (weighted 25%).

The final results dynamically adjust the FDR for fixtures based on the latest trends and historical performance, making the model adaptive and accurate.

1. **Result Analysis**
2. **Captaincy Prediction Model**

**Custom Accuracy Metric**

To evaluate the performance of the captaincy prediction model, a **custom accuracy metric** was developed. Unlike conventional metrics that only measure overall error, this custom metric focuses on the **closeness** between the **predicted points** and **actual points** scored by the selected captain. This tailored evaluation better reflects the practical requirements of Fantasy Premier League (FPL), where even slight deviations can impact decision-making.

The metric works as follows:

1. **Perfect Accuracy (100%)**: If the actual points scored by the player are **greater than or equal to the predicted points**. Alternatively, if the difference between the predicted and actual points is within a **2-point margin** (e.g., predicted = 10, actual = 12 or 8).
2. **High Accuracy (70%)**: If the difference between predicted and actual points is within a **4-point margin** (e.g., predicted = 10, actual = 6 or 14).
3. **Moderate Accuracy (50%)**: If the difference is within a **6-point margin** (e.g., predicted = 10, actual = 4 or 16).
4. **Low Accuracy (30%)**: If the difference exceeds **6 points**, indicating significant deviation.

This approach rewards predictions that are close to or exceed the actual points, aligning well with the goal of identifying high-performing captains. The tiered structure ensures that small deviations do not overly penalize the model, while larger discrepancies are appropriately reflected with lower accuracy scores.

**Gameweek-Wise Analysis**

To validate the model, the custom accuracy metric was applied to the test data, starting from **Gameweek 8 (GW8)** of the 2023-24 season. The table below highlights the performance across selected Gameweeks:

A screenshot of a table

Description automatically generated

The analysis reveals that the model achieved a **remarkable mean accuracy of 77.42%** over 31 Gameweeks, demonstrating its reliability despite the inherent uncertainties in player performance. The model consistently identified high-performing captains, such as **Mohamed Salah (GW12)**, **Joško Gvardiol (GW37)**, and **Jean-Philippe Mateta (GW38)**, who delivered significant points. Additionally, the vice-captain selections provided a balanced and effective fallback option, ensuring robust performance even in cases where minor deviations occurred for the captain picks.

1. **Conclusion**

**Seasonal Performance Summary (Captaincy Selection Model)**

The Random Forest Regressor model successfully captured the underlying relationships between historical data, recent form, and player performance features. The **77.42% average accuracy** over 31 Gameweeks reflects the model's effectiveness in predicting points and assisting in captaincy decisions. Key highlights include:

* **100% accuracy** in multiple Gameweeks, showcasing the model’s ability to identify top-performing captains.
* Strong predictive performance during critical Gameweeks, ensuring reliable captaincy selections when points matter most.

The results validate the decision to use the Random Forest approach, as it successfully handles the non-linear relationships in the data, identifies feature importance, and prevents overfitting. Moving forward, the model can be further fine-tuned to incorporate additional features and dynamic weightings, enhancing its accuracy even further. This analysis demonstrates the robustness and practical utility of the captaincy selection model in FPL, offering significant improvements over heuristic or manual decision-making strategies.

1. **Acknowledgement**

We would like to acknowledge and thank both team members for their equally significant contributions to this project. Naishal utilized his extensive knowledge of FPL to lead the data preprocessing efforts, including handling missing values and creating features for the Captaincy and FDR models. He also worked on improving the accuracy of both models, making a substantial impact on the overall performance. Yashwi focused on designing the machine learning algorithms to achieve optimal predictions and model comparison. She also analyzed the data using visualizations such as scatter plots and bar graphs, which provided critical insights and enhanced the interpretation of the results. Both members' efforts were integral to the success of this project.

1. **References**

1. https://medium.com/@nedara\_98396/winning-with-data-my-approach-to-fantasy-premier-league-6c5dd4b84f55#:~:text=Final%20Thoughts:,the%20top%20of%20the%20table

2. Data Source : https://github.com/vaastav/Fantasy-Premier-League/tree/master/data

3.

4.