Fantasy Premier League - Predictive Modeling to Optimal Squad Selection

Zinan Aguirre
University of Southern California
3551 Trousdale Pkwy, Los Angeles, CA, 90089
zaaguirr@usc.edu

Abstract

Fantasy Premier League is an interactive game which users select a 15 player squad of Premier League players throughout the season. A premier league player's fantasy points contribute towards the user's overall season score based on their match performance. Managers(users) are faced with weekly decisions towards their squad selection as they want to pick the best performing players every week. These decisions present a difficult challenge for managers to select the best players consistently. This paper aims to predict the total amount of points every premier league player will produce in their upcoming matc hweek to help guide managers in their squad selection throughout the season. Models such as linear regression, random forest, and XGBoost were constructed to perform such predictions, in which the XGBoost model outperformed all. As a result, Mixed Integer Linear Programming with the predicted values of the XGBoost model was used to simulate the 2024-2025 season in order to optimize squad selection and transfers. This simulation which selected the best 15 player squad every week under the FPL constraints, produced a total score of 2467 points placing it among the top 2.26 percent of managers in the world.

1. Introduction

Fantasy Premier League is a fantasy sports game for the English Soccer Premier League. At the start of a new Premier League season, FPL invites people around the world to construct a squad of 15 premier league players. The objective of FPL is for managers to construct a squad that maximizes the total amount of fantasy points the squad produces in a game week. This weekly total is summed up across 38 game weeks of the entire Premier League season. At the end of the Premier League season, managers are ranked to a certain position against the other total points of competing managers in the world. For each player, their match performance is a direct relationship to the production of their total fantasy weekly points. The distribution of the fantasy points in FPL are shown in Table 1. Players get points based on the amount of goals, assists, minutes, or goals conceded in a match. They can also lose points based on negative actions in the pitch such as getting a yellow or red card. As a result, FPL managers must choose players that they predict will perform well in their match because if they do so it increases the potential amount of points they can produce.

Table 1. Fantasy Premier League Points System

Action	**Points**
Playing up to 60 minutes	1
Playing 60 minutes or more	2
Goal scored by GK	10
Goal scored by DEF	6
Goal scored by MID	5
Goal scored by FWD	4
Assist	3
Clean sheet by GK/DEF	4
Clean sheet by MID	1
Every 3 saves by GK	1
Penalty save	5
Penalty miss	-2
Bonus points (1-3)	1–3
Every 2 goals conceded by GK/DEF	-1
Yellow card	-1
Red card	-3
Own goal	-2

As FPL managers want to maximize their weekly total points there is a set of rules that they must follow from FPL which creates a variation of constructed squads by managers. These rules that managers must follow are that they must select two goalkeepers, five defenders, five midfielders, and three forwards in their 15 man squad. In addition, a maximum of three players from the same team can be chosen. To add on, FPL forces a budget of 100 million dollars for each manager in the selection of their 15 man squad. Each player's price differs, and this difference is due to a combination of perceived quality and demand. Premier League players who are considered are more world class players relative to their position and are of higher demand are priced at a premium cost in comparison to others. As the FPL season consists of 38 game weeks which spans over 10 months FPL allows one free transfer to be made every week to substitute players as long as such substitution follows the rules.

Initially at the start of a new season, a manager constructs a 15 player roster in the first week. Yet every week, FPL managers choose a starting eleven in which the points for those players will be counted to their weekly sum. The players who are put on the bench do not have their fantasy points calculated in the total sum, and only do so if a player in the starting eleven did not play. Then FPL will automatically swap the highest ranking bench player and add their points to the weekly sum. Every week FPL managers select a captain that produces two times the total amount of points the specific player produced in that game week.

These rules and decisions that FPL managers must follow create many challenges and constraints that they must keep in mind towards their overall goal in their selection of players. The main challenge in FPL for managers is to predict which players will perform the best in an upcoming match week as their performance affects the total amount of points they receive. One's ability to choose the best performing players before the game week occurs helps separate FPL managers to climb the ranks in their overall performance.

The decision that leads FPL managers to select players is a result of many factors. These can include their bias and favoritism towards certain players and teams. Their hope for such players to perform well can block the actual form and production of players negatively impacting their score. Another factor can be the lack of knowledge towards certain groups of players' ability especially as it is difficult for the average user to watch all ten matches in a game week.

Therefore all these decisions and constraints that FPL managers are faced with led to the purpose of this project to build models that predict every premier league player's total fantasy points in their upcoming match. The importance of creating this model accurately, would provide it to be a beneficial tool for FPL managers to utilize during their weekly decision making of team selection and transfers. It will be a factor towards these decisions that are not biased and based on each manager's opinion, but based on the statistics and form of a premier league player. This model will allow managers to interact with it each week in order to maximize their total fantasy points score accumulated as a manager.

2. Dataset and Data Exploration

In this section, we explore the data set obtained and an exploration of the features it consisted. Through this analysis, a construction of new features were created and added on to the dataset for additional relationships for the models to learn.

2.1. Dataset

The dataset that was used was of past FPL seasons and statistics from each Premier League player and team. These statistics were of match stats and also data specific to Fantasy Premier League. We used the FPL Historical Dataset [1] retrieved from a public GitHub repository that provides historical player data from previous Fantasy Premier League seasons. In this project, the 2022-2023 and the 2023-2024 season were used to train and build every model. In addition, the 2024-2025 season was used to further test the accuracy of the models and to optimize squad selection. For these three seasons, the data extracted were of the same format with the exact features for each player.

The data of the whole FPL season from the public GitHub repository was retrieved from folders named of the season such as 2023-2024. Inside the folder there is a csv file named ”mergedgw.csv” which consisted every single players statistics for each of the 38 gameweeks in that FPL season. For each player and game week, the data set recorded the statistics of their performance of that specific game week after the match had been played. These statistics included goals, assists, and minutes played. In addition, advanced metrics such as expected goals, assists,goals conceded, saves were recorded for every match based on the player's performance. There were also features that were specific to FPL. Such examples included a feature called the ICT index, which is a combination of a metric score of influence, creativity, and threat. This index value explains a players' overall performance in the match. The higher the impact a player had in the match then the higher the ICT index value was and vice versa. One final important variable in the dataset was fixture difficulty of the upcoming team in the next match. Every team in the FPL are expressed as a certain value which represents their status of difficulty for a team's ability to win against. Teams with a higher value are perceived as a difficult matchup while those with lower values are perceived easier to win against.

Table 2. Features in the VAAS FPL GW Statistics

Feature	Description
Goals	Number of goals scored
Assists	Number of assists
Minutes Played	Total minutes played
Opponent	Opposing team
Was_Home	Home/Away indicator
Saves	Saves made (GK)
Clean Sheets	Whether team kept CS
XG / XA	Expected goals / assists
ICT Index	Influence–Creativity–Threat metric

2.2. Feature Engineering

The features from the original dataset were a representation of the statistics of a players' performance after the match had concluded. In order to create usable models which did not leak data and also predicted a players' total fantasy points in the upcoming weekend, the features were changed to represent a player's form up to the upcoming match. It is impossible in sports to input stats of a player in a game that has not occurred. Instead one can input the stats and data of a player's form as it can provide insight to how such players can perform in the next fixture. The decision to focus on a player's form to use for prediction was based on one of the many factors in which FPL managers use to decide their player selection. They look at previous performances of a player to create options and decisions to the players.

Using the data of the original features of each game for each player, it was transformed to calculate the rolling statistics up to the current game week but not including. These features that were created seen from Table 3 as a few examples were goal scored in the last 5 matches, goals per 90 in the last 5, and goal per 90 thus far in the season. The ICT Index was also rolled to get an understanding of the impact a player has had on their matches as of recently. This was done for every feature that the original dataset provided to get a better understanding of a player's performance in the past few weeks. The rolling statistics counted the last 5 games as it represented the amount of matches on average played in a month, therefore it meant that the form was based on the player in the past month. This window of 5 matches provides an understanding of a true form that would not be as volatile in comparison to looking at past week's stats.

In addition, the dataset which included all the players in the FPL season was divided into four separate datasets, each containing players specific to their position. This division consisted of Goalkeepers, Defenders, Midfielders, and Forwards. The reason to separate the dataset into positions was to allow features such as goals be more prevalent and understood in models for forwards. While features such as saves and clean sheets could be understood better in their relationship of total points for the prediction models of goalkeepers to obtain as accurate predictions.

2.3. Data Exploration

Preliminary exploratory data analysis was done for each position in order to understand the relationship of the features engineered and the total fantasy points a player performed in a given match. In this exploration, we looked at the 2022-2023 FPL season that will later be used as part of our training dataset. One of the first topics explored was the impact that fixture difficulty had towards the total amount of points a player produces. This is a common question in FPL discourse that always impacts managers to substitute players with harder opponents to those with weaker opponents in the upcoming game week.

Table 3. Examples of Engineered Features Used in the Model

Feature	Description
Rolling Last-5 Features	
goals_last5	Total Goals scored in last 5.
assists_last5	Assist.
minutes_last5	Minutes
points_last5	FPL points
ict_last5	Mean ICT Index
threat_last5	Mean Threat metric
creativity_last5	Mean Creativity metric
Per-90 Features in last 5 games	
goals_per90_last5	Goals per 90 minutes
assists_per90_last5	Assists per 90 minutes
clean_sheets_per90_last5	Clean sheets per 90 minutes
xG/xA-Based Features in last 5 matches	
xG_last5	Expected goals averaged
xGI_last5	Expected goal involvements average
xGC_last5	Expected goals conceded
xGA_last5	Expected assists
xG_per90_last5	Expected goals per 90 minutes
xGI_per90_last5	Expected goal involvments per 90 minutes

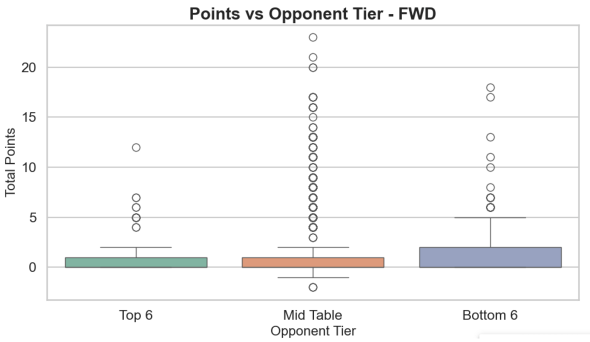

Figure 1. 2022-2023 Forward Total Points Performance based on fixture difficulty

In figure one, it is presented that forwards and midfielders produce more points on average against the bottom six teams based on their strength value in comparison to the top 6. While for goalkeepers and defenders on average fixture difficulty does not impact greatly their total points but there is a greater possibility for outlying performance of high total points in lower and mid difficulty teams.

The feature engineering emphasized an importance of

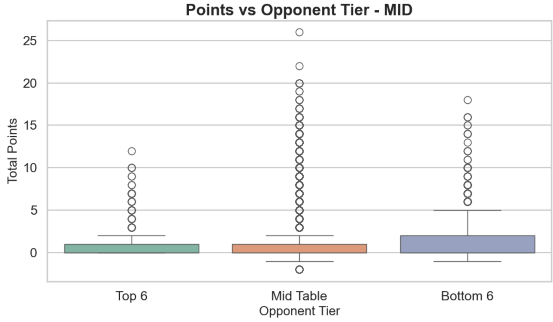

Figure 2. 2022-2023 Midfielder Total Points Performance based on fixture difficulty

form versus last week's performance. Therefore it was important to have a preliminary understanding between the relationship of the features towards the total fantasy points scored in a given game week. The relationships between each feature separately with the total points variable led to a non-linear relationship. Therefore Spearman's rank correlation coefficient was an optimal way to understand these relationships and how possible changes towards certain features impacted the total points.

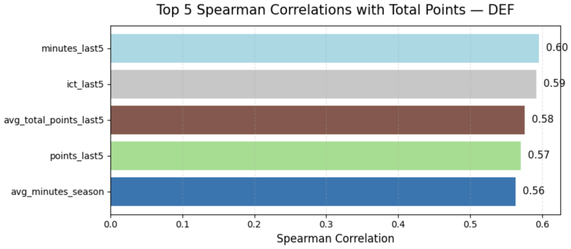

Figure 3. Defender Spearman Correlation

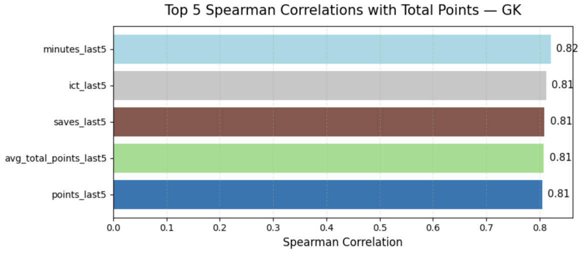

Figure 4. GK Spearman Correlation

The top 5 values of each feature for each position were recorded and in all positions two features that appeared consistently were ict index and total minutes played in the last 5 games. This relationship in respect to minutes is important as minutes played suggests whether a player played or not. The averages of points a player receives in a match is low thus the one of the most guaranteed and easiest way for players to get points consistently is by surpassing 60 minutes of playing time to receive two points. The other feature of ict index mean in the last 5 matches represent an importance of being impactful in a match. In this case, for all positions the data represent that the more meaningful or unimpactful a player was ultimately correlated greatly with their total points scored in a match. There were also features specific to each position such as goalies' total points were impacted by their saves in the last 5 matches. Saves in another avenue for goalkeepers to get points that are only specific to their role and position in the team. Features such as creativity total score in the last 5 games for forwards and midfielders were highly correlated with total points as influence in the attack increases a players' chances of assisting or scoring goals.

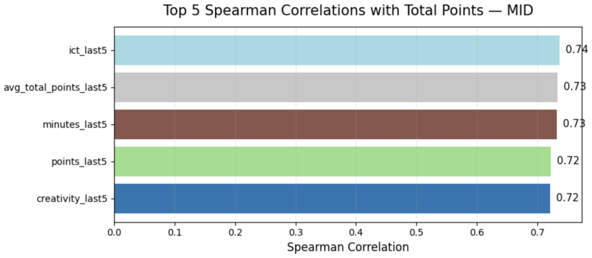

Figure 5. Midfielder Spearman Correlation

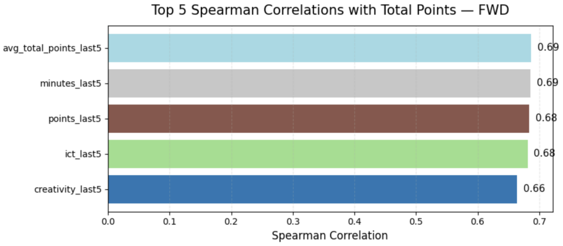

Figure 6. Forward Spearman Correlation

One important aspect in FPL is choosing players so that the overall cost of a team fits under the budget constraint. Therefore finding optimal players that return a positive amount of fantasy points in relation to their price is crucial for managers to maximize their squad points. In this season, we looked at the relationship between a player's value and their total points for any given match week. Players at the premium cost are seen in the public to have more potential to produce a high amount of total points, and this relationship is shown in figure 7. For many of the players in this dataset, they are priced at of similar values but we see that there are moments when such players can outperform others relative to their cost range. As well they can be able to compete against players that are of premium cost in comparison to their lower cost. This brings an importance to be able to predict and optimize the ability to choose those players based the prediction models accuracy that we produced.

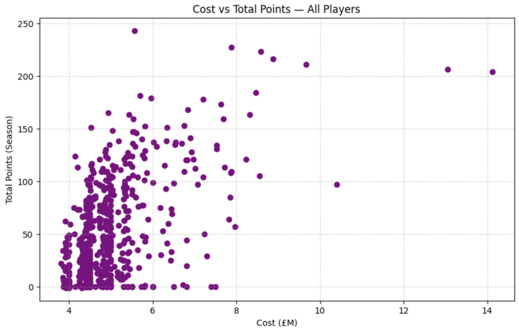

Figure 7. 2022-2023 Comparison of total cost and the total amount of fantasy points that was produced in the season for each player.

3. Methods

Based on the feature engineering of the dataset which transformed all the statistics to focus on a player's form up until that game week of the whole season, it led me to design models of forecasting methods. The form of a player is data that is historical and time specific thus when analyzed, models can make predictions of the future and in this case of the total amount of points a player will obtain in the upcoming match. In addition, in the development of the models we separated based on the dataset by position and therefore each position had their own prediction model. This allowed for certain features which were related to certain positions to have more impact and not be reduced if combining all players into one dataset. In addition, the improvement of metrics mean absolute error, R2 statistic, and root mean squared error was emphasized through cross validation in order to find the best parameters based on the values of each metric.

3.1. Logistic Regression

One important decision when it comes to FPL squad selection and prediction is determining whether a player is going to start a match. This factor is impactful as playing sixty or more minutes guarantees two fantasy points in comparison to zero or one point if the player comes off the bench. After seeing the low averages of points for each player in FPL, every point is crucial to get such an advantage. Thus creating a logistic regression model is important to predict if a player is going to play or not. The same features from the feature engineering are used to show an emphasis on the idea that if a player is on a good form and their rolling minutes are positive then there is sufficient data to create a probability whether they will start or not. Each logistic model was made for each position and the target variable was the binary feature, 'starts', which said if the player did or did not start that game week. The process of a five fold time series cross validation was used to find the best parameters within the logistic regression model for each position. Then it used the parameters found to fit each model by position from the whole dataset of players.

3.2. Linear Models

Building a linear regression model to predict the total amount of points a player will have on the upcoming weekend based on their form, home/away status, and opponent strength, established a base model of performance for this prediction. This approach was used to build fast and robust models that were interpretable at understanding further the relationship between the features and total points predicted. For each of the models, the 2022-23 and 2023-2024 datasets were combined and then separated from each position. In order to optimize the limited amount of data, the use of k-Fold time series cross validation was used for each of the regression models to use all the data and to find the best parameters that predicted as accurately as they could.

3.2.1 Ordinary Linear Regression

In the development of this model, as it represents the ordinary form of linear regression, it was used as a baseline for linear models to improve on. The importance of creating this model is that it will provide interpretable coefficients for each feature. Cross-validation was used to find the best parameter of the regularization strength.

3.2.2 Ridge Regression

Before applying the K-Fold cross validation and being fitted, the features were scaled through StandardScaler and this also occurred for Lasso Regression. The use of ridge regression was used to see if there was an improvement from the basic linear regression by dealing with multicollinear variables if they exist. In this situation many of the features in the statistics on the surface level are highly correlated with each other such as goals in last 5 matches and expected goals as usually a higher xG leads to higher goals scored.

3.2.3 Lasso Regression

The development of Lasso Regression was similarly constructed using Pipeline to avoid data leakage and perform such transformations to the features that had been done in Ridge. The number of iterations that this model used to ensure it converged was up to a maximum of 1,000. The goal of lasso regression is further minimize the number of features that were used in the predictions to a coefficient up to zero in order to understand which features truly impacted prediction.

3.3. Ensemble Methods

Linear regression models offer a baseline on how a simple model can perform using the available data and assuming a linear relationship to produce a prediction. In sports and specifically, in the Premier League Fantasy, many of the features' relationship with the total amount of points is non-linear. The features of the dataset emphasize form and it already creates a non-linear relationship with the points they output in a given week. The bonus point system in FPL is not as linear and robust as there are no set of rules and performances from the features that guarantees a player to get extra bonus points in a given week which then impacts their overall total points. In addition, features are performance based this means sports in general are non-linear as there are many complex systems and interactions that are impacted by human variability and action that is not consistent and highly variable. These methods will use regression trees that will capture the nonlinearity of the data and continuously improve on the weaker learner models to produce the best possible model.

3.3.1 Random Forests

Random forests is an ensemble method that can handle the nonlinearity and categorical features in the data set such as whether the player was playing at home and who they were facing against. As much of the dataset is in tabular data form with many features, random forests was a prime model to use to predict in comparison to the linear models. Soccer and the statistics for each week was highly variable due to its unpredictableness in sports. The use of random forests was used to be able to reduce this variation so that given any time during the season of historical or upcoming data, the model can still produced well in its predictions. For this model, and all other models the same dataset of the linear models were used and separated by position. A 5 fold time series cross validation was done to find the best parameters according to each position's dataset such as depth,learning rate,l2 leaf regularization, bagging temperature, random strength, border count, and iterations in the forest. After the parameters were found for each position, they were refitted on the entire dataset of the position to be later used for the 2024-2025 FPL season dataset.

3.3.2 XGBoost

This model focused on the gradient boosting algorithm that is able to handle the tabular data set for each position based on their performance. The motivation behind the XGBoost model was to continue to improve and understand the nonlinear relationships of the data. As this process builds trees sequentially from improving from the previous trees, it uses the L1 and L2 penalties in which from the regression models gave an extra understanding of which coefficient values from the features were minimized. As the datasets were separated by positions the size also reduced therefore XGBoost's capabilities of being able to reduce overfitting was crucial in order for it to perform well on different datasets of different seasons. In this model for each position the objective parameter was squared error and a 5 fold time series CV was used to find the best parameters. These parameters that were found were optimal colsamplebytree,learning rate, maxdepth, nestimators, subsample. These would then be refitted on each position dataset using the parameters found.

4. Evaluation and Results

In this section, we analyze the models constructed from the previous section and their results. All models were fitted for each position, and the data that were used were of the full FPL seasons from 2022-2023 and 2023-2024. Therefore, it was important to test these models in the 2024-2025 FPL season. This season which had just completed included all the data of the 38 game weeks thus it allowed us to understand how well these models would do from the start until the end of a new upcoming season. For the 2024-2025 FPL season dataset, the same feature engineering and transformations were made to produce the exact framework of features that our models were fitted based on a player's form up to an upcoming game week. It is important to note that through the 2022-2024 seasons, the rules and systems of FPL which determined how players obtain fantasy points were still applied and unchanged to the 2024-2025 season. Thus, there were no other features that could have been implemented in other ways to obtain fantasy points. In addition, the metrics used in the evaluations were MAE, RMSE and R2 statistics. It allowed us to understand the error of the model's prediction of a player's total fantasy points in comparison to the actual total points a player obtained in that game week. The R2 statistic is an important metric to understand how much our model was able to pattern and analyze the data to create their prediction because FPL can be very noisy and unpredictable in relation to the nature of sports.

4.1. Prediction to Start

We first analyze the logistic regression models that were built to predict whether a player would start in the upcoming match. In Table 4, the accuracy of each model is presented in which the goalkeeper model had the highest accuracy of 0.9211 and the AUC-ROC score. The higher values of the goalkeeper model can be understood through the total number of predictions it made. In general, there is a smaller pool of goalkeepers in a team, thus there is a small group of players that compete to start. On average a soccer team has 3 different goalkeepers and therefore it gives this model an advantage in correctly predicting whether a goalkeeper starts as there is less variability. A goalkeeper is more consistent within starting week in and week out in comparison to other outfield players.

For the other positions, the accuracy of their models decreased to the mid eighty percent. The forward logistic model was the highest with eighty eight percent. In relation to the forward model, it can likely be due to the similar situation seen in goalkeepers, where there is already a smaller pool of forwards in a team. This allows for a more consistent selection of players to start every game week compared to defenders and midfielders where there are plentiful of players. In addition, the accuracy for these outfield models is not as high as the goalkeeper models, as premier league players are susceptible to rotations in a season. These rotations come into play due to sudden injury or rotations within the squad in order to have players rest after playing a midweek competition game.

Position	**Accuracy**	**ROC-AUC**
GK	0.9211	0.9598
DEF	0.8530	0.9232
MID	0.8417	0.9168
FWD	0.8871	0.9422

Table 4. Model Accuracy and ROC-AUC by Position

Despite the slight decrease in the percentage of accuracy for the models based on midfielders and defenders, the AUC ROC score is consistent and strong with values above 0.916. This means that the logistic model when used in a new upcoming season was effective in ranking players by assigning higher probabilities to those who start versus those who do not start. It is important for the model to recognize these relationships and have high accuracy as randomly choosing a player to start or not will negatively impact the amount of points a manager obtains in their squad optimization. The logistic models by assigning the higher probabilities effectively showed that over the course of the FPL season they understood which players were regular starters and those who were not. The accuracy is due to many factors that the model may not predict but its power is shown to understand which players are starters in their teams.

In addition, tables five through nine represent the actual data of the number of predictions that were made for the entire season. For example, the number of players that were predicted to start but did not start for that game week. The number of players who were predicted to not start and did not start was the largest among all. This emphasizes the importance and ability of all position based models to correctly predict the non starters as there is a smaller pool of players to select who will start. This same argument is applied for the number of times during the whole season the model predicted a player to start and they were correct. The consistency of the models to predict starters and non starters correctly is shown in the data and is essential for FPL managers as they become a reliable tool for player selection during the season.

The moments in which the model's predictions were incorrect can be seen, for example, in the forward position based model. It predicted 146 times that a forward, not necessarily the same forward, will start in the next match but in actuality they did not. The vise versa is shown that for 185 times it predicted a player would not start but they actually did for the game week. These values show the limit that the forward model has in prediction as it cannot factor into other situations that could have led their prediction to be incorrect. These factors which are constant among all positions may be influenced by the squad rotations that occur in football as such substitution can cause a false positive and false negative incident to increment together. As assuming a player will not start and a player will start but due to rotation both predictions become incorrect. All together, the ability to predict whether a player will start in the logistic position based models was strong for the whole season of 2025 FPL based on the features of the dataset.

	Predicted Not to Start	Predicted To Start
Did not Start	1932	117
Did Start	103	635

Table 5. GK Confusion Matrix:

	Predicted Not to Start	Predicted To Start
Did not Start	5311	570
Did Start	735	2262

Table 6. Defender Confusion Matrix

	Predicted Not to Start	Predicted To Start
Did not Start	7361	899
Did Start	985	2658

Table 7. Midfielder Confusion Matrix

	Predicted Not to Start	Predicted To Start
Did not Start	2036	146
Did Start	185	564

Table 8. Forward Confusion Matrix

4.2. Linear Regression Model Performance

In the following section, the linear regression models that were constructed such as the Ordinary, Ridge, and Lasso regressions were compared with each other in their evaluation of the 2024-2025 FPL season data set. The importance of doing such a comparison among the linear models was to understand and find the most optimal baseline model of point prediction that would later be compared to one of the ensemble methods. As the data set contained a variety of features that were based on a player's form in soccer, it became important to understand which features were impactful towards the prediction of total points on a basic level.

Starting with Goalkeeper models, through the evaluation of the metrics of RMSE, MAE, and R2 statistics in the table below. It is seen that in terms of RMSE, the value decreased from ordinary regression of 1.6813 to 1.6717 for the Ridge regression and 1.6358 for the Lasso regression. The application of RMSE and for the rest of models represent the error of prediction per player from the actual points, but putting a larger emphasis on situations where the model's error was large. A large error means that the model's prediction either over predicted, suggesting that the player underperformed or the model under predicted, meaning that the player over performed.

Model / Metric	GK	DEF	MID	FWD
Linear Regression				
RMSE	1.6999	1.8712	1.9717	2.2433
MAE	0.9258	1.1292	1.0640	1.2719
R^2	0.3112	0.2361	0.3134	0.3223
Ridge Regression				
RMSE	1.6718	1.8584	1.9697	2.2419
MAE	0.8148	1.0744	1.0412	1.2354
R^2	0.3338	0.2466	0.3148	0.3232
Lasso Regression				
RMSE	1.6359	1.8606	1.9884	2.2413
MAE	0.8013	1.0615	1.0574	1.2368
R^2	0.3621	0.2448	0.3017	0.3235

Table 9. Fantasy Point Prediction Performance of Linear, Ridge, and Lasso Regression Models by Position for the 2024-2025 FPL season

A different perspective within the ability of the model to predict can be seen in MAE metric. MAE represents the absolute error of prediction without the emphasis towards large error values being significant and instead being equal among all data predictions. It shows the model's overall performance every week on average throughout the 2024-2025 FPL season. In this metric of comparison within the goalkeeper linear models, the trend of decrease in error is seen through Linear to Ridge to Lasso regression. This improvement within the goalkeeper models shows that the accuracy of prediction for a goalkeeper improved as certain feature's coefficient values were dropping closer to zero. This can be seen greatly within goalkeepers, as features that have an emphasis towards attacking statistics such as goals scored and assists have minimal impact for goalkeepers. Goalkeepers get their points from a defensive perspective. Features such as clean sheets and saves, and goals conceded result in the player's performance of total fantasy points and therefore Lasso regression is able to regularize the features. The Lasso regression model was able to learn the relationships and impact of features better due to its improved accuracy.

Compared to the R2 statistic scores for goalkeepers, the trend is the opposite, in which the values increased from linear models of Ordinary, to Ridge, to Lasso regression. Lasso regression was able to score a statistic of .3621, which meant that the variability of the actual points in the 2024-2025 season could be analyzed and explained due to the features of the dataset that the models were fitted on. This means that these actual total fantasy point values of the 2025 FPL season, around thirty six percent of it, the features could explain those differences. The reason why this value is not really high can be due to the great variation and unpredictability that comes with soccer. However, this value of R2 statistic is the highest among all other positions. This shows that the Lasso Regression model for goalkeepers was able to predict and as well explain the variability of the total fantasy points in goalkeepers better than the other models with respect to their position. The goalkeeper models had a higher value and it can be due to the fact that there is less potential of randomness of spikes in total points to occur in goalkeepers in comparison to forwards. A forward is able to have a higher spike in points on a week to week basis versus goalkeepers due to their ability to score and get assists. Therefore it leads to much more variability within the total points for forwards in comparison to goalkeepers.

We expand further to analyze the models for the other positions and from table 9, it is shown that in terms of the RMSE value they are all consistent and of the same value up to three decimal places. A bit of change can be seen in the Lasso regression models for defenders and midfielders as it does a bit worse in comparison to the ordinary and ridge regression models. One of the main situations with which the linear prediction models have to deal is the randomness and variance of a player scoring or assisting as it gets them points. There is more variance as we progress from defender to forwards, and this difficulty is shown in the values of the metrics. The RMSE values for the defender, midfielder, and forward regression models are higher than the goalkeeper's model. They increase by position since forwards have the highest potential and ability to score more goals and assists in a game week. For example, in the regression models for forwards they consisted of around 2.24 points which means that these linear models have moments where there is great error from their points prediction to actual total points. To add on, the model either over or under predicted a player's performance in the upcoming game week showing that it can be hard for the forward prediction model to correctly predict the weeks in which a forward spikes in total fantasy points.

In addition, from the tables for the three linear models, the MAE and R2 statistic values are very similar to each other in respect the defender, midfielder and forward data set. The MAE for defenders and midfielders is very close to 1, and for the forwards it is around 1.2. This value means that on average for each player per game in the 38 game season, there was a one fantasy point difference from their predicted to actual point value. This similar performance from each of the models towards point prediction for all

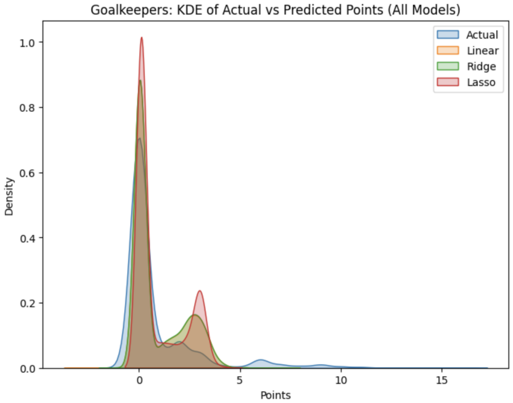

Figure 8.

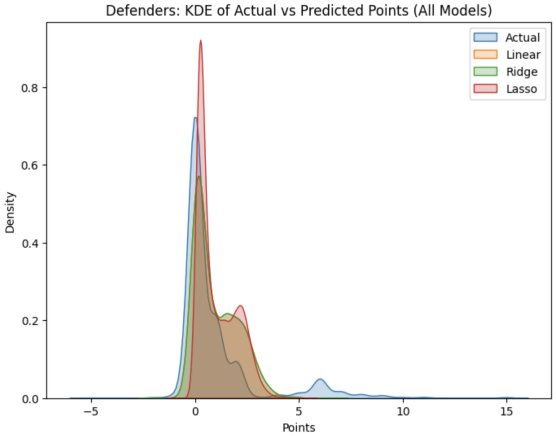

Figure 9.

positions other than goalie represents that smaller impact regularization had versus its impact on the goalkeeper.

The following figures are representations of the density of the Kernel distribution of each of the linear models prediction for the upcoming game week in comparison to the distribution of the actual performance of fantasy points in the 2025 FPL season. Ridge and Ordinary regressions were very similar to each other in their performance, which supports the reason that the values within the accuracy metrics were very similar to each other. For the Lasso regression prediction models in all the positions, its density is shifted a bit more to the right and it has more density in predicting high points for a given week. This shows its ability to predict more extreme and higher values of total points based on its ability to lower coefficient values to zero. This is an important feature within the model in terms of FPL since it shows that it is able to predict for certain weeks to certain players a spike in their total points. Whether their prediction is accurate or not, Lasso is able to make such predictions that allow FPL managers to understand how well a player is expected to perform.

Despite the Lasso Regression models being able to predict more spikes in total points in comparison to the other regression models, it is still not enough in comparison to the actual distribution of fantasy points. There is a portion of density towards the higher point values that is large in the actual points distribution, but there is very little density in the regression models. It is important for the models to have greater density in those areas in their prediction values as for example, a forward can many times throughout the season get really high total fantasy points in a match. Yet looking at the density graph for forwards, all the linear models rarely ever predict a forward's fantasy points to surpass five points. Therefore this is one of the limitations of the regression models for position.

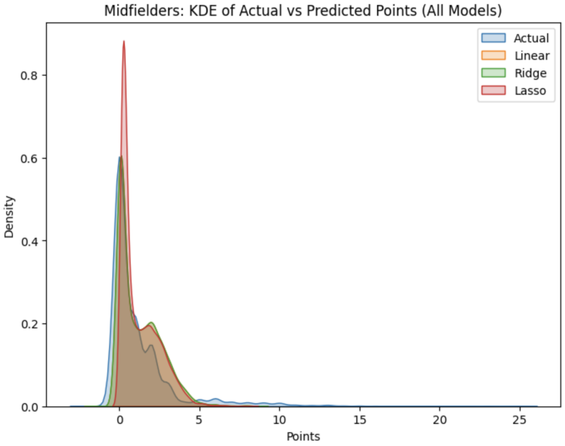

Figure 10.

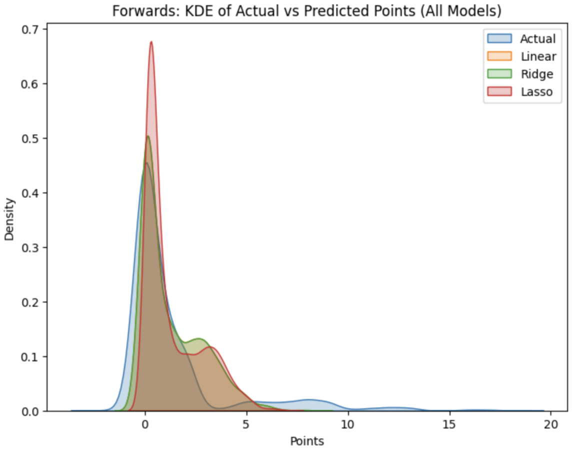

Figure 11.

4.3. Ensemble Model Performance

A random forest model and an XGboost model were constructed to understand the relationship of the features with total points assuming nonlinearity to improve the predicted precision compared to linear regression. Starting with the random forest models specific to each position, table ten shows the performance in regards to each metric in the 2024-2025 FPL season. Using the ridge regression model metrics compared to the random forest model, the RF model performs better in almost every metric for each position. In the MAE scores, as this explains the overall absolute accuracy of the model on average per player for each game week over the whole season. In ridge regression it decreased from 0.81 to 0.76 for goalkeepers, 1.07 to 1.01 for defenders, 1.04 to 1.02 for midfielders, and 1.23 to 1.24 for forwards. Therefore the only metric for which the ridge regression model outperformed the random forest models was for forwards very slightly. These improvements in MAE ultimately represent that the random forests model for each position is a superior model to the regression models, as it is more consistent in their predictions of fantasy points with regard to the actual points.

Table 10. Model Performance Comparison: Random Forest vs XGBoost

Position	Model	RMSE	MAE	R2
GK	Random Forest	1.6146	0.7641	0.3786
	XGBoost	1.6031	0.7524	0.3874
DEF	Random Forest	1.8387	1.0193	0.2624
	XGBoost	1.8315	1.0159	0.2682
MID	Random Forest	1.9591	1.0213	0.3221
	XGBoost	1.9429	1.0062	0.3333
FWD	Random Forest	2.2292	1.2425	0.3308
	XGBoost	2.2000	1.2095	0.3482

In addition to the improvement in the accuracy of the random forest models, there was an increase in the R2 statistic for each of the positions compared to the linear models. Therefore, the RF models were able to capture and explain the variance of total points in a game week better than regression models. This has to be due in part to the fact that ensemble methods are able to capture the nonlinearity of the data. Assuming a linear relationship leads to more restrictive understanding of the relationships that can limit the true relationships within the variance of fantasy points every game week. As well as RF is an ensemble method, its goal is to take an average from the trees constructed from its random splits thus creating a more flexible model that can be able to handle and understand such variance.

In figure 12, SHAP values are represented to help understand and give interpretability of the RF models that were constructed. SHAP values show which features of the data affected the prediction made by the random forest model. In this context, the feature value represents the value that a feature had for a specific player in that game week. The SHAP value in the x axis represents the value of predicted fantasy points that impacts the overall predicted fantasy points as a result of the feature's value.

Specifically, figure 12, is the SHAP model of the midfield random forest prediction model. The main feature that had the biggest impact to this model's prediction was the ICT mean value from the last 5 matches of a player. This model shows all the data on how this feature affected the prediction of points. For example, there was a player in the season in which their ICT last5 was a high value, so as a result this RF model produced a +1.5 in their predicted points for that game week. Another example is the feature of opponent strength. The RF model throughout the season in every game week given a midfielder's opponent difficulty it associated a high feature as a difficult team. This association led the model to many weeks in the season decrease a midfielder's predicted points of around 0.1 ot 0.5.

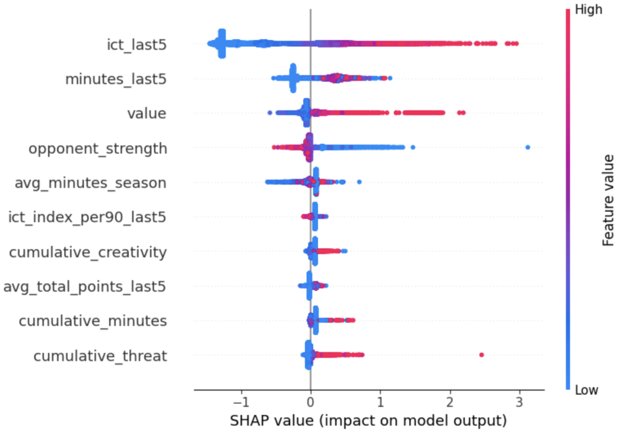

Figure 12. Midfielder Random Forest model SHAP Values

In relation to the high value of the feature in ICT last5, the midfield random forest model associated this feature as a positive indicator of their predicted points. This feature gives a score of this players impact on a match, thus on average over the last 5 games. Their impact is based on their creativity, threat, and influence in the match. Therefore the midfield random forest model believes that in fantasy point prediction being more involved and impactful in the last few matches for midfielders lead to a higher performance in positive fantasy points.

This improvement from linear regression to Random Forests continued to expand with respect to the XGBoost models that were built for each position. It can be seen in table 10, as XGBoost improved on both the linear and random forests models. We compare the XGBoost model's values of RMSE, MAE, and R2 statistic against the random forest, as it proved to be the superior model from earlier. Starting with the RMSE of XGBoost there were slight improvements for goalkeepers,midfielders, and defenders by an average of .01 decrease. Yet looking at the table, the improvement in RMSE for the forward XGBoost model increased to 2.20 compared to the 2.22 of the forward random

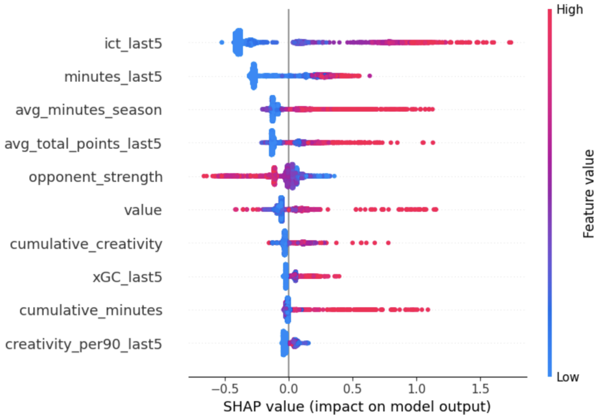

Figure 13. Forward XGBoost model SHAP Values

forest.

This is an important position for XGBoost to improve on, since RMSE is impacted by the great error values of the model's prediction it meant that XGBoost was able to lower the values of great error that existed in the other models. This meant that the XGBoost model was able to improve on the over or under predictions every game week that the earlier models presented. A decrease in RMSE is essential as this position has a high variance in the total points a forward performs per week. One week they can score 2 points and the other 10 weeks, leading models to be vulnerable to the over and under predicted values. Thus the XGBoost's prediction improved on this so that its error is not too impacted by this variance.

The overall accuracy of the XGBoost model for each position improved as the MAE scores decreased in comparison to random forest. This improvement is seen through the forward XG boost model as it obtained a value of 1.209 while the random forest model had 1.242. Therefore the XGBoost model not only improved the large error predictions from the other models but also the overall absolute error prediction from each forward in comparison to the actual amount of points they produced in the game week. This expands for every player in each position. In addition, for the XGBoost models, there was a slight increase in the R2 statistic for each position based model. In addition to their accuracy, these models expanded their flexibility compared to random forests to better understand and explain the variance of the total points every game week. Overall this consistent improvement across all metrics and positions prove that the XGBoost model is the best performing model in their prediction of fantasy points based on the features that factored greatly a player's form.

Analyzing the features of the XGBoost model with SHAP, it is to be noted that many of the top features of impact in the XGBoost model can be very similar to the random forests models in respect to their positions. This means that there are a group of features that are consistent among models as important for fantasy prediction points despite the different approaches the models took. Yet to add on, there is still a difference among these models in how much influence certain features have towards the model's prediction. One interesting difference between how two position based models handle the same features differently can be seen in comparison to the SHAP figures 14 and 15 of Forward XGBoost and RF.

There are similarities in the top features of impact for each of the two models but the impact towards prediction is different. For example, the opponent strength feature appears in both models. In random forest this feature influenced the prediction by increasing a player's predicted point by 1 if the opponent strength value was low. In comparison XGBoost uses the same feature to interpret that a higher value in opponent strength impacts the forward's predicted points negatively by up to 0.5 points. Therefore despite the same feature the XGBoost model was influenced more based on if the opponent strength was high rather than it being low as seen in RF forward model. Their approaches in relation to this feature are neither right nor wrong as in general FPL context, FPL managers have different beliefs about whether to add or remove players based on the difficulty of the next match. These two models show such decision and prediction in their final result.

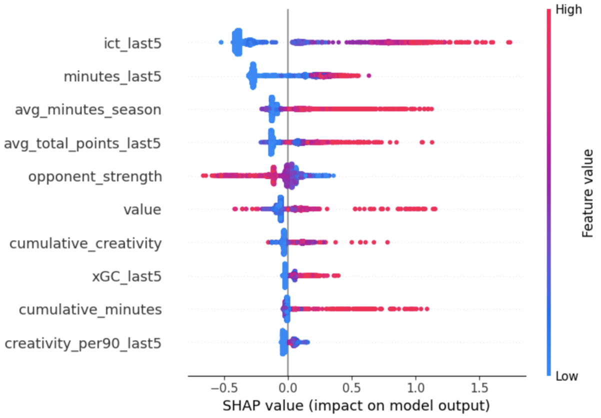

Figure 14. Forward XGBoost model SHAP Values

In figures 13,15,16,and 17, the SHAP values of the XGBoost prediction model for each position are presented. These figures represent the top 10 impactful features towards each model's prediction of total fantasy points of a player in an upcoming weekend. Throughout each position there are similarities between each model such as the feature ICTlast5 being one of the most impactful in prediction. Thus for the XGBoost models, this feature had a great impact towards their prediction value as they felt that a player's overall impact in a match impacts the fantasy points they will perform in the next game week. This feature like many

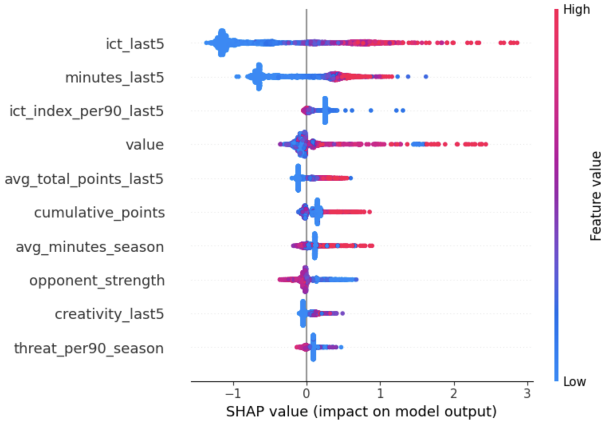

Figure 15. Forward Random Forest model SHAP Values

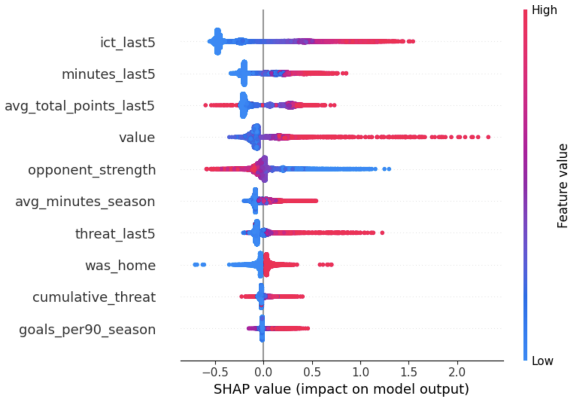

Figure 16. Midfielder XGBoost model SHAP Values

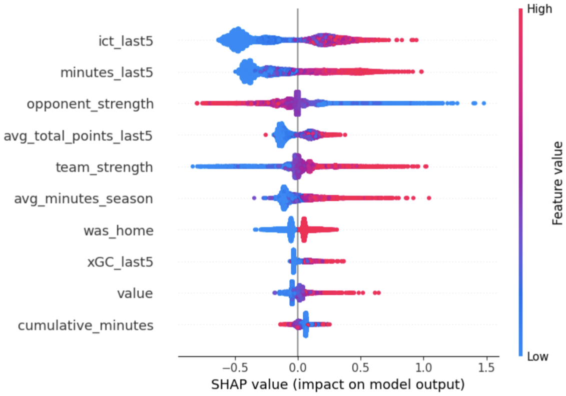

Figure 17. Defender XGBoost model SHAP Values

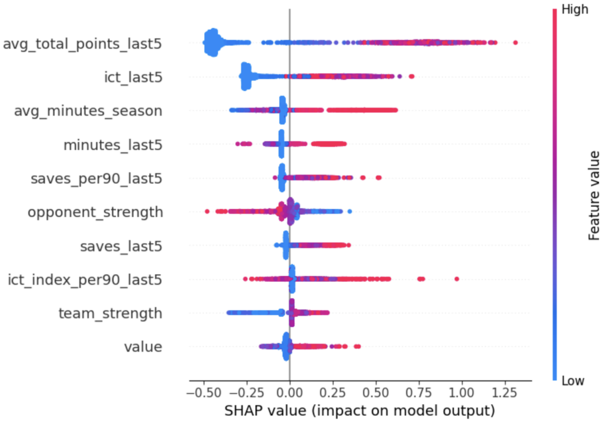

Figure 18. Goalkeeper XGBoost model SHAP Values

others from the figures are seen to be consistent among all positions in XGBoost models, but there are differences between each as different positions lead to different ways to consistently gain points.

For example, comparing the feature of team strength among the models shows that it is impactful for defenders and goalkeepers due to their ability to get clean sheets. While the midfielder SHAP model does not have this feature because it only gets 1 point for a clean sheet which is not as impactful as four points for defenders and goalkeepers. To further analyze, the impact of team strength further differs within goalkeepers and defenders. In the defender model, this feature has a balanced impact where a low value will lead the model to decrease a defender's predicted points and vise versa. In comparison, the use of feature was not as balanced for goalkeepers and it did not greatly penalize goalkeepers if they were in weaker teams. A potential reason for this can be due to the ability of goalkeepers getting points in saves. Thus there can be a beneficial impact according to the XGBoost model for goalkeepers to be a weaker team as they can face more shots so the goalie has many chances to do saves. The defenders cannot get such points and are then highly reliant on the team's overall strength for their fantasy points according to the XGboost model. This difference shows the importance of building separate models for each position because the model tackled each feature differently based on the position.

Continuing on the ensemble models, figures 19 and 20 represent player model points predictions that these models made throughout the 2024-2025 FPL season. The two players chosen were Mikel Merino and Erling Haaland as they both encompass prime examples of FPL players that many managers face in a season for each game week. Mikel Merino is an example of many common FPL players in which throughout the season the average amount of points per game is around the same range. Yet there will be moments where in a game he gets an assist or scores as he is a midfielder of Arsenal, a strong premier league team.

In the figure it is shown that the XGboost and RF model predict very similar amount of points each week for Merino and these predictions are close to the actual values. But both models have great error prediction in weeks in which Merino's points spike as they did not do well to predict such performance. It becomes a classic example of under predicting. Yet we can see the impact of form towards prediction in the end of the season. Merino can be seen to have much better form based on his actual values and both models were able to recognize this, and therefore give the highest predicted values seen in the whole season. Overall players such as Merino these two models are able to predict accurately in many of the weeks in which they perform in the similar range every week.

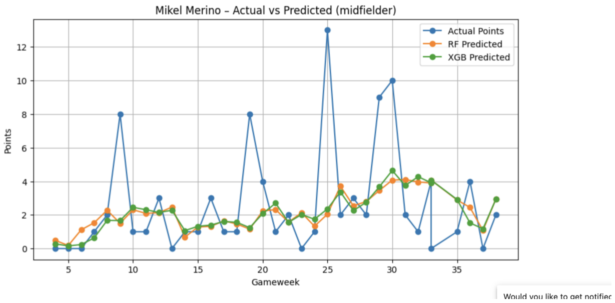

Figure 19. 2024-2025 Mikel Merino Predicted Points

Yet players such as Erling Haaland are prime examples of forwards with great variability and extreme performances in any given game week. He can for example produce 10 plus points but then the next week produce two points despite his recent good form. In these situations the figure show how different the XGBoost and RF models perform when dealing with such players. Both models by each game week were able to understand the data based on form as they decrease and increase predicted points based on Haaland's form. Yet they differ in how sharp and bound these predictions were. The random forest model gave higher points predictions in comparison to the XGboost model and as a result let the RF model to have larger error in many game weeks. The XGBoost model was already over-predicting in certain game weeks, but the RF model was over-doing it even more. The situations in which the random forest was seen to outperform in a given week with respect to error was when Erling Haaland perfectly performed such optimistic predictions from RF. Yet situations in which this occurred in a 38 game season is very little in comparison to weeks of over-prediction.

In the table 11, the total cumulative points of the 2025 FPL season is presented along with the total number of points the XGBoost and Random Forest models predicted for each position. Overall both models over the course of the whole season over-predicted in comparison to the actual total value, but the Random Forest model over-predicted more than the XGBoost model. It is interesting that both models over-predict as in the player predictions graphs there could be cases where the model severely under-predicts. Yet using the data of total cumulative points it shows that these moments are not often as the total amount of points would have been larger than the predictions. Therefore in Fantasy Premier League, a lot of the study and emphasis is towards predicting and optimizing the consistency of players overall in a season compared to predict such spikes in points as it is already very difficult and rare .

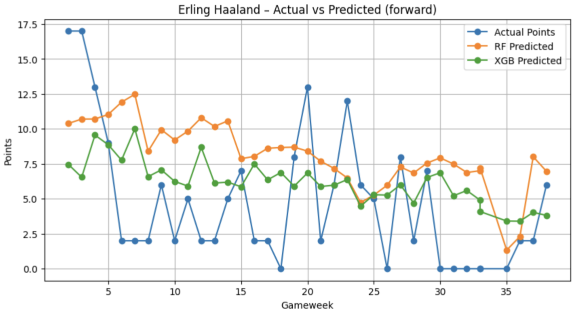

Figure 20. 2024-2025 Erling Haaland Predicted Points

Model	Total Points
XGBoost	31,257.07
Random Forest	32,514.40
Actual	30,420

Table 11. Total predicted points vs actual points for the whole 2024-2025 FPL season

Using all the models presented and the analysis of their predictions, the XGBoost model for each position was the more effective and superior model to predict a player fantasy point in the upcoming weekend. There was great consistency in improving the evaluation metrics as it obtained the lowest scores of in RMSE and MAE and a higher R2 statistic. The total predicted points in the whole 2025 FPL season was closer to the actual total cumulative fantasy points. As a result, this prediction model is used in the next objective in the project to optimize squad selection throughout the entire FPL season.

5. Squad Optimization

All of these models were built to make predictions about how many points a player will score in the upcoming match. These predictions and data sets are used to focus on the short term predictions, yet the season in Fantasy Premier league is long as there are 38 matches in a season. As the season is long, FPL managers have to keep up and do constant weekly updates to their squad in order to maximize their total points sum. In the first match week, FPL managers have many questions and decisions to make, such as choosing 15 players from a large selection of players that fit squad constraints and under a budget of 100 million dollars. Their selection of players are those they believe will perform individually well to get the most points.

After the selection of the 15 players to start the season, every week FPL managers are conflicted in choosing the best 11 players to start while leaving 4 players on the bench whose points are likely to not be accounted for. In addition, FPL managers have to select one specific captain in which they believe will perform the best in the upcoming match week as a captain's points points are doubled. These questions and decisions in which FPL managers are confronted are the purpose of the creation of the models as they predict how many points players will perform in the next match. These models can help in factoring which of the 11 of the 15 players a manager should be starting based on the players that have the highest predicted points and in choosing the highest predicted points player as their captain.

Therefore in this section, the XGBoost model came into use for the 2024-2025 fantasy premier league season. As the model was trained on 2022-2024 seasons of FPL there was no data leakage in this model's performance. The purpose was to use optimization within the XGBoost model as a tool to operate in a way similar to how FPL managers operate during the FPL season. An optimization algorithm will use the XGBoost predicted values of players to be informed and make decisions about squad selection that will attempt to optimize the maximum number of fantasy points weekly.

As a result, an optimization algorithm was constructed which implemented Mixed-Linear Programming since in FPL there are many constraints under squad selection such as only 15 players can be chosen and the total sum of the value of the squad must be under 100 million. In addition there are constraints in the starting 11 of valid formations based on the minimum and maximum players that can be in the starting 11. All of these constraints needed to be factor while making the algorithms that optimizes the amount of points a squad outputs in a given game week based on the XGBoost values of predicted points.

5.1. Algorithm 1: MILP Formulation for Optimal Squad Selection

The set of all available players is P. For each player $i \in P$ we are given XGBoost predicted points and value:

$$\text{points}_i \geq 0, \qquad \text{value}_i \geq 0, \qquad \text{position}_i \qquad \text{team}_i.$$

We will use a binary decision variable

$$x_i = \begin{cases} 1, & \text{if player } i \text{ is selected into 15–man squad,} \\ 0, & \text{not selected.} \end{cases}$$

Step 1: Optimization to create a 15-Man Squad

Objective: Maximize total predicted points:

$$\max \sum_{i \in P} \text{points}_i \, x_i.$$

Squad Size Constraint.

$$\sum_{i \in P} x_i = 15.$$

Position Constraints.

$$\sum_{i \in P:\text{position}_i=\text{GK}} x_i = 2,$$

$$\sum_{i \in P:\text{position}_i=\text{DEF}} x_i = 5,$$

$$\sum_{i \in P:\text{position}_i=\text{MID}} x_i = 5,$$

$$\sum_{i \in P:\text{position}_i=\text{FWD}} x_i = 3.$$

Budget Constraint.

$$\sum_{i \in P} \text{value}_i \, x_i \leq B,$$

where $B = 100$ (the FPL budget).

Team Constraint. For every club t,

$$\sum_{i \in P:\text{team}_i=t} x_i \leq 3.$$

The solution of this MILP yields a 15–player set

$$S = \{\, i \in P : x_i = 1 \,\}.$$

Step 2: Pick best Starting 11

Binary variables used again

$$s_i = \begin{cases} 1, & \text{if player } i \in S \text{ is selected to start,} \\ 0, & \text{bench player.} \end{cases}$$

Objective.

$$\max \sum_{i \in S} \text{points}_i \, s_i.$$

Total Starters.

$$\sum_{i \in S} s_i = 11.$$

Formation Constraints

$$\sum_{i \in S:\text{position}_i=\text{GK}} s_i = 1,$$

$$\sum_{i \in S:\text{position}_i=\text{DEF}} s_i \geq 3,$$

$$\sum_{i \in S:\text{position}_i=\text{MID}} s_i \geq 2,$$

$$\sum_{i \in S:\text{position}_i=\text{FWD}} s_i \geq 1.$$

It produces the optimal starting lineup

$$L = \{\, i \in S : s_i = 1 \,\}.$$

Step 3: Captain Selection

Let

$$c = \max_{i \in L} \text{points}_i.$$

The captain's actual points after the match are doubled.

Step 4: Substitution Rules

After a game week concludes, we know players who played 0 minutes, let minutes_i be the actual minutes played.

A substitution of a benched player j for a starting player i is made if:

$$\text{minutes}_i = 0,$$

and the resulting formation after replacing i with j still satisfies

$$\#\text{DEF} \geq 3, \qquad \#\text{MID} \geq 2, \qquad \#\text{FWD} \geq 1.$$

The official FPL bench ordering rules are used:

- GK must be placed last on the bench.
- Outfield players substituted based on ordering .

The first algorithm constructed used the XGBoost's predicted values of every game week to create a new squad every time. To further explain, in an upcoming game week, the XGBoost model produces predictions of fantasy points for each player. Under the constraints of 15 man squad selection,number of players per position, and budget, it finds the best 15 player squad selection from the many total combinations based on the predicted values. With the 15 players, the algorithm goes through another optimization problem as it outputs the most optimal starting 11 that fits under the formation constraints. The rest of the four players are inserted into the bench. They are ordered from highest predicted points to lowest as these players can be substituted for those who did not play in the starting 11. In addition, from the starting 11 it chooses the player with the highest predicted points to be chosen as the captain. The algorithm does this for every week in the 2024-2025 season and it creates a whole new squad every match week using the XGBoost prediction model.

The results of this algorithm led to a total of 2,467 points for the whole FPL season of 2025. This would have ranked the XGBoost model with the optimization algorithm of around the 260,000 rank among the 11.5 million FPL managers that participated in the 2025 FPL season. This would put the model in the range of the top 2.26 percent of the best performing managers. This result shows that the XGBoost model is an effective model that can be used as a tool to rank players weekly based on their predictions. As the algorithm used the prediction of XGBoost to help guide its optimization of squad selection, the XGBoost model was able to consistently present players that actually performed well in the next match as the total season points was high.

A second optimization algorithm was created that took a more realistic approach to FPL manager simulation. As FPL managers, are given one free transfer per week, they can only substitute a player not a whole squad that was done in the first algorithm. In this algorithm, it ran the same optimization constraints as the previous algorithm to construct the squad for the first week and pick the best starting 11, best captain, and ordered bench. After this squad selection, every week the algorithm decided whether a transfer was made for the next match. In order to make this decision, a greedy approach was used. Every single player in the 15 man squad and its predicted value for the next match was compared with all the non-selected players in respect to their position. It computed the predicted gain which was the predicted points of the non selected player minus the predicted points of the player in the squad. While this search was made it still checked to follow the constraints of the FPL rules. It then picked the maximum predicted gain in hopes that it would improve the overall total points of the squad for the upcoming game week.

This greedy approach to perform transfers every week resulted to an overall performance of 2216 points in the 2025 FPL season which is decrease to the earlier model. It would have been placed in the 2,500,000 rank among the top 21.73 percent. There are many factors that would have led to a decrease in performance in this simulation. One factor would be due to the greedy transfer that was made. It only looked at predicted points for the next game but not for the next few games. Another potential player who may have not had the highest predicted value in the next match could have had higher predicted points over the course of the next four matches for example, representing that they will perform better in the long run. Therefore this limitation of the greedy transfer leads another factor which is the limitation of the XGBoost model. As this model only predicts fantasy points for the next game week, it does not further predict in the future such as the average predicted points in the next four games. Thus doing a transfer can only look at predicted points in the next game week and limiting the usability and impact of the XGBoost model for long term squad planning.

The result of both algorithms show the power but also limitations that the XGBoost model presents in their predicted fantasy points. The points predicted for the next match have shown to perform well showcasing players every week that actually perform well in their next match based on their predicted values. This showcasing can help in features in FPL such as Triple Captain and Free hit that are only used for the next match. Yet its limitations are shown for FPL managers for transfers as they cannot forecast points that further in the future for long term management.

6. Limitations

Despite the performance of the XGBoost and random forest models, they are faced with difficult situations related to sports causing them to have certain limitations. One limitation that these models faced was the high degree of randomness and variability in football, especially as these models predicted points for every game week. The randomness and variability can average out over a few games but in any given week a player's performance can drastically change. The use of features that focused on form was a way to tackle this problem but it cannot fully resolve it, therefore there is led to error greater than one in the RMSE metric. Models are limited to decrease this error even further as sports do not follow a formula that guaranties fantasy points.

Another factor is that these models cannot predict sudden rotations or injuries in football. An injury just before the game week cannot be learned based on the features as they would have not played a game so there would be no indication. It would take at least a week for the models to recognize an association with injury based on the player's minutes being equal to zero but there is no feature that states before the game week if a player is injured or not. To add on, rotations occur during a long season in football to give players rest. A coach can place consistent starters on the bench for one match and the model cannot predict this as these are game time decisions from the managers.

Finally these models only predict player points for the upcoming match and are limited in the fact that they do not predict over the course of the next few games. This type of prediction would allow a different use for FPL managers that prioritize long term management over their squad.

7. Conclusion

The study of this project was to build models that predicted a player's total fantasy points for the upcoming game week in order utilized as tools for FPL managers in their decision making for squad selection. The first thing that was done in order to prepare the models was to separate the datasets of FPL statistics based on player positions. This forced every position to have their own dedicated models. The three main models created were Linear Regression, random forest, and XGBoost. In this order, the improvement of the models were seen through metrics of accuracy such as MAE and RMSE, in which the random forest improved on Linear Regression, and finally the XGBoost models improved over random forest. The XGBoost models predicted the whole 2024-2025 FPL season and produced RMSE of 2.2 for forwards, 1.942 for midfielders, 1.83 for defenders, and 1.603 for goalkeepers.

The construction of models for each position were important for this prediction seen through the SHAP values of the ensemble methods. In the top 10 important features based on the SHAP values of each position XGBoost model, it was presented while there were features that were shared, there were also features not as impactful in other positions. For example, creativity was a top 10 feature for forwards that was not in the SHAP midfeld model despite both positions being reliant on points based on their offensive productions. The improvement of XGBoost models over random forest was further seen in certain FPL players and overall value of predicted points in the season. Random Forest was over predicting fantasy points for players more than the XGBoost while XGBoost even though also over predicted was more balanced and closer to the actual fantasy points. Therefore, it proved to be the best model for fantasy points predictions compared to Linear regression and RF.

As it predicted every player's performance for every game week, the purpose of the model was to help managers use it weekly in their squad selection and transfer decisions. Therefore, through the MILP and optimization algorithm we simulated a full season of 2024-2025 Premier League using the XGBoost model's predicted values. In our first algorithm of creating the best squad every week it produced 2,467 points, while the other algorithm which found the best greedy transfer produced 2,216 points. This showed the XGBoost model's power to effectively every week understand the set of players that will actually perform well based its predicted values. Yet using this model for long term decisions led to more difficulties and limitations because of the lack of ability to do many changes and instead have really small gains.

The result of this project shows there is great potential in expanding predictive modeling of FPL based on the promising results that these models produced. There was value within using these model's predictions to create optimal squad selections in the upcoming weekend which led to positive results. There are many different avenues and approaches to expand this project depending on the overall goals. One approach is to assign probabilities within ranges of predicted fantasy points. A model can give a player different ranges of predicted points and the probabilities associated to provide another viewpoint of how a player will perform in the upcoming match. This would put an emphasis on using the model to analyze players based on their distribution and analyzing them as risk versus reward. Another approach would to improve the problems in the second algorithm, which is to expand the XGBoost's predictions over a couple of games. This would allow for FPL managers to be able to choose transfers based on long term investments. The accuracy of the models can improve as the player's actual points will have less variability in comparison to a week in and week out performance. Altogether these approaches can be used to expand the usability and accuracy of the XGBoost model to help FPL managers around the world in their decision making every week.

References

[1] Vaastav Anand. *FPL Historical Dataset*. Retrieved November 2025 from `https://github.com/vaastav/Fantasy-Premier-League/`.

[2] Tamimi, M., and T. Tran. “Players’ Performance Prediction for Fantasy Premier League, Using Transformer-based Sentiment Analysis on News and Statistical Data.” *International Journal of Computer Science in Sport*, vol. 24, no. 1, 2025, pp. 133–147.

[3] P. Pokharel, A. Timalsina, S. Panday, and B. Acharya. *Fantasy Premier League - Performance Prediction*. Proceedings of the 12th IOE Graduate Conference, vol. 12, October 2022. ISSN 2350-8914 (Online), 2350-8906 (Print).