How to Build your Initial Fantasy Premier League Squad via Machine Learning

Ahmed Almezail

05/09/2022

Abstract

There are 8 million people playing Fantasy Premier League (FPL), and there is a huge interest in data analysis from FPL manager to outperform their rivals. It is always hard to build your 15 players squad or to choose which player to transfer in or out in a weekly basis. There have been several websites who offer predictive modeling to their subscribers to predict which player to select. Thus, in this report, our main goal is to build a predictive model that could be used by FPL managers to create their initial squad before the start of the season based on the last season data. We were able to predict the total points of players in FPL by using Random Forest model. The training data was for 19/20 season then we tested the performance on 20/21 season data. The root mean squared errors were quite low and show that we could rely on the fitted values. In addition, we built new metrics based on our predictions to present the data in a a table and graphs to infer it. Table 1 present the mean and standard deviations of different metrics by the player position. On the other hand, the graphs recommended the best options players to choose for every position based on different metrics. The graphs are flexible to the FPL manager constraints and preferences by filtering out players that do not meet the preference conditions. Then we suggest that this report could be extended in the future by using a KNN model for the week by week analysis to recommend the transfer in and out options. We believe KNN models could capture the "hot hand" effect of players, or in other words, the inform players.

Introduction

Background

Fantasy sports games are in which participants assemble an imaginary team of real-life footballers (athletes) and score points based on those players' actual statistical performance or their perceived contribution on the field of play. Usually, in a particular fantasy game, players are selected from one specific division in a particular league. However, Fantasy Premier League (FPL) is the Fantasy Soccer game for the English Premier League in England. FPL is a global online-based platform that allows you (the manager) to select players from the Premier League, who you expect will perform well in a certain "Gameweek" (weekly round of games) and would generate the maximum points to you. The points of the players are measured based on different statistics such as goals, assists, and clean-sheets, which the points are awarded accordingly.

The original version of what is now Fantasy Premier League was created in England by Bernie Donnelly on Saturday 14 August 1971. However, the recent growth for the game itself have been increasing sharply; from around 3.5 million players during the 2014/15 season to 8 million in 2021/2022 season. That crowns the Fantasy Premier League to be the biggest Fantasy Soccer game in the world and second biggest fantasy sport in general.

A very understandable question here would be: What makes this game so attractive and addictive? Well, first of all, it is the inherent allure of soccer, the English Premier League and the fantasy of managing your own team in a weekly basis for a whole season, all combined. Add to this, mini-leagues that you could play with your friends and the bragging rights that come if you can win these mini-leagues or the FPL itself (if you can manage it, that is). To top it all off, with the amount of money that revolves around the English Premier League, the Premier League decided they could afford to invite the winner of the game for a 3-day fully funded trip to England with the honor of watching a live match of their favorite team to go along with it. However, only one in a 8 million will get this prize. So, apparently the rivalry in the mini leagues and the media coverage are the main growth to FPL.

Rules of the Game

Now let's demonstrate the rules of the game to get the full picture of it. First of all, you will play a full season which consists of a 38 gameweeks (rounds), and generate points in a weekly basis based on your selected players. Your will start with a constraint budget to build a squad which consist of 15 players which are:

- 2 Goalkeepers (GK)
- 5 Defenders (DF)
- 5 Midfielders (MID)
- 3 Forwards (FWD)

and you will select your starting 11 who you think will generate the maximum points with a minimum number of players in each position (i.e minimum of 3 defenders), while the other 4 would be in the bench and their points would not be counted to your weekly points. In addition, you will be able to pick a captain in which his point will be doubled. another important rule is you will start with £100.00 million squad budget and every player has his own price, knowing that prices changes every week based on FPL managers transfers activities. You can't pick more than 3 players from the same team. Finally, you can do 1 transfer after every Gamweweek. For the detailed rules, please look at appendix I.

Why we are doing the project

There are plenty of data scientists who capitalize on the game. For example, Fantasy Football Fix! and FPL Review are websites with a paid subscription that provide a tool-set for FPL managers such as useful data, strategies, and predictive models to predict the points of the players in a weekly basis. So, since we have been FPL managers for several seasons, we decided to utilize our data science skill-set to create a model that predict the total points of players based on previous season, and do summarize some useful statistics to help FPL managers build their initial squad before the start of a new season.

Methods

Data

We used a well known free source for the data which is used usually by FPL data analysts. It is a github account that has data from 16/17 season to 21/22 github.com/vaastav. Since the current season (21/22) has not ended yet, we decided to use 19/20 and 20/21 seasons data for our model. Our plan is to use 19/20 season data as training set, and to test our accuracy based on the 20/21 season data. The data for each season is on aggregate level, which means it sums the 38 Gameweeks data (week by week) and present it on full season aggregate form.

First, we used `cleaned_data_20` and `cleaned_data_21` data sets to do some data wrangling and mutate new columns from other data sets. Then, we break down each season data based on players' position, one includes only goalkeepers, and the other one includes the 3 other positions. We did that, since goalkeepers most likely would not generate points from scoring or assisting. So, to have better fitting, we let goalkeepers to be in a separate set to use it for a different model.

Second, we did some feature engineering, and create new variables which shown to be important in our model. we created the return on investment `ROI`, and return on minutes `ROM`. The first metric is basically `total_points/value` which evaluates if the player worth his price. On the other hand, `ROM` is the `total_points/minutes` which will help us to find the super-sub players.

Finally, after wrangling our data, we used 19/20 season data sets as training sets, and 20/21 season data sets as testing sets for our both models.

Training sets:

- `cleaned_data20_xgk`: 19/20 season data set for FW, MID, and DF
- `cleaned_data20_gk`: 19/20 season data set for GK

Testing sets:

- `cleaned_data21_xgk`: 20/21 season data set for FW, MID, and DF
- `cleaned_data21_gk`: 20/21 season data set for GK

Below are the variables in those data sets.

1. `goals_scored`: Total goals scored by the player
2. `assists`: Total assists by the player
3. `total_points`: Total points generated by the player
4. `minutes`: Total of minutes played by the player
5. `goals_conceded`: Total of goals conceded when the player is on the pitch
6. `creativity`: The player performance in terms of producing goalscoring opportunities for others
7. `influence`: The degree to which a player has made an impact throughout the season
8. `threat`: Value that examines a player's threat on goal
9. `bonus`: Total of Bonus points awarded to the player throughout the season
10. `bps`: Total of Bonus Points System which is the metric used to award players bonus points
11. `ict_index`: An index that combines `creativity`, `influence`, and `threat`
12. `clean_sheets`: Total of clean sheets by goalkeepers
13. `red_cards`: Total of red cards received by the player
14. `yellow_cards`: Total of yellow cards received by the player
15. `selected_by_percent`: Percentage of FPL managers selected the player at the end of the season
16. `value`: FPL value of the player at the end of the season
17. `ROI`: Return on investment `total_points/value`
18. `ROM`: Return on minutes `total_points/minutes`
19. `full_name`: The full name of the player
20. `id`: ID of the player used as a reference to connect between the data
21. `position`: 1 for Goalkeepers (GK), 2 for Defenders (DF), 3 for Midfielders (MID), 4 for Forwards

Table 1: Summary of the Averages and Standard Deviations by Position

	Residuals		ROI		Points above Replacement		ROI above Replacement	
Position	Average	ST DEV	Average	ST DEV	Average	ST DEV	Average	ST DEV
GK	3	11	23	48	72	34	15	7
DF	1	6	17	30	40	30	9	6
MID	4	7	15	25	43	36	7	5
FWD	2	8	16	27	61	44	8	5

Model

So we used the above data to predict the `total_points` of the players so you can build your initial squad for the next season upon that. In order to do our prediction, we used Random Forest model on the training sets, to predict the `total_points` of the testing sets. first, we used all the features as dependent variables to check the importance for each one. Then, based on that we adjusted the variables in our models according to their importance level (Figures 5 and 6 in Appendix II). In fact, we found that the root mean squared errors (rmse) for the training sets to be 2.3 (w/o GK) and 2.7 (GK), while the results of the predictions were 5.1 (w/o GK) and 6.1 (GK) for the testing sets. In simple words, it means on average, the margin of error for predicting the `total_points` of an FPL DF, MID, or FWD is 5.1 points and 6.1 for a goalkeeper. We have reported in Table 1 the averages and standard deviations of different variables for each position based on the predicted values to give us a clearer idea on the robustness of the model. However, the table would be discussed in the results section. In addition, we plotted the predicted values of 20/21 season on their true values (observations) and could be found in Figures 7 and 8 in Appendix III. A perfect prediction would plot the points as diagonal straight line. So, we were able to get very close to that since our rmse was relatively low.

We have also created 2 new metrics which going to be handy in the selection approach. The metrics are inspired by the Baseball statistic wins above replacement (WAR). There are many ways or formulas in how to calculate that, but basically it gives an idea on how many wins a team could get beyond an average baseball team. So, our first metric is basically how many predicted total points could a player generate over his replacement (an average player with the same position). On the other hand, the second metric is how more ROI we will get from a player over his replacement. So, we could say that we demeaned the values without standardizing it.

- `ppt_above_replacement`: Points Above replacements
- `ROI_above_replacement`: ROI above replacement

We used those metrics in presenting our final results eventually in figures 1, 2, 3, and 4. It could be also used for all Gameweeks data by showing how many times a player outperformed the average. However, our main goal is to help building the initial squad by using the aggregate data to predict the total_points. Thus, the week by week data is not in our interest for now.

Results

In this section we will present the most valuable information to the FPL manager based on our predictions. After predicting the points of the players for 20/21 season, we want to check how accurate we were for every position. First, in our summary analysis, we included only player who has `ppt_above_replacement` or `ROI_above_replacement` above zero to avoid biaseness of the statistical summary toward redundant players. In fact, the number of players shrunk from 713 players to just only 307 players. Then we calculated the average and standard deviation (stdev) for every position of different metrics as shown in Table 1. The average would help us understand where the central value lay, and the stdev will measure the volatility or how disperse the data from its average Therefore, standard deviation would be handy to understand whether players in a certain position clustered around the mean (low stdev) or the vice versa.

Table 1 Statitical Inference

First, let's analyze goalkeepers. The residuals average is pretty good, however its standard deviation is relatively high. This means that there are few good goalkeepers to choose from since the disperse is high from an average goalie. The same information could be derived when we look at the standard deviations of ROI which means they are few goalkeepers who worth their value and to be in your squad.

Second, looking at the defenders. We can see our model has the best prediction accuracy for them relative to the other positions. However, the ROI standard deviation is still relatively higher which would eliminate quite good number of defenders who worth their value. The average points and ROI for a defender would give us 40 FPL points and 9 returns above the average defender. Also, their standard deviations are quite relatively low, which means based on those metrics we will have plenty of defenders around the average.

Next, when we look at the midfielders, they have the highest average residuals, but it still consider to be low since it is suggesting that we have an average of prediction errors by 4 FPL points. Also, the dispersion from the mean is only 7 FPL points. The average ROI is the lowest relative to the other positions with the lowest standard deviation. This means on average, midfielders are overvalued compared to the other positions. However, that's because most the players in FPL are midfielders, so many of them do not play regularly because of the players rotations. In addition, midfielders could get you an average of 43 FPL points and 7 ROI above an average midfielder player with not high dispersion from the mean.

Finally, forward prediction shown solid precision with still good accuracy. The average points above replacement was the very high but with high standard deviation. That suggests you should be more careful on choosing your forwards. Only few would be a good pick for the initial squad which is quite the same to the goalkeepers.

Potential Players to Select Visualizations

In this section, we will show an example of how could an FPL manager select players based on our predictions and metrics. We produced a graph for every position that includes the name of the player, predicted points above replacement, ROI above replacement, and their value. The graphs could be modified based on the FPL manager preferences. It could be filtered to include only players below a certain value or any other preferences or constraint. However, in order to show our results, we have not filtered the values of the players, but only the predicted points above replacement to nominate only the best picks for an FPL manager's initial squad.

Figure 1 plots the best goalkeepers according to predictions and metrics. As we mentioned earlier in Table 1 Inference section, there would be only few good selections for a goalkeeper. According to our results, Emiliano Martínez and Illan Meslier are the hot picks here. They are the only goalkeepers with predicted points above replacements higher than 100 FPL points. In addition, the ROI above replacement for Martínez is higher than Meslier. However, Meslier price is lower, so it would be a matter of the FPL manager preferences and constraints. The point of the graph is to limit the choices based on statistical analysis.

Moving to Figure 2, we got a diverse selection choices based on various variables. Since an FPL squad requires 5 defenders, many veteran FPL managers follow a strategy to select 1 super defender who is going to be costly, 2 medium cost defenders, and 2 low with low cost. This graph would help to visualize the players based on the preferences and constraints which could be applied by filtering the players according to that. For example, Trent Alexander Arnold would the best choice for a super defender. Stuart Dallas and Matt Targett could be solid picks of defenders for a medium level budget. Luke Ayling is the best option for a budget defender. More filters could be applied to suit the FPL manager strategy.

Figure 3 nominated the best selection options for midfielders based on our analysis. There is a wide range of good options to choose from. For example, Mohamed Salah, Bruno Fernandes and Heung Min Son are the best super players options for your squad, which would elimenate the other players who are not shown in the plot or even Sadio Mané and Kevin De Bruyne. The graph also helps us to predict players which are going to receive similar points to more expensive players. For example, James Ward Prowse and Ilkay Gündogan are relatively cheap midfielders and could get the same amount of points as fancy players such as Marcus Rashford, Mason Mount and even Kevin De Bruyne. Tomas Soucek and Matheus Pereira are the hidden gems for this position. Their cost are ridiculously low compared to the predicted points they could generate.

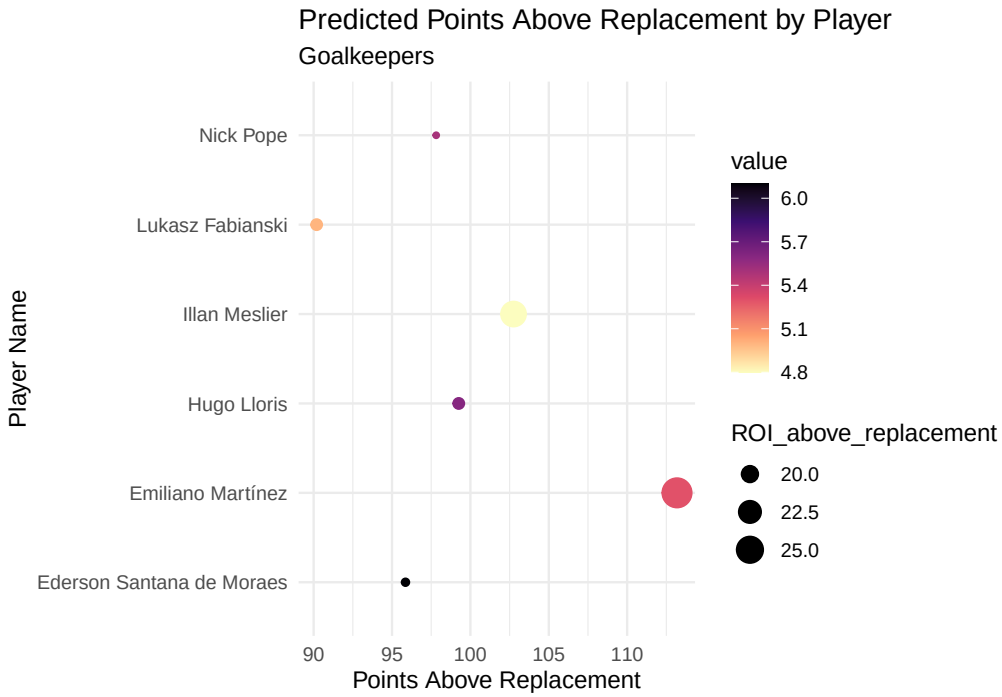

Figure 1: Predicted Points Above Replacement by Player (GK)

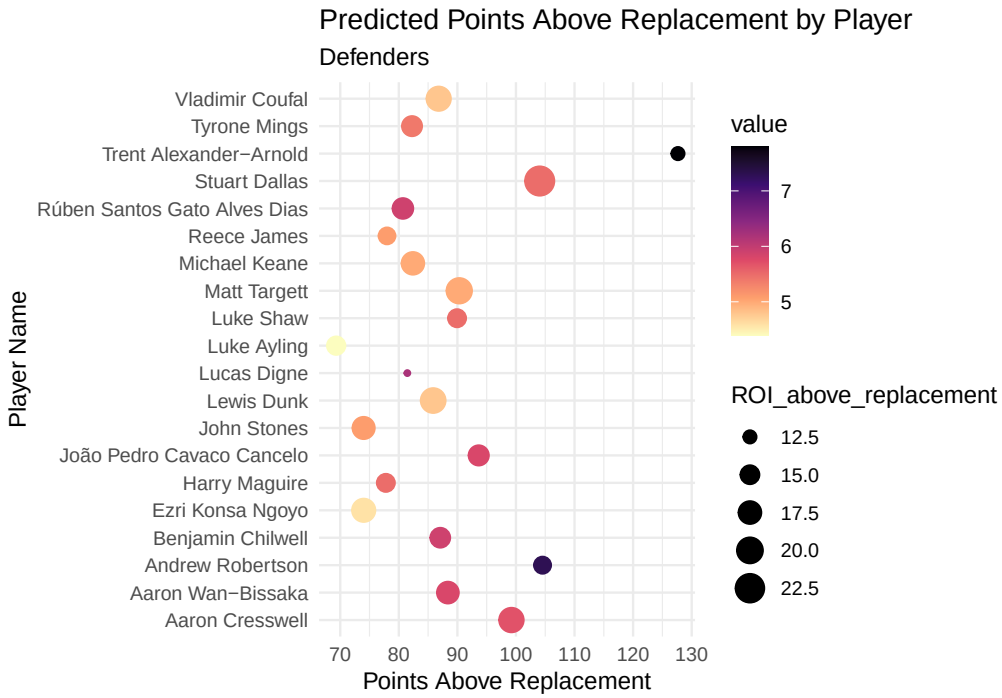

Figure 2: Predicted Points Above Replacement by Player (DF)

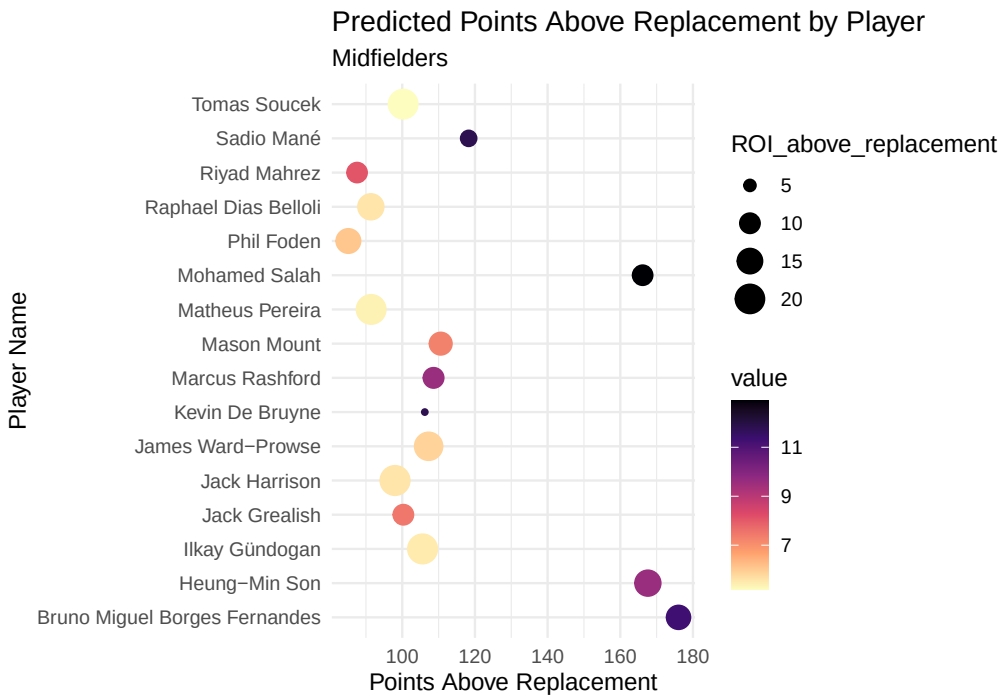

Figure 3: Predicted Points Above Replacement by Player (MID)

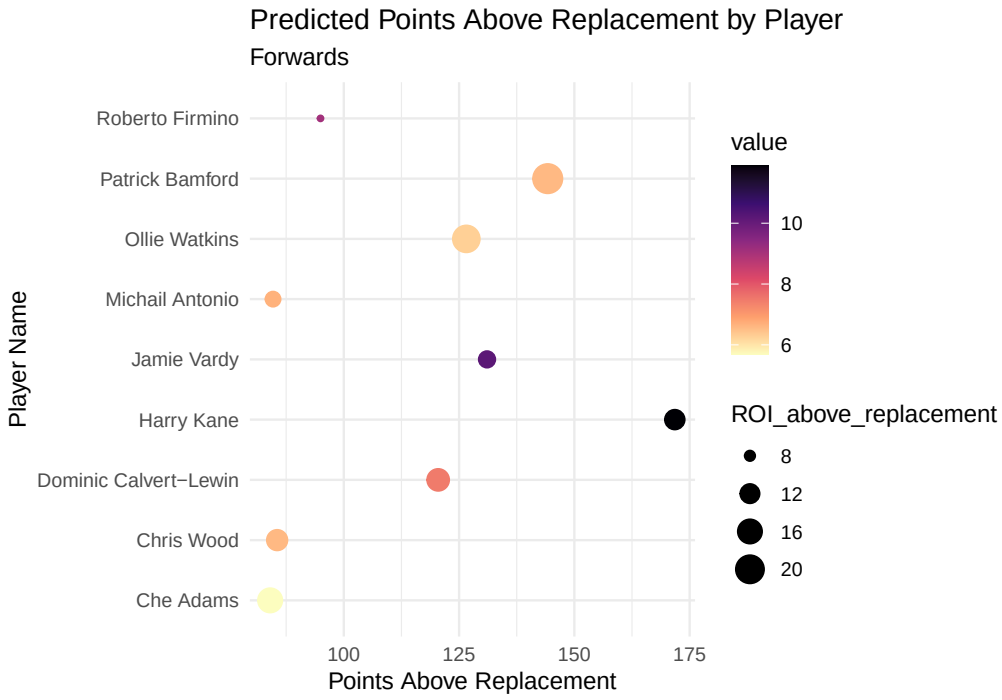

Figure 4: Predicted Points Above Replacement by Player (FWD)

Finally for Figure 4, as we thought earlier, the best forwards would be few since the standard deviations of their metrics were relatively high. Harry Kane would be the best option for a super forward. We can see that some lower cost forwards could generate the same amount of points or higher than a more expensive forwards. For example, Patrick Bamford and Ollie Watkins would be the favored options over expensive players like Jamie Vardy Roberto Firmino. Moving to the hidden gems, Che Adams would be the winner.

As we can see from all of these graphs, they provide very solid information from just having a look to it. However, we though about an FPL manager unable to take an educated decision in some situation because of the limited budget. Should you upgrade your chosen player or not? which position should you upgrade for? For this case, we developed a helpful metric that could be useful for these kind of situation. We call it the marginal returns per £ or the Premium. So, in this case you would take the difference between the best player and the 2nd best in terms of `ROI_above_replacement` dividing by the difference of their `value`. The higher the number would be the better option or the tie breaker.

Conclusion

In conclusion, this report helps to predict the total points of FPL players based on previous season, and infer the results and the statistical numbers. So, this model would help FPL managers to get ready for the new season by utilizing the results and methodology used in this report. First, we used the 19/20 season data as a training set, and tested the performance of our model on 20/21 season data. The performance of the model was pretty solid and was able to get low rmse for the both models. Finally, we were able to present our outcome in a summary table (Table 1) and in for different graphs for each position which help to improve the decision making of the squad building for an FPL manager. We note that the performance of the players could vary across the season and the difficulty of the game could differ as well. However, this model would be great to use as a helpful tool to the FPL managers when building the initial squad taking under consideration factors not captured in the model such as new players coming to the premier league, injuries, management changes in teams, or even the one season wonder players. After that, a KNN model could be built to predict the points on a weekly basis based on certain number of previous matches. The KNN model would capture the “hot hand” effect, which could be critical for recommending which player to transfer in and out.

Appendicies

Appendix I: Detailed Rules of FPL

1: The premise:

(a) The player is given a budget of £100 million to buy a squad of 15 players consisting of 2 goalkeepers, 5 defenders, 5 midfielders and 3 forwards with the added rule of being able to select a maximum of 3 players from any single team.

(b) The team can play in any formation providing that 1 goalkeeper, at least 3 defenders and at least 1 forward are selected at all times.

2: Captain and Vice-Captain:

(a) From the starting 11 a captain and a vice-captain should be nominated.

(b) The captain's score will be doubled. If the captain plays 0 minutes in the Gameweek, the captain will be changed to the vice-captain.

(c) If both captain and vice-captain play 0 minutes in a Gameweek, then no player's score will be doubled.

3: Prioritizing Bench for Automatic Substitutions:

(a) The substitutes provide cover for unforeseen events like injuries and postponements by automatically replacing starting players who don't play in a Gameweek. Playing in a Gameweek means playing at least 1 minute or receiving a yellow / red card.

(b) Based on the priorities assigned, automatic substitutions are processed at the end of the Gameweek as follows:

i. If the Goalkeeper doesn't play in the Gameweek, he will be substituted by replacement Goalkeeper, if he played in the Gameweek.

ii. If any of the outfield players don't play in the Gameweek, they will be substituted by the highest priority outfield substitute who played in the Gameweek and doesn't break the formation rules (e.g., If the starting team has 3 defenders, a defender can only be replaced by another defender)

(IV) Transfers:

(a) There is an option to transfer one player out from the team and bring in another one of the same position for free. Any additional transfer incurs a 4-point penalty.

(b) If free transfer is not used, you an additional free transfer is allowed the following Gameweek. If the saved free transfer is still not used in the following Gameweek, it will be carried over until it is used. Maximum of 1 saved transfer is permissible.

(V) Player Prices:

(a) The price of a player is predetermined by the game developers, based on the player's popularity and performance in the last football season. However, price of the players changes after very Gameweek based on fantasy managers transfer activity.

(b) A point worth noting is that a player cannot rise in price by more than £0.3m, in a single Gameweek.

The objective of the game is to have the best-scoring team of players possible every week given the budget and other constraints so that in the long-term, you can accumulate the most points among other players.

Appendix II: Importance of the Variables

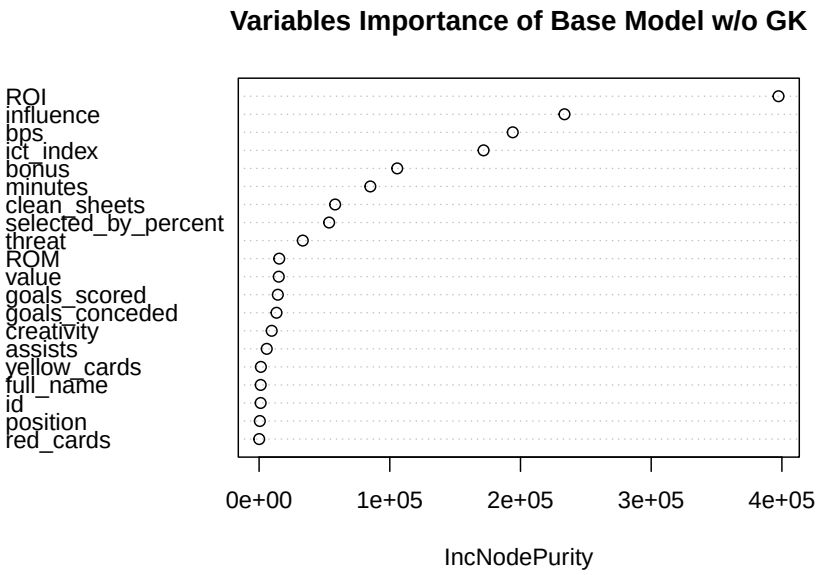

Figure 5: Random Forest Base Model (DF, MID, FWD)

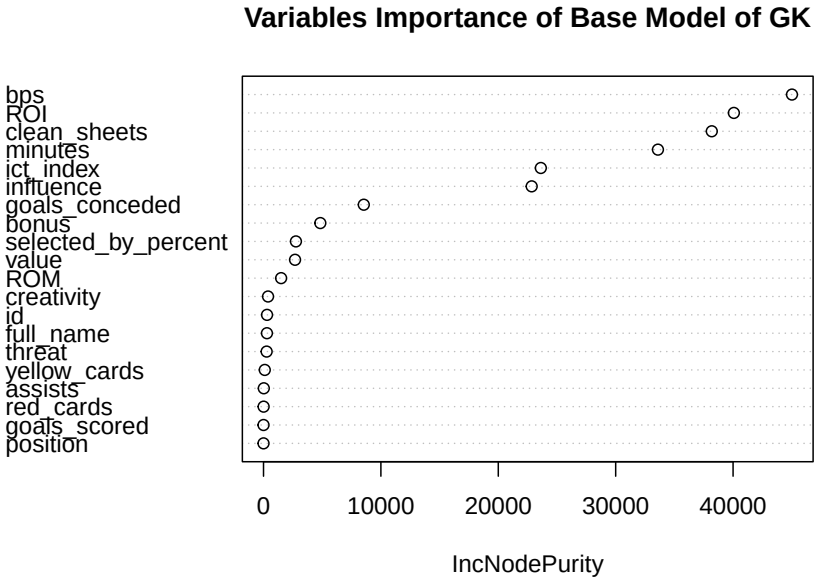

Figure 6: Random Forest Base Model (GK)

Appendix III: Plot of Predictions vs Observations

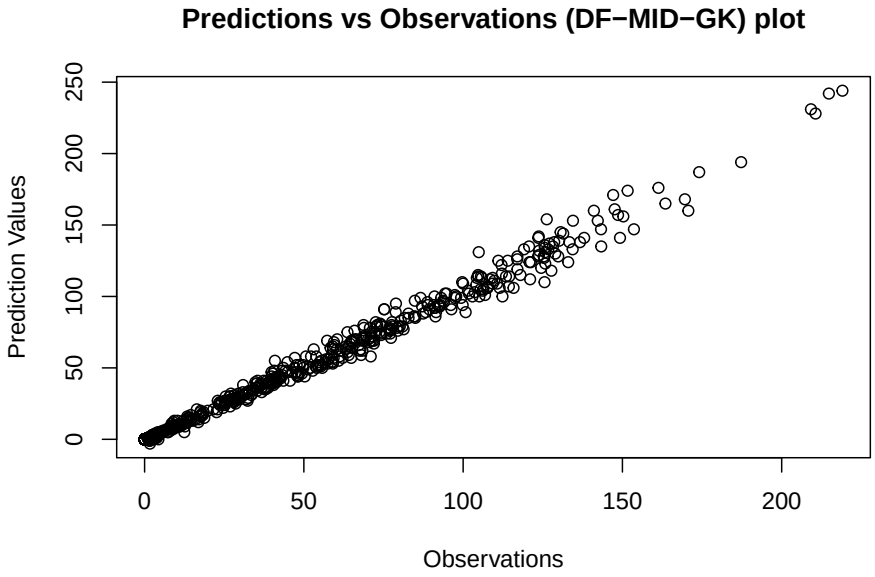

Figure 7: Random Forest Base Model (DF, MID, FWD)

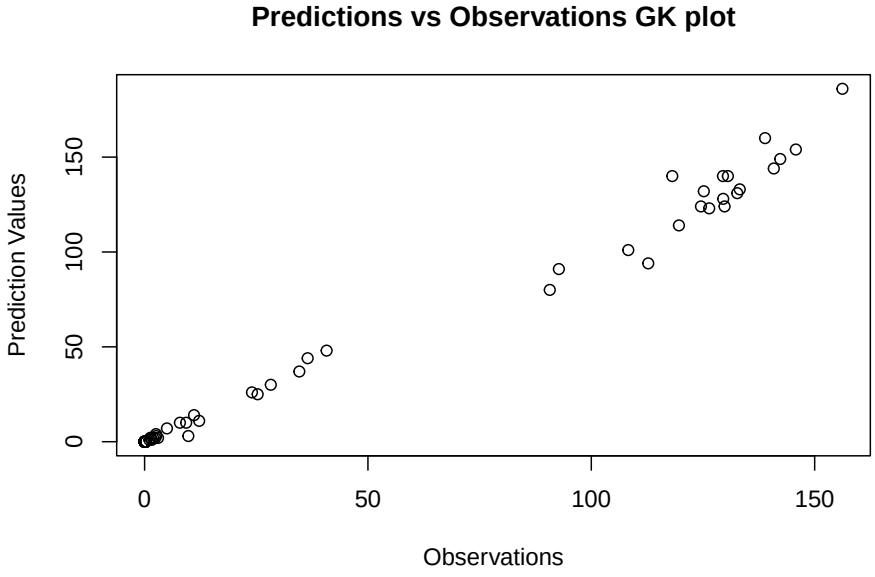

Figure 8: Random Forest Base Model (GK)