Predicting the best 2022/23 fantasy premier league team per game week

Austin Byrne[a]

[a]*Stellenbosch University, South Africa*

Abstract

This project aims to predict the best possible fantasy premier league 11 for each game week of the 2022/23 premier league season. The machine learning model that will be used is a random forests model. To perfect the model, hyper parameter tuning will be conducted to obatain the optimal parameter values.

Table of Contents

Email address: 22582053@sun.ac.za (Austin Byrne)

1. Introduction

The aim of this project is to create predictions on how many fantasy premier league points premier league players will score for each game week. Using these predictions I will be able to select the best fantasy premier league team which consists of 11 players. The fantasy premier league team restrictions make this more difficult than just selecting the top 11 players with the highest predicted points in a game week. The restrictions consist of only being allowed to have a maximum of 3 players from the same premier league team, your fantasy premier league team can not have a combined total value that exceeds 100 million and your team has too stick to certain formations which constrains the amount of players you can have in each position.

The machine learning model that will be used for these predictions is a random forests model. The layout of the project consists of data analysis section. Next, a base line random forests model will be created and evaluated by calculating the mean absolute error. Next using hyper parameter tuning the optimal values for the mtry, k-fold and ntree parameter values will be found. Next a new, tuned random forests model will be created and it's model performance will also be evaluated via the mean absolute error. Next, the best model will be used to create predictions on the amount of points that will be scored by each player in the 2022/23 premier league season. From these predictions a function will be created and used that will select the best fantasy premier league team to select per game week for the 2022/23 premier league season given the fantasy team restrictions.

2. Data exploration

The data sets used

The data used for this project are two data sets one for the 2021/22 premier league season and another for the 2022/23 premier league season. The data used is from the github of vaastav/Fantasy-Premier-League. The data consists of variables such as the name of each player in the premier league for that season, which team they play for, which position they play in, each players individual creativity scores, threat scores, how many goals they scored, how many goals they conceded, the game week, value etc.

The 2021/22 data set will be used to train and test the model and the predictions will be made on the 2022/23 data set.

2.1. Correlation plot between player attributes and total points

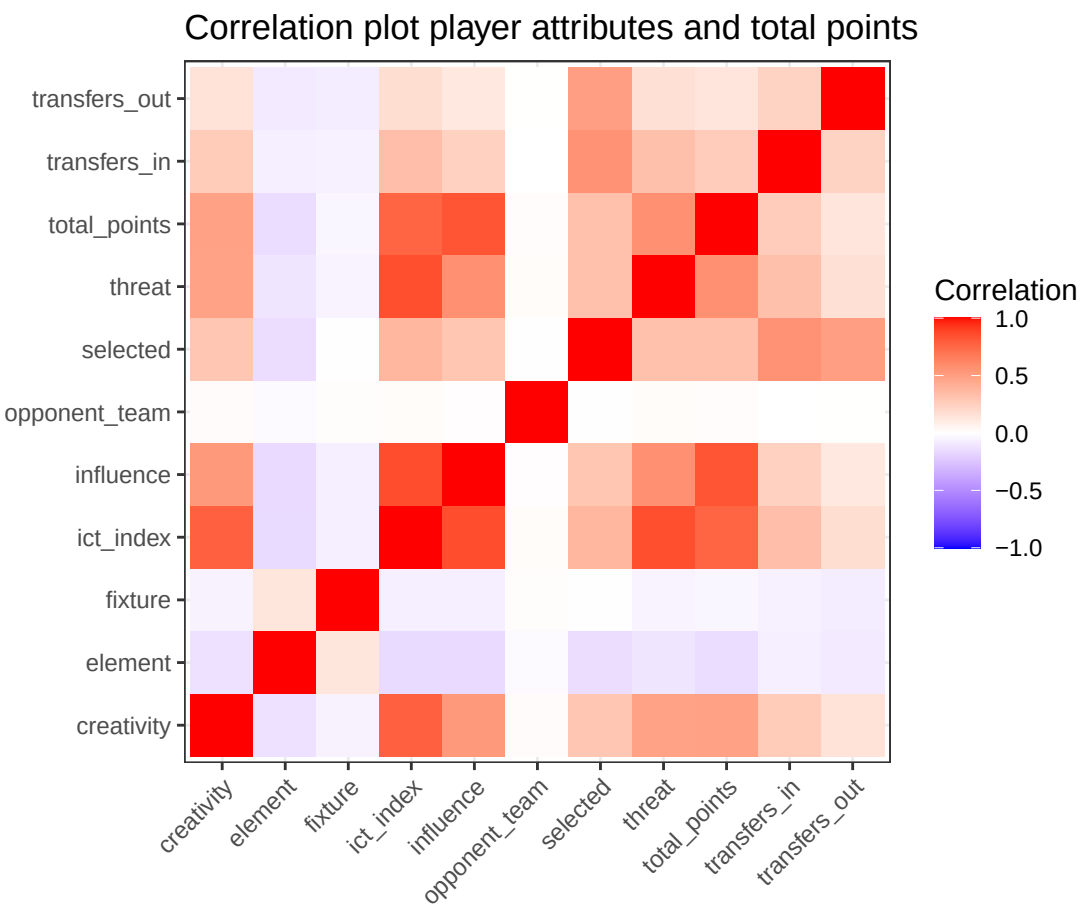

Figure 2.1: Player attributes correlation plot

The above 2.1 represents a correlation plot of player attributes for the 2021/22 premier league season for the purpose of identifying which player attributes have the biggest impact on total points earned by players. This is valuable information as we can then add these most prominent variables into our random forests model as features in an attempt to obtain accurate prediction results. It is important to understand that only variables that can be obtained before a game week must be evaluated.

From 2.1 it is evident that a players influence score, ict_index, threat, selected, transfers in and creativity scores are positively correlated to the total points variable, which is the variable we are trying to predict (the target variable). Thus, when building our random forests model these variables must be added. Furthermore, we must now look at the correlation between total points and team attributes.

2.2. Correlation plot between team attributes and individual player points

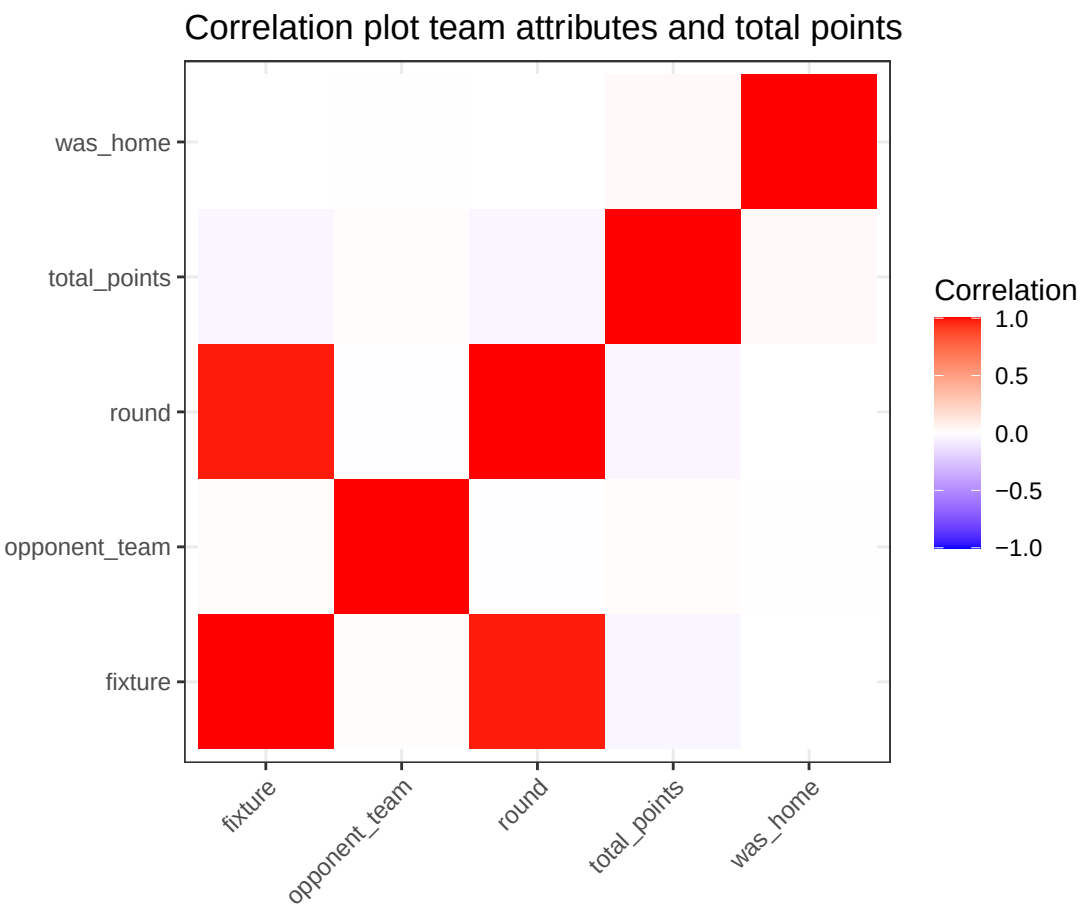

Figure 2.2: Team attributes correlation plot

2.2 represents the correlation plot for team attributes for the 2021/22 season such as whether the team is playing at home that game week, the team they are playing against and which game week it is for that specific match.As can be seen from 2.2 the team attributes actually don't have much of a correlation with the total points scored by players other than the slight correlation between the round and fixture with total points. However, this correlation seems so small that it will irrelevant to add into our model.

2.3. Plotting the average points scored per position per gameweek:

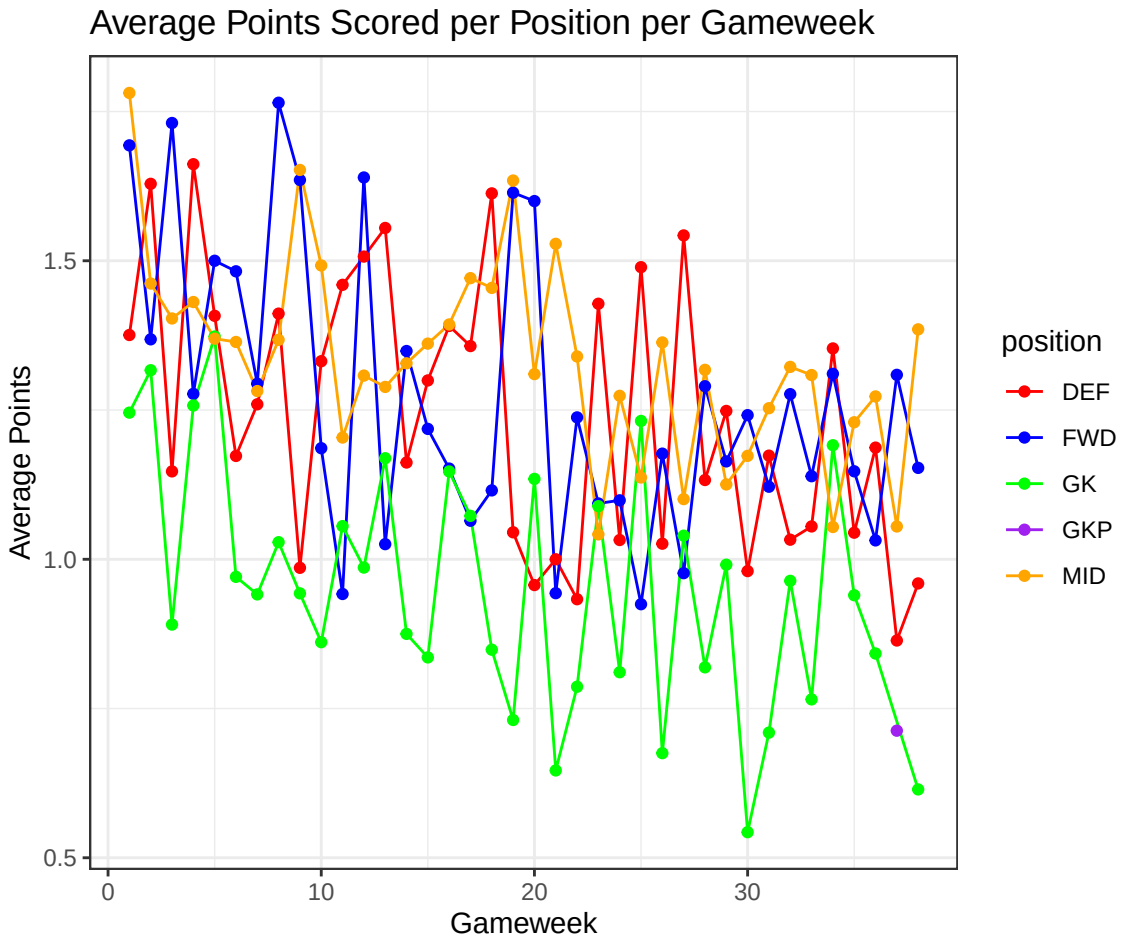

Figure 2.3: Average points per position combined plot

The above figure 2.3 provides some very valuable information on how many points are scored per position per game week throughout the 2021/22 premier league season. Each point on the above graph represents the average points scored for a position (either DEF which represents defenders, FWD which represents forwards, GK which represents goal keepers, or MID witch represents midfielders) for that particular game week. Some important takeaways can be found in this graph.

Goalkeepers on average seem to score the lowest points per game week, while strikers and defenders seem to have the highest variation and unpredictability. Midfielders seem to have on average the highest average points throughout the entire 2021/22 premier league season with the lowest variation. In order to obtain a clearer picture of the distribution of points per position the following figure will split the positions into their own respective plots.

2.4. Now plotting the average points per postion on differetn axis and plotting the overall average points scored per postion.

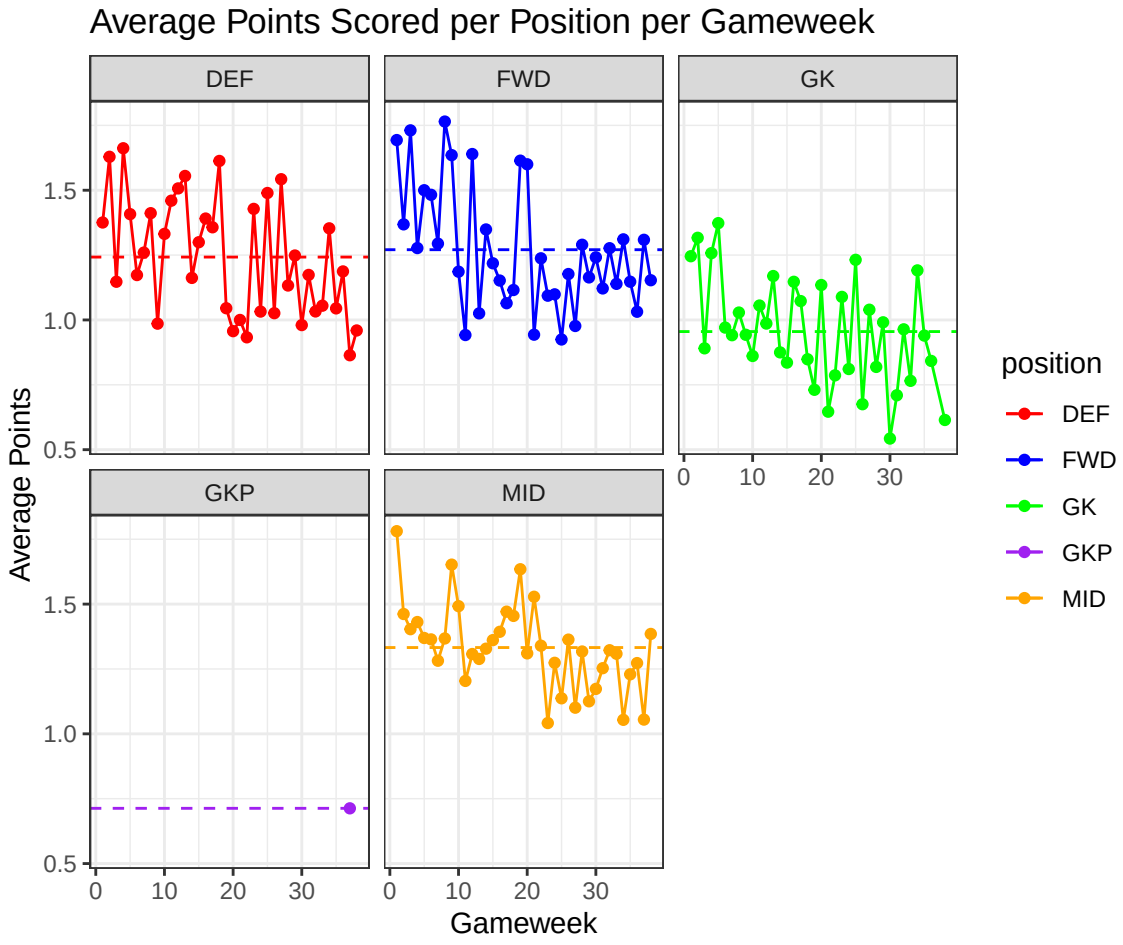

Figure 2.4: Average points per postion individual plots

2.4 provides a clearer picture of the distribution and confirms the takeaways made in response to 2.3. 2.4 places each positions average point distribution throughout the 2021/22 season onto their own plot but with the same axis and places a horizontal line on the plot that represents the overall average for that position. It can be confirmed that goal keepers do in fact score on average the lowest points per game week, defenders and strikers have a higher variation and midfielders seem to be the best bet to obtain a higher probability of receiving higher points throughout the season.

The takeaway that can me made from the evidence of 2.3 and 2.4 is that you will want to have as many midfielders in your team as possible to take advantage of their higher average points and lower variance.

Next lets look at the most valuable players.

2.5. Scatter plot for points per value:

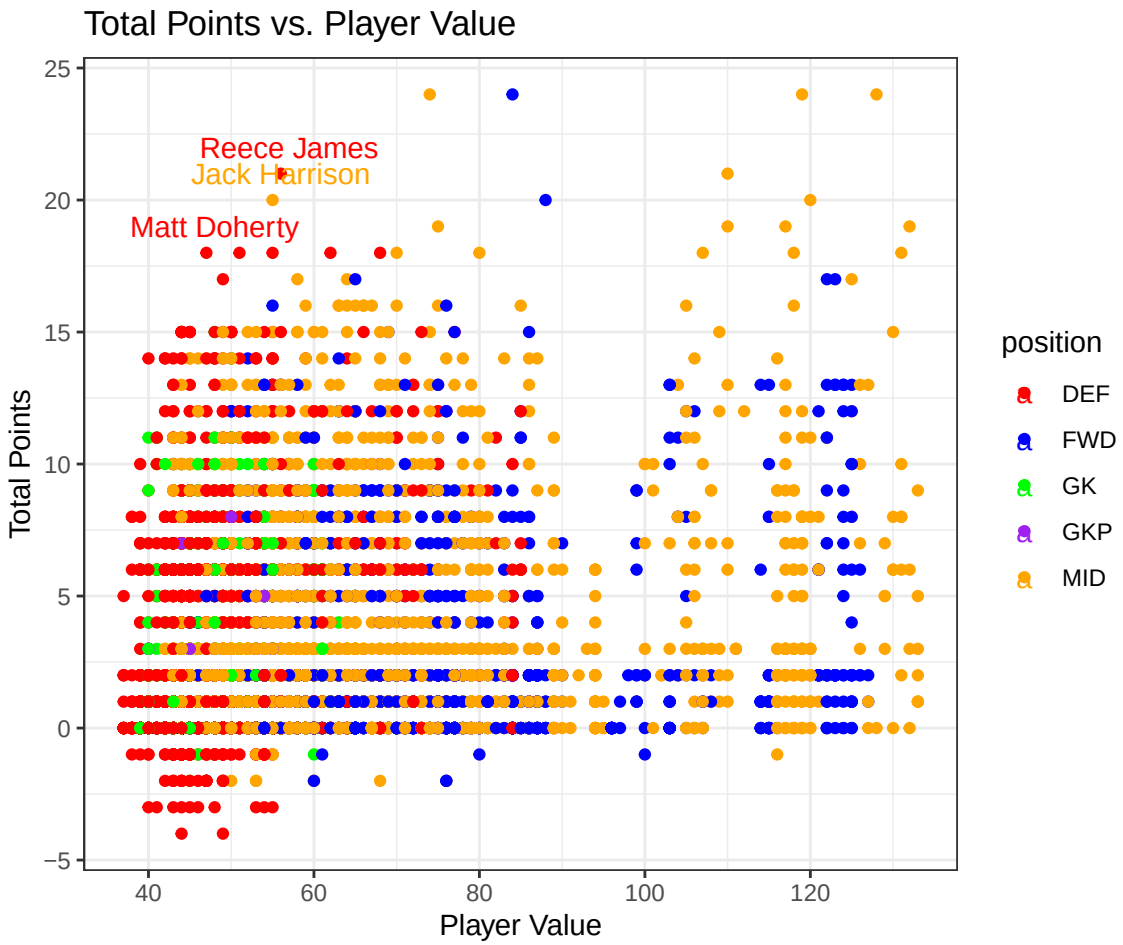

Figure 2.5: Scatter plot of players points per value

The above scatter plot 2.5 provides us with some more valuable insight. On the y axis we have the average total points per player and on the x axis we have the players value. Each point on this scatter plot provides us with the points per value statistic. In your fantasy premier league team you will want players with high points per value ratios. In the plot it is visible that defenders are much cheaper than midfielders and forwards. Thus, getting some good cheap defenders in your team may be beneficial.

Furthermore, this plot prints the names of the three players with the highest points per value ratios. These players are namely, Reece james (defender), Jack Harrison (midfielder) and Matt Doherty (defender). It is important to find players that are both cheap and provide good points to your fantasy team.

Finally, lets evaluate whether having a certain premier league team dominate your fantasy team is a viable option. This analysis will be done in the following figure.

2.6. Min/mean and max points per team for a gameweek:

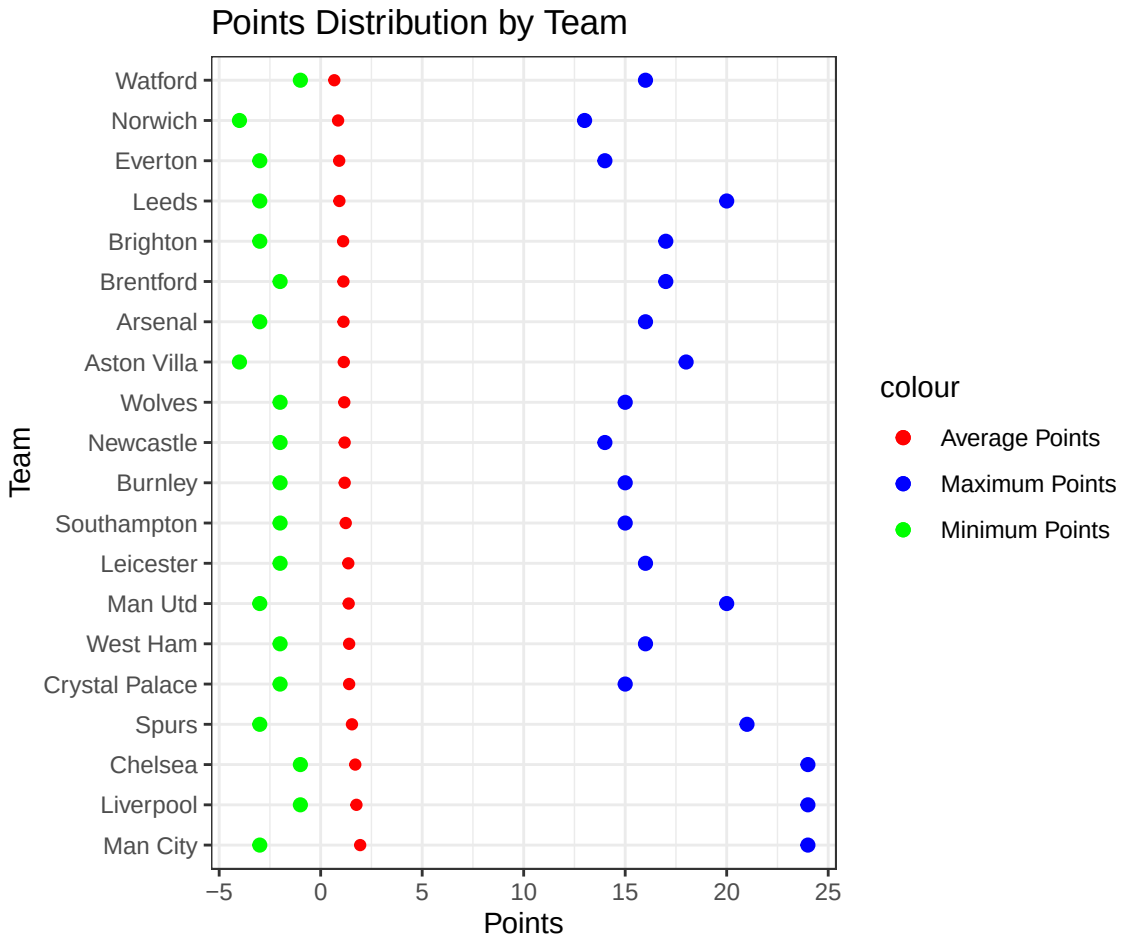

Figure 2.6: Team points min average and max distribution plot

2.6 provides insight into the minimum, average and maximum points scored by premier league players per team in the premier league. From this figure it can be seen that you would rather want to avoid having players from Watford, Norwich and Everton due to their low average points and low maximum points. The teams that you should look to have players from are Man city, Liverpool and Chelsea. These teams posses the highest average points and highest total points. Although the fantasy premier league restriction of a maximum of three players per team are allowed makes this difficult. This restriction ensures that we cannot flood our fantasy team with just Man city players or just Liverpool players. There may however be a case to look to obtain players from Wolves, Newcastle, Burnley,

Southampton and Leicester due to lower overall range in points, If you are a fantasy premier league player that values lower risk which is found in lower point variation, these teams may suit your risk profile.

2.7. Conclusions made from data analysis

The important notes made form this data analysis is that, creativity, ict_index, influence, threat, selected and transfers in are important variables to place in the random forests model as they have explanatory power over the total points variable. Furthermore, when choosing your formation of your fantasy premier league team you should look to have as many midfielders as possible and look to obtain players from Man city, Liverpool or Chelsea.

3. Machine learning model using Random forests

The machine learning model that has been selected for predicting the amount of fantasy points premier league players will score is the random forests model. The reasoning behind this choice is the ability of a random forest model to handle large data sets, reduced risk of over fitting, the results are robust to noise and outliers Breiman (2001).

Handling large data sets is important for this project due to the sheer amount of information available. There are 38 game weeks with over 500 premier league players. Secondly, reduced risk of over fitting is essential for this project as we are predicting the amount of points scored by premier league players over different game weeks. Lastly, it is critical that the model used is robust to outliers since outliers will be very prevalent in the premier league data sets due to shock player performances.

This next section will now walk through the setting up process of the random forests machine learning model.

3.1. Setup process of machine learning model

Firstly, before we begin creating the base line random forests model we need to preprocess the data to ensure the variables are in the correct format. Upon analyzing the variables in the first section it was found that the "value" variable did not read into R correctly and needs to be divided by 10. Next a new data set was created that only obtains the feature variables that will be used, these variables are, "creativity", "ict_index", "influence", "selected", "threat", "transfers_in" and also contains the target variable "total_points". These feature variables were chosen through the analyses done in the data analysis section.

3.1.1. Creating the base random forests model

In this section a base line random forests model will be created, once created an evaluation of the mean absolute error will be conducted. Following this, hyper parameter tuning will take place where the model will be adapted according to the results found from the hyper parameter tuning. Then the new tuned model will be run and the mean absolute error will be compared to that of the base line model. The model with the lowest mean absolute error will be used for the predictions made in the next section.

In setting up the base line model the 2021/22 premier league data will be split into a training set and a test set with a 70/30 split. The model will use repeated cross validation with 5 folds and repeated at each fold 3 times. Furthermore, the model will use 50 trees. This model is then run on the 2021/22 data set with the features being, creativity, ict_index, influence, threat, selected and transfers in and the target variable being "total points". The model is first trained using the training data set. The model then using what it has learnt from the training data set to create predictions on the test data set which it has not yet seen. Next the mean absolute error will be calculated using the yardstick method.

3.1.2. Evaluating the performance of the base line model

Using the yardstick method to calculate the mean absolute error of the base line random forests model the following results are established. The mean absolute error when using the training data is 0.3231536 and when using the test data 0.6033058. These results imply that for the training set, the predicted total points are on average 0.3231536 points off the true total points value and 0.6033058 points off when the test data set is used.

```
## [1] 0.3231536
```

```
## [1] 0.6033058
```

Now hyper parameter tuning on the amount of folds in the cross validation, the value for mtry and the amount of trees will be conducted. Once conducted the new model will be run and the mean absolute error will be evaluated.

3.2. Hyper parameter tuning

3.2.1. mtry hyper parameter tuning

Firstly, we will be tuning the mtry hyper parameter of the model. The tuning grid consists of (1, 2, 3, 4, 5). In the process of tuning this parameter the model will run 5 different times changing just the mtry value from 1 through 5. The model will then pick the mtry value which is associated with the lowest mean absolute error.

After completing the tuning process the mtry value of 2 is associated with the lowest mean absolute error. This mtry value is the same as the one used in the base line model and thus we do not need to change this parameter value.

3.2.2. k-fold Hyper parameter tuning

Secondly, evaluating how many folds to be used in the cross validation process will be calculated. A fold range of (5, 10, 15, 20) will be evaluated. Thus, again the model will be run 4 different times with differing k-fold values in an attempt to find the most appropriate k value that results in the lowest mean absolute error.

After running through the range of k-fold values a k value of 5 is found to be the most optimal.Like that of the mtry parameter value, this is the same k-fold value that was used in the baseline model and thus, the base line model is yet to be adapted.

3.2.3. ntree hyper parameter tuning

The last parameter value to be tuned is the amount of trees used in the random forests model. A tuning grid of (50, 100, 150) will be used. The model will run 3 times iterating through the different tree values and will look for the tree parameter that is associated with the lowest mean absolute error.

After running through the tuning grid of differing tree values, a tree value of 150 is found to create the lowest mean absolute error. This value of ntree = 150 differs from the base line model which used a ntree value of 50. Thus, the base line model needs to be adapted.

3.3. Best model after hyper parameter tuning

After conducting some hyper parameter tuning on the parameter of mtry, k-fold and ntree, the new tuned model is created. This model still splits the 2021/22 data into a training set ad a test set with a 70/30 split. This new random forests model uses repeated cross validation with 5 folds and is

repeated three times at each fold and uses 150 trees. The model is trained using the training set and creates predictions on the test set which it has not yet seen. Once again the features are, creativity, ict_index, influence, threat, selected and transfers in and the target variable being "total points".

After this model is run, the mean absolute error will be calculated and compared to that of the baseline model.

3.4. Evaluating the performance of the tuned model

Using the yardstick method to obtain the mean absolute error of the tuned model we get the following results of 0.3210782 when the training data is used and 0.6019206 when the test data set is used. This means that on average when using the training data set the predicted total points value is off by 0.3210782 points and when using the test set the predicted values are off by 0.6019206 points.

When comparing these mean absolute error values to that of the once obtained in the base line model, the tuned model performs slightly better. With respect to the training data predictions the tuned model performs on average better than the base line model by 0.0020754 points and with respect to the predictions made on the test data, the tuned model performs on average better than the base line model by 0.0013852 points. Thus by increasing the amount of trees from 50 to 150 the model only improves slightly. This provides evidence that although increasing the amount of trees further than 150 may decrease the error it will not be worth the additional computational time.

Thus the tuned model will be used for the predictions that will be made in the following sections.

```
## [1] 0.3210782
```

```
## [1] 0.6019206
```

4. The important restrictions to note when building your fantasy premier league team

There are three important restrictions that need to be taken into account when building your fantasy premier league team. The first being that your are only allowed a maximum of three players per premier league team in your fantasy premier league team. Secondly, your total team cost cannot exceed 100 mil. Lastly, you need to pick which formation you want to play. This means that you have to have at least one goal keeper, between 3-5 defenders, between 3-5 midfielders, and between 1-3 forwards. Due to the analyses made in the data analysis section I will be choosing a formation that has as many midfielders as possible and seeing that you can get cheap defenders that can score lots

of points I will be going for more defenders than strikers to keep the cost down. Thus the formation chosen for this project is 1 defender, 3 defenders, 5 midfielders and 2 strikers.

5. Time for predictions

5.1. Predictions on the 2022/23 data set

The new tuned random forests model is used to create predictions on the 2022/23 data set. Firstly, the 2022/23 data set must be cleaned to ensure that the 2022/22 and 2022/23 data sets contain the same variables. Once this was done, the predictions on the 2022/23 data set could commence.

Once predictions were made, this prediction variable is added back to the main data set so that we could extract variables such as the name of the player the actual points scored, the predicted points scored, what team they play for, which potion they play in and their value.

5.2. Game week predictions

5.2.1. Beggining of the season predtions

5.2.1.1. Game week 9 predictions and actual best team. The following 11 players are apart of the predicted fantasy team for game week 9:

```
## # A tibble: 11 x 6
##    names_22_23              predicted_points total_points position team    value
##    <chr>                               <dbl>        <dbl> <chr>    <chr>   <dbl>
##  1 Erling Haaland                       19.4           23 FWD      Man Ci~  12.1
##  2 Phil Foden                           16.3           19 MID      Man Ci~   8
##  3 Leandro Trossard                     16.0           20 MID      Bright~   6.6
##  4 James Maddison                       15.3           18 MID      Leices~   8
##  5 Roberto Firmino                      13.9           12 FWD      Liverp~   7.9
##  6 Miguel Almirón Rejala                13.9           15 MID      Newcas~   5
##  7 Jarrod Bowen                         10.6           14 MID      West H~   8.1
##  8 Conor Coady                           9.12           9 DEF      Everton   4.8
##  9 Thiago Emiliano da Silva              7.71           6 DEF      Chelsea   5.4
## 10 Vitalii Mykolenko                     6.16           2 DEF      Everton   4.5
## 11 Illan Meslier                         4.89          11 GK       Leeds     4.5
```

The following 11 players are apart of the actual best team for game week 9:

```
## # A tibble: 11 x 6
##    names_22_23           predicted_points total_points position team      value
##    <chr>                            <dbl>        <dbl> <chr>    <chr>     <dbl>
##  1 Erling Haaland                    19.4           23 FWD      Man City   12.1
##  2 Leandro Trossard                  16.0           20 MID      Brighton    6.6
##  3 Phil Foden                        16.3           19 MID      Man City    8
##  4 James Maddison                    15.3           18 MID      Leicester   8
##  5 Miguel Almirón Rejala             13.9           15 MID      Newcastle   5
##  6 Jarrod Bowen                      10.6           14 MID      West Ham    8.1
##  7 Roberto Firmino                   13.9           12 FWD      Liverpool   7.9
##  8 Illan Meslier                      4.89          11 GK       Leeds       4.5
##  9 Thilo Kehrer                       2.73          10 DEF      West Ham    4.5
## 10 Conor Coady                        9.12           9 DEF      Everton     4.8
## 11 Timothy Castagne                   2.27           8 DEF      Leicester   4.4
```

In game week 9 the prediction model correctly predicted 9 out of the 11 players with a total of 149 points out of a maximum of 159 points obtained by the actual best team.

5.2.2. Middle of the season predictions

5.2.2.1. Game week 16 predictions and actual best team. The following 11 players are apart of the predicted fantasy team for game week 16:

```
## # A tibble: 11 x 6
##    names_22_23           predicted_points total_points position team      value
##    <chr>                            <dbl>        <dbl> <chr>    <chr>     <dbl>
##  1 Ivan Toney                        15.7           13 FWD      Brentford   7.4
##  2 Rodrigo Moreno                    13.9           13 MID      Leeds       6.3
##  3 Darwin Núñez Ribeiro              13.6           13 FWD      Liverpool   9
##  4 Martin Ødegaard                   13.0           16 MID      Arsenal     6.4
##  5 Rodrigo Bentancur                 12.5           14 MID      Spurs       5.4
##  6 Crysencio Summerville              9.50           7 MID      Leeds       4.4
##  7 Phil Foden                         9.42           9 MID      Man City    8.3
##  8 Andrew Robertson                   9.02           9 DEF      Liverpool   6.7
##  9 Ben Davies                         8.96           7 DEF      Spurs       4.9
## 10 Mathias Jorgensen                  5.36           2 DEF      Brentford   4
## 11 Danny Ward                         5.00          11 GK       Leicester   4.1
```

The following 11 players are apart of the actual best team for game week 16:

```
## # A tibble: 11 x 6
##    names_22_23          predicted_points total_points position team      value
##    <chr>                           <dbl>        <dbl> <chr>    <chr>     <dbl>
##  1 Martin Ødegaard                13.0             16 MID      Arsenal     6.4
##  2 Rodrigo Bentancur              12.5             14 MID      Spurs       5.4
##  3 Christian Eriksen               0.304           13 MID      Man Utd     6.3
##  4 Rodrigo Moreno                 13.9             13 MID      Leeds       6.3
##  5 Ivan Toney                     15.7             13 FWD      Brentford   7.4
##  6 Darwin Núñez Ribeiro           13.6             13 FWD      Liverpool   9
##  7 Joe Willock                     9.12            11 MID      Newcastle   4.9
##  8 Danny Ward                      5.00            11 GK       Leicester   4.1
##  9 Andrew Robertson                9.02             9 DEF      Liverpool   6.7
## 10 Benjamin White                  3.11             8 DEF      Arsenal     4.6
## 11 Fabian Schär                    4.22             7 DEF      Newcastle   4.9
```

In game week 16 the prediction model correctly predicted 7 out of the 11 players with a total of 114 points out of a maximum of 128 points obtained by the actual best team.

5.2.3. End of the season predictions

5.2.3.1. Game week 35 predictions and actual best team. The following 11 players are apart of the predicted fantasy team for game week 35:

```
## # A tibble: 11 x 6
##    names_22_23               predicted_points total_points position team  value
##    <chr>                                <dbl>        <dbl> <chr>    <chr> <dbl>
##  1 Dwight McNeil                         17.1           21 MID      Ever~   5.1
##  2 Ilkay Gündogan                        13.9           13 MID      Man ~   7.3
##  3 James Ward-Prowse                     13.4           13 MID      Sout~   6.1
##  4 Abdoulaye Doucouré                    13.3           13 MID      Ever~   5.3
##  5 James Maddison                        12.9            9 MID      Leic~   7.9
##  6 Toti António Gomes                     9.64          14 DEF      Wolv~   3.8
##  7 Harry Kane                             9.13           8 FWD      Spurs  11.4
##  8 Lyanco Silveira Neves Voj~             7.78           6 DEF      Sout~   4.4
##  9 Benoît Badiashile                      7.71           7 DEF      Chel~   5
```

```
## 10 João Félix Sequeira                  5.84            7 FWD      Chel~   7.2
## 11 Nick Pope                            4.96            2 GK       Newc~   5.4
```

The following 11 players are apart of the actual best team for game week 35:

```
## # A tibble: 11 x 6
##    names_22_23         predicted_points total_points position team        value
##    <chr>                          <dbl>        <dbl> <chr>    <chr>       <dbl>
##  1 Dwight McNeil                  17.1            21 MID      Everton       5.1
##  2 Toti António Gomes              9.64           14 DEF      Wolves        3.8
##  3 Abdoulaye Doucouré             13.3            13 MID      Everton       5.3
##  4 Ilkay Gündogan                 13.9            13 MID      Man City      7.3
##  5 James Ward-Prowse              13.4            13 MID      Southampton   6.1
##  6 Pedro Porro                     6.08           12 DEF      Spurs         4.8
##  7 Saïd Benrahma                   9.86           11 MID      West Ham      5.5
##  8 Virgil van Dijk                 6.99           11 DEF      Liverpool     6.6
##  9 Aaron Ramsdale                  3.55           10 GK       Arsenal       4.8
## 10 Harry Kane                      9.13            8 FWD      Spurs        11.4
## 11 João Félix Sequeira             5.84            7 FWD      Chelsea       7.2
```

In game week 35 the prediction model correctly predicted 7 out of the 11 players with a total of 113 points out of a maximum of 133 points obtained by the actual best team.

6. Conclusion

TO conclude, by using the above fantasy premier league prediction model I was able to predict up to 9 out of the 11 players that would make the best team of the week with on average 6/7 players correctly predicted. However, this model can be adapted if it were to actually be used for selecting a fantasy team throughout the season. This is because what this model can do is predict a brand new team each game week fairly well, which is not allowed in the actual fantasy game, since you can only make a few substitutions per week.

References

Breiman, L. 2001. Random forests. *Machine learning*. 45:5–32.

Appendix

Appendix A

Some appendix information here

Appendix B