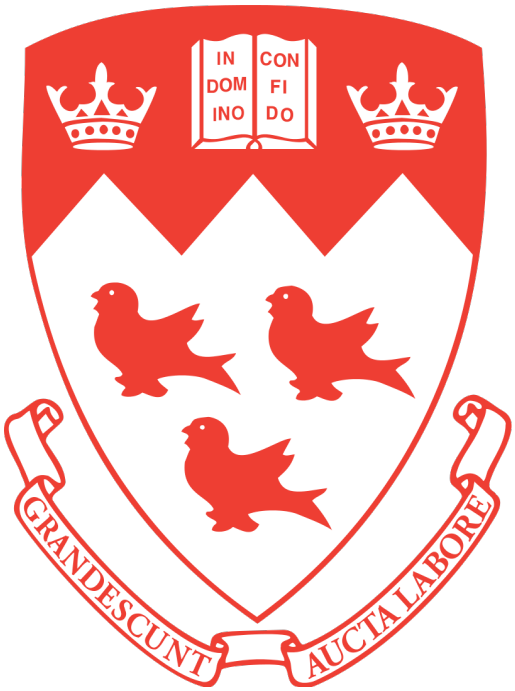
Soccer Analytics

December 16th, 2020

Master of Management in Analytics
McGill University



Summary

A Introduction

In 2002, Oakland Athletics defied all odds and won the Major League Baseball with a record breaking 20-game winning streak and their campaign went on to become one of the most famous victories in the league's history. A seemingly impossible task at the time, their success was spearheaded by one man with his creative yet unpopular approach of using technology and analytics to sign undervalued players. The immediate success of the Oakland Athletics spurred other sports teams to replicate the model pioneered by Billy Beane. Known popularly as the Moneyball Approach, this was a new methodology to player recruitment in baseball using Sabermetrics, that measure in-game activities, and is used to derive insights. With Billy Beane as general manager, the A's went on to win four west division titles with around 96 average wins per year. According to Billy Beane, the most important part of their success was the ability to know why they were successful, with the help of numbers.

Today, data and analytics are becoming an integral part of the soccer industry. With records on thousands of actions for just a single game, coaches and tacticians are relying more heavily on technology for scouting and pinpointing transfer targets, analysing performance of individuals, efficiency and impact of tactics deployed, developing young talents and analysing the styles of play. Clubs have started to collect huge amounts of data through third party vendors, such as Opta. Originally collected for fans and media outlets, they have trickled down to the actual football departments, where statisticians use them to reconstruct football matches and tell a story. But that is where things get interesting. Football is a complex sport, compared to Basketball which is high scoring, or Baseball which is segmented. It is time varying, continuous, low scoring and subjective, meaning that people could have different opinions for the same event. The avalanche of information collected from every match is used to answer questions like how a team pressed, how effective the formation was, how often were teams exposed to counter attacks, or how good a team was on set-pieces. The stakes are very high for the big clubs, but even higher for smaller ones. If they make a mistake in one of their signings, it could be catastrophic to the entire club. To reduce the risk and maximize the outcome from every recruitment, they have turned to Beane's sophisticated Sabermetrics Approach. An example of this is Southampton FC. To compete against the mega-rich teams in the Premier League, they created a live database collecting player metrics from every major league in Europe. This enabled them to acquire underrated players and sell them on for a profit. It is no coincidence that superstars of today, like Sadio Mane, Virgil Van Dijk and many more, were bought by super-clubs for big money from Southampton.

As analytics started to evolve in Soccer, new metrics began to rise. One of the most widely used and accepted metrics is the Expected Goals (xG). It essentially measures the quality of a chance. In terms of statistics, it is the probability of a shot becoming a goal, depending on various factors like the method of assist, body part used, location and distance of shot from goal. Analysts aggregate thousands of shots taken historically from different leagues and develop a machine learning model that predicts the probability of the shot being a goal. This is a decisive tool used by many of the top clubs in Europe today, because it gives us an idea of how good teams or players are based on the quality of chances created. For instance, in the 2016-2017 Premier League season, Chelsea finished top of the table, but their expected

goals were much lower than other teams. This is because they massively overperformed in their attack and resulted in goals which were less likely. Consequently, Chelsea replaced their star striker Diego Costa with a less prolific Alvaro Morata, which resulted in Chelsea finishing 5th in the subsequent season as Expected Goals had predicted the season before.

In this project, I will be developing a classification model using both Logistic Regression and Tree-Based Methods, to predict the expected goals of specific in-game events in Europe's top leagues from the year 2011 to 2017 and find out which teams under-performed and over-performed for these seasons.

B Data Description

The primary dataset, I would be using here, contains information on over 9000 matches that took place on the top five leagues of Europe from the year 2011 to 2017. Specifically, the dataset contains variables like the type of event leading to a shot, the time of shot, players and teams involved, location and situation of the shot, method of assist and outcome of the shot. This information was derived from the actual text commentary for each individual match taken from ESPN. After the cleaning process, the dataset contained over 2 lakh shots, on which subsequent analysis were done. I also worked on a secondary dataset which contained the betting odds for matches in the primary dataset.

B.1 Independent Variables

The aim of this project as mentioned before, is to measure the quality of a shot taken. Therefore, the independent variable here would be the outcome of the shot. In the dataset, the variable 'is_goal' provides this information where a value 1 signifies that a goal was scored and 0 says otherwise. Lets now look at how the goals are distributed in the dataset.

From graph 1, we could see that only around 25000 out of the two lakh shots resulted in a goal. This means that only around 12.5% of the shots were converted. On further analysis, we could infer that on average, around 2.77 goals were scored per game for this dataset. This goes on to say how low-scoring football matches can be. Now let us look at how location of a shot varies with outcome.

B.2 Variable Relationships

From graph 2, we could see that a large number of shots were taken from the centre of the box, and outside the box. However, majority of the goals were scored from the centre of the box and from very close ranges. On the other hand, majority of the shots were missing the end product when executed from outside the box. From this, we could understand that it is much easier to try a shot from outside the box, but the probability of the shot resulting in a goal is small. In contrast, it is difficult to take shots from the centre of the box and even harder to get close range opportunities, but these locations are the most ideal for taking a shot. Thus, we can say that the closer a shot is taken, the greater the probability of scoring a goal. It is also worth mentioning that majority of the penalty shots taken during these games had been converted to goals.

The method of assist leading to a goal could also have a significant impact when determining the quality of a shot. In this dataset, there are mainly five assist methods used. There are passes, through balls, headed passes and crosses. There is another category of assists named none, which means that those shots could be taken without any assists. For instance, shots from set-pieces or penalty spots does not require assistance from another player. From graph 3, a majority of the shots taken resulted from normal passes and crosses. A large number of shots did not require assists as well. Conversion rates were highest for shots that resulted from passes, followed by crosses. Shots from set pieces and freekicks were also effective since a large percentage of these shots had positive results. It is important to note that through balls are surprisingly difficult to execute as we could see from this graph, the shots resulting from through balls are very limited. It requires players with great vision and technical ability to execute these, let alone convert them to goals.

The next plot shows us how different leagues across Europe stacked up against each other. From graph 4, the number of shots taken were lowest for the Premier League (England) followed by the Bundesliga (Germany) while it is highest for Serie A (Italy), followed LaLiga (Spain) and Ligue 1 (France). The number of goals scored among these leagues is fairly the same, with a slightly lower total for the Premier League. From this analysis, we could deduct that the level of competition in the premier league is the highest since it is the hardest among the leagues to take a shot and convert them to goal. This could either mean that the Premier League teams are more defensively oriented, or they had a lower attacking power compared to other teams from across the continent. On the other hand, Serie A boasts the highest number of shots and one of the highest total number of goals scored. This means that teams in this league adopt a more attacking style but lacks the defensive solidity.

Another important factor to consider would be the venue in which teams play, which could have a psychological impact on the players. From graph 5, we could see that the total number of shots taken and the shots to goal conversion rate was higher for home teams. This further reinforces that teams benefit from the support of their fans and home conditions.

Now, let us see how shot outcome varies with time. From graph 6, we could see that a higher number of shots were attempted as the game progressed. Also, there is a significant increase in the number of shots close to the stoppage time, that is, the 90-minute mark. Also, the highest number of shots on target took place when matches were almost finished. This goes on to say that teams tend to be more aggressive during the tail end of the game, and mistakes during this time could prove costly. Therefore, managers should prepare the players to be extra focused during the later parts of the game and should try to change tactics depending on whether they are trailing or leading the match during these time periods.

Now, let us look at the top performing players from 2011 to 2017. Graph 7 and graph 8 shows the top 10 goal scorers and top 10 assist providers across Europe. Unsurprisingly, the player who outperformed everyone in terms of goals and assists is Lionel Messi, with a grand total of 186 goals and 73 assists. This means that roughly, he has scored or assisted a goal for Barcelona every 80 minutes, which is a truly mind-boggling statistic even for the greatest player in the world. The player who finished a close second is Cristiano Ronaldo with 175 goals and 51 assists under his belt.

As mentioned before, I have also worked with a dataset containing market odds of each match, which was aggregated from oddsportal.com. The dataset contains odds of the home team winning a match, the away team winning the match, and the match resulting in a draw. The odds are usually lowest for clear favourites, whereas highest for highly unlikely outcomes. Majority of odds of home winning the match ranged from 1.76 to 3.08 with an average of 2.93, whereas a majority of away team odds of winning had a higher range of 2.74 to 6, with a mean odd of 5.5. On the other hand, odds of teams drawing were fairly neutral with an average of 4.27. This correlates with our findings from the goal distribution for home and away sides. In other words, generally, it is more likely for home teams to win.

C Model Selection and Methodology

I have built three models for this project, a logistic regression to measure the expected goals of a shot, and two tree-based models to predict the outcome of a match.

C.1 Logistic Regression

As mentioned before, the project aims to analyse and measure the quality of a shot. For this I have used the Expected Goals metrics, which is essentially the probability of a shot leading to a goal. I have used the logistic regression to build this model.

After trying out several combinations of variables, I quickly realized that location from where a shot was taken had the most significant impact on the outcome of the shot. There were around 18 locations defined in the dataset from which shots where taken, and all the categories had very high significance for the model. As we inferred before, as the distance increased from the goal, the probability of shot leading to goal decreased. Also, scoring a goal is highest if the shot was taken from the left side or the right side of the six-yard box, followed by penalty shots and shooting from the centre of the box.

Match venue also had a significant predictive power with home teams usually having a higher chance of scoring a goal. Another predictor used in the model is the body part used for scoring. Shots taken with foot had a higher chance of resulting in a goal compared to headers. This could be because headed goals are less in number and require a higher physicality and height for proper execution. Therefore, players efficient in using their head for scoring a goal could be limited. Situation of shot taken also had a high influence in the model. Here, we see that shots from open play had a lower probability of resulting in a goal compared to shots taken from set-pieces. It is also worth noting that shots from free kicks also have a lower probability of goal. This result shows us that when teams get awarded free kicks, it is better to use them as set-pieces rather than directly trying to score goals.

Other influential predictors used in the model are, method of assist, season in which the match took place, if the shot was taken by a top player or provided by a top assist provider and if the top teams where involved in the event. I have also added an interaction term with the body part used for scoring and the method of assist.

On running the model with the above-mentioned predictors, it was seen that the algorithm took 15

iterations to find the optimal solution, with an R-squared of 25.9%. The reason for a lower R-Squared could be because of the high-class imbalance between the shots leading to goal and shots leading to a miss. An important note here is that typically in the football industry, if a shot has an expected goals value of 0.38 or higher, it could be termed as a big chance. This is because several studies conducted by top companies like Opta show that around 38% of penalties were converted into goals over the past years. Therefore, if shots received an expected goals probability of above 38%, it means that there is more probability of converting the shot compared to penalty kicks.

One thing to note here is that Expected Goals models are typically not used for prediction purposes. However, in order to determine the efficacy of the model, I have calculated the in-sample and out-of-sample error rates for the two classes wherein if the expected goals were greater than 0.5, then the shot would result in a goal.

The in-sample error rate for this model was 0.087. On performing the K-Fold cross validation with 5 folds, the model yielded a similar error rate 0.088.

C.2 Tree-Based Models

Once the Expected Goals (xG) were calculated for each shot in the previous dataset, this was used to engineer several new features along with the betting odds to predict the outcome of a match. I have decided to build a classification model with three separate classes, home team win, home team loss and draw for Random Forests, and two classes, home team win and home team loss, for Gradient Boosting method.

As a preliminary analysis, I have run the random forest and gradient boosting with their respective classes given above as outcomes, with predictors originally given in the dataset, that is, home win odds, away win odds and odds of draw. From the results, Random Forests yielded an Out-Of-Bag score of 51.3% and the Gradient Boosting method yielded an in-sample error rate of 0.278. I would be using these models as benchmarks for the subsequent models that I would be testing later on with the newly engineered features.

C.2.1 Feature Engineering

For properly predicting the outcome of a match, several things come into play. Some of them could be the quality of the team and players, the current form of the team before the match, total number of goals scored and conceded leading up to the match, total points earned and so on. Since I have data on almost all matches that took place over the 5-year period, I was able to derive all these additional features through data manipulation. In addition, I have also used the Expected Goals metric to measure the level of performance of teams in the previous matches.

First, I started by grouping individual matches so that I could calculate the total expected goals aggregate for each home and away team during those matches. This would then be joined to the odds table by their respective match IDs, which gave us the home team and away team expected goals for each match. However, this could not be used in the prediction model since the expected goals could only be calculated

after the match was over. Therefore, we should look at the stats of each individual match for the past games leading up to the match in question.

Next, I calculated the difference in goals scored and goals conceded for both home and away teams. This is known as Goal Difference. Along with this, I also calculated the number of points gained by both home and away teams for each match. If a team won, they would receive 3 points whereas if the team lost, they would receive no points. If the teams had drawn the match, each team involved would receive 1 point apiece.

Now that we have all the required attributes for each match, next step is to calculate the performance of teams based on these newly created variables for the previous 5 matches leading up to a match. For this, I have first extracted the Home Team and Away Team attributes into a separate data frame. Then, I moved on to combine both these data frames so that we could calculate the past performance of each team over the entire season instead of limiting to just home or away matches.

Next, I determined the matchday in which each match occurred for a single season. This was done by converting the date of the match into week of year, since league matches are usually played out every week. Also, I knew that football seasons normally started around the month of August. Therefore, I assigned those dates as the first matchday for each team and iterated this process through the entire dataset to determine the matchdays. After this step, I ordered all the observations according to team names and match days to find out the level of performance for the previous 5 matches. This was done by taking the moving sum of goals, goals conceded, and points earned, as well as the moving average of the expected goal and expected goal conceded for the previous 5 observations. This was then merged with the odds dataset. In addition to this, I have also calculated the goal differences for each home and away team for the past 5 matches.

Finally, I checked for correlation between the newly created variables and found out that none of them exhibited multicollinearity. Thus, I was able to engineer a total of 12 variables which could be used along with the odds data to predict the outcome of a match.

C.2.2 Random Forest

The random forest model was built using all the newly created variables along with the odds of the match. This resulted in an OOB score of 47.76%. Thus, I was able to reduce the error rate from the previous model by around 4%. From the confusion matrix, I could see that the model was most efficient in predicting Home Team wins but suffered heavily when it came to predicting draws. Again, this could be a result of high imbalance in the dataset, since the number of draws is lower compared to other classes. On analysing the variable influence from Graph 19, predictors like Top Home and Away Team, Home and Away Team points gained for past 5 matches and Home and Away team Goal Difference in the past were able to increase the accuracy the most among the newly created variables. On the other hand, variables like Home and Away Team Average Expected Goals along with Average Expected Goals conceded for the past 5 matches, where able to increase the purity of the leaf nodes the most.

I have also run a trace test with a step size of 50 to determine the optimal number of trees required. The

test revealed that a tree count of 500 was optimal for the random forest to operate most effectively for the dataset.

C.2.3 Gradient Boosting Method

For this model, I used the same set of features from the random forest, with the independent variable being the only difference. As mentioned previously, here, the classes used are Home Win and Home Loss. With a lower number of classes compared to random forest, the model was able to perform significantly better with an in-sample error rate of 18.2%. This is a significant improvement from the base model built with just the odds data, with a decrease in error of almost 10%.

On analysing the relative influence of each variable from Table 1, it is seen that the Home and Away Average Expected Goals and Average Expected Goals Conceded for the last 5 matches, had a higher influence compared to Odds of away win and odds of draw, whereas odds of home win had the highest influence. Along with these, the league of the match, number of goals scored, and points earned leading up to the match and goal differences also had a good influence on the predictive power of the model. Also, the number of trees constructed in the gradient boosting method is set to 1000 since this gave the optimal results.

From this, we could see that analysing the past performances of teams could significantly benefit us when trying to predict the future outcome of the teams involved. However, one drawback of using this approach would be that we could not predict the outcome of the first 5 matches in a given season for the team. On the other hand, including them might decrease the accuracy of the model even further.

D Results

Now that we have information on the team statistics across different seasons, let us look at how our model could be used by managers for performance analysis.

For this, I have combined performance metrics like Goals scored, goals conceded, expected goals and expected goals conceded for all teams into a single table. This would help us see which teams were underperforming or overperforming and for which seasons. We could measure the performance of teams over three aspects of their game. These are attack, defence, and overall performance. These could be measured by taking the difference of the actual goals scored and Expected goals for attack and difference between the expected goals conceded and the actual goals conceded by the team for defence. If the values of attack and defence are higher, it means that teams are outperforming their expected goals and results might not be sustainable in the long term. Overall performance of the teams could be computed by adding the attack and defence metrics of a team.

First, let us take a look at the overall team performances for teams in the Premier League. From Graph 9, the team that overperformed the most in the premier league between these years is Leicester City in 2016-2017 season when they unprecedentedly won the Premier League ahead of teams like Arsenal and Chelsea. This was one of the most astonishing feats achieved in the Premier League to date. At a time

when all the mega rich clubs of the premier league where signing on big players for millions of dollars, a team like Leicester who had been struggling to stay out of relegation zone in the previous seasons, had virtually no chance of winning the premier league. Now, let us take a closer look at Leicester's run of form from 2014 to 2017 and see the peaks and troughs of their performances.

From the graph 13, we can see that there is a big gap between the actual stats and the expected metrics of goals and goals conceded in the year 2016 when they won the premier league. However, on a closer look, the season leading up to that and the season after their miracle win, had results much more representative of their performances.

From this graph, we could see that Leicester had massively overperformed in their defence as the Defence metric was close to 20. This means that they are conceding 20 fewer goals than expected. This could mean that the Leicester goalkeeper Kasper Schmeichel had an unbelievable season as Leicester had just conceded around 0.95 goals per game. However, as our model rightly pointed out the inconsistent defensive performance, Leicester had conceded over 1.66 goals per game over the 2016-2017 season.

Now, let us take a look at the over performing teams in the Spanish top division. From Graph 11, we see that Barcelona topped the table as the most overperforming team with a performance metric of 36 for the season 2012-2013. It is no coincidence that they won the La Liga in 2013 with a record of scoring in all 38 games in a single season. The next team with a high-performance metric is Real Madrid, in 2014-2015 and 2015-2016. From the graph, we see that Barcelona and Real Madrid has consistently outperformed over these seasons. Therefore, let us compare the two teams.

From the graph 14, we could see that there are some clear differences between the expected goals and expected goals conceded over these seasons. However, when it comes to the defensive performance, both teams perform at the level indicated by the performance metrics. Therefore, one could assume that both teams are exceptionally attack oriented. On further analysis, we see that the two most prolific goal scorers over these seasons as we saw in graphs, played for these clubs. Coming back to the graph, we see that for Real Madrid, there is a clear rising trend in the expected goals conceded. This signifies that the team is giving away better-quality chances to opposing teams and it signals defensive reinforcements are required. On the other hand, for Barcelona, stats had varying trends across seasons with an under-performing defence in the years 2012 and 2013, and an over-performing defence in the year 2015, when they famously achieved the treble of La Liga, the continental Champions League, and the Spanish Cup. However, after the 2014-2015 season, they had a dip in performance as the expected goals conceded and actual goals conceded had a significant increase compared to previous seasons. This performance dip corelates with them losing to Real Madrid in La Liga and Atletico Madrid in the Champions League.

We could also use this method of performance analysis for individual players and for measuring their levels of goal scoring and assist. Let us take a look at the goal statistics of individual players. The graph 15 shows the top 10 over-performing players according to the Expected Goals metrics for them. Here, we can see that in 2013, Lionel Messi scored 41 goals but had a performance metric of 14.61. This means that Messi had scored 14 more goals than expected. We could also take a closer look at individual player performances over different seasons. Let us take a look at Gonzalo Higuain. From graph 16, we can see

that Higuain had outperformed the expected goals by around 9 and his scoring had become almost equal to expected goals by 2015. However, we could also see that his Expected Goals metrics had been rising throughout these years along with the number of shots on target. This means that his performances have improved over the years and as a result, he had another exceptional season in the year 2016.

In the same way, we could also derive insights about expected assists from the top assist makers in Europe in these seasons. If we look at the top 10 assist providers in Europe from Graph 17, we could see that Andreas Iniesta of Barcelona FC performed at the highest level and outperformed himself by 7.82 assists. Second most over-performing player would be Angel Di Maria with a total of around 15 assists. Let us take a closer look at Di Maria. From graph 18, we see that Di Maria had a couple of seasons with very volatile performances. In 2012, we could see a significant difference between assists and expected assists with the player outperforming by around 7. However, in the subsequent seasons, his performances became more in line with the Expected Assists. One thing to note here is that Expected Assists not only depends on the player but also on the teammates on the receiving end. Therefore, if his teammates were inconsistent or performing poorly, that would have a negative impact on the assist providers as well.

E Conclusion

Finally, we come to the end of the project. In the previous couple of sessions, I have tried to show how the performance analysis done in football had evolved over the years. Today, Expected Goals metric has revolutionized the game with many top-level managers and football clubs having analytics departments of their own to come up with efficient tactics based on different opposition. With some of the brightest minds working behind the scenes in the game, athletes could stay consistent and at the top of their game longer than ever before.

Here, I have made a simple logistic regression based on a couple of attributes of shots taken over the duration of various matches. Even with minimal information, the models were able to effectively analyse the player as well as team performances over the duration of individual campaigns. Effective usage of the Expected goals model could enable teams to identify their weak points as well as strong points. This would be crucial for top level management when it comes to additional investments and transfer market activities. Clubs should no longer rely on scout information and their gut feelings to buy players from other clubs or even free agents. Instead, they should look at the performance metrics along with the other actual attributes to get a better picture of the player potential.

Another important aspect is the style of play. With Expected Goals, we could easily identify how teams are oriented. If they are attacking oriented, we could see from which side or flank of the pitch, quality chances were created, or the method of assist that led to the best chances. On the other hand, we could go to the other side of the pitch to see how defenders are contributing to the team. If the Expected Goals conceded are higher for the team, it means that more shots are being attempted from high quality locations and defenders are not doing enough to give protection. We could also analyse the performances of defensive midfielders through tackles attempted or blocked attempts. Defensive midfielders are the

cornerstone of almost any team in the world and they are one of the most under-appreciated players in the pitch. There are a couple of high-profile examples of midfielders changing the scenario. One of them would be N'golo Kante. As mentioned before, during the historic race to the premier league by Leicester City, Kante stood out among the pack. He was seen by many as the main factor to the premiership as he made a large number of tackles and interceptions consistently over the entire 2015-2016 season. As a result, he earned a move to Chelsea FC in the next season and won the Premier League again. Subsequently, he was given the Premier League Player of the Year award in 2017.

Expected Goals metrics could also be used by betting companies to provide more information regarding teams and players to assist people in placing their bets. In this project, I attempted to build a predictive model using Expected Goals metrics along with the betting odds as the primary independent variables. Even with the unpredictable nature of the game, the model was able to predict outcomes of games with satisfactory results. This shows us that with additional information like the player squads, player line ups for the match (could be collected one hour before the game), weather conditions and location of the match among a few examples, the potential capabilities of the predictive models in soccer are staggering.

Even with the sophisticated performance models used in the industry today, people are still sceptical as to how effective these could be when determining the level of performance. A primary reason for this is that Expected Goals treat all players as equal. For instance, one might think that finishing ability of players is important when determining the probability of scoring a goal. However, Expected Goals were surprisingly effective when it came to player analysis and finishing turned out to be not as important as perceived. For instance, Cristiano Ronaldo, one of greatest finishers of all time, had outperformed the expected goal metric only twice since 2012. It was the same case with Robert Lewandowski, another great finisher of the modern game. So what makes Expected Goals metrics so special is that it showed us what we should actually value in a striker. Ronaldo and Lewandowski are world-class strikers, but what makes them special is their amazing ability to get high quality chances closer to goal, and not finishing them at an above average rate.

Another problem with xG is that it does not know where defenders are. This is because most xG models do not have tracking data built into them, which means that when it judges a shot location, it could not know how many defenders are there between the player taking the shot and the goal. This is the reason why there are through balls and fast breaks built into the model as a loose way of saying how set the defense is and how much pressure a player is likely to have.

Having said that, Expected Goals are not built to replace the eye test. Instead, they should be used to whittle down a huge number of players we should be scouting to one or two. This is kind of a myth in football where people think that statisticians do not really watch the game and only recommend players based on their statistics. However, with so many players playing in different parts of the world, we should have a shortlist which we could send scouts out to watch. Metrics like Expected goals, shots, key-passes and many others help us achieve this. This is why big clubs like Arsenal, PSG and Borrusia Dortmond are known to use data both in player recruitment and assessment of their own teams.

As a final note, Expected Goals and advanced statistics are going to be a part of the future of the game,

and the sooner we get accustomed to it, the better. As mentioned before, this does not mean that we do not have to watch the game. It is just a tool to get us a little bit more information.

F Appendix

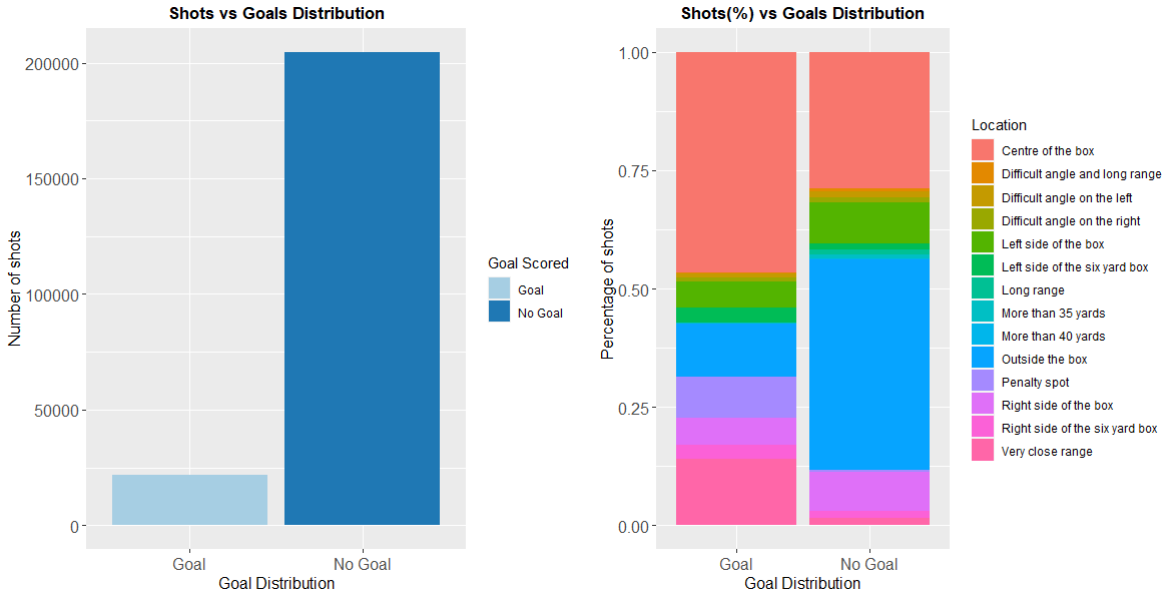


Figure 1: Graph 1 and 2

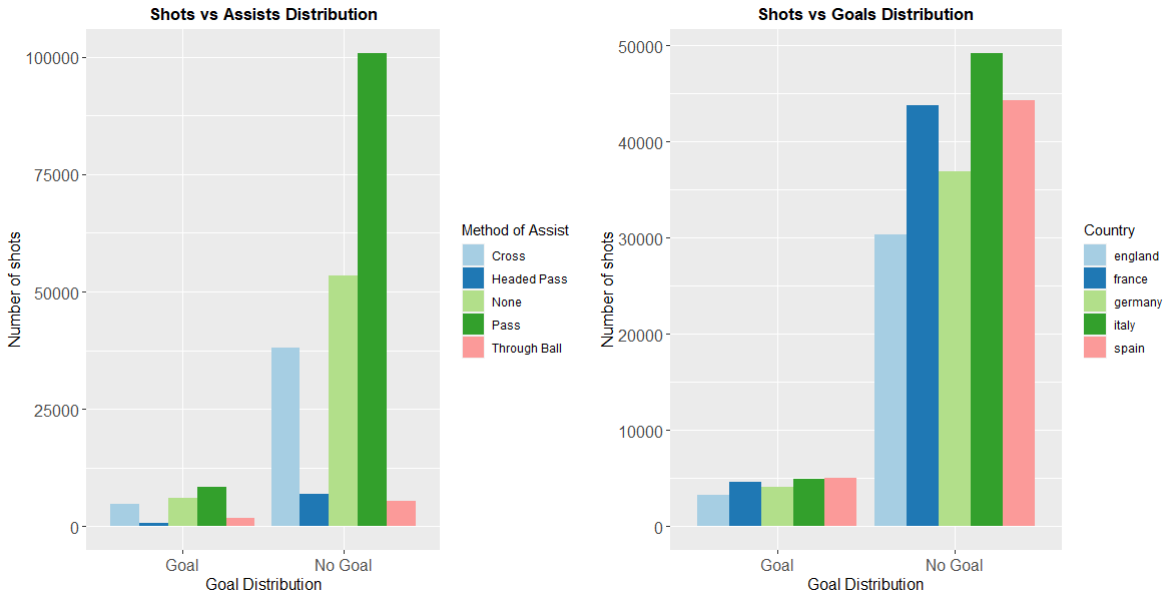


Figure 2: Graph 3 and 4

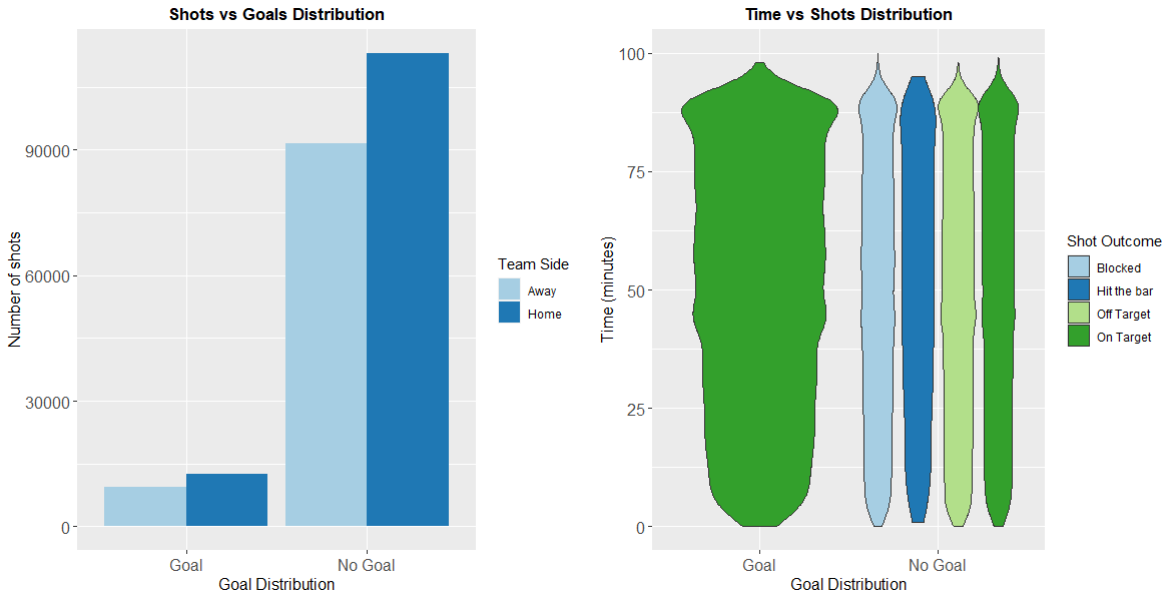


Figure 3: Graph 5 and 6

Variable	Relative Influence
odd_h	12.4106522
AT_avg_xGC_past	11.8842576
AT_avg_xG_past	11.7861756
HT_avg_xG_past	11.7322482
HT_avg_xGC_past	9.9209093
odd_a	9.8061762
odd_d	6.9202220
country	3.0651294
AT_Pts_past	2.9696019
AT_GD_past	2.8771722
HT_Pts_past	2.6963311
AT_Goals_past	2.4545975
HT_GD_past	2.4195683
HT_GC_past	2.3723378
HT_Goals_past	2.3481911
AT_GC_past	2.1923795
season	1.6380940
top_ht	0.2695878
top_at	0.2363683

Table 1: Random Forest

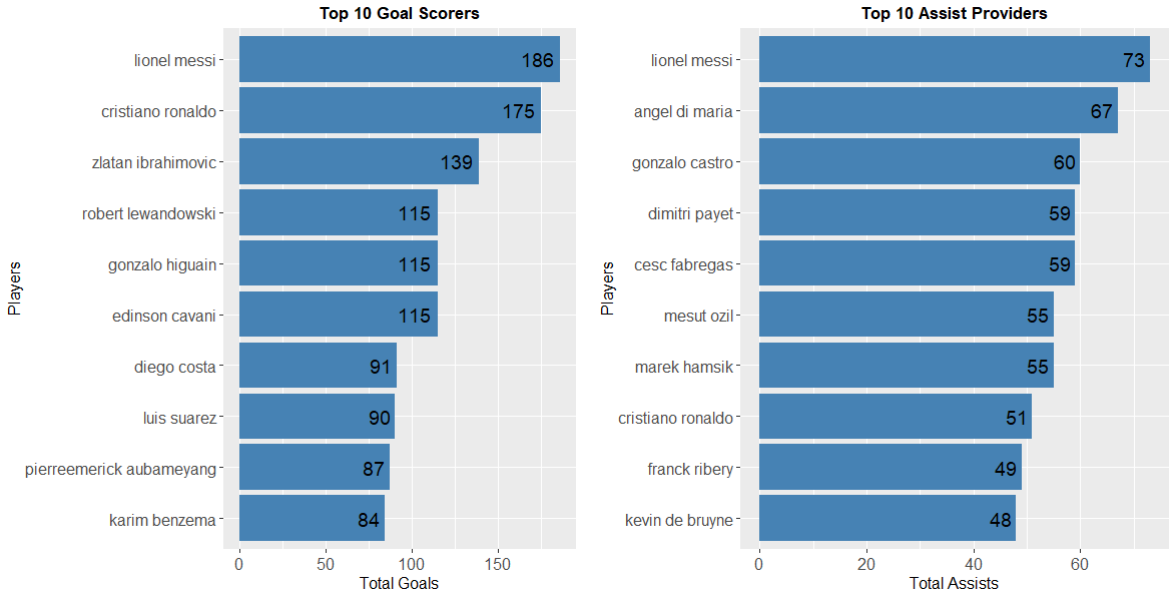


Figure 4: Graph 7 and 8

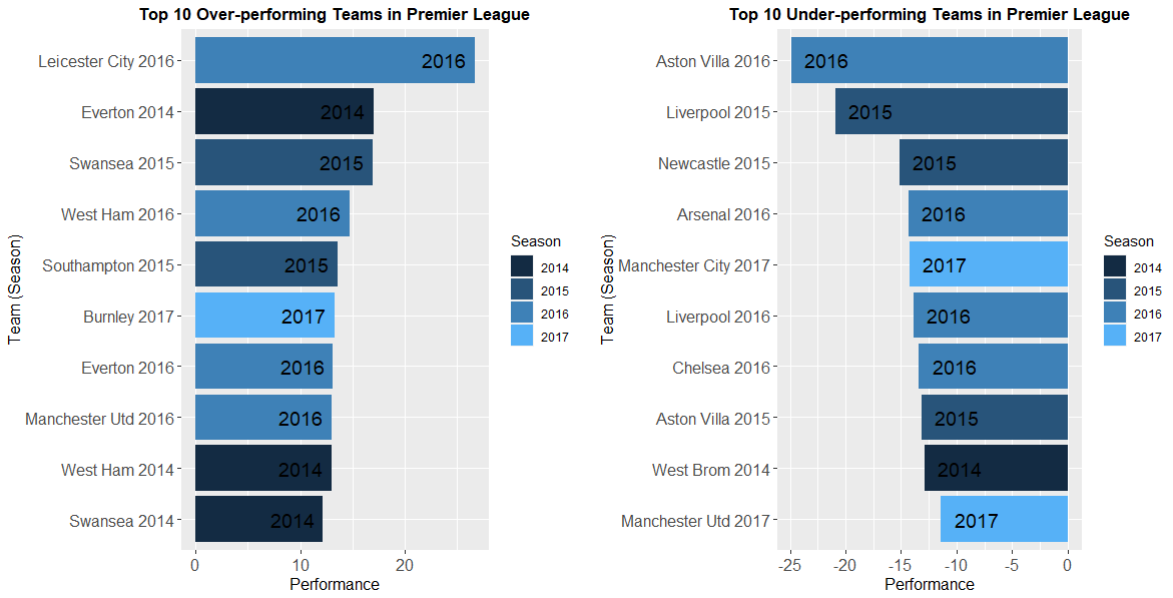


Figure 5: Graph 9 and 10

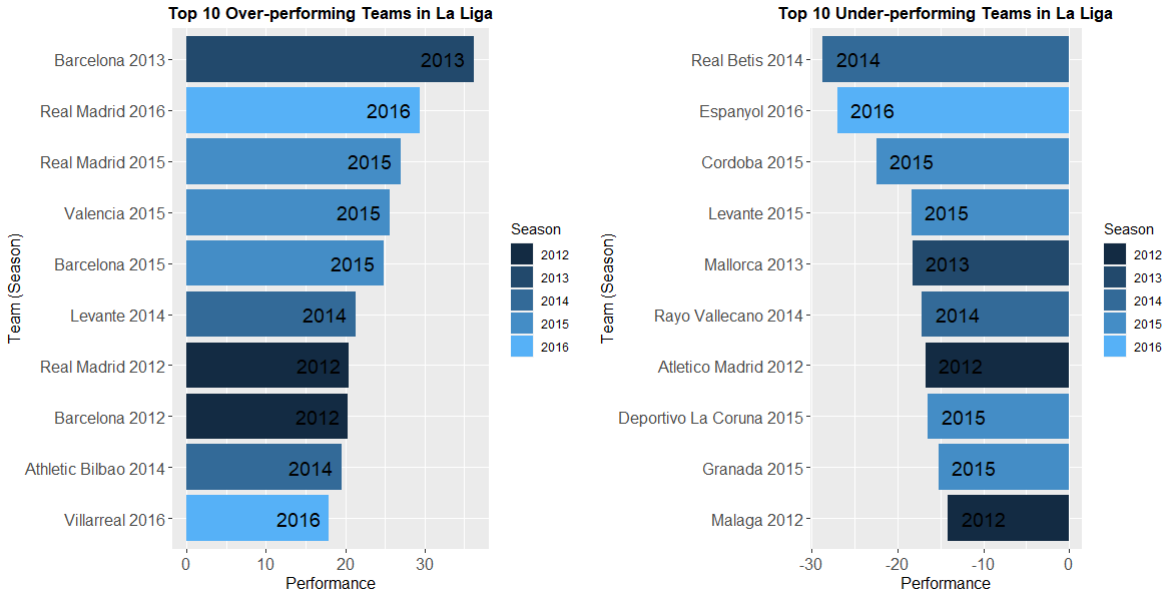


Figure 6: Graph 11 and 12

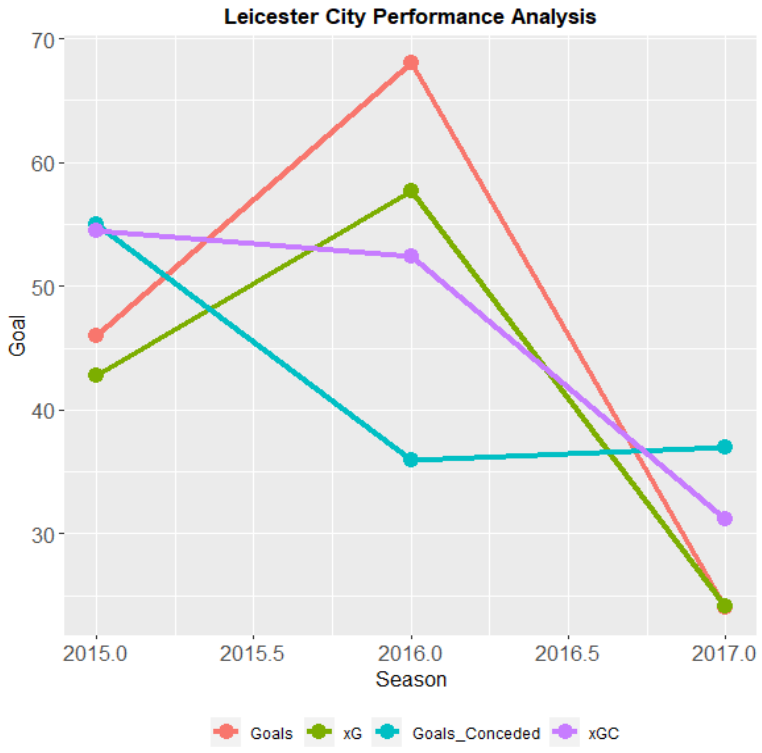


Figure 7: Graph 13

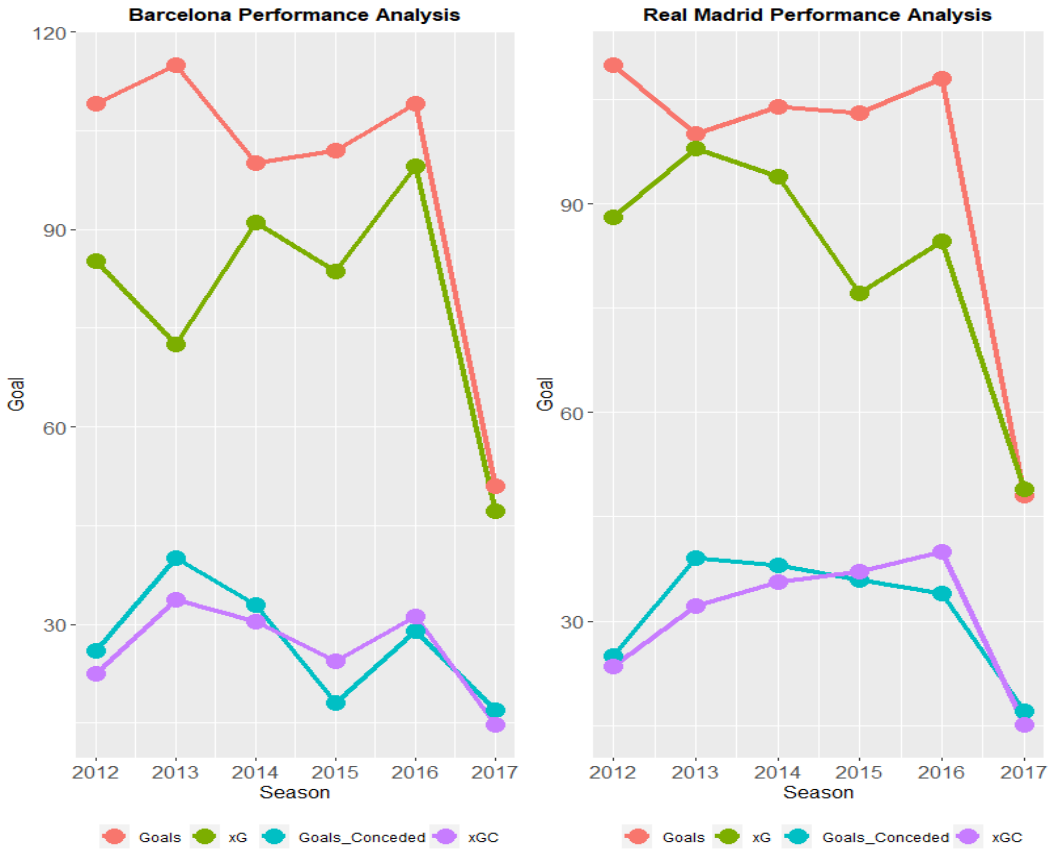


Figure 8: Graph 14

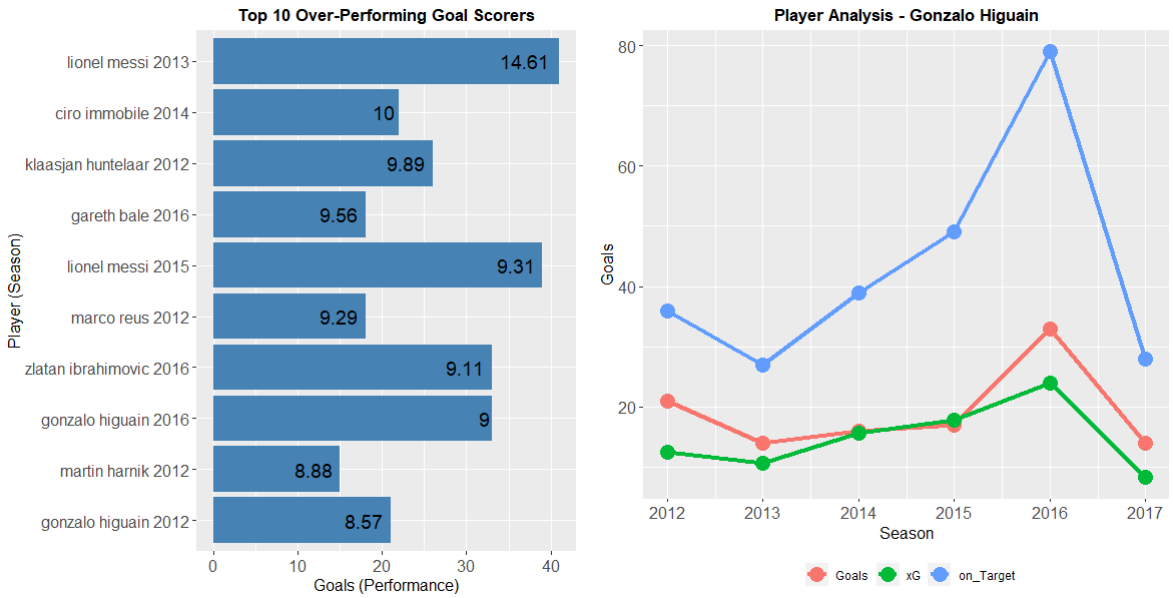


Figure 9: Graph 15 and 16

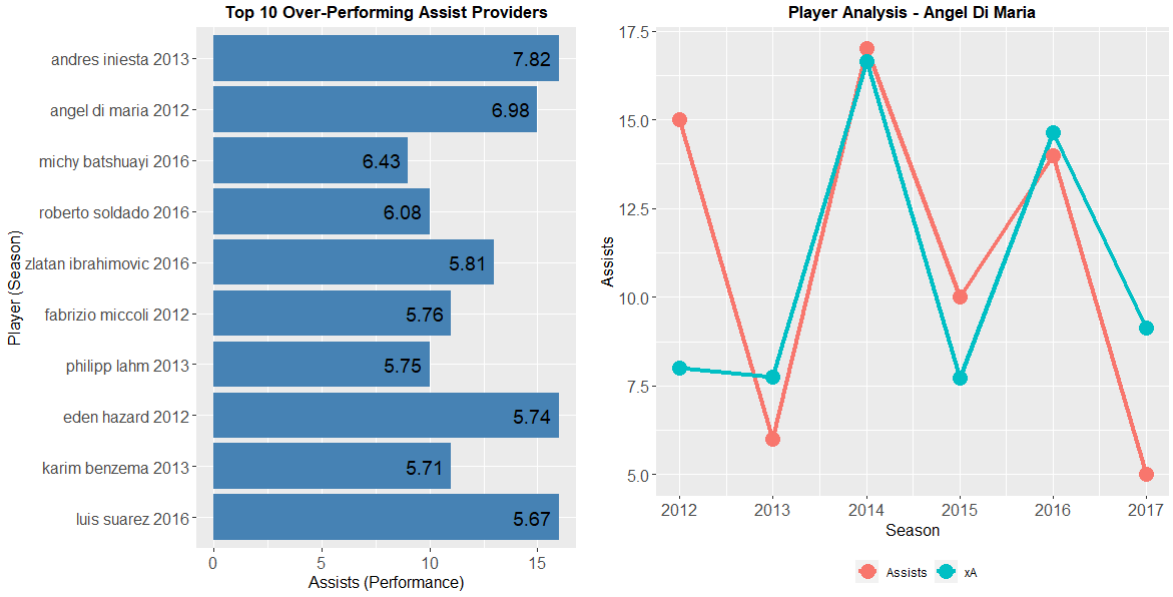


Figure 10: Graph 17 and 18

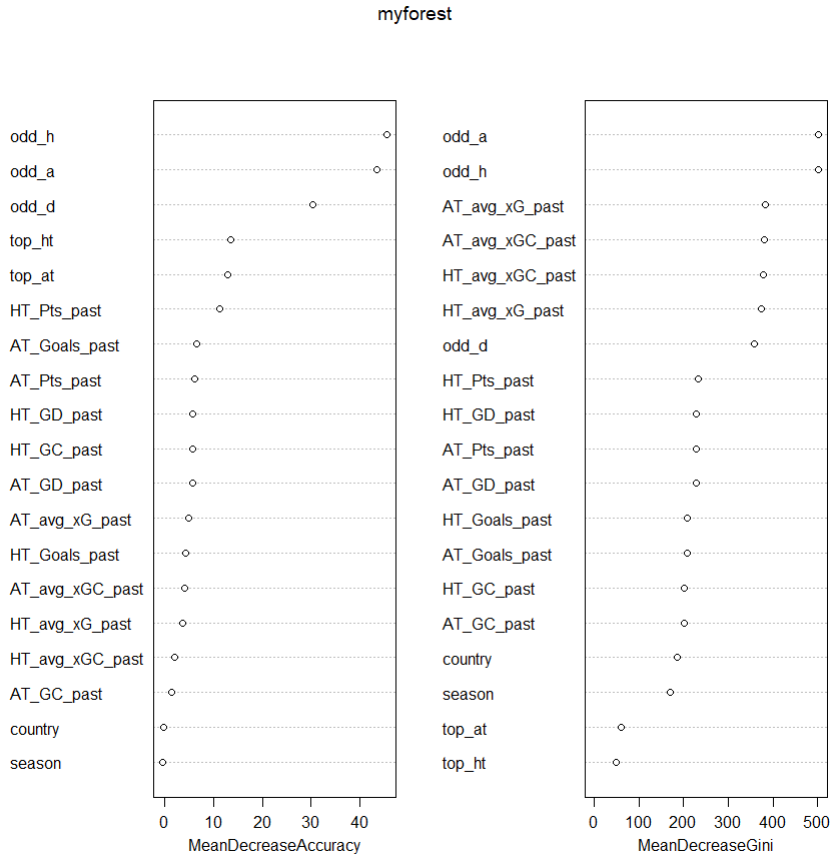


Figure 11: Graph 19