Leeds United 2021/2022

Club Performance: Olof Ekborg-Tanner, Byron O'Gorman, Filip Frostelind, Emilio Jorge
Set Piece Analysis: Amogh Avadhani, Nils Rörby, Liban Muse

January 28, 2024

1 Introduction

Last season Leeds United had their first outing in the Premier League since 2004. They defied the odds and generally low expectations to place a remarkable 9th place, scoring 62 goals and conceding 54, of which 15 came from non-penalty set pieces. Leeds United's manager, Marcelo Bielsa, had the team playing an unorthodox style of aggressive, attacking football. We have analysed the club's performance, using underlying statistics and league data, to predict the performance in the upcoming 2021/2022 season. We have also delved into a deep analysis of our set pieces, to gain an understanding into this weakness and to devise possible solutions.

2 Predicting Season Performance

Going in to a new season, it is important to have reasonable expectations on the team's performance. In this section we will simulate the coming season, based on historic data, repeatedly to get an understanding of what a good or a bad season might look like. We also include two alternate scenarios that could change the season outcome.

2.1 Our Approach

The English Premier League is the most popular football league on the planet, attracting millions of viewers from across the globe. Each season is made up of 380 matches played by 20 different teams, with multiple data providers in attendance to collect almost all of the actions. We had a plethora of data resources available, and we used three different sources in our approach. Firstly, we obtained the match results and betting odds of several seasons from Football-Data.co.uk. Underlying statistics such as xG were sourced from Understat's API and finally, team market value was extracted from Transfermarkt.

To predict the performance of the upcoming season, we simulated multiple seasons and obtained the probabilities of European qualification, a narrow survival from relegation, and relegation. To simulate a single season, 380 unique games were simulated. Match simulation was achieved by predicting the goal scoring rate of both teams, and then simulating the result. These rates were calculated by training a Poisson GLM on the 2020/2021 season's data, with the following formula:

Team's xG + Opponent's xGC + log(Team's Market Value) + log(Opponent's Market Value)

The two underlying statistics used were Expected Goals (xG) and Expected Goals Conceded (xGC). The reasoning behind choosing Expected Goals rather than actual goals is that the former can be seen as a more fair reflection of a team's goal scoring threat, while the latter can be deceptive because of an over and/or under performance in goal scoring. Finally, a minor adjustment was applied to each team's scoring rate based off of their form in the current season.

We can see in Figure 1 that the model has a quite wide distribution over likely outcomes of the season and that there is a reasonable risk of relegation but most likely the team will end up in the middle of the table.

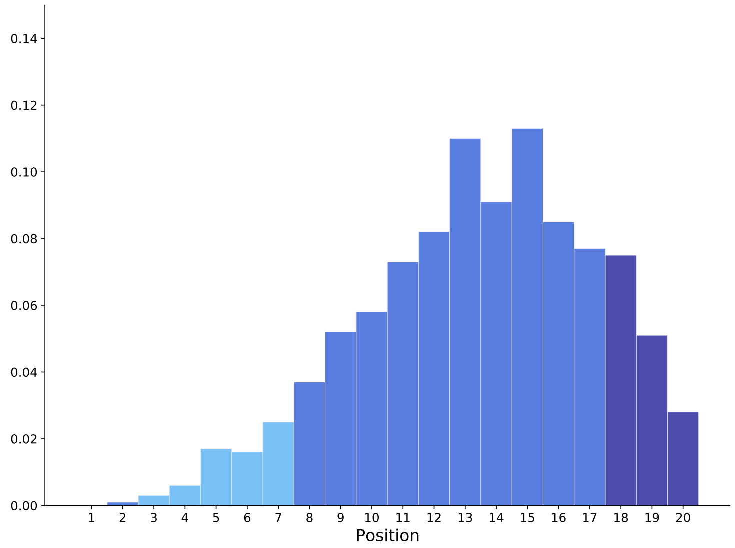


Figure 1: Expected placement for Leeds in the 2021/2022 season using 1000 simulations.

Table 1: An overview of the probability of the different placements in the base model and the alternate scenarios. The different placements have been divided up into Top 7 (qualification to European competition), 15-17th (close to relegation but still safe) and relegation (18-20th).

	Regular model	Injured Bamford	Improved defensive set pieces
Top 7	6.8%	2.4%	17.8%
15-17	27.5 %	33.6%	13.2%
Relegation	15.4%	21.5%	3.7%

2.1.1 How could alternate scenarios impact a Season

We will focus on two possible scenarios that could impact a season. Firstly, what would happen if Patrick Bamford gets injured? Bamford had an xG per 90 of 0.54 over the season, playing in every match with a total of 3085 minutes. If Bamford gets injured and only plays a third of next season and gets replaced by a player with an xG per 90 of 0.40 (such as Rodrigo, who is in the squad and sometimes plays as a forward), this results in a reduction of 3.2 xG over the season. Secondly, we look at the scenario where Leeds sign a new set-piece coach. Their introduction will reduce xGC from set-pieces to the league average of 7.5, from last season's tally of 15.5.

In Table 1 we can see how the alternate scenarios impact the probabilities of ending up in the different places of the table. A full visualization of the distributions can be found in Appendix A. It is clear that the quite significant change in xGc from the defensive set piece coach has a very large impact on the season. Relegation becomes very unlikely and a significant change of ending up in a top 7 placement (and European qualifications). On the other hand, if Bamford gets injured a significant part of the season, then Leeds are in great trouble with a significant probability of being relegated. A good backup striker could therefore be a solid strategy for the transfer market.

2.1.2 Alternative methods for simulating a Season

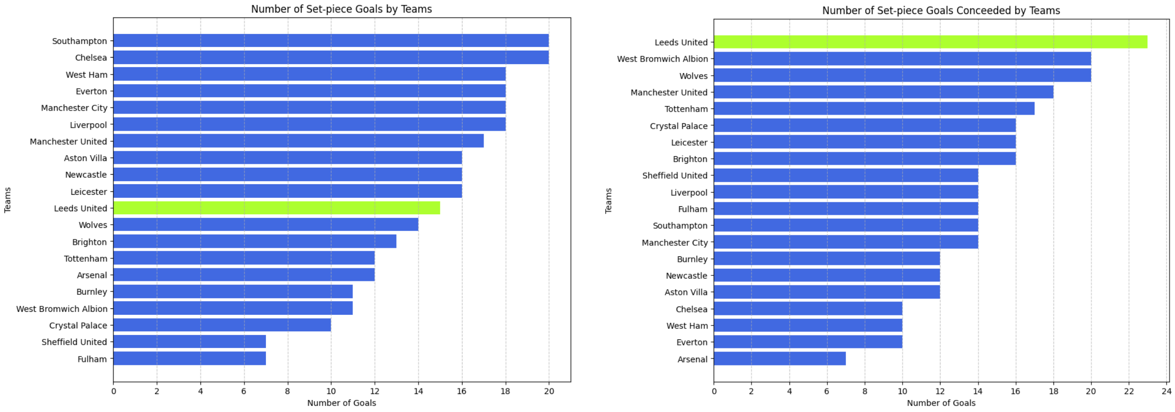
An alternative to building a simulator the season yourself is to use simulators that already exist. One alternative for this is to use games such as Football Manager or and let the game play through the season. Football manager has a very complex model that incorporates everything from player transfers to injuries which makes it much more flexible than a simple Poisson model based on last seasons stats.

Considering that Football manager is accurate enough to be used as a scouting tool[3] we can assume that the stats they use are representative of reality.

Another alternative is to base the model on the assumption that the betting markets are reasonably accurate. While this is not always perfect, there is a drive for betting companies to have accurate odds since they would otherwise lose money and the betting markets will tend to a "fair" price for event. In odds from July of 2021[1], we can find that the odds for relegation are 11/1, with 10 teams being more likely to be relegated. This indicates that a relegation for Leeds would be a significant upset. Additionally, the odds for winning were 100/1, making it very unlikely. These odds seem to match decently with the results obtained using simulations in Figure 1.

3 Set Piece Analysis

Set pieces, including free kicks, penalties, corners, and throw-ins, are crucial moments in a football match where teams can either capitalise on or be vulnerable to scoring opportunities. They require a blend of tactical preparation and individual skill. Analysing performance in these scenarios helps clubs identify strengths and weaknesses, enabling them to adjust strategies to maximise scoring chances and minimise risks.



(a) Set Pieces Scored

(b) Set Pieces Conceded

Figure 2: Comparison of Set Pieces Scored and Conceded (including penalties)

Here we define set pieces as corners, free kicks and throw ins. Penalties are uninteresting to analyse when they have been awarded. The 20/21 season Leeds let in the most goals in the league from set pieces with 15, 5.5 set piece goals above the league average. 11 of these came from corners. Clearly, this is something which can and probably should be improved upon. We will be looking at corner defence and potential causes for the many goals let in, divided in to two sections, tactics and players.

3.1 Marcelo Bielsa and Tactics

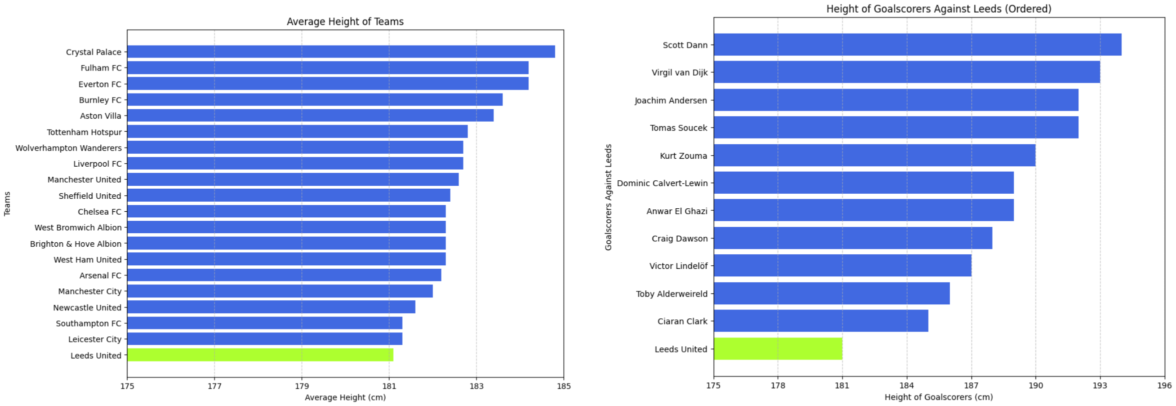
The Argentinian managed Leeds United this season and he is a very respected manager. First, we take a qualitative look at his other teams, to see if the problem has a tactical root. In Table 2 we see how most of his other teams are better than or about average, except for an *Annus horribilis* at Athletic Bilbao in 12/13 and Lille in 17/18. In Lille, Bielsa managed only 13 games. From this we can deduce (unsurprisingly) that Bielsa's tactics aren't completely terrible and that there are other aspects to Leeds' misfortune. It can also be noted that Leeds in 20/21 did not let in more goals than the Championship average the season before.

3.2 Players and Aerial Presence

In Table 3 it is made clear that Leeds, while defending corners, is more prone to let in goals to dominant aerial players than the average team. It is no surprise that aerially strong players are strong in corner

Season	Team	SPGs	League avg	Δ
11/12	Athletic Bilbao	17	13.5	+26%
12/13	——— " ———	25	13.75	+82%
13/14	——— " ———	11	11.25	-2%
14/15	Marseille	5	9.45	-47%
16/17	Lazio	9	11.6	-22%
17/18	Lille	13	9	+44%
18/19	Leeds	13	14.4	-9.7%
19/20	——— " ———	12	15.2	-21%
20/21	——— " ———	15	9.3	+61%

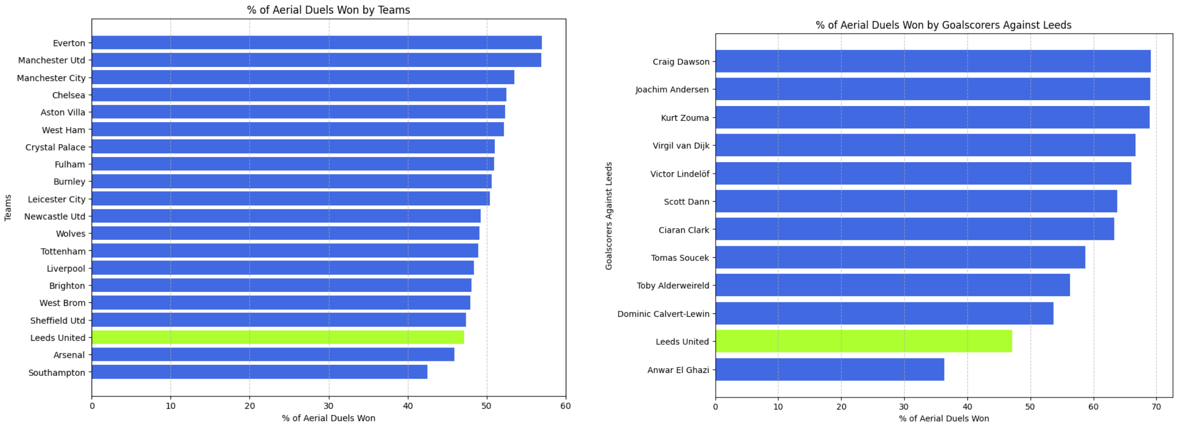
Table 2: Set Piece Goals (SPG) let in by teams managed by Bielsa, with league averages, (https://www.whoscored.com)



(a) Average Team Height

(b) Height of Goalscorers

Figure 3: Comparison of average team heights and heights of players that scored by corners



(a) Aerial Duels Won by Teams

(b) Aerial Duels Won by Goalscorers

Figure 4: Comparison of Aerial duels won by Team and Aerial duels won by players that scored by corners

situations, but they are even stronger against Leeds than other teams. The weak aerial play of Leeds has become an exploitable weakness, which strong opposition capitalise on.

Against Leeds	61.08%
Against all teams	53.46%

Table 3: Percentage of aerial duels won for corner scorers

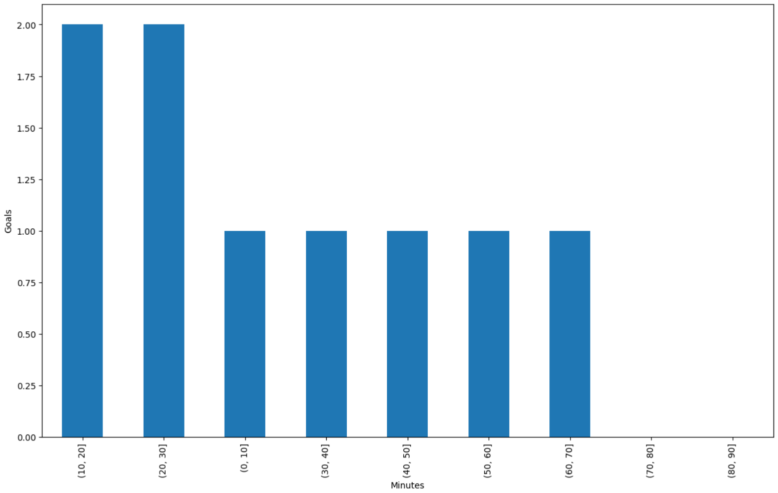


Figure 5: Corners conceded by Leeds by minutes

3.3 Corners Analysis

To find out the number of corners conceded by Leeds in terms of game minutes, the game minutes are divided into equal parts called bins in increments of 10. The conceded corners are then grouped into these bins accordingly. Figure 5 shows that Leeds United are more prone to conceding corners in the earlier stages of the game.

3.4 Pitch Control

Marcelo Bielsa uses a man-marking strategy to defend corners. To understand the spaces which are controlled by Leeds and the opposition, a pitch control model is devised to analyse the happenings during corners using tracking and event data. Pitch control models help us understand which team/players control the spaces at any given moment.

Figure 6 shows one incident of defending corners by Leeds. Leeds United are the team in blue while the opposing team, in this case is Aston Villa, are in red. From analysing pitch controls for different corner circumstances, it is clear that for a man-marking strategy employed by Bielsa, the players are adequately positioned to control the spaces.

It is easy to deduce, however, that a man-marking system will fail if all of your players are weaker than their respective attacker. This is what seems to be the case for Leeds.

3.5 Set Piece Summary

Leeds conceded the most goals from set-piece situations in the league. However, this was their first year in the top division and the 15 set piece goals would not have been much of an outlier had they stayed in the championship. It is not Leeds who are bad at defending set-pieces, rather the other teams who are good at it. However, if Leeds want to stay in the top flight, they will probably have to improve. A short squad which is weak in aerial duels combined with a man-mark defence is to blame. While transfers in is one way to improve, it is probably easier to alter the tactics to a zonal variant.

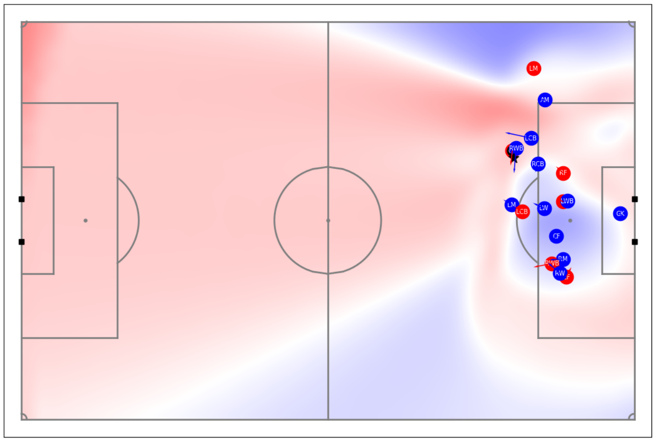


Figure 6: Pitch control for Leeds United vs Aston Villa - corner

4 Summary

This project has had two parts, simulating the next season and analysing set pieces. From set piece analysis it was found that Leeds concede the most goals in the league from set piece situations and that this is a consequence of a short squad and defensive tactics not suited for this. If set pieces were to be improved upon, the simulations give Leeds a much lower risk of relegation next season and a higher chance of a top 7 finish. The importance of Leeds' striker Bamford was also discovered to be significant when looking at the probabilities of a top 7 finish and avoiding relegation.

References

[1] *Premier League 2021/2022 odds: Man City title favourites but Liverpool and Chelsea will challenge, Watford and Norwich tipped for immediate relegation to Championship.* https://talksport.com/football/896263/premier-league-2021-2022-odds-bookmakers-man-city-title-favourites-chelsea-liverpool-watford-relegation/. Accessed: 2023-10-20.

[2] *Simulating Matches Example.* https://soccermatics.readthedocs.io/en/latest/gallery/lesson5/plot_SimulateMatches.html.

[3] *Why clubs are using Football Manager as a real-life scouting tool.* https://www.theguardian.com/technology/2014/aug/12/why-clubs-football-manager-scouting-tool. Accessed: 2023-10-20.

A Technical Appendix - Club Performance Model

A.1 Model Description

The model used to simulate and predict results of football games is a generalised linear model (GLM), using Poisson regression. Poisson regression is useful if the dependent variable is discrete, that is for example goals in football games. It assumes the events of the variable to be independent of each other, meaning if one goal has been scored it is neither more nor less likely that another goal is scored. One could of course argue that goals affects the state of football games and therefore goals are not completely independent of each other, but for simplicity the assumption of independent events is made in our model.

When building our model we tried using four input parameters as predictors for the number of goals scored by a team in a game, and in extension the outcome of a game. The output is a goal scoring rate for a team, which is used to sample the number of goals scored from a Poisson distribution. The model generates a scoring rate for both the home and the away team, and their respective goals scored are then sampled. Since the numbers of goals are sampled from a distribution the outcome is not deterministic and the outcome of a game will vary from sample to sample, just like a real game between two teams would have different results if they played each other multiple times.

The four input parameters to the model are *xG* for a team, *xG conceded* (xGC) for the opposing team, the *market value* (MV) for the team of interest and the *market value of the opposing team.*

- The *xG* and *xGC* values for each team are set as their average xG/xGC per game from the previous Premier League season, as a simple measurement of how good a team offensively and defensively. To acknowledge that many teams performs differently when playing at home compared to away, the *xG* and *xGC* values are separated for home and away games. Hence each team gets assigned a value for xG_{home}, xG_{away}, xGC_{home} and xGC_{away}. As the model only takes *xG* and *xGC* as variables, we set xG_{home} and xGC_{away} for the opposing team as values if a team is playing at home, and vice versa xG_{away} and xGC_{home} if they play an away game.

- The *market value* (MV) of all teams are set to the value (in £M) for the entire squad at July 1st the summer before the season to simulate, according to Transfermarkt. The model doesn't update the market values during the season, so transfers made in the January transfer window are not considered.

Before simulating a match, the generated scoring rates of each team are adjusted according to their form in the previous six games in the current simulated season. This adjustment is done by multiplying the scoring rate by a form value, which is accomplished by a linear transformation of the team's point tally from their previous six games to a value between 0.9 to 1.1. In other words, the team's scoring rate can be increased or decreased by a maximum of 10%. Intuitively, a team's scoring rate is only adjusted once they have played six games.

A.1.1 Promoted Teams

Since there are three newly promoted teams in the Premier League each season, they must also be assigned xG, xGC and MV values. For MV this is done in the same way as for every other team, using Transfermarkt. But xG and xGC is a bit more tricky, since it would be incorrect to use values from their previous season in the Championship where the competition is not as good. For each season simulated this is handled by sampling a xG_{home}, xG_{away}, xGC_{home} and xGC_{away} from the first season in the Premier League for one of the nine teams promoted in the last three seasons. That is Leeds, West Brom and Fulham for 2020/21, Norwich, Sheffield and Aston Villa for 2019/20 and Wolves, Cardiff City and Fulham for 2018/19. For each simulated season these values are sampled, giving the model some different parameter variations in each simulation.

A.1.2 Scenarios

To implement the scenarios we tried multiple alternatives. First we tried adjusting the *xG* and *xGC* values that went into the GLM model. An issue with this is that the model does not have a large weight on these parameters (as market value seems to be a better predictor, more on this in chapter

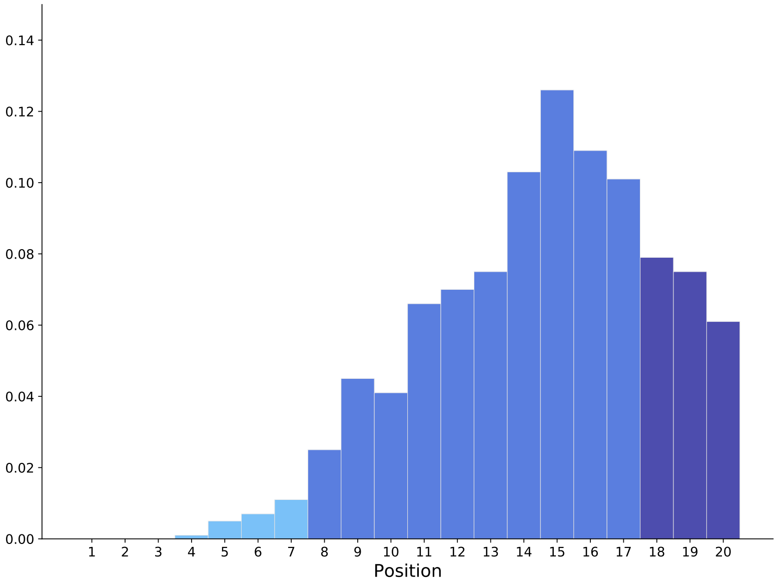


Figure 7: Simulated league position with Bamford injured during the season with 1000 samples.

A.2.1) such that a change of 0.1 in xG from the previous season will give much less of an impact on the next season. Instead we chose to make the impact directly in the Poisson sampling, i.e. the model prediction is unchanged but in the sampling step we add the change. Intuitively, this makes sense as the mean of the Poisson distribution is the parameter in question and as such it can be interpreted as the expected goals. If Bamford is injured for the game, it is reasonable to adjust the expected goals for Leeds in the game accordingly. The same goes for reduced xGC from set pieces.

The full prediction of how Leeds will perform in the scenarios can be seen in Figures 7 and 8.

A.2 Evaluating Performance

One way to evaluate the performance of a developed model is to look at the proportion of games it predicts correctly. To narrow this down a bit the focus here is on the final outcome of a game, that is if the home team won, if it was a draw or if the away team won (1X2), instead of looking at the exact number of goals. So for each game simulated the outcome is compared to the true outcome of a game. To get the proportion of correctly predicted games during a season, the number of correct predictions are divided by the total number of games.

A.2.1 Model Comparison

To decide on the most suitable combination of parameters, we ran 250 simulations of each combination of variables on the 2020/2021 Premier League season and compared our models predicted result of a game to the actual outcome to get an accuracy score. We also added a baseline model based on the *Simulating Matches* example from the course webpage [2], which simply uses the team and opponent as input variables, as well as the home advantage. The relegated teams are replaced with the promoted teams, so the baseline model basically assumes the upcoming season will be very similar to the previous season.

As seen in Table 4, the baseline model still performs better than the model using only xG and xGC, implying that these statistics alone don't contain enough information to predict the outcome of games as well.

The model using *market values* performs slightly better. This is not surprising as the value of each player is a combined measure based of their overall ability and quality and will increase if they and

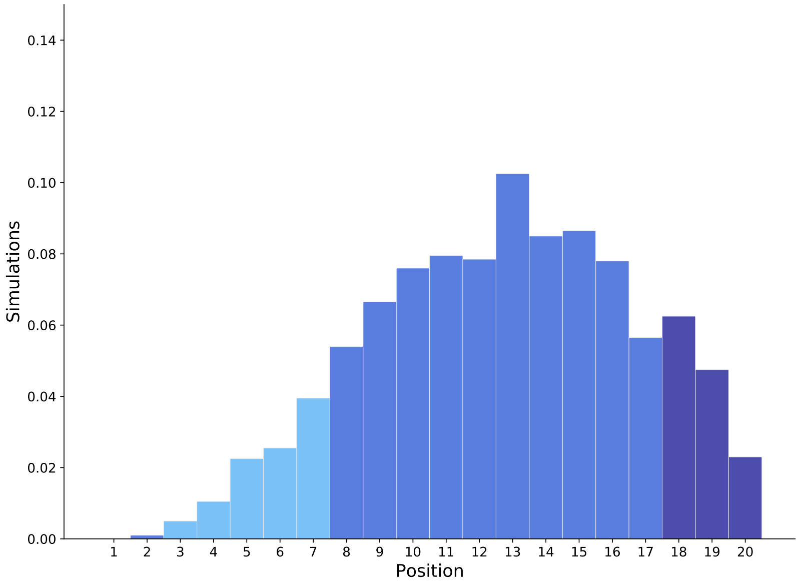


Figure 8: Simulated league position with improved set piece defending with 1000 samples.

Model	Avg. Accuracy	Std
Baseline	40.08%	2.36%
$xG_{team} + xGC_{opp}$	36.17%	2.40%
$MV_{team} + MV_{opp}$	40.23%	2.29%
$xG_{team} + xGC_{opp} + MV_{team} + MV_{opp}$	39.72%	2.25%
$xG_{team} + xGC_{opp} + \log(MV_{team}) + \log(MV_{opp})$	**40.45%**	**2.47**%

Table 4: Average accuracy and standard deviation for each model, based on 250 simulations comparing the predicted results of each match to the actual results of the 2020/2021 Premier League season.

their team have performed well recently. Using the *market values* is hence equal to using an overall estimate of each team's quality (while player's values are also being adjusted for age etc., which may not be as relevant for predicting performance). Also the correlation between a team's finances and results are well established in elite football.

What surprised us was that the performance decreased slightly when adding xG/xGC to the model. The reason for this could be that the market values are more than 100 times as big as the average xG values are, making the market value the dependant feature and xG more of a noise. To account for this we tried to use the logarithm of the market values, to decrease its impact on the model and to better capture that impact of expenditure is not linear. The performance of this model increased and was the best one, however still marginally, and it also had the largest standard deviation so the performance might be slightly more fluctuating.

By digesting the summary statistics of the chosen model, shown in Table 5, we can however see that even with a logarithmic value, *market values* are still the statistically significant variables whereas we cannot conclude the same for the xG and xGC values. Otherwise the coefficients tells us that a higher xG the previous season increases the goal scoring rate for a team, whereas a higher xGC value for the opponent decreases it. Similarly if a team has a high market value it's expected to score more goals, and if they're facing an opponent with a high market value it will be more difficult to score. All of this is of course very expected and seems like common sense, but it's still encouraging that the model performs accordingly. Furthermore it's interesting that the absolute coefficient values for xG_{team} is higher than the one for xG_{opp}, and similarly $\log(MV_{team})$ is larger than $\log(MV_{opp})$. This indicates that a teams own factors for scoring goals is stronger than the factors of the opponent (according to this model).

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0085	0.422	-0.020	0.984	-0.835	0.818
xG_{team}	0.1048	0.083	1.265	0.206	-0.058	0.267
xGC_{opp}	-0.0744	0.104	-0.715	0.474	-0.278	0.129
$\log(MV_{team})$	0.2478	0.050	4.987	**0.000**	0.150	0.345
$\log(MV_{opp})$	-0.2092	0.044	-4.750	**0.000**	-0.296	-0.123

Table 5: Summary statistics of model with logarithmic market values and xG/xGC.

It's noticeable that accuracy for our model is only a bit higher than 33%, which would be the expected result if we attributed the results randomly between win, draw and loss (assuming they are equally probable). This might be an indicator of our model being poor, or that predicting individual football matches is a complex task, where the statistically better team doesn't always win. To benchmark our model against a more established predictor we compared it against the odds market.

A.2.2 Comparing to Odds

To know if the performance of the model is reasonable it would be ideal to have a benchmark to compare to. It lies in betting companies' interest to have as good models as possible in order to predict games as well as they can beforehand. If they would not be state-of-the-art, then they could lose out on a lot of money. Hence the betting odds from bet365 have been used for each game of the 2020/21 season (https://www.Football-Data.co.uk), to evaluate how well the model would have performed during the past season. European odds are simply transformed to probabilities by

$$p = \frac{1}{odds}. \tag{1}$$

But the betting companies also adds a margin in their odds in order to assert profitability. Hence the fraction in equation (1) would need to be normalised in order to actually represent a probability if using bet365 odds. Assuming the margin is equal for each outcome, this gives

$$p_\alpha = \frac{\frac{1}{odds_\alpha}}{\frac{1}{odds_1}\frac{1}{odds_X}\frac{1}{odds_2}} \tag{2}$$

where α could be 1, X or 2.

These probabilities could then be used to simulate what outcome bet365's model predicts. This is done by a simple Monte Carlo simulation. The predictive performance of the developed model and the bet365 odds can be seen in Figure 9.

As can be seen in Figure 9 the model and bet365 have the same mode, but bet365 has a slightly higher mean value of 0.414 compared to the model's 0.405. Their model is hence better but not by a large margin. It should also be noted that their odds for each game are updated close to the start of each game, meaning they have access to additional information such as injuries and suspensions. It turns out that our accuracy of only 40.5% isn't as bad as was initially assumed.

A.3 Potential Model Improvements

A general problem with these types of simulations is that they can only be as good as the model, as this governs the outcome of each match. One suggested improvement could be to also use the change in Transfermarkt values between seasons rather than just the current value, as this is some way of indicating a change compared to the previous seasons squad who achieved the xG and xGC used in the model.

An inherent problem though is that it is very hard to predict how a team will perform in the next season, there can be large changes to a squad but even in cases where everything seems the same the teams performance can change greatly.

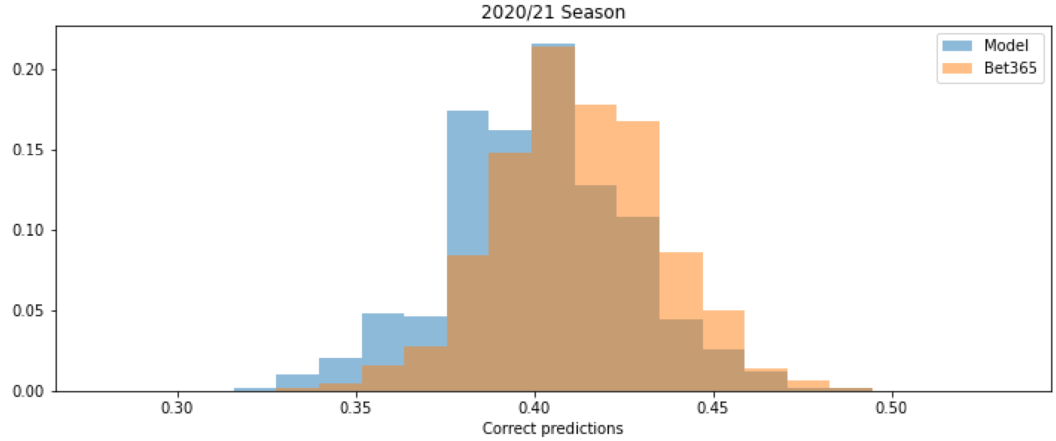


Figure 9: Normalised histograms over amount of correctly predicted outcome of games per season for 500 simulations of the 2020/21 season. Outcome is defined as home team win, draw and away team win (1X2). Exact goals scored are not accounted for. Bet365 predictions are based on Monte Carlo simulations using the bet365 odds for each game.

B Technical Appendix - Set Pieces

B.1 Aerial duels

Percentage of aerial duels won is calculated simply by taking won aerial duels over total aerial duels for every player. An ELO-model could be used instead, where each player has a score which is adjusted after every duel depending on the difference between the player's scores. Winning a duel against someone who usually wins duels will give you more points. Such a model would need polishing to actually tell more than just the percentage won. If a player consistently takes duels against weak opponents he will score higher here than he maybe should, but not many do that. Only tall strikers taking all duels against tall centre backs will get a percentage which is significantly lower than their actual performance.

B.2 Pitch control model

Pitch control at a location in the field is defined as the probability that a player/team will gain control of the ball if the ball moves to that location.

The probability of which player/team will control the ball is calculated for every location on the pitch. Below is a breakdown of the items to be calculated. For every location:

- Calculate the time taken by the ball to reach that location.
- Calculate the time taken by each player to reach that location.
- Compute the total probability that each team will control the ball at that location.

The arrival time can be calculated by:

$$Arrival\,time = \frac{End\,position - Start\,position}{Speed}$$

where Speed can be the Ball speed or Player speed.
The following assumptions are made in the calculations,

- Average ball speed = 15m/s
- Max speed of players = 5m/s
- Max acceleration of players = 7m/s^2

Based on the above values, using a sigmoid distribution model for uncertainty in player arrival time with a standard deviation of 0.45, the probabilities that each player will control the ball is calculated. These probabilities are determined for each point in the field. The pitch grid is then plotted based on these calculated probabilities for each team (players).

B.3 Potential improvements

The pitch control model could be improved by:

- Adjust the model to account for aerial strength of the players, but this may be complicated since it depends on the path of the ball
- Considering players in offside positions. Since players who are standing in an offside position do not matter, they may be excluded from the space control analysis.
- Consider goalkeepers differently. Since goalkeepers can also use their hands, the way they control space is different than outfield players, especially during corners.