An insight into Chelsea's performance as of summer 2021

Group Project of Mathematical Modelling of Football, Autumn 2022

David Andersson, Edd Webster, Leyla Abdul Kader, Lukas Schmid,
Mattias Holmström, Paula Borst, Tim Solig, Yash Karle

November 4, 2022

Contents

1 Introduction

1.1 Problem statement

Sections 2-4 will serve as a season report breaking down Chelsea's performance during the 20/21 Premier League season. We set out with the aim of answering the following two questions with the overarching aim of Chelsea managing to lower the gap to Man City and Liverpool in terms of their league position for the next season.

1. How can Chelsea be unpredictable in the final 3rd and unlock low-block opponents?
2. How can Chelsea manage to put games to bed (2+ goals lead) and avoid high pressure situations at the end?

Chelsea finished 4th behind Manchester City, Manchester United and Liverpool FC and therefore qualifying for the Champions League. After a mixed first half of the season under coach Frank Lampard, in which Chelsea finished ninth, 12 points behind the lead, Tuchel was appointed as the new head coach (January 19th). Under Tuchel Chelsea managed to get back on track by gaining 38 points in the second half of the season finishing 19 points behind the leading team Manchester City. Due to the coaches' replacement and the resulting changes in the game plan, the first and second half of the season are considered independently in some parts of the analysis.

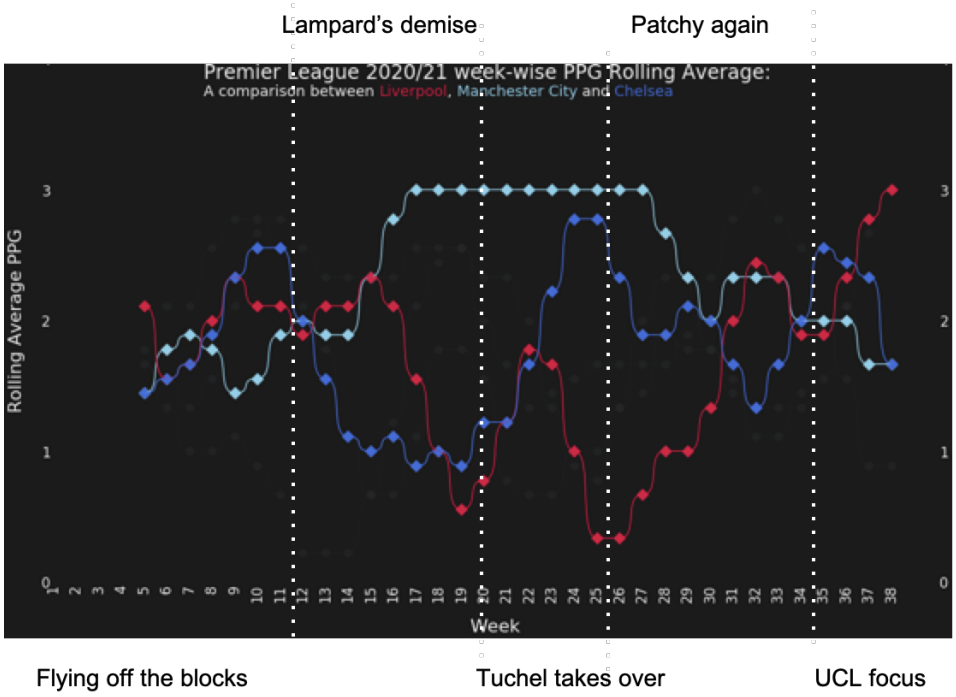

Figure 1: EPL 20/21 points per game rolling average

2 Club Performance Chelsea: Yash Karle, Lukas Schmid, Tim Solig

2.1 Narrowing down the focus

Keeping those high-level objectives in mind, we then drill down further to identify specialised areas of the team where we can focus in order to extract the most potential value.

2.1.1 Offense vs defense

Firstly, Chelsea's offensive and defensive performances will be compared to that of Liverpool's and Manchester City's across the whole 20/21 season. In order to achieve this, we build a naive first pass regression model to gauge offensive (goal scoring rate) vs defensive (goal conceding rate) performances for each team. It was observed that Chelsea had the 2nd least goal conceding rate behind only Manchester City indicating that in a short time period Tuchel had managed to strengthen the defensive performance of the team. On the contrary, Chelsea had only the 8th best goal scoring rate behind 5 out of the other big 6 teams and even Leeds and Leicester. Assuming this hypothesis, Chelsea's main focus area for improvement was identified to be towards maximising quality and quantity of attacking output at a team level. See appendix for a detailed analysis where we define team level KPIs and feed further insights gained from that into the following sections.

2.1.2 Analysing Chelsea's playing style

Next, we do a deep-dive on Chelsea's playing style for identifying how Chelsea create chances by analysing situations where most xG was generated. It was evident that under Tuchel, Chelsea had managed to generate significantly higher open-play xG compared to the first half of the season. Still, there remains scope to improve even further and hit the levels of Manchester City and Liverpool. See appendix for charts.

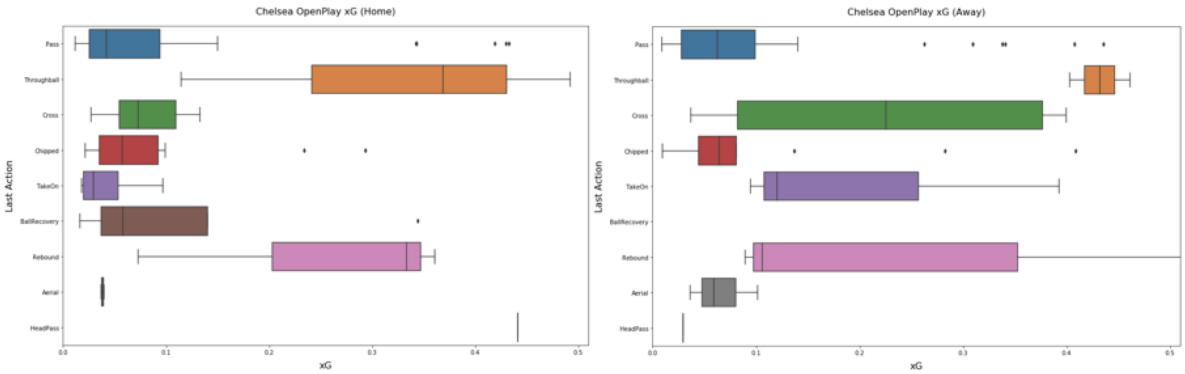

Figure 2: Open-play chance creation Chelsea home/away

Additionally, we then try to identify signals within Chelsea's xG creating avenues to help us spot differences between home and away games for the 2nd half of the season. This was motivated by the fact that Chelsea drew a total of 10 times, out of which 6 were at home and were often accused of honing predictable attacks against low-block opponents (for instance in the draws against Southampton, Villa, Wolves, Brighton). As seen in the plot above, at home, Chelsea tend to have a slower build-up play with more emphasis on through-balls and ball recoveries. Whereas for away matches, Chelsea had a direct attacking approach involving more crosses and take-ons in the final 3rd. This feeds into fine-tuning the recommendations of the tracking data scouting.

2.2 Simulation model

In order to identify areas of improvement within the set-pieces and tracking data scouting, a Poisson Regression model, that simulates the outcome of upcoming season based on relative goal scoring rates, was built. Given that we would have an estimate of how the league could turn out the following year, we can then know more about the areas to improve from a recruitment perspective and which aspects of the team's tactics to tweak from a coaching point of view.

2.2.1 Baseline model

When simulating the league, we first built a baseline model to feed in features related to chosen team KPIs like quality (xG) and quantity of shots and shots on target, corners and set pieces chance creation and variables to account for home advantage as well as factor in relative strengths of teams and opponents. For instance, the model learnt that playing against Manchester City or Liverpool had a negative impact on your goal scoring rate. Another such variable which took into consideration the superior attacking prowess of the big six clubs and favoured them with a higher coefficient. Teams playing at home had a better goal scoring rate than than teams playing away. As expected, teams with more shots on target and a higher xG value had a higher scoring rate as these variables were found to be directly correlated. Lastly, it was interesting to see that having more corners or xG accumulated from set pieces wasn't often a trait of high scoring teams and possibly suggests that as an avenue for teams in the lower half of the table to score goals from. See appendix for more details on the data used, coefficients for the baseline model and simulation output for the 21-22 season expected points league table [1].

2.2.2 Alternative scenarios

Next, we used the baseline model to simulate for three alternative scenarios to gauge the effect of different variables to feed into improvements to Chelsea's attacking output. In the three simulations we tried to answer the question what would happen if:

1. Scaling up home team's xG by 0.1% of stadium capacity in 1000s for all teams. e.g. Chelsea playing at Stamford bridge (capacity 40k) expect to have +0.4xG added for their home matches
2. If Chelsea had a 10% increase in overall shots on target per match
3. If Chelsea had a 10% increase in xG per match against the teams outside of top six

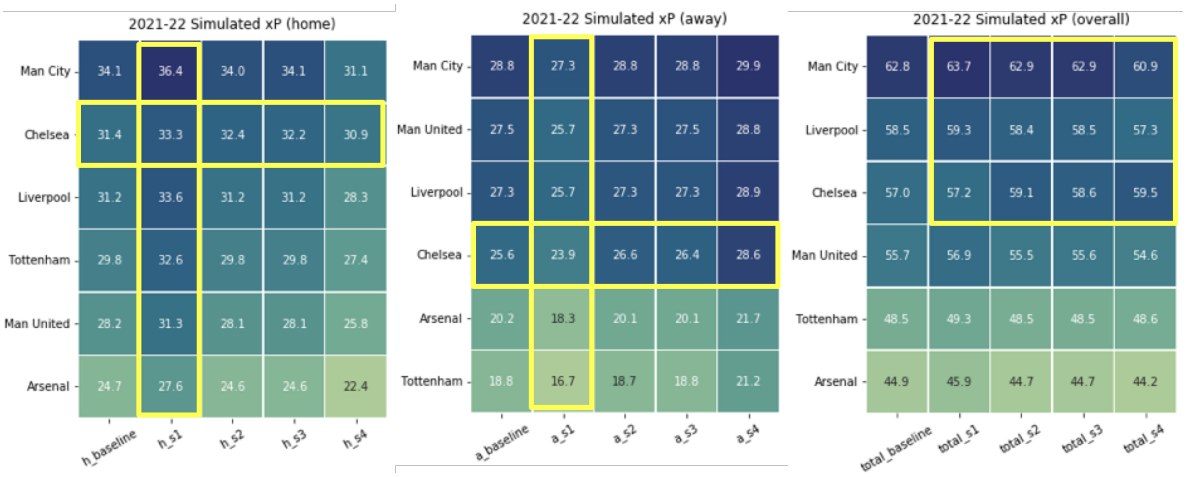

Figure 3: Alternative scenario simulation results

As seen in the image above, s1, s2, s3 correspond to each of the three alternate scenarios and we look at their impact on teams' conditional probabilities of winning home and away matches reflected by their expected points tally. Effect of s1 was quite obvious and we successfully managed to see the proportionate (based on stadium capacity) impact of crowds returning to the stadiums for the 21/22 season on the home expected points (positive uplift of +6% for Chelsea) and away expected points (negative uplift of -6.6%). This looks reasonable given that Chelsea's stadium capacity is only 10th best in the league. Next, for each of s2 and s3, Chelsea improved their expected points tally and especially had a relatively higher uplift in Chelsea's away expected points (+3.9% and +3.1%) compared to home(+3.1% and +2.5%). Lastly, we look at the combined home and away table where we can clearly see that if all of the above 3 scenarios simultaneously come true, Chelsea had a 65%(up by +35%) chances of a top 2 finish. Therefore, this would help the team not only catch-up with Liverpool in the final standings but also challenge Manchester City for the title[3].

3 Set pieces: Leyla Abdul Kader and David Andersson

Set pieces or set plays as they're often referred to, can be defined as plays where the ball returns to open play after a stoppage. Corners, free-kicks and throw-ins are three types of set pieces, which we will investigate further in this section. In the 2020/2021 season of the premier league, Chelsea scored twelve out of fifty-eight goals from a set piece initiated play [9]. Thus, around 20% of Chelsea's goals were related to a set piece, highlighting their relevance in football.

3.1 Objectives

The objectives of this section were to analyse set pieces; free-kicks, throw-ins and corners, and how they contributed to Chelsea's success in the 2020/2021 season. Identify possible weaknesses within set pieces. Compare Chelsea's set piece performance (mainly focused on the offense) with the rest of the league, also focusing on the top 6, based on expected goal (xG) and expected threat (xT) values. Additionally, we aimed at answering the question of whether switching from Lampard to Tuchel mid-season had any impact on set piece performance.

3.2 Set Piece Analysis

To investigate Chelsea's set piece performance, Expected Goals and Expected Threats were calculated based on set piece initiated possession chains (both metrics and set-piece chains explained in 6.3). With this, line graphs were generated to observe Chelsea's set-piece xG performance as in figure 4 over the course of the season, displayed in game weeks. Chelsea's management switch started with Lampard leaving after game-week 16. Tuchel took charge shortly after, ahead of game week 17.

For this part we will use these notations when grouping together teams in the Premier League:

- "Top 5" - Top 6 teams in the Premier League with Chelsea excluded i.e. Manchester City, Liverpool, Arsenal, Tottenham and Manchester United
- "ROTL" - All other teams in the Premier League with "Top 5" teams and Chelsea excluded.

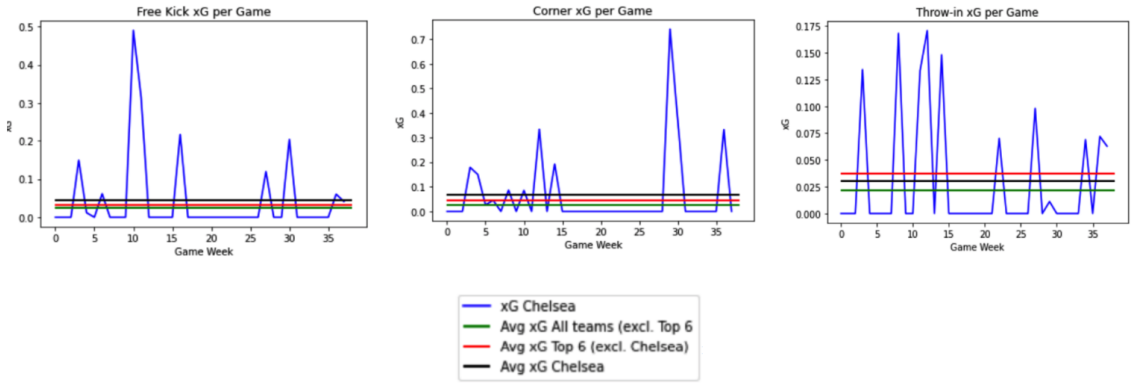

Figure 4: Average set-piece xGs compared to Chelsea set-piece xG per gameweek

From figure 4, it can be observed that Chelsea's free-kicks leading to direct or indirect shots reduced in xG value after the departure of Tuchel. Indicating either that these shots had a lower chance of being converted into a goal or the frequency of shots from freekicks decreased. Additionally, the average xG values indicate that Chelsea outperformed the average for the Top 5 teams as well as the average for ROTL teams. As for corners, shots seem to be taken less frequently but became more dangerous once they occurred. Similarly as with free-kicks, Chelsea outperformed the average for the Top 5 teams and the average for the ROTL teams. Finally, the xG for shots generated from throw-ins nearly halved with the new management. For the throw-in chains, Chelsea does outperform the average for the ROTL teams, however, falls a bit short compared to the average for Top 5 teams.

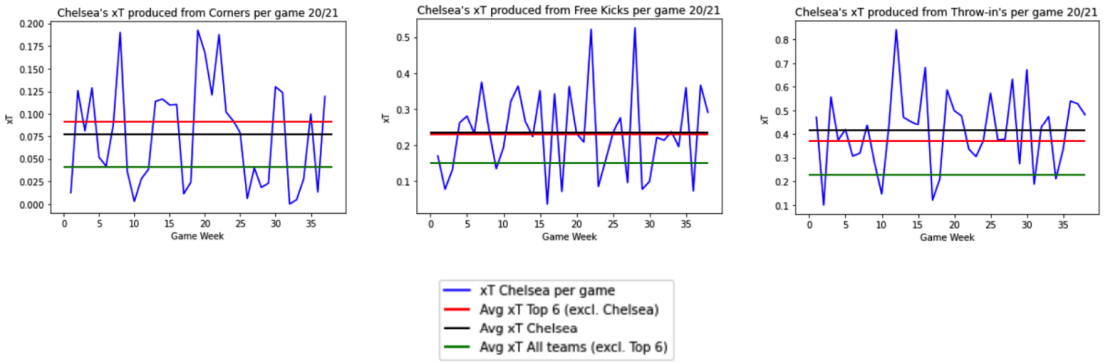

Figure 5: Chelsea set-piece xT per gameweek compared to average set-piece xTs

Figure 5 above, showing the Expected Threat for the generated possession chains starting from each set piece. The graphs give no clear indication that they had been affected by the management change mid-season. According to the graphs, Chelsea performs above average in xT from Free Kicks and xT for Throw in's compared to most teams in the league. They are, however, under-performing in corners. This in turn, could be attributed to direct corners; corners going straight into the box, as these result in either a shot or a clearance.

Lastly, players were ranked based on total xGChain and xT from set-pieces throughout the season. Where xGChain is the xG value that gets assigned to each player involved in the possession chain leading up to a shot. For this we got the top ten best players from the league and then top ten from Chelsea.

	player_id	xG	PlayerName		player_id	xG	PlayerName
263	141746	16.186447	Bruno Fernandes	20	184341	12.732077	Mason Mount
102	60551	14.146523	Ashley Westwood	17	172850	7.186730	Ben Chilwell
314	169187	14.069247	Trent Alexander-Arnold	24	225796	6.434027	Reece James
237	114283	13.269937	Jack Grealish	7	91651	5.446151	Mateo Kovacic
357	184341	12.732077	Mason Mount	16	165153	5.059755	Timo Werner
253	122798	12.665500	Andrew Robertson	5	85955	5.041397	Jorginho
107	61366	12.040358	Kevin De Bruyne	1	41328	4.915851	CÃ©sar Azpilicueta
77	55459	11.263566	Aaron Cresswell	13	124183	4.700083	Hakim Ziyech
315	169359	10.606970	Matt Targett	19	176413	4.472233	Christian Pulisic
306	166989	10.385418	Youri Tielemans	12	116594	3.933868	N'Golo KantÃ©

Figure 6: Top 10 players for set-piece xGChain for league (left) and for Chelsea (right)

	player_id	xT	PlayerName		player_id	xT	PlayerName
263	141746	11.703309	Bruno Fernandes	20	184341	7.846725	Mason Mount
253	122798	11.432426	Andrew Robertson	17	172850	6.076113	Ben Chilwell
314	169187	11.166984	Trent Alexander-Arnold	24	225796	5.463850	Reece James
102	60551	11.089565	Ashley Westwood	16	165153	4.413832	Timo Werner
315	169359	9.191580	Matt Targett	1	41328	4.322621	CÃ©sar Azpilicueta
188	101178	9.059363	James Ward-Prowse	19	176413	3.673523	Christian Pulisic
237	114283	8.491716	Jack Grealish	7	91651	3.672549	Mateo Kovacic
306	166989	8.303798	Youri Tielemans	5	85955	3.489482	Jorginho
357	184341	7.846725	Mason Mount	13	124183	3.248819	Hakim Ziyech
190	101188	7.775756	Lucas Digne	12	116594	3.221781	N'Golo KantÃ©

Figure 7: Top 10 players for set-piece xT for league (left) and for Chelsea (right)

Figure 6, shows Mason Mount as the only Chelsea player to make the top 10 for set piece xGChain with Bruno Fernandes leading the league quite comfortably. From figure 7 it can be observed that Mason Mount is still the only Chelsea player to make it into the top ten players for the whole league. For both metrics we can see Mason Mount, Ben Chilwell and Reece James being the top three players for Chelsea. Since they are the usual set-piece takers this does make sense.

4 Tracking data scouting: Edd Webster, Mattias Holmström and Paula Borst

By analyzing Chelsea´s performance in sections 2 and 3 for the 2020/21 season, it is apparent that Chelsea are lagging behind the top teams in the league with regards to their offense. Chelsea showed deficiencies in their attacking play and several players under-performed with regards to their xG ouput. With the help of tracking data, the cause for those deficiencies has been investigated.
Unlike the previous analysis derived from event-level data, this section features the application of tracking data, a dataset that measures the positions of the players and the ball 25 times per second. This richer, more granular dataset enables the analysis of actions that take place off-the-ball, an example being the effectiveness of passes in the offense to "take out" opponent defenders. Under Tuchel, who is known for his efficient "gegenpressing" and possession-based play [4], Chelsea might lack this effective passing and struggle when trying to advance quickly in the final third and create room in the center of the pitch. This raises the question as to how direct the Chelsea players are actually playing; how fast players accelerate in the offense; and, if no space is found in the center of the pitch, how well are the strikers able to generate chances from less favorable positions? To investigate this, the following three metrics are proposed and are subsequently applied as part of a recruitment analysis exercise.

4.1 Metrics

Directness of possession chain before a shot
This metric is based on pass-value packing, which assigns a value to a pass for every defender of the opposite team that has been bypassed by the pass in question. As opposed to pass-value packing, our metric only involves the passes in possession chains that are leading up to a shot. To incorporate the directness of the possession chain, the generated packing values are adjusted for the number of passes that were required, leading up to the shot i.e. if a possession chain led to a shot consisting of 4 passes and 8 players were passed in this chain, each contributing player is assigned a value of $8/4 = 2$. The value for direct play is calculated per 90 minutes.
Runs at high speed in the final third
The second metric evaluates how often a player is running at a high speed within the final third, while on the offense. In this case, all runs exceeding 7 m/s are considered to be high speed runs. The number of high speed runs are calculated per 90 minutes.
Ratio of difficult shots on target
The third metric calculates the ratio of a player´s shots that have been taken from difficult positions. Here, a difficult shot is considered to be one that is taken from an acute angle relative to the distance to the goal. Difficult shots are defined as those in which the position of the shot exceeds a deviation of 9 meters from the center of the goal. Consequently, shots that are closer to the goal might have a larger angle than long-distance shots, but are considered equally difficult due to their proximity to the goal. The ratio of difficult shots to normal shots is calculated per 90 minutes.

4.2 Chelsea´s player performances

Based on the defined metrics, three of Chelsea´s players that did not show a high level of performance during the season have been identified: Kai Havertz, Timo Werner and Tammy Abraham. In figure 8, their percentile rank compared to the rest of the league has been visualized in the form of radar plots.

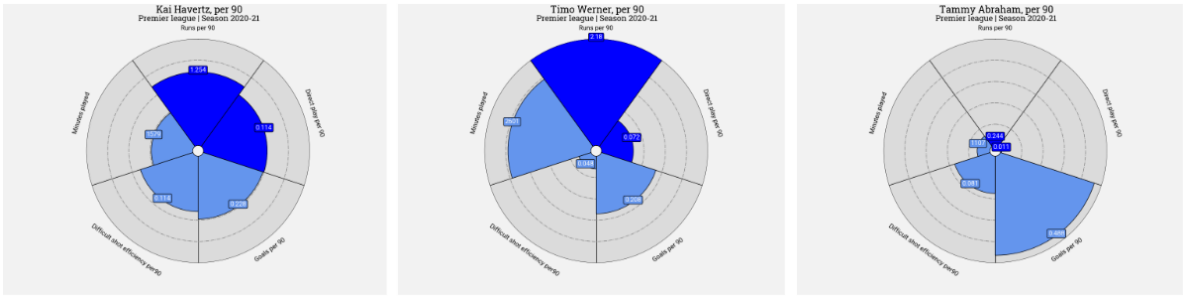

Figure 8: Radar plots for Havertz, Werner and Abraham

As it can be seen from these plots, Kai Havertz displayed mediocre performance for all the selected metrics. Although no clear weaknesses can be identified, he does not stand out in any metric either. As it was his first

season at Chelsea, he might need more time to fit in with the team and may show more convincing performances in the following season.
Timo Werner is the top player in the Premier League when it comes to high speed runs into the final third. Werner has previously been one of the top scorers in the Bundesliga but for Chelsea, he did not even feature in the top 40% of goal-scorers in the league. Therefore, it could be suggested that his high-speed runs in the final third are not resulting in successful actions. When playing for RB Leipzig, Werner's effectiveness in front of goal came when he was able to find space centrally behind the defensive lines of the opponent. However, Chelsea´s possession-based style forces the defensive line of most of their opponents to be pushed back deep into the final third, reducing the possibility for runs in-behind to take place. As Werner also was also not succeeding in scoring from more difficult positions, he was unable to live up the expectations of being the striker that Chelsea needed [5].
Tammy Abraham displayed below-average performances for high speed runs and difficult shots but was among the players at Chelsea with the best goal-scoring record for the number of minutes played. However, the limited number of minutes might skew the values when compared to players that played a substantial amount of minutes across the season.
Based on the analysis of the defined metrics, of the three players, Timo Werner is the one that is suggested for replacement. Based on the number of minutes played during the season, it would be suggested that Werner was considered to be part of the first XI, however, he did not perform at the level that might have been expected from him when he signed.

4.3 Scouting a player

To find a suitable replacement for Werner, the strikers of other teams were investigated for their performances using the identified metrics. The goal this exercise is to find a player with similar strengths to Timo Werner, but with a more solid overall performance. Using the data, three players have been identified:

- Neal Maupay: Brighton & Hove Albion (market value: 20 m. €)
- Marcus Rashford: Manchester United (market value: 85 m. €)
- Gabriel Jesus: Manchester City (market value: 60 m. €)

Their respective radar plots are shown in the following figure 9.

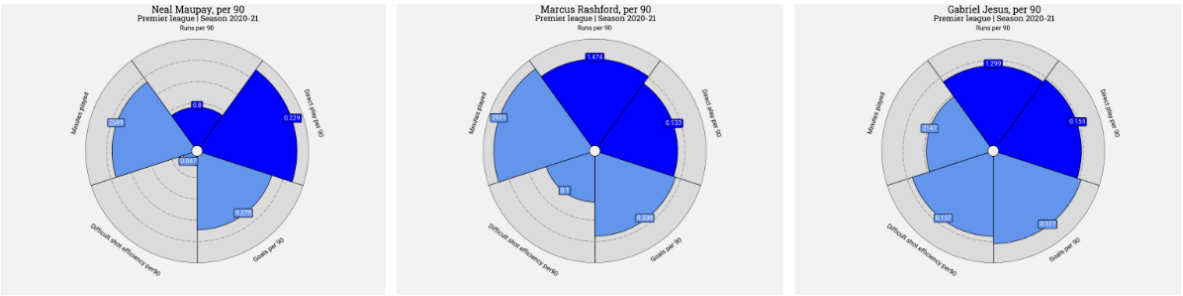

Figure 9: Radar plots for Maupay, Rashford and Jesus

Although Neal Maupay did not perform well with regards to high-speed runs and difficult shots, his output regarding direct play is among the best in the league. This is all the more impressive considering he is playing for one of the lower-half teams of the Premier League table and as such, his performances make him a player for consideration.
Marcus Rashford performs very well in every metric other than difficult shots. As Rashford plays for a strong side in Manchester United, it is also more likely that the areas of the field where he takes his chances are closer to the goal and in the center of the field and therefore, he is less likely to take on difficult shots. His capability of high-speed runs combined with a much better goal-scoring record, would suggest that he would be a good replacement for Werner, from a data perspective.
Gabriel Jesus is close to the best 20% of the league for each of the three metrics and displayed the most balanced set of performances overall. It has been shown in the club performance analysis that the team he plays for, Manchester City, have been superior to all other teams of the Premier League. With the extremely high level for team performance, Jesus' resulting values for the identified metrics could be argued to be more expected than other players and he is likely benefiting from the high-quality chances provided by his teammates.
Considering all the aspects discussed, the suggestion of this analysis is that Chelsea should look to sign Marcus Rashford as a new striker. He shows similar traits and strengths to Werner but furthermore, he often plays as a

left winger, which is why he is not expected to face the same difficulties as Werner in performing well only from the center, but also from the wings.

5 Conclusion

With help of the analysis in the sections club performance, set-pieces and tracking data, it was possible to identify Chelsea´s improvement areas in the season 2020/2021.

Within the club performance analysis, it was first determined whether the analysis should focus on improvements in the offense or the defense. This analysis showed that Chelsea was mainly lacking skill in the offense, and within that, performed worse in open play.

The hypothesis that the lack of offensive skills could be traced back to Chelsea´s open play was tested within the set-pieces section. Here, it was confirmed that Chelsea was performing according the expectations in set-pieces.

With the results of the previous analysis, in the tracking data section, metrics to evaluate different player´s skills in the offensive open play have been developed. This enabled identifying players within Chelsea that were not performing according to what would be required. To overcome these deficiencies, players from other Premier League teams were investigated regarding their suitability to be scouted for Chelsea for the next season. As a result, it was suggested to scout Marcus Rashford to help overcome Chelsea´s weakness in the offense and the open play.

6 Technical Appendix

6.1 Data

The data used in this project has been extracted from the following sources:

1. Football-data.co.uk recorded results data
2. Understat aggregated player and team performance data
3. Transfermarkt player bio and evaluation data
4. Opta Event data (Stats Perform)
5. SkillCorner Broadcast tracking data

6.2 Club performance

6.2.1 Team KPIs

Since Chelsea's relatively weak performance within attacking a more detailed analysis will be conducted. Therefore, five metrics have been deployed and compared with other teams:

1. Expected Goals (xG):xG measures the quality of a created chance in respect to scoring a goal. This metric is summed over all games of every team and over the whole season. To ensure comparability a mean xG per game is calculated for every team. For this model headers, shots, distance and angle have been taken into account
2. Shots: For every team all shots over the whole season are counted and the mean number of shots per game was calculated.
3. xG/shot: This metric calculates the ratio between xG and shots per team. A high ratio implies that shots are taken from more promising positions.
4. goals/xG: This metric calculates the ratio between the scored goals of a team and expected goals . A ratio higher than 1 indicates a better than leagues average ratio.
5. Box cross %: This metric measures the ratio of successful flat-passes to successful crosses into the box. A ratio above 0.5 displays that more successful crosses than passes have been played into the box.

To set Chelsea's performance into perspective the above metrics for Chelsea are compared to the top five teams (excluding Chelsea) and the whole league (excluding top five and Chelsea). The results are presented in figure 10 below and show that Chelsea outperforms the leagues as well as the top five teams average in expected goals generating the third highest value in the Premier league. Nevertheless, Chelsea has not been able to convert the high xg values into goals where they ranked only in the upper midfield of the league (8th). This can also be derived from the below leagues average goals/xG ratio, where the top five teams clearly outperform Chelsea.

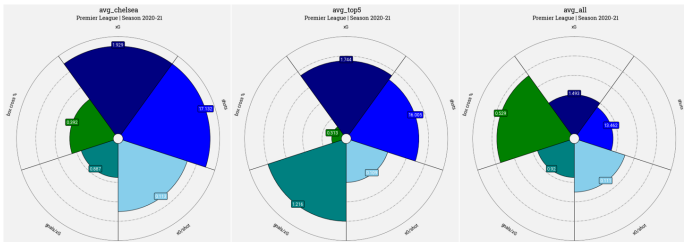

Figure 10: Radar plots offensive metrics for Chelsea, top five teams and all teams

6.2.2 Offense vs defense

Please see below:

1. Regression coefficients for scoring and conceding rates for the top 3 teams
2. Comparing team chance creations for Chelsea under Lampard vs Tuchel

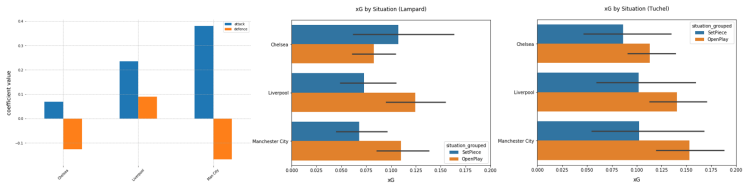

Figure 11: Naive regression model for attack/defence comparison

6.2.3 Simulations

Please see below for the full list of features and their respective coefficients for the baseline model. All features except 'SetPiece_xG' have a significant p-value and low std err. Also see the final league table for the baseline model simulation. Teams are ranked by their final expected points. Chelsea sit comfortably in 3rd spot but have room for improvement in terms of catching-up with Manchester City and Liverpool.

	coef	std err	z	P>|z|
Intercept	-0.4484	0.095	-4.745	0.000
shots	-0.0380	0.010	-3.797	0.000
shots_ot	0.1395	0.017	8.070	0.000
corners	-0.0252	0.012	-2.020	0.043
xG	0.3424	0.046	7.435	0.000
SetPiece_xG	-0.0862	0.147	-0.587	0.557
home	0.1715	0.063	2.703	0.007
big_six	0.2048	0.069	2.960	0.003
vs_city_liverpool	-0.3087	0.130	-2.370	0.018

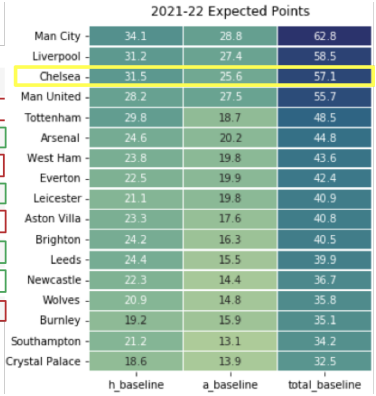

Figure 12: Baseline simulation model

6.2.4 Limitations

1. All variables used are static in terms of their values in the past seasons. Time discounting techniques could be used to make them dynamic where we are considering team forms and updating values on the fly to reward teams on a winning or unbeaten streak and similarly penalise teams on a bad streak. [8]
2. Another data limitation would be the lack of availability of consistent historical data. For example, we could only simulate the league with 17 teams given that 3 teams get relegated and promoted each season. Predicting team metrics for these newly promoted teams is difficult as they were playing in a different league with varied opponent strengths. There is scope to calibrate the predictions in this regard for a fair comparison.
3. The simulation model does not take into consideration relative team strengths. This could be achieved by weighting in club market values or payroll numbers or even using publicly available ranking indexes like UEFA club ranking points.
4. The model does not look into player granularity or effect of player attributes on team performances. As a future scope, we can consider looking at metrics like average age, % of players in peak age by position or metrics to rate players relative to other players.
5. Lastly, as a known limitation of similar models built in the past, performances of best teams are underestimated. Draw, as a match outcome, is also underestimated. There is potential to explore factoring in wisdom of crowds that has the most predictability when it comes to betting markets.

6.3 Set pieces

6.3.1 Data Preparation

To analyse Chelsea's performance within set pieces, first, the data was prepared by singling out possession chains that were initiated by one of the three set pieces mentioned previously. Then, the data was filtered so that all possession chains within the dataset consisted of plays that were no more than 25 seconds long. Additionally, a filter was applied to the possession chains that ended with a shot, making sure that the shots were no longer than 30 metres in distance. The result was a dataframe that consisted solely of relevant event chains which we call "set-piece chains".

6.3.2 Set piece metrics

The metrics used to investigate Chelsea's performance in set pieces are expected goal (xG) and expected threat (xT) as mentioned previously. In this section, the metrics will be explained in more detail:

1. Set-piece Expected Goals (xG): Expected Goals (or xG), is a measure of quality for a shot taken and its probability of resulting in a goal. Using data from the 2020/2021 season and from the opta data set, xG was calculated for each shot taken using the following data:
   (a) Shots
   (b) Distance to the goal and coordinates
   (c) Angle to the goal
   (d) Body part used: which was divided into two categories, namely headers and all other shots, namely non-headers.

   As already mentioned in 3.3, we concluded that Chelsea are performing well in generating xG from set-pieces compared to the rest of the league.

   To further investigate the xG values from set-pieces, heat-maps were plotted to investigate whether there were any particular trends that could be observed to explain why Chelsea performed well. The heat-maps for free-kicks were limited to the penalty area to filter out the short free-kicks in the middle of the pitch.

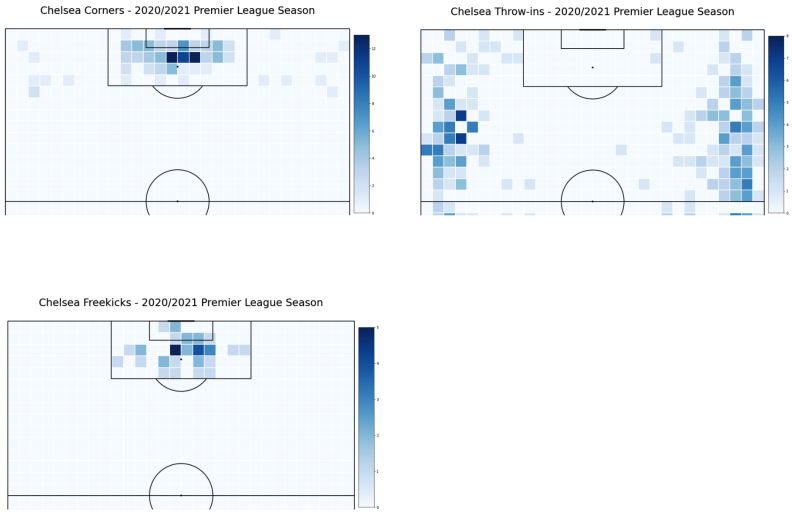

Figure 13: Heatmaps for where the ball ends up after set-piece is taken

The heat-maps could be a good visual explanation as to why Chelsea has above average xG values for corners and free-kicks. It can be observed that corners and free-kicks were placed most frequently into the penalty area, at a direct angle to the goal. This in turn, proves to be a good goal chance since the probability of scoring is the highest in the centre, right in front of goal. For throw-ins, the heat-map shows a much more random situation. One possible trend with throw-ins however, is that the throw-ins were all close to the edges of the pitch, only very few throw-ins were closer to the centre of the pitch and even less straight into the box. Albeit, this could be due to the high arm strength it requires to throw-in

a ball to a long distance.

As mentioned in 3.2, the shots that had a distance over 30 metres, were filtered out. With this we could calculate the Expected Goal metric using a linear regression model, as done in the examples shown in the course [7].

2. Set-piece Expected Threat (xT): Regular Expected Threat (or xT) is a metric that gives value to actions other than shots such as passes, crosses and dribbles. These actions are determined to either increase the chance of scoring (generating a positive value) or decrease the chance of scoring (generating a negative value).

   For calculating Expected Threat for an event the following metrics are needed:

   (a) Starting position of the ball

   (b) Ending position of the ball

   (c) Indicator on whether a shot has been taken or not

   (d) Expected Goals value if a shot has been taken

   With this we followed the example given from the course, for calculating Expected Threat for passes [6] and applied it to our data.

Since we have defined our set-piece chains to contain all events that are relevant to set-pieces only, we now have our two metrics Set-piece Expected Goal and Set-piece Expected Threat added to it.

6.4 Tracking data scouting

The tracking data team sought to identify attacking players who could improve Chelsea offensively and scout players who would fit into Thomas Tuchel's style of play. This was done by the previously mentioned metrics.

Structure of the data

The Skillcorner data is divided into event data and freeze frame data from each match of the 2020-21 season of the Premier League. The event data is organized into a single CSV file for the whole season. From the tutorials provided, JSON files for each match of the season were generated. The tracking allows looking at player positioning at each captured frame of the match. A limitation of the data is that about 70% of frames are reliable. There is also issues with players or the ball not being captured by the camera. Leaving us with a positioning of $(x, y) = (-1, -1)$.

6.4.1 Models and metrics

Measuring direct play

Implementation

The process of valuing possession chains by the directness of play was started by looking through each freeze and identifying how many defending players are behind the ball, so, closer to goal than the ball. The difference to the next frame after is the amount of defenders passed by that pass/event. This value could be used as what is referred to as pass-value packing. This value can be also negative. In order to measure the packing value of each possession chain, the difference between defending players in the first and last frame of each possession chain were calculated. By dividing this number by the length of the possession chain, a measure of direct play was generated. A higher number is therefore regarded as a better possession chain. The score of the metric is given by

$$directness = \frac{x_0 - x_n}{n},$$

where x is the amount of defending players behind the ball at chosen event and n the number of passes required before the shot.

In figure 15, an example of freeze frame can be seen, where the defenders are marked as blue. In this case, there are 7 defenders behind the ball. If after one pass, only 3 defenders would be left, and the next event would be a shot, the directness of this possession chain would have a value of 4 divided by 1. All players that are part of the possession chain are then awarded this value. In order to analyse the season performance for a player, all of these values are summed up and adjusted for the number of minutes played. Possession chains with negative values are removed from the data. With that, players are no longer penalized for being part of those possession chains. Furthermore, including negative values would most likely weight the metric even more towards skilled teams.

Results

The results show that heavily possession based teams are favoured by this metric. It also favours midfielders, which can probably be traced back to their central positioning on the pitch. The fact that players like Kevin De Bruyne are ranked high gives the metric credibility. Notable is also the featuring of Brighton players among top rated players from this metric. Brighton is a team known for their will to play a possession based game despite limited resources under Graham Potter[2]. The fact that not only players from more skilled teams are highly ranked speaks to the fact that a certain style of play has been identified. This would be interesting to investigate further.

Difficulties and limitations

The metric is an attempt to add context to passing in order to see how play is conducted. When adding context or not adding enough context to a metric, there is a risk of information loss. As such, a possession chain where the ball is played backwards in order to create a better shooting opportunity will not be highly valued if enough defenders have tracked back. This shows that the metric could benefit from further context. Furthermore, the metric was proven to be computationally expensive, which led to running the code for several hours.

Runs at high speed in the final third

Implementation

A run is selected if the speed that the player is running is above 7 m/s. Speed is calculated for each player of each frame with

$$speed = \sqrt{vx^2 + vy^2},$$

where vx and vy is the velocity in x and y direction. Runs are selected for offensive players who are in the final third.

Results

The results are promising as they show several players who are known for their pace and movement on the pitch. The top player by this metric is Timo Werner, which shows that he is taking initiative with those runs. The results of top players are shown in figure 14.

Difficulties and limitations

The main issue with this metric is that the team was not able to implement pitch control and it was therefore not possible to gauge the quality of each run. Furthermore, it could not be assessed what other options the player had. This type of context would be interesting to implement in further analysis.

Ratio of difficult shots on target

Implementation

The metric calculates the amount of shots players are able to get on target whilst they are at offset of 9 meters to the right or left of the center of the goal, which is the length of the goalkeepers area.

Results

As it can be seen in the results, the highest ranked players are generally players of weaker teams. This could be due to the fact that players of stronger teams are being put in better positions and are not forced to take those bad shots. This result is noteworthy as it is not something we expected to find and something that would be interesting to analyse further. The metric could also benefit from added context of pitch control. With that, one could assess the positioning of players and award heavily pressured players some added value. Results are shown in figure 14.

Difficulties and limitations

Since the players of weaker teams are over-represented in the results, this metric was not highly weighted in scouting players, as it was not regarded to be as significant when scouting a player for a big club like Chelsea. If one were scouting for a weaker team, this would be a metric to consider more valuable.

#	Player
1	Steven Alzate
2	Dennis Praet
3	De Bruyne
4	Connolly
5	Phil Foden
6	Ross Barkley
7	Bernardo Silva
8	Tyler Roberts
9	Adam Lallana
10	Neal Maupay

#	Player	Runs per 90
1	Timo Werner	2.17
2	Sterling	2.15
3	Firmino	2.06
4	Matej Vydra	2.0
5	Edinson Cavani	1.9
6	Nathan Redmond	1.85
7	Jamie Vardy	1.83
8	Erik Lamela	1.78
9	Salah	1.69
10	Danny Ings	1.68

#	Player	Difficult Shots
1	Ryan Fraser	0.61
2	Dennis Praet	0.49
3	Nathan Tella	0.41
4	Joe Bryan	0.37
5	Ayoze Pérez	0.34
6	Jordan Ayew	0.33
7	Ashley Barnes	0.28
8	Martial	0.25
9	Steven Alzate	0.24
10	Batshuayi	0.23

Figure 14: Top rated players for direct play (left) high-speed runs (middle) and difficult shots (right).

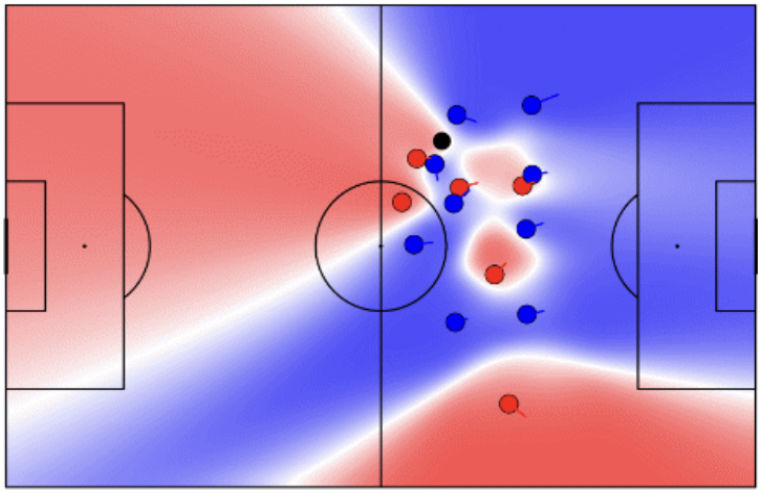

Figure 15: Snapshot of a highly valued possession chain that ends in a shot by Marcus Rashford.

References

[1] Jay Boice. How our club soccer predictions work, fivethirtyeight. 2020.

[2] Owen Bullman. Why is graham potter a future 'top six' manager? – tactical analysis. 2022.

[3] Michael Caley. Premier league projections and new expected goals. *cartilagefreecaptain.sbnation.com*, october 2015.

[4] Nizaar Kinsella. Tuchel and klopp speak the same footballing language - so why are their tactics so different? *Goal.com*, March 2021.

[5] Aniket Rai. Opinion: Why has timo werner not been able to replicate his bundesliga form in england? March 2021.

[6] soccermatics. Calculating xt (action-based).

[7] soccermatics. Fitting the xg model.

[8] Ben Torvaney. What can time-discount rates tell us about xg and goals? *statsandsnakeoil.com*, june 2021.

[9] whoscored.com. Premier league team statistics.