Analyzing Soccer Data: Can Statistics Accurately Predict Future Results?

By:

Muhammed Hanan

Submitted to the Department of Mathematics and Computer Science

School of Natural and Social Sciences

in partial fulfillment of the requirements

for the degree of Bachelor of Arts

Purchase College

State University of New York

October 2022

Sponsor: Athar Abdul-Quader

Second Reader: Lee Tusman

Abstract

This paper looks to see if specific soccer statistics can accurately predict the results of future matches. This research shows what soccer statistics or prediction methods are the most reliable in predicting future soccer matches. The methods used in this research involve using three different models, model one uses home and away statistics, model two uses expected goals, and model three uses machine learning. The research concluded that machine learning is the most accurate in predicting the results of future soccer matches. Along with this, the paper touches on the history and future of soccer analytics, the collection and use of soccer data, and the limitations of statistics.

Background Information

For those unfamiliar with the format of the English Premier League. The English Premier League (EPL) is considered to be the biggest and most popular soccer league in the world. There are 20 teams in the EPL and each team plays 38 matches. Each team plays the 19 other teams twice, both home and away. There are a total of 380 matches played during the season which takes place between August and May.

A team's ranking is determined by how many points they have and at the end of the season, the team with the most points wins the league. Ranks one through four qualify for the Champions League, rank five qualify for the Europa League and the bottom three teams are relegated to the second division of English Soccer. In contrast, the top three teams of the second division are prompted. ("Premier League Competition Format & History | Premier League")

The UEFA Champions League (UCL) is the top-tier European cup competition and English teams can only qualify by placing in the top four of the league. The UEFA Europa League (UEL) is the second-tier cup European competition and an English team can qualify by getting 5th place or by winning the FA Cup. The UEFA Conference League is the 3rd tier European Cup competition and an English team can qualify by winning the EFL Cup. A team can only qualify and play in only one of these three European competitions. (Crebolder)

1. Introduction

Statistics and data are a big part of modern professional soccer and without them, soccer would not be what it is today. Data is collected from the first second of a match to the last. In fact, as Monika Linder describes in her project report “A soccer game really is a field day for mathematicians.” (Linder, 4)

The collected data is used to assess a team’s performance and identify areas in need of improvement. This translates into the recruitment of new players for problem areas. Data provides insight for coaches/managers. The data also influences a team’s tactics for a game. Statistics tell us if a player is performing well or not. Data helps analyze the competition and helps fans place bets.

Even with the massive influence of data and statistics, some argue that the stats do not tell the whole story. Some players are even dubbed to be not blessed by the stats. A classic example of this is now retired soccer legend Zinedine Zidane, who is quite commonly referred to as one of the greatest soccer players of all time. Still, his stats tell a completely different story and are very underwhelming. (“Zinédine Zidane - Career Stats”)

With these two ideas in mind, I’m building three different models to predict the results of the 22/23 English Premier League season. Each model will have different soccer stats as variables and parameters. Each model’s accuracy will be measured to determine the reliability of the model and the variables used for that model. These models will be able to answer if statistics can accurately predict the result of a match.

Model one uses home and away stats to predict the result of a match. Model two uses the expected goal statistic to predict the result of a match. Model three involves machine learning to predict the result of a match.

2. Literature Review

2.1 History of Soccer Analytics

Charles Reep is widely considered to be the first soccer analyst. Charles Reep was a Royal Airforce pilot and was a massive supporter of Arsenal Football Club in the 1930s due to their play style, which was revolutionary at the time. When returning to England after service he found that Arsenal's revolutionary play style had not been embraced to its full extent across England.

During a professional match he attended, he grew tired of the teams' slow play style and build-up play, pointless passes, and poor scoring attempts. Due to this he grabbed a pen and paper and started recording the team's actions. Reep would continue taking notes and watched and recorded notes on around 40 matches per season. After analyzing the data he had recorded, he found that only two goals out of nine came from moves that included more than three received passes.

Reep would go on to work with a few different soccer clubs and their managers. With this, he developed a theory on how soccer should be played which is direct through long balls forward also known as "route one football". He came to the conclusion that passing too much was waste of time and risky and that most goals are scored through short moves. From then onwards the basics of English soccer were based on Reep's theory. (Pollard) However, Reep's statistical analysis was flawed due to its exclusion of crucial contextual factors such as the caliber of the opposing team, player positioning on the field, and the timing of passes. As a result, Reep made invalid assumptions about certain soccer strategies.

For instance, Reep believed that teams should strive for as many long passes as possible; however, this frequently led to lost possession and susceptibility to counter-attacks. He also advised teams to shoot from outside the penalty area, which yielded wasted chances and low conversion rates. Ultimately, Reep's notions disregarded the value of individual talents like dribbling, an effective method for disrupting defenses and generating scoring opportunities. (Sykes)

In the 90s, England was very disappointing on the national level. This caused a change in the game and the English Premier League was formed and so was the Champions League. Due to this different cultures started to mix in the game, and new ideas of how the game should be played were formed. This caused an evolution in ideas of how the game should be played and an evolution in technicalities and analytics.

A major statistical breakthrough came in 1996 when Opta began collecting Premier League game data. Premier League sponsor Carling paid for the Opta Index, but clubs and media received the data for free. Each club received an Excel report with some basic stats. The clubs learned facts they had never considered before. For example, the distance each player ran and how many tackles and passes they made per game. This data was revolutionary at the time but primitive by today’s standards.

With access to a lot of new information in the form of data, a lot of teams and their managers were prone to misinterpreting it often leading to errors. Despite this former and legendary Arsenal Football Club manager Arsene Wengner was very faithful in match data. With Wengner’s understanding of the numbers, he caused a lot of changes within Arsenal. So much so, it seemed Wegner and Arsenal were way ahead of their time and that caused a period of great success. With so much success other teams started studying Wengner's use of data, tactics, and reforms and started implementing changes within their own teams.

Over the past decade, data analytics has evolved rapidly in the game and has become an integral part of modern football club operations. Clubs obtained an immense amount of data, which led to new positions and sometimes departments. Complicated metrics have emerged and they give us more thorough insights into the teams' and players' performances. The use of these metrics has become essential not only in scouting and recruiting but also in opponent analysis. To this day, Liverpool Football Club is considered the most innovative club in modern football in this regard. (XFB ANALYTICS)

2.2 How is Soccer Data Collected?

There are many different ways data is collected in the sport of soccer. Some private companies like Opta have professional analysts who monitor, record and annotate every touch of the ball. Other methods include automatic video analysis and fully automated systems. ("Data Collection and Statistics", Rob et al.)

ChyronHego's TRACAB Gen5 incorporates distributed camera architecture that allows you to capture the action from four angles, with cameras positioned on either side of the pitch and behind each goal. The flexibility of the system to provide multiple camera views allows the system to track anything in the field with a much higher resolution. In a typical soccer match, with 22 players and the position of the ball, the system can record 25 times in one second, giving about 135,000 records of data. ("ChyronHego Introduces TRACAB Gen5")

Ball and player position data is automatically tracked by cameras or sensors. The position is usually measured by an x,y, and sometimes z coordinate system related to the pitch. It uses a distributed system of cameras installed around the pitch and advanced image processing technology to acquire and provide real-time tracking data on each player, referee, and ball movement. ("ChyronHego TRACAB Gen5 Updates Tracking Algorithms, Adds AI Recognition")

The official 2022 World Cup match ball has revolutionary data-collecting technology. Inside the ball is the Adidas suspension system which includes a 500-hertz measurement unit motion sensor that sends out information 500 times per second. The ball is powered by a rechargeable battery and the new technology does not interfere with ball performance and is unnoticeable to players. (Reiser)

Offside is a rule in soccer in which any part of an attacker's body which can be scored with is closer to the goal than the last defender or ball and if this is the case the player is called offside. Since offside can be very difficult to tell in some cases, FIFA has implemented semiautomatic offside technology and the new ball helps to improve the speed and accuracy of the semiautomatic offside technology. The technology also uses 12 dedicated cameras mounted under the roof of the stadium to track the ball and track up to 29 data points from each player 50 times per second. Using artificial intelligence, player, and ball tracking data automatically provides video match officials with offside alerts if an attacker is offside. (Mather)

Another data collection method for players is wearable technologies. Most professionals can be seen wearing black vests during training, these vests are used for data collection. An example of these vests is the Catapult Vector which features a heart rate sensor, accelerometer, gyroscope, magnetometer, and antennas. The vests are able to collect data such as position on the field relative to other players, heart rate, distance covered, speed, impacts, and more. The Catapult Vector comes with software for data analysis, a sensor, a Bluetooth app, and a charging case.

There are other consumer-level technologies that make data collection accessible for non-professional players. There are quite a few different sensors that involve putting a sensor in your sock or around the calf area, these are known as calf sensors, such as the Zepp Play Soccer and FootBar Meteor. Calf Sensors have Bluetooth technology and offer a mobile app as well. The devices are able to collect and present physical data and data on how a player's performance was on the ball.

Another interesting type of product is foot sensors. With foot sensors, the sensor goes directly into a player's cleats usually in the form of insoles. Examples of foot sensors are Xampion and Playermaker. Foot sensors also feature Bluetooth mobile apps and are able to collect and present movement data and data on how a player strikes or moves with the ball. (Chua)

2.3 Soccer Data Explained and Applied

There are a lot of soccer statistics and data that cover a wide variety of areas in the game. But, the data can be broken down into three types of data: event, tracking, and physical data. Event data is essentially just what happens on or with the ball such as passes, shots, tackles, etc. Tracking data is the positions of each player on and off the ball on the pitch. Physical data measure different physical aspects of a player such as height, weight, speed, etc.

The data that is collected in soccer influences a number of different aspects of soccer clubs and their players. Data influences scouting and player recruitment. By working with the data, clubs are able to find players in surprising places who will improve their teams and are available at an inexpensive cost.

Data also plays a role in almost all aspects of player fitness and performance. A player's diet, sleep, training regime, and recovery are all tweaked based on data. Data also helps with injury recovery and injury prevention. Data influences the tactics of a team and how a team prepares for a specific match based on the data they have from the opposition's team. (Carling, Mashinchi)

3. xG

3.1 Importance of xG

Expected Goals (xG), a statistic, is gaining popularity among soccer analysts and supporters as a reliable indicator of a team's or player's offensive prowess. By evaluating the likelihood that a shot will result in a goal based on a number of variables, it is a tool that assesses the average of scoring opportunities in soccer games. These elements include the shot's placement, angle, style of play that came before it and other pertinent elements. Instead of just looking at the total number of goals or shots, xG provides a more complex perspective of the game by taking these elements into account.

Utilizing xG in soccer analysis offers multiple advantages. It provides a clearer vision of a team's or player's offensive proficiency as opposed to counting goals or shots only. Additionally, it allows one to recognize players who are performing well but lack conversions due to below-par luck, as well as others who score from substandard opportunities.

By taking into account the grade of scoring chances, xG gives an elaborate evaluation of teams' and individuals' attacking performance. For this reason, it has become beloved among experts and supporters alike, distinguishing the probability of success for a particular kick depending on distinct factors. (Kelly)

3.2 Calculating xG

xG is not calculated by hand but through models. Each xG model tends to have their own specific characteristics but they do have some basic things in common. For instance, some variables xG models consider are distance to the goal, angle to the goal, quality of chance, body part, assist, and play pattern. Free kicks, penalties, and dead ball situations are considered constants. Considering all these variables, xG models use historical information from thousands of shots with similar characteristics to estimate the likelihood of a goal on a scale between 0 and 1. To clarify, there is no one specific formula for calculating xG, each model has its own unique characteristics, but all models do share some basic variables that help with the calculations. (Kelly)

3.3 Other xG-Related Metrics

Expected goals (xG) in soccer are not the only statistics that can supply valuable insight into team and player performance. A more nuanced analysis can be conducted by using additional expected stats. For instance, expected goals per 90 minutes (xGp/90) measures a player's average number of anticipated goals within each game interval. Like xG, it takes a number of factors into account, including the position of the shot, the angle, and the play that preceded it. Normalizing xG with respect to playing time through xGp/90 provides a superior method of evaluating attacking performance as compared to just examining total xG because it eliminates bias introduced by varying amounts of playing time for different players. (TedTalksFPL)

Expected assists (xA) gauge the probability that a pass will result in a goal. Like xG, xA considers a number of variables, such as the pass's position, the nature of the play that preceded it, and the anticipated caliber of the finish. Despite the fact that not all of their teammates' scoring chances result in goals, players with high xA values help their teammates score high-caliber goals. (Whitmore)

Expected goals against (xGA) is a statistic that gauges a team's likelihood of giving up a goal based on the caliber of the opposition's scoring possibilities. The shot's position, angle, kind of play preceding it, and defense quality are just a few of the variables that xGA takes into account. Teams with a low xGA value are better at restricting high-quality scoring opportunities for the opposition. ("Revealed: The Premier League's XG and XGA over and under Performers")

The post-shot expected goals (PSxG) metric calculates the likelihood that a shot will score after it has been fired. It considers the ball's trajectory, the shot's velocity, and other important variables. As it provides a more accurate assessment of goalkeepers' shot-stopping skills than simply looking at the number of goals they have given up, PSxG is a valuable metric for assessing goalkeeper performance. (Cook)

Overall, soccer matches and player performance can be studied using expected statistics. Expected statistics, which take into consideration the quality of scoring opportunities, offer a more accurate and nuanced assessment of the game than merely focusing on the number of goals scored or shots taken. More expected statistics will likely be created as soccer analytics develops to give an even deeper insight into the game.

3.4 Predictive Power of xG

The success of xG as a soccer predictor has generated a lot of discussion among soccer pundits and fans. As indicated, while xG is beneficial for gauging a team's or player's future performances offensively, it cannot guarantee an outcome. Soccer entails a multitude of variables that can influence the results of upcoming fixtures and seasons, making accurate predictions difficult. Several factors ranging from injuries to psychological mindsets to tactical changes can have unforeseeable impacts on success rates going forward. In light of all this information, relying solely on xG is not feasible, and other insights need to be considered before making any concrete assertions about the prospective outcomes.

Additionally, while xG takes into account the caliber of scoring opportunities, it ignores elements like a team's defense or a goalkeeper's ability to make saves. These elements, which xG might not completely account for, can also have a substantial impact on how a game or a season turns out.

Despite these drawbacks, xG is still a valuable tool for predicting soccer results. When used in conjunction with other statistical and analytical techniques, xG can give useful insights into a team's or player's performance and assist pinpoint areas for growth. Additionally, xG can be used to evaluate the efficacy of different tactics or plans as well as measure the development of a team or player through time.

xG is a good tool for examining performance and offering insights into the game even though it is not a perfect prediction of future soccer results. xG will probably continue to be a significant indicator for assessing the game and predicting future outcomes as soccer analytics develops and new statistical techniques are created. (beatthebookie2017)

4. Methods and Results

4.1 Failed Machine Learning Model

My attempt at making a machine-learning model with Python was a failure but still, I was able to learn a few things from this failure. I started working on this model by first scraping English Premier League data from the past two seasons. The data was scraped from a website called FBREF, which has essentially all the soccer statistics one could need. The data was scraped using BeautifulSoup, Pandas, and Requests.

BeautifulSoup is a Python library that is used for getting data from web pages. The Pandas library is used for cleaning up, sorting, grouping, and just overall making data easier to work with. The Requests library makes it possible to communicate with web servers. So I scraped through the HTML of the web pages using the Requests library. Then, I used Beautiful Soup to parse through the web pages. After, I got game statistics using Requests and Pandas.

Then, I cleaned up the data with pandas. Afterward, I used a loop to repeat the same process for two seasons. All the scraped data was exported as a CSV file. The CSV file was inputted into the Pandas data frame and then cleaned up by removing unnecessary data for machine learning. Next, I created predictors for machine learning which in my case was the venue, opponent, time, date, and result. (Paruchuri)

For machine learning, I used a library called Scikit-learn and from the library, I used the Random Forest Classifier model which in this application was to account for the non-linearities in the data. Then from there, the data was split into a training set and a testing set. Initially, the model had an accuracy of 32% but after retraining and improving precision the model got an accuracy score of 37%. For the purposes of this project, this was not a high enough accuracy score. (Dataquest)

Looking back at the model now, I can see now where I could have made a few changes to the model that might have improved the accuracy. For example, I could have changed up the predictors. Also, I could have scraped a larger data set and trained the model on a larger data set.

4.2 Home/Away and xG Model

The home and away model was made using Microsoft Excel and makes predictions on data from the first sixty matches of the 22/23 season. The model uses poisson distribution and poisson distribution gives us the probability of an event happening a certain number of times within a given interval of time or space. In this case, the event is scoring a goal and how many times it happens. The model uses a team's goals for and goals against stats for both home and away and compares it against the leagues as a whole. From there I am able to get the attack and defense matrix, which is essentially the strength of a team's attack and defense. By using the Poisson distribution, I'm able to get the predicted result of a match and also get the probability of different amounts of goals being scored or conceded. ("Build an Amazing Soccer Prediction App in Excel: 8 Step")

The xG model follows the same principles as the home/away model, however, I replaced goals for and goals against with expected goals and expected goals against and the model used 190 matches, or about half the season as data. The data from the xG model is from understat.com. I divided the xG and xGA for each team, both home and away, by the number of matches each team had played.

The people of understat.com have trained their own neural network prediction algorithms with a dataset of over 100,000 shots and ten parameters for each. ("EPL XG Table and Scorers for the 2019/2020 Season | Understat.com")

Both models were tested on fifty matches and they both had an accuracy of 40%. Below, is the Excel formula structure I used in the model to return the possibilities of the outcome of their being x amount of goals, mean is the predicted result, and cumulative is false as shown in the picture below. The table below shows that xG is able to predict the scoreline more accurately, therefore, providing more context also that the models struggle when there is an unexpected result because in both cases city was the clear favorite.

POISSON.DIST(x, mean, cumulative)

	Home Team	Away Team
PICK TEAMS--->>	Tottenham	Chelsea
Attack Matrix	2	1
Defend Matrix	1	1
Avg	2	1
Predicted Result	2	1

Probablity of Scoring "X" # of Goals		
# of Goals	Home	Away
0	18%	46%
1	31%	36%
2	26%	14%
3	15%	4%
4	6%	1%
5	2%	0%
6	1%	0%
7	0%	0%
8	0%	0%

Model and Match	City vs Forest	Forest vs City
Home/Away	4-1	1-3
xG	6-1	1-4
Actual Result	6-1	1-1

4.3 Machine-Learning Model - ProphitBet

ProphitBet is a machine-learning model on GitHub that can predict the results of soccer matches. ProphitBet gives users the ability to train their own machine-learning model based on their choice of data and machine-learning algorithms such as Neural Networks, Random Forests, and Ensemble models. It also provides the user, with different parameters to tune. The model has a lot of different soccer leagues from around the world. ProphitBet actively scrapes the latest soccer data from a website called footy stats which has a lot of different soccer data. The model also provides different analysis methods such as feature correlation and an ability to evaluate models. (Kochliaridis)

I trained a model using the RandomForest algorithm in the program and I got about a 60% accuracy when testing 50 matches. The model was trained on over 3000 matches and the data it was trained on for each match is below.

1	X	2	HG	AG	Result	HW	HL	HGF	HGA	HGDW	HGDL	HW%	HD%	AW	AL	AGF	AGA	AGDW	AGDL	AW%
2.1	3.6	3.4	4.0	3.0	H	1	1	4	4	1	1	35	35	0	2	4	7	0	1	35

4.4 Observations

Even though there is a significant difference in the accuracy of the Excel models and the machine-learning model, I still expected the machine-learning model to have an accuracy of around 70%. This could be due to how I trained the model. I'm assuming if I trained the model on more data and given it more time to train the model, the accuracy would have been higher. The machine-learning model was not able to predict draws very well. Overall though, the models were right when a match had a clear favorite, however, the models were not able to predict unexpected results. In terms, of predicting the results of a match the xG and home/away models had the same level of accuracy but the xG model was able to predict the scoreline more accurately. To conclude, the research reveals, that xG might not be a better predictor than goals for and goals against but it provides more context and that for the most accurate results machine learning is the most promising.

5. Limitations of Statistics

Through looking at the methods talked about in this paper and the history of Charles Reep, one can conclude that statistics do indeed have limitations. From looking at the history of Charles Reep, one can conclude that the context of statistics is very important and that statistics can be prone to misunderstanding. Reep's statistics did not take into account the context of a specific play. For example, Reep counted all forward passes equally and did not take into account whether a pass was made under pressure or not, since if a pass was made under pressure it is more likely to fail. Furthermore, Reep's statistics could not account for the complexities of soccer and therefore lead to an oversimplified interpretation of the game. (Sykes)

Looking at the research done with the models, one can conclude that statistics cannot necessarily predict the unexpected without context. For example, the models cannot account for individual player errors or red cards which lead to a player's dismissal in a game where one team has to play with one less player when making predictions.

Going back to Zinedine Zidane, his numbers show why statistics have limitations. As said previously, Zidane is widely considered to be one of the best of all time, however, his overt statistics do not back that. For example, Zidane has 125 career goals, and Frank Lampard played in a similar position to Zidane but, has 211 goals, however, Lampard is not considered to be one of the best. (“Frank Lampard - Career Stats”) Zidane's influence on the game cannot be justly quantified by overt statistics. What set him apart was his ability to change the game's tempo, his amazing touch and dribbling skills, his brilliant passing, scoring late and significant goals, and his presence on the field.

Zidane's leadership abilities and capacity to improve his teammates' performances also served as defining characteristics of his game. His unique talent and qualities cannot be quantified by overt statistics but led him to numerous titles and achievements including the World Cup and the Balon d’Or (best player of the year award). (“Zinedine Zidane”)

6. Future

6.1 Future Research

I will continue my research on this project and study how soccer analytics continues to grow and evolve with the appearance of new technologies, as well, as become more familiar with Python and machine learning to build a model I would consider ideal.

I want to build a model that is easily accessible online. The model would be a machine learning model that uses the random forest algorithm because machine learning shows the most promise and the random forest is considered to be the most accurate algorithm for sports predictions. The model would be trained on as much data as possible and would be retrained yearly on newer data. The model would be able to predict matches from different leagues from around Europe and the world, rather than just the English Premier League. Lastly, I would look for a way to have xG as a parameter in this ideal model.

6.2 Future of Soccer Analytics

Soccer analysis and sports analysis, in general, are rapidly growing and changing. I'm keen to see what role AI and machine learning will play in the future of soccer analytics as well as what new technologies will be developed to help with analysis and data collection. As for the immediate future of soccer analysis, there is a heavy emphasis on wearable devices to collect player data, GPS tracking, and smart broadcasting to provide real-time and in-depth player and match data to fans. (GoldCleats Scout)

7. Conclusion

Ultimately, it can be said that although statistics can be highly useful in determining a team's performance and potential, it's important to remember that soccer is a complex game with many other aspects at play. Even though relying solely on soccer statistics for forecasting the future may seem ideal to start, one should not completely depend on them. The results of a game can be influenced by coaches, players, and additional elements including injuries, team morale, and individual playing styles. The results from the models show statistics by themselves are not perfect predictors but statistics combined with other methods can lead to some positive implications such as providing valuable insights into a team's or player's performance, however, on the other hand, some negative implications can be potential biases, such as overreliance on one statistic that leads to a lack of context. Furthermore, I believe training a machine-learning model on different xG statistics can prevent certain negative implications. In the end, soccer's beauty rests in its unpredictable nature, and while statistics might offer useful information, they cannot by themselves reliably forecast future outcomes.

Works Cited

beatthebookie2017. “The Predictive Power of XG.” *Beat the Bookie*, 6 June 2021, beatthebookie.blog/2021/06/07/the-predictive-power-of-xg/. Accessed 9 May 2023.

“Build an Amazing Soccer Prediction App in Excel : 8 Step.” *Tinytutes*, 2019, tinytutes.com/tutorials/soccer-prediction-excel/.

Carling, Chris. “FOOTBALL DATA & TECHNOLOGY for MATCH ANALYSIS.” *Medium*, 8 Aug. 2019, medium.com/@ccarling_82663/football-data-technology-for-match-analysis-a19db1a788a4.

Chua, Julian. “Wearable Soccer (Football) Sensors That Track Shots, Passes and More.” *Sports Technology Blog*, 1 June 2022, sportstechnologyblog.com/2022/06/01/wearable-soccer-football-sensors-that-track-shots-passes-and-more/.

“ChyronHego Introduces TRACAB Gen5.” *TVTechnology*, 15 May 2019, www.tvtechnology.com/the-wire-blog/chyronhego-introduces-tracab-gen5. Accessed 9 May 2023.

“ChyronHego TRACAB Gen5 Updates Tracking Algorithms, Adds AI Recognition.” *Digital Media World*, www.digitalmediaworld.tv/broadcast/2391-chyronhego-tracab-gen5-updates-tracking-algorithms-adds-ai-recognition.

Cook, Sollie. “Post-Shot Expected Goals (PSxG): What Is It and Why Does It Matter? — the Sporting Blog.” *Thesporting.blog*, thesporting.blog/blog/post-shot-expected-goals-what-is-it-and-why-does-it-matter. Accessed 9 May 2023.

Crebolder, Finley. “The Premier League’s European Qualification Rules.” *90min.com*, 26 Feb. 2023, www.90min.com/posts/the-premier-league-european-qualification-rules. Accessed 9 May 2023.

“Data Collection and Statistics.” *Coaching American Soccer*, coachingamericansoccer.com/administration/data-collection-and-statistics/. Accessed 9 May 2023.

Dataquest. “Predict Football Match Winners with Machine Learning and Python.” *Www.youtube.com*, 9 May 2022, www.youtube.com/watch?v=0irmDBWLrco. Accessed 9 May 2023.

“EPL XG Table and Scorers for the 2019/2020 Season | Understat.com.” *Understat.com*, understat, 2019, understat.com/league/EPL.

“Frank Lampard - Career Stats.” *Transfermarkt.us*, 2017, www.transfermarkt.us/frank-lampard/leistungsdaten/spieler/3163. Accessed 11 May 2023.

GoldCleats Scout. “Revolutionizing Youth Soccer Analytics: IoT and Biometric Data Unveil Unique Player Insights.” *Goldcleats.substack.com*, 13 Apr. 2023, goldcleats.substack.com/p/revolutionizing-youth-soccer-analytics. Accessed 9 May 2023.

Kelly, Ryan. “What Is XG in Football & How Is the Statistic Calculated? | Goal.com US.” *Www.goal.com*, 27 Oct. 2019, www.goal.com/en-us/news/what-is-xg-football-how-statistic-calculated/h42z0iiv8mdg1ub10iisg1dju. Accessed 9 May 2023.

Kochliaridis, Vasileios. “ProphitBet - Soccer Bets Predictor.” *GitHub*, 6 May 2023, github.com/kochlisGit/ProphitBet-Soccer-Bets-Predictor. Accessed 9 May 2023.

Linder, Monika. *Football and Mathematics a Search for a Relation between Passing Structure and Successful Results*. June 2017.

Mashinchi, Navid. “The Current State of Major League Soccer Analytics.” *Medium*, 26 Jan. 2021, towardsdatascience.com/major-league-soccer-analytics-9fc76891c440. Accessed 9 May 2023.

Mather, George. “A Step to VAR: The Vision Science of Offside Calls by Video Assistant Referees.” *Perception*, 13 Nov. 2020, p. 030100662097200, https://doi.org/10.1177/0301006620972006.

Paruchuri, Vik. “Project-Walkthroughs/Football_matches at Master · Dataquestio/Project-Walkthroughs.” *GitHub*, 1 Sept. 2022, github.com/dataquestio/project-walkthroughs/tree/master/football_matches. Accessed 9 May 2023.

Pollard, Richard. “Charles Reep (1904-2002): Pioneer of Notational and Performance Analysis in Football.” *Journal of Sports Sciences*, vol. 20, 2002, www.semanticscholar.org/paper/Charles-Reep-(1904-2002)%3A-pioneer-of-notational-and-Pollard/5f7df9915a1e5649c1394c0ff4e0c985ec57ab2d. Accessed 9 May 2023.

“Premier League Competition Format & History | Premier League.” *Www.premierleague.com*, www.premierleague.com/premier-league-explained#:~:text=The%20league%20takes%20place%20between.

“Premier League Fixtures 2022/23 (England) | FootyStats.” *Footystats.org*, footystats.org/england/premier-league/fixtures. Accessed 11 May 2023.

"Premier League Stats." *FBref.com*, fbref.com/en/comps/9/Premier-League-Stats.

Reiser, Anya. "Technology Innovations at the 2022 FIFA World Cup Qatar." *TOMORROW'S WORLD TODAY®*, 21 Nov. 2022, www.tomorrowsworldtoday.com/2022/11/21/technology-innovations-at-the-2022-fifa-world-cup-qatar. Accessed 9 May 2023.

"Revealed: The Premier League's XG and XGA over and under Performers." *Football News, Views & Transfer Rumours | Football Whispers*, 18 Jan. 2021, footballwhispers.com/blog/premier-league-over-under-performer-xg-xga/.

Rob, et al. *Automated Soccer Data Collection from Videos*.

Sykes, Joe. "How One Man's Bad Math Helped Ruin Decades of English Soccer." *FiveThirtyEight*, 27 Oct. 2016, fivethirtyeight.com/features/how-one-mans-bad-math-helped-ruin-decades-of-english-soccer/.

TedTalksFPL. "Which Minutes See the Most Goals and XG in Premier League Games?" *Best FPL Tips, Advice, Team News, Picks, and Statistics from Fantasy Football Scout*, 12 Aug. 2022, www.fantasyfootballscout.co.uk/2022/08/12/which-minutes-see-the-most-goals-and-xg-in-premier-league-games/#:~:text=Whenever%20you%20use%20an%20.

Whitmore, Jonny. "What Are Expected Assists (XA)?" *The Analyst*, 24 Mar. 2021, theanalyst.com/na/2021/03/what-are-expected-assists-xa/.

XFB ANALYTICS. "The History of Football Analytics - Part 2." *Xfb Analytics*, 28 Apr. 2022, www.xfbanalytics.hu/blog/blog-post/32.

"Zinedine Zidane." *Wikipedia*, 1 Apr. 2020, en.wikipedia.org/wiki/Zinedine_Zidane.

"Zinédine Zidane - Career Stats." *Www.transfermarkt.us*, www.transfermarkt.us/zinedine-zidane/leistungsdaten/spieler/3111. Accessed 9 May 2023.