A machine learning approach for predicting outcomes of Premier League soccer matches based on a team's chance creation and quality of finishing.

LANDEN FOGLE, University of Nebraska-Lincoln, USA
MAX SIEVENPIPER, University of Nebraska-Lincoln, USA
TAGE ZERBY, University of Nebraska-Lincoln, USA

This research employs a machine learning approach to predict outcomes of Premier League soccer matches, focusing on the significance of a team's ability to create scoring opportunities and their efficiency in finishing these chances. By analyzing data on chance creation and the quality of finishing, the study seeks to answer the question of which aspect is more crucial for a team's success in the highly competitive environment of professional soccer. Utilizing metrics such as Expected Goals (xG) and actual goals scored, the paper investigates the relationship between the creation of high-quality scoring opportunities and their conversion into goals, offering insights into effective strategies for enhancing team performance.

Key Words: Expected Goals, Soccer Analytics, Chance Creation, Machine Learning, Premier League

ACM Reference Format:
Landen Fogle, Max Sievenpiper, and Tage Zerby. 2024. A machine learning approach for predicting outcomes of Premier League soccer matches based on a team's chance creation and quality of finishing. . 1, 1 (May 2024), 9 pages.

1 INTRODUCTION

The English Premier League is comprised of the twenty top clubs within the English Soccer pyramid, which has over 57 leagues and 1000 teams within it.[1] Within each season, the top six teams are rewarded with the opportunity to play in international competitions,[2] while the bottom three teams are relegated to the second division. Financially the impact of international competition and relegation are enormous. Just making international competition earns clubs over $15 million, with the opportunity to earn near $100 million, whereas relegation is estimated to cost clubs around $150 million in valuation. Simply put, the stakes to win cannot be higher.

With the enormous pressure to win, teams must perform offensively to avoid relegation. When it comes to offense, there are two important factors that determine a team's offensive performance: the ability to create quality opportunities and the ability to score goals (finishing). There are numerous ways to define a "quality" opportunity, but one of the simplest ways is to use distance from the goal. The closer the shot is to the opponent's goal, the better the opportunity is.[3] Finishing is much simpler; it's whether or not the shot went in the goal. Yet, it is unclear if these are related, and whether they correlate with overall team success. A team can score many goals from "poor"

[1] The English Football (Soccer) League system is a series of interconnected leagues for men's association football clubs in England, with promotion and relegation between leagues at different levels.

[2] International competitions include the UEFA Champions League and the UEFA Europa League.

[3] Shot quality can also be influenced by factors such as the angle of the shot, the position of defenders, and the goalkeeper's position.

Authors' Contact Information: Landen Fogle, University of Nebraska-Lincoln, Lincoln, NE, USA; Max Sievenpiper, University of Nebraska-Lincoln, Lincoln, NE, USA; Tage Zerby, University of Nebraska-Lincoln, Lincoln, NE, USA.

ACM XXXX-XXXX/2024/5-ART

opportunities or fail to convert “good” opportunities into goals, leading to unexpected results. As such, the research objected for this study is to investigate the relationship between the two and determine whether shot quality, finishing, or some combination of both are more responsible for team success.

2 RELATED WORKS

Considerable work has been done in machine learning to analyze different aspects of sports, especially when it comes to the offensive side of sports. With the amount of data available, from both box scores and cameras, there is incredible potential for breakthroughs in sports strategy. One such way tracking data has been utilized is in a study done by Micheal Hamilton, which is focused on Major League Baseball statistics and mathematics behind pitching success [2]. Pitch analysis is useful for helping teams better understand the effect pitching has on the predicted outcome and strategy. This paper uses machine learning classification models to classify pitches and correlate them to other game statistics, a unique application we may look to apply in our model. Additionally, it highlights strategic use cases for machine learning in sports, since access to this data could significantly impact how managers prepare for the game. If something similar were developed in association football, there is no doubt managers would use it when implementing match day strategy.

Another application of machine learning in sports is Santhosh Narayan’s paper focusing on applying Principal Component Analysis [6] to different features of basketball, such as shot quality, shot-making ability, and height, to perform multiple analyses, mainly mapping player footprints and clustering player archetypes. From this analysis, Narayan has an interesting finding, which is that his model, which accounts for numerous factors, matches the dominant strategy in the NBA. That is to say, it predicts the highest-value shots are close shots and three-pointers, a strategy employed by almost every NBA team. For our application, it provides a good basis to work on. First, it highlights different facets of machine learning that can be used in tandem together (PCA and k-means), which would be impactful for our project as we move into building our model. From both of these studies, we can see how machine learning can be applied to look into strategies and the different ways these models have come to their conclusions.

For association football, machine learning algorithms have been continuously introduced into how we interpret the game, especially towards attacking. In a study by Kerys Harrop [3], Harrop focused on identifying performance indicators predictive of success in an English League One soccer team during the 2012-2013 season. It was found that successful passes, fewer dribbles, and a direct style of play, rather than possession, were associated with winning. Significantly, more passes and offensive actions were linked with losses, suggesting a more effective strategy might be to prioritize direct attacks over maintaining possession. This summary aligns with our research topic by emphasizing the effectiveness of executing scoring opportunities over merely creating chances. It supports the notion that in lower-league soccer, direct play and making the most out of fewer chances might lead to more success than a possession-heavy style. Having established a use case for machine learning algorithms, we can finally delve into the most important portion of our related research: Expected Goals.

In a study by Mat Herold [4], he discusses how machine learning has impacted football attacking play, and continued applications for machine learning models in football. Specifically, it addresses how machine learning models have already been applied in football attacking, including the creation of the expected goals metric. Expected goals are one of the key factors we intend to investigate for our project, so learning how the expected goals metric was created to evaluate attacking creation adds to the depth of knowledge and provides insights into possible takeaways we might have. With expected statistics, we can reduce the overall variance of datasets, since they take out much of the luck associated with sports.

Herold's point is expanded upon in a paper by James Heiwett [5] that explores the impact of Expected Goals (xG) in soccer, utilizing machine learning to adjust xG values for player and position specificities. It develops novel features, such as Goalkeeper positioning and Player Pressure Radiuses, to refine xG predictions. Through detailed analysis, it establishes that forwards are the most efficient shooters, with positional adjustments confirming their superiority in goal-scoring efficiency. This highlights the significance of not just creating chances but efficiently converting them into goals, as demonstrated through advanced xG metrics refined by player and positional adjustments. For our research, this development is huge, as it redefines a key measure we intend to use in our model, and provides key insight to potential answers to our research question. Additionally, it could lead to strategy shifts, as it highlights the importance of getting attackers the ball in the right position, which could be reflected in our Premier League Data from the 2023-2024 season.

3 METHODOLOGY

This study aims to predict outcomes of Premier League soccer matches using a machine learning approach, focusing on metrics related to scoring opportunities and finishing quality. The methodology involves data collection, data processing, feature engineering, model training, testing, and evaluation, as detailed below:

3.1 Data Collection

To investigate how a Premier League team's chance creation and finishing quality predict success in matches, we begin by gathering data on scoreline results [4] and expected goal (xG) [5] generation from both teams over a series of games. This data is sourced from reputable soccer statistics website [1] and compiled into a CSV file with the following columns:

Expected Goals	Over/Underperformance of Expected Goals	Result
2.7	-0.7	1
3.12	-0.12	1
0.87	0.13	0
2.24	0.76	1

Table 1. Example training data

- **Expected Goals (xG)**: The primary measure of chance creation during a match.
- **Over/Underperformance of Expected Goals (G - xG)**: Indicates the quality of finishing based on the difference between actual goals and expected goals.
- **Result**: Binary outcome (1 for a win, 0 for a draw or loss).

3.2 Feature Engineering

Additional calculated features included the goal difference from Expected Goals (xG) for both home and away teams:

[4] In soccer, draws are also possible and are not always, but generally viewed as a negative result as they count for 1 point, whereas a win counts for 3, and a loss for 0. A draw therefore is two possible points dropped.

[5] This is the primary measure used to analyze how well a team finished the chances they created.

DiffHxG = HomeG - HomexG

Difference between actual and expected goals for the home team.

DiffAxG = AwayG - AwayxG

Difference between actual and expected goals for the away team.

3.3 Model Training and Testing

After gathering this data, we will perform a KNN analysis. Once our K-value is determined, each value in our test data will be predicted as a win or loss given the euclidean distance to the nearest k points.

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

The feature sets and target variable for the model were defined as follows:

- X: Comprising xG and DiffxG, representing the features.
- Y: Indicating the match result, used as the target variable..
- The data was split into training and testing sets, with a 80/20 split.

After receiving the results of our KNN model, we will create a variety of tests for fake games and gather those predictions at certain percentile or tiers of chance creation and finishing. E.g. how does a 90th percentile xG and 30th percentile over/underperformance of xG game compare to the inverse and so one. The goal of these sorts of tests will help us to better answer our research question.

3.4 Predictive Forecasting

Our KNN model is able to take statistics for a game, and be able to predict the result for it, solely based on chance creation and chance conversion. While incredibly useful, this model fails to directly determine which of xG or difference in xG is more important in team success, therefore failing to answer our research question. As a result, we needed another model to fully answer our research question. To address this, we opted to create two forecasting models, one that predicts games based off of season long averages and another that predicts games based off of a much shorter rolling average. Our forecasting models were very basic, as we built them from scratch based off our data. For the season long model, we first decided to take the first 21 match weeks, or 75% of the data set, as our training split and the last 7 match weeks, or the remaining 25% of the data set, as our testing split. To "train" our model, we averaged the statistic in question, which was either xG or difference in xG, for both home and away games during that span. For our predictions, we split up home and away averages into two categories, since there is an established difference in how teams play. With these averages, we then needed to figure out a classifier to predict whether or not a game would result in a tie. To do this, we averaged the difference between the two-teams statistics in games that result in a tie over our training data (ex. Home xG - Away xG), and divided it by the square root of the number of matchweeks, or 21 as seen in Equation 1.

$$\text{AvgTieXg} = \frac{1}{\sqrt{N}} \left(\frac{1}{n} \sum_{i=1}^{n} \text{diffxGDiff}_i \right) \tag{1}$$

With all necessary statistics, we can finally forecast. To build the forecasts we built a very simple classifier to determine the result of a game. For every matchup in the testing data, we found the difference between the home team's average and the away team's average. If this difference was less than the tie classifier, it would be marked as a tie (T), otherwise the team with the higher average would be marked as the winner (H or A). Then, we would compare this with the actual result to determine the accuracy. Equations representing the forcasting models show in Equations 2.

$$\text{AvgHomeXg} = \frac{1}{n}\sum_{i=1}^{n} \text{HomexGDiff}i \qquad \text{AvgAwayXg} = \frac{1}{n}\sum i = 1^{n}\text{AwayxGDiff}_{i} \tag{2}$$

Our rolling average forecast works very similarly to our season-long forecast model, with a few key differences. For the purpose of continuity, we will still test our rolling average model on the test data set. Rather than training on all 21 weeks and predicting the rest of the season, the rolling average model will train on the last p weeks and predict the next match week. After it predicts the next match week, it will repeat the process, moving up a week and training on the last p weeks and predicting the next match week until the entire tet set has been predicted. The tie classifier works the same way, dividing by the square root of p instead of 21. Once all match weeks have been predicted, we compare with the actual result to determine the accuracy.

$$\text{HomeAvgxG}_{w} = \frac{1}{p}\sum_{i=w-p+1}^{w} \text{HomexG}_{i} \qquad \text{AwayAvgxG}_{w} = \frac{1}{p}\sum_{i=w-p+1}^{w} \text{AwayxG}_{i} \tag{3}$$

3.5 Expected Output

Our goal for this model would be that it could predict a result of 1 or 0, that is a win or a draw/loss, with high accuracy given the team's xG and G - xG. Given knowledge of the sport and an investigation into literature on the subject, we expect to discover from our analysis of our model: a team's volume of chance creation weighs more heavily on their success in a given soccer match than a team's ability to finish created chances. With these forecasting models, we hope to determine which statistic, xG or difference in xG, is more important to predicting team success. Likely, this will be the statistic with the higher overall accuracy, as this better 'predicts' a teams success throughout a season. After combining these results with the results of the KNN model, we should be adequately prepared to answer our research question.

4 RESULTS

4.1 Model Performance

To start, we performed a KNN analysis for each feature individually. That is, we trained the model on just xG and the game's result, and did the same for diffxG. The idea here was that if one model was substantially more accurate at predicting whether there was a win or a loss, we knew that this feature was more important to a team's success. Unfortunately, neither model was particularly accurate, or more accurate than the other, as one was 67% and the other was 69% accurate. So, we decided to perform a KNN analysis on both features and analyze the model from there. We performed a KNN analysis on a dataset including only matches from the currently ongoing English Premier League season. We tested K values 1-10 show in Table 2.

K	1	2	3	4	5	6	7	8	9	10
Accuracy	76%	74%	80%	75%	81%	78%	76%	76%	77%	79%

Table 2. Accuracy of Model Predictions by K-Values

4.2 Model Results

After analysis we decided that a K value of five was most appropriate as it had the highest accuracy all k-values. Additionally, five was the typical amount of games used to measure a team's form in soccer, so looking at the five most similar games seemed to predict any given match best. Our KNN analysis resulted in the scatter plot that can be seen in Figure 1. Additionally we created a resulting KNN Decision Boundary, which showcases where each game was predicted to be a win or not, and what the actual result was as shown in figure 2. As seen in figure 3, a graph of the fake data games played out by our KNN model, the results were pretty inconclusive. When analyzing each feature, we collected the model's predictions for a game with every percentile for that feature from the 10th to the 99th, with the other feature being from the 10th to 50th, to determine if one feature did better than the other when one feature was below average to average. Doing this resulted in 14 total predicted wins when analyzing the feature xG and 15 total predicted wins when analyzing the feature DiffxG. The graph also appears fairly symmetric.

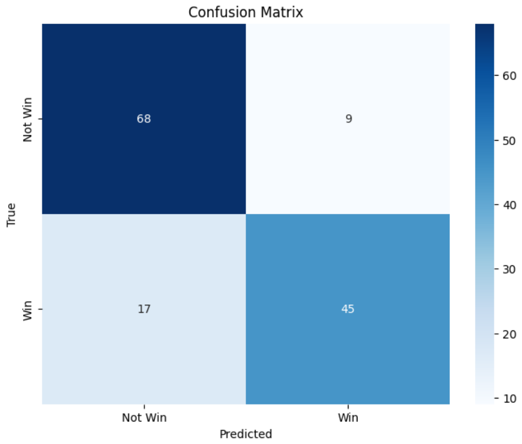

Fig. 1. KNN Confusion Matrix

4.3 Forecasting Results

From our forecasting model, we can determine more about the weight of our individual features of our research question. As can be seen from Table 3, on a season-long average, xG weighed significantly more in predicting a team's success than a team's chance conversion rate (G-xG).

Forecasting Accuracy	**xG**	**DiffxG**
Season Long Average	60%	35.7%
Rolling Average (r=3)	58.9%	62.8%

Table 3. Forecasting accuracy comparison between xG and DiffxG

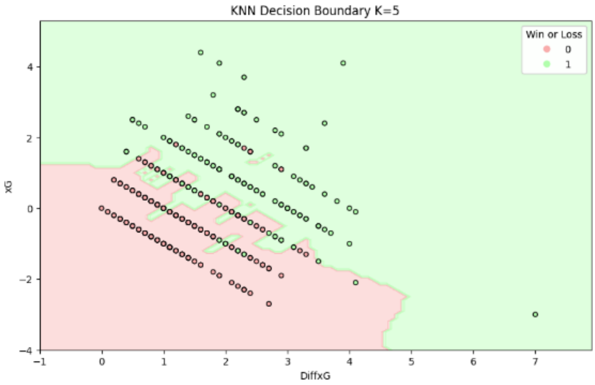

Fig. 2. KNN Model Output where K value is 5.

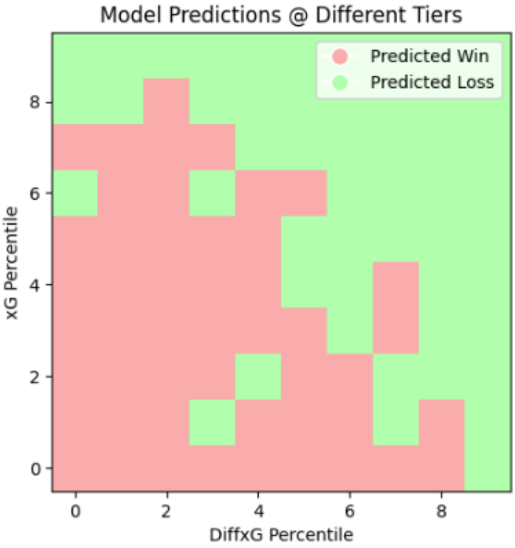

Fig. 3. Fake game predictions from KNN Model.

5 DISCUSSION

As illustrated in Figure 2, the points on the graph appear to align along a negatively correlated line. This behavior is due to the nature of the two metrics involved. Goals cannot be partial, and at any given number of goals, a higher expected goals (xG) leads to an equally lower over or under performance of that xG and vice-versa.

From the graph alone, it is challenging to draw significant conclusions to address our research question. However, there are a few observations worth mentioning:

(1) 1-2 goals generally don't secure a game win, but scoring 3 or more almost always does. This finding can guide further analysis, potentially leading to identifying feature tiers with expected wins.

(2) The observed frequency of games with 4 or more goals differs from the frequency of games with 4 or more xG. This can be due to the current Premier League season's context and trends. This season has already set the record for most goals scored, suggesting that overperformance of xG could result from more clinical finishing of chances.

One possible explanation for the discrepancy in the frequency of games with 4 or more goals compared to those with 4 or more xG is the context of this Premier League season. With dozens of games yet to be played, the season already has the highest goal tally in history. This trend could be due to improved accuracy and conversion of scoring opportunities. Another explanation is that games with higher goal counts tend to overperform xG. Each goal inherently represents an overperformance of its expected value, balanced by games where fewer goals are scored. For example, games with no goals would result in complete underperformance of xG, which is not uncommon.

Our forecasting model was able to predict with 60% accuracy whether a team won, lost, or drew a game based on their xG alone, whereas that same model only had a 35.7% accuracy for predicting game outcomes based on a team's finishing quality, which is practically random. Surprisingly at first, on a rolling average basis, the two features lined up fairly well, both at about 60% accuracy. In fact, finishing quality was a slightly better predictor in the short term. After some consideration, this seems to make sense. Finishing quality has a lot more to do with a team's form. If the team has been finishing their chances well in recent weeks, that is a solid predictor of them generally playing well, and vice-versa. This means that both features can predict games in the short-term, at a similar rate of success. However, long-term, sustainable measures are more significant for evaluating our research question. In this regard, xG is the only feature that can be considered a predictor for a team's long-term success.

6 LIMITATIONS

First, as mentioned previously, this Premier League season is the only one in our dataset, and it's unusually high scoring. This might result in us making conclusions that may be true for this EPL season, but aren't typical for professional soccer as a whole. Adding in different seasons and leagues might help us make more concrete conclusions. Second, we will consider potentially removing games with own goals, although this is a bit tricky. The reason behind this is how we currently model over/underperformance of xG, which is our metric for how well a team finished in the game. It's currently gathered by the goals the team scored - their xG. This results in any own goal scored as being 1.0 added to that number. Every own goal counts for 0.0 xG, as it wasn't a shot made by the team. An own goal doesn't say anything about how well a team finished their chances in a game, but currently it is being counted as an incredible finish from a chance that had quite literally 0 chance of being scored. It would be difficult to remove these games from our dataset though and it's also true the vast majority of own goals are still a result of the attacking team's creation, not purely error by the defense. Still, it's worth considering the possibility of removing these goals from our G-xG feature, if possible.

7 FUTURE WORK

If someone were to continue this research they would want to look at a team's season as the row of our dataset, not an individual game. First, comprise a dataset of many premier league seasons with features for a team's position in the table, total xG created over the course of the season, and their total over or underperformance of their expected goals. Next, train a machine learning model on every team's xG and position. Do the same for their G-xG and their position. From their test both model's accuracy, create confusion matrices, so on and so forth. If one of the model's is substantially better at predicting a team's position in the table (or general tier in the table if some feature-engineering is required), you know that this is a better predictor of a team's success. This data would not be easy to gather though. xG is a very new metric in soccer, as in, it's only really been a talking point for the last couple of years, and data on xG is difficult to obtain for a season even 10 years ago. Since each position in the table would only occur once a season, many, many seasons would likely be necessary for a model to be accurate, and there just isn't long term data on xG readily available. Additionally, data for a team's xG is typically only gathered on a game-to-game basis, not as a total across a team's season, meaning the research would likely have to compile that total themself. That said, if a dataset such as this was compiled, over a long enough period, it would probably yield a much better answer to our research question.

8 CONCLUSION

From our KNN model, we gathered that xG, which is a measure of a team's chance creation, as well as G-xG, called diffxG, which is a measure of a team's finishing quality or chance conversion ability, are both predictors of a team's success in an individual game. To put it simply, both are important to whether a team wins or loses. They are so important that our model, with just two thirds of one season of data, could predict a game's outcome with >80% accuracy, without knowing anything about what the other team created or any other metrics such as possession and field tilt. However, even after analyzing 100 different hypothetical games at various tiers of chance creation and finishing quality, our KNN model was unable to give us any concrete conclusion on which feature was more important to a team's success in soccer. This is why we also added our forecasting model. Our forecasting model demonstrated that xG was a feature that loosely predicted long-term success, whilst diffxG was not an accurate predictor. It also told us that in the short-term, both were loosely accurate predictors of a team's success. This leads us to our main conclusion. From both our models, we conclude that chance creation (xG) and finishing quality (G-xG) are related to a team's form, and thus both are loosely accurate predictors of a team's short term success. However, only chance creation (xG) can be said to be a predictor of a team's success in the long-term. Unfortunately, this conclusion is based on a loosely accurate model, and does have the limitations mentioned above.

REFERENCES

[1] FBref. 2024. *Premier League Scores and Fixtures.* https://fbref.com/en/comps/9/schedule/Premier-League-Scores-and-Fixtures

[2] Michael Hamilton. [n. d.]. https://www.researchgate.net/publication/326972628_Applying_machine_learning_techniques_to_baseball_pitch_prediction

[3] Kerys Harrop and Alan Nevill. 2014. Performance indicators that predict success in an English professional League One soccer team. *International Journal of Performance Analysis in Sport* 14, 3 (2014), 907–920. https://doi.org/10.1080/24748668.2014.11868767 arXiv:https://doi.org/10.1080/24748668.2014.11868767

[4] Mat Herold, Floris Goes, Stephan Nopp, Pascal Bauer, Chris Thompson, and Tim Meyer. 2019. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching* 14, 6 (2019), 798–817. https://doi.org/10.1177/1747954119879350 arXiv:https://doi.org/10.1177/1747954119879350

[5] James H. Hewitt and Oktay Karakuş. 2023. A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open* 4 (2023), 100034. https://doi.org/10.1016/j.fraope.2023.100034

[6] Narayan Santosh. 2019. Applications of Machine Learning: Basketball Strategy. Online; Accessed: 2024-04-25. https://dspace.mit.edu/handle/1721.1/123043

Received 11 March 2024; revised 08 May 2024