Analysis and Prediction using Premier League Football Data

Victor Brown, Darasimi Adeyemo, Harper Schwab
Data 11900, Professor William Trimble
Due May 13, 2022

Introduction and Dataset Background

The Expected Goals and Other Metrics dataset [3] is a compilation of parameters and metrics in European football leagues starting in 2014 and ending in 2019 collected by web-scraping data gathered and computed by the Understat organization. [4] Understat [6], specializes in a metric called expected goals (xG) which according to Analyst [7] "measures the quality of a chance [of a goal] by calculating the likelihood that it will be scored from a particular position on the pitch during a particular phase of play". The calculation often utilizes logistic regression with parameters such as: Distance to the goal, angle to the goal, one-on-one, big chance, body part etc. [7] Understat collects data on six football leagues and all teams within. This encouraged us to refine the dataset to one of the top five leagues, wherein we chose the English Premier League. This league was chosen for two primary reasons: Victor's extensive knowledge, watching around 50 EPL games per season, and the simple fact that Victor's favorite team, Tottenham Hotspur, competes in the Premier League. Within our focus on the English Premier League (EPL) we will focus specifically the top eight football clubs out of the twenty in the EPL: Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, Tottenham Hotspur, West Ham United, and Everton. In the English Premier League, teams compete against the other 19 teams once at home and once away for a total of 38 games per season. The dataset contains six years of data, for a total of 38 games * 8 teams * 6 years = 1824 perspectives. This will allow for faster computation by limiting the number of rows of our dataset from over 24500 to 1824. The subset will also allow us to hone the scope of our main question: Can classification accurately predict the result of a "top eight" English football team's match using common parameters such as excepted goals? This question will also provide insight into the accuracy of some of the most commonly used football statistics.

The Expected Goals and Other Metrics dataset [3] contains variables including but not limited to: home or away game, expected goals, team name, passes completed within roughly 20 yards of goal, passes allowed, and many different categories of excepted goals such as difference between actual goals conceded and expected goals conceded. One conflict we found with this dataset is that some metrics, if used for analysis and classification, seem almost counter-intuitive to what we are trying to classify. For instance, if we are trying to predict a win, loss, or draw, we wouldn't want to use elements that state the number of points a team earned from the match in question because if the number of points earned is directly correlated with if the match was a win, loss, or draw. If we were to use columns like this, we would essentially be predicting an outcome based on the outcome, which would be a form of self prediction, and unhelpful to our question.

Prior Work

The curator of the dataset, Sergi Lehkyi, previously performed an analysis on the dataset titled: *Football: Why Winners Win and Losers Lose* [2]. In this analysis Lehkyi focuses on expected goals, expected assists, and expected points and attempts to determine how much of football club's winning is determined by luck and skill. Expected assists is a measure of "the likelihood that a given pass will become a goal assist" [2], expected points is "the likelihood of a [certain] game to bring points to the team" [2], expected goals is defined earlier in the introduction. In their analysis Lehkyi used football expertise to make sense of the data collected from Understat. By comparing expected and actual points Lehkyi was able to determine how the performance of winning teams is compared to what was statistically predicted. For example, in Lehkyi's analysis of Bayern Munich

from the Bundesliga, it was determined that their out-performance of expected points won them the league, except for 2018 in which it was attributed to the fact that "competitors didn't take advantage of" the worse performing Bayern Munich that year. By plotting the difference between expected goals, assists, and points with their actual counterparts, Lehkyi was able to affirm the expected result that the top teams "score more, concede less and get more points than expected" [4]. This assertion lead him to search for outliers when comparing expected goals to actual goals. By finding the outliers Lehkyi was able to determine the probability of either succeeding in a significant number of unexpected goals or missing and significant number of expected goals. Being an outlier means finishing in the top four positions to guarantee entry to the Champions League, or being relegated to The Championship (second division) with a probability of 8.1% as these outliers perform better or worse than their expectations. This, according to Lehkyi demonstrates a level of luck that is apparent in all of football. Lehkyi determines the great importance of expected goals and other derivatives in football for analyzing successful vs unsuccessful teams, and provides explanation of seemingly unpredictable phenomena in football. To Lehkyi, "it is almost 100% chance that something weird will happen in one of the leagues". [4]

Analysis

To understand the accuracy and validity of statistics such as expected goals and it's derivatives we analyzed how the difference of actual goals and expected goals and how they correspond with various teams in the European Premiere League. By grouping all values by year and taking the mean of all the values in a corresponding year some interesting trends were uncovered.

Seen in figure 1, all of the top eight EPL football clubs have on average under-performed during the six years contained within this dataset. In other words, on average the teams have scored less than their expected goals. This could mean that expected goals is on average an overestimating statistic. To further investigate the usefulness of expected goals we plotted the yearly average expected goals of the three football clubs who won the Premier League between 2014 and 2019 which were Manchester

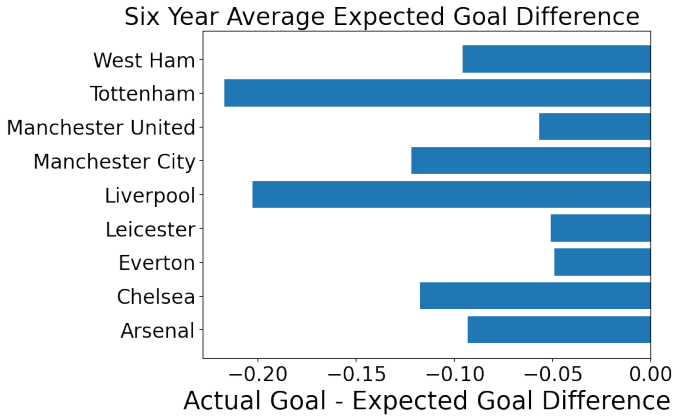

Fig. 1. Six Year Average

City, Chelsea, and Leicester. Shown in figure 2, all of the winning years had expected goal differences between -0.2 and 0.1, meaning that when the teams performed close to expectation (0 on the graph) they were more likely to win compared to their previously winning counterparts. This is surprising because especially when referencing the dataset analysis by Lehkyi [4] one would expect the average expected goal difference to be a larger positive value than is demonstrated in our analysis.

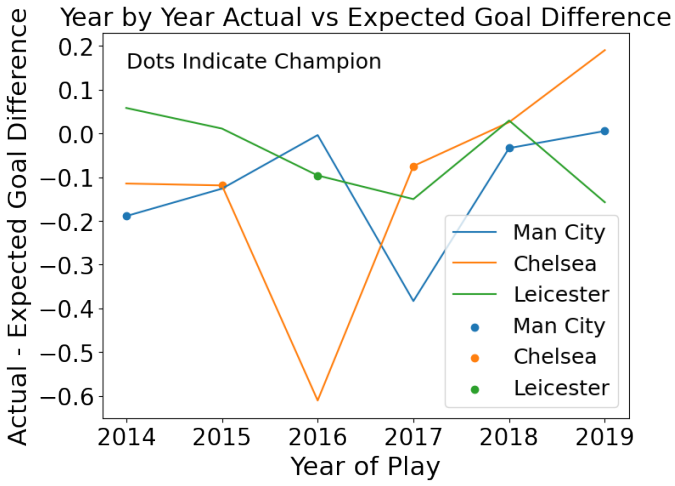

Fig. 2. Expected Goals of Winners

However, figure 2 does demonstrates a 'winning trend' of meeting expectations over the year average. This provides evidence to continue our analysis into the expected goals statistics and their helpfulness when predicting a team's success in a match. To further work on this question we decided to use classification.

Classification

Our group's main goal was to be able to make and analyze a prediction of whether the final result

was a win, loss or draw using expected values. After looking throughout the dataset provided by Lehkyi, we identified a few key parameters that we could use by a classification model. Of all possible parameters, we identified the following to be possible classification parameters.

xG: expected goals
xGA: expected goals conceded
npxG: **xG** without penalties and own goals
npxGA: **xGA** without penalties and own goals
scored: goals scored in the current game
conceded: goals conceded in the current game
npxGD: difference between **npxG** and **npxGA**
xG_diff: difference between **scored** and **xG**
xGA_diff: difference between **scored** and **xG**

After importing Lehkyi's dataset, we set out to refine this dataset to better classify the result of a game based on certain parameters. As previously mentioned, the subset that we selected only involved the "top eight" teams in the EPL. These teams were selected based off of their reputation for success and similar playing styles, which allowed us to ensure a somewhat normal, and similar, sample group for classification. We added a column called *w/l/t* which took in a 0, 1, or 2 representing win, loss, or tie respectively in order to be able to classify using machine learning.

Originally we wanted to try a new type of classification, so we attempted to use Gaussian Naive Bayes regression model and had successful output. However, upon further investigation of the model it was discovered that the Gaussian model assumes independence of the parameters passed into the model. This posed an issue because metrics such as expected goals and expected goal difference, and all expected goal or assist derivatives are not independent from each-other and therefore go against the assumptions of the model. In response we turned to the sci-kit learn python library [5], specifically the KNeighbors classifier, because of our familiarity of KNN algorithms and the modules use of Minkowski distance which allows for a multi-dimensional distance calculation instead of the euclidean two point distance.

After determining that we wanted to use the KNeighbors classifier, we defined a function called *model_machine* that took in two arguments: a dataset, and the classifying parameters. Due to the fact we didn't know which combination of the nine possible parameters would yield the best accuracy, we defined a function called *combination_maker* that took in two arguments, an array of all possible parameters, and the minimum number of items in each new array generated. *combination_maker* returned a 2D-array where each value in the array was a distinct combination of parameters of size minimum number of items in each new array - len(array of possible parameters). There was a third function defined called *check_best_params* which took in a dataset, all possible parameters, and the minimum number of parameters we wanted to look at each time. This generated a tuple of the list of parameters, and the accuracy of our model given these parameters. Finally, our fourth function was named *find_best_params* which has two arguments: the output of *check_best_params*, and an accuracy threshold that our parameters had to be greater than or equal to.

In short, we defined a series of functions that iterated through every possible combinations of the majority of numeric columns and returned the accuracy given each of these possible combinations. Then we sorted the output to only include parameters that generated an accuracy greater than or equal to the threshold provided.

Parameters	Accuracy
xG, xGA, scored, conceded, xG_diff, xGA_diff	0.9854
xGA, scored, conceded, xG_diff	0.9817
xGA, scored, conceded, xGA_diff	0.9817
npxGA, scored, conceded, xGA_diff	0.9817
xG, npxG, scored, conceded, xG_diff	0.9817

TABLE I
Table of Accuracy with Expected Parameters

Parameters	Accuracy
h_a, deep, deep_allowed, oppda_coef, oppda_def	0.540146
h_a, deep, deep_allowed, oppda_def	0.532847
h_a, deep, deep_allowed, oppda_coef	0.529197
h_a, deep, oppda_att, oppda_def	0.529197
deep, deep_allowed, ppda_coef, oppda_def	0.529197

TABLE II
Table of Accuracy Without Expected Parameters

This allowed us to create a dataframe with our conclusions. This dataframe had two columns: the

parameters, and the accuracy that they returned. We were able to create two dataframes, one with parameters that included expected values, and one with parameters that did not include expected values. The parameter that did not include expected values instead used:

h_a: home or away game
deep: passes completed within 20 yards of goal
deep_allowed: opposition (OP) **deep** passes
oppda_def: OP passes per defensive action
oppda_coef: **oppda_def** coefficient (Coeff)
oppda_att: **oppda_def** attempted
ppda_coef: passes per defensive action Coeff
scored: goals scored in the current game
conceded: goals conceded in the current game

The top performing values of these two tables can be seen in Table I, and table II .

Findings and Conclusions

Our maximum accuracy in prediction was 98.54% with the parameters: xG, xGA, scored, conceded, xG_diff, xGA_diff. This leads us to believe that classification using these parameters yields a result that could very effectively predict games based on a series of metrics. Using this accuracy, we would be able to accurately predict the results for 37 out of 38 of the games, which leads us to at most a 3-point margin of error. In order to find whether the specific parameters we chose were significant, we tested our functions with a set of parameters that didn't include expected values such as: home or away (binarized as 1 for home, 0 for away), passes completed within 20 yards of the opponents goal, passes allowed within 20 yards of their own goal etc. These parameters yielded a maximum accuracy in prediction of 54.01% which proves that the usage of expected values was useful in classifying the result as a win, tie, or loss. In figure 3 the accuracy value of the KNN model is shown on the y-axis, and the different combinations of parameters are shown as the graph increases on the x-axis with a different group every integer increase on the x-axis. Our findings demonstrate not only the accuracy of the usage of expected goals as a metric to predict the outcomes of football games but also the unreliability of other metrics that were commonly used before the expected goal metrics and it's derivatives became popular. For a league commonly known as the most unpredictable league in the world, our usage of expected metrics has cut through the unpredictability to give us an incredibly high prediction accuracy. In the future this model could be applied to attempt to predict an entire EPL season, or other league's season by using expected metrics which promises an even more exciting field of sport statistics in the worlds most unpredictable league. Although our model wouldn't be able to predict 5000-1 odds of Leicester City winning the Premier League in 2015/16, it is an exciting new step in the field of sports analytics.

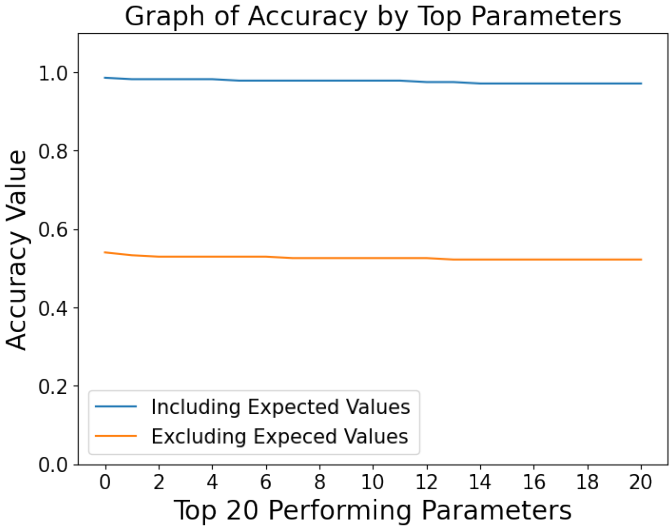

Fig. 3. Expected Goals of Winners

Group Statement

Within our group, we divided up this task into three equally important components: building the framework for this paper including previous knowledge, creating figures and methods of representation for data, and writing the code that enables us to use machine learning to classify a result as a win draw or loss depending on certain parameters. Victor Brown was in charge of developing the machine learning code for classification, Darasimi Adeyemo was in charge of creating figures and methods of representation for data, and Harper Schwab was primarily focused on the framework of the paper and statistical analysis, excluding the machine learning component.

REFERENCES

[1] How to classify data in python using scikit-learn. https://www.activestate.com/resources/quick-reads/how-to-classify-data-in-python/, journal=ActiveState, author=M, Remi, year=2021, month=Nov.

[2] Sergi Lehkyi. Football: Why winners win and losers lose, Sep 2019. =https://www.kaggle.com/code/slehkyi/football-why-winners-win-and-losers-lose/notebook.

[3] Sergi Lehkyi. Football data: Expected goals and other metrics, 2020. https://www.kaggle.com/datasets/slehkyi/extended-football-stats-for-european-leagues-xg/metadata?resource=download.

[4] Sergi Lehkyi. Web scraping football statistics 2014-now, Apr 2021. https://www.kaggle.com/code/slehkyi/web-scraping-football-statistics-2014-now/notebook.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Understat. Xg stats for teams and players from the top european leagues, May 2022. https://understat.com/.

[7] Johnny Whitmore. What are expected goals (xg)? July 2021. https://theanalyst.com/na/2021/07/what-are-expected-goals-xg/.

[8] Worldfootball. Premier league - champions. https://www.worldfootball.net/winner/eng-premier-league/, journal=worldfootball.net, publisher=HEIM:SPIEL, year=2021.