A Statistical Look into how Common Soccer Metrics Influence Expected Goal Measures in the Professional Game

Tristan Rumsey[a] *& Shaha Alam Patwary**[b]

[a] *Eli Lilly and Company, Indianapolis, IN and Department of Mathematical Sciences, Butler University, Indianapolis, IN*
[b] *Department of Mathematical Sciences, Butler University, Indianapolis, IN*

https://doi.org/10.33697/ajur.2025.161

Student: tristangrumsey@gmail.com
*Mentor and Corresponding Author: mpatwary@butler.edu**

ABSTRACT
The advent of sports analytics has ignited a fervor across all sporting disciplines, particularly soccer, where clubs are sprinting to harness vast data reserves to elevate team performance, spearhead effective marketing endeavors, and bolster financial gains crucial for club expansion. Much like Billy Beane's transformative "Moneyball" approach, soccer clubs are in pursuit of innovative strategies to transcend financial limitations and achieve triumph. In soccer, where goals are scarce commodities, heightened offensive efficacy becomes imperative. Presently, one metric stands out as pivotal in gauging a team's goal-scoring success: expected goals (xG). This metric quantifies the likelihood of a given shot or opportunity culminating in a goal, making it a linchpin in a team's offensive strategy. Maximizing expected goals becomes paramount for teams aiming to capitalize on limited scoring opportunities during matches. Crucially, the first step in reshaping tactical approaches hinges on identifying the most influential variables in predicting expected goals. This study employs an array of machine learning methodologies, including Ridge, Lasso, Elastic Net, and Group Lasso models. The objective is to unveil the key predictor variables that significantly impact team (offensive) performance, often delineating the thin line between championship glory and defeat. With the aim of predicting xG, this research also incorporates modified bootstrap techniques to compute prediction intervals for the regularized machine learning models. By delving into the intricate fabric of soccer analytics, this study seeks to empower clubs with actionable insights, fostering a new era of strategy and competitive edge on the field.

KEYWORDS
Soccer analytics; Expected goals; Managerial strategy; Statistical and machine learning methods; Bootstrap method; Prediction interval.

INTRODUCTION

In 2007, Liverpool Football Club faced Associazione Calcio (AC) Milan in the Champions League Final – a rematch of their dramatic 2005 encounter. While the final scoreline (2–1 to Milan) reflected a typical result, it did not capture the broader story. Milan had long prepared for this moment through its innovative "Milan Lab," where doctors and performance analysts examined players' physiological and biomechanical data such as jumping ability, heart rate, muscle weakness, and eye movement. The lab claimed that jump metrics alone could predict injuries with 70% accuracy.[1] This represented one of the earliest instances of using data to optimize player health and performance in soccer. The explosion of sports analytics began with the Oakland Athletics' revolutionary "Moneyball" strategy in the 2002 baseball season.[2] The Athletics' success demonstrated how data–driven strategies could allow underfunded teams to compete with wealthier clubs. Since then, analytics has evolved into a multibillion-dollar global industry, valued at more than $2.5 billion in 2022 and projected for rapid growth.[3] Despite initial skepticism from traditionalists such as Wilbon,[4] Rose,[5] and Cowher,[6] analytics now shape nearly every modern sport.

Compared to American sports, soccer was slower to embrace analytics.[7] Historically, managerial decisions were driven by instinct and short–term performance.[8] However, clubs now collect vast data on match statistics, training sessions, and player fitness. Jean–Pierre Meersseman of AC Milan likened analytics to a car dashboard–helpful information that "makes driving easier."[1] This "dashboard" of data empowers coaches, players, and executives to make informed, evidence-based decisions. Among modern soccer metrics, none has been more influential than expected goals (xG).[1] Expected goals quantify "the probability of a shot resulting in a goal."[9] For example, since 78% of penalties in professional soccer are scored,[10] each penalty carries an xG value of 0.78 rather than a binary outcome of 0 or 1. Summing these probabilities across all shots in a match gives a team's total xG, a key measure for evaluating offensive performance and tactical efficiency. This study focuses on identifying the variables most associated with a team's expected goal total.

Because soccer features few scoring opportunities, understanding which metrics most affect expected goals is crucial. A team's xG reveals what truly occurred on the field–far better than the final score alone.[11] A single goal can define a season; optimizing player performance and team tactics to maximize xG can mean the difference between winning and falling short. Recent advances in soccer analytics aim to model how both on-ball and off-ball actions contribute to success. The central question is how "any action changes the likelihood of scoring."[9] Statsbomb's models consider player positions, shot type, and shot quality, while Perl et al.[12] highlight pattern recognition and machine learning as emerging research frontiers. Data from cameras, passes, dribbles, and positional tracking have expanded the analytical scope of the game.[13]

Before developing this study's models, several existing approaches to goal prediction were examined. Sánchez Gálvez et al.[14] used Logistic Regression, Naive Bayes, Decision Trees, and SVMs to predict match outcomes. Decroos & Davis[15] applied a Generalized Additive Model (GAM) to estimate the probability of imminent goals. Inan[16] used Poisson regression to model goal frequency across teams, and Liu et al.[17] implemented transfer and vision learning methods with an Inflated 3D Network (I3D) model to predict goal likelihoods. These frameworks collectively informed the model selection process in this research. A substantial body of work also validates expected goals as a key metric for performance analysis. Historically, narratives in soccer were driven by outcomes rather than quality of play, but xG reframes performance by evaluating shot quality rather than results.[11] Expected goals capture the fairness of performance–highlighting when teams win by luck or fail despite dominance. Because xG incorporates nearly every event leading to a shot, it provides a nuanced view of team and player performance.

In a study of 5,020 matches, 1,366 matches had xG values matching the final score, and 3,443 matched within a one-goal margin.[8] This strong correlation underscores xG's reliability as a predictor of performance and its growing influence in the sport. Statsbomb's analyses further show that shooting from central areas, favoring foot shots over headers, limiting crosses, and improving finishing are all correlated with higher xG.[9] Over an entire season, these insights become powerful indicators of long-term team strength.[18] Opta, another leading analytics firm, uses an XGBoost-based model that incorporates contextual data such as distance to goal, shooting angle, goalkeeper position, defender pressure, and play type (e.g., fast break, set piece).[19] While both Opta and Statsbomb measure expected goals, their models differ in computation–Opta's inclusion of goalkeeper positioning being one example. Consequently, reported xG values can vary slightly depending on the data provider.

Still, expected goals alone cannot explain match outcomes. Shots account for only about 1.5% of all match events,[20] meaning that actions such as passing, dribbling, tackling, and transitions must also be included to model performance accurately. Thus, this study incorporates multiple offensive and defensive variables, extending beyond shot-based data to capture the full complexity of play. The growing adoption of analytics is transforming soccer management. Billy Beane, the pioneer of "Moneyball," now advises Dutch club AZ Alkmaar, which uses data–driven methods to compete successfully against much wealthier teams.[21] Similar success stories, including Brentford FC in the English Premier League, illustrate how analytics empower smaller clubs to achieve sustainable success through strategic efficiency.

This research explores the intersection of statistics, machine learning, and soccer, emphasizing how data–driven insights can improve decision-making. Players' livelihoods and team success depend on optimizing performance, and understanding which variables drive goal-scoring opportunities has both strategic and financial implications. This study contributes to that understanding by identifying the most significant predictors of expected goals, providing actionable insights for coaches, analysts, and executives.

This manuscript is organized into six sections. The **Introduction** outlines the background, motivation, and literature supporting the study. The **Materials and Procedures** section describes the dataset, variable definitions, data cleaning, and exploratory data analysis (EDA). The **Methods and Procedures** section details the statistical and machine learning methodologies used. The **Results** section presents model performance metrics, accuracy measures, and predictive comparisons. The **Discussion** interprets the key findings and implications, while the **Conclusion** provides a synthesis of insights and recommendations for future research.

MATERIALS AND PROCEDURES

The Dataset

The dataset used for this study worked was manually created with data from Football Reference (known as *FBref.com*). This site was created by Sports Reference to document many statistics in professional soccer matches around the world, dating back to 2017, for many top clubs in their various competitions. These competitions may include domestic leagues, domestic cups, intra-European/continental, or even friendly (exhibition) matches. The data stored on *FBref.com* is collected by Opta, who captures and shares real-time sports data with other companies, and professional teams.

This analysis studies the data for Arsenal Football Club, a well-known team in the English Premier League, over the past four league seasons, beginning with the 2019-2020 season. The English Premier League is widely regarded as the most competitive league in the world, and is currently recognized as such, as the Union of European Football Association (UEFA) has ranked it as having the highest league coefficient ranking. Throughout the duration of these four seasons, there have been a number of domestic opponents that Arsenal has faced, including the following clubs: Aston Villa, Bournemouth, Brentford, Brighton & Hove Albion, Burnley, Chelsea, Crystal Palace, Everton, Fulham, Leeds United, Leicester City, Liverpool, Manchester City, Manchester United, Newcastle United, Norwich City, Nottingham Forest, Sheffield United, Southampton, Tottenham Hotspur, Watford, West Bromwich Albion, West Ham United, and Wolverhampton Wanderers.

The complete dataset contains a total of twenty one variables (including the response variable), with a total of 152 records, accounting for thirty eight league matches per season, spanning four seasons. It is important to note that, while Arsenal plays many non-league matches throughout the course of a season, the number and frequency of additional matches is variable. Therefore, to keep each season consistent with the others, this study only analyzes data from the league-specific matches. Many of these records are quantitative data types, but qualitative data will also be present. The dataset was split into train and test datasets. The train dataset includes data from the first three seasons of analysis, including the 2019-2020 (1), 2020-2021 (2) and the 2021-2022 (3) seasons, and the test dataset includes data for the most recently completed season at the time of analysis, 2022-2023 (4). The train dataset will be used to train a model that will then be used to predict values for comparison to the test dataset to evaluate the performance of the adopted model.

Response and Predictor Variables

As the main objective of this research is to identify the metrics that most influence a team's total number of expected goals (xG) in a match, the variable *xG* will be focused on as a response. The response variable is continuous in nature. Additionally, the "Match" variable serves as a unique identifier for each observation. The nineteen remaining predictor variables for expected goals include:

Predictor Variable	Description
Season	Label representing the season that each record corresponds to (numbered 1 through 4)
Team formation	Formation used to start the game
Possession	Percentage of the game that a team has control of the ball
Passing accuracy	Number of passes completed divided by total number of passes
Short length passing accuracy	Accuracy for passes between 5 and 15 yards
Medium length passing accuracy	Accuracy for passes between 15 and 30 yards
Long length passing accuracy	Accuracy for passes greater than 30 yards
Dribble success rate	Success rate of taking on a defender while dribbling
Percentage of shots on target	Percentage of shots taken that would be a goal, if no goalkeeper were to be present
Opponent fouls	Number of fouls committed by the opposing team
Offsides	Number of times a team gets penalized for having an attacking player behind the last defender
Recoveries	An action that ends possession for the opponent and begins possession for the other team
Touches in the attacking third	Touches in the attacking team's final third of the field
Touches in the attacking penalty area	Touches in the attacking team's penalty box
Tackle win percentage	Proportion of tackles where the tackler's team won possession
Crosses	Medium-long range pass angled toward the center of the field near the goal
Corner kicks	Place kick taken by the attacking team, from the corner nearest the goal
Interceptions	Defensive player intercepts the ball from its intended target
Home/away status	Team's status of being home or away in a match

Table 1. Table of Predictor Variables and Description

In the raw dataset, there existed several variables that provided highly repetitive information, and required consideration for removal. A further discussion of those several variables is included in the following section, which discusses pre-processing steps taken. There were a multitude of pre-processing steps that were necessary prior to beginning analysis of the raw dataset.

Data Pre-Processing

Removal of Redundant Variables

Several of the predictor variables in the raw dataset are similar in nature and required removal due to their repetitive nature. One variable that was removed was passing accuracy. There was a total of four distinct variables that related to passing accuracy. Three of these four variables included short length, medium length, and long length passing accuracy. Short length passing accuracy is the accuracy of passes that are between 5 and 15 yards, whereas medium length is the success rate of passes between 15 and 30 yards, and long length pass completion is for passes that are greater than 30 yards. Because a single pass cannot be represented by multiple of these three pass length types, these three variables were all justifiably independent from one another. However, because the passing accuracy variable is a cumulative average of all a team's passing success (short, medium, and long) for a given match, it was best to consider removing the passing accuracy variable.

Additionally, the raw dataset included two variables relating to number of touches: touches in the opponent's final third of the field, and touches in the opponent's penalty area. In further thinking, it became clear that these variables were inherently correlated. If the attacking team happened to make a touch in the opponent's penalty box, there would have been a very strong likelihood that they had already had one or more touches in the opponent's final third of the field. Apart from long crosses or through balls, there are a limited number of ways where the team could ultimately attain a touch in the penalty area without already going through the attacking third of the field. Given this, it was de-

cided to remove the predictor variable of number of touches in the opponent's final third, and to keep touches in the penalty area.

Conversion to Factor Variables

There were several qualitative variables in the raw dataset, that required conversion to factor variables, prior to beginning modeling and analysis. These variables included season, formation, and home/away status. Each season was converted to be a number, labeled 1 through 4, to numerically represent each of the four different league seasons. Similarly, over the course of the four total seasons of data, Arsenal utilized a total of 11 total formations, that were each given a label. Finally, the home or away status variable had two possible values (home or away), and additionally required factor conversion. Upon the completion of this process, these three predictor variables were sufficient to be included in modeling.

Bivariate Vizualization and Analysis of Data

Below are scatterplots for each quantitative variable in the dataset, versus the response variable, expected goals (xG), as well as a table showing the correlation (r) value for each predictor variable versus the response variable. These figures (figure 1–4) and table 2 show the relationship between each numeric predictor variable and the response variable, giving us insight into how variables may or may not be correlated to expected goal values.

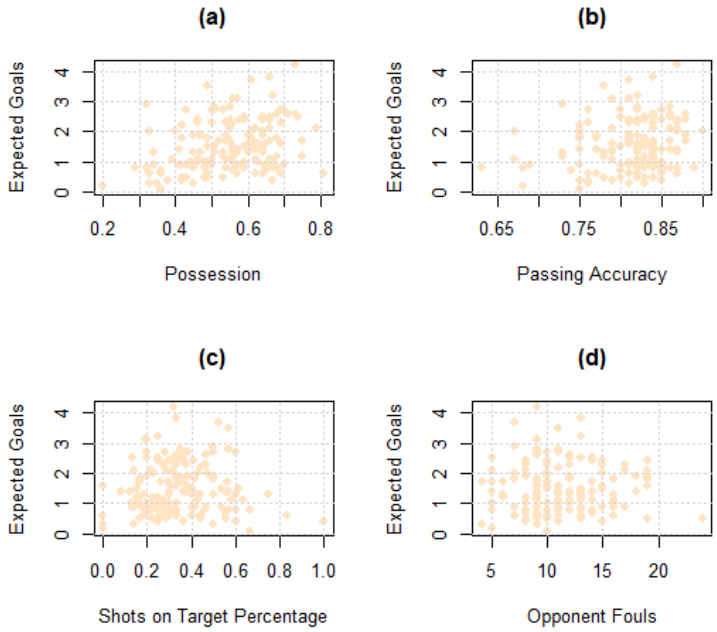


Figure 1. Scatterplots of Possession (a), Passing Accuracy (b), Percentage of Shots on Target (c), and Opponent Fouls (d) vs Expected Goals

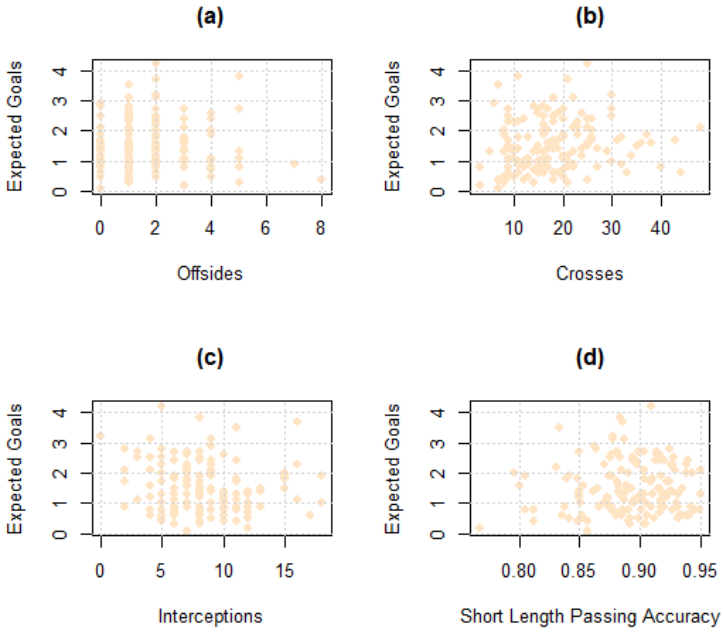


Figure 2. Scatterplots of Offsides (a), Crosses (b), Interceptions (c), and Short Length Passing Accuracy (d) vs Expected Goals

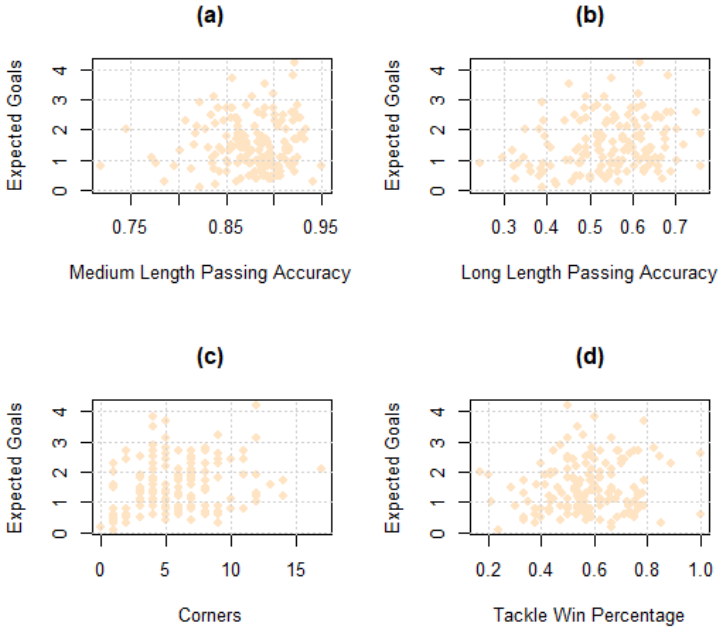


Figure 3. Scatterplots of Medium Length Passing Accuracy (a), Long Length Passing Accuracy (b), Corners (c), and Tackle Win Percentage (d) vs Expected Goals

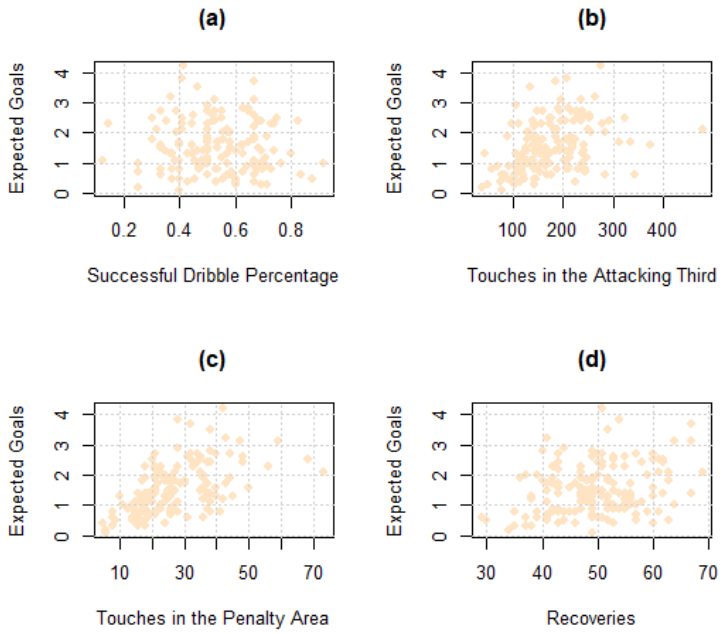


Figure 4. Scatterplots of Successful Dribble Percentage (a), Touches in the Attacking Third (b), Touches in the Penalty Area (c), and Recoveries (d) vs Expected Goals

Predictor Variable	Product-moment Correlation Coefficient (*r*)
Possession	0.31
Passing Accuracy	0.16
Percentage of Shots on Target	0.02
Opponent Fouls	-0.02
Offsides	-0.01
Crosses	0.15
Interceptions	-0.11
Short Length Passing Accuracy	0.05
Medium Length Passing Accuracy	0.06
Long Length Passing Accuracy	0.22
Corners	0.23
Percentage of Tackles Won	0.12
Percentage of Successful Dribbles	-0.06
Touches in the Attacking Third	0.40
Touches in the Opponent's Penalty Box	0.59
Recoveries	0.27

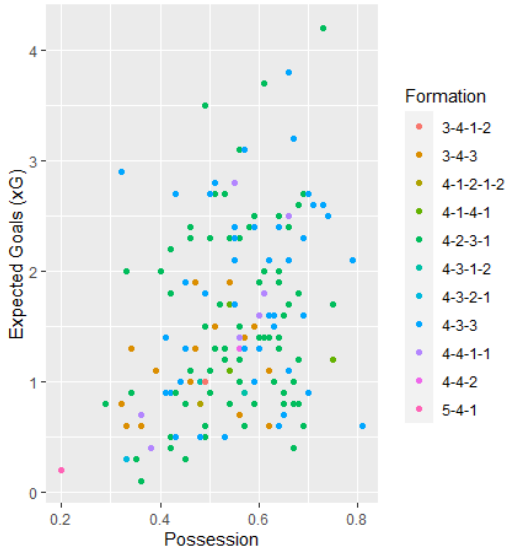
Table 2. Table of Pearsonian Product-moment Correlation Coefficient (r) for Predictor Variable vs Expected Goals (xG)

As seen in the scatterplots and table above, there are several variables that have some linear correlation with expected goals. Possession, long length passing accuracy, corners, touches in the attacking third and penalty box, and recoveries are all variables that had some correlation to the response variable. These correlations were studied in greater detail in the Results and Analysis section of this study, discussing multicollinearity issues in the dataset.

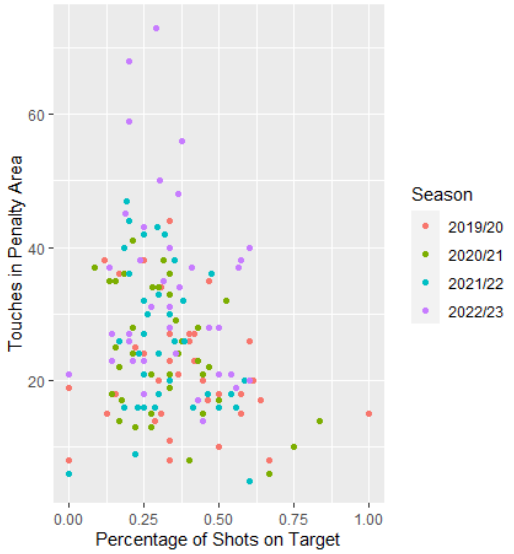
Additional Data Visualizations

From analysis of the scatterplots, it is seen that several variables, including possession, passing accuracy, percentage of shots on target, crosses, touches in the attacking third and penalty box, and recoveries were all correlated with expected

goals. To explore further graphical trends among these variables and several of the factor variables, several additional plots (figure 5–6) were created to highlight other interesting findings in the pre-processing stage. These plots, as well as a brief discussion for the findings from each, are given below.

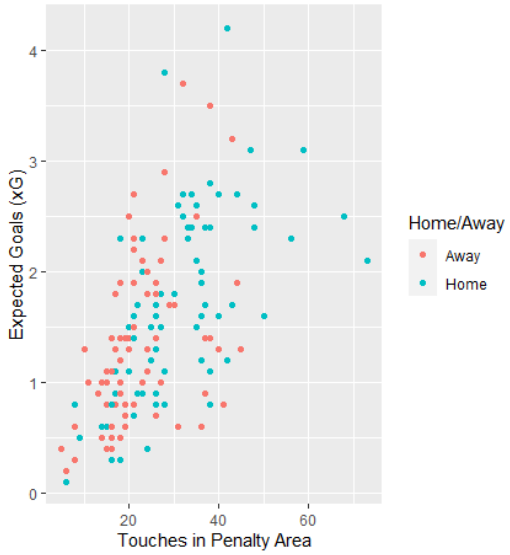


(a) Scatterplot of Possession vs xG, by Formation

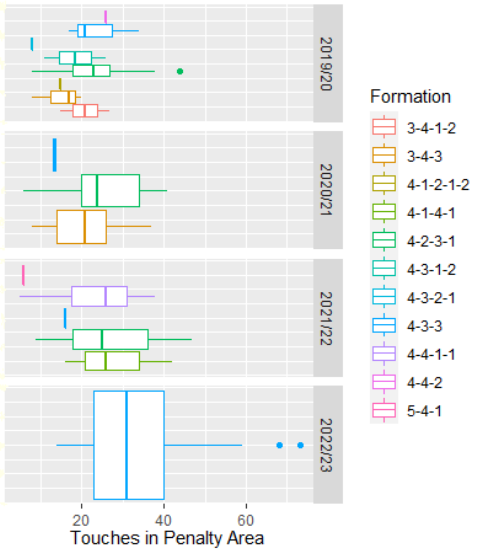


(b) Scatterplot of Shots on Target Percentage vs Touches in the Penalty Area, by Season

Figure 5. Plots Analyzing Possession, Shots on Target Percentage, and Touches in the Penalty Area



(a) Scatterplot of Touches in the Penalty Area vs xG, by Location



(b) Boxplot of Touches in the Penalty Area, by Formation and Season

Figure 6. Plots Analyzing Touches in the Penalty Area

In the first plot, **Figure 5 (a)**, it is clear that team possession was much higher for particular formations, including 4-3-3, 4-2-3-1, and 4-4-1-1. This implies that Arsenal achieved greater success in holding onto the ball, but does not yet confirm if possession is significant in predicting expected goals. The second plot, **Figure 5 (b)**, shows that there is no clear relationship between percentage of shots on target and number of touches in the penalty area. Additionally, percentage of shots on target generally varies by season, whereas touches in the penalty area is clearly higher in the fourth season. Knowing that Arsenal had performed the best in the most recent (fourth) season, it may have been expected that the team would have a higher percentage of shots on target, and also a higher number of touches in the penalty area. It was surprising to see that percentage of shots on target did not vary much throughout different seasons.

The next two plots focus more specifically on the relationship of touches in the penalty area and expected goals. Touches in the penalty area was highly correlated with expected goals based on its scatterplot, and it was of interest to explore further. First, in **Figure 6 (a)**, it can be seen that there tends to be a much greater number of total touches in the penalty area while playing at home. This result would be expected, as home field advantage does generally exist in most sports, but it was interesting to see just how much more pressure Arsenal was placing on their opponents while playing in North London. Finally, in **Figure 6 (b)**, the total number of touches in the penalty area was explored by season and by formation. It was peculiar to see how the number of formations used throughout the season continually decreased over time. Again, knowing that Arsenal performed well in the most recent season where only one formation (4-3-3) was used, it appears that experimenting with fewer formations may be a factor in improving team performance. This may be because certain formations are more comfortable and a natural fit for the players, and trying a new formation may throw off team chemistry and comfort.

METHODOLOGY

Methodology

This section details the several different regression modeling techniques used throughout this analysis. Due to the failure of several model assumptions, it was necessary to explore more complex techniques than multiple linear regression, to identify which predictor variables are most important in predicting expected goals. Each of these different models are described in the following subsections, including their formulas and purpose.

Multiple Linear Regression (MLR) Model

Prior to beginning a deep analysis, preliminary model assumption checking was required. This process began by creating a baseline multiple linear regression model that incorporated all original nineteen predictor variables. The initial multiple linear regression model was as follows,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \textbf{Equation 1.}$$

where $\mathbf{Y}$ is a $n \times 1$ vector of expected goals, $\mathbf{X}$ is a $n \times p$ design matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of random error components.[22, 23] Here, $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$. Therefore, $\mathbf{Y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$.

Residual Diagnostics and Multicollinearity Analysis

Using this initial multiple linear regression model, the key assumptions for linear regression were tested. The five key assumptions are for linearity, homoscedasticity, independence, normality, and multicollinearity. The first and second assumptions lie in checking the residual plot. The residual plot must show a random distribution of residuals, and also show a distinct horizontal band shape, which suggests that the variances of the errors are equal. The next model assumption that required checking was for a linear Q-Q plot, which determines how normal the initial dataset is. Finally, the remaining two plots focus on the independence of assumptions.[24]

Additionally, the initial multiple linear regression model is used to identify the Generalized Variance Inflation Factor (GVIF) values for each predictor variable. A GVIF value is a generalized version of the Variance Inflation Factor (VIF)

value. A VIF represents the amount of variance that is inflated, for each individual predictor variable in a multiple linear regression model. These coefficients have larger values when multicollinearity exists in the data. A VIF value of 1 means that the predictor variable in question has no correlation to any other predictor in the model. In contrast, a VIF value greater than 10 implies that the variable exhibits signs of multicollinearity and a correlation to at least one other predictor variable, necessitating correction.[25] A GVIF, in contrast, additionally accounts for variables that have greater than 1 degree of freedom, such as polynomials, or categorical variables with more than 2 levels. GVIF applies the appropriate change needed to properly analyze if a particular variable has issues relating to multicollinearity.[26]

Variance Stabilizing Transformation

Transforming the response, and/or the predictor variables, can play a great impact by improving model fit, and to achieve a standard closer to normality. To identify the optimal transformation (or lack thereof), the Box-Cox transformation method is often used. The Box-Cox method is a commonly used statistical technique that specifically changes the response variable, to resemble a normal distribution more accurately. This process is accomplished using a λ value, which represents the most optimal exponent to apply to the response variable data. This λ value can range from -2 to 2, with possible values including -2, -1, -0.5, 0, 0.5, 1, or 2; each value is associated with a transformation, including inverses, square roots, natural logs, or exponents. Additionally, a λ value of 1 results in a conclusion that no transformation is the best fit for the data.[27] The Box-Cox method is critical in determining if a transformation of the response variable is the best option to improve overall model performance. This decision and respective transformation must be made prior to modeling. The results of the Box-Cox test will be discussed in the Analysis section of this paper.

Likelihood Ratio Test

The likelihood ratio test is a form of hypothesis testing that enables one to choose the best model between two possible options, based on the ratio of their likelihoods.[28] The two models being compared are usually split, with one model being simple, and the other being complex. In this study, the simple model was the linear regression model without passing accuracy and touches in the attacking third (as both were found to have natural correlation to other predictors), and the complex model was the same model, but with the applied square root transformation, as was found optimal in the Box-Cox procedure. The formula for the likelihood ratio test represents the ratio between the log likelihood (L) of the simpler model and the more complex model. The result of this equation yields a chi-square value to use for testing the null hypothesis that the simpler model is the best fit, versus the alternative that the complex model is a better fit. The degrees of freedom for the test is equivalent to the difference in number of total parameters for the two individual models.[29] The likelihood ratio test procedure is important in determining if model complexities, such as additional variables and transformations are justifiable to implement. The results of the likelihood ratio test will be discussed further in the Results and Analysis section of the paper.

R^2 and AIC

The previously mentioned statistical and machine learning methods are all potentially applicable to the dataset for Arsenal Football Club, which has a sparse sample size and a large number of predictor variables. Machine learning is critical in finding conclusions from a wealth of data. Each of these methods (Ridge, Lasso, Elastic Net, and Group Lasso) create a model, and this study analyzes and compares the effectiveness of each model in predicting a team's total expected goals, by using adjusted R^2, Akaike Information Criterion (AIC), and other relevant criterion values. The adjusted R^2 value is different from the traditional R^2 value, as it more properly adjusts for a higher number of predictor variables in the model.[24] The adjusted R^2 value adequately penalizes for the number of predictor variables present, as they can implement bias. This method involves dividing the sum of the squares of the residuals and the totals, by their respective degrees of freedom, where n is the total number of observations in the dataset, and k represents the number of estimated parameters in the model, which includes a total of p variables and the intercept, β_0.

Another common statistic used to determine overall model success is AIC. AIC represents how well a model fits the data, for future prediction. Lower AIC values represent better model fit and are considered in conjunction with ad-

justed R^2 to determine which model is the most appropriate for analysis, as well as the appropriate variables to keep in the model,[30] ultimately to determine the goodness of fit of the model in representing the dataset.[31]

All these procedures, as well as pre-processing and analysis steps, have been completed in using the statistical programming language R, which has sufficiently allowed for the possibility of making conclusions from this study. The AIC function in R Studio is not compatible with certain advanced regression models, including those that are generalized linear models (GLM), such as Ridge, Lasso, Elastic Net and Group Lasso. However, AIC is still a highly effective and useful tool to compare initial models created for the dataset.

Variable Selection Methods

Stepwise Regression Methods

Stepwise regression is a procedure that builds a regression model from the given predictor variables and either keeps or removes predictors until they are no longer significant enough to consider.[32] There are two unique types of stepwise regression: forward and backward. A forward stepwise selection involves building a greedy algorithm, with a model beginning with only an intercept (no predictor variables), and adding predictors until they are no longer necessary to predict for expected goals. The model output ultimately performs dimension reduction, by never adding the variables that do not play a significant enough role into the model. In contrast, backward stepwise regression considers a complete model with all predictor variables included and eliminates non-significant predictors until a final model is chosen to represent the data, only including predictor variables deemed significant enough.[33] For both types of stepwise regression, each variable is evaluated against a particular criteria, most commonly Akaike's Information Criterion (AIC).[30]

Regularized Regression Methods

Ridge Regression

There are several other prominent machine learning procedures that can effectively handle sparse or correlated features in a dataset, including Ridge regression, Least Absolute Shrinkage and Selection Operator (Lasso) regression, Net Elastic regression, and Group Lasso regression. Ridge regression, also known as $\mathcal{L}_2$ regularization, begins by applying a penalty (the sum of the squared coefficients) to account for correlation in the dataset. This works by shrinking coefficients of correlated predictor variables, which helps achieve smaller variance. This is important in restricting the influence of predictors.[34] When determining the penalty, Ridge regression considers all predictor variables in the model, which therefore makes the method most useful in scenarios where there are many predictor variables present, and all have non-zero coefficients.[35–37]

Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) regression (known as $\mathcal{L}_1$ regularization) functions in a similar manner to Ridge regression, but has a key difference in the penalty term. Lasso methods also apply a penalty term, that is instead based on the magnitude of coefficients, rather than sum of squared coefficients. In contrast to Ridge regression, Lasso does not weight the coefficients of every predictor variable, but rather often picks one of the correlated predictor variables and ignores the rest, effectively performing variable selection.[35] It can remove predictors from the model, by shrinking coefficients to exactly zero, unlike Ridge.[38] Again, it is common to use machine learning techniques including Lasso when dealing with larger datasets and correlated variables.

It is common to compare the results of Lasso regression with those of Ridge, to determine which model produces superior results with minimized error, to reduce or eliminate the impact of correlated variables in the model.

Elastic Net Regression

Elastic Net regression is another popular machine learning technique, effectively combining elements of both Ridge and Lasso regression, as another way to perform variable selection and shrink predictor variable coefficients.[39] This method removes groupings of correlated variables, while also keeping several predictors in the model. A series of cross-validation steps are used to determine how much of the penalty is determined from Ridge regression, and how much is determined by Lasso regression. Elastic Net regression is traditionally viewed as a balance of both Ridge and Lasso regression, shrinking the coefficients, possibly to 0, but not as commonly as Lasso regression. Elastic Net regression is especially noteworthy for its ability to handle bias. With multiple variables that are highly correlated, it is possible for Lasso regression to introduce bias,[34] and therefore it is worthwhile to compare the performance of Elastic Net regression to both Ridge and Lasso.

Group Lasso Regression

Group Lasso regression follows a similar variable selection process as the Lasso method, but differs in its choice of important predictors. Lasso chooses several significant features among all the variables, whereas Group Lasso involves the splitting of variables into groups to complete variable reduction. Upon splitting the data into groups, if a variable is deemed important in a group, the Group Lasso procedure will include all variables in that same group in its model. If a group produces no important predictors, all its variables will be excluded from the model.[40–42]

Measures of Prediction Accuracy: MSE, RMSE, and MAPE

In addition to analyzing adjusted R^2 and AIC, models require the calculation of error values when determining overall model adequacy. There are three main measures that are routinely used, including mean squared error (MSE), root mean square error (RMSE), and mean absolute percentage error (MAPE).

MSE squares the distance of data points from the regression line (errors), to ultimately find the average of the set of errors. This procedure is popular, due to its squaring technique, which gives more weight to outlier data points. RMSE calculates the square root of the MSE, which effectively makes the larger values generated by outlier data points smaller, and therefore easier to interpret.[43] Finally, MAPE is a method to measure the accuracy of a predictive model. It is desirable to minimize the MSE, RMSE, and MAPE values to increase model success.[24] Each of these error calculations are important to consider, as they are critical in determining the optimal model with the best prediction accuracy. The formulas for each of these error values are listed below:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \textbf{Equation 2.}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad \textbf{Equation 3.}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \qquad \textbf{Equation 4.}$$

where n is the number of data points in the dataset, Y_i is the actual value of the response variable (square root of expected goals) for a particular data point, and $\hat{Y}_i$ is the predicted value of the response variable for a data point.

Bootstrap Method for Regression Models

In the analysis of these models, the creation of prediction intervals is hugely important to understanding the range that the estimated value of square root expected goals falls within, and the accuracy of model prediction. The creation of prediction intervals for each individual record in the test dataset requires bootstrapping, a procedure that resamples data from one sample in order to randomly generate a distribution. This procedure can be used to estimate standard errors, bias, obtain prediction intervals, and also to test a hypothesis.[44] The process of bootstrapping involves sampling the records in the test dataset (in this study, 38 total records) a selected number of bootstrap resamples, $B = 1000$. Often, B is chosen as a high number, to ensure better sample size in creating a prediction interval, from the enlarged sampling distribution.[45] The advanced regularized regression methods described above, including Ridge, Lasso, Elastic Net, and Group Lasso, and their functions and required packages in R, do not provide prediction intervals for individual records of the test dataset. Therefore, to generate these intervals, it became necessary to increase the sample size through a bootstrapping resampling technique.

To account for these issues, an alternative method was required to create a prediction interval using bootstrap resampling method. Each test record had a unique predicted square root expected goal value, $\hat{y}$, calculated by multiplying the $\mathbf{X}$ matrix with the $\vec{\beta}$ vector. $\mathbf{X}$ contained all values of the predictor variables for each record, which was then multiplied with $\vec{\beta}$, the vector storing the coefficients for each predictor variable for each model. In total, there existed 38 unique fitted $\hat{y}$ values, stored in $\vec{\hat{y}}$, the vector of all fitted values, for each model. After collecting these 38 fitted values, they were then bootstrapped, with 1000 resamples of size 38 being generated from the original 38 data vectors. Therefore, we have calculated 1000 predicted values by averaging out 38 predicted values from each of the bootstrap resamples. After this step, a prediction interval of desired confidence level $(1 - \alpha)$ for the expected number of goals was then able to be created for each of 1000 records. This process is unique from that of traditional bootstrapping, where random sampling is usually the first step of the procedure. However, this study first required collection of all fitted values, so that enough records were present to then randomly sample and create a distribution. The results of these bootstrapping procedures are discussed in the following subsection.

RESULTS

The following section describes the results of the aforementioned modeling procedures, including model assumptions, data transformation, fitted models, selection of most important predictor variables and of the best model to use, to predict expected goals along with prediction intervals.

Model Assumption and Multicollinearity Checking

As discussed earlier, inspection of preliminary model assumptions is required before using the model for inferential procedure. Upon fitting this multiple linear regression model of expected goals (xG) on the selected predictors, each key assumption referenced in Methodology section was checked. Four of the five assumptions of linear regression can be checked using the first three plots provided in figure 7, below, including (a) residual plot of the raw dataset versus the model's fitted values, (b) Q-Q plot, (c) scale-location plot, and (d) residual versus leverage plot.

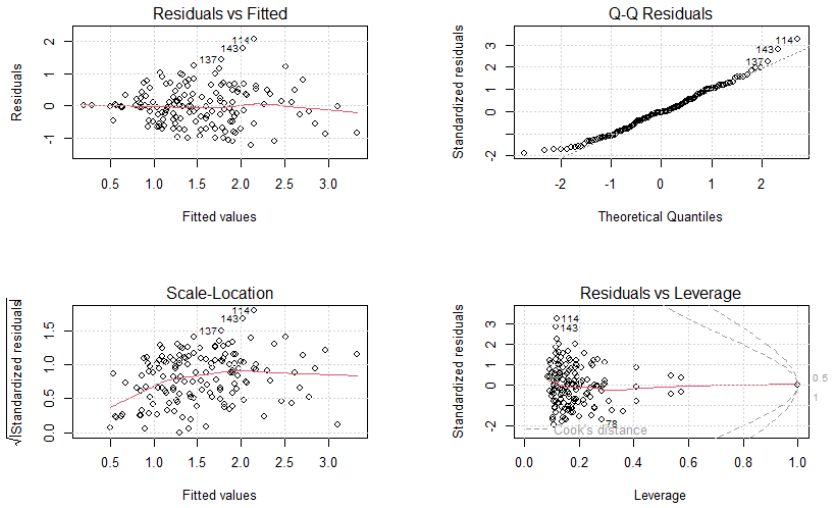


Figure 7. Plots for Simple Linear Regression Model Assumption Checking, Including Residual vs Fitted Values (a), Normal Q-Q (b), Scale-Location (c), and Residuals vs Leverage (d)

Since figure 7 (a) is not showing any specific direction or trend and the red line is very much horizontal, so, linearity and independence of observations assumptions seem reasonable. From figure 7 (b), the points are very close to the dotted straight line except a few points, indicating the normality assumption of error is reasonable. Finally, figure 7 (c) produces the red line as a curve and it has some trend of increasing error variance. Thus, a variance stabilizing transformation of the response variable is necessary for drawing inferential decisions. The multiple linear regression model is as follows:

$$\hat{xG} = 1.164 + 0.149(Season\ 2) + 0.215(Season\ 3) + 0.010(Season\ 4) - 0.424(Formation\ 3-4-3) - 0.965(Formation\ 4-1-2-1-2) - 0.352(Formation\ 4-1-4-1) - 0.277(Formation\ 4-2-3-1) - 0.258(Formation\ 4-3-1-2) - 0.378(Formation\ 4-3-2-1) - 0.220(Formation\ 4-3-3) - 0.263(Formation\ 4-4-1-1) + 0.007(Formation\ 4-4-2) - 0.786(Formation\ 5-4-1) + 1.017(Possession) + 0.874(Passing\ Accuracy) + 0.661(Percentage\ of\ Shots\ on\ Target) - 0.005(Opponent\ Fouls) + 0.048(Offsides) - 0.029(Crosses)^{*} + 0.024(Interceptions) - 1.167(Short\ Length\ Passing\ Accuracy) - 2.099(Medium\ Length\ Passing\ Accuracy) + 0.535(Long\ Length\ Passing\ Accuracy) - 0.011(Corners) + 0.649(Tackle\ Win\ Percentage) + 0.310(Successful\ Dribble\ Percentage) - 0.001(Touches\ in\ the\ Attacking\ Third) + 0.055(Touches\ in\ the\ Penalty\ Area)^{***} + 0.003(Recoveries) - 0.042(Home/Away\ H)$$

Additionally, GVIF values were also analyzed for assessing the last assumption, multicollinearity. Below is a table (table 3) documenting each predictor variable's respective GVIF values and degrees of freedom, in the multiple linear regression model. The predictor variables with GVIF values greater than 10 are considered muticollinear in practice.[24] However, it was plausible to leave all variables in the original multiple linear regression model, knowing that the machine learning algorithms used in modeling would perform variable reduction, and remove these problematic predictor variables at a later time.

Predictor Variable	GVIF	Degrees of Freedom
Season	17.528462	3
Formation	39.744686	10
Possession	5.009569	1
Passing Accuracy	19.563667	1
Percentage of Shots on Target	1.479036	1
Opponent Fouls	1.317865	1
Offsides	1.319266	1
Crosses	4.389041	1
Interceptions	1.607953	1
Short Length Passing Accuracy	4.742041	1
Medium Length Passing Accuracy	6.222260	1
Long Length Passing Accuracy	3.069269	1
Corners	2.914998	1
Percentage of Tackles Won	1.218060	1
Percentage of Successful Dribbles	1.261656	1
Touches in the Attacking Third	7.410396	1
Touches in the Opponent's Penalty Box	3.71142	1
Recoveries	1.601722	1
Home or Away Status	1.379950	1

Table 3. Table of GVIF Values and Degrees of Freedom for Each Predictor Variable

Variance Stabilizing Transformation of Response Variance

To account for the violations of the model assumptions described above, it was necessary to explore possible transformations of the response variable in the dataset, expected goals (xG). As mentioned in the "Variance Stabilizing Transformation" subsection in the Methodology section, the Box-Cox method is a common technique to use to identify the type of data transformation when either of the linearity, normality, and constant variance assumptions about the fitted model is questionable. The Box-Cox procedure allows us to produce a plot of transformation parameter (λ) on horizontal axis and value of likelihood function of the model associated with λ value on vertical axis. The plot for the original multiple linear regression model is shown below, where the confidence interval bounds of the optimal λ value intercept 0.5, but do not intercept values of $\lambda = 0$, nor $\lambda = 1$.

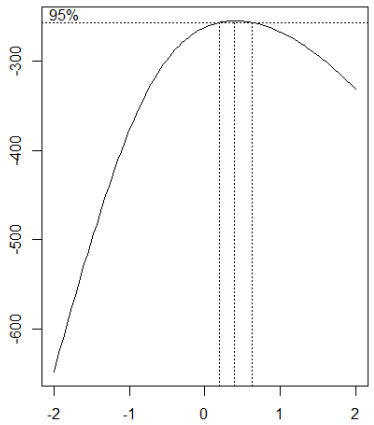


Figure 8. Box-Cox Plot to Identify Type of Variable Transformation

This plot implies that a λ value of 0.5 is the best fit, and therefore the response variable requires a square root transformation to improve normality. Going forward, future modeling would be completed using a transformed response variable: the square root of the expected goal value. The following multiple linear regression model, with response variable transformation, is shown below:

$$\hat{xG_t} = 0.836 + 0.043(Season\ 2) + 0.066(Season\ 3) - 0.003(Season\ 4) - 0.158(Formation\ 3-4-3) - 0.404(Formation\ 4-1-2-1-2) - 0.099(Formation\ 4-1-4-1) - 0.118(Formation\ 4-2-3-1) - 0.093(Formation\ 4-3-1-2) - 0.315(Formation\ 4-3-2-1) - 0.097(Formation\ 4-3-3) - 0.082(Formation\ 4-4-1-1) + 0.048(Formation\ 4-4-2) - 0.507(Formation\ 5-4-1) + 0.446(Possession) - 0.043(Passing\ Accuracy) + 0.218(Percentage\ of\ Shots\ on\ Target) - 0.001(Opponent\ Fouls) + 0.019(Offsides) - 0.011(Crosses)^{*} + 0.013(Interceptions) - 0.085(Short\ Length\ Passing\ Accuracy) - 0.706(Medium\ Length\ Passing\ Accuracy) + 0.246(Long\ Length\ Passing\ Accuracy) - 0.005(Corners) + 0.267(Tackle\ Win\ Percentage) + 0.124(Successful\ Dribble\ Percentage) - 0.001(Touches\ in\ the\ Attacking\ Third) + 0.023(Touches\ in\ the\ Penalty\ Area)^{***} + 0.001(Recoveries) - 0.037(Home/Away\ H)$$

As was done in the previous section, each of the five assumptions must be checked, following this transformation of the response variable. It can be seen that, while there is not perfect normality, there also is no distinct funnel shape, indicating that the overall fit of the model on square-root transformed response (xG) might be more reasonable than the model with the original xG. The same four plots are shown below, for the purpose of comparison. These plots are providing indication of reasonable model assumptions. Using this transformed response variable will be the basis of subsequent analysis in the coming sections.

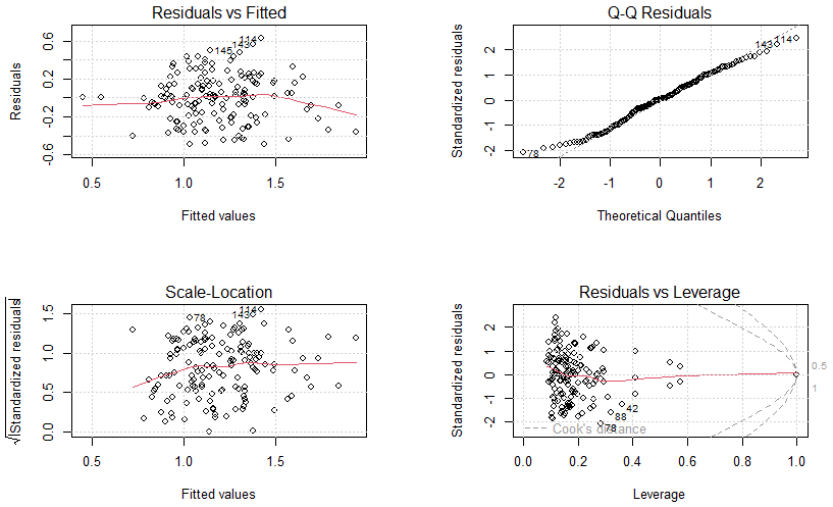


Figure 9. Plots for Model Assumption Checking with the Transformed Response Variable, Including Residual vs Fitted Values (a), Normal Q-Q (b), Scale-Location (c), and Residuals vs Leverage (d)

Results of the Likelihood Ratio Test

The likelihood ratio test is a key step to determine the validity of adding complexities in terms of the predictor variables to the original regression model. The test was specifically used in this study to confirm the results of the Box-Cox transformation, and to suggest that the more complex, transformed model would be a better fit for the data. Using the `lrtest` function in R, through the `lmtest package`, a likelihood ratio test was completed to confirm the use of a square root transformation. This test compared two models, including the "simple" model with all variables–except for passing accuracy and touches in the attacking third, as both were found to be redundant variables–versus the same model but with the square root transformation applied to the response variable, expected goals. This yielded a chi-square test statistic of 272.25, and a p-value of 2.2×10^{-16}. This p-value is practically 0 which is less than a designated α value of 0.05, which therefore suggests rejecting the null hypothesis, meaning that the two models are significantly different from one another. This conclusion means that the square root transformation is satisfactory and justifiable, in order to improve overall model adequacy. These results were important steps to enhance the validity of the subsequent

modeling that has been performed and that will be documented below, which uses the same square root transformation for the response variable.

Fitted Models Under Different Regression Methods

This subsection details results for each of the previously mentioned regression modeling techniques, and provides the fitted model, adjusted R^2, most important predictor variables chosen, and if applicable, the test statistic and p-value from each model. For the regularization methods, the glmnet() function used to create the models (Ridge, Lasso, Elastic Net, and Group Lasso) in R does not generate p-values or confidence intervals for each predictor variable. This limitation required the use of error values and adjusted R^2 to compare overall model performance.

Stepwise Regression Methods

Backward Stepwise Regression We have fitted the model using backward step-wise regression method. The outputs include each selected variable's estimated parameter (β), standard error, test statistic value (t), and p-value, as well as the overall model's residual standard error, degrees of freedom, multiple and adjusted R^2 values, F statistic, and p-value.

The fitted model for the backward stepwise regression model for the transformed response variable are as follows,

$$\hat{xG_t} = 0.412^{**} + 0.280(Percentage\ of\ Shots\ on\ Target)^{*} - 0.009(Crosses)^{**} + 0.021(Touches\ in\ the\ Penalty\ Area)^{***} + 0.006(Recoveries)^{*}$$

F = 28.17

Regression Degrees of Freedom = 4

Residual Degrees of Freedom = 147

P-value = 2.2e-16

Adjusted R^2 = 0.419

AIC = 30.98115

where $\hat{xG_t}$ represents the predicted (square root) value for expected goals. The stars (*) attached to each coefficient represents the significance at different nominal α levels. One star represents that the parameter is significant at the α = 0.05 level, two stars mean that the parameter is significant at the α = 0.01 level, and three stars mean that the parameter is significant at the α = 0.001 level.

The fitted model produced an adjusted R^2 value of 0.4185, meaning that approximately 41.85% of the total variation in the square root of expected goals is explained by these four predictor variables. Considering the fitted model, all four predictor variables selected through this machine learning technique–percentage of shots on target, crosses, touches in the penalty area, and recoveries–are all individually significant in predicting expected goal values. Specifically, percentage of shots on target, recoveries, and touches in the penalty area had a positive impact on a team's predicted total xG in a match, whereas crosses had a negative impact on predicting a team's total xG. This will be contrasted with the variables deemed significant from each of the subsequent models.

Further, the overall model achieved an F statistic value of 28.17, with a regression degree of freedom equal to four, and a residual degree of freedom equal to 147. This test statistic value resulted in a p-value of 2.2×10^{-16}, implying that, in addition to each individual predictor variable, that the overall model is also significant in predicted expected goals. Finally, the backward stepwise regression model generated an AIC value of 30.98115, which will be used to compare to the model generated by forward stepwise regression.

Forward Stepwise Regression The fitted model for the forward stepwise regression model along with the necessary outputs for the transformed response variable was as follows,

$$\hat{xG_t} = 0.412^{**} + 0.280(Percentage\ of\ Shots\ on\ Target)^{*} - 0.009(Crosses)^{**} + 0.021(Touches\ in\ the\ Penalty\ Area)^{***} + 0.006(Recoveries)^{*}$$

F = 28.17

Regression Degrees of Freedom = 4

Residual Degrees of Freedom = 147

P-value = 2.2e-16

Adjusted R^2 = 0.419

AIC = 30.98115

where $\hat{xG_t}$ again represents the predicted (square root) value for expected goals. This fitted model, the corresponding significance levels, and other measures of model success are the exact same as the values found from completing backward stepwise regression. The adjusted R^2 value represents that 41.85% of the total variation in the square root value of expected goals is explained by this forward stepwise regression model, containing these four variables (percentage of shots on target, crosses, touches in the penalty area, and recoveries), which were again found to be individually significant at the α equals 0.05. Further, the overall fitted model was found to be significant, with the same F statistic of 28.17, and a p-value of 2.2×10^{-16}. Again, each of these four predictor variables, in addition to the overall model, were significant in predicting the square root transformed value of expected goals, with percentage of shots on target, recoveries, and touches in the penalty area having a positive impact on xG, and crosses having a negative impact. The AIC of this model was also equivalent to the backward stepwise model, with a value of 30.98.

Regularized Regression Methods

Ridge Regression Model Ridge regression keeps all predictor variables in the fitted model (including those that were deemed to have a natural correlation to other predictors), but appropriately penalizes the less important predictor variables. The fitted model is as follows,

$$\hat{xG_t} = 0.512 + 0.031(Season)0.002(Formation) + 0.333(Possession) + 0.077(Percentage\ of\ Shots\ on\ Target) + 0.002(Opponent\ Fouls) + 0.002(Offsides)0.004(Crosses) + 0.005(Interceptions) + 0.004(Corners)0.136(Short\ Length\ Passing\ Accuracy)0.234(Medium\ Length\ Passing\ Accuracy) + 0.192(Long\ Length\ Passing\ Accuracy)0.003(Percentage\ of\ Tackles\ Won) - 0.005(Percentage\ of\ Successful\ Dribbles) + 0.012(Touches\ in\ the\ Penalty\ Area) + 0.006(Recoveries)0.001(Home\ or\ Away\ Status)$$

Adjusted R^2 = 0.694

where $\hat{xG_t}$ represents the predicted square root value for expected goals. The coefficients in the fitted model have been penalized where appropriate, so that the variables that are most important in predicting expected goals, including the four variables discussed earlier (percentage of shots on target, crosses, touches in the penalty area, and recoveries), all have a higher weight and importance in calculating the square root value of expected goals.

It is important to note that the regularization methods (ridge, lasso, elastic net, and group lasso) do not use an F test statistic, degrees of freedom, nor AIC values, unlike both backward and forward stepwise selection processes, due to the differing methodologies previously described. However, adjusted R^2 is used across all models to determine overall model success. The ridge regression model achieved an adjusted R^2 value of 0.694, which is a step above the results

achieved by the forward and backward stepwise regression models. Approximately 69.4% of the total variation in the square root of expected goals is explained by this model, which is a quite high percentage, indicating that the model fits the data well.

Lasso Regression Model Lasso regression removes predictor variables that were deemed to not be important in the overall success of the model, by penalizing them and reducing their coefficient to 0. This change is shown in the following fitted model for predicting the square root of expected goals, which contains only several of the nineteen total predictors.

$$\hat{xG}_t = 0.699 + 0.015(Touches\ in\ the\ Penalty\ Area) + 0.002(Recoveries)$$
$$\text{Adjusted } R^2 = 0.693$$

This model indicates that Lasso regression only found two predictor variables to be sufficient for predicting the square root expected goal values. These two variables are touches in the penalty area and recoveries–both of which have been deemed significant in other modeling techniques, such as forward and backward stepwise regression. Both predictors positively impact a team's total expected goal count in a match. The Lasso regression model had an adjusted R^2 value of 0.693, which is just below the adjusted R^2 of the ridge model (0.694).

Elastic Net Regression Model The net elastic regression model serves as a compromise of the ridge and lasso models, with the ability to penalize variables to a coefficient value of 0, but it may keep the coefficient of enough important predictor variables. The following fitted model shows the coefficients generated through elastic net regression.

$$\hat{xG}_t = 0.702 + 0.013(Touches\ in\ the\ Penalty\ Area) + 0.002(Recoveries)$$
$$\text{Adjusted } R^2 = 0.696$$

While this fitted model only kept the same two predictor variables as found in lasso regression, it is important to note the slightly different coefficient values, indicating that each variable was penalized and weighted differently through this method. However, these two variables were again determined to be the most important in predicting the square root value of expected goals, both positively impacting the count of expected goals. The elastic net model had a higher adjusted R^2, informing that the model and its predictor variables chosen account for a greater proportion of the total variation in the transformed response variable.

Group Lasso Regression Model The group lasso model was the final model chosen to represent the dataset, splitting predictors into groups. In R, group lasso is performed using the same glmnet() function, but instead using adding the multinomial parameter to the code, indicating a "grouped" type, instead of the default (no grouping). Similar to the lasso regression model, the group lasso model can penalize insignificant predictor variables to the value of 0, keeping only the predictors that are significant enough. The fitted model below shows the coefficients generated through group lasso regression.

$$\hat{xG}_t = 0.699 + 0.015(Touches\ in\ the\ Penalty\ Area) + 0.002(Recoveries)$$
$$\text{Adjusted } R^2 = 0.693$$

The fitted model has again selected touches in the penalty area and recoveries to be two particularly strong predictors to model the square root value of expected goals. These coefficient values are similar to those generated from the

lasso and elastic net models. This implies that all the models found touches in the penalty area and recoveries to be especially important in prediction of square root expected goals, each positively impacting the expected goal tally. The adjusted R^2 value generated by this model was 0.693. These results are less successful than those of elastic net regression. The group lasso model had slightly higher error values, indicating slightly worse accuracy, and had a lower adjusted R^2 than the Elastic Net model. This informs us, that while all models performed well, the Elastic Net model was marginally better than the other models employed in analysis.

Results and Accuracy Measures of the Fitted Models

The prediction accuracy measures of each of the six non-transformed regression and machine learning models are shown in the table below. It can be seen that many of the models, especially the regularization methods, produced the best accuracy measures. Out of all, the elastic net model performed the best, as it contained both the highest adjusted R^2 value, and produced the lowest error values.

	MSE	RMSE	MAPE
Backward	0.569	0.7546	44.88%
Forward	0.615	0.7843	47.35%
Ridge	0.596	0.7718	45.57%
Lasso	0.620	0.7877	43.58%
Elastic Net	0.616	0.7846	42.71%
Group Lasso	0.620	0.7877	43.59%

Table 4. Prediction Accuracy Measures by Model

Prediction Intervals

Following modeling procedures, it became necessary to produce prediction intervals for the 38 predicted expected goal values from the test dataset. This process yielded a 95% prediction interval for each of the predicted expected goal values, comparing the predicted value to the lower (2.5%) and upper (97.5%) bounds of the interval. To generate the prediction interval bounds for each of the 38 predicted values, bootstrapping was required because of the relatively small size of the number of predicted values, as well as the inability of these machine learning models to generate their own prediction intervals. Each of the 38 records in the test dataset were randomly bootstrapped a total of 1000 times in the creation of a prediction interval, to test if each predicted expected goal total value fell inside or outside of the 95% interval.

This procedure was repeated for each modeling technique–Ridge, Lasso, Elastic Net, and Group Lasso–on the non-transformed response variable. The bounds of the prediction intervals, as well as the predicted values for the test dataset generated by each model, are shown in the tables (table 5–6) below, with a further discussion taking place afterwards.

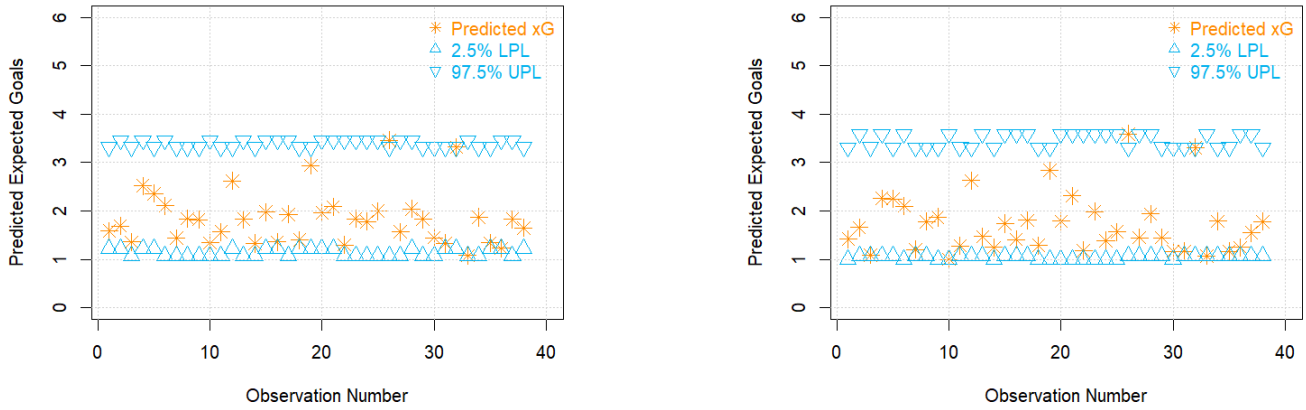
Ridge			Lasso		
Predicted xG	2.5% LPL	97.5% UPL	Predicted xG	2.5% LPL	97.5% UPL
1.59	1.22	3.32	1.42	1.00	3.31
1.69	1.23	3.46	1.66	1.07	3.59
1.36	1.08	3.32	1.09	1.07	3.32
2.52	1.23	3.46	2.26	1.07	3.59
2.36	1.22	3.32	2.25	1.07	3.31
2.11	1.08	3.46	2.10	1.00	3.59
1.44	1.08	3.32	1.20	1.07	3.31
1.84	1.08	3.32	1.78	1.07	3.31
1.82	1.08	3.32	1.87	1.00	3.31
1.34	1.08	3.46	1.00	1.00	3.59
1.57	1.08	3.32	1.28	1.07	3.32
2.62	1.23	3.32	2.63	1.07	3.31
1.83	1.08	3.46	1.47	1.07	3.59
1.33	1.08	3.32	1.25	1.00	3.31
1.98	1.23	3.46	1.74	1.07	3.59
1.36	1.23	3.46	1.41	1.07	3.59
1.93	1.08	3.46	1.81	1.07	3.59
1.40	1.22	3.32	1.29	1.00	3.31
2.94	1.23	3.32	2.84	1.00	3.31
1.96	1.23	3.46	1.79	1.00	3.59
2.09	1.22	3.46	2.31	1.00	3.59
1.30	1.08	3.46	1.18	1.00	3.59
1.84	1.08	3.46	1.99	1.00	3.59
1.78	1.08	3.46	1.39	1.00	3.59
2.00	1.08	3.46	1.57	1.00	3.59
3.46	1.08	3.32	3.59	1.07	3.31
1.57	1.08	3.46	1.45	1.07	3.59
2.04	1.23	3.46	1.94	1.07	3.59
1.83	1.08	3.32	1.45	1.07	3.32
1.45	1.08	3.32	1.17	1.00	3.31
1.33	1.22	3.32	1.17	1.07	3.31
3.32	1.23	3.32	3.31	1.07	3.31
1.08	1.08	3.46	1.07	1.07	3.59
1.87	1.08	3.32	1.79	1.07	3.31
1.34	1.23	3.32	1.17	1.07	3.32
1.23	1.23	3.46	1.25	1.07	3.59
1.83	1.08	3.46	1.55	1.07	3.59
1.65	1.23	3.32	1.78	1.07	3.31

Table 5. Ridge and Lasso Regression to Predict Expected Goal Values and 95% Prediction Intervals, on the Response

Elastic Net			Group Lasso		
Predicted xG	2.5% LPL	97.5% UPL	Predicted xG	2.5% LPL	97.5% UPL
1.42	1.05	3.05	1.42	1.00	3.31
1.61	1.08	3.30	1.66	1.07	3.59
1.13	1.08	3.06	1.09	1.07	3.32
2.14	1.08	3.30	2.26	1.07	3.59
2.12	1.08	3.05	2.25	1.07	3.31
1.98	1.05	3.30	2.10	1.00	3.59
1.22	1.08	3.05	1.20	1.07	3.31
1.72	1.08	3.05	1.78	1.07	3.31
1.78	1.05	3.05	1.87	1.00	3.31
1.05	1.05	3.30	1.00	1.00	3.59
1.30	1.08	3.06	1.00	1.07	3.32
2.44	1.08	3.05	2.63	1.07	3.31
1.50	1.08	3.30	1.47	1.07	3.59
1.25	1.05	3.05	1.25	1.00	3.31
1.72	1.08	3.30	1.75	1.07	3.59
1.39	1.08	3.30	1.41	1.07	3.59
1.74	1.08	3.30	1.81	1.07	3.59
1.29	1.05	3.05	1.29	1.00	3.31
2.67	1.05	3.05	2.84	1.00	3.31
1.74	1.05	3.30	1.79	1.00	3.59
2.16	1.05	3.30	2.31	1.00	3.59
1.19	1.05	3.30	1.18	1.00	3.59
1.87	1.05	3.30	1.99	1.00	3.59
1.40	1.05	3.30	1.39	1.00	3.59
1.54	1.05	3.30	1.57	1.00	3.59
3.30	1.08	3.05	3.59	1.07	3.31
1.45	1.08	3.30	1.45	1.07	3.59
1.89	1.08	3.30	1.94	1.07	3.59
1.44	1.08	3.06	1.45	1.07	3.32
1.21	1.05	3.05	1.17	1.00	3.31
1.17	1.08	3.05	1.17	1.07	3.31
3.05	1.08	3.05	3.31	1.07	3.31
1.08	1.08	3.30	1.07	1.07	3.59
1.74	1.08	3.05	1.79	1.07	3.31
1.19	1.08	3.06	1.17	1.07	3.32
1.25	1.08	3.30	1.25	1.07	3.59
1.52	1.08	3.30	1.55	1.07	3.59
1.69	1.08	3.05	1.78	1.07	3.31

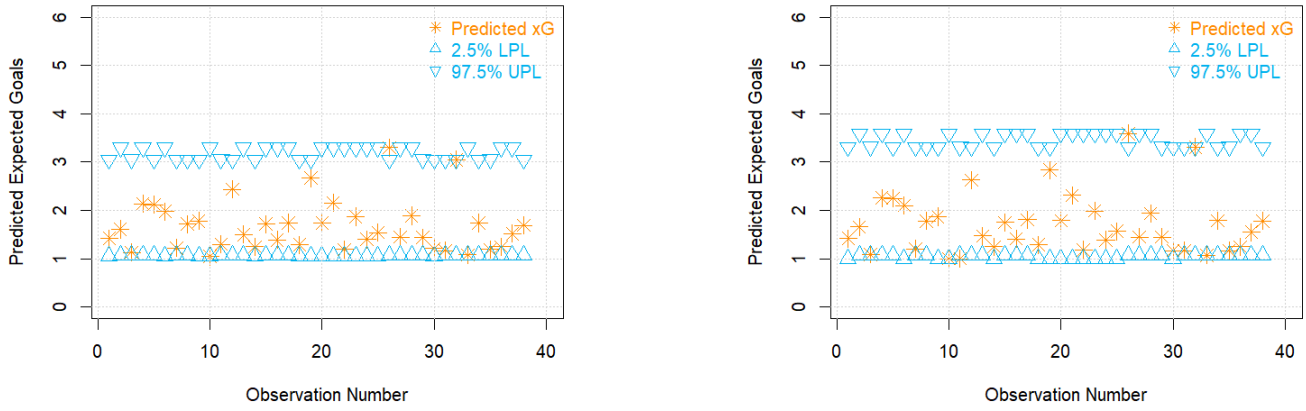
Table 6. Elastic Net and Group Lasso Regression to Predict Expected Goal Values and 95% Prediction Intervals, on the Response

The above tables highlight several interesting features. Each of these advanced regression techniques produced predicted values of expected goals that were relatively similar to one another. This confirms the findings from previous testing, where it was seen that these models were performing with similar success to one another, with a high adjusted R^2 value. Additionally, the vast majority of these predicted values of expected goals fall within the 95% prediction intervals for each test record, as generated through bootstrapping. These results are reflected in the figures (figure 10–11) below, depicting the 95% prediction interval for each modeling technique. The prediction interval for each technique includes its 2.5% lower prediction limit (2.5% LPL) and 97.5% upper prediction limit (97.5% UPL).



(a) Ridge Predicted xG Prediction Interval **(b)** Lasso Predicted xG Prediction Interval

Figure 10. 95% Prediction Intervals for Ridge and Lasso Predictions of Expected Goals



(a) Elastic Net Predicted xG Prediction Interval **(b)** Group Lasso Predicted xG Prediction Interval

Figure 11. 95% Prediction Intervals for Elastic Net and Group Lasso Predictions of Expected Goals

Again, as expected, most of the predicted responses are within the 95% prediction limits. The few predictions that fell outside the bounds of the corresponding prediction interval, however, are deemed to be multivariate outliers. The results of each of these graphs, combined with the results of modeling, show that Ridge, Lasso, Net Elastic, and Group Lasso regression all predicted the expected goal values with acceptable accuracy.

In addition to bootstrapping and creating prediction intervals for each record in the test dataset, each modeling technique required a prediction interval for its overall predicted value of expected goals. A 95% global bootstrap prediction interval (GPL) for each of the advanced regression and machine learning techniques is listed below:

Method	2.5% GPL	97.5% GPL
Ridge	1.658	1.963
Lasso	1.518	1.854
Elastic Net	1.497	1.786
Group Lasso	1.518	1.854

Table 7. 95% global bootstrap prediction interval (GPL)

Each of these 95% prediction intervals represent the predicted intervals for the true value of the team's expected goals scored per match, based on each of the four separate advanced regression techniques that yielded the highest accuracies and adjusted R^2 values. After analyzing these prediction intervals–both for the true value of expected goals, and for the bootstrapped test dataset–it's confirmed that these machine learning methods are strong, and that the variables chosen are significantly important indicators for predicting expected goals.

DISCUSSION

This section details the results and conclusions based on the data and analyses provided above, including the most important variables in predicting expected goals, and what this analysis means for the broader soccer and sports analytics communities.

In analyzing this data, a plethora of different models, ranging from general multiple linear regression method to complex regression methods, were used in determining the variables that were most impactful in predicting expected goals. Among the most advanced modeling techniques–namely, Ridge, Lasso, Elastic Net, and Group Lasso regression–there was very little difference in the success of each model in predicting for expected goal values. Each of these four models produced very similar results, including low error values, and an adjusted R^2 of close to 0.70. While the differences in error and R^2 values among these models is minimal, it does appear that Elastic Net regression is marginally better than the rest, as it produced lower error values (lower MSE, RMSE, and MAPE), and a higher adjusted R^2 value. The choice of an Elastic Net model means that it was most appropriate to penalize non-significant variables to play a lesser impact in the prediction of the square root value of expected goals. This superior model ultimately showed that there were several key variables, which are discussed below. The results of Net Elastic regression, and other similar models, are significant enough to determine that these advanced regression techniques succeed in predicting the square-root transformed expected goal value for a given soccer team.

The modeling procedures discussed earlier are key in identifying the variables that are most important in predicting expected goals. In these advanced regression models, as well as earlier models, such as forward and backward stepwise regression, there are four variables that continually are deemed to be important. These four variables include: percentage of shots on target, number of crosses, number of touches in the penalty area, and number of recoveries. In particular, the number of touches in the penalty area and the number of recoveries are especially significant, as they had the lowest p-values among the four variables from forward and backward stepwise regression, and also were the only two predictor variables remaining (alongside the intercept) in advanced techniques such as Lasso, Elastic Net, and Group Lasso.

These four variables were found to be the most impactful in determining a team's predicted expected goal value. By maximizing the number of goals that a team is expected to score in a match, there is a greater chance that the team will end up winning. Therefore, there is great potential for team success in training with a focus on taking quality shots on goal more often, by crossing the ball into the box from the wings, putting the ball into the box more often, and playing a press on defense, to record higher numbers of recoveries. Practicing these concepts in drills, to maximize the

percentage of shots on target, number of crosses, number of touches in the penalty area, and the number of recoveries ultimately is important in helping overall team success in a match. All four variables have a positive relationship with expected goals, indicating that an increase in any will help the team (be expected to) score more often. When combined with strong defense, a potent offense that threatens many goals is very powerful. A tactical focus on getting the most from your players, including focusing on these specific parts of the match, have significant ramifications in setting a team up for the most success possible.

CONCLUSIONS

The effectiveness of these models represents another step taken to continue to analyze sports and its data in a deeper manner. The largely untapped potential of statistics in sports is boundless. Teams are finally beginning to realize that having a focus on analytics can be revolutionary to the success of the team, in matches, in maintaining physical health, in marketing, in making money, in increasing fan engagement, and much more. Soccer clubs are beginning to employ analytics teams to find the angles to get any advantage possible in each match. An extra advantage can be the difference between winning or losing, and ultimately between winning trophies or nothing at all.

The results of this research can impact how teams prepare for matches and strategize to win. By determining the variables that are most significant in predicting expected goals, there are clear metrics for teams to focus on maximizing. Achieving higher levels of expected goals has a direct correlation to actual goals scored in a match, and therefore determining how to get the best out a team can be related to these several key predictor variables. Additionally, it is clear how complex statistical models can be used to represent a wealth of sports data. As mentioned in the introduction, shots only represent a very minimal proportion of the total events occurring in a soccer match. However, using machine learning models, it is possible to account for a much greater proportion of the overall number of events occurring in a match, beyond considering only the total number of shots. This modeling enables a team to make decisions on a greater amount of data than only one or two unique predictor variables.

This study had several limitations, that would be key focus areas in the expansion of this work. Firstly, it would have been relevant to consider additional predictor variables–especially those that are more defensive in nature, or reflect transition periods in the match. It would also be beneficial to collect a larger dataset, expanding just beyond one team's records, and rather opening the door to expand into other teams, or even more broad, other international leagues. This could elicit insight on how different variables are of importance across borders, reflecting unique styles of play. With this larger dataset, a truly random train and test split could be used, further benefiting the study. The implementation of these changes would be exciting and important for future continuation of the study.

Advances in the world of sports analytics are very exciting, as they open the door to further innovation in a relatively new industry. New measures for team success are being found, and have the ability to continue to change the way players, coaches, and fans think about and analyze matches. This study reflects the growth of analytics opportunities in sports, and particularly soccer, and how much more can be discovered through deep learning methods. Key variables to improve team performance and health, and routes to commercial success are imperative for sports franchises around the world, and there is an opening to make meaningful change in this community. Our goal in writing this paper is to be a part of the sports analytics revolution. To push this research forward, we plan to extend the work to the spectrum of more sophisticated machine learning methods, such as deep learning (DL), neural networks (NN), and artificial neural networks (ANN) on big data related to professional soccer leagues.

ACKNOWLEDGMENTS

This study was made possible through the support of several parties, and we are sincerely grateful for their contributions. We extend our heartfelt thanks to the reviewers and the editorial team for their invaluable feedback, which significantly enhanced the quality of this manuscript.

DATA AND CODES AVAILABILITY

The data and R code are available upon request from the corresponding author.

CONFLICTS OF INTEREST

The authors have no known conflicts of interest in this research and manuscript submitted to the American Journal of Undergraduate Research (AJUR).

REFERENCES

1. Kuper, S., & Szymanski, S. (2018). *Soccernomics: Why England loses; Why Germany, Spain, and France Win; and Why One Day Japan, Iraq, and the United States Will Become Kings of the World's Most Popular Sport*. New York, NY, Nation Books.
2. Lewis, M. (2004). *Moneyball*. WW Norton.
3. Strategic Market Research. (2022, April). Industry Report and Statistics (Facts & Figures) - Number of Deployments & New Installations, Installation Cost & Demand by Sports Type. Sports Analytics Market Size, Global Trend Forecast – 2030. Retrieved January 31, 2023. *https://www.strategic marketresearch.com/market-report/sports-analytics-market*
4. Schilling, D. (2016, May 25). Sports analytics don't exist in "BlackWorld"? That's ridiculous. *https://www.theguardian.com/sport/blog/2016/may/25/sports-analytics-african-americans-michael-wilbon-article*
5. Chotiner, I. (2019, June 6). Jalen Rose has a Problem with Basketball Analytics. *https://www.newyorker.com/news/q-and-a/jalen-rose-has-a-problem-with-basketball-analytics*
6. Kirshner, A. (2022, December 15). Football has Found its New Bogeyman. *https://www.theatlantic.com/technology/archive/2022/11/sports-analytics-nfl-football/672275/*
7. Sommi, G. (2020, December 2). A Crash Course in Soccer Analytics. Retrieved February 16, 2023. *https://www.samford.edu/sports-analytics/fans/2020/A-Crash-Course-in-Soccer-Analytics*
8. Eggels, H. P. H. (2016). Expected Goals in Soccer: Explaining Match Results using Predictive Analytics (thesis). Eindhoven University of Technology, Eindhoven. *https://pure.tue.nl/ws/files/46945853/855660-1.pdf*
9. Statsbomb. (2022, November 15). *What is xG? How is it calculated? – Statsbomb – Data Champions*. What Are Expected Goals (xG)? Retrieved February 2, 2023. *https://statsbomb.com/soccer metrics/expected-goals-xg-explained/#:~:text=Put%20simply%2C%20Expected%20Goals%20(xG,scale%20between%200%20and%201*
10. Muller, J. (2022, March 4). Penalties are too generous a reward. we have a solution... and it involves running. *https://theathletic.com/3161748/2022/03/04/penalties-are-too-generous-a-reward-we-have-a-solution-and-it-involves-running/*
11. Tippett, J. (2019). *The Expected Goals Philosophy*. Self-published.
12. Perl, J., Grunz, A., & Memmert, D. (2013). Tactics Analysis in Soccer – An Advanced Approach. *International Journal of Computer Science in Sport*, 12(1), 33–44.
13. Muller, J. (2021, September 14). Possession Is The Puzzle Of Soccer Analytics. These Models Are Trying To Solve It. Retrieved February 1, 2023. *https://fivethirtyeight.com/features/possession -is-the-puzzle-of-soccer-analytics-these-models-are-trying-to-solve-it/*
14. Sánchez Gálvez, A. M., Álvarez González, R., Sánchez Gálvez, S., & Anzures García, M. (2022). Model to Predict the Result of a Soccer Match Based on the Number of Goals Scored by a Single Team. *ComputaciÃ¸n y Sistemas* 26(1). *https://doi.org/10.13053/cys-26-1-4172*
15. Decroos T., & Davis, J. (2019). Interpretable Prediction of Goals in Soccer. *http://statsbomb.com/wp -content/uploads/2019/10/decroos-interpretability-statsbomb.pdf*
16. Inan, T. (2021). Using Poisson model for goal prediction in European football. *Journal of Human Sport and Exercise*, 16(4), 942-955. *https://doi.org/10.14198/jhse.2021.164.16*
17. Liu, D. Z., Bratko, A., Prevc, J., Pataky, L., Noori, M., Sathyanarayana, S., & Lipovsek, U. (2020, December

8). Predicting Soccer Goals in Near Real Time Using Computer Vision. *https://aws.amazon.com/blogs/machine-learning/predicting-soccer-goals-in-near-real-time-using-computer-vision/*

18. Tweedale, A. (2022, June 24). Expected goals: Explained. *https://www.coachesvoice.com/cv/expected-goals-xg-explained/#: :text=Although%20the%20expected%20goals%20metric,such%20as%20an %20entire%20season*
19. Whitmore, J. (2023, August 8). What is expected goals (XG)? *https://theanalyst.com/na/2023/08/what-is-expected-goals-xg/*
20. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019, October 28). A public data set of spatio-temporal match events in soccer competitions. *https://www.nature.com/articles/s41597-019-0247-7*
21. Soccerment Research. (2019, November 21). AZ Alkmaar: Creating success through Data Analytics. *https://soccerment.com/az-alkmaar-success-data-analytics/*
22. Johnson, R. A., & Wichern, D. W. (2020). *Applied Multivariate Statistical Analysis* 6th ed. Pearson.
23. LaMorte, W. W. (2016, May 31). The Multiple Linear Regression Equation. *https://sphweb.bumc.bu.edu /otlt/mph-modules/bs/bs704-ep713_multivariablemethods/bs704-ep713_multivariablemethods2.html*
24. Mendenhall, W., & Sincich, T. (2011). In *A Second Course in Statistics: Regression Analysis* 7th ed., p. 333. Pearson.
25. Penn State Eberly College of Science. (n.d.). Detecting Multicollinearity Using Variance Inflation Factors. *https://online.stat.psu.edu/stat462/node/180/*
26. University of Cambridge Cognition and Brain Sciences Unit. (2003, November 27). *Collinearity: Its origins, effects, signs, symptoms and cures*. FAQ/Collinearity - CBU statistics Wiki. *https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Collinearity*
27. Dean, A., Voss, D., & Draguljić, D. (2017). *Design and Analysis of Experiments* 2nd ed. Springer.
28. Patrone, C. (2022, August 11). The Likelihood Ratio Test. *https://towardsdatascience.com/the-likelihood-ratio-test-463455b34de9*
29. Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications* 7th ed. Thomson Brooks/Cole.
30. Walczak, B., & Massart, D. L. (Eds.). (2000). *Data Handling in Science and Technology* Vol. 22. Elsevier.
31. Zajic, A. (2022, November 29). What is Akaike Information Criterion (AIC)? *https://builtin.com/data-science/what-is-aic*
32. Zach. (2021, August 25). A Complete Guide to Stepwise Regression in R - Statology. Retrieved February 16, 2023. *https://www.statology.org/stepwiseregression-r/*
33. Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.
34. Kenny, A., Solomon, D., Patwary, M. S., & Das, K. P. (2022). Sparse Regression with Clustered Predictors. *Journal of Statistical Research*, 56(1), 37–53. *https://doi.org/10.3329/jsr.v56i1.63945*
35. Friedman, J., Hastie, T., & Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, Articles 33 (1), 1–22. *https: //doi.org/10.18637/jss.v033.i01.*
36. Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55–67. *https://doi.org/10.1080/00401706.1970.10488634*
37. Penn State Eberly College of Science. (n.d.). Ridge Regression. *https://online.stat.psu.edu/stat857/node/155/*
38. Zach. (2020, November 13). Lasso Regression in R (Step-by-Step). Retrieved February 16, 2023. *https://www.statology.org/lasso-regression-in-r/*
39. Zou H. & Hastie T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society* B, 67(2), 301–320.
40. Yuan, M. and L. Lin (2006), Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society*, Series B 68, 49–67.
41. Lee, S.-H. (2023, January 12). Exclusive Lasso and group Lasso using R code. *https://ibkrcampus.com /ibkr-quant-news/exclusive-lasso-and-group-lasso-using-r-code/#: :text= While%20group%20lasso %20selects%20all,each%20other%20within%20each%20group*

42. Meier, L., van de Geer, S., & Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1), 53-âĂŞ71. *https://doi.org/10.1111/j.1467-9868.2007.00627.x*

43. Hodson, T. O. (2022, July 19). Root-mean-square error (RMSE) or mean absolute error (mae): When to use them or not. *https://gmd.copernicus.org/articles/15/5481/2022/*

44. Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician*, 69(4), 371–386. *https://doi.org/10.1080 /00031305.2015.1089789*

45. Penn State Eberly College of Science. (n.d.). Bootstrapping Methods. *https://online.stat.psu.edu/stat500/lesson/11/11.2/11.2.1*

ADDITIONAL REFERENCES

Arora, A. (2022, July 15). The Most Important Things You Need To Know About Elastic Net. *https://analyticsarora.com/the-most-important-things-you-need-to-know-about-elastic-net/*

Box, G. E. P., and Cox, D. R. (1964). An Analysis of Transformations (With Discussion). *Journal of the Royal Statistical Society B*, 26, 211–252.

Datasciencebeginners. (2020, May 17). *https://www.r-bloggers.com/2020/05/simple-guide-to-ridge-regression-in-r/#: :text=Overview,and%20tries%20to%20minimize%20them*

Glen, S. (2023, April 17). Akaike's Information Criterion: Definition, Formulas. *https://www.statisticshowto.com/akaikes-information-criterion/*

Hastie, T. J. & Pregibon, D. (1992). Generalized linear models. Chapter 6 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Heuer A. & Rubner O. (2012). How Does the Past of a Soccer Match Influence Its Future? Concepts and Statistical Analysis. *PLoS ONE* 7(11). 10.1371/journal.pone.0047678

Maher, M. (1982) Modelling association football scores. *Statist Neerland* 36: 109.

National Institute of Standards and Technology, U.S. Department of Commerce. (n.d.-a). Likelihood Ratio Tests. *https://www.itl.nist.gov/div898/handbook/apr/section2/apr233.htm*

Nopp S., Memmert D., Kempe M., & Vogelbein, M. (2012). Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer. *International Journal of Sports Science*. German Sport University: Cologne. 10.5923/s.sports.201401.05

Penn State Eberly College of Science. (n.d.). The Lasso. *https://online.stat.psu.edu/stat508/lesson/5/5.4*

Rue H., & Salvesen O. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician* 49, 399–418.

The Investopedia Team, Potters, C., & Eichler, R. (2023, September 28). R-Squared vs. Adjusted R-Squared: What's the Difference? *https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp#: :text=What%20Is%20the%20Difference%20Between,and%20R%2Dsquared%20does%20not*

Tianqi Chen and Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Van Roy M., Yang W., De Raedt L. & Davis J. (2021). Analyzing Learned Markov Decision Processes using Model Checking for Providing Tactical Advice in Professional Soccer. *https://eos.cs.kuleuven.be/sites/eos.cs.kuleuven.be/files/Defensive-Tactical-Advice-IJCAI-AISA2021.pdf*

Wallstreetmojo Team, & Vaidya, D. (n.d.). Elastic Net. *https://www.wallstreetmojo.com/elastic-net/*

Zhou, Y., R. Jin, and S. Hoi (2010), Exclusive Lasso for Multi-task Feature Selection. *In International Conference on Artificial Intelligence and Statistics*, 988–995.

ABOUT THE STUDENT AUTHOR

Tristan Rumsey is an alumnus of Butler University, having graduated in May 2024. Tristan majored in Statistics, and had minors in Business Administration and Data Science. While at school, and as a requirement of graduating with honors, Tristan wrote a thesis under the same name, and is now looking to publish the work with a prestigious under-

graduate journal. Currently, Tristan works in the industry in downtown Indianapolis. Tristan worked on this research under supervision of Dr. Shaha Patwary who is a co-author of this research.

PRESS SUMMARY

This work discusses the growth of analytics in sports within a soccer context. Specifically, the "expected goal" (xG), a key metric in modern soccer analytics, is the focus of the study, which seeks to analyze the most important predictor variables in maximizing xG. This analysis will give us key insights into how a team can practice, strategize, and plan appropriately to give themselves the best chance of winning, even if they might not have the same resources as larger clubs or organizations.