Analysis on “Football Player Statistics (Premier League from 2021-2022)”

1. **Introduction**

Football Analytics has become more and more obvious these days and has originated since the 18th Century. The main concept of bringing this data is to check the output of Goals and determining the players who perform well, gaining that competitive edge over others that has continued since World War II. Football managers and Directors spend a lot of time and money before they bring a particular player in their squad and build a team who could lead them to the League Cups and Titles there are to win.

Before Scouting a player, they must be aware of the position, type of player, strengths and weaknesses, and the past record of achievements that the player has. To portray a player as Excellent player and add him to the roster, the attributes of a player are analyzed by the desired team managers before they are recruited by other teams and have a good chance in the transfer market to aim to be the Top Clubs in the English Championship.

The ability to make sense of the gathered performance player data to scout the player’s style that also matches the coach, maintains a good relationship with the team enhancing mutual trust and respect among them. The signing of players based on their ratios helps them to discover innovative, counter-intuitive, and winning strategies with the results of the previous matches thereby improving granularity of their overall stats and performance.

The Expected Goals ratio has been one of the most revolutionary metrics to calculate the output of Goals for a team and the probability of scoring them is based on several factors like distance, angle of the shot, weak foot or strong foot, type of attack, direction of shot taken etc. Again, Goals are not the only measure of a players worth since, the one who passes, and the way that pass had been made for the scoring player, who has created the chance for the Goal scorer, is of the highest recognition and prominence. Some players score less Goals but the way they give away the ball so that a Goal is scored is called an Assist and it has the same value as a Goal in Football.

Another important metric that is considered as a key factor in Football is Defense or the ability to win the ball back from the opposition and not allow the other team players to score a goal by becoming the shield other than the Goalkeeper. In a technical aspect, football is generally categorized into Attack, Mid-field, and Defense where the stats of a player in detail can be seen below as follows:

Chart

Description automatically generated with medium confidence

Thus, the **Attack, Passing and Defense** constitute the main factors of a player’s profile. The way of playing of a player matters during analysis which includes **Vision** (seeing the formations and passing the ball, creating chances for more goals), **Dribbling** (using various skills needed to bring the ball into the Goal zone by getting past the defenders), **Heading** (scoring goals with the head and passing the ball using the head), **Crossing** (the ability of passing the ball the from far side to the center above the defenders), **Tackles** (the number of successful times player has won the ball from the opposition), **Cards obtained** (the yellow and red cards that are obtained when a player fouls another), and **Physical condition** ( the stamina and speed of the player to continue playing for longer duration of the game). Having objectives can help speed up the learning processes and create virtuous development cycles, making data analytics a powerful tool in Football to predict, identify and cultivate a players’ potential.

**About the Data Set:**

This dataset contains the Statistics of football players who played in the Premier League from (2021-2022). It has 692 rows and 29 columns consisting of various attributes and contributions of the players in detail.

Y = The Best Premier League Player with Goals. (New)

The 29 Column names listed below are the key variables (X's) to conclude the Y variable and are Abbreviated as follows:

Player: Player's name.  
Team: Played club during 2021-2020.  
Nation: Player's nation.  
Pos: Position that one plays in.  
Age: Player's age.  
MP: Matches played.  
Starts: Matches started in the playing 11.  
Min: Minutes played.  
90s: Minutes played divided by 90.  
Gls: Goals scored or allowed.  
Ast: Assists.  
G-PK: Non-Penalty Goals.  
PK: Penalty Kicks made.  
PKatt: Penalty Kicks attended.  
CrdY: Yellow Cards.  
CrdR: Red Cards.  
Gls 90: Goals scored per 90 mins.  
Ast 90: Assists per 90 mins.  
G+A 90: Goals and Assists per 90 mins.  
G-PK 90: Goals minus Penalty Kicks made per 90 mins.  
G+A-PK 90: Goals plus Assists minus Penalty Kicks made per 90 mins.  
xG: Expected Goals.  
npxG: Non-Penalty Expected Goals.  
xA: Expected Assists.  
npxG+xA: Non-Penalty Expected Goals plus Expected Assists.  
xG 90: Expected Goals per 90 mins.  
npxG 90: Non-Penalty Expected Goals made per 90 mins.  
xA 90: Expected Assists made per 90 mins.  
npxG+xA 90: Non-Penalty Expected Goals plus Expected Assists made per 90 mins. Link: <https://www.kaggle.com/datasets/omkargowda/football-players-stats-premier-league-20212022>.

**Linear Regression Analysis:**

A screenshot of a computer

Description automatically generated with low confidence

The First step is Converting the Gls variable to numeric and making a new column which consists of numbers, and this will be our new Y variable from now on. The summary statistics of fbplst.df data frame with 31 variables is shown above.

fbplst.df$New\_Gls <- as.numeric(fbplst.df$Gls)

Now, the Linear model is run on the data frame by selecting only the desired or most wanted variables that effect the Goal value of a Player and how it effects indirectly. The variables that tell us the number of Goals a player has scored constitute of:

Player (Name of the player), Position (Pos), Age, Matches Played (MP), Starts, X90s (Min per game), Assists (Ast), GPK (Goals without Penalties), PK (Penalty Kicks), CrdY,CrdR (Yellow and Red Cards obtained), G.A90 (Goals and Assists combined per 90 mins of a game), xG ( Expected Goals ), npxG (Non-Penalty Expected Goals ), xA ( Expected Assists ), xG90, xA90, npxG90, npxG.xA90. (The ratios of expected ones per 90 mins of the game).

## Selecting only the necessary fields or columns as mentioned above for our Regression model

selected.var <- c (1, 4, 5, 6, 7, 9, 11, 12, 13, 15, 16, 19, 21, 22, 23, 24, 25, 28, 30)

Our next step will be partitioning the data into training data and validation data where 60% will be Training and the rest 40% will accommodate to Validation.

# Partition data

set.seed(1) # set seed for reproducing the partition

train.index <- sample(c(1:546), 328)

The value 328 is mentioned here since 60 % of 546 is 328 and this can also be obtained from the code below by considering the length of the total observations and multiplying it with 60 for our training data.

##train.index <- sample(C(1:length(fbplst.df)), 0.6 \* length(fbplst.df))

train.df <- fbplst.df[train.index, selected.var]

After the training data has been set, the rest of the data is assigned to the validation by using the negative index and applied to all the selected variables.

valid.df <- fbplst.df[-train.index, selected.var]

Finally, the Linear model for our data frame and its selected variables is performed with the command ‘lm’. The selected variables are added as well.

fbplst.lm <- lm(New\_Gls ~ MP + Age + Pos + Ast + X90s + Player + GPK + PK + CrdY + G.A90 + G.A.PK90 + xG + npxG.xA + xA + xG.xA90 + npxG.xA90 , data = train.df) # lm can account for categorical variables.

The output of the above Linear Model is not accurate as we have large number of regressors, and the Players names are many.

Table

Description automatically generated

As per the result, we have NaN, and we did not get any kind of R-square, p-values or the F-statistic because of the many regressors that were present.

Now, the regression model is run with only 10 regressors rather than 20 regressors and optimizing our result to get the R-square and the corresponding p-values to show the significance of the variables.

As per our observation, the residuals, p-values, and R-square have been generated and residual standard error is only 0.3 % for 310 degrees of freedom. The Adjusted R-square is 99% indicating that our model is very good and these regressors are worth the Y. The Variables, Pos Forward, Assists, Goals minus Penalty Kicks, Expected Goals, Non-Expected Goals and Assists, Expected Goals and Assists per 90 minutes of the game and Non – Expected Goals per 90 minutes are the major factors to determine the Goals that the Player has scored, and the rest of the variables have greater p-values showing less significance.

Table

Description automatically generated

Thus, now prediction of residuals for the first 20 values in our data frame is done to check if the values are proper.

Table

Description automatically generated

As per the above observations, many of the actual residuals seem to be zero and the others are 5 or less with only a few being 10 and greater. The predicted value and actual value do not have much difference and gives us the reason why there is such a high R-square.

Text

Description automatically generated

The figure is obtained after running the Linear Model with the ‘Best Subset method’ which shows the Adjusted R-square and the variables that are required to make our model precise. According to the final 0.99, Forwards have scored the most goals, followed by Midfielders and the Goals minus Penalty Kicks, Expected Goals, Assists and Non-Expected Goals per 90 minutes are the prominent ones to give us our Y and the Players importance.

Chart, histogram

Description automatically generated

This is the BIC diagram of the same regressors run via ‘Backward selection’ method where Assists, and X90s are not significant when compared to the (above) Best subset method.

Next, ‘Forward selection’ is run on the Regressors where all the variables are included in the Last model except the Yellow Card that is given to the player. A picture containing text

Description automatically generated

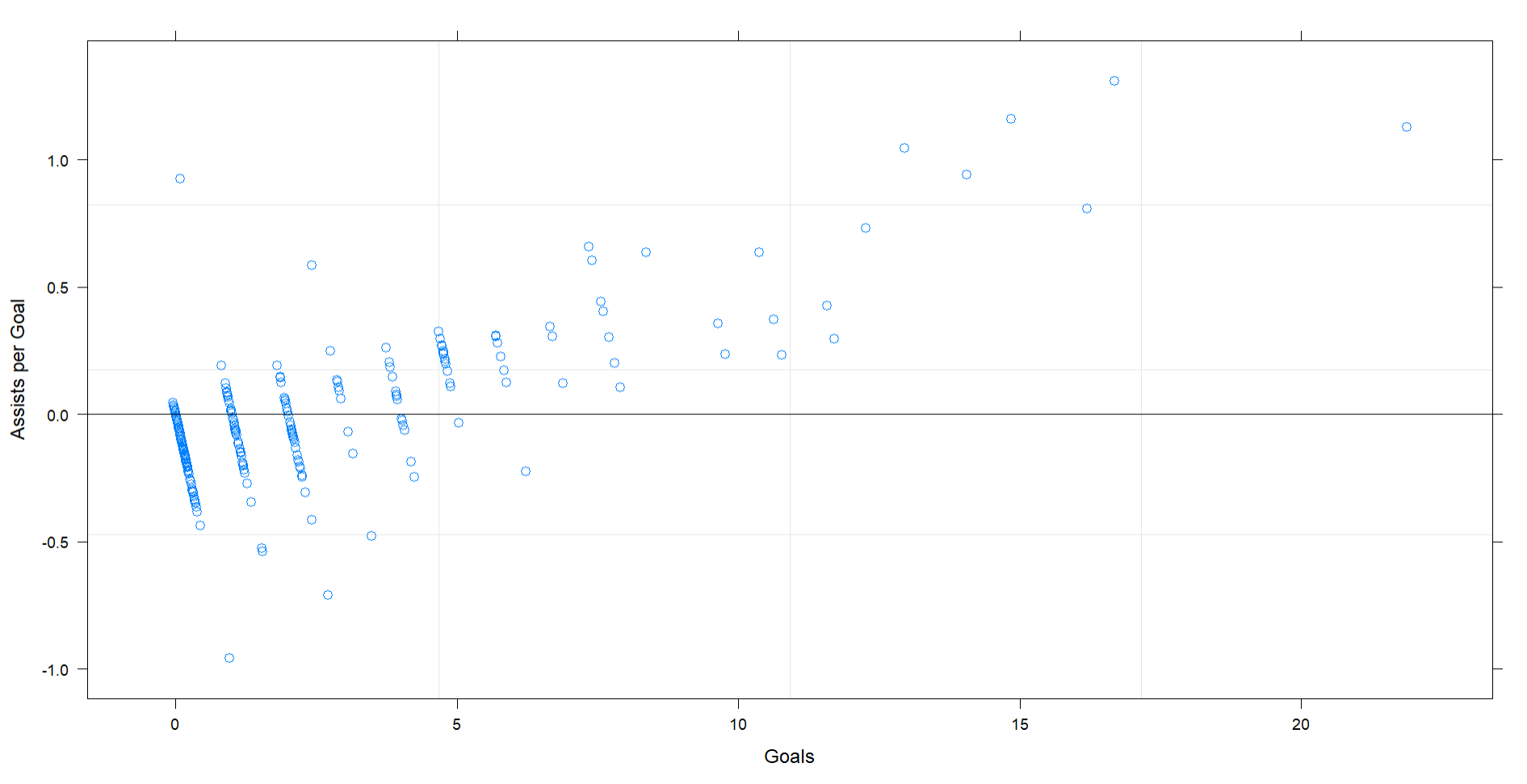
It also indicates that the Goal Keepers and Defenders have less impact than the forwards since they are not in range of the Goal.

Now, when the ‘Sequence Repetition’ method is run to check if we get the same result or a different one, the value tends to shift in between the regression but reach the same result as before showing us that Forwards, Mid fielders, Assists, X90s and the rest of the variables are significant except the Yellow Cards given.

Text

Description automatically generated with low confidence

Thus, after comparing all the variables, charts and methods, the most significant ones have been obtained that effect our Y and play a major role in its depiction.



The above figure tells us that how many Players have scored Goals as many Assists they have to give during the season 2021-22. There is only one player who has scored more than 20 Goals and given assists at a good rate while most of them are found in the 0 to 5 region of Goals and Assists respectively.

Chart, line chart

Description automatically generated

This QQ plot tells us that there are only a few players who have scored this season and the points that are away from the Regression line (zero /0) tell us about it. While some players have improved their quantity, some players have had a bad season than the previous one thereby dropping down from the measured Theoretical value.

Table

Description automatically generated

The training data error values and the correlation between them is shown in the above table. It tells us that Matches Played, Assists, Position of the Player (Forward and Mid field) are the major factors in determining the Goal count and telling the value.

Chart, histogram

Description automatically generated

The above histogram shows us that there are only a few players who scored more Goals and there are less outliers in the Linear Model.

Text

Description automatically generated

The accuracy measure of our Goal variable (Y) is very good as the RMSE is only 0.16 and the Mean Absolute Error is 0.07. The Percentage error is displayed as NaN since there are a lot of zero Values in the player columns as many have not scored even a single goal or assist and we cannot divide by zero to get its value. MAPE is Infinite as Something by 0 is as we all know not possible and thus, we have our values and predictions from the Linear Regression that we ran.

A picture containing logo

Description automatically generated

This is another example of the Regression Model where the Position of the Player has not been taken as a Regressor and the overall R-square has lowered by 10% indicating that the position of the player matters to consider the number of Goals that will be scored and every other factor here is important because of very less amount of regressors in-turn showing that all these are significant.

**Summary and Conclusion:**

Thus, after all these Models and Methods have shown that the Goals variable is directly depended on the Number of Assists, Matches, Expected Goals and Play time per 90 minutes of the game. The Residuals are very low, and the Histogram tells us that there are very less outliers with few exceptional players scoring Goals more than 15. The Position of the player and its importance is shown along when tested with all the different methods. A lot of factors were considered before Analyzing the Player position with most Potential.

There are still a lot of options to explore on what else can be considered to see the Best in the Player.