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. These are refined to 546 rows and 32 columns where we add our Y with 0/1 and the New.Gls variable.

Y = The Goals variable 0/1 (New) where ‘0’ is for Players with less than the mean value of Goals and ‘1’ is the value of the Players more than mean.

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.

**Logistic Regression Analysis:**

Table

Description automatically generated

The First step is Converting the New.Gls variable to numeric and making a new column which consists of numbers, and this will be our new Y variable with 0/1 from now on. On running the regression model, we get an AIC of 654 with the same variables that were taken during Linear regression Analysis.

The Estimates, Standard error and Deviance residuals for the rows are observed selecting Binomial as our type of Logistic Model considering the training data.

Most of the Residuals are close to zero with 25 Fischer Iterations and the names and the AIC can be seen in the diagram below.

The Logistic Regression model is run based on 60% Training data while the rest 40% is Validation data and the results are compared with the predicted and actual values to see if the context matches and gives us the accurate result as required.

Graphical user interface, text

Description automatically generated with medium confidence

The Co-efficients, residuals, fitted values, methods are saved in the logit.reg model and our job is to obtain the AUC curve and the Cumulative curve and show he efficiency of our model with the True and False Positives that our model has.

The above model does not show us p-values since the Player variable was added and it can’t process each player with its attribute differently.

Thus, running the Logistic regression model without the Player variable, the results obtained are as follows:

Table

Description automatically generated

The Deviance residuals are almost close to zero as obtained in the previous model, but the p-values of every variable are 1 which implies there is a defect with our model. The estimates are thrown off and the Standard error is very much which is not good. The AIC is 48 for 25 Fischer Iterations with Null and Residual Deviance being the same for 326 and 302 degrees of freedom.

Text

Description automatically generated

The next figure shows us the stats of the predicted model i.e., validation model which has the type ‘response’ and gives us the mean, variance, null errors etc.

Table

Description automatically generated

Though the null values have already been eliminated, the Sum and Median of our model is quite low and the standard deviation is not even 1. The rage and max values are 0.99, very close to 1 which gives us the optimal point of our Logistic regression as 1. The Minimum value is followed by 10 zeroes and can be rounded off to 0 without any doubt and the coefficient variance is also not more than 1.4.

Now, the accuracy of the model is determined and the table for the Classification matrix is generated to give us the True and False positives of our data. The accuracy is close to 70 percent while the False positives are 146. The values are under the probability of 0.5. The True Positives are only 7 and True Negatives are around 57 as seen below.

Graphical user interface, text, application, email

Description automatically generated

The Misclassification error in our validation data for these 16 variables on whom the Logistic regression was run on is 30 percent which makes sense with our accuracy being close to 70 percent. The Optimal Cut-off is 0.9999, very close to 100 percent along with the FPR and TPR with the High Specificity Index for the first 5 variables as seen below.

Text

Description automatically generated

Text

Description automatically generated

The above figure shows the first 16 predicted values and the actual values indicating how close they are to the original ones in the validation data. Further, proving that the Specification of 0.99 in our model is accurate.

Now, The ROC curve is presented which covers 63 percent of the graph as per the accuracy. The Sensitivity has a lot of breaks and no gradual uprising.

Chart, histogram

Description automatically generated

The code below shows the gains and the number of observations for it to reach the Optimal value of the Cumulative graph which is 7. The greatest number of observations that it could reach at optimal value was 219 and the Cumulative graph is shown below them.

Text

Description automatically generated

There is a gradual and sharp incline in the graph from 0 to 80. After 100 observations, the graph is stabilized, and the value is at its optimum.

Chart

Description automatically generated

Table

Description automatically generated with medium confidence

From the above values, the mean of the gains and height of the Mean Logistic value is obtained. The height of the max value does not cross 3 and the min is 0. The gains for the starting observations are zero and significantly make their way to 70 and 100 percent respectively.

This can be proven with the Decile chart and the underlying percentile values.

Chart, histogram

Description automatically generated

Finally, the last result of our Model-1 would be the step AIC to check the significance and correlation of variables. The Deviance is zero with low AIC values indicating less significance when compared to the Goals minus Penalty Kicks ratio. The results can be seen below.

Table

Description automatically generated

What we infer from our first Model is that the Accuracy can be increased and the number of regressors can be reduced from 16 to a single digit number. 63 % is not even close to being a good model and we need to run a lot of different models to compare and see if it gets any better than the present.

**Second Model:**

In this model, only 10 variables are our regressors. They are:

MP + Pos + Ast + GPK + PK + CrdY + G. A90 + G.A.PK90 + xG. xA90 + npxG.xA90.

Table

Description automatically generated

As seen above, there is not much change in the p-values of the variables or the Deviation residuals, degrees of freedom of the variables and the AIC which is 48.

The only little change is in the estimates and the reduction in the number of regressors.

Table

Description automatically generated with medium confidence

NO change can be observed in the Max and Min value but there is a little reduction in the sum from 80 to 64 and variance from 0.48 to 0.20 with a minute increase of coefficient from 1.3 to 1.5.

A picture containing calendar

Description automatically generated

The confusion matrix shows us that the lesser the regressors, the less is the accuracy and the number of zeroes matter as well.

Text

Description automatically generated with medium confidence

As per the accuracy, the True and False positives are at around 50% and the ROC curve is estimated to be near that value. The Specificity of the model has reduced drastically from 99% to 56% in Model-2. There is an increase in the Misclassification error from 30 to 47 percent indicating a depreciation of the Model with no improvements.

Graphical user interface, text, application

Description automatically generated

These values are the ones that are got from the prediction of the validation data after comparing the actual values with the predicted ones. A lot of False positives and True Positives are obtained with zero False negatives in Model-2.

Chart, histogram

Description automatically generated

The ROC curve value has been reduced from 63 to 57 percent with the reduction of regressors, specificity and sensitivity. The curve goes randomly until the 0.5 region and gains a good incline after 0.75. Since, more Area under the ROC indicates a better model, a look at the cumulative graph is not mandatory as it is the same with the same decile chart so this time let’s change the variables again to get a better ROC value without running the Step AIC since it will be low anyway.

**Logistic Model-3:**

The reduction in variables is only reducing the accuracy and making our Model worse thus, this time let us consider all 16 variables and filter only the ones suited to our Y and increase the Area under the ROC curve.

After the comparison of the Model stats, the results are as follows:

Graphical user interface, text, application, email

Description automatically generated

In this model, there is a reduction of False positives and a little increase in True Positives and out of the 16 regressors run, I have selected only 14. The False negatives have increased as well but not that much.

A picture containing table

Description automatically generated

As observed from the above figure, the misclassification error has been reduced to 22 percent from the 1st model of 30 percent. The False positives have significantly reduced with just a drop of 20 percent Specificity. The True Positive values are 89 percent, the highest value obtained so far.

Chart

Description automatically generated

The ROC curve above shows a 14 percent increase in the area and is near to 80 percent. With the increase of True Positives and Elimination of False negatives, the area under our curve has significantly gone up. This Model is by far the best model among the three.

Chart

Description automatically generated

The Cumulative curve is almost the same with just a sharp incline from 0 to 60 and then a bump at the 65 mark which shows that there was some residual there.

The observations and the sum remaining the same, has not impacted much on the curve and the Model is seemingly good.

Chart, funnel chart

Description automatically generated

This is another ROC curve which is obtained by removing some regressors from the 1st model. The highest area of around 85 percent shows us that this is the best model of all. Removal of variables like the Card obtained, Matches Played or the Expected Goals ratio is removed.

The Actual and Predicted values are just the same when it comes to zero or one and the Validation data almost matches the logit.reg3 predicted values with less standard variation, no deviation residuals and good amount of Specificity by Sensitivity.

**Summary and Conclusion:**

Thus, all these Models and Methods have shown that the Goals variable is directly depended on the Play time per 90 minutes of the game, Goals minus Penalty Kicks, Position of the Player, and the rest in balance. The Residuals are very low, and the Cumulative graph tells us that there are very good chances for predicting few exceptional players scoring Goals. 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.

Link: <https://www.kaggle.com/datasets/omkargowda/football-players-stats-premier-league-20212022>.