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 30 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.

**Best Model in Logistic Regression Analysis:**

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.

This model tells us the classification matrix and its values based on the New.Gls(Y) variable during classification.

Graphical user interface, text, application

Description automatically generated

This matrix gives us the highest Accuracy value of 81 percent which is not seen in any of the above confusion matrices. The increase of True Positives along with the False Positives with minimum number of the Negatives is the proof for such accuracy.

Text

Description automatically generated

As seen in the above figure, the Optimal Cut-off is as 26 percent with a misclassification error of only 19 percent when compared to 30 in the first one.

Text

Description automatically generated

After including the regressors with the Y variable, this model gives us a True Positives of around 88 percent while the False Positives are only around 15 percent.

Chart

Description automatically generated

The ROC curve has an Area of 92 percent, the highest seen so far, and is the best model with very less misclassification and good amount of True Positive values. The values keep going till 0.5 and have gradual breaks until the 0.75 mark. After a sharp rise again till 0.85, there is only one break until it reaches 1.00 and flows continuously after the 0.25 x values.

Chart, line chart, scatter chart

Description automatically generated

The Cumulative graph above indicates a steep incline with more bent angle than the previous one. The observations instead of starting at 80, seems to be starting around 92-95 of the Cases and then going on until 219.

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

**Best Model in Neural Network Analysis:**

The Neural Network 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.

A picture containing logo

Description automatically generated

According to the above values, there are only single digit negatives, and the True positives are 69, much greater than 54 compared to Model-2 with the False positives topping at 141. Our Model has significantly improved, and the error has also been reduced as seen below.

Diagram

Description automatically generated

This NN-plot has an error of less than 1 percent and a greater number of steps (44) than any other models making it the best model of all.

The ROC Curve below has 99 percent area covered under it indicating it has been better than the training data and can be called as the Best.

The dips of the curve are at the top left at the 0.90 to 1.00 point but can be discarded as it does not matter compared to the rest of the 99 percent of the curve.

The Sensitivity and Specificity of the Graph would be more since the True and False negatives are in single and not more than 9.

Chart, treemap chart

Description automatically generated

Now, the histogram shows us how the values have been predicted and what is the range of the most predicted values. As seen, most of them are at 0 and the others are at 0.8 while the rest are scattered in between.

Chart, histogram

Description automatically generated

The Boxplot below shows us that the median of the values ranges at 0.2 while the minimum and maximum stand at 0 and 0.9 respectively showing just how good our model is.

Chart, box and whisker chart

Description automatically generated

The Lift chart is almost as same as the validation data in Logistic Regression model and does not need much explanation.

Chart, line chart

Description automatically generated

**Best Model in Tree Classification Analysis:**

The Tree classification 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.

As observed in the output below for the prediction of the classification table, New.Gls our Y variable which shows the Goals scored by a player at the top. The most important variables that impact the Y are GPK (Goals minus Penalty Kicks), xG (Expected goals ratio), G.A.PK90 (The overall Goals, assists and Penalty kicks per 90 minutes), npxG (Non -expected Goals scored), G.A90(Goals and Assists per 90 mins of a game). The rest have a mean goal less than 5 so they can be considered less important. The ones at zero are not significant at all.

A picture containing diagram

Description automatically generated

According to the above output, the tree ends with only the GPK node as the all the values are either classified as false (220) or true (108) and there is no need for the tree to traverse the next nodes for the correct classification. Thus, the tree is run without the GPK and checked for other important nodes and variables other than GPK.

Pruned Tree:

The Tree after removing certain unimportant variables and having only the major ones to support the Tree model is obtaining after a process called as Pruning. This not only helps the Tree to be cut down to a smaller size but also removes the misclassifications of the bad variables present in it.

A screenshot of a computer

Description automatically generated with medium confidence

After running the pruned tree with 10 variables, the error percentage is brought down to 0.33 and the most important ones are displayed after the print(pruned.ct) command is run for this model. The root node for the pruned tree is same with only 3 others being prominent.

Text

Description automatically generated

The pruned tree can be seen clearly and the classified nodes in the right corner being 93 while the False positives are 196 with only 12 players less than the 0.6 ratio of GAPK90. The 3 most important variables are derived from the pruned tree as shown below:

Diagram

Description automatically generated with low confidence

As seen in the below pruned tree for validation data, there is another important observation that was missing in the training data. The Expected Goals ratio after pruning in the validation data has increased from 1.4 to 1.6 respectively. There is an increase in True positives from 59 to 72 compared to Model 1 and 2 of the normal and deeper trees while the False positives remain the same. The negatives are in single digits indicating that the Accuracy of the model has increased to be more than 85 percent in the pruned tree.

Diagram

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

The above table shows the error percentage and the variables used in tree construction with the relative error values of the first 5 data points.

**Random Forest:**

Text

Description automatically generated

The random forest has been run for the previous 10 variables to compare and check if the output is better than the deeper tree and the pruned tree. There are 204 False positives with 96 True positives, the same as Pruned tree but the Negatives are a little more when compared to the previous one. The error rate is only 8.6 percent meaning our model is very good and is not under the tendency of overfitting which was seen in the Deeper Tree.

Text

Description automatically generated with low confidence

As seen in the above table, the True and False positives are more in number with Accuracy and Specificity being more than 80 percent. The misclassifications are in single digits only which is better than the pruned tree.

Thus, among all the models that were observed, the random forest and the pruned forest were the Best.

**Bagging and Boosting (For Trees):**

This model is used to reduce the mean decrease value and the variance in the Tree classification thereby promoting a better version of the Pruned Tree with subsets of the important data points and variables. They try to reduce the complexity of the Tree and remove the extra steps in our Model.

1. Bagging:

Graphical user interface, text, application, email

Description automatically generated

As per the table above, there are good amount of False and True positives with only 30 False negatives. The bagging method have improved our tree in one way and impacted the rest by increasing the Negatives amount from mere 15 to 30 doubling them. The False positives have reduced in the training and validation data and converted to Negatives with a small increment in True positives. The Accuracy and Sensitivity is still only under 80 percent, and this would not be good or the best model.

1. Boosting

Graphical user interface, text, application

Description automatically generated

Boosting has helped us reduce the number of False negatives but still seems not that much improved when compared to the pruned tree. The True and False positives are almost the same and have no actual increment or reduction in the error classification. Thus, the pruned tree is the only Best Model with less errors and no overfitting which is seen in the deeper trees.

**Summary and Conclusion:**

Thus, all these Models and Methods have shown that the Goals variable is directly depended on the Expected Goals variable, Goals minus Penalty Kicks, Non-Expected Goals, assists per 90 mins of the game, Assists, and Matches played. The rest of the variables did not seem that important and including or removing variables does not make any sense as the most important variables masked the others during classification. The categorical variables are out of scope as models cannot run them with the numbers, and we cannot generate a Model with the Player names or their Positions and classify.

Out of the Best Models that were obtained in the Logistic Regression model, the highest number of True positives were only 51 when compared to 72, the highest in the Tree model. The error misclassification was 14 percent and Optimal cut-off was 26 percent making it the least considerable. Though the True positives and Sensitivity of the model is more than 85 percent, it still has only 91 percent area in the ROC and is not up to the mark when compared to the Trees or the Neural Network models.

The Second-Best model out of the three is the Tree classification which is subjected to Pruning and considered as Random Forest whose True positive values are more than 90 percent with the Negatives being in single digit and very less. The misclassification error was 8.54 percent, and the tree was cut down to only a limited number of variables thereby not giving us a complete idea of which variable is significant and masking the others.

Finally, the Best Classification model among the three would be the Neural Network which has the number of True and False positives with Negatives being only 3,5 which is very very less. The Area under the ROC was stupendous 98 percent with a perfect gains plot. The misclassification error was only 0.97 percent with over 44 steps(iterations) making the model that much precise and accurate. The Sensitivity of the Model is more than 95 percent making it the best model among all. The point where NN gets an edge over tree is that the classification does not mask the insignificant variables and tells us the accurate measure of all the variables.

A lot of factors must be considered before Analyzing the Player with most Potential. Thus, for now the Pruned Tree is our perfect model with the 4 variables. There are still a lot of options to explore on what else can be considered to see the Best in the Player.