Project

DIT862 - Statistical Methods for Data Science

January 2022

Contents

1 Introduction

Have you ever bet on a match either on an official site or, with your friends? Would it not be convenient to have a model that can inform you which team has the highest probability to win? This analysis will take care of this issue as the aim is to find a model that can give a hint or even better, predict, a future game outcome based on analyzing football data from the past years.

1.1 Purpose

The intention of this analysis is to examine if one can utilize old data to predict future outcomes of football games using football data from the past five years. Moreover, the aim is to contribute to studies within sports analytics focusing on the five best football leagues in Europe.

1.2 Hypothesis

This analysis focus on the relationship between performed goals and game wins. The collected data is from the following leagues, Bundesliga, La Liga, Premier League, Serie A, Ligue 1 and the Russian Premier League with the considered time period, or ”season”, fall 2014 to spring 2019. Within the field of football, the stated leagues are the five top leading ones in the world. Therefore, it is of our interest to analyze this data as there is a lot of different actors that are affected by the performance in these leagues. To evaluate whether more goals in a game indicates a win, following hypothesis are constructed:

1. With increasing number of scored goals, the probability of winning a match will increase.

2 Method and Data

The data set used for this analysis is downloaded from Kaggle, where the data is collected from 2014 to 2019. The included football teams are from the five top leagues in Europe, Bundesliga, La Liga, Premier League, Serie A, Ligue 1 and the Russian Premier League. Furthermore, the statistical tests used in the analysis will be linear regression, logistic regression and random forest for classification. Using these three different methods will be helpful in order to determine whether to reject the null hypothesis or not.

Table 1: Summary of the variables

Variable	Label	Unit
Wins	Win or loss	Binary(1/0)
Scored	Number of goals	Integer
xG	Expected number of goals	Continuous
xGA	Expected number of goals against	Continuous
Deep	Deep passes made	Integer
Deep allowed	Deep passes allowed	Integer

Shown in Table 1, a total of three variables from the data set are considered whereas two of them are the input variables; scored goals and expected goals. On the other hand, the target variable that is being analyzed is the match result, win or loss. In order to test the data correctly, a three part train, validation and test split is made to separate the model training from the actual test.

For the purpose of studying the distribution of the data, four Q-Q plots are presented below where the left-sided plots are train data and the right-sided ones are validate data. Illustrated in all of the eight graphs on the next page, there appears to be right-skewed data where each variable do not follow a linear pattern indicating non-normal distribution which is the case for both the training and the validation set. This pattern is reasonable due to the nature of the data being football goals, or other football metrics, where the most number of goals probably will be centered close to zero.

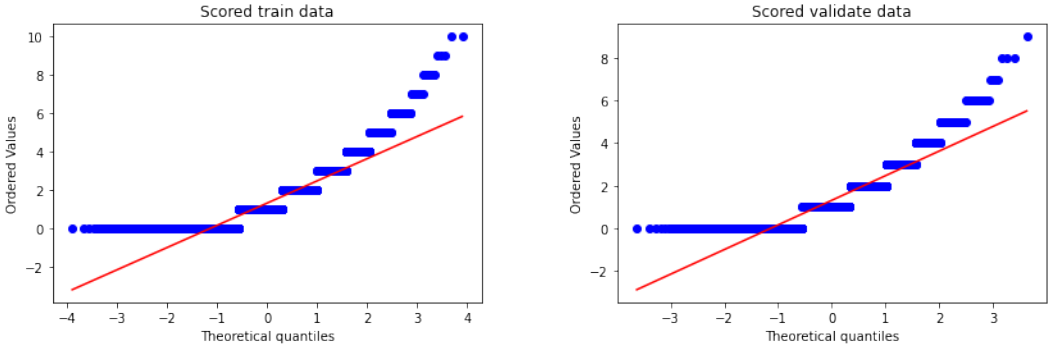


Figure 1: Probability plot of scored goals

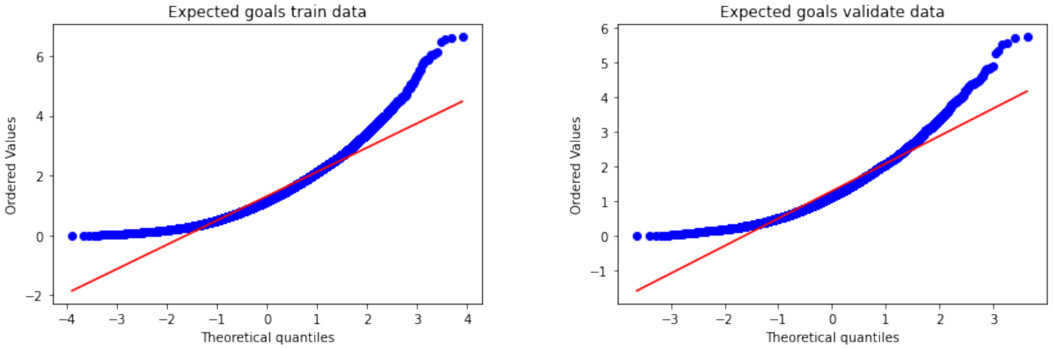


Figure 2: Probability plot of expected goals

3 Descriptive Analysis

Table 2: Descriptive statistics

Variable	N	Mean	St.Dev.	Min	25%	50%	75%	Max
Wins	14748	0.373	0.484	0.000	0.000	0.000	1.000	1.000
Scored	14748	1.341	1.244	0.000	0.000	1.000	2.000	10.000
xG	14748	1.309	0.841	0.000	0.683	1.152	1.767	6.630

Displayed in the table above, there is a total of 14 748 observations for each of the three variables. The mean of scored goals and expected goals are almost the same, indicating that the actual number of goals scored is closely related to the estimated variable, expected goals. Interpreting the standard deviation of the independent variables, scored goals has a higher value than expected goals which means that there is a larger data spread around the mean of scored goals. Furthermore, the quantiles for the scored goals displays that up to the 25th percentile there are not any goals made in the games. However, for the 50th percentile and above, there is one or more goals.

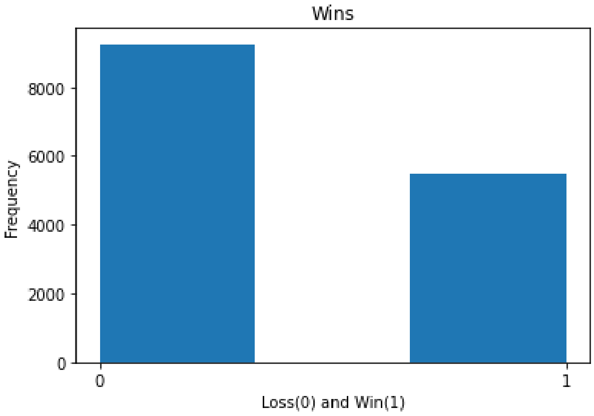


Figure 3: Histogram of win and losses

Figure 3 represents the amount of wins and not wins which includes draws and losses for the team that plays at the home field. Represented on the histogram, there is a higher frequency of not wins than wins whereas approximately, 9000 matches result in a draw or a loss while, 5000 ends up as wins.

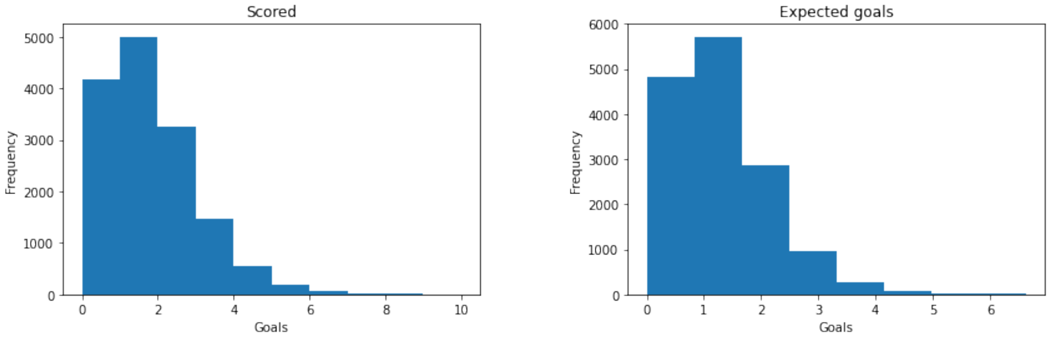
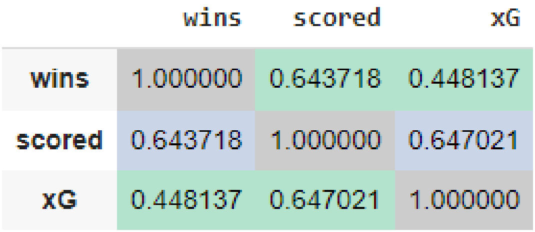


Figure 4: Histograms of scored and expected goals

The histograms for both scored goals and expected goals in Figure 4 are skewed as there is more data on the left side of each diagram. This indicates that each game has usually not more than three goals and that is rare that a game consists of more than four goals. This pattern is reasonable when comparing to the general notion of football games having few goals on average. The low amount of actual goals can be reflected in the expected goals where goals above two is rare.



	wins	scored	xG
wins	1.000000	0.643718	0.448137
scored	0.643718	1.000000	0.647021
xG	0.448137	0.647021	1.000000

Figure 5: Correlation matrix

In order to investigate how the selected variables are correlated, a correlation matrix is computed in Figure 5. Shown in the table, there is a strong correlation between scored goals and wins. The chosen independent variables are of interest because of how they are related to each other and the outcome of a match. Scored goals and if the match result is a win, can be seen as highly correlated. Moreover, the relationship between scored goals and expected number of goals is also strongly related. These key points are of importance as it is assumed that more goals on average leads to a win. If expected goals are to be used as a proxy for actual goals, the two variables must be related in order to utilize the properties of the continuous variable expected goals. Additionally, there is a strong relationship between scored goals and expected goals which is strengthen by the fact that these two variables has a closely related mean value which was mentioned above for Table 2. Lastly, expected goals do not indicate a high correlation with the variable of win thereby, one cannot state that these two variables have a reliable relationship.

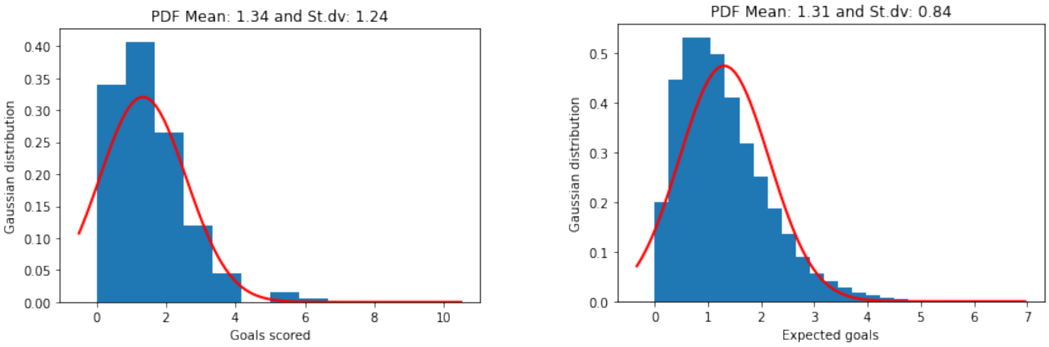


Figure 6: Probability density function for scored and expected goals

Probability Density Function (PDF) is represented as a graph, where the x-axis is a representation of a set of random and continuous values. The y-axis helps with distinguishing the probability for given random value. A histogram can be converted to a PDF by dividing all the buckets in the histogram with the frequency of all buckets respectively. The total sum becomes one, which gives the probability distribution. PDF's are probabilistic and histograms are statistical. PDF is the area under the curve, integral. An integral uses two parameters, a and b, to define a specific range where the function $f(x)$ can be utilized to calculate the area (Skiena, 2017). Hence, the formula for PDF is the following:

$$\int_a^b f(x)\, \mathrm{d}x \tag{1}$$

4 Predictive Analysis

To begin with, a simple linear regression have been computed to make a prediction if the football team is winning a game or not. The linear regression is stated with the following mathematical formula $y = a + xb$, where the parameter a is the intercept and b is denoted as the slope of the line. Interpreting the slope, chances to win a game occurs to increase with 25 percent for each goal. Studying the graph, there is an intersection of the line at four scored goals indicating that the probability of winning the game is 100 percent at this point. On the contrary, two goals indicate that the game will result in a draw whereas less than two goals will result in a loss for the given probability.

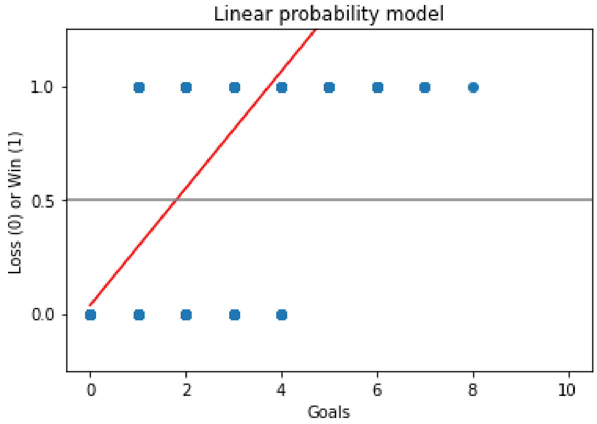


Figure 7: Linear regression of wins and losses

A logistic function, also named Sigmoid function, is illustrated as an S-shape where the values between zero and one are denoting the probability range. Instead of predicting the number of goals, you calculate the probability, or likelihood, of the game resulting in a win. This model is constructed to work with binary values, 0 and 1. As it is a binary model, there is only two possible outcomes where the probability of the output is based on the given input. The parameter for the logistic regression curve is the x-value, which is used in the Sigmoid function. Following mathematical functions represents the logistic regression:

$$Logisticfunction = \frac{p}{1-p} \tag{2}$$

$$Lossfunction, Sigmoid = \frac{1}{1+e^{-x}} \tag{3}$$

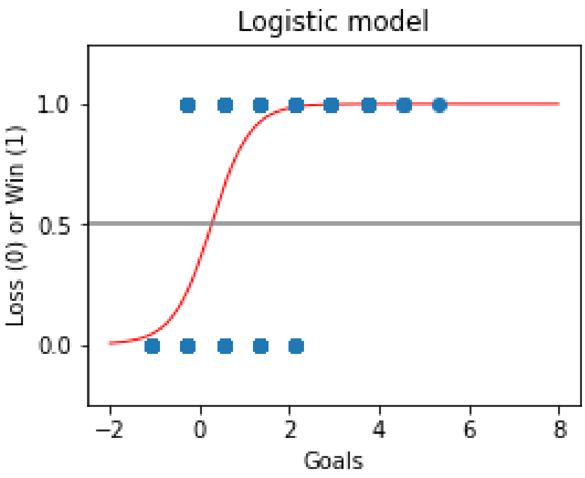


Figure 8: Logistic regression

Studying the logistic graph above of football data, the probability of a game outcome being a loss or a win are classified using the number of goals as a parameter to identify the possible game result. As goals increase, the probability of winning the game increases too. However, there is no clear separation between the data as overlapping appears to exist. Goals up to approximately two, are not perfectly separated between winning or losing. In other words, in a game where two goals have been succeeded, there is almost a 100 percent chance that one of the teams is winning but also a risk to get a draw which is indicated by the horizontal line in the middle of Figure 8. Nevertheless, if three or more goals are scored during a game the probability of winning is almost 100 percent which is also feasible in reality.

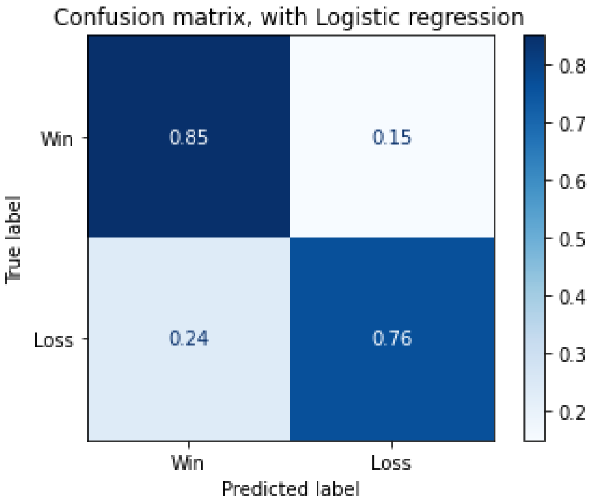


Figure 9: Confusion Matrix of the logistic regression

For the purpose of receiving more detailed insights of how the logistical regression classifies the outcome of football games, a confusion matrix is computed. Illustrated in Figure 9, there appears to be misclassification in the two different outcomes; wins and loss which was also recognized in the logistic regression in Figure 8 where data points overlapped. Predictions are plotted against the true labels where the darker shades on the diagonally indicate that the classifier works well whereas lighter shades tell that there exists some misclassifications. Shown in the heat map, 85 percent of the win predictions are true labels while, 15 percent are classified to be a loss when its true label is a win. On the other hand, 76 percent of the games are predicted to be a loss while 24 percent are falsely predicted as wins.

Lastly, in order to investigate the classification problem further, a Random Forest method is employed to predict the outcome of the football games. The random forest prediction is based on decision trees where leaf nodes aid in the classification of the data. After merging several decision trees, the method returns the average outcome. The main hyper parameters used when applying the random forest method are the number of trees as well as the number of splits made to each tree when analyzing the nodes. Parameters in this model is the level where split is supposed to take place for each node. The equation for this model is expressed below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{4}$$

Mean squared error (MSE) is used to calculate the distance between each node and the predicted value respectively. This supports the decision making of which branch is the most optimal for the specified forest.

$$Gini = 1 - \sum_{i=1}^{c}(p_i)^2 \tag{5}$$

When classifying data using Random Forest, a common classification function is the Gini index. The function uses multiple parameters like probability and the class it self to classify the index for each respective node within the decision tree. The outcome of the Gini index function is deciding which of all branches has the highest likelihood to occur.

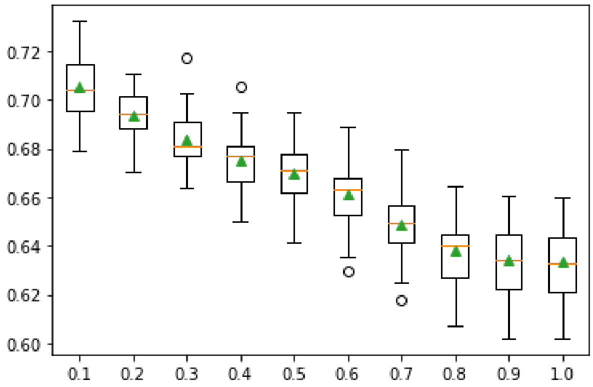


Figure 10: Box plot of Random Forest accuracy score

Accuracy scores for the samples are provided in Figure 10, where the data set is divided into several parts, from 10 percent up to 100 percent shown on the x-axis. Observing the boxplot, it is possible to conclude that the accuracy stabilizes at approximately 0.64 where at least 80 percent of the data is used.

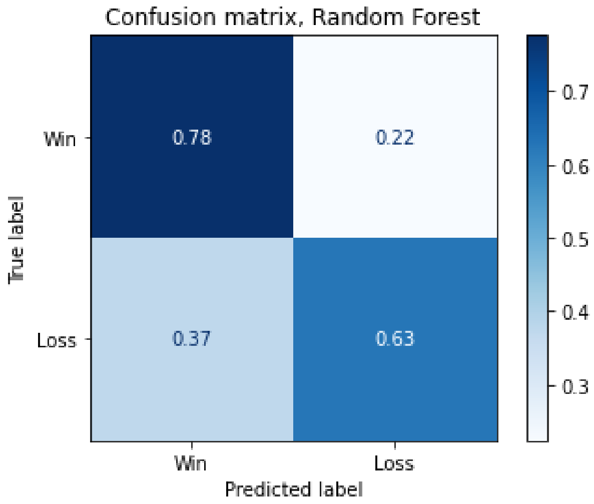


Figure 11: Confusion Matrix of the Random Forest

Similar to the heat map that was computed for the logistic regression, a confusion matrix are also inserted here in order to analyze how well the classifiers made with Random Forest are performing. In comparison to the logistic function, the confusion matrix for this method here is showing slightly different values. The ratio of misclassified outcomes in games, are more frequent in Random Forest classifier. 78 percent of the win predictions are true labels while, 22 percent are classified to be a loss when its true label is a win. On the contrary, 63 percent of the games are predicted to be a loss while 37 percent are falsely predicted as a win indicating a high misclassification as its true prediction is a loss.

Evaluating the random forest method against logic regression, the optimal outcome is to minimize the number of false predictions made by the respective method. It is desired to obtain as many true predictions of both wins and losses as possible thus, as few misclassifications as possible. A stable and accurate model will result in a clear distinction between the included classes. Given the predictions from both methods, it is possible to find the logistic regression to be slightly more successful in classifying the data correctly with fewer misclassifications.

Furthermore, computing a linear regression is a good way to start an analysis of this sort to receive an overview of the relationship between the dependent and independent variables. The method is used for forecasting where a prediction of the dependent variable, game outcome, is made for a value of the independent variable, scored goals. While studying the used method and observing its graph in Figure 7, a clear observation of data being overlapped is displayed. From that, one cannot distinct directly if two goals in a game result in a win or a loss. Consequently, this does not explain the data well as the predictions are not perfectly separated from each other. Applying linear regression could be a solid starting point in order to express one variable as a function of another. However, this does not explain the data further as one does not obtain insights about complex data of this sort. To summarize, the results indicate that the logistic regression is the preferable model in this case because of the higher number of true predictions found.

5 Conclusion

For this report, the main problem being investigated is the ability to predict the outcome of football games with the intention of providing several models for accurate predictions. The hypothesis states if an increasing number of scored goals will increase the chances of winning a game. The used predictive machine learning models, Linear regression, Logistic regression and Random Forest classification indicated similar results. In conclusion, from the analysis in this report, it is possible to predict football game outcomes based on scored goals and expected goals. These results give useful insights to the sports analytic side of statistics for those interested in match outcomes.

For future investigations, improvements could be made regarding the Random Forest classifier where making additional adjustments such as tweaking the hyperparameters could be beneficial. Observing the three methods used in this report, Random Forest obtained the most potential and possible room for improvement. In addition, looking into other variables provided in the data set and possibly classifying not only wins and losses but also, tied matches is of great interest. Furthermore, it would be interesting to investigate if COVID-19 has affected the outcomes of games as the teams have been playing in front of empty stands. Analyzing if the affect of fans and supporters present in the arena encourages players to perform better would be a interesting factor to explore in this context.

References

Football Data: Expected Goals and Other Metrics (2021). URL: https://www.kaggle.com/slehkyi/extended-football-stats-for-european-leagues-xg?select=understat_per_game.csv (visited on 11/21/2021).

Football Data: Expected Goals and Other Metrics (2021). URL: https://www.kaggle.com/slehkyi/extended-football-stats-for-european-leagues-xg?select=understat.com.csv (visited on 11/21/2021).

Skiena, S. S. (2017). *The Data Science Design Manual.* 1st. Springer Publishing Company, Incorporated.