Soccer Competitiveness Using Shots on Target: Data Mining Approach

Neetu Singh[1(✉)], Apoorva Kanthwal[2], and Prashant Bidhuri[3]

[1] University of Illinois at Springfield, Springfield, IL 62703, USA
nsing2@uis.edu
[2] SEI Investments, One Freedom Valley Dr, Oaks 19456, USA
akanthwal@seic.com
[3] Enterprise Cloud Solutions, 600 Third Avenue, 2nd Floor, New York, NY 10016, USA
prashant.bidhuri@eclouds.co

Abstract. This paper presents the model for the competitiveness of soccer matches played in the top four European soccer leagues. Every soccer match in every league holds some importance and contributes towards the overall performance of the league compared to other leagues. These individual results constitute a single season. A lot of aspects of a team and a season are attributed to their final positions in the league. These positions, however, do not detail the competitiveness of a single match. This research aims to highlight the competitiveness in each match without any relation to how the season may have ended. A match gives out a lot of details towards how it was approached by a team. A win may not constitute competitiveness, but the approach does. The idea is to look at individual statistics of a match and use them to construct a model using SEMMA approach of data mining, that classifies the matches based on how competitive they were. This research constructs various models for classification as each model provides its own variant based on the different methodologies used in the individual models. Our analysis is mainly depended on, but not limited to, the number of attempted shots on goal and on the number of those shots that were on target. An important characteristic of the attempts on goals is that they are subjective to the performance of a team and its ability to try and secure a win in a match. This performance formulates competitiveness which is the basis of our research.

Keywords: Competitiveness · Data mining · Shots on target · Performance evaluation · Misclassification rate

1 Introduction

Competitiveness in sports, especially in leagues, is often characterized as a measure of the toughness of the sport. Sports leagues throughout the world have multiple measures of competition and success. In soccer, a team's success in the league is determined by the points scored at the end of a season. These points are a total of the points accumulated based on the result of each match, where a win gives 3 points and a draw gives 1 point. The difference between these points and a team's league position at the end of a

F. F.-H. Nah and K. Siau (Eds.): HCII 2019, LNCS 11589, pp. 141–150, 2019.
https://doi.org/10.1007/978-3-030-22338-0_11

season is considered as the level of competition during that season, which are some measures of competitive balance used in sports [1]. This research is focused more towards the competitiveness of each match played in the top 4 soccer leagues of Europe and not the inter-league competitiveness which is one of the main reasons why the measure of 'competitive balance' was not considered to measure the competitiveness in the study. The different approach to the competitive analysis led to the research question: "How competitive are matches in Top European Leagues?".

Competitiveness of a soccer match can be based on many factors such as possession of the ball, chances created, total goals scored, total shots attempted. Although crucial to a soccer match, the total points accumulated, the final position in the league, or number of goals scored provide a singular approach to measure competitiveness [2]. An important statistic in a team's attacking prowess is the number of shots on target that team attempted in a match. Combining the total shots on target of the two teams in a match gives the information about how much both teams attacked in a match. The difference of these shots on target by both teams gives the actual difference of the level of performance of both the teams in a match. As, this research is not focused on the outcome of a match or a season, but the performance of individual matches.

Another reason for this approach is because, in soccer, and in any team sport, a team that dominates the match in terms of statistics may not always have the result in their favor at the end of the game. Due to which, the number of goals or number of points were not considered as a measure of competitiveness. Significant research has been done in competitiveness in soccer in Europe using competitive balance and many other factors [1]. However, extant literature does not emphasize on competitiveness in soccer matches based on difference between shots on target.

2 Literature Review

Competitiveness in sports often give rise to a term called "Competitive Balance". It assesses the wins of teams within a season and performance within a team [1]. Although it is great for percentages and numbers, it only gives out competitiveness of leagues with respect to the teams playing in it [3]. Lower the balance, bigger the difference between stronger and weaker teams. What is usually missed out in this comparison is the competitiveness of each match played in these soccer leagues. A single strong team can alter the competitive balance of a league if it is far superior to the other teams. But that does not necessarily mean that the rest of the teams are not equally competitive amongst themselves.

Additionally, the competitive balance measures the competitiveness within a season. One single season does not define a league or a team. Consistency matters in sports and that's what defines the strength or weakness in a league. As Brad [3] concludes by providing his alternative solution to competitive balance, "CBR: This computationally simple statistic scales average team-specific variation in won-loss ratio during a number of seasons by the average within-season variation in won-loss percentage during the same period", the gap remains on measuring competitiveness based on the way the matches were played and not the outcome.

Other methods devised to measure competitiveness in soccer follow the age old competitive balance where the competitive balance was either used as the base to formulate a different method of measurement [4, 5] or there are attempts to overcome the shortcomings of the original method [6]. The method used in this research attempts to provide the measurement of competitiveness in European soccer a new approach in this field of study.

While researchers argue the frailties of the older method, few have provided ways to measure competitiveness of each match played over the years in the top European leagues. This research attempts to fill that gap with the method of measuring the difference between shots on target of two teams in each match played over the period of 12 years in the top European leagues.

However, newer methods have started to come up in the measurement of competitiveness. The most notable being the "Situational Score Line" devised by Wibowo [7]. He proposes two new methods to consider the outcome of a match apart from the score line:

- Away Rating: The impact of performances of a team in away matches to negate home field advantage.
- Opponent Rating: Quality of the opponent team in a match.

Another method to measure teams' performances in soccer was used by Julen [8] by comparing the statistics of matches played between teams to identify the success of teams playing in the football world cup. Although, the intended outcome was not competitiveness, it emphasizes the importance of individual match statistics.

This research employs similar methodology in the way that it uses match statistics of over 17000 matches played over the period of 12 years in the top four European leagues, but the intended outcome is not to measure success, but to measure competitiveness.

3 Data Analysis

The data was acquired from the website 'Football Data' which contains details of football matches being played in all European soccer leagues. Finding a data set that is in multiple formats, locations and files is an obstacle that is most common with researchers and analysts [9]. We collected information of matches of top 4 European leagues, English Premier League, German Bundesliga, Italian Serie A and Spanish La Liga. The initial sample consisted of data between 2000 and 2017.

The data set included 64 variables, out of which 42 were the betting odds of each match, hence we excluded those from the data set completely. Keeping in mind the main focus of analysis was to analyze competitiveness, we also excluded the count of penalty yellow cards and red cards given by the referee during the match as these do not factor in when considering competitiveness of a match in terms of game play. The final sample had 22 variables which included some derived variables as well with 17030 records.

3.1 Data Cleaning

The data gathered for the analysis is generally raw form of data that has lots of discrepancies including but not limited to missing and inconsistent values. Data set like this can lead to misdirected analysis, false conclusions and biased results. The data errors are mainly caused by mistakes in data entry, different methods of measurement used, integration errors etc. [10]. There are two types of data quality problems, single source problems and multi-source problems and both types were encountered in the data. In case of single source problem, duplicate rows were deleted, and null values were removed. In case of multi-source problems, the inconsistencies like extra columns were removed. The data was formed by combining multiple files into one and some additional fields had to be removed to keep the dataset consistent. The data obtained from a European website and hence needed date format corrections as well.

3.2 Data Preparation

Once the data is cleaned and the formats are uniform and appropriate, the next stage is preparing the data to get the best result out of the available data. Data preparation is a pre-requisite for successful data mining process and it has been generally seen that data preparation takes up 80% of the total time in analysis [11]. The different steps involved in data preparation were, extracting appropriate data, checking integrity of data, integrate data, create new variables if required [12]. Another important part of data preparation is preparing the score data. In this case 20% of the total data was used as score data. Score data included data from all the four leagues to maintain the homogeneity of the data. 52 different csv files containing one soccer season each from each league were collected and then command prompt was used to successfully combine data set into one csv file.

3.3 Variable Selection

In any kind of data analysis, once the data is cleaned and prepared the next step involves selecting the variable that will be used for the analysis. Identification of the appropriate predictors and the target variable is of utmost importance [13]. The variable selection will determine the results of classification, prediction and data analysis [14]. As explained before, as to why shots on target is an important factor in determining competitiveness, DST (Difference between shots on target) was the first choice of target variable. While exploring the dataset we observed wide range of values for DST ranging from 0 to 21. It would have been difficult to make a list of the values and then determine the result based upon these values. Thus, the values were categorized into ‘0’ and ‘1’. The value of DST from 0 to 5 was categorized as ‘0’ being the most competitive and the value of DST 6 or greater than 6 was categorized as ‘1’ being least competitive. Categorizing the DST values into 0 and 1 had the advantage as it made the target variable binary. The values ‘0’ and ‘1’ were in the variable ‘Class’ which is a

derived variable. Observing the variables DST and Class, they came out to be highly collinear as Class was derived from DST. To benefit the analysis, the variable DST was rejected, and Class was selected as the target variable. One of the cases in multicollinearity is when two variables are correlated and could lead to greater standard errors. More indifferent results are possible by avoiding collinearity, thus, it was important to reject and select the appropriate variables [15].

3.4 Data Mining Techniques

The SEMMA (Sample, Explore, Modify, Model, and Assess) data mining approach is used to develop the classification model using SAS® Enterprise Miner. The choice of model/technique depends on many factors such as how big the data is, type of target and predictors, and also on the end result that we are looking forward to i.e. whether we are trying to classify or predict in our analysis [16]. Based on the type of data that is selected there are two types of techniques that can be used to analyze the data, namely supervised learning technique and unsupervised learning technique [17]. When we consider supervised learning techniques, we are already aware of the various groups that exist (target variable is known), and this information is used for the analysis. Whereas, in unsupervised learning technique such as clustering we have no information about the target variable and try to look for patterns [18].

As our target variable is 'Class' which has two categories '0s' and '1s' and our predictors were continuous (interval variables), our analysis fits in the category of supervised learning technique with categorical response and continuous predictors. We decided to use three techniques which are Logistic Regression, Neural Networks, and *K*-Nearest Neighbors (Memory Based Reasoning) for the analysis of our data set.

Regression because of its flexibility in application is most widely used analysis technique. Further using logistic regression, it is possible to represent the chances of target variable falling in one category or class as compared to the other [19]. As we have a categorical target variable 'class' and we are trying to classify the percentage of matches of European League which fall in the respective classes '0' and '1', thus logistic regression with the help of iterations helps in identifying the predictors that will have the maximum likelihood of identifying the desired outcome. For Logistic regression we developed exhaustive regression, forward regression, backward regression, and stepwise regression.

The second technique that we used was Neural Networks. It is also known as Artificial Neural Networks and is one of the techniques that is used for classification and prediction. As neural network tries to understand and mimic how the human brain works, so it is possible to train a neural network efficiently thereby using it for pattern recognition and problem solving [20]. Neural network consists of input node, output node, hidden layers, and the weights are determined for the networks. Iterations keep improving the knowledge of the neural network thereby helps in giving nearly accurate predictions [21]. Neural networks have some disadvantages, with the main disadvantage being the inability to explain the networks and their relationships [22]. Hence, to avoid this black box problem, we also used a decision tree model to compare with the

neural network models. We used Neural Network with hidden units as 3 named 'Neural' and second with hidden units as 5 and named it as 'Neural 2'. The decision tree node was also kept to default as it was added to avoid the black box problem because of the ease of interpretation of the leaf nodes of the decision tree model.

Apart from logistic regression and neural networks, *K*-Nearest Neighbors also known as Memory Based Reasoning is used for the analysis. By using this technique with the help of mathematical formula like Euclidean distance it is possible to determine the k nearest neighbors, and class having majority is put under the values of k determined [18]. By using this technique, we can find k records in the training data set which are identical to the records that we are trying to do classification for [16]. All the above-mentioned techniques helped us in analyzing our data and gave an insight how the various predictors were affecting the target variable which is 'class'. MBR was developed for two different values of K ($K = 16$ *and* $K = 7$).

When the target variable is categorical in nature then one of the factors on which the models or techniques are compared is misclassification rate of the validation data set. If misclassification is high in the model it makes the analysis bias and also the statistical efficiency of the model is reduced [23]. To see that whether the model is performing good or bad we compared the misclassification rate of validation data set with the baseline misclassification [24]. This baseline misclassification also referred to as the naïve rule helps to examine how well the model is performing [16]. Thus, it is important to calculate the baseline misclassification rate, which in our case is 17.47%. In addition, Receiver Operator Characteristics (ROC) curve is used to evaluate the performance of our classification models. To further our analysis, final model comparison was conducted where the best models out of all the three techniques were compared to obtain the overall best model.

4 Results and Discussions

After model comparison Memory Based Reasoning with $K = 7$ turned out to be the best model with misclassification rate of 0.069094 or 6.9%, which is also less than the baseline misclassification rate of 17.47%. Thus, this also supported that the model was performing better than the baseline misclassification. Also, misclassification rate of neural network also improved to 0.079272 or 8%, and for logistic regression (stepwise) the misclassification rate was 0.138188 or 13.8%. The accuracy of the best model came out to be 93.09%, and ROC curve (Fig. 1) was further used to see how well the model is performing. As it is the graph plotted between true positive and false positive so more the graph is towards the upper left (towards true positive) better is the model. From the ROC curve shown in Fig. 1 it is evident that the model performs best for Memory Based Reasoning (MBR).

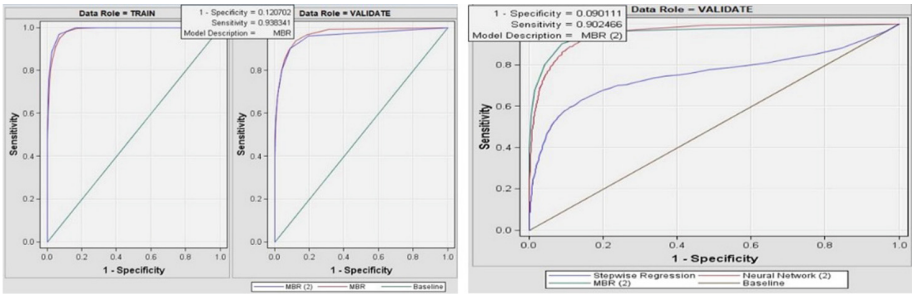

Fig. 1. Receiver Operator Characteristics (ROC) for model performance

Therefore, with final model comparison and taking all parameters into consideration Memory Based Reasoning with k = 7 came out to be the best model for our analysis.

The classification chart for the best model MBR (k = 7) with target variable 'class' was also plotted to further see how well the model classified, which in our case is the percentage of matches which fall in class '0' which is competitive and class '1' uncompetitive. The classification chart for the two classes is shown in Fig. 2.

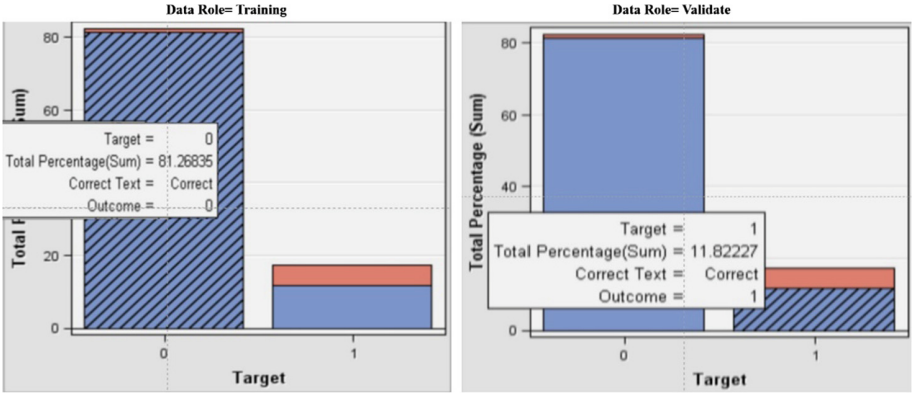

Fig. 2. Classification chart for MBR (k = 7)

The graph depicts that in target class 0, the total percentage is 81.26 which shows that the competitiveness is more as class 0 depicts more competitiveness. Similarly, target class 1, shows 11.82% of the matches fall under this category. This class 1, in our case is less competitive. It can also be summed as; that for class '0', the model classifies 81.26% of matches as competitive, and on the other hand for class '1', 11.82% of matches are classified as non-competitive by the MBR model.

We wanted to conduct extensive analysis, so we decided to score our data and used 20% our main data as score data. The results obtained after scoring also validated our

analysis. The results (Table 1) depicted that class 0, which in our case is the class depicting competitiveness, accounts for 82.15% of the total number of matches in the top European Leagues, whereas Class 1 which depicts less competitiveness accounts for mere 17.84% of the matches. This results clearly answers our question and validates that matches in the top European Leagues are highly competitive.

Table 1. Class variable summary statistics output

Class variable summary statistics				
Data role = SCORE Output type = CLASSIFICATION				
Variable	Numeric value	Formatted value	Frequency count	Percent
I_Class	.	0	2795	82.1576
I_Class	.	1	607	17.8424

5 Conclusion

This research offers a new dimension to the measurement of competitiveness in soccer. The variables used in the research have been used in previous research, but the target variable used has not been used in classifying competitiveness in soccer before. The findings were based on the shots attempted and not on the points accumulated, or the goals scored. The results show that the statistics of a soccer match are important to each match played even if the research and analysis have different targets.

The models constructed in the research concluded that 82% of the matches in our data set were competitive based on the classification of competitiveness. The model comparisons however, improved our analysis based on the accuracy of the models where MBR (2) turned out to be our best model for the research. The results of our models also validated the techniques that we chose for our analysis where each model had a positive output and gave more rigor to the analysis.

While this research is based on classifying matches that have already been played, further research into the data can be done to forecast competitiveness of the matches between the teams that are in this data set.

6 Implications

Competitiveness in soccer matches is rarely done based on matches. The general discussion of competitiveness has been taking place based on domestic leagues. This comparison, although extensive, is done using “Competitive Balance”, which is a common method to compare competitiveness in team sports. We recommend that “Difference Between Shots on Target” should also be considered as an important factor in assessing competitiveness.

This research also adds “Shots on Target” as a new dimension to the debate of determining competitiveness in soccer. Any future research on the topic should keep in mind that there is big difference in a final outcome of a match and how a match actually

unfolds. Predefined factors such as points, goals and league standings will predict outcomes but cannot tell the actual story of the match, which we believe our methodology of analyzing a soccer match will. Teams who have played other teams previously can predict the approach of their opponent based on the previous level of competitiveness in matches between them and assert the competitiveness of the upcoming match with that team and strategize for accordingly.

References

1. Pawlowski, T., Christoph, B., Hovemann, A.: Top clubs' performance and the competitive situation in European domestic football competitions. J. Sports Econ. **11**(2), 186–202 (2010)
2. Jessop, A.: A measure of competitiveness in leagues: a network approach. J. Oper. Res. Soc. **57**(12), 1425–1434 (2006)
3. Humphreys, B.R.: Alternative measures of competitive balance in sports leagues. J. Sports Econ. **3**(2), 133–148 (2002)
4. Criado, R., García, E., Pedroche, F., Romance, M.: A new method for comparing rankings through complex networks: model and analysis of competitiveness of major European soccer leagues. Chaos **23**(4), 043114 (2013)
5. Owen, P.D.: Limitations of the relative standard deviation of win percentages for measuring competitive balance in sports leagues. Econ. Lett. **109**(1), 38–41 (2010)
6. Eckard, E.W.: The NCAA cartel and competitive balance in college football. Rev. Ind. Organ. **13**(3), 347–369 (1998)
7. Wibowo, C.P.: Clustering seasonal performances of soccer teams based on situational score line 1, vol. 1, no. 1, May 2016
8. Castellano, J., Casamichana, D., Lago, C.: The use of match statistics that discriminate between successful and unsuccessful soccer teams. J. Hum. Kinet. **31**, 139–147 (2012)
9. Brown, J.G.: Using a multiple imputation technique to merge data sets. Appl. Econ. Lett. **9** (5), 311–314 (2002)
10. Hellerstein, J.M.: Quantitative Data Cleaning for Large Databases. United Nations Economic Commission for Europe, February 2008
11. Zhang, S., Zhang, C., Yang, Q.: Data preparation for data mining. Appl. Artif. Intell. **17** (5/6), 375 (2003)
12. Refaat, M.: Steps of data preparation. In: Data Preparation for Data Mining Using SAS. Morgan Kaufmann, San Francisco (2007)
13. Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E.W., Khalili, D.: A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J. Clin. Epidemiol. **71**(Supplement C), 76–85 (2016)
14. Trappenberg, T., Ouyang, J., Back, A.: Input variable selection: mutual information and linear mixing measures. IEEE Trans. Knowl. Data Eng. **18**(1), 37–46 (2006)
15. Yoo, W., Mayberry, R., Bae, S., Singh, K., (Peter) He, Q., Lillard, J.W.: A study of effects of multicollinearity in the multivariable analysis. Int. J. Appl. Sci. Technol. **4**(5), 9–19 (2014)
16. Schmueli, G., Bruce, P.C., Patel, N.R.: Data Mining for Business Analytics, Third. Wiley, Hoboken (2016)
17. Asheibi, A., Stirling, D., Sutanto, D.: Analyzing harmonic monitoring data using supervised and unsupervised learning. IEEE Trans. Power Delivery **24**(1), 293–301 (2009)
18. Baxter, M.J.: A review of supervised and unsupervised pattern recognition in archaeometry. Archaeometry **48**(4), 671–694 (2006)

19. Stoltzfus, J.C.: Logistic regression: a brief primer. Acad. Emerg. Med. **18**(10), 1099–1104 (2011)
20. Boritz, J.E., Kennedy, D.B., De Miranda e Albuquerque, A.: Predicting corporate failure using a neural network approach. Int. J. Intell. Syst. Account. Finan. Manag. **4**(2), 95–111 (1995)
21. Ince, H., Aktan, B.: A comparison of data mining techniques for credit scoring in banking: a managerial perspective. J. Bus. Econ. Manag. **10**(3), 233–240 (2009)
22. Tsai, C.-F., Chiou, Y.-J.: Earnings management prediction: a pilot study of combining neural networks and decision trees. Expert Syst. Appl. **36**(3), 7183–7191 (2009). Part 2
23. Barron, B.A.: The effects of misclassification on the estimation of relative risk. Biometrics **33**(2), 414–418 (1977)
24. Kayhan, V.O.: SAS Enterprise Miner Exercise and Assignment Handbook for Higher Education, Second. Valor Onur Kayhan (2016)