The Magic Formula to Score Goals: an Analysis of Football Events.

Pasquale Formicola, Pasquale Gravante, Angelo Limone, Antonio Mastroianni
June 3, 2023

Abstract

Scoring goals is a fundamental aspect of football, and teams constantly seek ways to enhance their goal-scoring capabilities. This paper explores the application of association analysis to a comprehensive dataset of football match events in order to uncover the elusive "magic formula" of scoring goals. By analyzing variables such as player positions, shot locations, and other event-specific data, association analysis provides a data-driven approach to identify significant associations and rules that contribute to goal-scoring success. The aim of this research is that findings offer valuable insights for coaches, analysts, and strategists in optimizing team formations, player roles, and tactical decisions to improve goal-scoring efficiency and overall team performance.

1 Introduction

Scoring goals lies at the heart of football, and teams continually strive to uncover the optimal combination of events that can maximize their goal-scoring potential. In recent years, the availability of comprehensive event-level data in football matches has opened up new avenues for analysis and exploration. This paper aims to investigate the application of association analysis techniques on football match events data to unveil the combination of events that increases or even maximizes the chance of scoring goals. By leveraging association analysis, we can uncover meaningful associations and rules among different events that provide insights into the patterns and strategies associated with goal-scoring success.

Traditionally, analyzing football matches has relied on aggregated statistics such as possession, shots on goal, and passing accuracy. While these metrics offer some level of understanding, they often fail to capture the nuanced interactions between individual events that contribute to scoring goals. By shifting our focus to event-level data, which encompasses detailed information about passes, shots, tackles, interceptions, and other specific events occurring throughout a match, we can gain a more granular understanding of the game dynamics [1].

The potential benefits of discovering the magic formula of scoring goals through association analysis are significant. Coaches, analysts, and strategists can utilize this knowledge to optimize team formations, player positioning, and tactical decisions during matches. By identifying event combinations associated with goal-scoring success, teams can refine their offensive strategies, improve decision-making on the field, and ultimately enhance their ability to score goals consistently.

This paper aims to contribute to the existing body of research by delving into the use of association analysis on football match events data for the specific purpose of uncovering the combination of events that increase or maximize the chances of scoring goals. Through a systematic analysis of event-level data and the application of association analysis techniques, we seek to unveil the hidden patterns and relationships that exist within the vast amount of information generated during a football match. By doing so, we hope to provide valuable insights that can revolutionize the way teams approach goal-scoring strategies and ultimately contribute to the overall advancement of football analytics.

2 Data Description

The idea behind the analysis is to analyze football matches in order to discover hidden rules in the sequence of events that lead to goals. The data chosen for this purpose is available on Kaggle [2] and consists of the entire series of events for each football match of the top-5 leagues in Europe (England, Spain, Germany, Italy and France) between 05/08/2011 and 22/01/2017. To each event corresponds information about the teams playing the match, the players involved, the event type, the type of pass or shot and its outcome, the pitch location in which it happens and which part of the body was used. Further details concerning the considered data will be discussed in the Appendix.

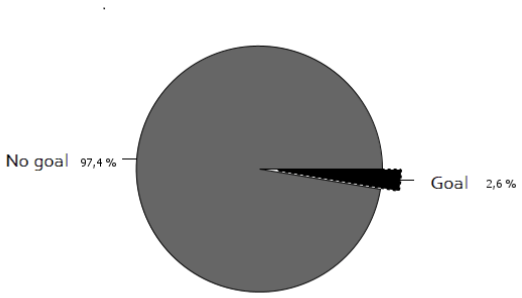
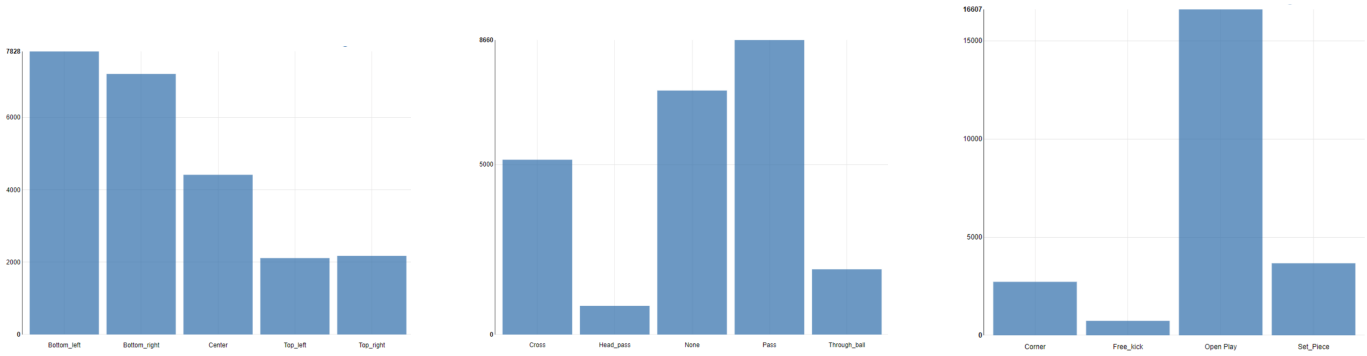


Figure 2: Percentage of goals

3 Pre-processing

Before diving into the processing part, multiple filter on the data have been applied in order to capture only the necessary information that comply with the purpose of the analysis.



(a) Shot Location (b) Assist Type (c) Situation

Figure 1: Bar charts showing the frequency of some of the considered variables

The proposed work has been developed on the KNIME [3] software. The considered data consists of more than 940000 rows representing events that happened during football matches. The details related to the these events are provided by twenty-two columns which give information about the playing time the event occurred, the kind of event that has occurred (offensive attempts, fouls and cards, substitutions, ecc.), the teams playing and the players involved in that event, the side of the pitch. Moreover, information about offensive plays are provided such as the location of the pitch in which the shot was taken, its outcome, the part of the body used to score, the type of assist and the situation in which the play has happened.
Anyways, since the aim of the research is to discover the "magic formula" to score goals, only offensive plays are considered for the analysis thus reducing the number of rows to 229135 thus including both failed attempts and goals. Moreover, it is necessary to remove all the rows representing attempt events that certainly do not resemble intention and thus are not possible to consider for the analysis as they would compromise the results. For this reasons, all the events concerning own goals are removed. Moreover, among the twenty-two available variables, six are chosen:

- Situation: state of the game during which the attempt happened.
- Assist Method: whether there is an assist and how was it performed.
- Location: area of the pitch from where the ball was shot
- Body part: indicates whether head, left or right foot was used.
- Shot Place: the destination of the ball after the shot.
- Goal: a binary variable specifying whether the attempt follows a goal or not.

After dealing with the selection of the appropriate variables for the analysis, the missing values problem is also considered. The only column containing missing values is Shot Place with 1034 missing values. Since it is reasonable to predict the shot location according to the other considered columns, Naive Bayes missing value imputation is employed through the use of the Weka NayveBayes node.
Since our scope is to perform association analysis of football events, the final aim of the pre-processing procedure is to generate a new structured column in the data that holds for each row a vector containing all the information related to the event related to that row. This specific column is usually referred to as "transaction" column. In order to achieve this structure, first of all, the encoding representing the categories of the each of the considered columns must be different among each column. In our case, some attributes of different columns were encoded with the same value. For example, one of the rows had value 3 for two columns indicating a player shooting from the center of the box (Location) towards the bottom left corner of the goal(Shot place). Thus, it is necessary that for these kind of rows identically encoded information are edited differently such that when they are aggregated in the transaction column they assume different values or names. This problem was solved by creating a dictionary for each column through the Table Creator node containing the new encoding for each of the columns categories. Right after that, a Cell Replacer node was used for each column to replace the old encoding with the new ones. From a set of numbers indicating the attribute of each column, the new encoding provides a string summarizing the attributes specific for each column. In particular, the encondings for shot taken with left and right foot are joined together due to the imbalance of the two classes.
Finally, the transaction column is created by means of a Column Combiner and a Cell Splitter nodes. It is noted that two different transaction columns are considered for the analysis. One considering all of the offensive attempts which is used to learn the association rules, while the second table contains only rows related to scored goals and it is used to investigate frequent itemsets in order to capture the highest-frequency combination of events that occur when a goal is scored.

4 Association Analysis

The primary objective of this chapter is to delve into the domain of football event data analysis using association analysis techniques. Association analysis, also known as market basket analysis or affinity analysis, originated in the retail industry to identify associations among frequently co-occurring items purchased by customers. However, its application has transcended traditional retail settings and found valuable use cases in diverse domains, including sports analytics. By applying association analysis techniques to football event data, it is possible to uncover meaningful associations between events, uncover frequently occurring itemsets, and derive actionable association rules. We aim to explore the inherent patterns and dependencies within football events and uncover the relationships that contribute to the scoring of goals.

4.1 Frequent Itemsets

In association analysis of football match events data, frequent itemsets play a crucial role in discovering meaningful relationships between various events occurring during a game. A frequent itemset refers to a set of items that frequently co-occur together in the dataset, exceeding a predefined minimum support threshold. The required parameters are the minimum itemset size which indicates the minimum number of items in an itemset to be considered and the minimum support which is the threshold for an itemset to be considered frequent where support is given by

$$s\{X \rightarrow Y\} = \frac{\sigma(X \cup Y)}{N}$$

and corresponds to the percentage of the total transactions (football match events) that contain a specific itemset. In the case of the proposed analysis, the choice was to set the minimum set size to 3, indicating that we are interested in finding associations between groups of at least three aspects related to the same offensive play. This helps ensure that the discovered itemsets are meaningful and not based on chance occurrences of individual events. Moreover, the minimum support has been set to 0.20, indicating that an itemset must occur in at least 20% of the total match events. In addition to the frequent itemsets, other to types of itemsets are considered for he analysis:

- Closed Frequent Itemset: it is a frequent itemset that does not have any superset with the same support count.
- Maximal Frequent Itemset: it is a frequent itemset that is not a proper subset of any other frequent itemset. It represents the largest possible itemset that satisfies the minimum support requirement.

These itemsets are achieved by means of a set of Itemset Finder Nodes using the Apriori algorithm [4], which is known for its efficiency in mining frequent itemsets.

4.2 Association Rules

Association rules are a fundamental technique used in association analysis, a machine learning method that seeks to uncover interesting relationships or patterns in large datasets. Association rules consist of two components: an antecedent, which represents the set of events or conditions that act as a precursor, and a consequent, which represents the event that is predicted or implied by the antecedent.

One crucial measure associated with association rules is confidence. Confidence quantifies the predictive strength of a rule and is given by:

$$c\{X \rightarrow Y\} = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

The learning of the association rules is performed through the Association Rule Learner (Borgeit) on the data including all the offensive plays. The aim of the analysis is to highlight only the rules including events with goals. Hence, the chosen parameters are 4 for the minimum set size (increased by 1 due to the inclusion of the consequent in the count); the chosen minimum support to 0.01 implying that a rule should appear in the dataset at least 2% of the time. It may appear like a low value, but since the goals belong to a small percentage of the total events, this value seems appropriate. Moreover, the chosen level of confidence is 0.4.

5 Rules Evaluation

For association analysis, it is crucial to evaluate the importance and significance of the discovered association rules. This evaluation allows us to identify the rules that have the most relevance and impact in the context of goal scoring. To achieve this, we employ a rules evaluation criterion or measure. In the proposed work, we have chosen lift as the primary evaluation criterion. Lift is a widely adopted evaluation criterion in association rule mining and is given by

$$Lift = \frac{c(A \rightarrow B)}{s(B)}$$

Lift is a widely adopted evaluation criterion in association rule mining, and we have selected it as our primary measure for several because it is able to measure the strength of the association between the antecedent and consequent items.

6 Results

Frequent, closed frequent and maximal frequent itemset finders discovered on goals data all discovered the same sets and are showed in Fig.2

Itemset	Support %
Pass, Open Play, Foot	32.2
Center of the Box, Open Play, Foot	30.0
Bottom Left, Open Play, Foot	20.3

Table 2: Frequent, Maximal Frequent and Closed Frequent Itemsets

Event	Coefficient	SE	P>|z|
Free Kick	0.91	0.17	0
Open Play	2.96	0.11	0
Set Piece	0.42	0.137	0.10
Constant	5.82	1.87	0.002

Table 3: Most relevant Logistic regression coefficients

It signifies a strong association between events involving a open play, and the use of the foot. Given that passing is a fundamental aspect of football, it is not surprising to see this item appearing frequently in the data. The high support value suggests that this combination of event attributes is a common occurrence during matches. Moreover, the center of the box is a crucial area in football, often leading to goal-scoring opportunities.

On the other hand, the association rules having goal as consequent item are derived from all the offensive plays and are showed in Tab.1. All of the selected rules are noteworthy as they have a lift score higher than 4.46, indicating a strong positive association with scoring a goal. In particular, the first association rules suggests that the 47% of the set piece plays in which there is no assist and the ball is shot by foot, a goal is scored. The interpretation for this rule is that it is likely to score when a play starts from a set piece and there is no assist. The situations in which this conditions might happen are multiple. For example, these combination of events might happen when goals are scored from free kicks or when a goal taps in the ball after the goalkeeper pushes it away. Moreover, another important rule that emerges from the data is that the important way a secure way to score during an open play, is to shoot the ball from the center of the box towards the bottom right or left angles of the goal where the confidence is respectively 54.10% for the bottom left and 48.70% for the bottom right.

6.1 Logistic Regression

In addition to the association analysis, logistic regression can be employed to further validate the results obtained. By analyzing the coefficients of the logistic regression model, we can gain insights into the significance and impact of different event attributes on goal-scoring. The goal-scoring event was considered as the outcome variable, while the event attributes identified in the association analysis, were used as predictors. Table 3 reports the most relevant coefficients obtained by the model.

Open Play generally has a substantial positive impact on the likelihood of scoring a goal and the multiplicative effect it has on the odds of scoring is bigger with respect to the other coefficients concerning free kicks and set pieces. Additionally, the constant term with the attribute combination of, between the others, foot shot from the center box has a coefficient of 6.0, with a standard error of 1.96 and the p-value is statistically significant. This constant term represents a baseline effect, capturing the influence of the Foot attribute and events occurring at the center box. The high coefficient value suggests that the aforementioned combination of events significantly contributes to the probability of scoring a goal.

7 Conclusions

In this study, we conducted an association analysis and utilized logistic regression to explore the relationship between various event attributes and goal-scoring in football. The results obtained from both analyses provide valuable insights into the factors influencing goal-scoring outcomes. The association analysis revealed several interesting patterns and association rules. Specifically, events involving free kicks and shots from the center of the box during open plays were found to be positively associated with goal-scoring. These findings highlight the importance of set pieces, targeted play in specific areas of the field, and the execution of foot-based actions in creating goal-scoring opportunities. To further validate the results, logistic regression was employed, and the coefficients of the model were analyzed. The logistic regression analysis reinforced the findings from the association analysis, strengthening our understanding of the impact of event attributes on goal-scoring. It is important to note that this study is based on the analysis of football events data and the interpretation of association rules and logistic regression coefficients. While the results provide meaningful insights, other factors such as team dynamics, player abilities, and tactical considerations should also be taken into account when formulating comprehensive strategies. Other information concerning the considered events such the ball and

	Antecedent		Consequent	Support %	Confidence %	Lift
1	Set Piece, No assist, Foot	→	Goal	1.02	47.40	4.56
2	Bottom Left,No assist, Foot	→	Goal	1.00	46.0	4.46
3	Bottom Left, Center of the box, Open Play	→	Goal	1.20	54.10	5.20
4	Bottom Right, Center of the box, Open Play	→	Goal	1.20	48.70	4.6

Table 1: Association rules having goals as consequent item

the player's, the use of the strong or weak foot or perhaps even the type of grass of the field could be valuable information that would have increased the reliability of the analysis. The results of this kind of analysis could offer valuable insights that can inform decision-making processes related to tactics, player positioning, and set-piece plays. Future research can further explore these findings and expand the analysis to encompass additional factors that contribute to goal-scoring in football.

A Appendix

In order to make the selection of certain more clear variables used for the analysis, the complete list of attributes selected from the original dataset is shown below

A.1 ORIGINAL VARIABLES

- id osp = unique identifier of game (odsp stands from oddsportal.com).
- id event = unique identifier of event (id odsp + sort order).
- sort order = chronological sequence of events in a game.
- time = minute of the game.
- text = text commentary.
- event type = primary event. 11 unique events. We have considered only the "attempt"'s event.
- event type 2 = secondary event. 4 unique events.
- side = if the team involved plays at home or away.
- event team = team that produced the event. In case of Own goals, event team is the team that beneficiated from the own goal.
- opponent = team that the event happened against.
- player = name of the player involved in main event.
- player 2 = name of player involved in secondary event.
- player in = player that came in (only applies to substitutions).
- player out = player substituted (only applies to substitutions).
- shot place = placement of the shot (13 possible placement locations).
- shot outcome = the result of a shot. 4 possible outcomes.
- is goal = binary variable if the shot resulted in a goal (own goals included).
- location = location on the pitch where the event happened (19 possible locations).
- bodypart = the part of body used by a player to make a shot. Three possible outcomes.
- assist method = in case of an assisted shot, 5 possible assist methods.
- situation = the start of the action, 4 possible outcomes.

A.2 DICTIONARY OF USED VARIABLES

- Shot place: Bottom left - Bottom right - Center - Top left - Top right.
- Location: Center box - Diff angle long range - Diff angle left - Diff angle right - Left box - Left six y box - Right box - Right six y box - Close range - Penalty spot - Long range - 35+ yards - 40+ yards - Unknown - Outside box.
- Bodypart: Right foot - Left foot - head.
- Assist method: None - Pass - Cross - Head pass - Through ball.
- Situation: Open play - Set piece - Corner - Free kick.

References

[1] Process Mining of Football Event Data: A Novel Approach for Tactical Insights Into the Game - P. Kröckel and F. Bodendorf

[2] https://www.kaggle.com/datasets/secareanualin/football-events, Football Events

[3] https://www.knime.com/

[4] Fast Algorithms for Mining Association Rules - R.Agrawal, R.Srikant