Pervasive Decision Support to predict football corners and goals by means of data mining

João Gomes[1], Filipe Portela[1,2], Manuel F. Santos[1]

Algoritmi Research Centre, University of Minho, Portugal
[2]ESEIG, Porto Polytechnic, Porto, Portugal
joaogomes0991@gmail.com; {cfp, mfs}@dsi.uminho.pt;

Abstract. Football is considered nowadays one of the most popular sports. In the betting world, it has acquired an outstanding position, which moves millions of euros during the period of a single football match. The lack of profitability of football betting users has been stressed as a problem. This lack gave origin to this research proposal, which it is going to analyse the possibility of existing a way to support the users to increase their profits on their bets. Data mining models were induced with the purpose of supporting the gamblers to increase their profits in the medium/long term. Being conscience that the models can fail, the results achieved by four of the seven targets in the models are encouraging and suggest that the system can help to increase the profits. All defined targets have two possible classes to predict, for example, if there are more or less than 7.5 corners in a single game. The data mining models of the targets, more or less than 7.5 corners, 8.5 corners, 1.5 goals and 3.5 goals achieved the pre-defined thresholds. The models were implemented in a prototype, which it is a pervasive decision support system. This system was developed with the purpose to be an interface for any user, both for an expert user as to a user who has no knowledge in football games.

Keywords. Data Mining, Bets, Pervasive Decision Support, Football, Corners, Goals

1 Introduction

Betting on sporting events these days is a fashionable activity. The sport that arouses more interest and it has more fans in the world is football. There are several bookmakers such as is Betfair, Bet365, and Bwin that allow you to perform a wide range of betting, and if you can bet on the outcome, you can bet on the number of goals, number of corners, etc. The number of bookmakers have grown greatly in recent years, leading to the conclusion that this is a profitable business for them at the expense of its users. This project appears with the aim of increasing the betters' profits.

This project is focused on the induction of data mining models. After evaluation these models a Pervasive Decision Support System prototype was implemented. This project distinguishes itself from other platforms due the use of several Data Mining techniques. This article focuses on the release of the first results obtained in the forecast number of goals and corners in the 2013/2014 English Premier League season.

The methodology used to develop this project was the Design Science Research. This methodology is applied when the goal is to develop technology-based solutions to important and relevant business problems [1].

The best models achieved an accuracy between 78% and 82% to predict 7.5 and 8.5 corners and 1.5 and 3.5 goals.

The article is divided into six sections. The first section contains a brief introduction of the project. In the second section is presented a bibliography review. In the third section is presented the methodology used to develop the project. In the fourth section is displayed the development of the completely practical work. In the fifth section is conducted a discussion of the obtained results in the realized tests to the prototype and in the last section is presented the conclusions and suggestions for future work.

2 Background

2.1 Knowledge Discovery, Data Mining, Decision Support and Pervasive Data

Knowledge Discover in Database (KDD) is a modelling and automatic exploratory analysis of large data repositories. It is an organized process that aims to identify useful patterns, which can be understandable, in large and complex dataset [2]. It is an interactive and iterative process where interaction of a responsible for making decisions is required at various stages [3]. The basic framework is divided in five main steps: Selection, Pre-processing, Transformation, Data Mining and Evaluation [4].

Data Mining is the process of discovering patterns and interesting knowledge in large amounts of data [5]. It is considered a key process to any organization [6]. DM contains technical activities that can be subdivided into two major focuses of research, according to the analysis to be achieved, it can be interpretive or predictive analysis [4].

Decision Support Systems (DSS) can be described as an interactive computer system that supports managers to make decisions related to attributes, goals and objectives, to solve semi-structured and unstructured problems [7]. The purpose is giving support to problems solving them by following the development stages of the decision-making process [8]. Simon [9] defines the decision-making process as having only three phases, Intelligence, Design and Choice. Years later, Simon [10] and many other authors defined a fourth phase, Implementation and a fifth phase Monitoring.

Pervasive computing focus on taking the technology from centre stage to the "background" [11], abstracting the user from its complexities. In order to bring the technology to the background a characteristic named "invisibility" is necessary. This concept means that technology is used unconsciously, removing the need for adaptation or understanding of how to utilize it.

Pervasive Data is the possibility of putting the knowledge achieved by means of Artificial Intelligence techniques (e.g. Data Mining) available anywhere and anytime, running in background being the process totally hidden to the user.

2.2 Football Gambling Support Systems and related work

The activity "bet" is an industry that is expanding. There are more and more bookmakers. The activity focus occurs online, each online betting company has their own odds and betting exchanges. Bets on football matches are the most common. Being the result the bet that moves more money. The bookmakers tend to innovate and other markets have emerged, such as the number of corners and number of goals. This is an interesting area to develop research works. However, it is very difficult to control the game variables. A little change in a game can modify the bet result. Due to this fact, the number of gamblers with big winners is lower. The idea of earning money by making bets is a very interesting subject, but at same time it is very dangerous (the gambler can lose a lot of money).

For that reason, there are several suggesting system on this area. There are some web platforms with the same goals. However, they are not using DM techniques. There are also mobile platforms using mathematical calculations which is the case of applications, “KickOff”, “Smart BET Prediction “and “FootWin”.

Some scientific studies were conducted in this area. Owramipur et al [12] intend to make the prediction of the results of the Barcelona games in the 2008/2009 season. Joseph et al [13] Effected identical to the previous work, but the team under study was Tottenham. Rotshtein et al [14] effected a study that aims to predict the results of the Finnish League. Tsakonas & Dounias [15] through its study were intended predict the results of the Ukrainian league and what would be the winner of the championship. Nunes & Sousa [16] created a model that predicts the results for the Portuguese league. Ulmer & Fernandez [17] aimed to make the prediction of the English Premier League results. Hucaljuk & Rakipovic [18] did a study in order to predict the results of games in the Champions League. And finally Suzuki et al [19] did a job that has the objective of predicting the outcome of the 2006 world championship.

3 Pervasive Intelligent Decision Support System

3.1 Phase 1

The main purpose of this phase was to identify the problem or opportunity that could be exploited. In this case has emerged one opportunity, to support the gamblers in football games on the decision about which it is the bet more "safe" to carry out in a certain game. This opportunity has been identified after check the increase of the number of bookmakers in the last decade. This reality shows that it is a profitable market for the bookmakers, and therefore detrimental to its users, being in some cases the user profit equal to null. To explore the context a research was carried out in order to understand the business and the environment by gathering information about the business.

First, there was an effort in finding an open-access database containing a high number of statistical variables associated to football matches. After a depth research a database were found (http://www.football-data.co.uk/). This database contains a relevant number of football games variables. Some other variables can be used (e.g.

rest time, players ratings) however there is not a database containing this information with the same detail and frequency.

3.2 Phase 2

After a review of the existing information related the statistical data related to football games the dataset was created using the data found on the website "football-data-co.uk", the dataset only contains continuous records between 2000 and 2014 football games involving 41 distinct teams, the variables related to the half-time were not considered. The variables contained in that dataset are Match Date (dd/mm/yy); Home Team; Away Team, Full Time and Half Time Result (H=Home Win, D=Draw, A=Away Win); Crowd Attendance; Name of Match Referee and Betting odds data from several bookmakers. For each team (home and away) the dataset contains: Full Time Goals; Half Time Goals; Team Shots; Shots on Target; Hit Woodwork; Team Corners; Fouls Committed; Offsides; Yellow Cards; Red Cards. This dataset contains information from 5320 games.

After the data are collected, a treatment and data processing phase was executed. For this, it was used an Extract Transforming and Loading (ETL) process which it is presented in the Fig. 1.

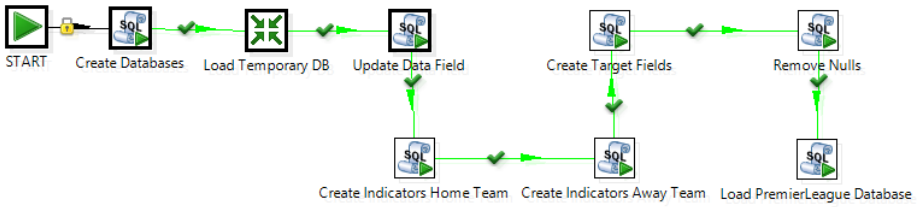

Fig. 1. ETL Process

The “PremierLeague” database was fully treated and it only contains the data that will be needed to the DM models (variables present in the final table). This variables are, “season”, “day, “month, “year”, “home team (HT)”, “away team(AT)”, and it is also composed by several variables one for each team (Home and Away): “AVG Goals”, “AVG Goals Conceded”, “AVG Shots”, “AVG Shots Conceded”, “AVG Shots Target”, AVG Shots Target Conceded”, “AVG Corners” and “AVG Corners Conceded”.

Also in this phase, the Data Mining (DM) models were induced. The models used the data previously processed, which are on the table "PremierLeague", through four distinct DM techniques: Naive Bayes (NB), the Support Vector Machines (SVM), the Decision Trees (DT) and Lazy Learning (LL). To apply these techniques in the induction of the models the Weka tool was used. This tool allows running several algorithms like NaiveBayes, LIBSVM, J48 and Kstar, each one of these algorithms were applied for the above techniques respectively. For the development of the models two different sampling methods were used: Holdout Simple (HS) that uses 66% of the data for training and 34% for testing and the sampling method 10-Folds Cross-Validation (10FCV).

The variables loaded in the table "Premier League" were grouped into different groups to define different scenarios. For this, it was necessary to focus on the characteristics and existing processes in each football game. Each of these groups is composed for two sets, one related to the variables associated to the home team and other by the indicators of the away team. The following groups and the respective variables are Attack (AVG Goals, AVG Shots, AVG Shots Target, and AVG Corners) and Defence (AVG Goals Conceded, AVG Shots Conceded, AVG Shots Target Conceded, and AVG Corners Conceded)

Then it was necessary to define the scenarios through which the DM models would be induced. Eleven scenarios were defined: SA (All Variables), SC (Attack HT and Attack AT), SD (Defence HT and Defence AT), SE (Attack HT and Defence AT), SF (Defence HT and Attack AT), SL (SD+ SE), SO (AVG Corners HT, AVG Corners AT, AVG Corners Conceded HT and AVG Corners Conceded AT), SP (AVG Corners HT and AVG Corners AT), SQ (SC+SD), SR (AVG Goals HT, AVG Goals AT, AVG Goals Conceded HT and AVG Goals Conceded AT) and SS (AVG Goals HT and AVG Goals AT). Therefore, the DM Models (DMM) are composed by:

- Eleven scenarios (SA, SC,…,SS);
- Two sampling methods: 10FCV and HS;
- Four DM techniques: NB, DT, SVM and LL;
- Seven Targets: More or less than “7,5C”, “8,5C”, “9,5C”, “10,5C” corners and more or less than, “1,5G”, “2,5G” and “3,5G” goals.

Initially 416 models were induced. 192 related to the target variables related with the number of goals, the “G1,5”, “G2,5” and ”G3,5” variables and the remaining 224 models were related with the number of corners, “C7,5”, “C8,5”, “C9,5” and “C10,5”. Then for target variables "C7,5", "C8,5", "C9,5", "G1,5 and “G3,5” (which had unbalanced values of the number of instances of each class) 24 more models were induced for each, using the oversampling technique. For each target attribute a total of 121 new models were induced. In total, 537 DM models were induced.

To oversampling an existing function in WEKA was executed, namely SMOTE. This function doubles the number of class instances that contains fewer occurrences and it can be used repeatedly until the classes contain a number of similar occurrences.

A DMM can be represented by the following tuple:

$$DMM = <\Delta, \alpha, DMT, DMSM, DMTG, SCENVAR> \quad (1)$$

Where Δ is the DM rules, α is the DM model configuration, DMSM is the sampling method, DMT is the DM technique, DMTG is the target and SCENVAR are the variables that can be used by each scenario (SA-SS)

For example, if the model chosen is composed by the scenario SO, using as sampling method CV, as DM technique DT and the target which is intended to predict was “3,5G” this tuple can be represented as:

$$DMM=<\Delta, \alpha, DT, CV, \text{“3,5G”}, HomeTeam, AwayTeam, AVGCornersCHT, AVGCornersConcededHT, AVGCornersAT, AVGCornersConcededAT> \quad (2)$$

3.3 Phase 3

The third phase of the project is the combination of three distinct phases of the methodologies used for the development of this project. The phase "Evaluation" of the CRISP-DM and the phase "Choice" of the decision-making process. The evaluation of the DM models induced was made in this phase in order to choose which it is the best model to be used. To evaluate all the induced models, the metrics contained in the confusion matrix were used

Using the confusion matrix several metrics can be calculated such as sensitivity, specificity, accuracy and area under curve (AUC). So, to evaluate all the DM models induced four metrics were used.

For each target a set of thresholds were defined to ensure the quality of the models and at the same time facilitate the choice of the best model for each target. If the models do not meet the parameters, it is possible to conclude that the models do not have the quality required to support gamblers in that particular bet. Based on the performed literature review and as it was not possible to contact an expert in football games betting in order to understand what would be the thresholds that models should achieve, the quality parameters were defined with a minimum value of 65% in metrics accuracy, specificity, AUC and sensitivity. Accuracy was considered the most relevant metric to perform the evaluation of the induced models.

In the Table 1 are present the best models obtained for each previously defined target. In the table are only the targets that meet all thresholds

Table 1. Best DM Models (percentage)

Target	DMSM	Scenario	DMT	Specificity	Sensitivity	Accuracy	AUC
7,5C	HS	SO	LL	89,16	71,34	80,99	0,90
8,5C	10FCV	SO	LL	87,89	68,40	78,32	0,89
1,5G	10FCV	SQ	LL	90,08	71,87	81,65	0,91
3,5G	10FCV	SQ	LL	90,08	71,87	81,65	0,91

These four models were obtained after application of the oversampling technique to the dataset. All models can be represented by an expression, for example, to the target “7,5C” the expression is:

$$\text{DMM}=<\Delta,\alpha,LL,10FCV,'C7{,}5',\text{HomeTeam},\text{AwayTeam},\text{AVGCornersHT},\text{AVG CornersCOncededHT},\text{AVGCornersAT},\text{AVGCornersConcededAT}> \quad (3)$$

3.4 Phase 4

The fourth phase is composed by the combination of two phases, the "Development" phase of CRISP-DM and "Implementation" phase of the decision-making process. In this phase, the development of the prototype was initialized. This prototype allows the user to make intelligent predictions of various events in real time at football games anywhere and anytime [20], [21], being the system designed following some of pervasive features as is scalability, context awareness and ubiquity.

It was decided to create a web platform, because this allows easy access from any location in different devices. In the Fig. 2 is presented the architecture by which the prototype can be represented.

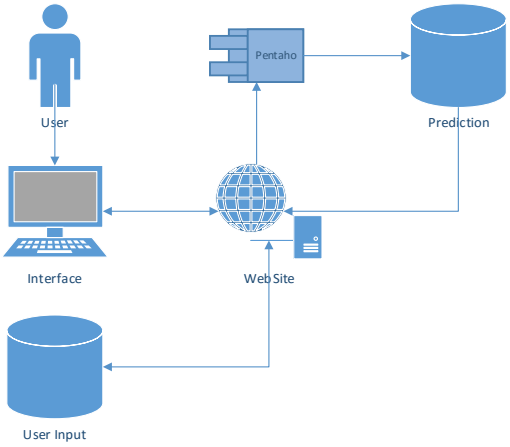

Fig. 2. Prototype architecture

To use the prototype, the user starts by entering information necessary for the system know of which the game is intended to make a prediction. This information is stored in a database that is shown in Fig. 2 by the "User Input". The platform will then use this information to make a request through a *.bat* file that automatically starts the process designed in Pentaho tool.

After starting the job in Pentaho, the information previously entered by the user is used by the model DM previously created, "Weka Scoring", which is an existing process of Pentaho, to generate a prediction that it is stored in the database "Prediction".

The generated prediction is then sent to the web platform to be used by the better. In Fig. 3 is the prototype main menu.

Fig. 3. Prototype

In this prototype, the user starts by selecting one of two groups of predictions, a group that includes the predictions related to the number of corners and another with the number of goals. Clicking, for example, in the "Number of Goals" button will emerge two new buttons, the "More or Less than 1.5 Goals" and "More or Less than 3.5 Goals". If you click on one of them emerges the form that the user needs to complete to pass the information to the DM model. After the user fill form it is submitted, automatically and the prediction is presented for the user. The predictions are presented as the possible result and the probability of it occurs. For example, the output can be: *There is 95% probabilities of the number of goals be "More or Less than 1.5 Goals".*

5 Discussion

To induce the models, for all targets two sampling methods were used, the 10-Folds Cross-Validation (10FCV) and Simple Holdout (HS). Four Data Mining (DM) techniques were also explored: Naive Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT) and Lazy Learning (LL).

The targets that have unbalanced classes, for example, the target "More or less than 1.5 goals" (G1,5) have 74% examples of more than 1.5 goals, which means that there is an imbalance in the model. In these situations, the oversampling technique was used to balance the dataset records.

For each target defined related to the number of corners, 56 DM models were initially induced and in the case of targets associated with the number of goals 64 DM models were induced, also for each target.

The values obtained in metrics do not have a significant variation associated to the sampling method and the DM technique used in the induction of DM models.

The first results obtained in the metrics in each target were weak and did not meet the defined quality parameters. It was then applied the oversampling technique into the dataset to balance the classes of each target. The metric values obtained in the models have substantially improved after the application of this technique having four targets that hit the defined quality parameters, "C8.5", the "C9.5", the "1.5G" and "3.5G". The DM technique that stands out was the LL, with this technique the models obtained best values in the evaluated metrics as can be observed in table 1.

6 Conclusion and Future Work

The objective of this project was to obtain predictive models that will support gamblers to increase their profits. In particular when they are betting on the number of goals or number of corners in a specific football match in order to reduce the risks that have on each placed bet.

Several targets were defined within these two groups. After the DM models were induced, they were evaluated according to the defined thresholds, the models that have value to be entered in the prototype are "C8,5", "C9,5", "G1,5" and "G3,5". These models were obtained after the application of the oversampling technique to the dataset; this technique has substantially improved the values obtained in the evaluated metrics.

Future work will pass for adding new variables to these models, to try different scenarios in order to obtain models with even greater precision to be added later to the prototype. In parallel, the prototype will be converted into a system able to disseminate all the probabilities anywhere and anytime in mobile or situated devices. This prototype also will incorporate the other predictions made in this field related to the final result [22, 23, 24].

Acknowledgments

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

References

1. A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, (2004).
2. L. Maimon, Oded; Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. (2010).
3. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining : Towards a Unifying Framework," *Kdd*, (1996).
4. C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. (2009).
5. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
6. E. Turban, R. Sharda, and J. Aronson, "Business intelligence: a managerial approach," *Tamu-Commerce.Edu*. (2008).
7. H. R. Nemati, D. M. Steiger, L. S. Iyer, and R. T. Herschel, "Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing," *Decis. Support Syst.*, vol. 33, pp. 143–161, (2002).
8. J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, present, and future of decision support technology," *Decis. Support Syst.*, vol. 33, pp. 111–126, (2002).
9. H. A. Simon, *The New Science of Management Decision*. (1960).
10. H. a. Simon, *The new science of management*. (1977).
11. M. Weiser, "The Computer for the 21st Century," *Scientific American*, vol. 265, no. 3. pp. 94–104, (199)1.
12. F. Owramipur, P. Eskandarian, and F. S. Mozneb, "Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team," *Int. J. Comput. Theory Eng.*, vol. 5, no. 5, pp. 812–815, (2013).
13. a. Joseph, N. E. Fenton, and M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," *Knowledge-Based Syst.*, vol. 19, no. 7, pp. 544–553, (2006).
14. A. P. Rotshtein, M. Posner, A. B. Rakityanskaya, M. Lev, and V. National, "Football predictions based on a fuzzy model with genetic and neural tuning," *Cybern. Syst. Anal.*, vol. 41, no. 4, pp. 619–630, (2005).
15. a Tsakonas and G. Dounias, "Soft computing-based result prediction of football games," *First Int. ...*, vol. 3, no. May, pp. 15–21, (2002).
16. S. Nunes and M. Sousa, "Applying data mining techniques to football data from European

championships," *Actas da 1ª Conferência Metodol. Investig. Científica*, no. December 2005, (2006).
17. B. Ulmer and M. Fernandez, "Predicting Soccer Match Results in the English Premier League," p. 5, (2013).
18. J. Hucaljuk and A. Rakipovic, "Predicting football scores using machine learning techniques," *2011 Proc. 34th Int. Conv. MIPRO*, vol. 48, pp. 1623–1627, (2011).
19. a K. Suzuki, L. E. B. Salasar, J. G. Leite, and F. Louzada-Neto, "A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup," *J. Oper. Res. Soc.*, vol. 61, no. October 2015, pp. 1530–1539, (2010).
20. Filipe Portela, Manuel Filipe Santos, Pedro Gago, Álvaro Silva, Fernando Rua, António Abelha, José Machado and José Neves. Enabling Real-time Intelligent Decision Support in Intensive Care. ESM 2011 - 25th European Simulation and Modelling Conference. Guimarães, Portugal. EUROSIS. (2011).
21. Portela, F., Santos, M. F., Silva, Á., Machado, J., & Abelha, A.. Enabling a Pervasive approach for Intelligent Decision Support in Intensive Care. Communications in Computer and Information Science - ENTERprise Information Systems. Volume 221, Part 4, pp 233-243. ISBN: 978-3-642-24351-5. Springer. (2011).
22. João Gomes, Filipe Portela, Manuel Filipe Santos, José Machado, António Abelha. Predicting 2-way Football Results by means of Data Mining. ESM - 29th European Simulation and Modelling Conference. Leicester, UK. EUROSIS. (2015). (accepted for publication).
23. João Gomes, Filipe Portela, Manuel Filipe Santos. Decision Support System for predicting Football Game result. Computers - 19th International Conference on Circuits, Systems, Communications and Computers - Intelligent Systems and Applications Special Sessions. Series 32, 2015, pp 348-353. ISBN: 978-1-61804-320-7. INASE. (2015).
24. João Gomes, Filipe Portela and Manuel Filipe Santos. Real-Time Data Mining Models to Predict Football 2-Way Result. Jurnal Teknologi. Penerbit UTM Press. (2016). (accepted for publication).