Decision Support System for predicting Football Game result

João Gomes, Filipe Portela, Manuel Filipe Santos

***Abstract* —** there is an increase of bookmaker's number over the last decade, leading to the conclusion that the bet houses have obtained profitability in the detriment of its users. Based in this principle arises an opportunity to explore a set of artificial intelligence techniques in order to support the user betting decision. The development of this project aims to support bookmaker's users to increase their profits on bets related to football matches, suggesting to them which bet that they should carry out (home win, draw or away win). To this, it was collected several statistical information related to football games from the Premier League. It was developed a dataset and applied data mining techniques to create a model with good predictive capability. This model was then integrated in a decision support system which allows complement the machine intelligence with human perception. The model developed allowed to have profits of 20% in relation to an initial bankroll.

***Keywords*—** Decision Support Systems, Data Mining, Football Games Prediction, Decision Support Systems for Football Betting, Knowledge Discovery in Database, Football Bets

I. Introduction

THIS paper presents the first step of a project that is being developed in the area of Decision Support Systems using Data Mining applied to football betting.

In the last few years a growing number of bookmakers is been observed in this industry, leading to the conclusion that this area presents itself as a profitable activity for the bookmakers themselves and consequently unfavorable for its users.

The project consists in analyzing football games statistics, trying to identify patterns used in these data and then leading to a suggestion for which bet should be held in a given game.

The goals of this project are divided in two groups:

- Get models with good predictive capabilities, aimed to support the users decision in a way that they can perform the best bet in a particular football game;
- Develop a prototype of an Decision Support System that incorporates the developed models.

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

João Gomes is with Information System Department, University of Minho, Portugal.

Filipe Portela is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

The first one can be divided into the following steps:

- Collect statistical data of football games (number of goals, number of shots, etc...);
- Make a treatment of data collected;
- Create predicting data mining models;
- Make an assessment of the metrics models;
- Choose the model that satisfies the work requirements.

The second objective can be divided into:

- Create a prototype of a Decision Support System;
- Integrating the forecasting models previously created in the prototype.

After this work it was possible to find a path in order to achieve the defined goals. In this first step logical blocks were developed combined with data mining models in order to predict the better bet. In terms of model assessment, the results were not totally satisfactory being notorious the need of adding other variables and continuing the work. Although the obtained results were not the expected, it was possible to obtain models that are able to achieve good profits. By making a model evaluation, it was also possible to obtain upper than 50% correct results (more than 1/3 of the possibilities (home win, draw or away win) and profits of 20%. The achieved results gave enormous confidence to proceed with this type of research.

This paper is divided into seven topics. The first topic presents itself as an introduction to the project made. The second is called "Background" and focuses on the structuration of the theoretical context, in such way that we can ease the understanding of the project. The third topic is called "Methodologies" and in this chapter is made a description of the methodological environment concerning the development of the carried out project. The following topic is "Intelligent Decision Support System" and describes the practical work developed. The fifth topic is titled "Discussion" and it shows the results obtained through tests made evaluating the performance of the system. In the sixth topic is stated a conclusion of the theme. Finally, the last topic discusses the work that is needed to carry out in the future.

II. Background

A. Knowledge Discovery in Data

Knowledge Discovery in Database (KDD) is described as the process of identifying and understanding incomprehensible patterns in datasets, being these patterns valuable novelties, and potentially useful [1]. This is an organized process. It is

performed automatically with an easy exploratory analysis and modeling, occurring in a given data repository [2]. KDD is an iterative process, due the fact that it may be necessary to go back one or more steps to get the results that best suit project requirements [1]. The process is composed by nine steps that can be compressed in the next five [3]:

- Selection: In this first phase a dataset selection is made to be used and sometimes it is still necessary to create a new dataset. In this phase the objectives to be achieved are also defined and it is chosen what data will be used throughout the process;
- Pre-processing: this is the phase in which is performed a cleaning and pre-processing of data;
- Transformation: At this stage the data transformation is performed in order to improve the dataset quality in order to facilitate finding data dependencies;
- Data Mining: This topic will be addressed in the topic II.B;
- Evaluation: this is the phase in which is performed an interpretation of the found patterns at the previous phase, the models are evaluated and then it is found a model that meets the objectives defined in the first phase [2].

B. Data Mining

Turban *et. al* [4] stated that the use of Data Mining (DM) from databases would become an activity that would be fundamental for organizations in the near future. This will be an important technique, so the organizations cannot waste any information regarding to their business and customers, having the risk of being overcome by its competitors.

DM is a process aiming to find patterns and interesting information in a given dataset [5].

This process was initially defined by Turban *et. al* [4] as a pattern discovery process, later they improved the concept and define DM as a process that can also be used as a way to analyze the data with the objective of increasing the efficiency and effectiveness of organizations.

The DM contains activities that can be divided into two dimensions, forecasting activities and interpretation activities [6]. In the case of this project the goal is to predict the outcome in a football game, so it will be using classification techniques.

The objective of this task was to predict the value that any random variable will assume or else estimate the probability of a future event occurs. Estimating therefore the value of a variable called 'dependent', 'target', 'response' [7].

Depending on the target class it is intended to provide the prediction of an activity. It can divided them into two specific groups: classification - if what you want is to predict the label of a class, for example "Victory", "Draw" or "Defeat", or regression if you what to want is to predict a real number "0", "1" or "2" [8].

C. Decision Support Systems

Decision Support Systems (DSS) are like an interactive computer system able to supports managers in the decision process by connecting attributes, goals and objectives, in order to solve the semi-structured and unstructured problems [9]. These are, therefore, systems aiming to support decision makers by providing alternative analysis, studying previously made decisions and what influences these decisions had in an organizational context in order to better support the decision [10].

DSS based their development in the phases of the decision-making process. Herbert Simon [11] is the author of the methodology that gathers a greater consensus among the community. Initially he defines the decision-making process as having only three phases: Intelligence, Design and Choice. Later, together with other authors and gathering consensus they defined a fourth phase Implementation, therefore, argued that the implementation of what had been previously decided was important enough to create an individual stage for the several authors [12].

D. Similar Systems

After making a research on systems that have the objective to predict football matches results, it was possible to verify that there are platforms aiming to support the gambler decision in what will be the best bet to make in a football matches.

The DSS in development by this work has a particular feature not found in any of the existing platforms. The operation of these platforms is based on some mathematical calculations.

From the study made, the following web platforms were found:

- http://soccervista.com/
- http://vitibet.com/
- http://pt.zulubet.com/
- http://www.footwin.net/
- http://www.predictz.com/
- http://www.forebet.com/
- https://www.statarea.com/predictions
- http://www.windrawwin.com/predictions/
- http://www.prosoccer.gr/
- http://spotwin.net/football-betting-system-7.html

There is also applications for smartphones available in the AppStore or PlayStore which have the same purpose of previously described web platforms.

- KickOff – Smart Betting Made Simple
- Smart BET Prediction
- FootWin – Sports Prediction

The system that is being developed will be distinguished from these by the use of DM techniques.

III. Methodologies, Material and Methods

The main project follows the Design Science Research (DSR) Methodology. DSR should lead to the production of a viable artefact in the form of a building, a model, a method, or an instantiation. The main goal is to develop a technology-based solutions able to solve important and relevant business problems [13]. To complement this methodology it was used a combination of two other methodologies, the CRISP-DM and the Simon phases of decision making. For example, the first phase is composed by a combination of the Intelligence phase of decision-making and Business Understanding phase of CRISP-DM methodology. The complete methodology was presented in Table 1.

Table 1 - Combined Methodology

		Combined Methodology - DSR			
		Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
CRISP-DM	Business Understanding	X			
	Data Understanding		X		
	Data Transformation		X		
	Modelling		X		
	Evaluation			X	
	Deployment				X
Decision-making	Intelligence	X			
	Design		X		
	Choice			X	
	Implementation				X

IV. Decision Support System

As previously mentioned this project was developed using a combination of three methodologies, CRISP-DM, the phases of decision-making and the DSR. At this topic is present the work developed in each phase which has been generated by the combination of the three methodologies.

A. Phase 1

In the first step was performed the identification of the problem. After making an analysis of the bookmakers market it was verified an increase of them in recent years. This situation leads to the conclusion that the profit made by them is lucrative, on the opposite only a small percentage of their users get in long-term considerable profit.

Then came the idea of creating a Decision Support System (DSS) that has the ability of help its user to make the best decision at the time it will make a bet on a certain football game, suggesting which bet must be made.

This problem corresponds to a semi-structured decision because it is necessary to complement the existing data (already collected) in the dataset with information possessed by users.

B. Phase 2

In this second phase, the project goals were defined. An analysis was made and the risks and possible restrictions that could exist for its development were determined.

The development of a DSS was made through the *Exsys Corvid* tool that would integrate data mining models developed from the *WEKA* tool, creating a system that combines intelligence machine with human perception.

Also in this phase it was carried out a dataset considered interesting and valuable for the development of the project. After making the dataset collection a detailed analysis (statistics) of it was performed in order to better understand the data that it contains and their distribution.

The data collected are detailed in Table 2. The data are from 13 seasons of the English Premier League until the season 2012/2013, which corresponds to statistical information of 4940 games.

Table 2 - Original Variables

Original Variables	Description
Date	Match Date (dd/mm/yy)
HomeTeam	Home Team
AwayTeam	Away Team
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
FTR	Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result (H=Home Win, D=Draw, A=Away Win)
Attendance	Crowd Attendance
Referee	Name of Match Referee
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HHW	Home Team Hit Woodwork
AHW	Away Team Hit Woodwork
HC	Home Team Corners
AC	Away Team Corners
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HO	Home Team Offsides
AO	Away Team Offsides
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards
ODDS	Betting odds data from several bookmakers

After collecting the data, it was necessary to make the respective analysis. For it, an extract, transform and loading (ETL) process was designed. This process is presented in the Figure 1. In this figure is possible observe all the process since database creation, passing by the prediction phase and concluding with the dataset preparation which will allow its use in *Exsys Corvid* tool.

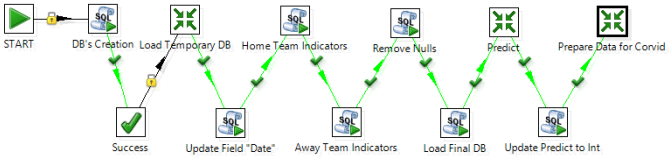

Figure 1 - ETL Process

This ETL is a simple process that begins with the creation of two databases, one temporary database and another one that will be later filled with all the data processed. The original dataset was uploaded to the temporary database.

Then the field “Date" was updated. This field was separated into three other fields, “Day”, “Month” and “Year” in order to understand for example the weekday game.

The data presented in the dataset are mostly information that can only be obtained at the end of each game. Therefore it was necessary to identify other variables that could be known before the game starts. In this step several indicators for the home team

ISBN: 978-1-61804-320-7

and for the away team were created. These indicators (table 3) are averages of the variables presented in the original dataset. Finally and before loading the final database the existing null fields were removed.

Table 3 - New Variables

New Variables	Description
AverageGoalsHomeTeam	This variable stores the goal average of the home team, in the matches disputed at home.
AverageGoalsAwayTeam	This variable stores the goal average of the away team, in the matches disputed out.
AverageShotsHomeTeam	This variable stores the average shots of the home team, in the matches disputed at home.
AverageShotsAwayTeam	This variable stores the average shots of away team, in the matches disputed out.
AverageShotsTargetHomeTeam	This variable stores the average shots on goal the home team, in the matches disputed at home.
AverageShotsTargetAwayTeam	This variable stores the average shots on goal of away team, in the matches disputed out.
HomeWinLastFive	This variable stores the number of victories in the last five games of the home team in home games.
AwayWinLastFive	This variable stores the number of victories in the last 5 games of away team in matches disputed outside.
HomeWinLastFiveConfrontation	This variable stores the number of victories in the last 5 home team's games in confrontation with the away team in home games.
AwayWinLastFiveConfrontation	This variable stores the number of victories in the last 5 games of away team matched up against the home team in matches disputed outside.
Predict	This variable stores the result predicted by the data mining model.

The next step, after loading the final database, was focused in the induction of classification models with the data obtained after them been processed and transformed. The DM models were induced using three different DM techniques: Naive Bayes (NB), Decision Trees (J48) and Support Vector Machine (LibSVM) and two sampling methods, 10-Folds Cross Validation (CV) and Percentage Split where 66% of data was used to carry out training and 33 for testing the models

The data mining models was created and tested using a scenario. The variables used in this scenario were

- Home Team;
- Away Team;
- AverageGoalsHomeTeam;
- AverageGoalsAwayTeam;
- AverageShotsHomeTeam;
- AverageShotsAwayTeam;
- AverageShotsTargetHomeTeam;
- AverageShotsTargetAwayTeam;
- HomeWinLastFive;
- AwayWinLastFive;
- HomeWinLastFiveConfrontation;
- AwayWinLastFiveConfrontation.

The target variable was:

- FTR (Final Time Result).

Then a transformation was made to the field predicted by the models, making it a field containing only numeric data. This transformation was important because only in this way the field was recognized by *Exsys Corvid* tool. Finally the database was extracted to a text file, in order to be imported in *Exsys Corvid*.

C. Phase 3

As result is possible to get three distinct predictions (Home Win- "1", Draw- "2" and AwayWin- "3"). To analyze the models, the accuracy metric provided by Confusion Matrix was used. The six models (1 scenario x 3 techniques x 2 sampling methods) were created using all the data available. In this first phase of the project (exploration phase) it was not defined any extra scenario (combination of different variables).

The obtained confusion matrix through these models has a size of 3x3 as can be seen in the following tables (table 4 to 9). Each table is corresponding to each created model.

Table 4 - Model 1

NB	1	2	3	Accuracy
1	505	188	152	0,597633
2	140	140	169	0,311804
3	105	124	282	0,551859
Accuracy Average				0,487099

Table 5 - Model 2

J48	1	2	3	Accuracy
1	701	52	92	0,829585799
2	249	87	113	0,19376392
3	232	78	201	0,39334638
Accuracy Average				0,472232033

Table 6 - Model 3

LibSVM	1	2	3	Accuracy
1	701	52	78	0,84852071
2	251	87	11	0,249284
3	218	75	218	0,426614481
Accuracy Average				0,50813962

Table 7 - Model 4

NB	1	2	3	Accuracy
1	1477	612	395	0,594605
2	435	476	455	0,348463
3	259	422	779	0,533562
Accuracy Average				0,49221

Table 8 - Model 5

J48	1	2	3	Accuracy
1	2127	145	212	0,85628
2	797	232	337	0,169839
3	666	209	585	0,400685
Accuracy Average				0,475601

Table 9 - Model 6

LibSVM	1	2	3	Accuracy
1	2122	120	242	0,854267
2	788	229	349	0,167643
3	592	203	665	0,455479
Accuracy Average				0,492463

Table 10 presents an overview of the achieved results for each model, the sampling method and algorithm used.

Table 10 - Models Evaluation

Model	Sampling Method	Algorithm	Accuracy
Model 1	Percentage Split	NaiveBayes	0.487
Model 2	Percentage Split	J48	0.472
Model 3	Percentage Split	LibSVM	0.508
Model 4	10-Folds CV	NaiveBayes	0.492
Model 5	10-Folds CV	J48	0.476
Model 6	10-Folds CV	LibSVM	0.492

The accuracy obtained was identical in all models, being model 3 the best. So model 3 was chosen with an accuracy of 50.8%. Even though this value be not high, it is better than the 33% (random probability). This work represents an early stage of the project which leads to believe that it is a great starting point.

D. Phase 4

The goal of this phase was to ensure that what it was defined in the previous phases it is applied. In this phase was performed all the development process from the data collection, data treatment and data transformation. Then the decision support system was developed.

In this task the first prototype of the project turned up. The tool *Exsys Corvid* was used to develop the prototype.

In this phase several system development designs were tested. Below it is presented the "models" which allowed to achieve best results (user profits).

The first step was used to import the data obtained in the ETL process. The variables that the system would use were also defined. Besides the already existing variables, four new variables were defined. Two for each team, because in the moment of decision-making, there are not game data. These data are normally only known a few moments before the games start.

For that the system did the following questions to the user:

- "Which is the present classification of the home team?"
- "Which is the present classification of the away team?"
- "How many holders, from the user's opinion, are unavailable of home team?"
- "How many holders, from the user's opinion, are unavailable of away team?"

Moreover, four logic blocks were created, one was responsible for setting the value of the variables associated with the home team (Figure 2), and the other was responsible for defining the variables associated with away team. Another block was designed to do the calculation of each team score. Finally the last block makes a relationship between the scores of each team in order to suggest the best bet for the user, in this phase the data mining models are induced. It was also created one command block to the system knowing the sequence that must take (Figure 3).

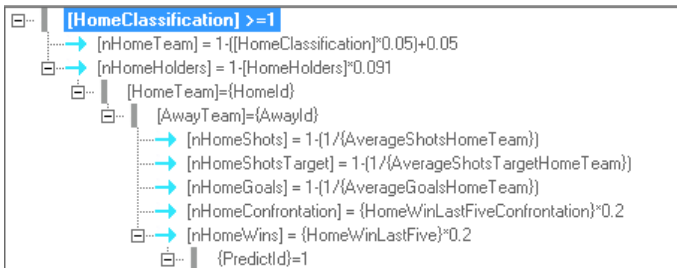

Figure 2 - Logic Block Home

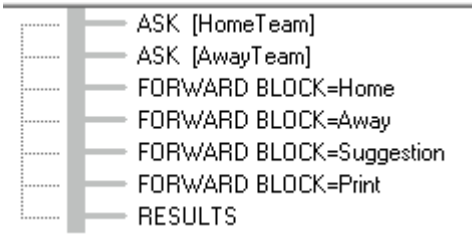

Figure 3 - Command Block

ISBN: 978-1-61804-320-7

V. Discussion

After the system be created, it was necessary to make an assessment of it. Regarding this fact, tests were performed at seven rounds of the English Premier League matches. They were simulated betting of € 100 in each of the 10 games by each round of matches. The obtained results are shown in Table 11.

Table 11 - System Performed Tests

Round	Percentage of Correct Bets	Return (Bets of 100€ by game)
Round 5	80 %	689 €
Round 10	30 %	-418 €
Round 15	40 %	11 €
Round 20	70 %	713 €
Round 25	60 %	480 €
Round 30	40 %	-245 €
Round 35	60 %	179 €
Total	**Average = 54,29%**	**1409 €**

These results shown that even though the model accuracy results are around 50% it was possible to have good profits. In total in this simulation was bet € 7000 and obtained a return of € 1409, about 20%, which is certainly a value to be taken into account.

In the Table 12 is presented in more detail the results obtained in the fifth round of the English Premier League in season 2013/2014.

Table 12 - Return in fifth round

	Game	Result (1,2,3)	Corvid Output (1,2,3)	Return (Bets of 100€)
Round 5	Norwich x A. Villa	3	3	220€
	Liverpool x Southampton	3	1	-100€
	Newcastle x Hull City	3	2	-100€
	West Brom x Sunderland	1	1	110€
	West Ham x Everton	3	3	140€
	Chelsea x Fulham	1	1	33€
	Arsenal x Stoke	1	1	40€
	C. Palace x Swansea	3	3	130€
	Cardiff x Tottenham	3	3	91€
	Man City x Man Utd	1	1	125€
			Total	**689€**

VI. Conclusion

This work is a first step in order to develop an Intelligent Decision Support System to predict football game results. Although the achieved results are not totally satisfactory it was possible to prove the possibility of increasing the profits on football games betting through the use of data mining classification models.

The system used data of fourteen seasons of English Premier League. Although the accuracy obtained was not very good, it was upper to 33% (probability if it was a matter of luck). The obtained profit proves that the system has enough value to continue their research and development. Scientifically this type of models has a high level of evolution and shown to be a good option to the researchers which wants to explore this area.

In conclusion this is a project that has the potential to make the test of creating different data mining models with different types of target class to be able to make the decision support with the highest accuracy possible.

VII. Future Work

In the future the objective is to improve data mining models accuracy by increasing system reliability and consequently obtaining a higher profit. For that it will be considered the following aspects:

- To explore different types of data mining techniques;
- To create a new result variable, instead of there being three possibilities (home team win, draw or away team win), only have a chance to bet in favor of a team or against it, for example;
- To create new variables, as the weather, the players rest time;
- To use other variables which are in original dataset and which have not been used;
- Develop an Intelligent Decision Support System combining rules with the data mining engine.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–53, 1996.

[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] E. Turban, R. Sharda, and J. Aronson, "Business intelligence: a managerial approach," *Tamu-Commerce.Edu*. 2008.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[6] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. 2009.

[7] S. Tuffery, *Data Mining and Statistics for Decision-Making*. 2011.

[8] E. Turban, *Decision Support and Business Intelligence*, vol. 1968. 2010.

[9] 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.

[10] 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.

[11] H. A. Simon, *The New Science of Management Decision*. 1960.

[12] H. a. Simon, *The new science of management*. 1977.

[13] 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.