Recommendations for sports games to bet on

Toon De Pessemier, Bram De Deyn, Kris Vanhecke, Luc Martens
imec - WAVES - Ghent University
{toon.depessemier,bram.dedeyn,kris.vanhecke,luc1.martens}@ugent.be

ABSTRACT

The outcome of sports games, such as football, is non-deterministic since it is subject to many human actions of players and referees, but also injuries, accidents, etc. Betting on the outcome is becoming increasingly popular which is reflected by the growing sports betting market. This research tries to maximize profit from sports betting on football outcomes. Predicting the outcome can be considered as a classification problem (Home team/Draw/Away team). To decide on which games to bet (betting strategy) and the size of the stake of the bet (money management), recommendations can be provided based on personal characteristics (risk taking/risk averse). Profitable ternary classifiers were found for each of the five major European football leagues. Using these classifiers, a personal assistant for bettors was engineered as a recommendation tool. It recommends the betting strategies and money management systems that were the most profitable in recent history and outputs the game outcome probabilities generated by the classifier.

CCS CONCEPTS

• **Information systems** → **Information systems applications**; *Data analytics*; Data mining;

KEYWORDS

Sports betting, Recommendation, Classification, Data mining

ACM Reference Format:
Toon De Pessemier, Bram De Deyn, Kris Vanhecke, Luc Martens. 2018. Recommendations for sports games to bet on. In *Proceedings of ComplexRec 2018 Second Workshop on Recommendation in Complex Scenarios*. ACM, New York, NY, USA, Article 4, 5 pages.

1 INTRODUCTION

Football, also called association football or soccer in some countries, is the most popular sport internationally. But, it is also a sport that can be very difficult to predict because of a whole series of factors that can influence the outcome: current performance and motivation of the 11 players per team on the pitch, and some additional substitutes, decisions made by the players, interactions between players, decisions of coaches and referees, injuries, etc. Therefore, an increasing share of the multi-billion dollar gambling industry is directed to betting on the outcome of football games. Both academical researchers and industrial organizations have a growing interest in the football odds to predict the outcomes thereby profiting from potential market inefficiencies. Most of them focus on football game forecasts and the main objective is often the accuracy of the prediction model, i.e. the fraction of correctly predicted outcomes of football games.

In commercial applications, bookmakers take their share before paying out the winning bets, i.e. the profit margin. In case of a balanced book (e.g. approximately the same amount of money is bet on both outcomes of a fifty-fifty bet) their profit is assured. In case of unbalances, bookmakers might have to pay out more than what was staked in total, or they earn more than was expected. To avoid unbalances, bookmakers allow their odds to dynamically change in proportion to the amount of money staked on the possible outcomes to obtain a more robust book. However, if the bookmakers' odds are significantly deviating from the true event probabilities, these faulty odds provide opportunities to make profit from the bets. Research has shown that the odds of individual bookmakers suffer from these deviations, implying that the gambling market is inefficient [5].

This paper goes further than finding a model with a good accuracy and also considers the profitability of the prediction model (relative to market odds) as an assessment tool . To achieve a profitable model, market odds have to be sufficiently less accurate relative to those generated by the prediction model so that the bookmakers' profit margin can be overcome [7]. The expected profit is calculated based on the discrepancy between the output of the prediction model and the market odds. Only a few research initiatives consider profitability, on top of that, this paper proposes a personal assistant that provides recommendations for betting (which game and which stake), instead of predicting every game.

2 RELATED WORK

Game results and scores have been modeled since the eighties. Maher [15] proposed a Poisson model, in which home and away scores are modeled independently. Lots of improvements on this model have been published since then, such as incorporating time, giving weights to different kinds of scores etc [8]. Besides, the influence of home advantage on the outcome of the game has been proven [7]. Many researchers have investigated features and models to figure out the dependencies between the historic statistics and game results [2, 16, 17]. It has been proven that accuracies above 54% can lead to guaranteed net profit [19], if bettors use an adequate betting method and money management system - assuming that the bookmakers use a moderate profit margin. Neural network, naive Bayes, random forest, and multinomial logistic regression classifiers succeed to achieve accuracies up to 55% [2, 20].

However, it is necessary to consider both accuracy and profit to get the full informative and practical sense of a model's performance [4, 6]. To calculate the possible profit a model could generate, its predicted outcome probabilities have to be compared to the published odds. Bookmakers' published odds have multiple times been shown to be good forecasts for game outcomes [9, 19, 20] and

have been called "the golden odds" for exactly that reason. Studies suggest that bookmakers acquire extra information that exceeds the historical game data available for the public, improving their published odds [6].

Multiple ways exist one can go about betting and choosing the size of the stakes - so called betting strategies and money management systems [1, 13, 14], each with a potential profit and an associated risk. Many studies have been focused on prediction accuracy, thereby neglecting the betting decision users have to make. This paper goes further and proposes how recommendations can assist bettors in their choices: on which game to bet? (betting strategy), how much to bet? (money management), and on which team to bet? (outcome prediction).

3 DATA

Historical data about football games can be retrieved through a sports data provider. For this research the API of SportRadar.com was used. Historical data was fetched for five major national professional European leagues: the Spanish LaLiga, the English Premier League, the Italian Serie A, the German Bundesliga, and the Belgian Pro League. For each league, data were fetched for all games since the 2011/2012 season until the end of 2017 (middle of the 2017-2018 season). Besides statistics about the football games, the data provider also specifies probabilities about the outcomes of games without profit margins. In this research, 109 different features were considered. Most of them are available for both the home playing team (HT) and the away playing team (AT). In addition, some features are specific for the past confrontations of the two teams, also called head-to-head games (H2H).

- *Recent Game Results.* To predict the outcome of a game, a set of obvious features represents the outcome of the most recent game(s) of the teams: win, draw or loss. For each prediction of a game between a specific home team and a specific away team, the most recent games played by the home team (HT Recent Games) as well as the most recent games of the away team (AT Recent Games) are considered. In addition, the most recent confrontations between home and away team are a feature (H2H Recent Games).
- *Goal (Difference).* The difference in goals (scored goals minus against goals) during the most recent games is used to estimate the effectiveness of the team. A strictly positive number means that the considered team won, whereas a strictly negative number indicates a loss. Zero stands for a draw. Large differences in the number of goals reflect large performance differences between the teams. Besides, for each team also the absolute number of scored goals is a feature.
- *Ranking.* The number of points that the team won in the national league during the current season is a measure for its performance (win=3,draw=1,loss=0 points). To compensate for a different number of games played by different teams, the number of points is divided by the number of games played by the team in that league.
- *Fatigue.* Consecutive games might exhaust a team, and cause a poor performance in the next game. The number of games played by the team in the last couple of weeks is used as an indicator for the fatigue of the team. Also the distance that the away team has to travel is used as a feature, since long trips may fatigue the team.
- *Historical game statistics.* Many game statistics of the previous games can be an indicator of a well or poorly performing team. The following were considered: ball possession, free kicks, shots on target, shots off target, shots saved, offsides, yellow cards, yellow-red cards, red cards, corners, successful passes, successful crosses, successful duels, and created chances.

  Many of these statistics (such as recent game results, goal difference, or historical game statistics) can be aggregated over a longer period of time, or aggregated over multiple games to obtain a more reliable value. The results of the 5 most recent and 10 most recent games were considered (older games are considered as less relevant). E.g. a feature can aggregate the amount of goals made by the team during the last 10 games.

4 FEATURE SELECTION

Football games are characterized by a rich set of features, and for each team the past performance is available as the outcome of previous games. An important research question is: Which of these features (based on historical records) are correlated to the outcome of the game that has to be predicted? For feature selection, four algorithms of the WEKA workbench [10, 18] were used: OneR, InfoGain, GainRatio and Correlation. The OneR algorithm assesses the importance of each feature by evaluating the corresponding one feature classifier, and ranking all these classifiers. InfoGain evaluates the worth of a feature by measuring the information gain with respect to the class. GainRatio is similar but measures the information gain ratio. So, both algorithms are evaluating how much a feature reduces the entropy. The Correlation ranker calculates the Pearson correlation between the feature and the result class, i.e. a linear relationship between both is searched.

Table 1 shows the features with the highest information gain according to the InfoRatio algorithm. The results of the other algorithms are consistent. These results show that features derived from the goal differences in the recent past are the most important. In addition, the recent game results of both teams, and the results of the H2H games provide a significant information gain. Also the ranking of both teams in the national league can be used to predict the game result. The absolute number of goals scored by both teams has a lower information value as well as the results of the last H2H games. Noteworthy, none of the historical game statistics and none of the features reflecting the fatigue of the team was found to have a significant information gain.

5 GAME OUTCOME PREDICTION

The goal of the model is to predict which team wins the game (Home team/Draw/Away team). This is tackled as a classification problem with unbalanced classes because of the home advantage [15] (Approximated probabilities based on historical data: 45% Home team, 25% Draw, 30% Away team). While cross validation is considered to be the standard evaluation method, it does not reflect a realistic scenario for sport predictions where the date of the game is important. Cross validation would allow the classifier to find dependencies that cannot be replicated outside the training and evaluation phase. Therefore, a more realistic evaluation approach is adopted by splitting the labeled data thereby remaining the chronological order of the games. An 80% split is used, which means the most recent

Attribute	Information Gain
HT Goal Difference - 10 most recent games	0.0398
H2H Goal Difference - 10 most recent games	0.0379
H2H Goal Difference - 5 most recent games	0.0361
HT Ranking	0.0358
HT Recent Game Results - 10 most recent games	0.0356
AT Goal Difference - 10 most recent games	0.0352
HT Goals Made - 10 most recent games	0.0312
HT Goal Difference - 5 most recent games	0.0310
H2H Recent Game Results - 10 most recent games	0.0303
AT Ranking	0.0297
AT Goals Made - 10 most recent games	0.0293
H2H Recent Game Results - 5 most recent games	0.0284
AT Goal Difference - 5 most recent games	0.0277
AT Recent Game Results - 10 most recent games	0.0275

Table 1: Features with the highest information gain.

20% of the games is predicted with models that were trained on the oldest 80%.

To avoid overfitting, three *feature reduction* methods are tested to reduce the number of features to 25. The first method is based on the Pearson correlation. Features are ranked by their correlation with the game outcome, and only the 25 features with the highest correlation values are used for classification. The GainRatio method works similar and only keeps the 25 features that have the highest information gain. Principal Component Analysis is a more advanced method and transforms the 109 features into a reduced set of 25 new features which are a combination of the original features.

To measure the accuracy improvement of complex classifiers, two simple, *baseline predictors* were used. ZeroR uses none of the features and predicts the majority result class for every record. So, ZeroR always predicts the home team to be the winner of every game. OneE is a predictor based on one feature, the feature that produces the smallest error for the training set. For the other predictors, different classifiers available in WEKA and LibSVM [3] are used.

- *Support Vector Machines (SVM)* are non-probabilistic binary linear classifiers. The used SVMs of WEKA are trained using Sequential Minimal Optimization (SMO). In addition, the C-SVC (Support Vector Classifier) type of LibSVM is used. Different kernels are evaluated: linear kernels, polynomial kernels (standard and normalized version), sigmoid kernels, RBF (radial basis function) kernels, and PUK (Pearson function-based universal) kernels.
- *Naive Bayes Classifiers* are probabilistic classifiers with the assumption that features are independent (WEKA).
- *Multi Layer Perceptrons (MLP)* are feedforward neural networks utilizing backpropagation as supervised learning technique (WEKA).
- *Random Forest* is an ensemble technique using multiple learning algorithms to obtain less overfitting and better predictive performance (WEKA).
- *Bagging* is a bootstrap ensemble method. As a base learner, it uses REPTree, a decision tree based on information gain.
- *Simple Logistic Regression* estimates the probability of a binary outcome using a logistic function. For fitting the logistic models, LogitBoost (ensemble algorithm) with simple regression functions as base learners is used (WEKA).

Table 2 lists the accuracy of the different prediction models based on data of the five national football leagues. Each predictor was evaluated with the full set of 109 features (Full), and with reduced sets of 25 features. These reduced sets are generated using the GainRatio (GR), Correlation (Corr.) or Principal Component Analysis (PCA) technique. All results above 53.50% are in bold, since these models are useful in view of generating profit [19]. Simple classifiers, such as OneR, provide already an accurate baseline, as is common for classification problems [11]. The best result (54.37%) was obtained using RandomForest.

Model	Full	GR	Corr.	PCA
ZeroR	47.46%	-	-	-
OneR	52.29%	(`ht_goal_diff10`)		
SMO (PolyKernel)	52.54%	52.95%	53.001%	**53.81%**
SMO (Norm.PolyKernel)	48.62%	53.15%	53.10%	52.84%
SMO (RBFKernel)	52.14%	52.54%	52.54%	47.46%
SMO (Puk)	48.16%	**53.51%**	**53.81%**	49.44%
C-SVC (sigmoid)	**53.71%**	53.20%	53.45%	**53.71%**
C-SVC (polynomial)	**53.76%**	52.54%	52.44%	46.38%
C-SVC (radial)	52.95%	**53.81%**	**53.71%**	**53.96%**
C-SVC (linear)	53.15%	52.79%	52.84%	**53.76%**
NaiveBayes	27.00%	49.79%	50.40%	26.75%
MLP	51.93%	53.30%	53.20%	52.89%
RandomForest	**53.96%**	**53.71%**	**54.37%**	52.54%
Bagging	47.45%	47.45%	47.45%	47.45%
SimpleLogistic	52.89%	52.89%	52.89%	52.84%

Table 2: Accuracy of the predictors for data of all leagues.

The analysis was repeated for each league separately, since most teams do not play (often) against teams of other leagues. Support vector classifiers (C-SVC of LibSVM) showed to be the most consistent models over the leagues. Table 3 shows the most accurate model per league, together with the Kernel, the optimal value of the complexity parameter C, and the used technique to reduce the number of features. Large accuracy differences were witnessed over the different leagues. The highest accuracy was achieved for the Premier League, followed by the Serie A and LaLiga.

6 BETTING RECOMMENDATIONS

The accuracy results of Section 5 are calculated as if a bet was placed on every game. However, better results, in terms of accuracy and profit, can be achieved by holding off on some of the more uncertain bets. Therefore, different *betting strategies* can be considered. Bettors typically have their own preferences or decision rules to decide on a bet. Often these decisions are driven by the risk users are willing to take.

- *Published favorites.* This simple, baseline strategy is to always bet on the team that is the favorite, according to the published odds.
- *Predicted favorites.* This strategy always bets on the team that is the favorite, according to the predicted odds. If the model is more accurate than the published odds, this strategy can be profitable.
- *Predicted safe favorites.* A bet will only be placed if one of the teams is the clear favorite. In this experiment, betting is done if the probability that the favorite wins is at least 10% higher than

League	Accuracy	Kernel	C	Reduction technique
LaLiga	55.30%	linear	0.5	PCA
Premier League	60.09%	radial	1	GainRatio/Corr.
Serie A	57.83%	linear	0.125	None
Bundesliga	51.87%	sigmoid	4.0	None
Pro League	49.52%	radial	2.0	GainRatio

Table 3: The LibSVM parameters with the highest accuracy.

the other game outcomes. This strategy is more robust and often recommended to risk-averse users.

- *Playing the odds.* This is a commonly-used term in the betting jargon. If users suspect that the odds of a game published by the bookmaker are not correct, they bet on the game. This incorrectness can be estimated by comparing the published odds with the estimated probabilities of the model. Bets will be placed on outcomes that are underestimated by the bookmaker, but only if the probability of the prediction model is at least 10% higher than the probability of the bookmaker. Since this strategy also bets on underdog teams, it is recommended to users who are willing to take more risk with the perspective of a higher profit.
- *Home underdogs.* Bets are made if the away playing team is the favorite and the difference in probability between the home and away playing team is at least 10%. Because of the bookmakers' bias towards higher ranked teams (favorites), this strategy can be profitable [5]. Bookmakers often overestimate the odds of the favorite team, and underestimate the effect of the home crowd of the underdog. Since this strategy always bets on underdog teams, it is recommended to users who take big risks.

Besides recommendations for deciding on which games to bet (betting strategy), users can get recommendations for the size of the stake of the bet (money management). The output of the different *money management* (MM) strategies is a real number between 0 and 1, which can be multiplied by the maximum amount of money the user wants to spend per bet.

- *Unit bet (UB).* In this simple strategy, every bet gets the same stake, 1. This is a high risk, high reward MM strategy, since bets with a high risk get a high stake and thus a high potential profit.
- *Unit return (UR).* This strategies determines the stake size based on the odds to obtain equal unit sized returns. So, each winning bet yields the same amount of money, 1. UR is recommended to risk-averse users since risky bets receive a lower stake.
- *Kelly Ratio (KR).* This strategy is typically used for long term growth of stock investments or gambling [12]. The strategy is based on the difference between the model's estimated probabilities and the bookmaker's odds, and is therefore similar to playing the odds. If the model's probability is much higher than the bookmaker's odd, the bet is placed with a high stake. This strategy focuses on a consistent profit growth in the long term.

To evaluate the betting and MM strategies, a simulation is performed based on historical data. A fixed profit margin of 7.5% (This is an upper bound for realistic profit margins) is used to calculate the bookmaker's odds from the probabilities without profit margin. Again, the most recent 20% of the games are used for evaluation. Figure 1 shows the results of the different betting and MM strategies obtained with the best model for the Premier League (SVM with SMO and RBF Kernel). Playing the odds as betting strategy and unit bet as MM strategy showed to have the highest profit. The total profit after about 340 bet opportunities is 29.48 times the unit stake. However, this combination of strategies is characterized by strong fluctuations. A more risk-averse user can be recommended to use playing the odds in combination with UR or KR. Voting for the underdog was not profitable.

This analysis was repeated for the other leagues as well. Playing the odds and UB showed to have the highest profit potential; but for some seasons/leagues also big losses were made. This indicates that another strategy might be optimal for each league, but also that the results are strongly influenced by the game outcomes.

To demonstrate the prediction models, an interactive tool (called the betting assistant) generating rule-based recommendations for sports betting was developed. Users first specify their risk profile, which determines their matching betting and MM strategy. Optionally, they can specify their betting preferences such as the league. Subsequently, users get recommendations for football games to bet on, together with a recommendation for the size of their stake (value ranging from 0 to 1). Then, it is up to the user to accept the betting advice or not.

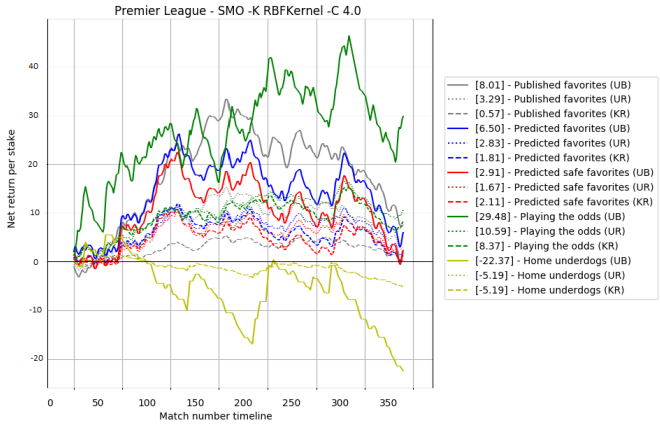

Figure 1: The evolution of different betting strategies.

7 CONCLUSIONS

Predicting the outcome of football games is a research topic with a growing interest. Various prediction models are assessed for this classification problem based on data of five European leagues. The game predictions are used in a prototype recommendation tool that suggests users on which game to bet (betting strategy), on which team (prediction outcome), and how much to bet (money management) depending on their personal preferences regarding risk and profit potential. These prediction models might be applied to other domains as well, such as predicting stock prices, or the outcome of elections. In future work, we will investigate the causality between game features (such as number of offsides, free kicks, etc.) and the game outcome in order to identify the drivers of the game's outcome. These drivers may expose the weaknesses of a team, which can be used by the team's coach to focus on specific tactical aspects during training sessions.

REFERENCES

[1] Georgi Boshnakov, Tarak Kharrat, and Ian G. McHale. 2017. A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting* 33, 2 (2017), 458–466.

[2] Maurizio Carpita, Marco Sandri, Anna Simonetto, and Paola Zuccolotto. 2016. Discovering the drivers of football match outcomes with data mining. *Quality Technology and Quantitative Management* 12, 4 (2016), 561–577.

[3] Chih-Chung Chang and Chih-Jen Lin. 2018. LIBSVM - A Library for Support Vector Machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4] Kong Wing Chow and Karen Tan. 1995. The use of profits as opposed to conventional forecast evaluation criteria to determine the quality of economic forecasts. *Applied Economics Research Series* 1 (1995), 187–200.

[5] Anthony Costa Constantinou and Norman Elliott Fenton. 2013. Profiting from arbitrage and odds biases of the European football gambling market. *Journal of Gambling Business and Economics* 7, 2 (2013), 41–70.

[6] Anthony Costa Constantinou, Norman Elliott Fenton, and Martin Neil. 2012. Pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems* 36 (2012), 322–339.

[7] Anthony Costa Constantinou, Norman Elliott Fenton, and Martin Neil. 2013. Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks. *Knowledge-Based Systems* 50 (2013), 60–86.

[8] Mark J. Dixon and Stuart G. Coles. 1997. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46, 2 (1997), 265–280.

[9] David Forrest, John Goddard, and Robert Simmons. 2005. Odds-setters as forecasters: The case of English football. *International Journal of Forecasting* 21, 3 (jul 2005), 551–564.

[10] Eibe Frank, Mark A Hall, and Ian H Witten. 2016. The WEKA Workbench. *Morgan Kaufmann, Fourth Edition* (2016), 553–571. http://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

[11] Robert C Holte. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine learning* 11, 1 (1993), 63–90.

[12] John L Kelly Jr. 2011. A new interpretation of information rate. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*. World Scientific, 25–34.

[13] Siem Jan S.J. Koopman and Rutger Lit. 2015. A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 178, 1 (2015), 167–186.

[14] Helge Langseth. 2013. Beating the bookie: A look at statistical models for prediction of football matches. In *Frontiers in Artificial Intelligence and Applications*, Vol. 257. 165–174.

[15] Michael J Maher. 1982. Modelling association football scores. *Statistica Neerlandica* 36, 3 (1982), 109–118.

[16] Byungho Min, Jinhyuck Kim, Chongyoun Choe, Hyeonsang Eom, and R. I. (Bob) McKay. 2008. A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems* 21, 7 (oct 2008), 551–562.

[17] Joel Oberstone. 2009. Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success. *Journal of Quantitative Analysis in Sports* 5, 3 (2009).

[18] The University of Waikato. 2018. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. https://www.cs.waikato.ac.nz/ml/weka/

[19] Martin Spann and Bernd Skiera. 2009. Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting* 28, 1 (jan 2009), 55–72.

[20] Niek Tax and Yme Joustra. 2015. Predicting the Dutch football competition using public data: A machine learning approach. *Transactions on Knowledge and Data Engineering* 10, 10 (2015), 1–13.