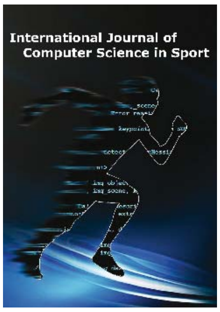
International Journal of Computer Science in Sport

Volume 18, Issue 1, 2019

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2019-0006



Forecasting Soccer Outcome Using Cost-Sensitive Models Oriented to Investment Opportunities

Talattinis, K.[1], Kyriakides, G.[1], Kapantai, E.[2], Stephanides, G.[1]

[1] School of Information Science, Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece

[2] School of Science and Technology, Department of Data Science, International Hellenic University, Thessaloniki, Greece

Abstract

Realizing the significant effect that misprediction has on many real-world problems, our paper is focused on the way these costs could affect the sports sector in terms of soccer outcome predictions. In our experimental analysis, we consider the potential influence of a cost-sensitive approach rather than traditional machine-learning methods. Although the measurement of prediction accuracy is a very important part of the validation of each model, we also study its economic significance. As a performance metric for our models, the Sharpe ratio metric is calculated and analyzed. Seeking to improve Sharpe ratio value, a genetic algorithm is applied. The empirical study and evaluation procedure of the paper are primarily based on English Premier League's games, simple historical data and well-known bookmakers' markets odds. Our research confirms that it is worthwhile to employ cost-sensitive methods for the successful predictions of soccer results and better investment opportunities.

KEYWORDS: SOCCER FORECASTING, COST-SENSITIVE CLASSIFICATION, BINARY DATA SET, GENETIC ALGORITHM, PREDICTION MODELS

Introduction

In the last few years, considerable attention has been paid to the forecasting of sporting event results. Sports betting attracts the interests of both casual investors and researchers and academics whose efforts have been devoted to proving that sports outcomes can be predictable. The problem with this approach is the fact that competitive sports such as soccer are inherently unpredictable. It is difficult to predict the outcome of any game that is played between two equally poised teams, particularly in a soccer match that is a low-scoring game. Thus, despite all the opposition and extensive research by numerous authors, there is still room for improvement in terms of developing more accurate forecast models.

Several publications have demonstrated various attempts to develop more accurate prediction models in sports analytics, including computational and statistical methods. Our investigations into such studies have shown that most of them try to either minimize the error rate (number of incorrect predictions) or maximize the accuracy ratio (proportion of the correct predictions). Relevant contributions include the publications by Buursma (2011), Odachowski & Grekow (2013) and Haghighat, Rastegari & Nourafza (2013). These authors disregard the differences between types of misclassification errors, giving them equal weight in their analysis. Nevertheless, in real-world problems, the mispredictions are not equally costly. In fact, such an assumption could have a considerable impact on one's decision-making process. Ignoring the differences in these costs could lead to a useless model, because only the most frequent types of mispredictions would be considered in the analysis, even though the less frequent types of mispredictions could also result in substantial cost. For example, in the medical diagnosis of cancer, misdiagnosing a health patient as a cancer-sufferer has less impact on the patient than diagnosing a patient as cancer-free when he or she is ill, because the latter error could result in loss of a life due to a delay in treatment.

To understand this issue, various approaches have been reported with a bias toward minimizing the total cost of misclassification. Cost-sensitive classification is the general method that is applied to minimize the expected cost. However, even it finds wide applicability in different sectors of daily life, in sports field slight research activity has been observed until now and our work is focused on filling this need.

Our contribution is twofold. First, we apply both multi-class and binary classification and explore the feasibility of a cost-sensitive approach in terms of predicting soccer match outcomes. Second, we examine the quality of our results and discuss the investment opportunities that may arise in comparison with prior works that mainly focus on the optimization of prediction accuracy.

The remainder of the paper is organized as follows. In Section 2, an overview of the related literature is conducted. We describe various studies associated with soccer-outcome predictions and potential profitability. In Section 3, we detail the theoretical background of machine learning methods that will be examined. In Section 4, the value of cost-sensitive approach along with its technical part is described. Section 5 provides a discussion of our problem and initial challenges and presents the standardized data sets and useful metrics of our research methodology. The experimental procedure along with the implementation strategy are introduced in Section 6. Sections 7 and 8 describe the results of our empirical investigation and present our concluding remarks respectively. Finally, in section 9 a discussion of our results is conducted considering the related work.

Related Work

After a thorough review of the literature, we found several scientific publications that focus on how someone can predict soccer results in terms of investment returns. These studies can be categorized based on the kind of data they use such as structured data (i.e. game/player data, odds data) or unstructured data (i.e. tweets), the type of the outcome that are about to predict (i.e. number of goals scored or the direct outcome "win-draw-loss") or the techniques they use for prediction (i.e. statistical models, machine learning, etc.).

In this section we provide research works that examine betting market's efficiency using structured data i.e. bookmaker odds, in terms of "win-draw-loss" before the game. As for the employed techniques and methods, there are various ways in which the quality of a forecast model can be assessed. We identify two types of approaches *i) accuracy-oriented* (how close the forecasts are to actual results) (Constantinou, Fenton & Neil, 2012) that seek to enhance the prediction ability of the suggested models (i.e. making the least possible mistakes) and *ii) profit-oriented* (how useful the forecasts are in a betting strategy basis) Constantinou et al. (2012), allowing greater errors on the premise that riskier matches yield higher profits.

Accuracy-oriented

A number of authors have attempted to model efficient processes in order to determine the outcome of soccer matches. The earlier focus was on econometric approaches developing statistical and probabilistic strategies such as Poisson distributions (Koopman & Lit, 2013; Karlis & Ntzoufras, 2003; Dixon & Pope, 2004), ordered probit (Goddard & Asimakopoulos, 2004), Markov Chain simulations (Crowder, Dixon, Ledford & Robinson, 2002), Regression models (Karlis & Ntzoufras, 2008; Haaren & Broeck, 2014; Goddard, 2005). Newer contributors adopted more computational methodology performing predictive ranking systems (Haaren & Davis, 2015) or data mining methods like Naïve Bayes, Bayesian Networks, Support Vector Machines, Neural Networks and combinations of various machine learning algorithms. For example, Buursma (2011) applied machine-learning techniques (ClassificationVia Regression (linear regression), Multiclass Classifier (logistic regression), RotationForest, Bayes Net, and Naive Bayes) using full-time scores, half-time scores and bookmaker odds as key features of his data set. The models achieved accuracy rates from 54.43% to 57.00%. Odachowski & Grekow (2013) took their research a step further using predictive machine learning algorithms (Bayes Net, VotedPerception, Ibk, Bagging, DecisionTable, and LADTree) over a novel binary classification approach that focuses on the prediction of each possible outcome individually. They managed to achieve an accuracy of 70%. Another study proposed by Eryarsoy & Delen (2019) assessed the predictive power of different models (Naïve Bayes, Decision Trees, and Ensemble models) based on their accuracy values. The results showed 74% accuracy rate in "Win/Loss/Draw".

Profit-oriented

The availability of multiple forecasting methods raised questions about their effective use and the potential of systematic profitability for investors in sport markets. Among the works that tried to come up with this challenge, Forrest, Goddard & Simmons (2005), Spann & Skiera (2009), Constantinou (2018), Kyriakides, Talattinis & George (2014), Kyriakides, Talattinis & Stephanides (2015) and Kyriakides, Talattinis & Stephanides (2017) proposed interesting approaches. Some of these cases showed abnormal positive returns from betting strategies.

Forrest et al. (2005) focused their interest on English Premier League and examined the effectiveness of forecasts based on published odds over the forecasts made using a benchmark statistical model incorporating. The findings challenged the consensus about the superiority of odds in terms of more accurate and profitable predictions.

Spann & Skiera (2009) forecasted the results of the German premier soccer league, examining the predictive accuracy and profitability capacity of three different methods: prediction markets, tipsters and betting odds. In terms of accuracy, tipsters seen to outperform the other methods. However, none of the forecasts leads to systematic monetary gains in betting market.

Constantinou (2018) designed a model to predict match outcomes from all over the world using a mixture of dynamic ratings and Hybrid Bayesian Networks. Return On Investment (ROI) metric was considered as profit evaluator. The model performed remarkably in the case of English Premier League managing to reach a 38% ROI.

Kyriakides et al. (2014) also worked on English Premier League to predict final outcomes and identify profitable methods. They compared the performance of linear algebra ranking algorithms (mHITS, Colley, Massey) to a machine learning approach (Neural Networks, Decision Trees, Random Forests). Concluding that accurate models are not always profitable, tried to improve them following another approach Kyriakides et al. (2015) and applying risk management as filters on their preferences. Finally, considering the performance of both machine learning and linear algebra ranking systems on prediction of upcoming matches' outcomes, Kyriakides et al. (2017) suggested a hybrid method, combining Colley's, mHITS and the novel AccuRate rank/rating systems with machine learning methods (Artificial Neural Networks, Decision Tables, Naive Bayes, Logistic Model Trees, Bagging, Stacking). They suggested that combination models performed the best and allow greater flexibility in terms of the desired goal (accuracy, profit).

Machine Learning Methods

Conventional classification procedures are fundamental to data mining. They have been used for decades in research areas such as machine-learning and statistics. The functions of a classifier are to be trained from a set of unknown objects for which class labels have been defined and to be used in making predictions of these classes to a new set of objects. In such cases, the results are assessed by analyzing the success rate (the probability of making correct predictions). There are various well-known and well-understood techniques and algorithms that are extensively applied for predictive modelling.

Naive Bayes

Naive Bayes classifier (Kyriakides et al., 2017; Hand & Yu, 2001) is a simple and easy to implement algorithm. It is also considered as a computationally fast and surprisingly powerful technique that performs well in most cases. It is based on two types of probabilities that can be calculated directly from your training data: 1) The probability of each class; and 2) The conditional probability for each class. The probability model can be used to make predictions for new data using Bayes Theorem. Naïve Bayes can be considered as a special type of Bayesian network, relying on the assumption that the attributes are independent and that no other attributes influence the predicted class. The technique is very effective on a large range of complex problems.

Decision Tables

Decision tables (Kyriakides et al., 2017) are one of the simplest and easily understandable algorithms for supervised learning. They are consisted of two main components: The schema, which is a set of attributes, and the body, which is a multiset of labeled instances from the space defined by the schema. The labels of instances outside the body can be predicted based on the instances inside the body.

Decision Tree

A decision tree (Bhargava, Sharma, Bhargava, & Mathuria, 2013; Kyriakides et al., 2014) is a tree-like graph that is used to describe and classify data. Algorithms usually construct decision trees top-down, by choosing the variable that best splits the data into appropriate sets. The structure includes a root node, links(branches) that represent a decision rule, leaf nodes that represent an outcome/class (categorical or continues value) and internal nodes that represent test conditions applied on attributes. The goal is the creation of a tree for the entire data and the process of a single outcome at every leaf (or the minimization of the error in every leaf). Decision Tree is a supervised classification method because the dependent attribute and the counting of classes (values) are given.

Random Forest

Random Forest (Breiman, 2001; Le, 2019) is a specific type of ensemble machine learning known as Bootstrap Aggregation (Bagging). It is a combination of tree predictors in a manner that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Contrary to traditional bagging where a random vector of instances is sampled from the dataset, Random Forests sample a vector which contains the whole dataset, but only a subspace of the available features. Random Forest is considered to be a robust method with respect to noise.

Support Vector Machine

Support Vector Machine (SVM) (Suykens & Vandewalle, 1999) is a pattern recognition method based on statistical learning theory. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space, where N is the number of the features, that distinctly classifies the data points. In order to select the optimal hyperplane, SVMs utilize data instances that lie on the edge of each class called Support Vectors. Following, they select the hyperplane that has the maximum possible margin from the support vectors.

K-Nearest Neighbor

The KNN algorithm (Le, 2018) is simple and effective. The model representation for KNN is the entire training dataset. A new data point is predicted by searching through the training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. In classification, this is usually achieved through majority voting, i.e. selecting the class that is most prevalent amongst the K nearest neighbors.

A major task of most ML techniques and methods including the above-mentioned ones is the process of classification. What is not considered though by these methods is the cost of misclassification of the different classes.

Overview of Cost-sensitive Learning

Cost-sensitive learning, as noted by Elkan (2001), surveys each type of emerging cost and replaces that amount with the average cost per prediction. The ultimate aim is to minimize the average cost per object. The cost in each case is given by a specific entry in the confusion matrix (Goddard & Asimakopoulos, 2004).

A confusion matrix is commonly used to assess the accuracy of a classification approach. As developed by Provost & Kohavi (1998), it contains information about actual and predicted classifications. The strength of a confusion matrix lies in the fact that it identifies the nature of the classifications errors as well as their quantities. The behavior of a classifier can be further interpreted by using a cost matrix that corresponds to the confusion matrix and provides the costs for each of the outcomes shown in the confusion matrix (McCarthy, Zabar & Weiss, 2005). In a classification problem with K classes, the misclassification costs can be represented by a K x K cost matrix. The rows of the matrix represent the classes, whereas the columns represent the predicted classes. The on-diagonals depict the costs of correctly classified instances, while the off-diagonals depict the misclassification costs (Ting, 1998). If the positive and negative classes are labeled 1 and 0, respectively, the cost matrix of a two-class case would be configured as in Table 1.

Table 1: Confusion/Cost matrix

	Predicted Negative class	**Predicted Positive Class**
Actual Negative Class	$C(0,0)$ or TN	$C(1,0)$ or FP
Actual Positive Class	$C(0,1)$ or FN	$C(1,1)$ or TP

Suppose that an entry of a matrix C is depicted by the vector *(i, j),* where i represents the cost of the predicting class and j represents the cost of the true class. We can produce a general rule to define the minimum expected cost of x if x belongs to the class i. Much of what follows is taken from Michie, Spiegelhalter & Taylor (2009) and Sheng & Ling (2009):

$$R(i|x) = j\ P(j|x)C(i,j) \tag{1}$$

where $P(j|x)$ is the probability of each class j to be the true class of x. A classifier will identify an example x as belonging to a positive class if:

$$P(0|x)C(1,0) + P(1|x)C(1,1) \leq P(0|x)C(0,0) + P(1|x)C(0,1) \tag{2}$$

Note that a positive class is more difficult to predict than a negative one. Thus, our efforts will focus primarily on the recognition of positive instances, as their misclassification values are higher than those of negative instances.

Another approach to understand the performance is by using ROC curves. A ROC graph is a plot in which the axis X represents the false-positive rate of the confusion matrix and the Y axis depicts the positive well-classified samples. Noting that, by definition, the true-positive rate represents the sensitivity of a system, while the false-positive rate indicates the probability that an (false) error is occurring, ROC curves are a measure of the sensitivity of the fall-out. The graph encapsulates the information contained in the confusion matrix and forms an efficient tool with which to evaluate both the classifier's ability to correctly identify positive cases and its ability to determine the number of negative misclassification cases.

ROC curves are generally used for binary classification problems; however, we use them to analyze pair-wise comparisons, which allows us to interpret our 3-class problems. Extensive analysis will be given in Section 5.

There are two commonly accepted cost-sensitive methods: the direct learning method and the meta-learning process. The former concerns the implementation of algorithms that are cost-sensitive oriented, while the latter concerns the generic approach with which to evaluate methods that work as a "wrapper," i.e., the methods that convert a cost-insensitive classification method into a cost-sensitive one (Sheng & Ling, 2009).

Our work employs the meta-learning procedure because it has broader applicability and can be further categorized into two general approaches. One approach focuses on the unbalanced costs prior to or during the learning phase of the model by assigning proportional weights to different training instances or by modifying the class distributions and applying cost-insensitive classifiers (sampling). The other approach consists of defining and implementing a decision threshold to classify examples derived from a model by cost-insensitive methods (thresholding). Meta-cost and cost-sensitive classifiers introduced by Domingos (1999) and Witten & Frank (2005), respectively, are examples of thresholding methods.

MetaCost Classifier

Aiming to transform a cost-insensitive classification problem into a cost-sensitive one, the MetaCost classifier combines the predictive ability of bagging with an accessible model for cost-sensitive prediction. Bagging is a powerful method because of its ability to produce very accurate probability estimates. The MetaCost function focuses on assigning new labels to the training data examples, because each new label reflects the least costly prediction for the example and the lowest risk to a base learner, even if the predicted outcome is incompatible with the true one. The new labels are defined based on the probability estimates of bagging that are used as an ensemble classifier. Next, the function discards these labels and learns a new classifier from the relabeled data. As the costs have been incorporated into the class labels, the newly generated model is able to make a cost-sensitive prediction.

Cost-Sensitive Classifier

In addition to the MetaCost classification technique, the cost-sensitive classifier is an alternative cost-sensitive learning method. There are two different approaches to this method: The first approach consists of changing the proportion of each class in the training data to reflect the cost matrix. The second approach concerns the prediction of the class with the minimum expected misclassification cost. This second process involves both a learning and a testing phase The use of a bagging classifier is helpful to the performance of the cost-sensitive classifier as well.

The fundamental structure of both the MetaCost and the Cost-sensitive classifiers is concerned with making its base learner cost sensitive. Thus, the MetaCost classifier is unique in that it produces a single cost-sensitive classifier of the base learner which results in fast classification and interpretable output. This cost utilizes all bagging iterations when reclassifying the training data, as discussed by Domingos (1999).

Binary classification vs. Multi-class classification

The processes discussed above were approached from the perspective of extending the analysis of cost-sensitive learning. However, these methods may also be examined as two-class or multi-class problems.

Binary classification examines problems entirely defined by two classes. On the other hand, multi-class classification is a process that uses more than two classes and aims to assign instances to one of the possible discrete classes. There are two well-known approaches in the literature that are suggested for the extension of the binary to the multi-class case.

The former entails the training of a classifier over a particular class, taking into account that the samples of that class are assumed to be the positives, while all the remaining samples are categorized as negatives (One-vs-All classification). The latter, defined as a K-class problem, considers K(k-1)/2 binary classifiers and assigns samples of two classes to each of them. The goal of the K-class problem is to learn to recognize this pair of classes (All-vs-All classification) (Bishop, 2006).

Our attention is focused on the investigation of a three-class problem through binary components and evaluating simpler problems with a comparable analysis. For every multiple class, a unique binary model that assess the individual class against all the remaining ones is going to be constructed. The evaluation of these binary classifiers in terms of prediction will be measured while considering the percentage-of-confidence score.

When assessing the binary problem from a cost-sensitive view, one sees that there are two types of costs that need to be addressed: the cost of misclassifying the first class as the second and the cost of misclassifying the second class as the first. Table 1 presents an illustration of a two-class problem using a cost matrix.

Portfolio Definition

In a soccer match with three possible outcomes, namely, the win of the home team or the draw or win of the visiting team, the binary method seems to play a crucial role in the decision-making process. Because a binary classifier focuses on one problem the time, it possesses a better performance in the classification process than does a method that handles three different classes. Thus, in developing a classifier for one of the three cases, it is assumed that the intended class remains unchangeable, whereas the other two are merged in order to form one integrated class. The assemblage of such combinations composes a portfolio. We describe the portfolio in more detail below.

The portfolio concerning our binary approach involves three data sets and three classifiers:

- Win of the home team. One class includes the matches that end up with a win for the home team, and the other class combines the matches that ended up either with a draw or with a win of the visiting team.
- Draw. One class is composed of the matches that end up with a draw, whereas the matches with a winning outcome for the home team or a win for the visiting team are combined, forming an integrated class.
- Win of the visiting team. This type of data set is characterized by one class with all the matches ending up with a win of the visiting team and another class with the remaining soccer matches.

Methods of Approach

Initial Challenges

To verify the validity of the methods and techniques described above, we carried out several experiments to develop our models. We began by investigating the behavior of different algorithms on our data space by implementing and analyzing individuals in each of the various sectors. The overall objective was to reveal the best-performing model in terms of prediction ability and profit regularity and continuity.

The precision and profit predictions of our models appeared to be quite promising; perhaps they were even more promising than the established ones.

Although in most surveys on machine-learning the determination of the best model is based on the proportion of accurate predictions, we concluded that this method is not always cost-effective.

For example, assuming a soccer game with average odds (1.2, 7.5, 12) for (Home Win, Draw, Away Win) respectively and stake of 100 monetary units. If we bet on Home Win correctly, we win 20 monetary units but in case of a loss we need 5 times of correct predictions to balance our initial capital. On the other hand, suppose that we bet on a Draw or Away Win, our net profit will be 650 monetary units and 1100 monetary units respectively giving us room for more unsuccessful predictions. Thus, the cost or the risk of misprediction of final outcomes for low valued odds is higher. Considering that models are biased on predicting low odds as they tend to improve the number of correct predictions, we apply cost-sensitive methodologies in order to direct our model to focus on high valued odds. In this way, the model is trained to be reluctant to predict low-valued odds for a fear of loss, emphasizing on predictions of high valued ones that allow unsuccessful predictions in a larger amount. We acknowledge the difficulty in this endeavor, however this is the matter about profitability.

Data used

The set of input data was collected from the online available database at http://www.football-data.co.uk (Football Results, 2018). We used 6 years of soccer games spanning the years 2010-2016 in the English Premier League. The historical data used consists of the wins, goals and shoots. The bookmaker odds are taken as an average of the 5 biggest online bookmakers (Bet365, Bet & Win, William Hill, SportingBet and Interwetten).

Each data point represents an independent game that is described by a set of different attributes. To ensure the best results, we selected only those features that were either publicly available or easily calculated. We found that there were many useful features with which to predict the outcome of the game with maximum accuracy. For each possible outcome, the most valuable attributes for the decision-making process were: the average odds (home-win, draw, away-win); the total number of shots-on-target; the total goals scored by each team prior to the itinerary match and the final outcome of the match. Example of an instance can be found in Table 2.

Table 2: Example of dataset instance

Avg. Home Odds	Avg. Draw Odds	Avg. Away Odds	Home Total Shots On Target	Away Total Shots On Target	Home Total Goals Scored	Away Total Goals Scored	Final Outcome
2.175	3.4	3.45	24	19	6	8	1

Money Management

The efficient (indicative) stake was defined in two different methods. The first method involved fixed placing bets, and the second stake was determined by the Kelly criterion. Before deciding on the stake amount, it was essential to look for a mathematical edge rather than rely on our impulses. Many papers recommend the use of the Kelly criterion, as it is assumed to hold a distinct advantage over other similar methods because of its lower level of risk. This model calculates the amount with which to make a bet based on an outcome with a probability of success that is higher than the given odds. A successful outcome of the model involves the long-term exponential growth of the fund.

The Kelly criterion is:

$$f^* = pb - \frac{1-p}{b} \tag{3}$$

where: f*= bankroll proportion per bet, b = betting odd, p = probability of success and q = (1-p) = probability of failure

To evaluate the results using the Kelly criterion, two different frameworks were defined and applied. The comparative analysis was performed to illustrate the graphical fluctuations, whereas the accuracy rate was used to define the ability of the model to accommodate misclassifications. Based on a voting scheme, a customized rating system with which to reveal the model with the best results was applied.

Model Evaluation Metric

To assess the models' performance, Sharpe ratio (Sharpe, 1994) was used as a measure. Financial investors widely tend to utilize Sharpe ratio, in order understand the return of an investment compared to its risk. In this paper, we investigate the feasibility of such methods on sports betting. Distributions with a higher mean and a lower standard deviation, i.e., a cumulative distribution to the right with a steeper curve, indicated the better model. The Sharpe ratio quantified an investment strategy's performance by combining its strategy's average returns with its standard deviation of returns. It ranked an investment's performance system by adjusting for its risk. In our case, it gave the amount that a gambler is compensated for with respect to the risk taken, i.e., the return on the investment. The formula is given below:

$$Rating = \sqrt{K}\, Sharpe\ Ratio = \sqrt{K}\frac{Mean(x) - Mean(f)}{Sd(x)} = \sqrt{K}\frac{Mean(x)}{Sd(x)} \tag{4}$$

where K is the total number of trades (bets) and x is the trade's return. We integrate the square root of K in our metric in order to force the system to generate models with a high trading frequency. This eliminates cases where a high but statistically insignificant Sharpe ratio is generated, due to a low number of observations (trades). We assume a zero mean of a risk-free instrument, as any other value will interact with $\sqrt{K}$ and affect the relative rankings of our models, by reducing its effect on the final rating.

Furthermore, $\sqrt{K}$ regulates the impact that negative Sharpe ratios have on the selection of models. As the rating also acts as a fitness function, models with negative Sharpe ratios and a high number of trades (and thus consistently negative returns) will have a lower fitness function than models with the same Sharpe ratio but lower number of trades (and thus more probable to be random).

Experimental Procedure

To solve the learning problem effectively, given the learning algorithms that we utilize, feature engineering is of paramount importance. As the features will determine the linearity and degree of classes' separability, good features will enable even simple classifiers to perform well. In order to automate the process, we employ a genetic algorithm, able to perform elementary operations on the (addition, subtraction, multiplication, division) as well as transform them by applying specific functions (logarithm, square root etc.). We used WEKA for this purpose (Hall et al., 2009; Frank et al., 2017).

Genetic Algorithms & Feature Extraction

Genetic algorithms are adaptive heuristic search algorithms that are used to solve optimization problems. They exploit information from the population to address the search process.

A genetic algorithm's strategy for solving betting problems is analogous to Darwin's "Survival of the Fittest" evolutionary philosophy. Thus, within a search space, each member of a population represents a possible solution. In general, evolutionary computational methods, including genetic algorithms, are created through the following steps: 1. Initialization of population, 2. Fitness evaluation of population, 3. Reproduction (a. Selection of parents, b. Application of crossover operator, c. Application of mutation operator, d. Definition of new population) and 4. Termination condition.

In our experimentations we use genetic algorithms as optimization method in order to automate the feature engineering procedure as we cannot have full knowledge of the best-performing attributes or combinations of them.

We note that models built on optimized data tend to be of higher quality because the data are characterized by more valuable information. In addition, one can learn important statistical information from certain types of transformations (Box & Cox, 1964). Depending on the demands of a given data set, the best mathematical transformation is selected to enhance the variables' distributions. In general, it is desirable for random variables to be distributed normally. Transformations can help to achieve this, as they are invertible and possess minimal impact on a variable's properties. For example, if one considers the logarithmic transformation and the fact that that the slope of logarithmic function is larger for smaller domain values, one observes that when values are small, one sees that the differences of small values are expanded, whereas the differences between larger values are reduced (Scibilia, 2012).

Weka Hall et al. (2009) and Frank et al. (2017) also discuss a variety of helpful and effective transformations. For example, MathExpression is a filter used to modify numeric attributes of a given expression. This transformation is an association between the value of the attribute being processed and support operators, such as the minimum or maximum value of the attribute, the standard deviation, the mean, etc.

Recall that our experiment involves a genetic algorithm in which each variable represents a different feature combination. Using the initial dataset, each member of the initial population was associated with either a unary operator (sin, cos, ln, exp, sqrt) that operated on one attribute or with a binary operator $(+, -, /, \cdot)$ that operated on two attributes, chosen at random. The crossover operator (Figure 1) involved splitting the representation of top-level binary operators of two individuals and exchanging the rest of their representation.

The mutation operator involved either applying a unary operator to an individual member of the population or applying a binary operator to an individual as the first operand and an attribute of the original dataset as the second operand. To avoid generating overly complex attributes, an upper limit of 5 unary nested operations was imposed.

Thirty percent of the population was used as a validation set. A learning algorithm was trained on 70% of the population and tested on the validation set. The classifier's accuracy on the remaining 30% was regarded as the individual's fitness value. Roulette wheel selection was utilized at the end of the evaluation to select the parents of the next generation.

The population consisted of 100 individuals with 0.9 crossover rates and 0.1 mutation rates. Each initial dataset and target combination were subjected to 100 different evolution runs, each with a duration of 2 minutes.

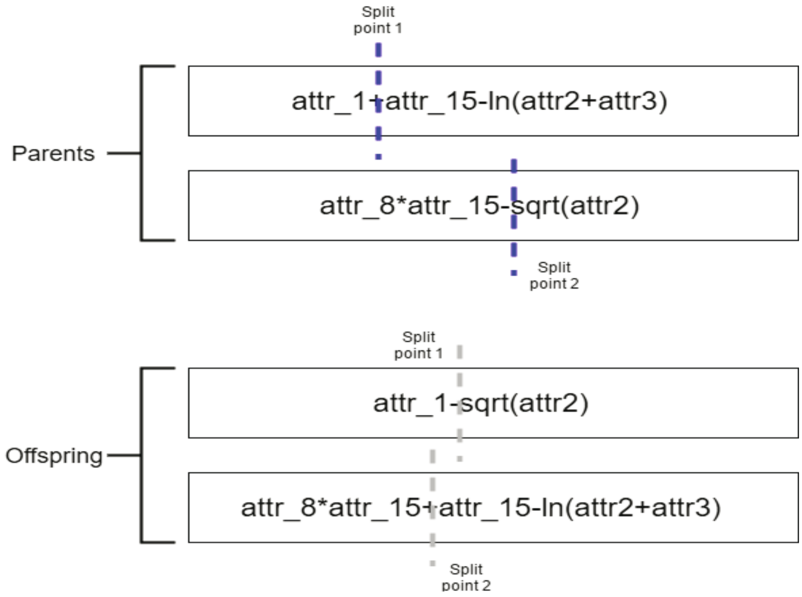

Split point 1
Parents
attr_1+attr_15-ln(attr2+attr3)
attr_8*attr_15-sqrt(attr2)
Split point 2
Split point 1
Offspring
attr_1-sqrt(attr2)
attr_8*attr_15+attr_15-ln(attr2+attr3)
Split point 2

Figure 1. Example of crossover operation

Experimental Design

Our experiments exhibit a step-wise behavior. In each step we compare two variations of our method and proceed with the best of the two. The comparisons entail: 1) Efficiency between fixed stakes and Kelly criterion over ternary classification. 2) Performance of ternary (all possible outcomes) and binary classification (home win or not). 3) Employment of cost-sensitive techniques. 4) Comparison of various machine learning algorithms. To demonstrate the system's capabilities, various algorithms were used to create predictive models using the generated features. Kelly criterion was implemented as a risk management technique and stake size calculator. Sharpe ratio was used to determine the individuals' fitness.

As already mentioned, the classifiers were trained on 70% of the data, using traditional loss functions, depending on the classifier (i.e. cross-entropy losses, misclassification rates etc.). The final assessment of each classifier involved generating predictions on the validation set and utilizing the predictions as betting tips (i.e. betting on the predicted outcome) and generating an equity curve, based on the outcome. A successful bet was defined as a successful prediction of the outcome and was rewarded as the average odds (of the 5 biggest online bookmakers).

In order to assess the method's ability to produce high-quality models, 100 individuals were evolved for two minutes, on a computer with AMD FX-6100 CPU and 16GB of DDR3 SDRAM. The experiment was repeated 100 times, yielding 100 best-performing individuals. Each individual was evaluated by calculating the Sharpe ratio of its equity curve, based on the evaluation (30%) set.

Our experimental procedure is represented by Figure 2 below.

Experimental Results

As stated in experimental design section the genetic algorithm (depicted in Figure 2) was executed 100 times for every variation/experiment (i.e. fixed-stakes/Kelly criterion) of each step of our methodological procedure, resulting in 100 best-performing models (one for each run). For these models a cumulative distribution of Sharpe ratio was constructed as they are presented in Figures 3,4,6,7.

For the comparisons, Fixed-Stake/Kelly criterion and Binary/Ternary we compare these distributions using Kolmogorov two sample test, in order to examine if the models originate from the same distribution. We further analyze them in terms of their performance using their mean Sharpe ratio metric. The one with the best average Sharpe ratio value outperforms. Tables 3,4,5,6 describe those values.

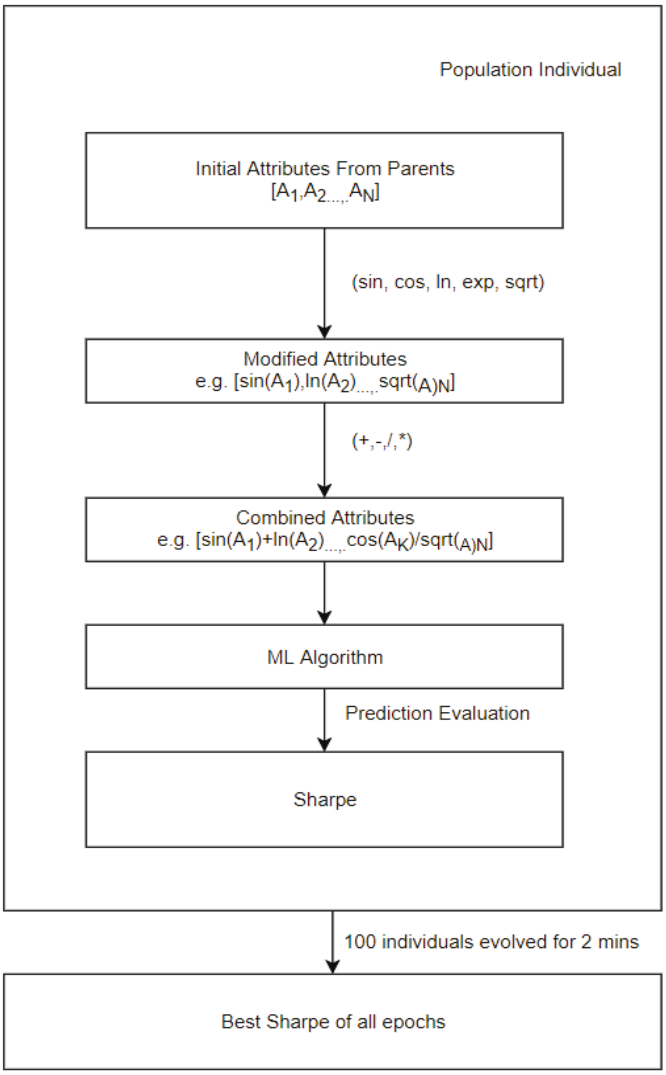

Population Individual
Initial Attributes From Parents
$[A_1,A_2...,A_N]$
(sin, cos, ln, exp, sqrt)
Modified Attributes
e.g. $[\sin(A_1),\ln(A_2)...,\mathrm{sqrt}(A)_N]$
(+,-,/,*)
Combined Attributes
e.g. $[\sin(A_1)+\ln(A_2)...,\cos(A_K)/\mathrm{sqrt}(A)_N]$
ML Algorithm
Prediction Evaluation
Sharpe
100 individuals evolved for 2 mins
Best Sharpe of all epochs

Figure 2. Design of the Experimental Procedure

Before we analyze our results, we should also mention that during the first three steps of our experimental process, Naïve Bayes classifier was used in order to examine the outperformance of the suggested methods in each step. Naïve Bayes classifier was selected among others as it is a simple and well-performed method. Also, newer studies (Dobravec, 2015; Godin, Zuallaert, Vandersmissen, De Nevem & Van de Walle, 2014) tend to use Naïve Bayes algorithm as works well with almost any kind of dataset. The results of our experiments are shown below.

The first experiment consisted of a comparison of a fixed stake size versus Kelly's criterion stake size. By utilizing the Naive Bayes classifier, the estimated probability and the predicted outcome's odds, either 1 monetary unit (fixed stakes) or less (Kelly criterion), was bet. As the average Sharpe ratio is negative for the fixed stake approach and positive for the Kelly's criterion approach, it was safe to assume that the variable stake approach was superior.

Using a two-sample Kolmogorov-Smirnov test, we rejected the null hypothesis that the two distributions were identical (p-value< 1E-3). The results are depicted in Table 3 and Figure 3.

Table 3: Fixed stakes over Kelly criterion comparing the Mean Sharpe ratio value

Model	Mean Sharpe ratio	Stand. Deviation
Ternary (Fixed stakes)	-0.0369	0.0394
Ternary (Kelly criterion)	0.0014	0.0366

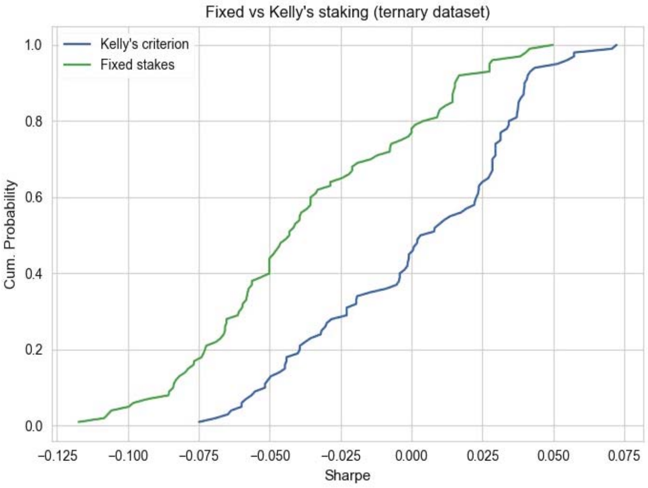


Figure 3. Performance comparison of fixed stakes and Kelly criterion

After identifying the advantage of using Kelly criterion to determine our stake size, we tested the feasibility of using a binary classification scheme. Specifically, we focused on the probability that the home team would win. The target variable was transformed into a home win 'H' or a no-home win 'N'. By using this binary classification and Kelly criterion as a staking scheme (as it was proven superior in the previous step), we observed an improvement over the ternary dataset. Using a two sample Kolmogorov-Smirnov test, we reject the null hypothesis (p<1E-3). By comparing the average Sharpe ratio of the two models, we observed that the binary dataset has a small advantage over the ternary dataset (Table 4 and Figure 4).

Table 4: Performance of Binary model over Ternary comparing the Mean Sharpe ratio value

Model	Mean Sharpe ratio	Stand. Deviation
Ternary	0.0014	0.0366
Binary	0.0043	0.0301

Additionally, after examining ROC curves for the Binary and Ternary datasets (Figure 5), we verified the small predictive advantage (in terms of accuracy) observed in Table 4 of using the binary, over the ternary, dataset. Another advantage was that the binary dataset generated simpler models.

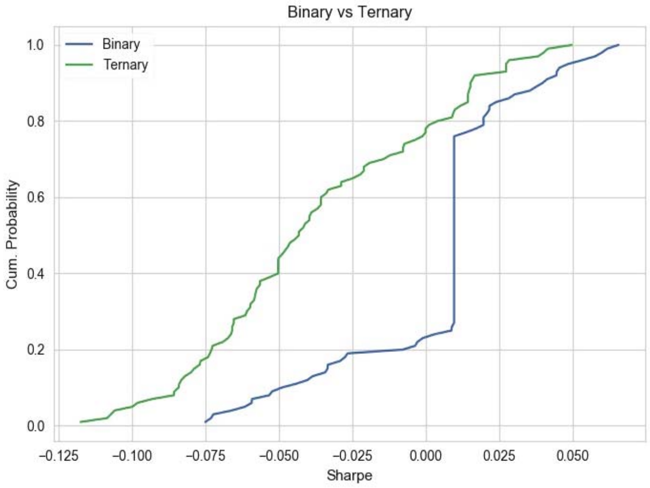


Figure 4. Comparison of Ternary and Binary sets

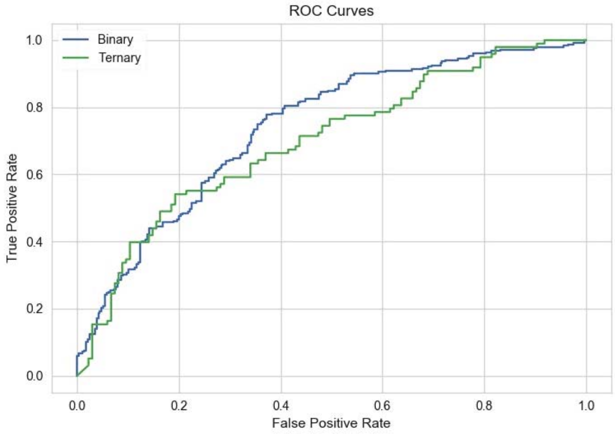


Figure 5. Comparison of Binary and Ternary sets using ROC curves

We next compared the performance of cost-sensitive learning over the binary data sets using Cost sensitive and Meta Cost classifiers. We concluded that the latter outperforms the former in terms of potential profitability. Table 5 and Figure 6 validate these results and confirm our initial assumption regarding the high performance of cost sensitive models.

Table 5: Performance of Binary models comparing the Sharpe ratio value

Model	**Mean Sharpe ratio**	**Stand. Deviation**
Binary	0.0043	0.0301
Cost sensitive (Binary)	0.0206	0.0320
MetaCost (Binary)	0.0338	0.0112

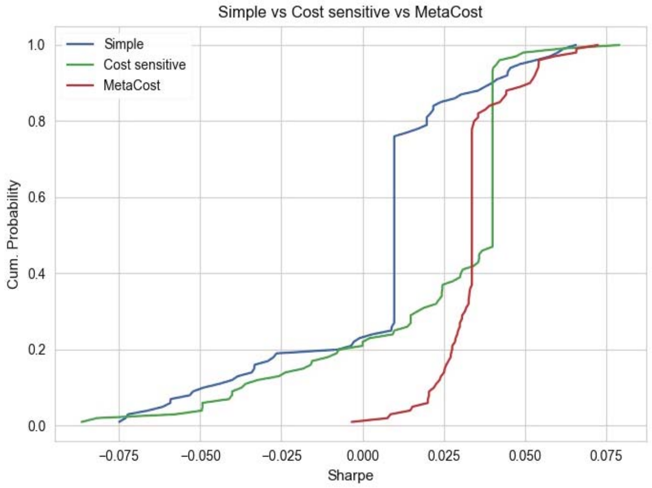


Figure 6. Comparison of simple, cost-sensitive and MetaCost binary models

Finally, considering the better feasibility of Meta Cost classifier, we compared many classifiers using Kelly criterion as money management, binary dataset and the MetaCost meta-classifier. The classifiers tested were Decision Trees (J48), Naive Bayes (NB), Decision Table (JRip), Support Vector Machines (SMO), Random Forest (RF) and K-Nearest Neighbor (IBk). Random Forests and Naive Bayes seemed to produce the best models (Table 6 and Figure 7).

Table 6: Classifiers comparison using Mean Sharpe ratio metric

Classifier	**Mean Sharpe ratio**	**Std. Deviation**
J48	0.022	0.027
JRip	0.006	0.034
SMO	-0.001	0.033
NB	0.034	0.011
RF	0.03	0.012
IBk	0.023	0.032

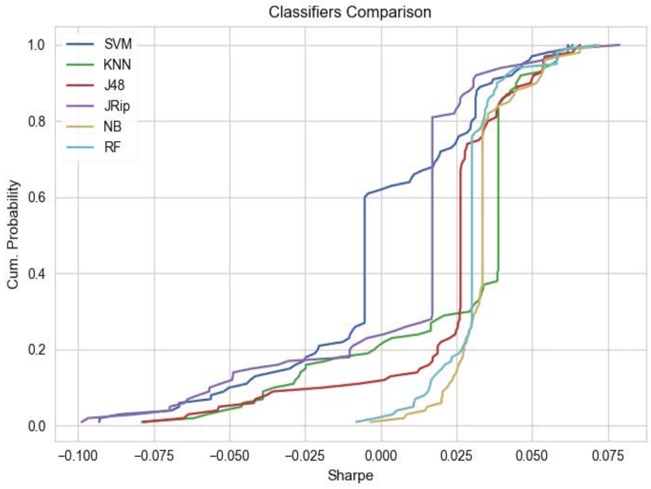


Figure 7. Performance Comparison of other classifiers

Figure 8 provides a graphical description of the overall results and shows that we proceed to every next step by taking into consideration the results of previous steps.

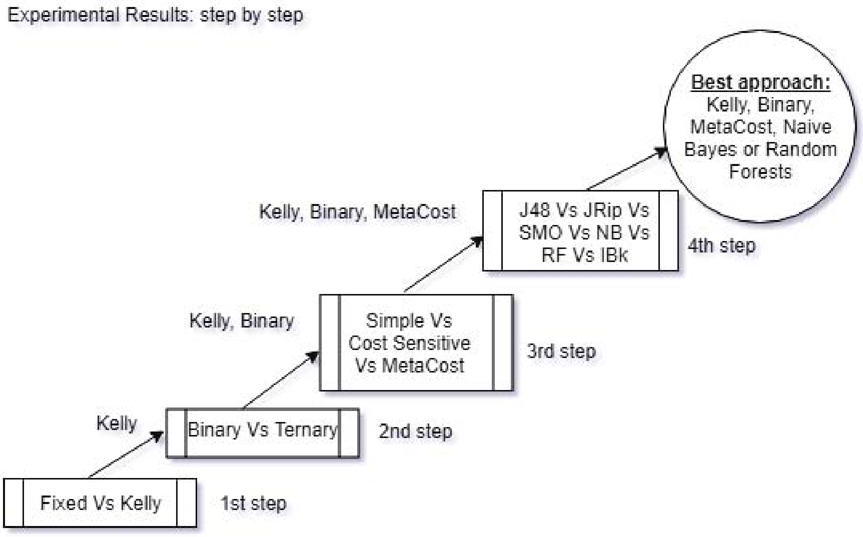

Experimental Results: step by step
Best approach:
Kelly, Binary, MetaCost, Naive Bayes or Random Forests
Kelly, Binary, MetaCost
J48 Vs JRip Vs SMO Vs NB Vs RF Vs IBk
4th step
Kelly, Binary
Simple Vs Cost Sensitive Vs MetaCost
3rd step
Kelly
Binary Vs Ternary
2nd step
Fixed Vs Kelly
1st step

Figure 8. Experimental Procedure and Results

Discussion

This work aims to provide a unique insight to the very challenging world of sports betting, examining the prediction of game results from an economical perspective. Evaluating the available literature, there are some evidence of inefficiencies in soccer betting market (Angelini & De Angelis, 2018), which means that many interesting approaches could be utilized to define efficient models and strategies in terms of profitability. In the core of all these approaches data mining was widely accepted as appropriate method to predict and explain events. Triggered by the absence of an alternative methodological context, our objective was not to extend the established studies and enhance already suggested models, but to propose a novel profit-oriented approach concluding to high quality results.

Our contribution both in academic and industry applications, was to examine the feasibility of a cost-sensitive approach in sports field in contrary to traditional machine learning methods along with the potential investment opportunities that may derived from our strategy. In order to examine the behavior of the cost-sensitive approach in soccer outcome prediction, we decided to avoid the use of exotic features and strategies or hybrid models that have been extensively performed by other studies for accuracy/profit optimization. Considering that those could affect the performance of our approach and distract us from a clear view of the results, we emphasized on the creation of a raw model using the simplest forms of the existing input attributes (i.e. shoots, wins, bookmaker odds, etc.), strategies (binary and ternary) and money management techniques (fixed stake size, Kelly criterion).

Among the challenges we met, was the difference in datasets and evaluation metrics used by previous studies that prevented us from a fair and clear comparison with our findings. As sports prediction and betting is a multi-disciplinary field, it is only logical that researchers utilize tools most relevant to their original field. Thus, direct comparison among studies is difficult. Nonetheless, we tried to describe our models with metrics widely utilized in machine learning and finance. Furthermore, due to the plethora of sports and sports associations, the datasets used vary among studies. Again, we tried to utilize a well-known league (English Premier League) from a widely popular sport (soccer) in order to have open and publicly available data in our dataset.

Although the proposed strategies greatly improved the performance of the models, we believe that there is still room for improvement. Taking into consideration the findings of relevant research studies, there are many potential extensions to our research work in terms of our model’s facilitation and sports outcome profitability. One of these extensions would be the inclusion of tipster data that are also used by Spann & Skiera (2009). Another extension would be the integration of a hybrid ranking method as proposed by Kyriakides et al. (2017) and Constantinou (2018). A third extension would be the use of betting exchanges odds regarding a potential improvement to our Sharpe ratio’s distributions, as Franck, Verbeek & Nüesch (2010) showed that these outperform classic betting odds. Lastly, it would be interesting to examine the performance of our approach exploiting unstructured data (i.e. tweets). According to Schumaker, Jarmoszko, & Labedz (2016) tweet sentiment outperforms wagering on odds-favorites, with higher payout returns versus odds-only but lower accuracy.

Conclusion

In this paper we employed a genetic algorithm in order to generate better features for machine learning algorithms. The generated models were used in order to virtually bet on soccer outcomes and calculate the Sharpe ratios of their equity curves. We used this approach in order to perform binary comparisons of various methods of generating predictive betting strategies.

We compared a fixed stake size with a variable stake size approach, a ternary versus binary classification scheme, a cost-sensitive versus non-cost-sensitive approach as well as a comparison of various learning algorithms. Cost-sensitive approaches seem to work better on this domain. This can be explained by the inherent imbalance in the dataset. In soccer matches, the home team wins almost 50% of all matches (Kyriakides et al., 2017). This means that the home-win class dominates the other two classes, having 100% more examples on average. This allows cost-sensitive classification to provide an informed rebalance of the classes.

We conclude that Kelly criterion as a money management and binary classification, focusing on the home team's win or loss, combined with a cost-sensitive classification, seemed to be the best approach. Further experiments suggested that the use of Naive Bayes or Random Forest as a base classifier performed best in the betting prediction task, evaluated by the Sharpe ratio of the produced equity curves.

Despite the room for improvement we feel that our experiments capture a different perspective in soccer forecasting. This paper could consist the inception of a new investigation sequence in sports betting.

References

Angelini, G., & De Angelis, L. (2018). Efficiency of online football betting markets. *International Journal of Forecasting, 35*(2), 712- 721. doi: 10.1016/j.ijforecast.2018.07.008

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3*(6).

Bishop, C. (2006). Pattern recognition and machine learning. *New York: Springer.*

Box, G., & Cox, D. (1964). An Analysis of Transformations. *Journal Of The Royal Statistical Society Series B*, *26*(2), 211-252.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Buursma, D. (2011). Predicting Sports Events From Past Results Towards Effective Betting On Football Matches. *In 14Th Twente Student Conference On IT*, Twente, Holland (Vol. 21).

Constantinou, A. C. (2018). Dolores: A model that predicts football match outcomes from all over the world. *Machine Learning, 108*(1), 49-75.

Constantinou, A., Fenton, N., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, *36*, 322-339. doi: 10.1016/j.knosys.2012.07.008.

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English Football League matches for betting. J*ournal Of The Royal Statistical Society: Series D (The Statistician)*, *51*(2), 157-168. doi: 10.1111/1467-9884.00308.

Dixon, M., & Pope, P. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal Of Forecasting, 20*(4), 697-711. doi: 10.1016/j.ijforecast.2003.12.007.

Dobravec, S. (2015, May). Predicting sports results using latent features: A case study. *In 2015 38th International Convention On Information And Communication Technology, Electronics And Microelectronics (MIPRO)* (pp.1267- 1272). IEEE. doi: 10.1109/mipro.2015.7160470.

Domingos, P. (1999). MetaCost: a general method for making classifiers cost-sensitive. *Proceedings of the 5th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining (KDD'99),* (pp.155-164). doi:10.1145/312129.312220.

Elkan, C. (2001). The Foundations Of Cost-Sensitive Learning. *Proccedings of the 17th international joint conference on Artificial Intelligence* (pp. 973-978). Seattle, WA, USA.

Eryarsoy, E., & Delen, D. (2019, January). Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods. *Proceedings of the 52nd Hawaii International Conference on System Sciences.* doi: 10.24251/HICSS.2019.136

Football Results. (2018). *Football-data.co.uk*. Retrieved 7 September 2018, from http://www.football-data.co.uk/.

Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal Of Forecasting*, *21*(3), 551-564. doi: 10.1016/j.ijforecast.2005.03.003.

Franck, E., Verbeek, E., & Nüesch, S. (2010). Prediction accuracy of different market structures—bookmakers versus a betting exchange. *International Journal of Forecasting*, *26*(3), 448-459.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I., & Trigg, L. (2017). Weka. *Data Mining And Knowledge Discovery Handbook*, (pp. 1305-1314). doi:10.1007/0-387-25465-x_62.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal Of Forecasting*, *21*(2), 331-340. doi: 10.1016/j.ijforecast.2004.08.002.

Goddard, J., & Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal Of Forecasting*, *23*(1), 51-66. doi: 10.1002/for.877.

Godin, F., Zuallaert, J., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2014). Beating the bookmakers: leveraging statistics and Twitter microposts for predicting soccer results. *Workshop on Large-Scale Sports Analytics, Proceedings.* Presented at the Workshop on Large-Scale Sports Analytics (KDD 2014).

Haaren, J., & Broeck, G. (2014). Relational Learning for Football-Related Predictions. *Latest Advances in Inductive Logic Programming*, 237-244.

Haaren, J., & Davis, J. (2015). Predicting The Final League Tables Of Domestic Football Leagues. *Proceedings of the 5th International Conference On Mathematics In Sport*, (pp. 202- 207).

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A Review Of Data Mining Techniques For Result Prediction In Sports. *Advances In Computer Science: An International Journal*, *2*(5), 7-12.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10- 18. doi: 10.1145/1656274.1656278.

Hand, D., & Yu, K. (2001*).* Idiot's Bayes: Not So Stupid after All?. *International Statistical Review / Revue Internationale De Statistique*, *69*(3), 385. doi: 10.2307/1403452.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal Of The Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381-393. doi: 10.1111/1467-9884.00366.

Karlis, D., & Ntzoufras, I. (2008). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal Of Management Mathematics, 20*(2), 133-145. doi: 10.1093/imaman/dpn026.

Koopman, S., & Lit, R. (2013). 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), 167-186. doi: 10.1111/rssa.12042.

Kyriakides, G., Talattinis, K., & George, S. (2014). Rating Systems Vs Machine Learning on the context of sports. *Proceedings Of The 18th Panhellenic Conference On Informatics - PCI '14*. doi: 10.1145/2645791.2645846.

Kyriakides, G., Talattinis, K., & Stephanides, G. (2017). A Hybrid Approach to Predicting Sports Results and an AccuRATE Rating System. *International Journal Of Applied And Computational Mathematics*, *3*(1), 239-254. doi: 10.1007/s40819-015-0103-1.

Kyriakides, G., Talattinis, K., & Stephanides, G. (2015). Raw Rating Systems and Strategy Approaches to Sports Betting. *In 5th International Conference on Mathematics in Sport* (pp. 97-102). Loughborough.

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes?. *Proceedings Of The 1st International Workshop On Utility-Based Data Mining - UBDM '05* (pp.69-77). doi:10.1145/1089827.1089836.

Le, J. (2019). A Tour of The Top 10 Algorithms for Machine Learning Newbies. Retrieved 20 June from https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies.

Michie, D., Spiegelhalter, D.J., Taylor, C.C. (eds). (2009). *Machine Learning: Neural and Statistical classification*. London: Overseas Press.

Odachowski, K., & Grekow, J. (2013). Using Bookmaker Odds to Predict the Final Result of Football Matches. *Lecture Notes In Computer Science*, (pp. 196-205). doi:10.1007/978-3-642-37343-5_20.

Provost, F., & Kohavi, R. (1998). Glossary of terms.*Machine Learning, 30*(2-3), (pp. 271-274).

Schumaker, R. P., Jarmoszko, A. T., & Labedz Jr, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems*, *88*(C), 76-84. doi: 10.1016/j.dss.2016.05.010

Scibilia, B. (2012). How Could You Benefit from a Box-Cox Transformation?. [The Minitab Blog.] Retrieved September 10 2018 from http://blog.minitab.com/blog/applying-statistics-in-quality-projects/how-could-you-benefit-from-a-box-cox-transformation.

Sharpe, W. (1994). The Sharpe Ratio. *The Journal Of Portfolio Management*, *21*(1), 49-58. doi: 10.3905/jpm.1994.409501

Sheng, V., & Ling, C. (2009). Cost-sensitive learning. In J. Wang, *Encyclopedia of Data Warehousing and Mining* (2nd ed.), (pp. 339-345).

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal Of Forecasting*, *28*(1), 55-72. doi:10.1002/for.1091.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293-300.

Ting, K. (1998). Inducing cost-sensitive trees via instance weighting. *Principles Of Data Mining And Knowledge Discovery*, (pp. 139-147). doi:0.1007/bfb0094814.

Witten, I., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.