Available online at www.sciencedirect.com

ScienceDirect

International Journal of Forecasting 26 (2010) 518–536

www.elsevier.com/locate/ijforecast

Alternative methods of predicting competitive events: An application in horserace betting markets

Stefan Lessmann[a,*], Ming-Chien Sung[b], Johnnie E.V. Johnson[b]

[a] Institute of Information Systems, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany
[b] Centre for Risk Research, School of Management, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Abstract

Accurately estimating the winning probabilities of participants in competitive events, such as elections and sports events, represents a challenge to standard forecasting frameworks such as regression or classification. They are not designed for modeling the competitive element, whereby a specific participant's chance of success depends not only on his/her individual capabilities but also on those of his/her competitors. In this paper we consider this problem in the competitive context of horseracing and demonstrate how Breiman's (2001) random forest classifier can be adapted in order to predict race outcomes. Several empirical experiments are undertaken to demonstrate the features of the adapted random forest procedure and confirm its effectiveness as a forecasting model. Specifically, we demonstrate that predictions derived from the proposed model can be used to make substantial profits, and that these predictions outperform those from traditional statistical techniques.

Keywords: Probability forecasting; Classification; Random forest; Sports forecasting

1. Introduction

This paper considers the task of predicting the outcomes of competitive events like political elections or sport events. These events are competitive in the sense that a given candidate's or sport team's chances of winning depend not only on variables concerning his/her own 'credentials', but also on factors related to the opponents' skills. Such characteristics are evident in a variety of situations, and are examined in the context of forecasting sports events throughout the paper; specifically, in forecasting the outcomes of horseraces.

From a forecasting perspective, the presence of competition introduces particular modeling challenges. As will be explained in detail below, these limit the applicability of standard techniques such as regression and discriminant analysis. Consequently, dedicated forecasting methods are required to accurately model the outcomes of competitive events, and horseraces in particular.

[*] Corresponding author. Tel.: +49 40 42838 4706; fax: +49 40 42838 5535.
E-mail addresses: lessmann@econ.uni-hamburg.de (S. Lessmann), ms9@soton.ac.uk (M.-C. Sung), jej@soton.ac.uk (J.E.V. Johnson).

0169-2070/$ - see front matter © 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2009.12.013

The conditional logit (CL) model (McFadden, 1974) has become the conventional method for horserace forecasting (see, e.g., Bolton & Chapman, 1986; Chapman & Staelin, 1982; Figlewski, 1979; Johnson, Jones, & Tang, 2006), because it maintains the connections between individual runners in a given race. That is, CL can extract information from the composition of a race, and considers opponents' strengths when assessing a given runner's likelihood of winning, whereas many of the alternative techniques fail to account for such *within-race* competition. However, despite this conceptual advantage, CL also suffers important limitations. For example, the estimation is based on distributional assumptions which may be violated in practical applications. In addition, it is unable to capture nonlinear relationships between the dependent and independent variables, unless they are specified explicitly by the modeler. Support vector machine (SVM) based procedures have been proposed to overcome these problems (Edelman, 2007; Lessmann, Sung, & Johnson, 2007, 2009), since they do not rely on any distributional assumptions and can approximate any nonlinear function to an arbitrary level of accuracy. However, SVMs are black box procedures and do not enable the predictive contributions of individual variables to be discerned. Furthermore, due to the computational complexity associated with constructing SVM-based prediction models, it is difficult to accommodate the large datasets often encountered in modeling horserace results.

The random forest (RF) procedure, a state-of-the-art machine learning method proposed by Breiman (2001), has features which enable the limitations of both CL and SVM-based prediction models to be overcome. First, its ability to form highly accurate predictions has been verified in several empirical studies (see, e.g., Caruana & Niculescu-Mizil, 2006; Chen, Liaw, & Breiman, 2004; Hamza & Larocque, 2005; Leon, Zaharia, & Gâlea, 2004; Lessmann, Mues, Baesens, & Pietsch, 2008). Second, RFs can identify the importance of individual variables and are capable of discerning nonlinear interactions between variables. Third, the RF procedure is fast and scalable, so that large datasets can be handled with ease. Finally, it might also be hypothesized that RFs are well-suited for predicting the results of events which are associated with betting markets (e.g., horseraces), as there are similarities between the ways in which RFs and betting markets aggregate information.

Despite their advantages, RFs do not account for competition between participants directly. Consequently, the aim of this paper is to develop a methodology for adapting the RF procedure in order to make it suitable for predicting the outcomes of competitive events. We will demonstrate the effectiveness of the proposed procedure through an empirical evaluation using a newly acquired dataset of 1000 races (involving 12,902 horses) run in Hong Kong. In line with Leitch and Tanner's (1991) argument that a model's profitability is the primary indicator of forecasting accuracy, we demonstrate that it is possible to make substantially greater profits based on the model's predictions than are achievable using predictions from a well established benchmark model. Statistical indicators of forecasting accuracy are also used to contrast the model's performance with that of the benchmark.

The paper makes a number of important contributions: first, it is the first to apply a RF procedure in the context of competitive event prediction in general, and to horserace betting in particular. Second, it introduces a methodology which better accounts for within-race competition within a RF framework and provides evidence of the value of such an adaptation. Third, it examines the robustness of RF in detail with respect to setting of its parameters, and demonstrates how these may be determined in practical applications. Fourth, it demonstrates how measures of variable importance can be obtained from the RF classifier, in order to clarify the types of information that influence the model's predictions.

The paper is organized as follows: an introduction to horserace forecasting with CL is given in Section 2. This is followed by a description of the RF methodology for ordinary discriminant analysis and an elaboration of our proposed adaptation of RF for horserace prediction. In Section 3, the results of several empirical experiments assessing the RF model's effectiveness are presented. Conclusions are drawn in Section 4.

2. Horserace forecasting methodology

The task of effectively predicting the outcomes of horseraces has been examined extensively, particularly in studies with the objective of shedding light on the informational efficiency of horserace betting markets (see Sung & Johnson, 2008, for a summary). Considering Fama's (1970) well-known market efficiency hypothesis, a market's efficiency determines the extent to which market prices discount publicly available information. Therefore, building a horserace forecasting model from such information and verifying that betting on its predictions generates abnormal returns facilitates the drawing of conclusions regarding the betting market's informational efficiency (Johnson et al., 2006). Consequently, the forecasting objective in horserace prediction is to estimate all runners' probabilities of winning the race as accurately as possible.

2.1. Predicting winning probabilities with conditional logit regression

CL regression has been successfully applied in several marketing applications for modeling choice decisions (see, e.g., Chapman & Staelin, 1982), and was introduced as a suitable technique for forecasting horserace results by Bolton and Chapman (1986). Its distinctive advantage for this task stems from the fact that CL maintains the connections between the alternatives of a choice set (e.g., between runners in a race). That is, CL enables the winning probability of one horse to be estimated in conjunction with those of the other runners in a race, and thus accounts for within-race competition.

Let $S = \{\boldsymbol{x}_i^j, \boldsymbol{y}_i^j\}_{i=1,j=1}^{n_j,R}$ denote a dataset of R past races, where $\boldsymbol{x}_i^j \in \Re^m$ represents a single horse i in race j that is characterized by m independent variables, and n_j is the number of horses in race j. The aim of a CL horserace forecasting model is to predict the vector of winning probabilities $\boldsymbol{p}^j = \left(p_1^j, p_2^j, \ldots, p_{n_j}^j\right)$ for race j, where the component p_i^j represents the estimated model probability of horse i winning race j. Let W_i^j be a binary win/lose indicator variable, defined as:

$$W_i^j = \begin{cases} 1 & \text{if } \boldsymbol{\beta} \cdot \boldsymbol{x}_i^j + \varepsilon_i^j > \boldsymbol{\beta} \cdot \boldsymbol{x}_l^j + \varepsilon_l^j \\ & \forall i, l = 1, \ldots, n_j, i \neq l \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of coefficients that measure the relative contributions of the independent variables contained in the vector $\boldsymbol{x}_i^j$, and the error term ε_i^j represents unperceived information. It can be shown that, if the errors are independent and distributed according to the double exponential distribution, the probability of horse i winning race j is given by the following CL function (McFadden, 1974):

$$p_i^j = \frac{\exp\left(\boldsymbol{\beta} \cdot \boldsymbol{x}_i^j\right)}{\sum_{i=1}^{n_j} \exp\left(\boldsymbol{\beta} \cdot \boldsymbol{x}_i^j\right)}. \tag{2}$$

The model's coefficients, $\boldsymbol{\beta}$, are estimated by means of the maximum likelihood procedure (see, e.g., Johnson et al., 2006).

The particular suitability of the CL model for competitive event forecasting is emphasized when comparing its probability estimates (Eq. (2)) to those of an ordinary binary logit model (BLM).[1] Specifically, BLM estimates the *a posteriori* probability of the event *win*, given a set of independent variables associated with horse i:

$$p(win|\boldsymbol{x}_i) = 1/1 + \mathrm{e}^{-\boldsymbol{\beta} \cdot \boldsymbol{x}_i}. \tag{3}$$

It is important to note that the model's coefficients, $\boldsymbol{\beta}$, are estimated by maximizing the likelihood over the *whole* dataset, without considering which runners compete in a given race. In other words, a horse's individual probability of being a winner (in some race) is estimated. This notion is stressed in Eq. (3) by dropping the race index j. Consequently, one may argue that this modeling paradigm assumes that winning horses are conceptually different from nonwinning horses, and that all (available) information which could possibly affect a runner's chances is contained in the independent variables. This assumption is violated in competitive scenarios, and in horseracing in particular, since a single runner's chance of winning clearly also depends upon its competitors' strengths. Therefore, it is preferable to forecast a runner's winning probability in relation to those of its opponents, as is achieved in CL.

[1] It should be noted that multinomial logit models (MLM) are not applicable for horserace forecasting, since the number of runners varies across races. As a consequence, the number of 'groups' varies, which prohibits the use of MLM.

Despite CL's advantage with respect to accounting for within-race competition, it suffers from some important limitations. In particular, the application of CL assumes that errors are distributed according to the double exponential distribution, which is unlikely in practice (Gu, Huang, & Benter, 2003). In addition, the CL model is unable to capture nonlinear interactions among independent variables and between independent variables and the target. That is, an independent variable's weighted (i.e. through $\boldsymbol{\beta}$) influence on the target is linear and additive. Furthermore, arguments from statistical learning theory (Vapnik, 1995) indicate that a maximum likelihood based estimation of the coefficients may be unstable if a large number of independent variables are used, as is commonly the case in horserace forecasting (Edelman, 2007).

2.2. *Random forests*

The RF procedure (Breiman, 2001) is a state-of-the-art machine learning method that has proven its potential for constructing accurate forecasting models in a variety of applications (see, e.g., Burez & Van den Poel, 2007; Caruana & Niculescu-Mizil, 2006; Coussement & Van den Poel, 2008; Guo, Ma, Cukic, & Singh, 2004; Hamza & Larocque, 2005; Lariviere & Van den Poel, 2005; Leon et al., 2004; Lessmann et al., 2008). Furthermore, RFs have various features which make them well suited for competitive event forecasting. In particular, they avoid the potential problems of CL indicated above. That is, RF is a non-parametric technique which is not based on any distributional assumptions and is able to automatically extract nonlinear relationships from data (Breiman, 2001). In addition, RF is able to provide comprehensible estimates of the importance of different variables, whereas most alternative nonlinear methods, such as artificial neural networks or SVMs, function as black-boxes. RF is also preferable to these alternatives from a computational perspective. Specifically, model building is fast and can easily be organized in a parallel fashion. On the other hand, standard RFs are restricted to the discriminant analysis setting, and are therefore unable to account for within-race competition. Consequently, RF can be considered as an approach which complements rather than replaces CL-based horserace forecasting. Therefore, the following sections will first introduce the ordinary RF approach for discriminant analysis, and subsequently explain how an integration of RF and CL for horserace forecasting may be achieved.

2.2.1. *Discriminant analysis*

Formally, the discriminant analysis setting involves predicting the membership of objects in *a priori* known groups. This is related to the BLM approach, where the objects are runners and the groups are winning and non-winning horses. In particular, BLM can be used for discriminant analysis by introducing a threshold value (say, 0.5), such that horses whose *a posteriori* probability (Eq. (3)) exceeds this value are classified as winners. Other discriminant analysis models can produce such a discrete prediction directly.

RF is based upon the well-known CART (classification and regression trees) methodology (Breiman, Friedman, Olshen, & Stone, 1984) for constructing *decision tree models*. Such models involve a recursive partitioning of the dataset S to separate out examples (i.e. horses) of different groups. This principle is illustrated in Fig. 1 using artificial data of two four-runner races, where horses are characterized by two independent variables.

Generally, a decision tree can be described as a collection of *nodes* which are arranged in a hierarchical fashion. Nodes serve as containers for the objects to be classified (i.e. horses). In a first step, all runners enter a *root* node, which is the first node in the hierarchy. This node exhibits high heterogeneity, in the sense that it contains both winning and non-winning horses (objects of different groups). The tree building algorithm strives to reduce the heterogeneity by branching a node into two more homogeneous sub-nodes, which are commonly referred to as inner-nodes (see Fig. 1). In particular, a specific value of one independent variable is selected and used to define a *splitting rule* that determines which runners should enter which sub-node in order to maximize the homogeneity within the sub-nodes. For example, Fig. 1 illustrates a split based on the variable *average performance rating*, depending on whether or not runners achieved a rating greater than or equal to seven. This branching is continued until all nodes contain only winning or non-winning horses. This process leads to a hierarchical, tree-like structure,

	Winner	No. of past wins	APR*
Race 1			
Horse 1	**1**	**13**	**8**
Horse 2	0	8	6
Horse 3	0	6	8
Horse 4	0	5	4
Race 2			
Horse 5	0	6	4
Horse 6	0	1	3
Horse 7	**1**	**8**	**7**
Horse 8	0	0	1

* APR = average performance rating of horse in its previous races

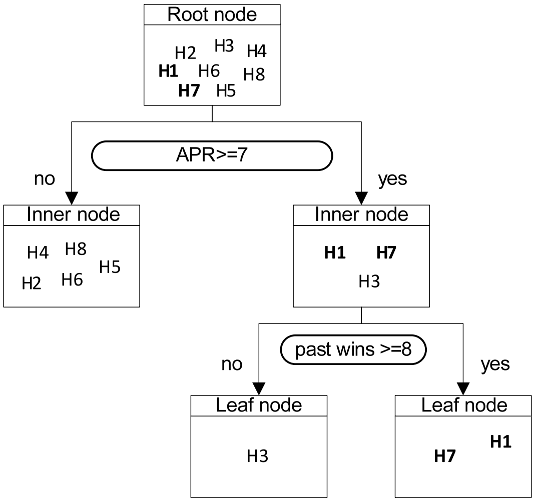

Fig. 1. Illustration of the decision tree approach to building forecasting models for discriminant analysis using artificial data from two four-runner races. Each runner is characterized by two independent variables. The binary dependent variable is denoted by *winner*, and bold face is used to highlight the horse which finished first in each race.

where the terminal nodes (i.e., those that are not branched) are termed leaf nodes.

The final tree can be converted into a set of rules that facilitate prediction. For example, Fig. 1 would suggest that horses which achieved an average performance rating of seven or above and had won at least eight races in the past should be classified as winners, whereas all other horses are predicted to lose. This illustrates the way in which a tree model can be used to generate discrete horserace forecasts. However, the example also emphasizes the inability of discriminant analysis to capture the competitive element of horseracing (e.g., a model could predict that a race has multiple ‘winners’). We explain the way in which this limitation can be overcome in the context of the proposed forecasting model in Section 2.2.2.

A key ingredient of any tree model is the procedure for identifying the independent variable and the respective ‘split value’ for branching, and therefore distinguishing between objects of different classes. This is achieved by means of the *Gini*-Index (*GI*) within the CART framework. For the case of binary discriminant analysis, *GI* can be computed as follows (see for example Duda, Hart, & Stork, 2001):

$$GI_{Node} = 1 - p\,(win)^2 - p\,(lose)^2\,, \tag{4}$$

where $p\,(\cdot)$ represents the prior probability of a class within a given node. Clearly, the minimum of *GI* is zero, when a node contains only winning or non-winning horses, whereas a value of 0.5 (i.e., both groups are equally probable) indicates maximal impurity. Hence, the reduction of impurity is used to govern the model building. That is, any time a node is divided, all independent variables are assessed in terms of their ability to decrease the impurity through branching at a particular value. The split which achieves the largest impurity reduction is conducted, and the resulting two new (sub- or inner-) nodes are inserted into the tree. This procedure is continued until all leaf nodes (final nodes) contain only examples of a single class.

RF embraces the CART methodology and extends it by constructing multiple tree models, whose predictions are averaged to form a final forecast. Such a combination and aggregation of different models is beneficial for increasing the (*meta*-) model’s forecasting accuracy, as well as its robustness toward distortions in the training data (see, e.g., Breiman, 1996; Freund & Schapire, 1997). However, these advantages can only be achieved if the individual base models truly complement each other. That is, a combination of forecasting models whose predictions are mostly correlated is futile. Therefore, RF incorporates two mechanisms to construct *diverse* CART models (Breiman, 2001): first, each individual decision tree is derived from a *bootstrap* sample of size *N*, drawn from the original training dataset *S* with replacement. This results in randomization, because each sample could include some examples multiple times, while others (approximately one-third of the data) are ignored (see, e.g., Efron & Tibshirani, 1993). In addition, the diversity is increased further by modifying the node splitting procedure within CART. Specifically, instead of searching among all m independent variables, a subset $Z < m$ is selected at random any time a node is split. Ho (1998) was the first to propose this *random subspace* procedure for increasing the diversity among a collection of tree models. As a side benefit, this approach also significantly improves RF's computational efficiency by accelerating the search for an appropriate split.

In summary, a *random forest* is a forecasting model that comprises a collection of decision trees (representing a forest), each of which has been randomized by means of *bootstrapping* and the *random subspace* procedure in order to achieve diversity. To predict the group membership of a novel example (e.g., to decide whether or not a given runner should be classified as a winning horse), all CART models cast a vote on the example's group using their internal rule base (Fig. 1). Then, the final prediction corresponds to the group which receives the most votes.

In comparison to alternative approaches for combining multiple forecasting models, such as bagging (Breiman, 1996) and boosting (Freund & Schapire, 1997), RF has the advantage of embracing the former and being less susceptible to noise than the latter (Chen et al., 2004; Dietterich, 2000; Hamza & Larocque, 2005). Furthermore, RFs also provide the opportunity to easily measure the importance of different variables after all trees are fully developed. In particular, each split upon an independent variable within a single tree produces two nodes whose *Gini*-impurities are less than that of the preceding node. Consequently, adding up the decrease of *GI* over all nodes in the forest (i.e., over all nodes per tree and for all trees) that are split on a particular variable gives an estimate of this variable's relevance for forming class predictions (Breiman, 2001).

2.2.2. Forecasting horserace outcomes with random forests

The key objective in horserace forecasting is to estimate a runner's chances of winning. Consequently, in employing RF, the first step involves augmenting the discrete model predictions (i.e. winner/loser) in order to enable a comparison of the scores of competing runners. This may be achieved by counting, for each runner, the number of CART models within RF that predict it being a winner. The resulting figure can then be interpreted as the RF model's estimate of a horse's ability (see the Appendix for a formal description). Furthermore, the ability scores could be considered as a 'forecast' of the full finishing order of a race (i.e., the horse with highest ability score is predicted to win, the one with the second highest score is predicted to finish second, etc.). Such a modeling approach is closely related to regression (i.e., forecasting a continuous dependent variable, such as the runner's finishing position, as was employed by Benter, 1994, and Edelman, 2007), and may enable further information to be extracted from the rank-ordered finishing data. On the other hand, the reliability of this information is questionable (particularly for horses which finish beyond third, since jockeys have little incentive to run a horse to its full potential once it becomes obvious that they are not going to secure a prize). Such a bias could severely influence regression-based models, whereas binary discriminant analysis remains unaffected (see Lessmann et al., 2008; Sung, Johnson, & Bruce, 2005, for empirical evidence). Consequently, the approach employed in this study strives to integrate the respective advantages of discriminant analysis and regression, while avoiding their disadvantages. To that end, an indicator of ranking accuracy, known as the normalized discounted cumulative gain (NDCG) (Järvelin & Kekäläinen, 2000), is defined, which ensures that only those placings which are associated with prize-money are incorporated in the model development. A detailed example of using NDCG in the context of horseracing is given by Lessmann et al. (2007), and the implementation of NDCG is discussed in Section 3.2.

The major challenge in effectively applying RF to forecasting the outcomes of horseraces is overcoming the general limitations of discriminant analysis (as well as regression) concerning the modeling of within-race competition. To achieve this, the CL model may be used to post-process a RF model's predictions, so as to enhance its ability to forecast competitive scenarios. Therefore, a *two-stage* forecasting methodology is proposed, based on the framework initially presented by Benter (1994) and Edelman (2007). They suggest conducting a regression to model the relationship between the variables which capture a horse's fundamental ability (e.g., its average running speed in previous races) and its normalized finishing position. This forms a *first stage model* which generates forecasts of runners' normalized finishing positions in future races. These forecasts can be interpreted as an aggregation of a horse's strength (*ability index*), and are then fed into a *second stage model* using a CL function. Consequently, this second stage CL model incorporates two input variables: (1) the *ability index*, and (2) market prices (odds), which represent the public's opinion of a horse's prospects of winning a race.

The two-stage modeling framework offers two major advantages over a single stage forecasting model. First, the need to account for within-race competition is postponed to stage two, and a number of advanced forecasting algorithms can be employed to capture the subtle and sophisticated relationships between the fundamental variables concerning a horse's ability and the outcome of races in the first stage. For example, Edelman (2007) proposed a support vector regression model for processing a large number of highly correlated variables (as is generally encountered in horserace predictions), to help reveal the nonlinear relationships between the input variables. Similarly, Lessmann et al. (2009) demonstrated that SVM classification models, which categorize horses as either winners or losers, are a promising alternative to regression. Second, the segregation of fundamental variables and market odds has been shown to be beneficial (Sung & Johnson, 2007) because the latter are known to be powerful predictors of race outcomes (see, e.g., Bruce & Johnson, 2000). Consequently, a model which simply combined the fundamental variables and market odds in one stage would be geared towards market odds, and the independent variables would receive less attention. This is undesirable because odds reflect the public's opinion of a horse's chance of winning, and to a substantial extent comprise the publicly available information. Therefore, profitable betting only becomes possible if a forecasting model succeeds in distilling additional information (e.g., from fundamental variables) over and above that which has already been discounted in the market odds.

The forecasting model proposed in this study is similar to that of Edelman (2007), but employs the RF methodology in the first stage for the reasons outlined above. In addition, it is hypothesized that the integration of diverse base models, as postulated by the RF framework, makes the proposed model particularly appropriate for forecasting horseraces. This is because the first stage mimics, to some extent, the decision-making processes that ultimately lead to the formation of market odds. That is, a given horse is evaluated by several bettors with different levels of expertise (akin to the individual CART models in RF). The bettors presumably consider different sets of publicly available information (different independent variables) when making their decisions. They then place their bets (vote) based on their assessment of a horse's ability. Aggregating these assessments produces the market prices.

In summary, the proposed forecasting paradigm consists of the following three steps. First, a subset of R_1 races is drawn from the dataset S, and is used to build a RF forecasting model (with market odds being excluded from the list of independent variables). Second, RF is used to produce an ability forecast for the horses in the remaining $R - R_1$ races. Third, the coefficients of a CL regression model are estimated by means of maximum likelihood over the same $R - R_1$ *second stage races*, considering two independent variables: market odds and the RF-based ability forecasts. This step is a post-processing procedure that relates the RF predictions of individual runners to the race context (i.e., CL models races, rather than runners, as entities). Consequently, it reintroduces within-race competition and ensures that the predicted winning probabilities for a given race sum to one.

Having built the RF model and estimated the coefficients of the second-stage CL, novel races may be forecast in the same way (i.e., by generating ability forecasts by means of RF and pooling them with market odds through CL). The overall architecture of the proposed model is illustrated in Fig. 2, and a formal description is provided in the Appendix.

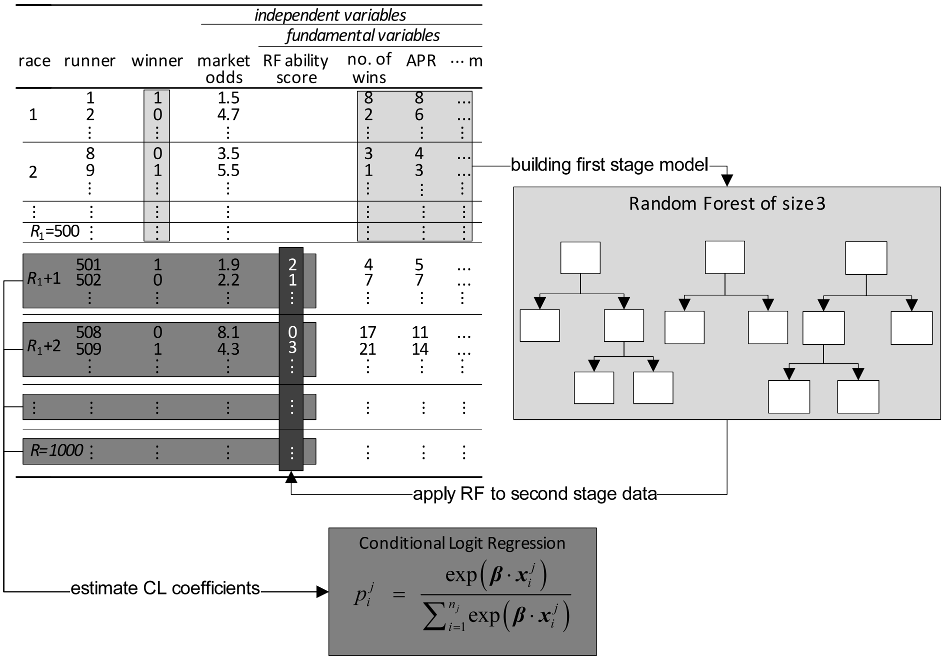

Fig. 2. Architecture of the two-stage RF forecasting model for predicting the outcomes of competitive events, assuming a RF of size three, a dataset containing 1000 races, and the same number of races being used for the first and second modeling stages.

3. Empirical analysis of the RF-based forecasting model

3.1. Experimental setup

The empirical analysis is based on a dataset of 1000 races run at Hong Kong racetracks between 1st January 2005 and 26th December 2006. The database contains details of 12,902 horses, and the number of runners per race varies between seven and fourteen. Each runner is described by a set of forty fundamental variables that are related to its performance in previous races, as well as to its preferences in relation to the current race (e.g., distance, going, etc.). The variables represent a more sophisticated set of attributes than those used in Bolton and Chapman's (1986) seminal paper on horserace forecasting. A brief description of the explanatory variables used in this study is given in Table 1.

The dataset is split evenly into a training set for model development and a holdout test set for out-of-sample evaluation. It has been argued that the true value of forecasting models lies mainly in their profitability (see, e.g., Granger & Pesaran, 2000; Leitch & Tanner, 1991). This is particularly true in a financial market context, and is consistent with the predominant research objective in horserace forecasting, namely, to shed light on the informational efficiency of betting markets. Specifically, to demonstrate that the betting public fail to fully account for publicly available information, it has to be shown that betting on the model's predictions generates a profit. Consequently, to explore the value of the RF approach developed here, the model's predictions are employed in a Kelly wagering strategy (Kelly, 1956). Given the estimated winning probabilities (p_i^j) calculated from Eq. (A.2), the Kelly strategy determines how much should be bet on each horse in a race in order to maximize the expected log payoff in the long run. It has been shown to be optimal in the sense that it maximizes the asymptotic rate of wealth growth, with a zero probability of ruin (Breiman, 1961). Consequently, the rate of return of a forecasting model, as produced by simulating Kelly-betting over the holdout test races, is used as the primary indicator of forecasting accuracy in this study. In addition, a detailed analysis of a model's winning

Table 1
Definitions of the 41 independent variables employed in the empirical evaluations.

Independent variables	Variable definitions
Market-related variable	
	The natural logarithm of the normalized track probabilities.
Fundamental variable categories	
	Twelve variables related to a horse's past performances (HPP): variables to capture past race incidents (e.g., jockey loses whip), trackwork, strength of past races, and beaten lengths in previous races.
	Three variables related to a horse's most recent performance on and off the track (HMR): variables to capture its speed, and the distance and strength of its recent races.
	Nine variables related to a horse's specific preferences (HSP): variables to capture a horse's preference for blinkers, track, distance, surface, and jockey.
	Two variables related to a jockey's past performances (JPP).
	Three variables related to a trainer's characteristics and past performances (TPP).
	One post position bias variable (PPB): accounts for a bias at each track.
	Three variables related to the weight carried by the horse in current and past races (WGH): variables to capture the weight of horse, recency adjusted weight carried in previous races, and weight carried in the current race.
	Two variables related to the horse's age (AGE).
	Two variables related to the number of days since the last run (DSL).
	Two variables related to the conditions of today's race (CDR): variables to capture the distance and surface of the current race.

probability estimates is conducted, to give a better understanding of the origins of profitability.

A two-stage CL model (Sung & Johnson, 2007) is employed as a benchmark against which the performance of the RF-based two-stage model is assessed. This model adopts the two-stage paradigm, but also uses CL regression in stage one (instead of RF) to process the horses' fundamental variables and to generate an estimate of the horses' abilities. Sung and Johnson (2007) demonstrate that in a horseracing context, this type of model is superior to a standard CL procedure which combines the fundamental variables and market odds in one step.

In the following, the term '2RF model' is used to refer to the two-stage RF model proposed in Section 2.2.2, whereas '2CL model' refers to the two-stage CL benchmark model.

Before the forecasting accuracies of the 2RF model and the 2CL model can be compared, one must ensure that RF is properly adapted to the given dataset, meaning that suitable settings for the number of individual CART models (T) within the RF and the number of independent variables drawn at random each time a tree node is split (Z) have to be determined. Therefore, the empirical study begins by presenting an appropriate tuning procedure and examining the robustness of RF to different parameter values. This demonstrates how the model could be used in practice and scrutinizes its sensitivity. Subsequently, comparisons between the 2CL and the (tuned) 2RF model on the holdout test set are undertaken in order to examine the effectiveness of the latter in predicting accurate winning probabilities. Finally, we demonstrate how measures of variable importance can be extracted from the 2RF model.

3.2. Parameter selection and sensitivity

A popular technique for tuning the parameters of machine learning methods is the *grid-search* approach. It involves defining a set of candidate values for each parameter and empirically assessing all possible value combinations (see, e.g., Coussement & Van den Poel, 2008; Van Gestel et al., 2004). Candidate values for RF, based on those identified in the literature (e.g., Breiman, 2001; Hamza & Larocque, 2005; Lessmann et al., 2008), have been set to $T = \{100, 200, \ldots, 800\}$ and $Z = \{0.25, 0.5, \ldots, 2\} \cdot \sqrt{m}$, yielding an overall number of 64 combinations to be assessed.

The computational burden associated with evaluating several parameter values can be reduced when exploiting the nature of the bootstrap sampling underlying RF. Since approximately one third of the examples do not enter the training dataset for an individual

CART tree, a set of 'fresh' holdout sample cases is readily available to measure the forecasting accuracy of this tree (Breiman, 2001). Such instances are called *out-of-bag* (oob) examples, and, due to their availability, costly procedures like cross-validation can be avoided when tuning the parameters of the RF classifier.

To assess the different parameter settings, we compare the accuracies of the respective models. However, it is important to remember that winning probabilities are not generated until stage two, and, as a consequence, the rate of return cannot be used for model evaluation at this stage. To alleviate this problem, Lessmann et al. (2007) propose using the NDCG measure, which assesses the ability of a prediction model to accurately rank finishers. In particular, NDCG allows us to restrict the influence of rank orderings to the first three finishing positions (i.e. those who won prize-money), and therefore avoids unreliable rankings among minor placings. This approach is intuitively appropriate, because a model which fails to identify the winning horse but ranks it second should be considered superior to a model that assigns it a lower rank (see Lessmann et al., 2007, for a detailed description).

The results of evaluating the 64 combinations of T and Z values in terms of NDCG on oob examples are depicted in Fig. 3. A perfect prediction (i.e. ranking the top three finishers in the right order) is associated with a value of one (100%), whereas decreasing values indicate inferior performance. The best result (71.45%) is achieved with a forest of 600 trees, each of which uses 4 randomly selected independent variables for splitting. Consequently, this set of parameters is retained for measuring the profitability of the final 2RF model in later experiments.

It can also be seen from Fig. 3 that the variation along the Z axis appears to be greater than that along the T axis, indicating that the number of variables used to grow one CART model has a stronger impact on NDCG than forest size. This hypothesis can be scrutinized by means of the Friedman-test, which was

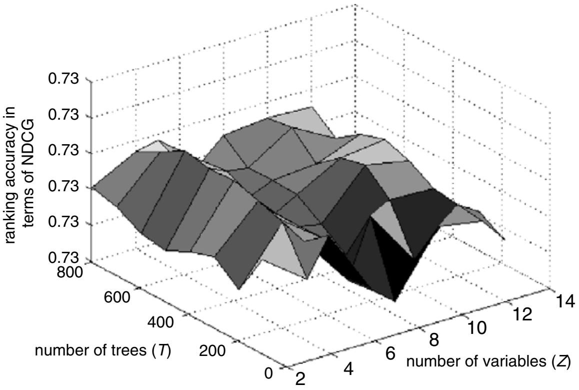

Fig. 3. Dependency of the parameter settings of the number of trees per forest (T) and the number of variables selected at random to branch a node during tree growing (Z) in terms of NDCG for the first-stage RF classifier.

Table 2
Results of the *Friedman*-test scrutinizing the significance of the impact of the number of variables per branch (Z) and the number of trees (T), respectively, on RF accuracy in terms of NDCG.

Source	Sum of squares	Degrees of freedom	Mean square	Chi-square	Prob >Chi-square
Number of variables per node branch (Z)	230	7	32.858	38.33	2.619×10^{-6}
Error	106	49	2.163		
Total	336	63			
Number of trees (T)	59.5	7	8.5	9.92	0.1933
Error	276.5	49	5.643		
Total	336	63			

recommended for classifier assessments by Demšar (2006). This test is a nonparametric alternative to the analysis of variance (ANOVA), and relies on less restrictive assumptions (see, e.g., Zar, 2007, for details). Since the Friedman-test can only consider one factor at a time, it is sequentially applied to check first the significance of the impact of Z, then that of T on NDCG. The results of the Friedman-test are shown in Table 2. The significant effect of Z is confirmed (p-value < 0.0000), while the p-value of 0.1933 for T suggests that the number of trees in the forest may not affect model performance significantly.

The results displayed in Table 2 also help to explain the influence of the RF parameters on NDCG for the data considered in this study. The insignificance of T is consistent with previous research, in the sense that RF is generally credited with being robust to the settings of its parameters (see, e.g., Breiman, 2001). However, it seems precipitate to conclude that this parameter does not significantly affect the performance, and should thus be set to a low value, with regard to computational efficiency for example. Similarly, the fact that Z is found to be significant does not allow any conclusions to be drawn regarding the superiority of the 'best' setting, $Z = 4$, over other candidate values. Therefore, further analysis seems desirable to improve our understanding of the behaviour of RF with different parameter settings.

To achieve this, a *what-if* analysis is undertaken to measure the profitability of the full 2RF model that would have resulted if a given parameter setting had been selected. This involves one RF parameter being fixed at its 'best' setting (as determined by grid-search) while the other parameter is varied over a large interval. An RF classifier is then constructed for each parameter value in order to classify the horses in the holdout sample. These predictions are post-processed by a second stage CL model to estimate the winning probabilities for all holdout sample runners, then the probabilities are employed in a Kelly-wagering strategy to determine the rate of return. These returns are displayed in Figs. 4 and 5 for different values of the parameters Z and T, respectively. It has to be emphasized that this type of analysis is strictly theoretical, in the sense that, in practical applications, some of the information used (i.e. the races of the holdout test set) would not be available at the time when parameters have to be determined. However, the experiment is a valuable enhancement of the statistical analysis. In particular, it allows an examination of the impact of parameter variations on profit, which enables us to verify the stability of the proposed model selection mechanism (i.e. grid-search) in determining suitable settings, and therefore the appropriateness of NDCG for guiding the search for predictive parameter values.

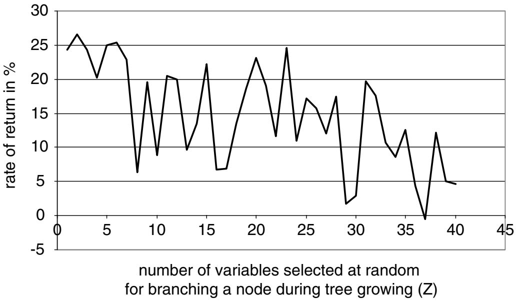

Fig. 4. What-if analysis regarding the profitability of RF models with a varying number of independent variables selected at random for branching a node during the growing of individual CART trees.

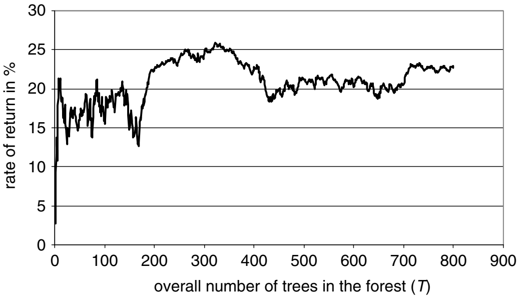

Fig. 5. What-if analysis regarding the profitability of RF models with a varying forest size.

When comparing Figs. 4 and 5, the higher impact of Z on the RF model is confirmed. In particular, larger variations across candidate values, as observed in terms of NDCG, are reflected in larger variations in terms of profitability. This is taken as evidence of the correlation between the Z value and profitability, and thus, the effectiveness of NDCG for approximating profitability. It is also clear from Fig. 4 that there is a significant oscillation of profit as the number of variables used to split the tree increases beyond seven. This pattern is explained by the fact that the value range considered for Z in this what-if analysis spans several settings, including many which are conceptually inappropriate for RF. In particular, limiting the number of independent variables used per tree is a key mechanism in RF for achieving diversity. Very large Z values increase the similarity of individual trees and render this mechanism meaningless.[2] Consequently, these results can be related to the proposed grid-search approach that considered candidate settings for Z up to a maximum of 14 (i.e. $2\sqrt{m}$). In view of Fig. 4, one may speculate that this range could be further reduced, since profit variations increase substantially when Z is greater than seven.

The results regarding parameter T, displayed in Fig. 5, confirm Breiman's (2001) suggestion of employing large forests with many trees. In retrospect, $T = 300$ would have been optimal, but the grid search failed to select this because the respective NDCG value was 71.15%, compared with an NDCG of 71.45% for the chosen setting of $T = 600$. However, it is important to note that in practice such knowledge would not be available at the time when parameter values must be selected. Consequently, the proximity of the two NDCG scores and the fact that the associated parameter settings turn out to generate remarkable profits, can be taken as further confirmation of the appropriateness of the grid-search and NDCG, respectively.

3.3. Forecasting accuracy

The effectiveness of the 2RF model with parameter settings $T = 600$ and $Z = 4$ (i.e. the 'optimal' settings from the grid-search) for predicting horses' winning probabilities is now compared with that of 2CL. In particular, a 2CL model is constructed and used to predict 500 holdout sample races run between the 1st January 2006 and 26th December 2006. The effectiveness of the 2CL model for forecasting the outcomes of horse-races has been verified in several experiments (see, e.g., Bolton & Chapman, 1986; Chapman, 1994; Figlewski, 1979; Johnson et al., 2006). Consequently, it represents a highly challenging benchmark.

The comparison between the 2RF and 2CL models confirms the suitability of the former for predicting race outcomes. A Kelly-wagering strategy (without reinvestment) based on the predicted winning probabilities of the 2RF model yields a significantly higher return than the 2CL model (20.26% and 8.84%, respectively). If the reinvestment of winnings is permitted, the 2RF model results in an increase in wealth of 528.56% over the 500 holdout races, whereas wealth increases by 187.1% using the 2CL model. Detailed results of this betting simulation are plotted in Fig. 6.

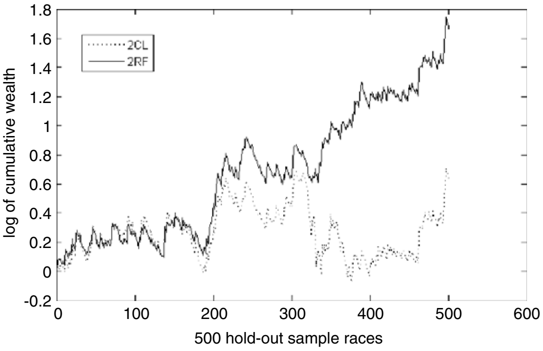

Fig. 6. Results of applying a Kelly-wagering strategy with the reinvestment of winnings to the 500 holdout sample races using the estimated winning probabilities of the 2RF and 2CL models.

In testing the significance of returns, it is common to consider the profit per pound bet. However, as was noted by Johnson et al. (2006), this measure suffers some limitations, so that it is preferable to examine the *factor* by which wealth has increased as successive bets are placed. In particular, let the cumulative wealth after bet j equal Y_{j+1}. Then, we can define $Y_{j+1} = (1 + y_j)Y_j$, where y_j is the factor by which bet j increases or decreases our cumulative wealth. Hence, the wealth after n bets is $Y_{n+1} = Y_0 \prod_{j=1}^{n}(1 + y_j)$. Taking logs, the multiplicative form of cumulative wealth is turned into an additive form, to obtain $\ln(Y_n) = \ln(Y_0) + \sum_{j=1}^{n} \ln(1 + y_j)$. Johnson et al. (2006) argue that it is sensible to compute the mean of $A_j = \ln(1 + y_j)$, because $A_j > 0$ implies that wealth increases as a result of the sequence of bets, whereas wealth decreases if $A_j \leq 0$. Using Kelly-betting, the mean of A_j is the asymptotic growth rate for wealth (per race). From the central limit theorem, the estimators $\mu = E(A_j)$ and $\sigma^2 = Var(A_j)$ are approximately normally distributed, which allows the null hypothesis $H_0 : \mu = 0$ to be tested against the alternative $\mu > 0$ by means of a one-tailed z-test. Consequently, using the previous results from betting

[2] Note that in the extreme case where $Z = 40$, all independent variables are considered in all trees. Consequently, the resulting classifier is no longer a 'random forest', but a simple bagging classifier.

over the 500 test set races, it may be concluded that A_j is significantly greater than zero for the 2RF model (p-value: 0.0085). However, the p-value of 0.1354 for the 2CL model indicates that there is a reasonable probability that its positive results occurred by chance.

The results presented above confirm that the 2RF model yields a significant profit. Given that all independent variables in the model are based on information which is publicly available, it may be concluded that the Hong Kong horserace betting market exhibited informational inefficiency during the period considered in this study (2005–2006). In particular, the fact that the 2RF model succeeds in outperforming the market indicates that it extracts information from the fundamental variables that has not been fully discounted in market prices by the betting public. Furthermore, the higher profitability of the 2RF model (relative to the 2CL model) can be taken as evidence of the former distilling additional information from the same variables. This result is confirmed by the fact that the (adjusted) R^2-statistic of the 2RF model exceeds that of the 2CL model by 1.6% (0.1296 c.f. 0.1276).

Although profitability is arguably the most important indicator of a forecasting model's success in economic environments (see, e.g., Granger & Pesaran, 2000; Leitch & Tanner, 1991), a detailed analysis of the 2RF model's forecasts is desirable to shed light upon the origin of its success. In particular, the profitability of a horserace forecasting model depends mainly upon the difference between its estimated winning probabilities and the public's estimates of runners' winning chances (as reflected in market odds).[3] The Kelly wagering strategy exploits differences between these two estimates to maximize the long-run rate of growth of wealth (see, e.g., Johnson et al., 2006). For example, in the betting simulation, wagers are only placed on runners where the 2RF model (or the 2CL model) predicts a higher chance of winning than that implied in the horse's odds. Furthermore, the difference between the estimated and odds-implied winning probabilities determines the size of the bets. Consequently, the profitability of the 2RF model and its superiority to the 2CL model may be understood by examining the respective estimates of the winning probabilities. The results of such a comparative analysis are presented in Table 3.

It is clear from Table 3 that the number of occasions where a model's estimate of a runner's chances of winning is higher than that implied by the horse's odds using the 2RF model is similar to the number using the 2CL model, resulting in similar numbers of bets (2RF: 4637; 2CL: 4500). Furthermore, the number of successful bets over the total number of bets is the same for the two models (5.1% in both cases), and they also succeed in correctly identifying the winner (i.e., estimating the highest winning probability across all runners in a race for the horse that eventually wins) on almost the same number of occasions. However, the appealing performance of the 2RF model in terms of profitability, and its advantage over 2CL in particular, can be understood when examining the differences between the two models' probability forecasts. In particular, there appears to be little difference between the proportion of occasions when one model's predicted probability for winning horses exceeds that of the other model (i.e. RF: 47%; CL: 53%). A two-sided sign test confirms that there is no significant difference between the probabilities of the 2RF and 2CL models' predicted probabilities for winning horses exceeding those of the other model ($p > 0.195$). However, a substantial difference between 2RF and 2CL can be observed for losing horses, with 2CL predicting a higher probability of winning on 59.3% of occasions. A two-sided sign test confirms that 2CL produces a significantly larger number of poorer predictions than 2RF for losing horses (i.e., it predicts a higher winning probability: $p < 0.0000$). Since the amount bet on a horse depends critically upon the difference between a model's estimate of the horse's chances and its odds-implied probability, the tendency for 2CL to forecast higher winning probabilities for losers more often than 2RF results in more money being lost on unsuccessful bets.

To augment the results of Table 3, the magnitudes of the differences between the 2RF and 2CL models' predicted winning probabilities are examined in Fig. 7. The differences between the winning probability forecasts of 2RF and 2CL are computed for all runners, and Fig. 7 depicts the means of this difference at different threshold values. That is, the first (i.e. leftmost) mean is computed across all runners, whereas

[3] In the following, the term *odds-implied* probability is used to refer to the public's estimate of a runner's chances of winning.

Table 3
Comparative analysis of the 2RF and 2CL models' forecasts over the 500 races holdout sample test set (6286 runners).

	2RF model	2CL model
Analysis of estimated versus odds-implied winning probabilities		
Runners where the estimated winning probability exceeds the odds-implied winning probability (and thus a bet is placed)	4637	4500
No. of true winner identifications[a]	125	126
No. of bets on winning horses	234	230
Rate of successful bets	5.1%	5.1%
Analysis of estimated winning probabilities of 2RF versus 2CL		
No. of runners where one model estimates a higher winning probability than the other model		
Across winners	235/47%	265/53%
Across non-winners	2356/40.7%	3430/59.3%

[a] The term 'true winner identification' refers to the case where a model estimates the highest winning probability for the horses that eventually finishes first.

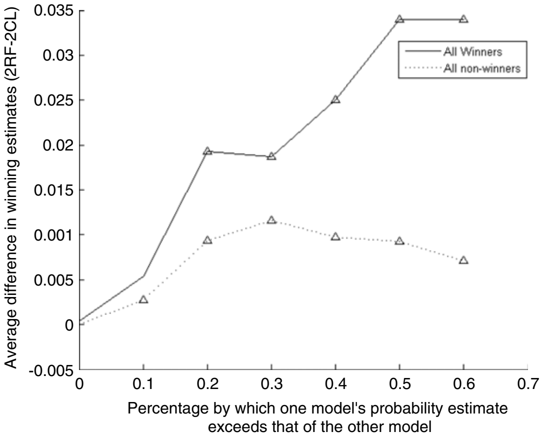

Fig. 7. Average difference in the estimated winning probabilities of 2RF and 2CL. Triangles indicate that an average difference is significantly different from zero at the 1% level based on a two-tailed t-test.

subsequent mean computations are restricted to horses where the (absolute) difference between the 2RF and 2CL estimates exceeds a certain percentage threshold. At each threshold value, a two-tailed t-test is conducted to explore whether the mean difference between the 2RF and 2CL estimated probabilities differs significantly from zero. A positive mean indicates that 2RF predicts higher winning probabilities (on average). We have already shown that 2RF does not forecast higher winning probabilities *more often* than 2CL (see Table 3), but we find that for all threshold values, for both winners and losers, the mean differences between the 2RF and 2CL estimated probabilities are positive (and mostly significantly so).

Importantly, as the absolute difference between the 2RF and 2CL predictions increases (i.e. at higher threshold values), the mean (positive) difference between the 2RF and 2CL predictions increases at a much faster rate for winners than for losers. Consequently, although 2RF predicts higher winning probabilities for both winning and losing horses than 2CL, the 2RF predictions for winners exceed those of 2CL by a considerably higher margin than those for losers. Such differences, in a Kelly betting strategy, translate into significantly larger sums being invested in these winners.

3.4. Variable importance

The greater profitability of the 2RF forecasting model implies that it produces winning probabilities which are more accurate than those of the 2CL model. This indicates that the 2RF model is able to capture more useful information contained in the fundamental variables concerning the race outcomes. Consequently, it is interesting to analyse the way in which independent variables are processed by these two models, and in which cases disagreements concerning the relevance of a variable or a set of variables can be identified.

The standard approach to appraising the importance of variables in linear models like 2CL is to examine their respective coefficients (i.e. $\boldsymbol{\beta}$ in Eq. (2)), which have been estimated from the training

Table 4
Ranking of variable categories according to the normalized mean importance score.

	RF	CL	DR[a]
Variables related to the horse's age	1	9	8
Post position bias variable	2	5	3
Conditions of today's race	3	1	2
Variables related to a horse's most recent performance on and off the track	4	8	4
Variables related to a jockey's past performance	5	2	3
Variables related to a horse's past performance	6	4	2
Variables related to days since last ran	7	3	4
Variables related to the weight carried by the horse in both current and past races	8	7	1
Horse's specific preferences (e.g., blinkers, track, distance, surface, jockey)	9	6	3

[a] Difference in rank.

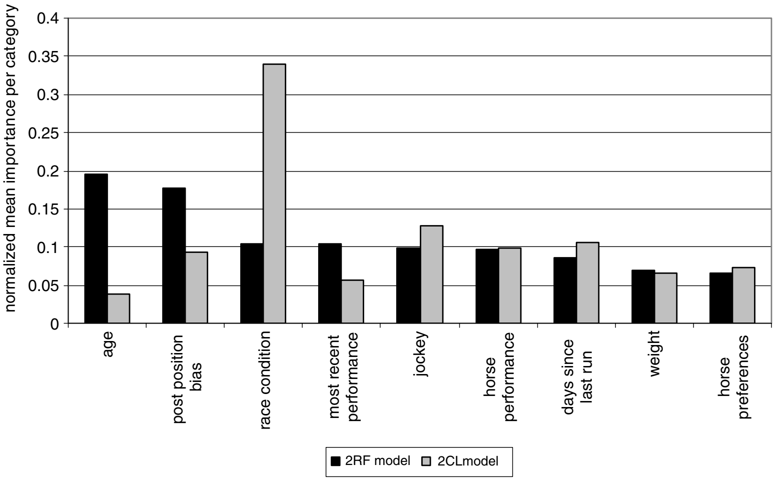

Fig. 8. A comparison of variable importance measures as produced by the 2RF and 2CL prediction models using all of the variables described in Table 1 except track probabilities.

sample. For the RF classifier, the cumulative decrease in *GI* over all nodes in the forest that split on a particular variable quantifies this variable's importance. These respective measures for the 2CL and 2RF models are extracted from the final forecasting models and are summarized in Fig. 8. To maintain comprehensibility when comparing a large number of variables, importance measures have been aggregated using the mean value per variable category (see Table 1). Furthermore, a normalization has been conducted to account for different value ranges across variables and models. A variable category with a higher importance score can be regarded as 'more' influential than others, and the categories are ordered according to their normalized mean importance scores (as produced by the 2RF model). For example, variables related to a runner's age (AGE) and a horse's post position (PPB) turn out to be particularly important when estimating winning probabilities by means of 2RF (cf. 2CL), whereas the conditions of today's race (CDR) are particularly important in the 2CL model (cf. 2RF). There is less difference between the two models in terms of the importance of the variables in other categories (see Table 4, which summarizes the observed ranking deviations).

One has to be careful when drawing conclusions regarding variable importance across models, because the respective importance measures of 2CL and 2RF capture significantly different aspects (absolute coefficient value compared to cumulated *GI*-decrease). It is therefore inappropriate to conclude that, for example, the relevance of PPB is approximately 'twice' as high in the 2RF model as in the 2CL model. However, comparisons of the relative orderings of the variables are feasible, and one may conclude that, for example, variables in the AGE group are most relevant in the 2RF model, whereas the 2CL model assigns the least importance to this category. This in turn indicates that the models disagree on the importance of the respective variable/category (i.e. they rely more or less heavily on this type of information).

In order to confirm the view that the higher profitability of the 2RF model can be attributed to it distilling more useful information from the fundamental variables, it is desirable to determine the significance of the observed differences in variable importance between the 2RF and 2CL-based models. This may be achieved by measuring the degree of correlation between the rankings in Table 4. In particular, Kendall's τ represents a relevant ranking correlation coefficient, and indicates that no significant correspondence exists between the orderings of variable importance ($\tau = -0.56$; p-value $= 0.92$).

In view of the fact that the use of the available information within the two models differs significantly, and given that the estimated winning probabilities of the 2RF model are demonstrably more accurate than those of the 2CL benchmark, strong evidence for the effectiveness of the RF classifier and its way of processing fundamental variables is provided.

An important question remains; namely, whether the observed results allow any further conclusions to be drawn, which could help to explain the higher profitability of the RF model. On the one hand, Table 4 suggests that the 2CL model failed to fully distil the information contained in the AGE variables. One could speculate that this might be due to the fact that the relationship between the AGE variables and race outcomes is nonlinear, and therefore impossible for a linear model like CL to detect. However, formally confirming such an assertion is difficult, if not impossible,[4] and we are confronted with the well-known statistical dilemma of correlation (i.e. higher profit and different use of information) versus causality. Consequently, further research is needed to enable us to fully understand the origin of the 2RF model's effectiveness in forecasting race outcomes, and winning probabilities in particular.

4. Conclusion

A forecasting model which is based on the recently developed RF classifier has been proposed for predicting the outcomes of competitive events, and horseracing results in particular. While RF has been found to be a highly accurate approach for solving standard classification problems in previous research, this forms the first attempt to use this method in competitive environments, and in horserace forecasting in particular. Several enhancements have been proposed to increase RF's awareness of within-race competition. In particular, an accuracy indicator that captures the subtle differences between runners who finish in the first three (and consequently win prize-money) and other horses has been employed to guide the search for predictive settings for the parameters of the RF classifier. Furthermore, a two-stage forecasting methodology has been adopted for post-processing the predictions of RF by means of a CL regression.

A number of experiments have demonstrated the steps to be taken when applying RF in competitive scenarios, and have confirmed the effectiveness of both the proposed model and its individual components in this context. In comparison to traditional statistical techniques like CL, the need to empirically determine two parameters (i.e., the number of trees in the forest and the number of independent variables selected randomly for splitting in each individual tree) may be considered an additional overhead. However, it has been shown that this task can be handled effectively and efficiently by a grid-search over some candidate values using oob examples. Furthermore, the significantly higher forecasting accuracy of the 2RF model over the 2CL competitor has demonstrated that this moderate additional effort is a wise investment. With respect to the comprehensibility of the forecasting model, it has been shown that variable importance measures similar to those in linear statistical models can easily be extracted from the 2RF model to gain an understanding of the type of information upon which the predictions are based, and which individual variables are most relevant. Finally, based on the computational experience gained during the evaluation of 2RF, it can be concluded that this type of model can

[4] In principal, one could employ an artificial dataset to scrutinise this issue, but it is questionable whether the respective findings could be generalized to real world data.

process a large dataset of 1000 races (12,902 runners with 41 variables each) with ease. The 2RF model is therefore suitable for applications in the social sciences where large databases are commonly available.

Replicating the results of this study with different datasets seems one fruitful direction for future research. For example, the accuracy of the estimated winning probabilities from the 2RF model beings previous conclusions regarding the informational efficiency of horserace betting markets into question. Consequently, a further analysis using different datasets would be beneficial to scrutinize whether betting – or trading – on the 2RF model's forecasts in other financial applications would also generate significant returns. This could shed light on the true degree of informational efficiency in these markets.

With respect to the forecasting methodology, this is, to the best of our knowledge, the first experiment to consider the ensemble paradigm for competitive event forecasting. Having confirmed that it is generally applicable in this context, one could envision a deeper integration between RF and CL. For example, the original RF approach involves aggregating the predictions of individual CART models by means of majority voting. CL could replace this relatively naive voting mechanism and model the relationship between the predictions of individual trees and the dependant variable within a competitive context. A related technique, known as 'stacking' in the machine learning community, could also be used to introduce early within-race competition. Modifications along this line may help to further improve the alignment between the forecasting methodology and the ultimate modeling objective (i.e. profit), and would presumably produce even better predictions.

Appendix. Formal description of the proposed forecasting model

This appendix provides a formal derivation of the proposed forecasting model. Using the notation introduced in Section 2.1, let $S = \{\boldsymbol{x}_i^j, W_i^j\}_{i=1,j=1}^{n_j,R}$ represent a horseracing dataset containing R past races, $\boldsymbol{x}_i^j \in \Re^m$ a single horse i in race j that is characterized by m independent variables, and n_j the number of horses in race j. Information on whether or not horse i won race j is given by $W_i^j \in \{0, 1\}$. Furthermore, q_i^j represents the market odds of runner i in race j, which are known to be important for predicting race outcomes and play a special role within the two-stage modeling paradigm (see, e.g., Bruce & Johnson, 2000; Sung & Johnson, 2007). Therefore, the symbol $\bar{\boldsymbol{x}}_i^j \in \Re^{m-1}$ is used to refer to a horse's fundamental characteristics, i.e. after excluding market odds from the set of independent variables.

To derive the two-stage model, the dataset S is first partitioned into two disjoint subsets, S_1 and S_2, which contain approximately the same number of races, R_1 and R_2 respectively. The dataset S_1 is used to build a RF forecasting model in stage one. Since the race context is not considered at this stage and the use of market odds is postponed to stage two, the respective data can be represented as $S_1 = \{\bar{\boldsymbol{x}}_i, W_i\}_{i=1}^{N_1}$, where N_1 gives the overall number of runners in S_1 across all races. Using S_1, a family of CART decision trees (f_{CART}) of size T is grown, employing the procedures outlined in Section 2.2.1. Afterwards, this family (i.e. the RF) facilitates an assessment of the runners in S_2. Remembering that $f_{CART}(\bar{x}_i) \to \{0, 1\}$, the RF forecast is given as:

$$\text{Stage one: } f_{RF}(\bar{\boldsymbol{x}}_i) = \sum_{t=1}^{T} f_{CART}(\bar{\boldsymbol{x}}_i) \qquad i = 1, \ldots, |S_2|, \tag{A.1}$$

where $|S_2|$ represents the cardinality of S_2 (i.e., the total number of runners in this set). Hence, the prediction of RF for an individual runner is obtained by counting the number of tree models which classify it as a winner. Consequently, the RF output may be interpreted as an assessment of a runner's skill (i.e. the ability index), which is based only on its individual fundamental characteristics.

In stage two, the predicted ability index is combined with the market odds using CL regression to predict horses' winning probabilities, $\boldsymbol{p}^j = \left(p_1^j, p_2^j, \ldots, p_{n_j}^j\right)$, for race j:

$$\text{Stage two: } p_i^j = \frac{\exp\left(\boldsymbol{\beta}_1 f_{RF}\left(\bar{x}_i^j\right) + \boldsymbol{\beta}_2 q_i^j\right)}{\sum_{i=1}^{n_j} \exp\left(\boldsymbol{\beta}_1 f_{RF}\left(\bar{x}_i^j\right) + \boldsymbol{\beta}_2 q_i^j\right)}. \tag{A.2}$$

The parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are determined by means of maximum likelihood estimation over the second stage dataset S_2. Having determined the RF model, as well as the parameters of the CL model, Eqs. (A.1) and (A.2) can then be employed to forecast the winning probabilities of runners in future races.

References

Benter, W. (1994). Computer based horserace handicapping and wagering systems: A report. In D. B. Hausch, V. S. Y. Lo, & W. T. Ziemba (Eds.), *Efficiency of racetrack betting markets* (pp. 183–198). New York: Academic Press.

Bolton, R. N., & Chapman, R. G. (1986). Searching for positive returns at the track: A multinomial logit model for handicapping horseraces. *Management Science*, *32*(8), 1040–1060.

Breiman, L. (1961). Optimal gambling systems for favourable games. In J. Neyman (Ed.), *Proceedings of the 4th Berkeley symposium on mathematical statistics and probability* (pp. 63–68). Berkeley: University California Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont: Wadsworth.

Bruce, A. C., & Johnson, J. E. V. (2000). Investigating the roots of the favourite-longshot bias: An analysis of supply and demand side agents in parallel betting markets. *The Journal of Behavioral Decision Making*, *13*(4), 413–430.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, *32*(2), 277–288.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In W. W. Cohen, & A. Moore (Eds.), *Proceedings of the 23rd international conference on machine learning* (pp. 161–168). New York: ACM.

Chapman, R. G. (1994). Still searching for positive returns at the track: Empirical results from 2000 Hong Kong races. In D. B. Hausch, V. S. Y. Lo, & W. T. Ziemba (Eds.), *Efficiency of racetrack betting markets* (pp. 173–181). New York: Academic Press.

Chapman, R. G., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, *19*(3), 288–301.

Chen, C., Liaw, A., & Breiman, L. (2004). *Using random forest to learn imbalanced data*. Technical Report, 666, Statistics Department, University of California at Berkeley.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Edelman, D. (2007). Adapting support vector machine methods for horserace odds prediction. *Annals of Operations Research*, *151*(1), 325–336.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*(2), 383–417.

Figlewski, S. (1979). Subjective information and market efficiency in a betting market. *Journal of Political Economy*, *87*(1), 75–89.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, *55*(1), 119–139.

Granger, C. W. J., & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*(7), 537–560.

Gu, M. G., Huang, C., & Benter, W. (2003). *Multinomial probit models for competitive horse racing*. Working paper. Chinese University of Hong Kong.

Guo, L., Ma, Y., Cukic, B., & Singh, H. (2004). Robust prediction of fault-proneness by random forests. In *Proceedings of the 15th international symposium on software reliability engineering* (pp. 417–428). Los Alamitos: IEEE Computer Society.

Hamza, M., & Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, *75*(8), 629–643.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844.

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In N. J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 41–48). New York: ACM Press.

Johnson, J. E. V., Jones, O., & Tang, L. (2006). Exploring decision makers' use of price information in a speculative market. *Management Science*, *52*(6), 897–908.

Kelly, J. L. (1956). A new interpretation of information rate. *The Bell System Technical Journal*, *35*, 917–926.

Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, *29*(2), 472–484.

Leitch, G., & Tanner, J. E. (1991). Economic forecast evaluation: Profit versus the conventional error measures. *American Economic Review*, *81*(3), 580–590.

Leon, F., Zaharia, M. H., & Gâlea, D. (2004). Performance analysis of categorization algorithms. In *Proceedings of the 8th international symposium on automatic control and computer science*.

Lessmann, S., Mues, C., Baesens, B., & Pietsch, S. (2008). Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Transactions on Software Engineering*, *34*(4), 485–496.

Lessmann, S., Sung, M., & Johnson, J. E. V. (2007). Adapting least-square support vector regression models to forecast the outcome of horseraces. *Journal of Prediction Markets*, *1*(3), 169–187.

Lessmann, S., Sung, M., & Johnson, J. E. V. (2009). Identifying winners of competitive events: A SVM-based classification model for horserace prediction. *European Journal of Operational Research*, *196*, 569–577.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

Sung, M., & Johnson, J. E. V. (2007). Comparing the effectiveness of one- and two-step conditional logit models for predicting outcomes in a speculative market. *Journal of Prediction Markets*, *1*(1), 43–59.

Sung, M., & Johnson, J. E. V. (2008). Semi-strong form information efficiency in horserace betting markets. In D. Hausch, & W. Ziemba (Eds.), *Handbook of sports and lottery markets*. North Holland: Elsevier.

Sung, M., Johnson, J. E. V., & Bruce, A. C. (2005). Searching for semi-strong form inefficiency in the UK racetrack betting market. In L. Vaughan Williams (Ed.), *Information efficiency in financial and betting markets* (pp. 179–192). Cambridge: Cambridge University Press.

Van Gestel, T., Suykens, J. A. K., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., De Moor, B., & Vandewalle, J. (2004). Benchmarking least squares support vector machine classifiers. *Machine Learning*, *54*(1), 5–32.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Zar, J. H. (2007). *Biostatistical analysis*. Upper Saddle River: Prentice Hall.

Stefan Lessmann received an M.Sc. and a Ph.D. in Business Administration from the University of Hamburg (Germany) in 2001 and 2007, respectively. He is currently employed as a lecturer in information systems at the University of Hamburg. His research interests include the development and application of forecasting methods in various domains, ranging from customer relationship management and empirical software engineering to financial markets.

Ming-Chien Sung is Lecturer in Management Science at the University of Southampton. She has presented papers at numerous international conferences and has published a range of papers which explore new approaches for modelling uncertain outcomes. She is currently examining the use of both parametric and non-parametric methods for predicting the outcomes of uncertain events in speculative markets with respect to a range of publicly available information.

Johnnie E.V. Johnson is Professor of Decision and Risk Analysis and Director of the Centre for Risk Research in the School of Management at the University of Southampton. He has published widely in the areas of risk perception, risk management, and decision making under uncertainty. Johnnie has a particular interest in developing new approaches for forecasting outcomes in speculative markets, such as prediction and betting markets.