The Journal of Prediction Markets (2007) **1** 3, 169–187

ADAPTING LEAST-SQUARE SUPPORT VECTOR REGRESSION MODELS TO FORECAST THE OUTCOME OF HORSERACES

Stefan Lessmann†, Ming-Chien Sung‡¶ and Johnnie E.V. Johnson‡§*

†*Institute of Information Systems, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany*

‡*Centre for Risk Research, School of Management, University of Southampton, Highfield, Southampton, SO17 1BJ, UK*

This paper introduces an improved approach for forecasting the outcome of horseraces. Building upon previous literature, a state-of-the-art modelling paradigm is developed which integrates least-square support vector regression and conditional logit procedures to predict horses' winning probabilities. In order to adapt the least-square support vector regression model to this task, some free parameters have to be determined within a model selection step. Traditionally, this is accomplished by assessing candidate settings in terms of mean-squared error between estimated and actual finishing positions. This paper proposes an augmented approach to organise model selection for horserace forecasting using the concept of ranking borrowed from internet search engine evaluation. In particular, it is shown that the performance of forecasting models can be improved significantly if parameter settings are chosen on the basis of their normalised discounted cumulative gain (i.e. their ability to accurately rank the first few finishers of a race), rather than according to general purpose performance indicators which weight the ability to predict the rank order finish position of all horses equally.

Keywords: Forecasting, Horseracing, Support Vector Machines

1. INTRODUCTION

It is widely accepted that horserace betting markets are well suited for testing market efficiency since they share many features in common with wider financial markets, including a large number of participants and a considerable amount of information which is available to assess a horse's (asset's) market values (Hausch and Ziemba, 1985; Johnson *et al.*, 2006; Law and Peel, 2002; Levitt, 2004; Sauer, 1998; Schnytzer and Shilony, 1995; Sung and Johnson, 2007; Vaughan Williams, 1999). In addition, horserace betting markets offer an important advantage over wider financial markets: They generate an unequivocal outcome (a winner) and an associated rate of return within a finite time frame (Law and Peel, 2002), and, hence, provide an objective benchmark against which to measure the quality of an investment decision (i.e. a bet). "As a result, wagering markets can provide a clear view of pricing issues which are more complicated elsewhere" (Sauer, 1998 p. 2021) and the value of studying bettors' decisions is reinforced by the fact that these markets are, in themselves, important. For example, the turnover of the UK horserace betting market in 2006 was £15,500 million[1].

*Corresponding author. Tel.: + 49-40-42838-4706. Fax: +49-40-42838-5535.
Email: lessmann@econuni-hamburg.de
¶Email: ms9@soton.ac.uk
§Email: jej@soton.ac.uk

Predictive modelling may help to shed light on the rationality of traders' collective decisions in such markets. In particular, a forecasting model can be derived from past race data and employed to estimate a runner's likelihood of winning a future race. If the model predictions can be used to secure a profit over a number of future races it may be concluded that market participants do not fully discount information contained in the model (e.g., Benter, 1994; Bolton and Chapman, 1986; Johnson *et al.*, 2006; Sung *et al.*, 2005). Clearly, models which are able to more fully capture information (contained in input variables) regarding the probabilities of horses' winning are those which are more likely to identify the true degree of inefficiency in a market.

Recently, Sung and Johnson (2007) have shown that, in a horserace context, a two-stage modelling procedure, as advocated by Benter (1994), outperforms a one-stage modelling procedure. The intuition behind such a nested model is that market odds (prices) of horses are an extremely powerful predictor and may thus mask the influence of other potentially informative variables. Therefore, a first stage model is developed to process fundamental variables to produce a score which reflects a runner's ability based on this fundamental information. Subsequently, this score is combined with market odds in a second stage to generate the final forecast.

The conditional logit (CL) model (Maddala, 1983) used to be the "gold-standard" in horserace prediction because of its ability to account for independent variables measuring a runner's potential *and* within-race competition (Bolton and Chapman, 1986). However, recent results suggest that the predictive power of traditional CL-based two-stage models can be further enhanced if modern machine learning algorithms, namely support vector machines (SVM), are employed in the first stage to extract more information from fundamental variables (Edelman, 2006). In particular, Edelman (2006) suggests the use of a SVM regression (SVR) model to forecast runners' normalised finishing positions in stage one. It is argued that a regression model makes full use of information contained in rank ordered finishing data and is, therefore, superior to a classification model which simply extracts information distinguishing winners from losers (Edelman, 2006). On the other hand, forecasting models which use the CL procedure in both stages have been demonstrably successful (Sung *et al.*, 2005; Sung and Johnson, 2007). This may be explained by the fact that prize money is generally only offered for the first few finishers of a race, so that jockeys have little incentive to continue riding a horse to its full potential when it becomes clear that they are not going to secure a prize. In fact, there may be a motivation for jockeys to secure a *poorer* finish position on non-winning horses than they might be able to achieve. This will have the effect of reducing the public's perception of the ability of the horse, which will result in higher odds being available on the horse in subsequent races; offering the prospect of sizeable betting gains to the owners in future races. Such effects would impair the reliability of rank order finishing data and may reduce the accuracy of regression-based forecasting models.

SVR is a semi-parametric model which requires so-called hyperparameters to be defined prior to model development. This is accomplished by a validation procedure which selects the optimal hyperparameters by iteratively assessing the performance of candidate values on holdout samples based on the mean-squared-error (MSE) between estimated and true (normalised) finishing position. It is well known that this model selection procedure has a vital impact on the predictive performance of the final model. Consequently, the potential problem of using the full range of rank ordered finishing data for developing horserace forecasting models is further increased when using SVR. However, the results presented in section 3.2 show that a novel approach towards model selection for predicting race outcomes, namely, normalised discounted cumulative gain (NDCG- inspired from the evaluation of internet search engines or recommender systems), enables a substantial improvement over an ordinary SVR-based forecasting model in terms of profitability.

The aim of this paper is to reduce the risk of selecting hyperparameter values based on potentially unreliable rank ordered finishing data in horseraces. To achieve this, the paper builds on the previous work of Edelman (2006) by proposing an augmented SVR-based two-stage model which introduces NDCG in the SVM model selection procedure. The paper is organised as follows: Section 2 explores the nature of SVM modelling, including the traditional and a more sophisticated means of selecting hyperparameter values (NDCG) for horserace forecasting. The two stage model employed in this paper is also introduced in this section. The experimental setup used to explore the advantage afforded by the NDCG approach to hyperparameter selection is outlined in section 3. The experimental results are also reported in this section. Some conclusions are drawn in section 4.

2. FORECASTING HORSERACE RESULTS WITH TWO-STAGE MODELS

2.1. Support vector machines for regression

The support vector machine was introduced in 1992 (Boser *et al.*, 1992) as a learning algorithm that infers functional relationships from data following the structural risk minimisation induction principle (Vapnik, 1995). The original algorithm considers the task of classification (i.e. predicting discrete target variables (see Cristianini and Shawe-Taylor (2000) for a detailed overview). Subsequently, the method has been extended to allow forecasting real-valued target variables. This support vector regression (SVR) aims at inferring a functional relationship $y(\mathrm{x}) : \Re^N \rightarrow Y$ from an i.i.d. training sample $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{M}$ of M observations, whereby $\boldsymbol{x}_i \in X \subseteq R^N$ represents a vector of measurements, and $y_i \in Y \subseteq \Re$ denotes a continuous target variable. SVR assumes the linear model (1), with $\boldsymbol{w}$ and b representing the normal and intercept of the resulting hyperplane, respectively, and $\langle \cdot , \cdot \rangle$ is the

scalar product in X.

$$y(\boldsymbol{x}) = \langle \boldsymbol{w}, \varphi(\boldsymbol{x}) \rangle + b. \tag{1}$$

To allow for more complex, nonlinear relationships, the input data is transformed into a higher dimensional feature space via an a priori chosen mapping $\varphi(\cdot) : \Re^N \rightarrow \Re^{N_F}$. It is important to note that the dimension of the feature space, N_F, is defined in an implicit way and can, in fact, be of infinite dimensions.

According to statistical learning theory, the model parameters $\boldsymbol{w}$ and b should be determined such that the resulting function $y(\boldsymbol{x})$ exhibits high accuracy while at the same time being as *flat* as possible (Vapnik, 1995). In particular, flatness corresponds to having a model with low complexity (i.e. small $\boldsymbol{w}$ (Smola and Schölkopf, 2004) whereas inaccurate predictions are considered irrelevant as long as the deviation between the estimated and true values are less than ε, a user-defined constant. This idea motivates Vapnik's ε-insensitive loss-function (Vapnik, 1995), which is defined as follows:

$$|y - y(\boldsymbol{x})|_{\varepsilon} = \left\{ \begin{array}{ll} 0, & \text{if} |y - y(\boldsymbol{x})| \leq \varepsilon \\ |y - y(\boldsymbol{x})| - \varepsilon, & \text{otherwise} \end{array} \right\}. \tag{2}$$

Integrating the two goals of predictive accuracy and flatness leads to the following quadratic program:

$$\min_{\mathrm{w},b,\xi,\xi^*} \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w} \rangle + C\sum_{i=1}^{M}(\xi_i + \xi_i{}^*)$$

$$\begin{array}{lll} \text{s.t} & y_i - \langle \boldsymbol{w}, \varphi(\boldsymbol{x}_i) \rangle - b \leq \varepsilon + \xi_i & \forall i = 1, \ldots, M \\ & \langle \boldsymbol{w}, \varphi(\boldsymbol{x}_i) \rangle + b - y_i \leq \varepsilon + \xi_i{}^* & \forall i = 1, \ldots, M \\ & \xi_i, \xi_i{}^* \quad \geq 0 & \forall i = 1, \ldots, M. \end{array} \tag{3}$$

The slack variables $\xi_i, \xi_i{}^*$ account for the fact that it might not be possible to approximate all $\mathrm{x}_i \in S$ with precision ε. In other words, they represent the distance from points outside (above or below) an ε-tube around the regression function. The regularisation parameter C allows for a trade-off between flatness and accuracy, i.e. the amount up to which deviations larger than ε are tolerated (Smola and Schölkopf, 2004). The intuition behind program (3) is illustrated in Figure 1.

Program (3) is generally solved in its dual form, which can be obtained by substituting the conditions for optimality with respect to the primal variables $(\boldsymbol{w}, b, \xi, \xi^*)$ into the Lagrangian following from (3) (see, for example, Vapnik, 1995). This leads to the program (4), with $\alpha_i, \alpha_i{}^*$ denoting the Lagrangian

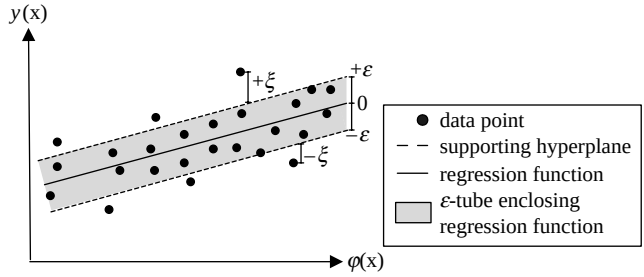

FIGURE 1. Linear SVR with ε-insensitive loss-function.

multipliers (Smola and Schölkopf, 2004):

$$
\begin{aligned}
\max_{\alpha,\alpha^*} = \quad & -\frac{1}{2}, \sum_{i,j=1}^{M}(\alpha_i - \alpha_i{}^*)\big(\alpha_j - \alpha_j{}^*\big)\big\langle \varphi(\mathrm{x}_i), \varphi\big(\mathrm{x}_j\big)\big\rangle \\
& -\varepsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i{}^*) + \sum_{i=1}^{M} y_i(\alpha_i - \alpha_i{}^*) \qquad (4) \\
\text{s.t.} \quad & \sum_{i=1}^{M}(\alpha_i - \alpha_i{}^*) = 0 \quad \forall i = 1, \ldots, M \\
& 0 \le \alpha_i, \alpha_i{}^* \le C \quad \forall i = 1, \ldots, M
\end{aligned}
$$

Program (4) includes the input data only in the form of scalar products. This feature enables an implicit transformation of the data by introducing a kernel function. The kernel K calculates the inner product $\big\langle \varphi(\boldsymbol{x}_i), \varphi\big(\boldsymbol{x}_j\big)\big\rangle$ directly in the input space and thereby avoids the need to compute $\varphi(\cdot)$ explicitly (Cristianini and Shawe-Taylor, 2000):

$$K\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) = \big\langle \varphi(\boldsymbol{x}_i), \varphi\big(\boldsymbol{x}_j\big)\big\rangle. \qquad (5)$$

Common choices for kernel functions include the linear kernel, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j$, the polynomial kernel with degree $d, K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^d$, as well as the Gaussian radial basis function (RBF) with spread parameter $\sigma, K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2/\sigma^2)$. The latter model has been employed by Edelman (2006) to predict horses' normalised finishing positions over a set of Australian races and is given as:

$$y(\boldsymbol{x}) = \sum_{i=1}^{M}\big(\alpha_i - \alpha_i^*\big) = K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \qquad (6)$$

Putting to one side the selection of a kernel function together with its respective parameters, SVR requires the user to specify two hyperparameters, $\big(C, \varepsilon\big)$ which enable the model to be adapted to different tasks. However, the hyperparameter ε may be eliminated when considering a least-square loss

function instead of the original ε-insensitive loss-function (Suykens *et al.*, 2002; Suykens and Vandewalle, 1999). The construction of this least-square Support Vector Machine (LS-SVM) requires that a set of linear equations is solved, which may be a much simpler task rather than solving the quadratic program (4). Furthermore, empirical results suggest that LS-SVMs are at least as accurate as conventional SVMs (see, for example, Baesens *et al.*, 2003; Van Gestel *et al.*, 2004). Most importantly, the reduced number of hyperparameters for the regression setting helps to simplify the task of model selection substantially; this is discussed more fully below. Therefore, the LS-SVM formulation for function estimation (LS-SVR) is adopted here.

LS-SVR also considers the functional model (7) and is based on the following optimisation problem (Suykens and Vandewalle, 1999):

$$\begin{aligned} &\min_{\boldsymbol{w},b,e} \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{M} e_i^2 \\ &\text{s.t.} \quad y_i = \langle \boldsymbol{w}, \varphi(\boldsymbol{x}_i)\rangle + b + e_i \quad \forall i = 1, \ldots, M. \end{aligned} \tag{7}$$

The error term is now denoted by *e* to emphasise that it represents the true deviation between actual values and forecasts in the LS-SVR formulation, rather than a slack variable which is needed to ensure feasibility (as in the SVR case). The optimal solution of the dual of (7) is given by a system of linear equations (e.g., Suykens, 2001) that can be solved efficiently, even for large scale problems (Suykens *et al.*, 1999). The resulting function estimation model then becomes:

$$y(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) + b. \tag{8}$$

2.2. Support vector machine model selection for horserace forecasting

The task of model selection aims at finding suitable settings for the hyperparameters of a predictive model. LS-SVR involves choosing a kernel function and kernel parameters, and determining the regularisation parameter *C*. The RBF kernel is predominantly used in SVM applications and has been shown to possess some desirable properties. For example, Keerthi and Lin (2003) show that it includes the linear kernel as a special case. Opting for this kernel leaves two free parameters, namely C and σ, which have to be determined prior to model development.

A number of strategies have been proposed to organise SVM model selection (e.g., Chapelle *et al.*, 2002; Chung *et al.*, 2003; Duan *et al.*, 2003; Joachims, 2000; Keerthi *et al.*, 2007) with grid-search being the most popular one (Hsu *et al.*, 2003; Keerthi and Lin, 2003). Grid-search involves predefining a set of candidate values for each parameter and empirically evaluating all possible combinations. The values which are shown to lead to

the best predictions are then used when developing the final model. Some modifications have been suggested to ease the computational burden of assessing a large number of parameter combinations. The approach adopted here resembles Van Gestel *et al*. (2004) and starts with a coarse grid that is subsequently refined in promising regions of the search space. In doing so, a useful balance is achieved between examining a wide range of parameter values and an intensive search of an area which contains the most promising candidates (Hsu *et al.*, 2003; Van Gestel *et al.*, 2004).

It is well known that the predictive performance of a SVM model depends heavily upon a suitable choice of hyperparameter values. However, the particular parameter combination which should be chosen during model selection depends, to some extent, on the concrete measure of predictive performance (some empirical evidence is provided, e.g., by Coussement and Van den Poel (2008)). This dependency is of pivotal importance when SVM-type models are employed for horserace forecasting. The accuracy of prediction for a regression-based modelling approach is most often assessed in terms of mean-square-error (MSE) between the estimated and true finishing position of horses over a number of out-of-sample races. Consequently, information provided within rank ordered finishing positions is utilised not only for model building but also for guiding the search for suitable hyperparameter values. These values, in turn, have a significant impact on the performance of the final forecasting model.

2.3. Model selection with normalised discounted cumulative gain

In view of the importance of finishing position information for SVR-based forecasting models, unreliability of finishing positions, which often occurs among the runners at the rear of the field, is unacceptable (see Sung and Johnson, 2007). Therefore, the approach suggested here replaces the MSE criterion with a more sophisticated performance indicator that better reflects the nature of horseracing and, thereby, facilitates more robust model selection. In particular, a performance measurement from information retrieval is adopted, which is commonly used to evaluate search engines or recommender systems.

It is generally accepted that users assess the accuracy of an internet search engine predominantly in terms of the number of relevant documents presented at the first results page. In other words, it is important for the topmost results returned as a response to a query (i.e. the ones obtainable without additional browsing) to contain a large number of documents the user judges as useful. Subsequent results presented at the second, third, and following pages of a summary dialog are rarely considered. Consequently, the primary objective for search engines is to maximise accuracy within the highest ranked sub-sample of retrieved documents (Cao *et al.*, 2006).

A similar rationale can be applied to horserace prediction. For instance, a forecasting model that predicts the finishing order of a race and confounds only the two horses finishing at the rear of the field would be regarded as

superior to another model which predicts the actual runner-up as winner, the true winner as second and all remaining runners correctly. However, these models are indistinguishable using the MSE criterion: both make one error by predicting the last two horses and the first two horses, respectively, in the wrong order. This is problematic since a horserace forecasting model should put special emphasis on accurately predicting the first finishers, whereas errors among the later finishing positions should be assigned minor importance. This view is also supported when considering which model produces predictions which could best be used to make profits from betting. Consequently, a model selection criterion for regression-based horserace forecasting should be appraised in terms of its ability to rank horses in accordance with their respective finish position, while putting special emphasis on the ability to forecast winners and placed horses. Hence, the situation in horserace modelling mimics the aforementioned search engine scenario.

A number of metrics that implement this evaluation strategy have been developed within information retrieval (e.g., for crediting search-engines or recommender systems: Breese *et al.*, 1998; Järvelin and Kekäläinen, 2000). The NDCG criterion (Järvelin and Kekäläinen, 2000) appears to be the most promising indicator for LS-SVR model selection and is adopted in this study.

To introduce NDCG, consider the example given in Table 1, which represents a race j with $m_j = 4$ runners ordered according to their actual finish position. Let $\boldsymbol{FP}_j$ and $\boldsymbol{NFP}_j$ be vectors containing the ordered finish positions and normalised finished positions of race j. Normalisation is undertaken in a manner to scale all finishing positions to the interval [0,1] with one representing the winner and zero the horse placed last. Let $\boldsymbol{FP}(\boldsymbol{j})$ and $\boldsymbol{NFP}(\boldsymbol{j})$ represent the model-based estimates of these values. To ensure that greater emphasis is given to accurate prediction of winners and near winners, a weight vector $\boldsymbol{v}$ is introduced. Its m_j components, v_i, represent the relative importance of predicting the finishing position of runner i correctly. As discussed above, it is reasonable to assume that the relevance decreases as finish position increases (i.e. it is most important to predict the winner correctly). Thus, the weights in vector $\boldsymbol{v}$ should decrease monotonically and can be defined as

TABLE 1
PERFORMANCE MEASUREMENT WITH NDCG

Runner	$\boldsymbol{FP}_j$	$\boldsymbol{NFP}_j$	$\boldsymbol{FP}(j)/\boldsymbol{NFP}(j)$
A	1	1,00	3/0,33
B	2	0,66	2/0,66
C	3	0,33	1/1,00
D	4	0,00	4/0,00

follows (Le and Smola, 2007):

$$v_i = 1/\left(\log_2(i) + 1\right) \quad \forall i = 1, \ldots, m_j. \tag{9}$$

Arranging the model-generated forecasts according to the true finishing order of the horses (column 4 in Table 1), the discounted cumulated gain (DCG) is given as the scalar product of $\boldsymbol{v}$ and $\boldsymbol{NFP}(\boldsymbol{j})$. This measure can be normalised by dividing by the DCG of an optimal prediction (i.e. the scalar product of $\boldsymbol{v}$ and $\boldsymbol{NFP}_j$ in Table 1). Consequently, NDCG is scaled between [0, 1] with one indicating a perfect prediction.

$$\begin{aligned} \mathrm{DCG}_{\mathrm{Model}} &= \langle \boldsymbol{v}, \boldsymbol{NFP}(\boldsymbol{j}) \rangle \\ \mathrm{DCG}_{\mathrm{Optimal}} &= \langle \boldsymbol{v}, \boldsymbol{NFP}_{\boldsymbol{j}} \rangle \\ \mathrm{NDCG} &= \mathrm{DCG}_{\mathrm{Model}} / \mathrm{DCG}_{\mathrm{Optimal}}. \end{aligned} \tag{10}$$

This indicator may be further refined according to the factors prevailing in the prediction task setting which determine the reliability of the rank ordered data. Given a threshold T which determines the number of ranks that are considered important for the prediction task (e.g., the number of search results that can be displayed on one web-page), $\boldsymbol{v}$ may be defined as follows (Le and Smola, 2007):

$$v_f = \left\{ \begin{array}{ll} 1/\left(\log_2(f) + 1\right) & \forall f \leq T \\ 0 & \forall f > T \end{array} \right\}, \quad f = 1, \ldots, m_j. \tag{11}$$

For horseracing data, as discussed above, the reliability of the rank ordered data is likely to severely decrease after rank three (because horses finishing worse than third are generally not awarded prize money). Consequently, to incorporate this consideration into NDCG, v_i can be set to zero for all $i > 3$. Referring to the earlier search engine analogy, finish positions which are not associated with any prize money represent results of a query which are presented on the second or later pages of a result dialog (see Le and Smola, 2007).

2.4 Architecture of two-stage models

Benter (1994) was the first to develop a computer model for predicting winning probabilities of horses in two-stages. He first conducted a multivariate linear regression of horses' finishing positions, employing different measurements of horses' past performances and physical abilities as independent variables. In the second stage, winning probabilities are estimated using a CL procedure which combines the estimates from stage one with odds implied probabilities (derived from the closing odds and representing the publics' opinion of the winning chances of each runner Johnson *et al.* (2006)).

Edelman (2006) revised this procedure by replacing the multivariate regression in stage one with a SVR model. In particular, in a horserace context, numerous factors are potentially relevant for assessing a horse's

chance of winning. Forecasting models, therefore, have to be robust towards a large number of, commonly highly correlated, inputs. This is a core feature of SVM-type methods which balance the conflicting goals of modelling the training sample with high accuracy whilst avoiding overfitting the noise in the data (Vapnik, 1995). The approach adopted in this paper builds on Edelman's (2006) work by employing a NDCG-based model selection procedure to tune the hyperparameters of the LS-SVR model in stage one.

In outlining the procedures used to develop the LS-SVR-based model it will be assumed that a horseracing database has the form:

$$D = \left\{ \left(\boldsymbol{x}_i^j, q_i^j, y_i^j \right) \right\} \quad i = 1, \ \ldots, M; \ j = 1, \ \ldots R, \tag{12}$$

where R is the number of races, M denotes the number of runners in the database, the vector $\boldsymbol{x}_i^j$ represents the fundamental variables associated with horse i in race j, and y_i^j represents its finishing position; used as a target variable in the first modelling stage. Note that, in order to account for the fact that the number of horses per race varies, the normalised finishing position here takes a value between 0.5 and -0.5 (i.e. the first half of the field takes positive values and second half of the field takes negative values) using the formula (13):

$$\hat{y}_i^j = -0.5 + \frac{\left(y_i^j - \mathrm{Min}_j\left(y_i^j \right) \right)}{\mathrm{Max}_j\left(y_i^j \right) - \mathrm{Min}_j\left(y_i^j \right)} \quad \forall i = 1, \ \ldots, M. \tag{13}$$

Previous studies have demonstrated that the closing odds of a horse u_i^j are closely related to its probability of winning. They may thus unduly influence the prediction model and mask the effect of other fundamental variables (Benter, 1994; Edelman, 2006; Sung and Johnson, 2007). Consequently, closing odds are not considered until the second modelling stage. In other words, the objective of stage one is to estimate a horse's normalised finish position solely on the basis of fundamental variables. Therefore, a LS-SVR model is constructed on a sub-sample $D_1 = \{(\boldsymbol{x}_i, y_i^j)\}_{i=1}^{M_1}$ of D, where M_1 gives the size of this first stage sample.

In the second stage model, closing odds of a horse u_i^j are first converted to the odds implied probabilities Q_i^j via equation Default (14) and these in turn are converted to 'normalised odds implied probabilities,' which sum to one in a race, via the normalisation process shown in (14).

$$\begin{aligned} Q_i^j &= \frac{1}{1 + u_i^j} \\ q_i^j &= Q_i^j \Big/ \textstyle\sum_{i=1}^{m_j} Q_i^j. \end{aligned} \tag{14}$$

A CL model is then developed using the remaining sub-sample (i.e. $D \setminus D_1$). This employs the output of the first stage LS-SVR model and the

natural logarithm of the normalised odds implied probabilities $\ln(q_i^j)$ as independent variables. Unlike the first stage model, the dependent variable of this second stage model is a discrete variable which takes the value one for a winner and zero otherwise. This CL model aims to predict a vector of winning probabilities $\boldsymbol{p}_i^j = \left(p_1^j, p_2^j, \ldots, p_{m_j}^j\right)$ for race j, where p_i^j is the estimated model probability of horse i winning race j. It can be shown that p_i^j is given by the following CL function (McFadden, 1974):

$$p_i^j = \frac{\exp\left(\alpha \hat{y}_i^j(\boldsymbol{x}_i) + \beta \ln\left(q_i^j\right)\right)}{\sum_{i=1}^{m_j} \exp\left(\alpha \hat{y}_i^j(\boldsymbol{x}_i) + \beta \ln\left(q_i^j\right)\right)}. \tag{15}$$

where α and β are estimated using maximum likelihood procedures. This second-stage model is designed to capture the subtle relationships between a runner's normalised odds implied probability, its SVR-based assessment, and the outcome of a race.

3. EMPIRICAL EVALUATION OF THE PROPOSED MODEL SELECTION STRATEGY

3.1. Experimental setup

To explore the potential of NDCG as a means of effectively guiding the search for suitable LS-SVR parameter settings, an empirical study is conducted using real-world horseracing data collected between May 1995 and August 2000 at Goodwood racetrack in UK. The fundamental variables describing a horse's ability and past performances (Table 2) mimic those included in Bolton and Chapman's (1986) seminal paper on horserace modelling. These variables are pre-processed in a similar manner to Lessmann *et al.* (2007) (i.e. they are standardised to zero mean and unit variance).

Overall, the data set includes 556 races which are partitioned into disjoint sub-samples for each individual modelling stage. In particular, the first 400 races (run before May 1999) are treated as a training set and are partitioned evenly into a first (S_1) and second (S_2) stage sample. The remaining 156 races are set aside for out-of-sample evaluation of the final forecasting model. This is achieved via a betting simulation, which adopts a 'Kelly wagering strategy' (Kelly, 1956).

The LS-SVR model is constructed on S_1, which comprises 200 races of 2116 runners in total. A 10-fold cross-validation setup (Stone, 1974) is employed to organise model selection. That is, S_1 is split into ten sub-samples of approximately equal size and LS-SVR models are recursively constructed on nine combined samples and applied to the remaining one (validation set). An estimate of the model's predictive accuracy is obtained by averaging over the individual performance assessments on the ten validation sets. This is undertaken for every parameter combination over an initial grid of $\log(C) = [-3; 5]$ and $\log(\sigma) = [-2.5; +2.5]$ with 10*10 cells. Then, the

TABLE 2
DEFINITIONS OF THE INDEPENDENT VARIABLES EMPLOYED IN THE ONE- AND TWO-STEP MODELS

Independent variable	Variable definitions
Market-generated variable	
$\ln(q_i^j)$	The natural logarithm of the normalised track probabilities
Fundamental variables	
pre_s_ra	Speed rating for the previous race in which the horse ran
avgsr4	The average of a horse's speed rating in its last 4 races; value of zero when there is no past run
disavesr	The average speed rating of the past runs of each horse at this distance; value of zero when no previous run.
go_avesr	The average speed rating of all past runs of the horse on this going; value of zero when no previous run.
draw	Post-position in current race
eps	Total prize money earnings (finishing first, second or third) to date/Number of races entered
newdis	1 indicates a horse that ran three or four of its last four races at a distance of 80% less than current distance, and 0 otherwise
weight	Weight carried by the horse in current race
win_run	The percentage of the races won by the horse in its career
jnowin	The number of wins by the jockey in career to date of race
jwinper	The winning percentage of the jockey in career to date of race
jst1miss	1 indicates when the other jockey variables are missing; 0 otherwise

grid is refined over the most promising region of the hyperparameter space to expand the search in this area, up to a maximal level of two refinements (Van Gestel *et al.*, 2004). Following this strategy, 300∗10 models are constructed and evaluated during LS-SVR model selection. The hyperparameter values yielding the overall "best performance" are maintained. These are used to construct an LS-SVR model on S_1. The whole procedure is traversed twice, using either MSE or NDCG for assessing the merit of a particular hyperparameter setting.

Having completed model selection, the resulting LS-SVR model is used to estimate normalised finishing positions of the runners in S_2. These are fed into the second stage CL-model. The parameters (α, β) of this model are obtained by maximising the joint likelihood (16) of observing the respective race results, assuming that p_i^j is as above (15):

$$(16) \qquad L(\alpha,\beta) = \prod_{i=1}^{|S_2|} p_i^j = \prod_{i=1}^{|S_2|} \frac{\exp\left(\alpha \hat{y}_i^j(\boldsymbol{x}_i) + \beta \ln\left(q_i^j\right)\right)}{\sum_{i=1}^{m_j} \exp\left(\alpha \hat{y}_i^j(\boldsymbol{x}_i) + \beta \ln\left(q_i^j\right)\right)}$$

Following the procedures outlined above, two final forecasting models are developed, which enable winning probabilities of horses in future races to be predicted on the basis of normalised odds implied probabilities and estimated normalised finishing positions. They differ in terms of the model selection

criterion employed in stage one, i.e. MSE or NDCG. In order to appraise the merit of these forecasting models, their respective profitability is measured by simulating a '*Kelly* wagering strategy' over the 156 out-of-sample races run after May 1999 (see Sung *et al.*, 2005; Sung and Johnson, 2007 for details). The *Kelly* wagering strategy identifies how much to bet on each horse. Let r_i^j be the return on a bet of one pound if horse i wins race j and let b_i^j be the fraction of current wealth that is bet on horse i,. Given that horse h wins race j, current wealth increases by a factor:

$$1 - \sum_{i=1}^{m_j} b_i^j + b_h^j \cdot r_h^j, \tag{17}$$

The *Kelly* strategy determines the bets to maximise the expected log payoff across all potential winners h:

$$\max_{b_h^j} \sum_{h=1}^{m_j} p_h^j \cdot \ln\left(1 - \sum_{i=1}^{m_j} b_i^j + b_h^j \cdot r_h^j\right), \tag{18}$$

It is important to note that the setup considered here employs NDCG only to guide the search for appropriate SVR hyperparameter settings during model selection. In particular, the first-stage SVR model with given hyperparameter values is applied to score the runners within a set of validation races. Subsequently, NDCG is used to measure predictive performance over each race in this sub-sample and the merit of the respective hyperparameter setting is given by the means NDCG over the whole validation sample.

3.2. Experimental results

The ability of the forecasting model combining LS-SVR (with traditional MSE-based model selection) and CL to extract information from the underlying data is confirmed when examining the performance of the holdout sample bets: A Kelly wagering strategy (without reinvestment) based on the predicted winning probabilities of the two-stage forecasting model yields a remarkable return of 10.96%. For comparison purposes a two-stage CL model is considered. This model processes fundamental variables by means of CL regression in stage one and then pools the resulting estimated winning probabilities with normalised odds implied probabilities in a second stage as in (15). As shown in Sung *et al.* (2005) and Sung and Johnson (2007), this procedure represents a very high benchmark. However, the respective rate of return when run on the same dataset is only 1.75%. When reinvestment of winnings is permitted, the LS-SVR model produces an increase in wealth of 112.20% over the 156 holdout races. On the contrary, wealth decreases by 16.53% when applying the model using CL in both stages (see Figure 2). As both techniques differ only in the first modelling stage, it can be concluded that the utilisation of LS-SVR offers a significant improvement. This may be attributed to the fact that the nonlinear

RBF kernel function which is employed enables LS-SVR to identify relationships among the fundamental variables which the linear CL procedure is unable to capture. This view is supported by the fact that the only variables used in the two models are those included in Bolton and Chapman's (1986) seminal paper. The results of this paper have been widely publicised and it is, therefore, likely that the betting public to a large extent discount the information contained in these variables in market odds; only complex, nonlinear interaction remaining concealed.

The previous results confirm the appropriateness of the forecasting paradigm initially proposed in Edelman (2006). However, despite the remarkable success of the LS-SVR/CL combination, this approach can be augmented by employing NDCG for LS-SVR model selection. Predictions from the NDCG approach applied to the same holdout sample used for testing the MSE approach, offers an additional 56% profit (without reinvestment), leading to an overall rate of return of 17.08%. Here, reinvestment of winnings produces a wealth increase of 172.48%. The results of a betting simulation using a Kelly wagering strategy based on the predictions from all three procedures over the 156 holdout races is illustrated in Figure 2.

When comparing the prediction performance of the two LS-SVR based forecasting models, it is important to note that they differ only in terms of the accuracy indicator used in model selection to guide the search for suitable hyperparameter values. Consequently, any observed difference between these two can be attributed to different hyperparameters and therewith the effectiveness of the respective model selection criterion to identify appropriate prediction settings. In fact, the optimal hyperparameters selected by MSE and NDCG vary by several orders of magnitude ($[C = 0.11; \sigma^2 = 18.85]$ c.f. $[C = 0.004; \sigma^2 = 1198.99]$, respectively),

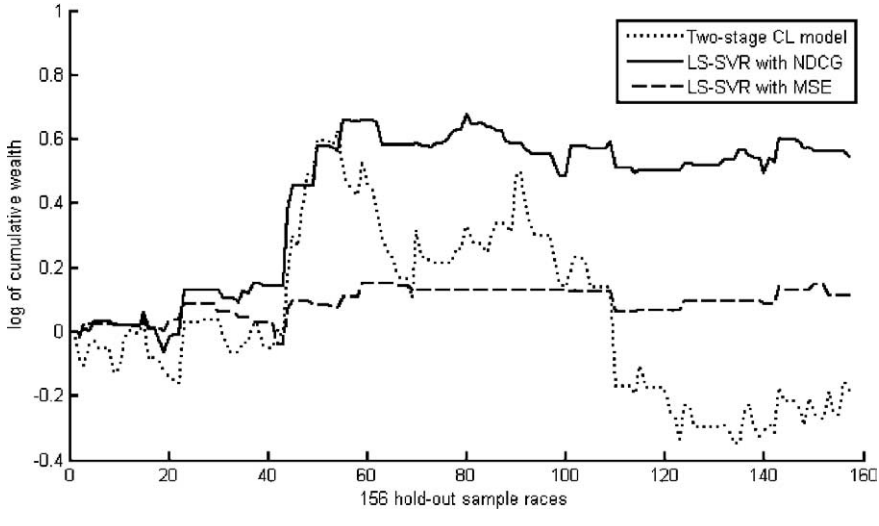

FIGURE 2. Wealth as a result of applying a Kelly-wagering strategy to the holdout sample.

thus leading to fundamentally different LS-SVR models. For example, the smaller C value resulting from NDCG-based model selection indicates that this approach favours less complex models (i.e. with a flatter estimation function). A higher tolerance towards deviations between estimated and actual finishing positions follows directly from the construction of NDCG, punishing prediction errors only on top-ranked horses. Therefore, the higher profitability of the model produced by NDCG demonstrates that this performance indicator better reflects the nature of horserace modelling.

This is further confirmed when examining the R^2 values associated with the two models. The R^2 of the NDCG-based model (0.10) is greater than that of the MSE-based model (0.08); indicating that the former model captures more useful information contained in the fundamental variables. The processing of these fundamental variables within the first stage model is, in turn, heavily affected by the smoothing parameter, σ^2, of the RBF kernel. In particular, NDCG-model selection leads to a significantly higher degree of smoothing. Consequently, MSE results in inferior out-of-sample performance because it produces overly sensitive models which are susceptible to over-fitting noise among fundamental variables. This may arise because the MSE approach attempts to fit the estimation function so as to predict the finishing position of *all* runners as accurately as possible (including those which are less relevant from a betting perspective). This effect may be exaggerated in horserace prediction since different performance indicators are necessarily used for model selection in stage one and for evaluation of the final forecasting model. To confirm this, both MSE and NDCG are adopted to assess predictive accuracy on the holdout sample of 156 races and the results of the two LS-SVR models are shown in Table 3. It is shown that both measures are well suited for hyperparameter selection. In particular, the resulting models achieve the best out-of-sample performance when measured in terms of the criteria which is used for model selection in stage one. Thus, the model tuned in terms of MSE in stage one produces estimates with a lower (better) MSE on the 156 holdout races than the model tuned in terms of NDCG. Similarly, the model tuned in terms of NDCG in stage one produces a higher (better) mean NDCG value over the 156 holdout races than the model tuned on MSE. However, in a horseracing context, the ultimate objective of model selection is to identify hyperparameters values for the first stage model which, at the end of stage two, enable the construction of profitable prediction models. Therefore, it may be concluded that the observed superiority of

TABLE 3
RESULTS ON THE OUT-OF-SAMPLE DATA IN TERMS OF MSE AND NDCG

	LS-SVR model tuned in terms of MSE	LS-SVR model tuned in terms of NDCG
MSE on 156 holdout races	0.406	0.414
NDCG averaged over 156 holdout races	0.654	0.673

NDCG over MSE indicates that the former is better suited to achieving this objective.

4. CONCLUSION

This paper builds upon previous results of Edelman (2006) by developing a two-stage modelling technique for forecasting horseracing outcomes with a novel means of parameter selection in stage one. The first stage consists of a regression of fundamental variables describing horses' abilities and past performances' on finishing positions by means of LS-SVR. In order to better reflect the objective of developing a model that enables profitable betting, a novel approach for selecting the respective LS-SVR hyperparameters values in stage one is proposed. The NDCG performance indicator emerges as a very promising candidate to guide the search for predictive parameter settings. An empirical experiment using data from 556 races reveals significant improvements of the augmented model over challenging competitors (i.e. two-step models that use either CL or LS-SVR guided by MSE in stage one).

Two-step procedures for horserace forecasting are much championed in the literature and represent the state-of-the-art in the field. In particular, the inclusion of modern machine learning techniques like SVM-type methods in stage one has been shown to enhance predictive accuracy. This has led to the development of hybrid prediction models which have been shown to be superior to approaches relying solely on traditional statistics. On the other hand, usage of machine learning methods has so far been restricted to stage one, whereas the task of predicting winning probabilities is predominantly left to CL. In fact, CL may be seen as the best approach available today to account for within race competition because the winning probability of a runner is estimated in relation to the chances of competing horses. As a consequence, the first modelling stage can only represent to a limited degree the objectives of horserace forecasting. Aspects like competition, profitability, and risk elude analysis at this stage. On the one hand, the success of these methods justifies their application in horserace modelling and demonstrates that they are capable of discerning information from fundamental variables which is not taken into account by the betting public. However, the results presented here indicate that there is further room for improvement. In particular, approaches which narrow the gap between traditional function estimation (e.g., LS-SVR) and the objectives of a successful horserace prediction model appear promising. NDCG achieves this goal at the model selection level. Thus, although a standard least-square loss function is optimised "inside" LS-SVR, the selected hyperparameter values may increase or decrease the models suitability for horserace forecasting tasks. This follows directly from the fact that hyperparameters selected by NDCG yield a significantly higher profit than those favoured by MSE. Consequently, this may suggest that further profitability gains are achievable if procedures are developed that embody the state-of-the-art machine learning knowledge (i.e. are based upon structural

minimisation) and, at the same time, optimise loss functions which reflect the particularities of horseracing. In other words, this paper demonstrates that there are a number of areas associated with model development which can influence the forecasting accuracy: for example, model type (e.g., CL and SVM), loss function (e.g., least-square and ε-sensitive), and model selection criterion (e.g., MSE and NDCG).

However, SVM-type models are black-boxes in the sense that they offer no structural understanding to explain their predictive performance. Consequently, the noteworthy profit produced by the LS-SVR/CL models cannot be traced back to particular relationships among variables. Additional methods are required to clarify these relationships and provide a deeper understanding of market phenomena. However, the immediate benefit of powerful modelling techniques like SVM is to demonstrate that the true relationship between independent variables in models to predict race outcomes has not been fully captured by traditional modelling approaches. In this respect, the results reported here suggest that previous findings based on conventional techniques may well have overestimated the degree to which market participants discount information in prices.

NOTE

1. Source: Estimate from Ladbrokes, the UK's largest bookmaking organisation.

REFERENCES

B Baesens, T Van Gestel, S Viaene, M Stepanova, J Suykens and J Vanthienen 'Benchmarking state-of-the-art classification algorithms for credit scoring' *Journal of the Operational Research Society* (2003) 54 627–635.

W Benter 'Computer based horse race handicapping and wagering systems: A report' in DB Hausch, VSY Lo and WT Ziemba (eds) *Efficiency of Racetrack Betting Markets* (London, Academic Press, 1994) pp 183–198.

R N Bolton and R G Chapman 'Searching for positive returns at the track: A multinomial logit model for handicapping horse races' *Management Science* (1986) 32 1040–1060.

B E Boser, I M Guyon and V N Vapnik 'A training algorithm for optimal margin classifiers' in D Haussler (ed) *Proc. of the 5th Annual Workshop on Computational Learning Theory* (Pittsburgh, Pennsylvania, USA, ACM Press, 1992) pp 144–152.

J S Breese, D Heckerman and C Kadie 'Empirical analysis of predictive algorithms for collaborative filtering' in GF Cooper and S Moral (eds) *Proc. of the 14th Annual Conf. on Uncertainty in Artificial Intelligence* (Madison, Wisconsin, USA, Morgan Kaufmann, 1998) pp 43–52.

Y Cao, J Xu, T-Y Liu, H Li, Y Huang and H-W Hon 'Adapting ranking SVM to document retrieval' in EN Efthimiadis, ST Dumais, D Hawking and K Järvelin (eds) *Proc. of the 29th Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval* (Seattle, WA, USA ACM, 2006) pp 186–193.

O Chapelle, V Vapnik, O Bousquet and S Mukherjee 'Choosing multiple parameters for support vector machines' *Machine Learning* (2002) 46 131–159.

K-M Chung, W-C Kao, L-L Wang and J Lin 'Radius margin bounds for support vector machines with RBF kernel' *Neural Computation* (2003) 15 2643–2681.

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

N Cristianini and J Shawe-Taylor *An Introduction to Support Vector Machines and other Kernel-based Learning Methods* (Cambridge, Cambridge University Press, 2000).

K Duan, S S Keerthi and A N Poo 'Evaluation of simple performance measures for tuning SVM hyperparameters' *Neurocomputing* (2003) 51 41–59.

D Edelman 'Adapting support vector machine methods for horserace odds prediction' *Annals of Operations Research* (2006) 151 325–336.

D B Hausch and W T Ziemba 'Transactions costs, market inefficiencies and entries in a racetrack betting model' *Management Science* (1985) 31 381–394.

C-W Hsu, C-C Chang and C-J Lin 'A practical guide to support vector classification' *Department of Computer Science and Information Engineering,* Working paper, National Taiwan University (2003).

K Järvelin and J Kekäläinen 'IR evaluation methods for retrieving highly relevant documents' in NJ Belkin, P Ingwersen and M-K Leong (eds) *Proc. of the 23rd Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval* (Athens, Greece, ACM Press, 2000) pp 41–48.

T Joachims 'Estimating the generalization performance of an SVM efficiently' in P Langley (ed) *Proc. of the 17th Intern. Conf. on Machine Learning* (Stanford, CA, USA, Morgan Kaufmann, 2000) pp 431–438.

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

S Keerthi, V Sindhwani and O Chapelle 'An efficient method for gradient-based adaptation of hyperparameters in SVM models' in B Schölkopf, JC Platt and T Hoffman (eds) *Advances in Neural Information Processing Systems 19* (Cambridge, MIT Press, 2007) pp 217–224.

S S Keerthi and J Lin 'Asymptotic behaviors of support vector machines with Gaussian kernel' *Neural Computation* (2003) 15 1667–1689.

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

D Law and D A Peel 'Insider trading, herding behaviour and market plungers in the British horse-race betting market' *Economica* (2002) 69 327–338.

Q Le and A Smola, Direct Optimization of Ranking Measures (electronic paper, http://www.citebase.org/abstract?id = oai:arXiv.org:0704.3359, 2007).

S Lessmann, M-C Sung and J E V Johnson 'A new method for predicting the outcome of speculative events' *Centre for Risk Research*, Working paper CRR-07-03, University of Southampton (2007).

S D Levitt 'Why are gambling markets organised so differently from financial markets?' *The Economic Journal* (2004) 114 223–246.

D McFadden 'Conditional logit analysis of qualitative choice behavior' in P Zarembka (ed) *Frontiers in Econometrics* (New York, Academic Press, 1974) pp 105–142.

G S Maddala *Limited Dependent and Qualitative Variables in Econometrics* (New York, Cambridge University Press, 1983).

R D Sauer 'The economics of wagering markets' *Journal of Economic Literature* (1998) 36 2021–2064.

A Schnytzer and Y Shilony 'Inside information in a betting market' *The Economic Journal* (1995) 105 963–971.

A J Smola and B Schölkopf 'A tutorial on support vector regression' *Statistics and Computing* (2004) 14 199–222.

M Stone 'Cross-validatory choice and assessment of statistical predictions' *Journal of the Royal Statistical Society (Series B)* (1974) 36 111–147.

M-C Sung, J E V Johnson and A C Bruce 'Searching for semi-strong form inefficiency in the UK racetrack betting market' in L Vaughan Williams (ed) *Information Efficiency in Financial and Betting Markets* (Cambridge, Cambridge University Press, 2005) pp 179–192.

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

M Sung and J E V Johnson 'The influence of market ecology on market efficiency: Evidence from a speculative financial market' *Journal of Gambling, Business and Economics* (2007) 1(3) 185–198.

J A K Suykens 'Support vector machines: A nonlinear modelling and control perspective' *European Journal of Control* (2001) 7 311–327.

J A K Suykens, L Lukas, P Van Dooren and B De Moor, V J 'Least squares support vector machine classifiers: A large scale algorithm' *Proc. of the European Conf. on Circuit Theory and Design* (Stresa, Italy, 1999) pp 839–842.

J A K Suykens, T Van Gestel, J De Brabanter, B De Moor and J Vandewalle *Least Squares Support Vector Machines* (Singapore, World Scientific, 2002).

J A K Suykens and J Vandewalle 'Least squares support vector machine classifiers' *Neural Processing Letters* (1999) 9 293–300.

T Van Gestel, J A K Suykens, B Baesens, S Viaene, J Vanthienen, G Dedene, B De Moor and J Vandewalle 'Benchmarking least squares support vector machine classifiers' *Machine Learning* (2004) 54 5–32.

V N Vapnik *The Nature of Statistical Learning Theory* (New York, Springer, 1995).

L Vaughan Williams 'Information efficiency in betting markets: A survey' *Bulletin of Economic Research* (1999) 51 1–39.