ARTICLES

The Journal of Prediction Markets 2024 Vol 18 No 1 pp 3-20, DOI:10.5750/jpm.v18i1.2093

CLASSIFICATION IN HORSE RACE PREDICTION THROUGH PRINCIPAL COMPONENT DECOMPOSITION

*Jason West**
Bureau of Meteorology,
Brisbane, QLD, 4000,
Australia

Vlad Kazakov
Racelab Global, Sydney,
NSW, 2000,
Australia

ABSTRACT

The established view for horse race handicapping and staking strategies is to model them as a classification problem using factors describing horse, jockey, trainer, and racing history coupled with public odds, solved via a logistic regression. Logistic regression probabilities are then normalised, and bets filtered by threshold, or anomalous pricing. However, published algorithms do not show systematic profitability, nor do machine learning approaches using algorithmic betting strategies. This deficiency is due to three factors. First, wins are rare and racing data are thus imbalanced. Second, racing factors are multicollinear. Third, the number of factors needed for accurate prediction is very large. We show that alternative methods using variants from principal component analysis produces sustainable profitability regardless of staking strategy through a reduction of factors to fundamental drivers. We apply a partial least squares regression methodology to Australian thoroughbred racing. This approach is shown to outperform logistic regression and machine learning methods in classifying winners for a profitable trading strategy. This method can be applied to multiple betting domains.

Keywords: partial least squares, logistic regression, horse racing, imbalanced data

1 INTRODUCTION

The established analytical approach for thoroughbred race handicapping and staking strategies is for odds estimation in betting to be modelled as a classification problem using factors describing the horse, jockey, and or trainer, as well as training and racing history, coupled with public odds, and solved

*Corresponding author: e-mail: jason.west@bom.gov.au using a logistic regression. Logistic regression probabilities are then normalized, and optimal bets can be chosen using a threshold filter or by searching for under-priced runners and applying a staking process aligned with the Kelly criterion (Kelly, 1956). Benter (1994) famously demonstrated the capacity to earn positive returns using a computer-generated betting strategy based on a variant of this approach. Well-known gambling identities Alan Woods and Patrick Veitch have described detailed implementation methods and factors used to rank thoroughbreds for odds estimation. However, the durability of these methods in practice has been limited.

To the best of our knowledge, most published algorithms fail to show systematic profitability. Similarly, none of the published solutions from the machine learning domain result in a viable betting algorithmic strategy. One can develop and apply any of them, adapt them for local data, generate bets and the resulting expectation is that one will lose money over the medium term. Sustained profitability based on persistent pricing anomalies from racing metrics is difficult to achieve in practice.

There are several reasons explaining the dislocation between a theoretically proven betting/staking strategy and profitability. First, wins are rare and racing data are imbalanced. Classifiers, including logistic regression methods, will underestimate the probability of rare events unless this imbalance is corrected. Popular approaches to rebalancing data are divided into pre-processing, post-processing, and hybrid methods. Pre-processing is a simple but effective technique that avoids issues of signal detection using arbitrary thresholds from model estimates. Any one of several pre-processing corrections can be applied, including up-sampling, down-sampling, and class re-weighting. In this analysis we adopt the pre-processing technique of up sampling (scaling) to address the imbalance which retains the benefits of a large sample size to maximise predictive power.

Second, many of the factors used to predict racing ability are multicollinear. To defend against model misspecification, regression methodologies demand the use of independent factors or at least the use of an alternative approach whereby factors are transformed into principal components. Finally, racing guides contain, and experts use, many more factors than the 20 or so factors often cited in published algorithms. Racing data is now so prevalent that potentially thousands of factors are available for analysis, many of which are at least weakly informative and thus can provide incremental information for prediction.

In this analysis we show how classification accuracy improves when adjusting for imbalanced data using a pre-processing approach. We then define principal components, using a variant of Partial Least Squares (PLS) estimation, from a large set of factors and align them with expert knowledge and publicly available odds. We show that by integrating these derived factors with a logistic regression approach greatly reduces prediction error and improves profitability regardless of staking strategy. This approach can accommodate any number of correlated factors from data with limited sample size and missing values, with only a minor impact on prediction accuracy.

2 METHODOLOGY

The application of conditional logistic regression to horse racing was popularised by Boltman and Chapman (1986) and extended by a range of authors (Benter, 1994; Edelman, 2006; Silverman and Suchard, 2013). The evolution of the research in this field has been aimed at improving the calculation of a horse's "strength" estimated using a conditional logistic regression on identifiable factors and combining this with a payoff dividend related to market odds. The likelihood of the conditional logit can be expressed as

$$\Pi_{r=1}^{R} \frac{e^{\alpha_{rh}^{w}\beta_1+p_{rh}^{w}\beta_2}}{\sum_{h\in r} e^{\alpha_{rh}^{w}\beta_1+p_{rh}^{w}\beta_2}}, \tag{1}$$

which combines previous model predicted "strength" α = $\{1,2,\ldots,A\}$ and historical limit A with the "odds implied probability" $p = \{1,2,\ldots,P\}$ for a runner. The payoff dividend for a horse is converted to an implied probability $p(x_h) = \frac{1}{d_h}$ or $\tilde{p}(x_h) = \frac{1}{d_h+\delta}$ where δ is the track take from a parimutuel pool. The value p is the bettor implied confidence rather than a genuine estimate of outcome probabilities which is an aggregate estimate of the betting public's opinion.

The expression above expression is formulated as follows. A race $r = \{1,2,\ldots,R\}$ is run between multiple horses $h = \{1,2,\ldots,H\}$ with a single runner winning. The racing characteristics of each horse is represented by a k dimensional vector X_{rh} with the covariates X represented by an $N \times K$ dimensional matrix where each row represents the covariates of each horse / jockey / trainer characteristics, β as a column of k regression coefficients, and p_i as the probability of winning an event (raw response variable). The probability of each runner h winning in race r is

$$p_{rh} = \frac{e^{X_{rh}\beta}}{\sum_{h\in r} e^{X_{rh}\beta}}, \tag{2}$$

where the winning probabilities of race r sum to 1, $\sum p_{rh} = 1$. Estimates for the regression coefficients β in $X_{rh}\beta$ are typically derived using logistic regression

$$\left(\frac{p_r}{1-p_r}\right) = e^{X_{rh}\beta}. \tag{3}$$

The formulation in (1) combines the set of regression coefficient estimates (the "strength" estimate) $\alpha_{rh}^{w}\,\beta_1$ with the odds implied probability $p_{rh}^{w}\,\beta_2$ to derive a fundamental probability estimate for each runner. Strength estimates can be obtained using various techniques including conditional logistic regression (Benter, 1994), support vector regression on a runner's finishing position (Edelman, 2006), support vector classifier coupled with distance from the hyperplane (Lessmann and Sung, 2007), CART (Lessmann and Sung, 2010), LASSO regression or variants of the Cox Proportional Hazards model (Silverman and Suchard, 2013). Despite different methods for calculating the "strength" of a horse, existing models rely on the standard conditional logistic regression as a key component for their prediction algorithm.

There are well-known shortfalls to this approach, and in many cases previous scholars have admitted to these issues as limitations to the generalisability of their results. When a large set of j explanatory variables x_j, $j \in \{1,2,\ldots,n\}$ are required to forecast a single response variable y using a sample of size n where all regressors contain information, the use of ordinary least squares (OLS) or logistic regression (LR) can be problematic when $j >> n$ or j is at least large relative to n. An OLS or logit model of this type may also be mis-specified when the variables x_j are collinear, and where the variance of the model is of order $j \approx n$ the prediction estimates may be biased. Importantly, even when $j > 3$, the OLS or logit estimator can be inadmissible using a mean square error (MSE) criterion (James and Stein, 1961) which can undermine the typical diagnostics applied to such models to test for validity.

These issues have motivated the use of shrinkage estimation methods. Some shrinkage estimators such as RIDGE and LASSO are relatively specific criteria-based choices in the space of OLS-consistent solutions. Recent algorithms have applied random forests and gradient boosting machines (GBM). Whereas random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. These can be combined such that the many weak successive trees produce a powerful ensemble algorithm. However, shrinkage estimation and machine learning derivatives still suffer from the same issues of collinearity and small sample size relative to factors that undermines the more orthodox methods.

We offer an alternative approach to existing approaches that entirely circumvents the limitations related to standard regressions. Partial Least Squares (PLS) techniques, along with other dimension reduction methods, have been used in chemometrics and related applications where $j >> n$. In an early PLS method Wold (1966) uses latent quantitative factor variables (akin to principal components) from the data to regress an outcome variable against many components. The contribution of each variable is evaluated using standardized model coefficients, with the outputs indicating the direction and magnitude of the effect. A positive correlation between the independent variable and the outcome variable is inferred for positive coefficients.

For a typical multiple linear regression approach the least-squares solution for

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}, \tag{1}$$

is

$$\boldsymbol{B} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}. \tag{2}$$

When $j >> n$ and/or in the presence of collinearities the estimate for $\boldsymbol{X^T X}$ becomes singular and unable to converge to a unique solution. The PLS approach averts this by decomposing $\boldsymbol{X}$ into orthogonal 'scores' $\boldsymbol{F}$ and 'loadings' $\boldsymbol{P}$

$$\boldsymbol{X} = \boldsymbol{FP}, \tag{3}$$

and regressing $\boldsymbol{Y}$ not on $\boldsymbol{X}$ itself but on the first n columns of the 'scores' $\boldsymbol{F}$. The PLS approach therefore incorporates information on both $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the definition of the 'scores' and 'loadings'. The PLS algorithm performs a simultaneous bilinear decomposition of the outcome variable and the regressors. Scores $\boldsymbol{F}$ and loadings $\boldsymbol{P}$ are orthogonal and the following decompositions are carried out concurrently

$$x = q_1' f_1 + \cdots + q_u' f_u + E_u, \tag{4a}$$

$$x = p_1' f_1 + \cdots + p_u' f_u + e_u, \tag{4b}$$

where f_i are the individual scores, and p_i and q_i are the loadings generated at each step. The suffix m represents the final step of the calibration process and by design, the algorithm converges in the sense that after p steps the factors will be identical and equal to zero. The recursive formulas for the scores and loadings provide a linear form like the predictions of the estimators

$$y_{PLS} = x'_{\mathrm{t}=0}\beta_{PLS}(m), \tag{5}$$

where

$$\beta_{PLS}(m) = \boldsymbol{W}_m\,(\boldsymbol{W}'_m \boldsymbol{\Sigma} \boldsymbol{W}_m)^{-1}\boldsymbol{W}'_m \sigma_{xy}, \tag{6}$$

and $\boldsymbol{W}_m = (w_1, \ldots ,w_m)$ is obtained after m recursions of the algorithm by stacking weights generated at each step, which are effectively the weighted covariances of the predictors and the response.

We extend the PLS approach to a generalised linear regression model (PLSGLR) which appropriately accounts for missing data. The PLSGLR regression of the response y_{PLS} on variables x_i is defined as

$$g(\theta)_i = \textstyle\sum_{h=1}^{H} (\mathbf{B}\sum_{j=1}^{u} w^*_{jx} x_{ij}), \tag{7}$$

with H components, where θ is a probability vector of the response variable y_{PLS} with a finite support. The components x_{ij} are built to be orthogonal and the link function $g(.)$ is a logistic function to fit the model to the data.

In conceptual terms, the PLS method projects the input and output variables in directions of maximum covariance, with the calibration performed between the orthogonal 'latent' variables. This approach has been shown to be stable with respect to collinearity (Esbensen, 2002) and is able to generate unbiased prediction equations from pre-processed datasets (Sundberg et al., 1999; Wold et al., 2001).

We validate the PLS method using a k-fold leave-one-out cross-validation approach to identify significant variables contributing to the best fit of the model. The advantage of PLS over other approaches is that it identifies only relevant predictor variables, while other linear models require pre-selection of potential predictor variables prior to regression analysis. The disadvantage of PLS is that the resulting components don't necessarily correspond to a specific 'factor' unlike typical regressions that identify predictive attributes directly related to the data.

Model estimation and subsequent prediction is affected by the presence of imbalanced data (many losers, few winners). Importantly, model diagnostics are also affected by imbalanced data. For classification, accuracy and its complement error rate used to assess performance when defined as

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \tag{7}$$

where the correctly classified number of outcomes for True Positive (TP) and True Negative (TN) are combined with incorrectly classified outcomes for False Positives (FP) and False Negatives (FN). High accuracy can be obtained by predicting a loss for all test races while winners are misclassified. This metric is of little use when predicting relatively rare outcomes is the objective.

An alternative metric for imbalanced data is the geometric mean G_{mean} which is defined as

$$G_{mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} = \sqrt{sensitivity \times specficity} \cdot \tag{8}$$

The G_{mean} computes the geometric mean of the accuracies of each class seeking to measure the relative balance between classification. These metrics are used to compare model accuracy in the analysis below.

3 DATA AND SETUP

We use a data set of all competitive thoroughbred horse races in Australia from January 2019 to December 2021 in both metropolitan and provincial grades (but excluding country-level races which are in remote locations with limited competitiveness and market liquidity). Race distances vary from 1000m to 3200m under all weather conditions and multiple field sizes. Hurdles and steeplechase races are excluded. To eliminate possible bias, maidens (runners who are yet to win) are initially excluded, but when added to the training and test data sets for model verification made little difference to accuracy. Events with fewer than five runners, more than one winner (a tie), and for which no pricing data is available were excluded.

Data is obtained for a range of primary variables which are provided in the Appendix. There are hundreds of potential variables that can be used but we constrain the list to 22 of the most promising for this analysis. Figure 1 provides a correlation matrix pictorial of the strength of the relationship between variables over the full period. Several variables are highly correlated (positive in blue, negative in red) which is known to bias regression coefficient estimates. The effects of high correlation are of no concern for the PLS regression when the dimensions of the data are reduced to orthogonal components but have deleterious effects on the applicability of regression methods.

We split the data into two halves with pre-July 2020 designated as the training data set and post-July 2020 as the test data set. This results in

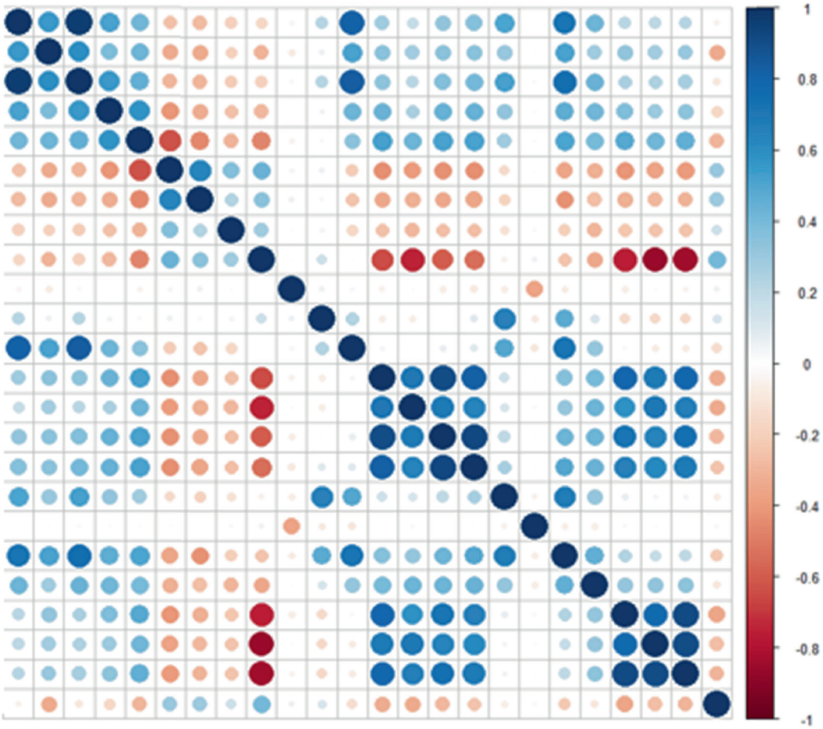

Figure 1. Correlation matrix between factors used in the analysis

Table 1. Chi-squared style comparison of win percentage between training and test data sets, against expected win probability implied by starting prices, Australian thoroughbreds, 2019–2021

Price Category	Training Data Win %	Test Data Win %	Expected Win %
$1 – $2	56.83%	57.46%	56.97%
$2 – $3	39.28%	37.86%	38.35%
$3 – $4	27.65%	28.39%	28.06%
$4 – $5	21.35%	21.64%	22.00%
$5 – $10	13.72%	14.02%	13.25%
$10 – $20	6.97%	6.92%	6.67%
$20 – $50	3.30%	3.38%	3.10%
$50 – $100	1.60%	1.36%	1.41%
$100+	0.45%	0.41%	0.37%

9,060 events in each of the training and test data sets. Table 1 provides a chi-squared style comparison of win percentage both training and test data sets, against expected win probability implied by the starting prices for each event. The χ^2 p-values for win percentages in both the training and test data sets is <0.001. A feature of these results is that no hint of favourite-longshot bias (i.e., overvaluing longshots and undervaluing favourites; Sobel & Raines, 2003) or other forms of bias appears in either data set.

To avoid issues related to bias from the presence of few winners relative to many losers, the training data set was 'up-sampled' using bootstrap resampling so that the number of winners exceeded 40 per cent of the data used to train each model, where the number of observations (individual runners) increased from 85,540 to 147,170. We use both logistic regression and PLSGLM regression to classify runners into winners and losers per race.

4 RESULTS

The logistic regression results using the up-sampled training data set are provided in Table 2.

A total of 14 variables are significant at a 95% confidence level with the AIC of 15,975 marginally improved over the AIC of the alternative formulation to use all variables of 16,025. Variables representing previous wins at the current track, jockey rating, horse rating, and price are strongly related to win probability.

The PLSGLM logistic regression was calibrated using the training data set and the results are provided in graphical form for ease of interpretation. Figure 2 indicates the explanatory strength of each coefficient (numbered 1 to 22 on the x-axis) represented by each regression component (first five components are shown). One variable (price at start of day) exerts a strong influence on the

Table 2. Logistic regression results (odds ratios, 95% confidence intervals, p-values) for thoroughbred races in Australia, Jan 2019-Jun 2020

Variable	Coefficient	Std Err.	Odds Ratio	95% CI	p-value
Start price last start	–0.041	0.012	0.94	(0.87, 1.01)	0.091
Track win (previous)	0.016	0.005	1.17	(1.06, 1.29)	0.002
Grade difference	–0.005	0.001	0.99	(0.98, 1.00)	0.007
Distance difference	0.000	0.000	1.00	(1.00, 1.00)	0.018
Finishing speed	–0.065	0.015	0.93	(0.86, 1.02)	0.110
Horse rating	0.032	0.004	1.03	(1.01, 1.06)	0.006
Jockey rating	0.223	0.012	1.32	(1.24, 1.40)	<0.001
Expected position	–0.002	0.014	1.07	(0.99, 1.17)	0.100
Barrier adjustment	–0.005	0.027	0.80	(0.67, 0.95)	0.009
Quick back up	0.009	0.002	0.98	(0.96, 0.99)	0.007
Handicap	–0.012	0.003	0.96	(0.93, 0.99)	0.002
Peak rating	0.014	0.003	0.94	(0.90, 0.98)	0.002
Race distance	0.000	0.000	1.00	(1.00, 1.00)	0.001
Price at start of day	0.385	0.003	1.57	(1.54, 1.61)	<0.001

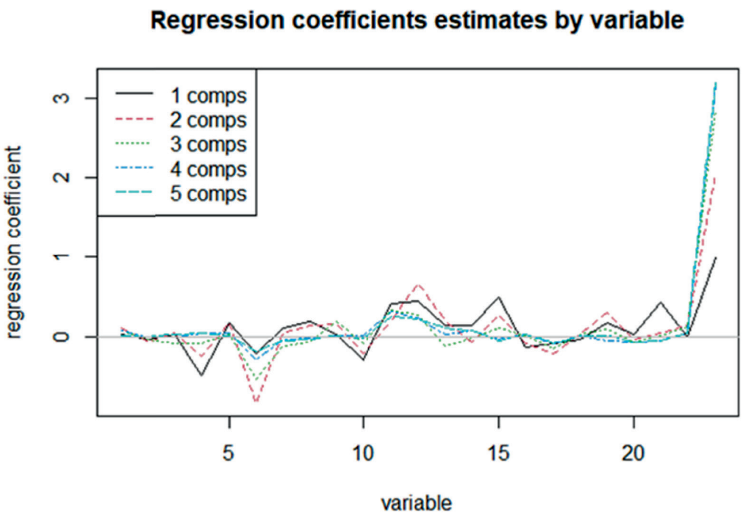

Figure 2. Partial least squares regression coefficient estimates by variable (Australian thoroughbreds Jan 19 – Jun 20)

decomposition relative to other variables in a similar way to its influence over the logistic regression model. The correlation loading between the first two most significant components are provided in Figure 3, which indicates that while the first component is dominant, it is also unbiased.

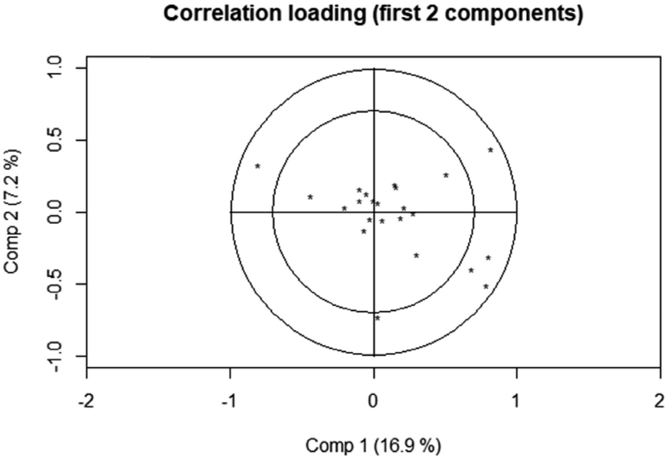

Figure 3. Partial least squares regression correlation loading, first two components (Australian thoroughbreds Jan 19 – Jun 20)

Figure 4 depicts the Root Mean Squared Error of Prediction (RMSEP) from the bias-corrected cross-validation estimate as the number of principal components is added to the model. The RMSEP reaches a stable minimum after 8 components are added. Prediction quality is provided in Figure 5 using the test data, which shows predicted probabilities relative to measured probabilities (implied by market odds). There are some differences at the lower 'scores' (low win probability) but the bulk of the predicted scores for runners at higher probabilities tracks measured values.

A comparison of the effect of up-sampling winners in the training data set is depicted in Figure 6. Estimated prices using the PLSGLM model derived from the up-sampled data are closely aligned with market observed prices (start prices) which have been shown to be relatively accurate (Table 1). Some divergence is present at higher prices (lower odds) however the fact that most trades are placed on runners at odds representing a winning probability greater than around 15 percent is the critical zone in which correct model estimates are important. Model estimates for the PLSGLM model using the non-scaled data indicates a divergence in estimates at higher prices (lower odds) which demonstrates the bias away from winning categories when data is imbalanced.

Diagnostics for the original test data and the up-sampled test data are provided in Table 3. The PLS model produces greater improved accuracy measures over the logistic regression approach as well as a superior G-mean metric.

To compare profitability profiles between models, we apply a simple flat staking betting strategy of $1 for each trade. We trade on the highest rating runner per race using the logistic regression model. We bet on runners whose

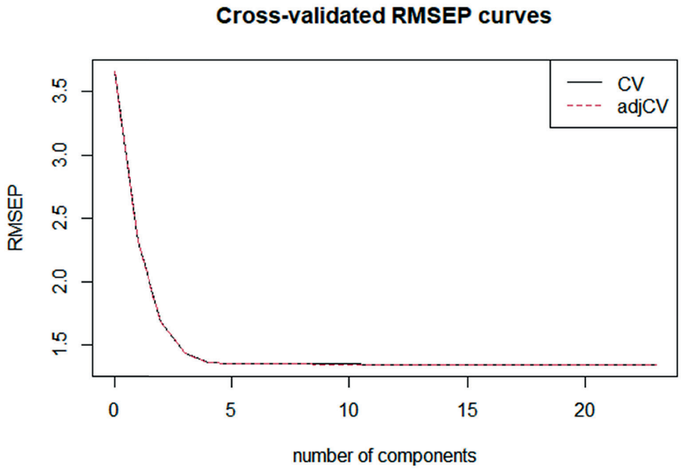

Figure 4. Cross-validated RMSEP curves for PLS regression (Australian thoroughbreds Jan 19 – Jun 20)

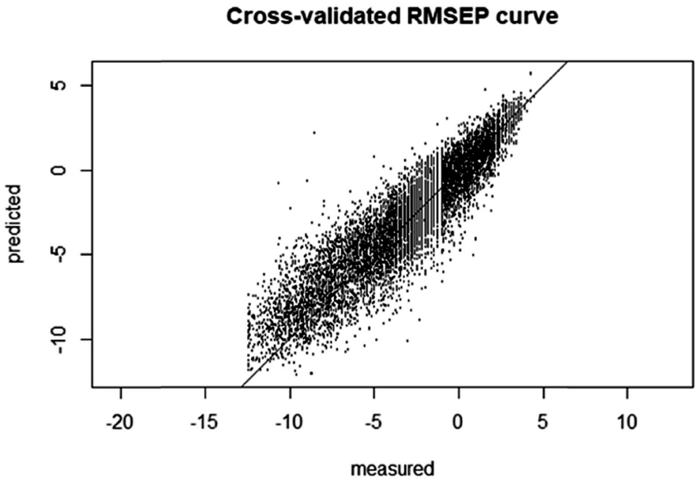

Figure 5. Cross-validated predictions for PLS regression (Australian thoroughbreds Jul 20 – Dec 21)

scores exceed a threshold equivalent to the 80th percentile for the PLS model. This means that for the PLS method, there are potentially multiple bets for some races and zero bets for other races. For comparison, we also use a naïve strategy of simply betting on the race favourite for each race. The trading results comparison is provided in Figure 7. Profitability from trades informed by the PLS approach clearly dominate alternative methods. Surprisingly, the

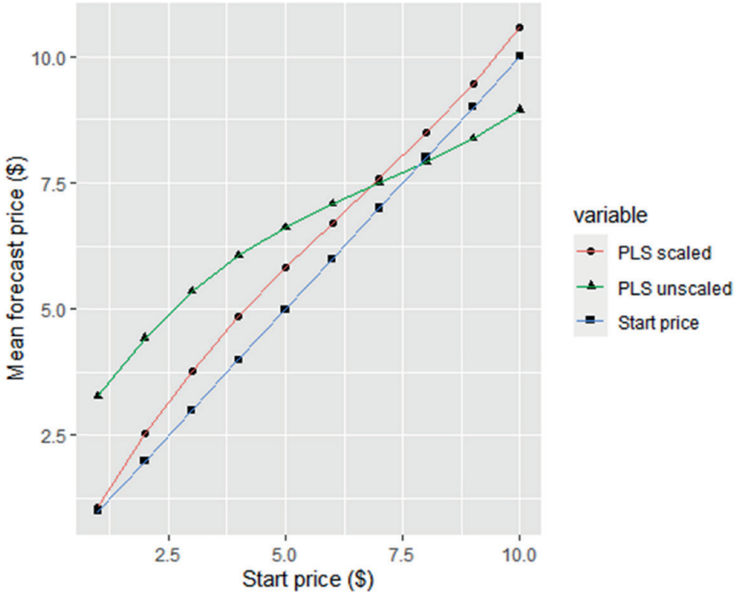

Figure 6. Forecast price against traded starting price for up sampled (scaled) and the unscaled test data sets (truncated at a start price of $10)

Table 3. Model diagnostics for logistic regression and partial least squares models using existing test data and up-sampled test data, Australian thoroughbreds 2019–2021

Test statistic	Log. regression (scaled)	Partial least squares (scaled)
Accuracy	0.857	0.871
(95% CI)	(0.855, 0.859)	(0.868, 0.873)
Sensitivity	0.920	0.925
Specificity	0.325	0.328
Prevalence	0.894	0.916
Detection rate	0.827	0.840
Balanced accuracy	0.623	0.613
G-mean	0.547	0.551

naïve approach generates a profit over the testing period however the resulting return on investment makes this approach infeasible.

The profitability breakdown by price category (i.e., inverse odds) are provided in Table 4. All strategies are unprofitable using flat staking for prices less than $3.00 (odds 4:1 against). While there is a higher chance of

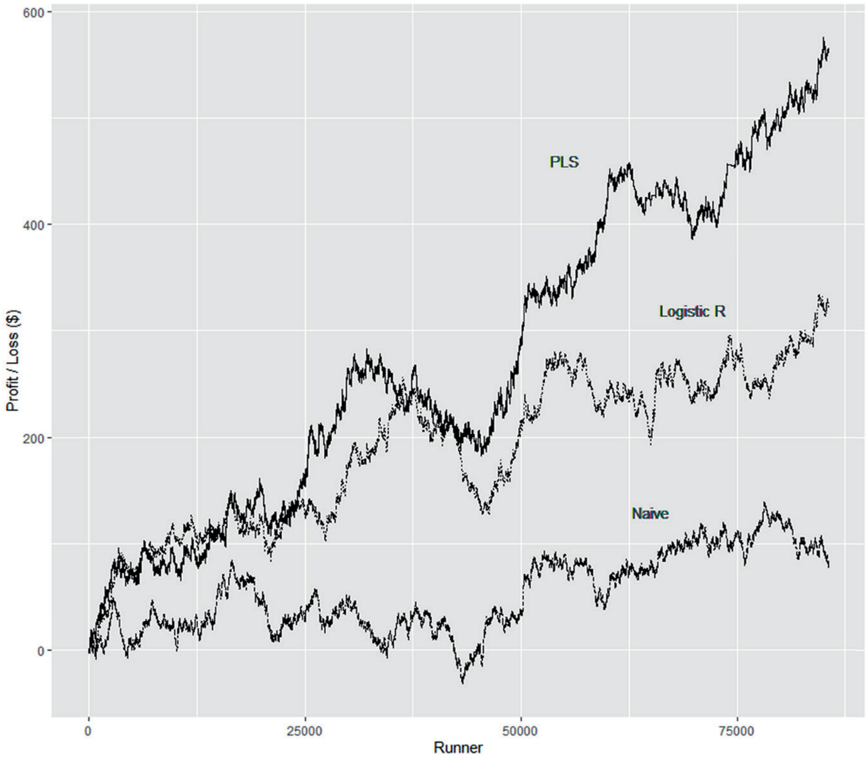

Figure 7. Profit/loss results for PLS versus logistic regression model and naïve methods using a flat staking strategy $1 per bet, Australian thoroughbreds Jul 2020 – Dec 2021

earning a return at lower odds, the ability of the models to distinguish between winners predicted by the model versus those predicted by the market is relatively poor. At higher price categories (i.e., odds greater than 5:1 against), profitability increases. Simply backing the favourite (the naïve strategy) is also unprofitable at lower odds and there are naturally lower profits available at higher odds.

The PLS model earns over half its profits in the $5–$10 price range (win probability of roughly 8.3–14.3 per cent) indicating that shorter priced favourites are often overpriced relative to the model choice. The logistic regression and naïve strategy profitability is highest in the $4–$5 price range (win probability of 14.3–16.6 per cent).

Concerns over model degradation over time resulting in a loss of accuracy for either approach are addressed through the accuracy plot provided in Figure 8. The plot depicts the positive predictive value (precision) as win percentage for each month from July 2020 – December 2021 in the test data defined as TP/(TP+FP). While precision varies from month to month,

Table 4. Profitability by price category for PLS versus logistic regression model and naïve methods using a flat staking strategy $1 per bet, Australian thoroughbreds Jul 2020 – Dec 2021

Price Category	Logistic R.	%	PLS	%	Naïve	%
$1 – $2	–$10.80	–3.3%	–$22.50	–4.0%	–$11.50	–14.9%
$2 – $3	–$18.00	–5.6%	–$22.80	–4.1%	–$97.20	–125.9%
$3 – $4	$63.80	19.7%	$140.80	25.1%	$18.60	24.1%
$4 – $5	$174.20	53.8%	$138.40	24.6%	$159.50	206.6%
$5 – $10	$110.50	34.1%	$327.80	58.4%	$7.80	10.1%
$10 – $20	$60.00	18.5%	$14.00	2.5%	$–	0.0%
$20 – $50	–$38.00	–11.7%	–$10.00	–1.8%	$–	0.0%
$50 – $100	–$12.00	–3.7%	–$3.00	–0.5%	$–	0.0%
$100+	–$6.00	–1.9%	–$1.00	–0.2%	$–	0.0%

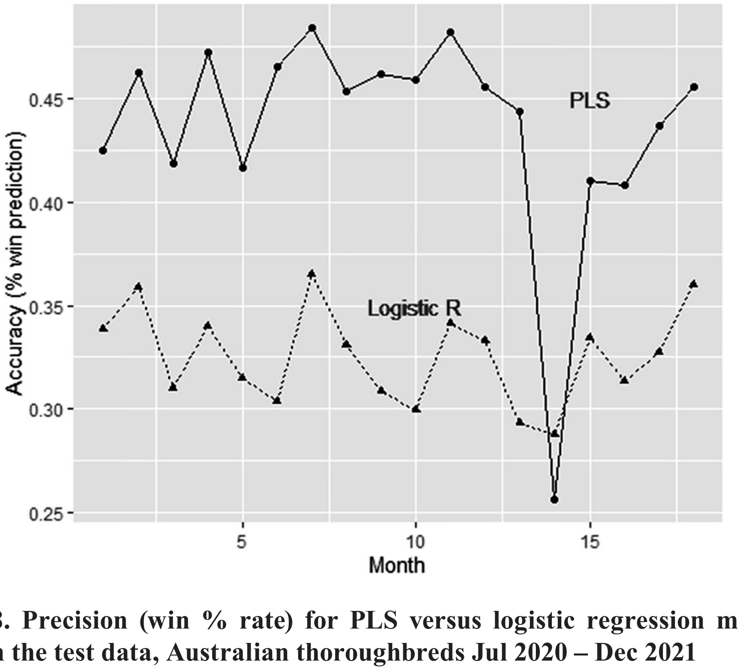

Figure 8. Precision (win % rate) for PLS versus logistic regression model by month in the test data, Australian thoroughbreds Jul 2020 – Dec 2021

there is no persistent rate of degradation over the 18-month test data series. The precision in the PLS model outperforms the logistic regression model in every month apart from August 2021 where the precision of the PLS approach fell to roughly 25 per cent of win predictions compared with 28 per cent for the logistic regression.

5 DISCUSSION

The benefits of the PLS approach extends to the use of many more variables than used in this analysis which can serve as ‘weak learners’ that incrementally improve the information content from large, complex, and interrelated data. The need to account for multicollinearity and interaction terms when using regression and other classifiers in complex data sets is avoided with the PLS approach. While PLS model outputs don’t directly attribute explanatory strength for features (variables) used to estimate principal components as they do in regression models, this is a small price to pay for the opportunity to improve model accuracy.

Probability estimates are greatly improved when accounting for imbalanced data using a relatively simple scaling (up sampling) process. Pre-processing data through bootstrapping the re-weighting towards winners avoids the arbitrary filtering techniques of post-processing methods such as adopting thresholds for selection or using hybrid methods that rely on learning ‘agents’ from alternative classifiers. While post-processing is a valuable tool for prediction in complex settings (e.g., weather forecasting, genomics, computational biology), the simple, but effective, method of up sampling winners achieves close alignment in estimated odds relative to market observed odds in thoroughbred racing. The simpler bootstrapping method also avoids issues of overfitting associated with post-processing methods.

The combination of principal decomposition and data scaling results in superior profitability and return on investment using a narrow data set for estimation. The accuracy can be shown to improve even further with data sets comprising many more features, but the incremental increase in accuracy rapidly diminishes. Profitability can also be enhanced through Kelly staking or proportional Kelly staking strategies, particularly when the PLS method is able to reliably identify winners offering higher market prices that represent a greater return on investment relative to a flat staking strategy.

6 SUMMARY

We have shown that the use of generalised principal component analysis to dissect data for trading strategies can deliver superior accuracy and profitability against existing methods. Also, simple up sampling of categorical data greatly improves accuracy at lower odds. A limitation of PLS in prediction is that its validity pertains to the conditions under which the observed data was obtained. While seasonality is apparent in racing (e.g., high profile runners prepare and compete for racing ‘carnivals’), its effects are generally assumed to be minimal. We did not detect degradation in accuracy due to seasonal effects in the test data. The persistence of factors and the accuracy of the PLS did not materially change when the model was trained on alternative periods or across seasons. Dimension reduction methods for large and complex data are robust and offer improved accuracy, which should be considered for algorithmic selection strategies in similar settings.

7 REFERENCES

Benter, W. (1994). Computer-based horse race handicapping and wagering systems: a report. *Efficiency of Racetrack Betting Markets* 183–198.

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

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

Esbensen, K. H. (2002). *Multivariate Data Analysis – In Practice: An Introduction to Multivariate Data Analysis and Experimental Design*, 5th Edn. Oslo, Norway: Camo Process AS.

Frank, I. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 35(2), 109–148.

Geladi, P., and Kowalski, B. R. (1986). Partial least squares regression: A tutorial. *Analytica Chimica Acta* 185, 1–17.

James, W., and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* 1, 361–379.

Kelly Jr, J. R. (1956). A New interpretation of information rate. *Bell System Technical Journal* 35, 917–926.

Lessmann, S., Sung, M. S., and Johnson, J. E. V. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. *International Journal of Forecasting* 26(2), 518–536.

Silverman, N., and Suchard, M. (2013). Predicting horse race winners through a regularized conditional logistic regression with frailty. *Journal of Prediction Markets* 7(1), 43–52.

Sobel, R. S., and Raines, S. T. (2003). An examination of the empirical derivatives of the favourite-longshot bias in racetrack betting. *Applied Economics* 35(4), 371–385.

Sundberg, R., Brown, P. J., Martens, H., Næs, T., Oman, S. D., and Wold, S. (1999). Multivariate calibration: Direct and indirect regression methodology. *Scandinavian Journal of Statistics* 26(2), 161–207.

Wold, H. (1966). "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis,* eds P. R. Krishnaiaah *(*New York Academic Press).

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58(2), 109–130.

APPENDIX

Factors used in the analysis of racing data used for the logistic regression and PLS models.

- Number of starts
- Days since last run
- Days since second last run
- Days since third last run
- Win rate (% wins by starts)
- Place rate (% places by starts)
- Log transform of starting price in previous race
- Number of times won at current track
- Differential rating between grades for this race against previous race
- Weight difference between this race and previous race
- Distance difference between this race and previous race
- Number of times won at current Grade
- Finishing speed in previous 5 races (averaged)
- Horse’s rating in previous 5 races (averaged)
- Current jockey rating computed as % wins in previous 12 months
- Jockey rating differential between this race and previous race
- Average rating in first up runs, if last run ≥60 days
- Number of times horse has competed against current class
- Highest rating over the last 3 races
- Highest rating over the last 8 races
- Expected position in race (adjusted for distance, turns, barrier)
- Adjustment factor for wide barriers over track and distance
- Quick back up (days since last run (DSLR) if DSLR≤10)
- Horse age if ≤3 years old competing in an open age race
- Average pace over last 200m (previous 8 races)
- Odds (price) listed at start of racing day