Data Mining the Harness Track and Predicting Outcomes

Robert P. Schumaker
Management Information Systems
Central Connecticut State University, New Britain, Connecticut 06050, USA
rob.schumaker@gmail.com

Abstract

We present the S&C Racing system that uses Support Vector Regression (SVR) to predict harness race finishes and analyzed it on fifteen months of data from Northfield Park. We found that our system outperforms the most common betting strategies of wagering on the favorites and the mathematical arbitrage Dr. Z system in five of the seven wager types tested. This work would suggest that an informational inequality exists within the harness racing market that is not apparent to domain experts.

Introduction

Racing, like many other domains including the stock market, trades on publicly available information (i.e.; racing histories). Therefore all participants behaving in a rational manner should have an equal chance of success. However, information markets have inequalities through withheld information, a human tendency to discount certain information or weighting it incorrectly. These informational inequalities lead to arbitrage opportunities that can be unlocked using data mining techniques.

Literature Review

Predictive algorithms have been adopted successfully in racing sports, such as greyhound and thoroughbred racing. These systems use machine learning techniques to train the system on historical data then make predictions on previously unseen data. Highlights of several studies are presented below.

The first is a study of greyhound races using ID3 (a decision-tree algorithm) and Back Propagation Neural Network (BPNN) on 100 races at Tucson Greyhound Park [1]. The authors limited themselves to ten race-related variables over a seven race history upon advice from greyhound domain experts: fastest time, win, place and show percentage, break average, finish average, the average finish

time over the last three and seven races respectively, competitive grade of the race and a modifier if the horse is competing is a less competitive race. Their system made binary win/no-win decisions for each greyhound based on their historic race data. If a dog was predicted to win the system would make a $2 wager. The ID3 decision tree was accurate 34% of the time with a $69.20 payout ($0.69 excess return per dollar wagered) while the BPNN was 20% accurate with a $124.80 payout ($1.25 excess return per dollar wagered). This disparity in decreased accuracy and increased payout is justified by arguing that the BPNN was selecting longshot winners. As a result accuracy decreases and higher payouts are gained from the longer odds. When comparing their machine learning techniques to track experts, the experts managed a lower 18% accuracy and a payout loss of $67.60 ($0.68 excess loss per dollar wagered). It was speculated that Chen's machine learning system was taking advantage of information inequalities by successfully predicting longshot wagers more often than chance, however, given the black-box nature of BPNN and the difficulties of interrogating its decision-making process, it is hard to be certain.

In a follow-up study that expanded the number of input variables to 18, Johansson and Sonstrod used a BPNN on 100 races at Gulf Greyhound Park and found 24.9% accuracy for Win with a $6.60 payout loss ($0.07 excess loss per dollar wagered) [2]. The improvement in accuracy had a corresponding decrease in payout and implies that either the additional variables or too few training cases (449 as compared to Chen's 1,600) confounded their ability to identify longshots. However, the exotic wagers performed better; Quiniela had 8.8% accuracy and $20.30 payout ($0.20 excess return per dollar wagered), while Exacta had 6.1% accuracy and $114.10 payout ($1.14 excess return per dollar wagered).

In a third study that focused on using discrete numeric prediction rather than binary assignment, Schumaker and Johnson used Support Vector Regression (SVR) on the same 10 performance-related variables as Chen et. al. [3]. Their study of 1,953 greyhound races employed selective wagering where only the races with the predicted strongest competitors were wagered upon. Of the 505 races selected, their system produced a $0.95 excess return per dollar wagered for Win.

In looking at prior research, we discovered a lack of study of machine learners versus the wisdom of crowds. Several studies offered insight between crowds and experts, but none could be found that

explored how well a machine learning platform could perform versus crowd wisdom. From our analysis we propose the following research question: C*an a Machine Learner predict Harness races better than established prediction methods?*

Crowdsourcing and Dr. Z methods have been well established within the racing domain. However, both of these methods are susceptible to human biases and risk avoidance tendencies. Machine learning is devoid of these human characteristics and should be able to outperform the established prediction methods in a bias-free decision-making environment.

System Design

The S&C Racing system consists of several major components: a web scraper to gather online odds and race history from race programs, the machine learning algorithm that learns the patterns from historical data, a betting engine to make different wagers and evaluations to measure system performance. For odds data, harness track odds are pari-mutuel where the track sets the odds to balance the amount of money transfer from losing to winning wagers, minus a commission. Thus if a particular horse is the favorite and is heavily bet upon, the odds are decreased, which decreases payout. To offset favorite betting, the track will increase odds on less favored horses to give bettors an incentive to wager on longshots. Odds are made for each wager type. The Win wager is where the bettor receives a payout only if the selected horse comes in first place. Place produces two differing payouts depending upon whether the selected horse comes in either first or second place. A Show bet has three differing payouts that depend on whether the selected horse comes in first, second or third place. For exotic wagers, an Exacta bet receives a payout by successfully picking both the first and second place horse. A Quiniela wager is like an Exacta except the order of finish does not matter, only that the selected two horses finish within the top two spots. Trifecta, or Trifecta Straight, is where the bettor wagers on the first three horses, in order. For a Trifecta Box wager, the bettor still wagers on the first three horses, but their order does not matter as long as all three finish within the top 3 spots.

The other critical system input is historical performance data for each horse. There are generally 14 races per day where each race averages 8 or 9 entries. Race-specific information includes gait (pacing versus trotting), race date, track, fastest time, break position, quarter-mile position, stretch position, finish position, lengths won or lost by, average run time and track condition.

Once trained on historical data, the system is tested with different wagers and results are evaluated for betting efficiency, (payout – wager) / number of bets.

Experimental Design

For our collection we automatically gathered data from Northfield Park between October 1, 2009 and December 31, 2010 and divided the first twelve months into the training set and the remaining three months into testing. In all, we gathered 5,777 useable training cases covering 698 races and tested our system on 194 testing races covering 1,653 testing cases which is comparable to prior studies [1, 2 and 3].

Using the work of Chen et. al. (1994) as a template, we limited ourselves to the following eight variables over a four race history: fastest time, win, place and show percentage, break average, finish average, and the average finish time over the last three and four races respectively.

For our machine learning component we used Support Vector Machines (SVM). SVM is a machine learning classifier that attempts to maximally separate the classes by computing a hyperplane equidistant from the edges of each class [4]. Aside from dividing classes, the hyperplane can also be used as a regression estimate where independent variables are projected on to the hyperplane in order to derive the dependent variable. This variation of SVM is called Support Vector Regression (SVR) which allows for a continuous numeric prediction instead of classification. This allows us the freedom to rank predicted finishes rather than perform a win/lose classification.

Because S&C Racing employs the same selective wagering engine as Schumaker and Johnson (2008) we briefly explain how the system choses which races to wager upon. Once the training data has passed through the SVR algorithm and a model of predicted behavior established, we then run the predictions on the same training data for each horse in the race. The predicted values in theory are

continuous numbers between 1 and 8 corresponding with the estimated finish for each horse. We then rank order the set for a particular race and focus on the strongest horse (the one with the lowest predicted finish). We then perform a sensitivity analysis for each wager type in the training set by simulating wagers on those animals with a predicted finish less than our sensitivity cutoff value and vary the sensitivity cutoff from 1 to 8 in increments of 0.1. This builds a sensitivity map for the particular wager as shown in Figure 1. Once complete we identify the maximum value in our training sensitivity set and hold that sensitivity cutoff value. We then use the same prediction model on the testing set and build a sensitivity map as well. We then use the held sensitivity cutoff value from the training set on the testing set to establish the excess returns per dollar wagered for that wager type.

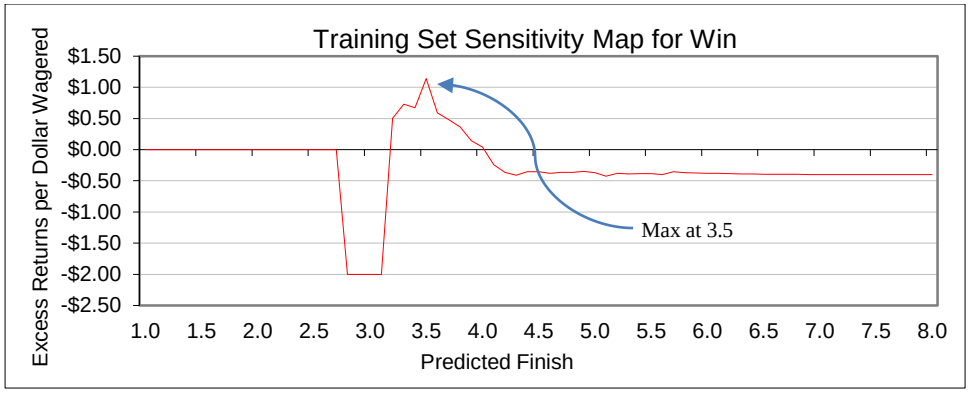

Figure 1. Training Set Sensitivity Map for Win wager

In order to best compare the S&C Racing results, we compare them against established prediction methods of crowdsourcing, Dr. Z bettors and random chance. For the crowdsourcing comparison we use pre-race odds where the crowd favorite, the animal with the lowest odds, was selected and wagered upon. For Dr. Z bettors, Place wagers were made on animals with win percentage to place percentage ratios greater than or equal to 1.15, and Show wagers were made with win percentage to show percentage ratios greater than or equal to 1.15. For random chance, we calculate the statistical odds of selecting an animal to wager upon given the number of race participants and use those excess returns.

Experimental Findings and Discussion

To answer our research question we analyze each wager in terms of betting efficiency in Table 1.

	Win	Place	Show	Exacta	Quiniela	Trifecta	Trifecta Box
S&C Racing	$1.08	$2.30	$2.55	$11.70	$4.23	-$1.24	-$4.23
Crowdsourcing	$0.68	$1.29	$1.42	$0.72	$1.02	$1.15	-$0.75
Dr. Z Bettors		$0.00	$0.06				
Random Chance	-$0.49	-$0.20	$0.25	-$4.50	-$4.93	-$4.84	-$11.19

Table 1. Comparing Betting Efficiency

From the data, S&C Racing outperformed Crowdsourcing in five of the seven wagers; Win, Place, Show, Exacta and Quiniela (p-values < 0.01). Our system also outperformed all the Dr. Z Bettors (p-values < 0.01) and all the random chance wagers (p-values < 0.01). While it could be argued that S&C Racing is wagering on the strongest races, Dr. Z is similarly selective and Crowdsourcing should be just as strong assuming rational bettors with the same access to information that S&C Racing uses. However, the discrepancies would appear to indicate that S&C Racing is exploiting an informational inequality within the harness racing market not apparent to other bettors. The two wagers with the highest returns, Exacta at $11.70 and Quiniela at $4.23 were performing in excess of crowd wisdom. We speculate that the inherent difficulty for crowds to correctly pick the combination winners may have contributed to this result which would in effect raise the odds and the corresponding payout. However, the same did not hold true for Trifecta and Trifecta Box, where we found that the training set peaked too early versus the testing set which led to losses, whereas the other wagers types were fairly uniform with little discrepancy between the sets allowing for more accurate prediction.

Conclusions

Comparing our system against established betting strategies we found that S&C Racing outperformed both random chance and Dr. Z Bettors in both accuracy and payouts. When compared against Crowdsourcing, S&C Racing fared well able to identify top contenders better than the crowds were.

References

1. Chen, H., et al., 1994. Expert Prediction, Symbolic Learning, and Neural Networks: An Experiment on Greyhound Racing. *IEEE Expert*, 9(6):21-27.

2. Johansson, U. and C. Sonstrod, 2003. Neural Networks Mine for Gold at the Greyhound Track, in *International Joint Conference on Neural Networks*: Portland, OR.

3. Schumaker, R.P. and J.W. Johnson, 2008. Using SVM Regression to Predict Greyhound Races, in *International Information Management Association (IIMA) Conference*: San Diego, CA.

4. Vapnik, V., 1995. *The Nature of Statistical Learning Theory* New York: Springer.