Journal of International Technology and Information Management

Volume 27 | Issue 2 Article 1

12-1-2018

Machine Learning the Harness Track: A Temporal Investigation of Race History on Prediction

Robert P. Schumaker
University of Texas at Tyler, rob.schumaker@gmail.com

Machine Learning the Harness Track: A Temporal Investigation of Race History on Prediction

Robert P. Schumaker
Computer Science Department
University of Texas at Tyler, Tyler, Texas 75799, USA
rob.schumaker@gmail.com

ABSTRACT

Machine learning techniques have shown their usefulness in accurately predicting greyhound races. Many of the studies within this domain focus on two things; win-only wagers and using a very particular combination of race history. Our study investigates altering these properties and studying the results. In particular we found a race history combination that optimizes our S&C Racing system's predictions on seven different wager types. From this, S&C Racing posted an impressive 50.44% accuracy in selecting winning wagers with a payout of $609.34 and a betting return of $10.06 per dollar wagered.

KEYWORDS: machine learning, support vector regression, data mining, harness racing

INTRODUCTION

Within the domain of racing, the ability to make accurate predictions has attracted gamblers and academics alike. Even in situations where accurate forecasts are possible, it is entirely possible to focus on unimportant aspects which can lead to crippled systems relying on unimportant data or worse, not based on sound science (e.g., basing predictions on the color of a horse as a performance measurement).

Before making a wager, a bettor will typically gather as much information about the horses as possible which may include data concerning a horse's physical condition and how they have performed historically, their breeding and bloodlines, who is their trainer or owner, and odds of winning. Automating this decision process using machine learning may yield as equally predictable results as greyhound racing, which is considered to be the most consistent and predictable form of racing (Chen, Rinde et al., 1994). Consistency lends itself well to machine learning algorithms that can learn patterns from historical data and apply itself to

  ISSN: 1941-6679-On-line Copy

previously unseen racing instances. The mined data patterns then become a type of arbitrage opportunity where an informational inequality exists within the market. However, like other market arbitrages, the more it is exploited the less the expected returns, until the informational inequality returns the market to parity.

Our research motivation is to create and test a machine learning technique that can learn from historical harness race data and leverage a hidden arbitrage through its predictions. In particular, we will focus on varying the amount of race history used and its effect on differing wager types.

The rest of this paper is as follows. Section 2 provides an overview of literature concerning prediction techniques, algorithms and common study drawbacks. Section 3 presents our research questions. Section 4 introduces the S&C Racing system and explains the various components. Section 5 sets up the Experimental Design. Section 6 is the Experimental Findings and a discussion of their implications. Finally Section 7 delivers the conclusions and limitations of this stream of research.

LITERATURE REVIEW

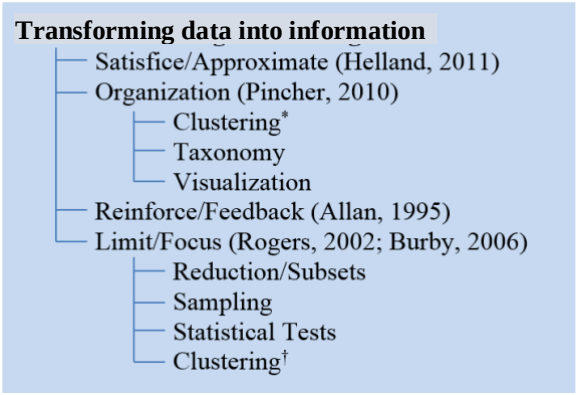
Harness racing can be thought of as a general class of racing problem, along with greyhound, thoroughbred, automotive and even human track competition. While each race subset enjoys its own unique aspects, all share a number of similarities in both format and goals. Participants behave independently of one another and are largely interchangeable. These similarities can lead to the successful porting of techniques from one race domain to another.

Converting Raw Data into Predictions

All of racing relies on data. A bettor takes available data and attempts to extract knowledge – predicted finishes – by using gut instinct or an algorithm. Automating this algorithmic process, the same steps can be applied to computer systems; feed data in, extract knowledge and predict finishes. The question then becomes how do we construct such a system to go from raw data to accurate predictions? The answer lays in a two-step transform, data to information and information to knowledge (Ackoff, 1989). Figure 1 demonstrates a taxonomy of data mining techniques for converting data into information.

ISSN: 1941-6679-On-line Copy

Figure 1. Taxonomy of techniques to transform data into information



[1][2]

From this figure, data conversion can fall into one of four major areas: satisfice/approximate, organization, reinforce/feedback and limit/focus. In Satisfice/Approximate, Helland argues that sometimes good enough is good enough and the level of precision should be tempered by the information needed (Helland, 2011). This definition can be traced back to Herb Simon whom argued that this decision-making strategy was preferred for finding adequate solutions amongst incomplete data or limited resources.

Another technique of data transformation is Organization. We can partition organizational techniques into spacial clustering, where the distance between cluster centers provides information about the level of relation between clusters; taxonomy, where the organization of data into a hierarchy provides information, and visualization techniques, where data is condensed into a visual depiction and the distances, shape, color and composition of data becomes information (Pincher, 2010).

A third technique is that of Reinforce/Feedback. With this technique, data is fed back through the system in the attempt to isolate weak relations and hence information (Allan, 1995).

Fourth is to Limit/Focus the data to sift out information. There are several sub-techniques such as reduction/subsets, where similar data is isolated or aggregated to form information (Rogers, 2002; Burby, 2006), sampling from a larger pool of data, statistical tests and clustering for data reduction such as topic classification.

---

1 Clustering for Euclidean distances
2 Clustering for data reduction

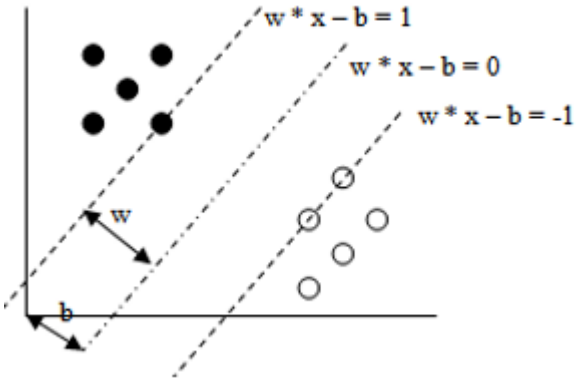
While many techniques are possible in converting data into information and most are simplistic in nature, the process of converting information into knowledge is more complex. With respect to data mining, this transformation can take one of three distinct paths; simulation, artificial intelligence and machine learning (Schumaker and Johnson, 2008). In Simulations, similar data is constructed to test various parameters. Applied to thoroughbred racing, simulations have been used to test theoretical sire/dame offspring combinations to determine the most potent racing colt (Burns, Enns et al., 2006). In Basketball, simulations can determine optimum player substitution patterns (BBall, 2008). However, simulations do not address the complexities amongst a large number of parameters.

In Artificial Intelligence, computers attempt to find solutions by applying iterative codified rules or cases. Heuristic solutions may not be perfect, however, the solutions generated are considered adequate (Schumaker, Solieman et al., 2010).

In Machine Learning, a system attempts to identify unknown patterns to add to the understanding of the dataset (Chen and Chau, 2004; DataSoftSystems, 2009). Examples of algorithms include both supervised and unsupervised learning techniques, such as genetic algorithms, neural networks and Bayesian probability. Machine learning systems are considered to be better able to generalize data into usable patterns (Lazar, 2004).

One of the better suited machine learning algorithms for sports data mining is the regression-based variant of the Support Vector Machine (SVM) classifier, called Support Vector Regression (SVR) (Vapnik, 1995). SVM is a classification algorithm that seeks to maximally separate high dimension data while minimizing fitting error as shown in Figure 2.

Figure 2. Support Vector Machine (SVM)



It does so by calculating a multi-dimensional hyperplane of n-1 dimensions, optimizing distances between the different classes. SVR differs from SVM by using the hyperplane as a regression estimator to return discrete values rather than classes. The SVR technique was used in a similar context to predict stock prices from financial news articles (Schumaker and Chen, 2008), greyhound (Schumaker and Johnson, 2008) and harness racing (Schumaker, 2013).

Relevant Prior Studies

To lay the groundwork of a majority of prior machine learning racing studies, we first need to discuss Chen et. al. (1994) whose contributions still reverberate through many follow-up studies. In a study of greyhound races, Chen et. al. tested an ID3 and Back Propagation Neural Network (BPNN) on ten race performance-related variables as determined by human domain experts, on 100 races at Tucson Greyhound Park (Chen, Rinde et al., 1994). These ten variables include:

Fastest Time – time difference between the subject and the winner in the last competed race

Win Percentage
Place Percentage
Show Percentage
Break Average – average position coming out of the gate
Finish Average

over the past seven races

Time7 – average race time over the last seven races

Time3 – average race time over the last three races

Grade Average – race grade or competitiveness of the field

UpGrade – assigns points based on race downgrades

From their work, the system made binary win/lose decisions on each greyhound, independent of the other race participants. If a greyhound was predicted to finish first, the system would make a $2 wager. The ID3 decision tree that they used was accurate 34% of the time with a $69.20 payout while BPNN was 20% accurate with a $124.80 payout. This disparity between accuracy and payout is justified by arguing that the BPNN was selecting longshot winners. By doing so, higher payouts were gained at the expense of accuracy because of the higher odds.

With these ten variables, two of them, GradeAverage and UpGrade, are specific to greyhound racing and have no equivalence in harness racing. From the remaining eight variables, only one, Fastest Time, is not dependent upon an arbitrary amount

of race history. The percentages of Win, Place and Show plus the Break and finish averages all depend on a seven race history. Time7 and Time3 depend on a seven and three race history respectively. The human domain experts chose these variables based on their experience.

In a related study, Johansson and Sonstrod expanded the number of variables studied from 10 to 18 and also used a BPNN (Johansson and Sonstrod, 2003). Their study of 100 races at Gulf Greyhound Park in Texas found 24.9% accuracy for Wins and a $6.60 payout loss. This seemingly improved accuracy came at the cost of decreased payout and would imply that either the additional variables or too few training cases hampered the ability to identify longshots.

In a third study that focused on using discrete numeric prediction rather than a binary (win/loss) assignment, Schumaker and Johnson used Support Vector Regregression (SVR) on the same 10 performance-related variables from Chen's study (Schumaker and Johnson, 2008). Their study of 1,953 greyhound races across the United States demonstrated a 45.35% Win accuracy with a $75.20 payout. To maximize payout, they had 23.00% Win accuracy with a $1,248.40 payout. They found the same tradeoff between accuracy and payout as Chen's work.

In a fourth study that examined crowdsourcing on harness race wagering, Schumaker varied the primary race variable, Time7, to maximize its impact on system accuracy and payout (Schumaker, 2013). From this study, a four-race history was found to have better accuracy than a seven-race history. In terms of wagering payout the Win and Place wagers lost money while Show, Exacta, Quiniela, Trifecta and Trifecta Box had positive returns. This study also required apriori knowledge of racing events (e.g., identified maximized values within the entire dataset and not just training) and was not systematic in treatment of both primary and secondary time variables.

Common Study Drawbacks

Much of the prior work used 10 race performance variables derived from greyhound track experts. These variables include a precise combination of race history to use; namely the primary race history variable uses the most recent seven races and the secondary race history variable uses the most recent three races. While this expert-derived race history may be appropriate to greyhound racing, none of the prior literature explored varying both history variables to optimize race prediction.

ISSN: 1941-6679-On-line Copy

Second, the one harness racing study that did evaluate varying primary race history, was not thorough enough and maximized vales within the entire dataset. For the purposes of robustness, we seek to break the dataset into distinct partitions; maximize the values in one set and observe the results in the other. This treatment will not only lead to better robustness, but also be better at generalizing our observations.

RESEARCH QUESTIONS

From our analysis we propose the following research questions.

- *What is the impact of race history on various wagers?*

Previous studies have all relied on using a Time 7-3 setup where the primary race history uses the prior seven races and the secondary race history uses the most recent three races and a win-only wagering system. Subsequent studies used this race history and wagering combination without question. We plan to investigate the manipulation of primary and secondary race history with respect to different wager types and study its effect on race prediction. In particular we are interested in how the different race histories and wagers perform and whether any patterns or similarities exist between them.

- *What is the optimal amount of race history for a machine learning system?*

Following up on the previous research question, we would like to further investigate whether Time 7-3 is optimal for the harness track, or whether other race history combinations might prove better. We suspect that if Time 7-3 is optimum it is for greyhound racing and possibly not for harness racing.

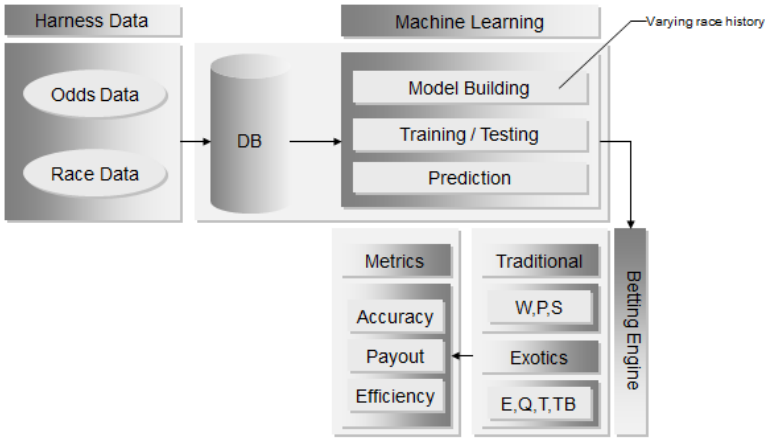
- *What wager combinations work best and why?*

Most prior studies only examined Win wagers. We plan to expand this to other traditional and exotic wager types. By looking at performance data between several potential models, the results should provide some predictive clarity.

SYSTEM DESIGN

To address these research questions, we built the S&C Racing system shown in figure 3.

Figure 3. The S&C Racing System



The S&C Racing system consists of several major components: the data gathering module, the machine learning aspect, a rudimentary betting engine and evaluation metrics. Odds data is composed of the individual race odds for each wager type (e.g., Win, Place and Show). Race data is then gathered from a race program.

Each race program contains a wealth of data. There are generally 14 races per program where each race averages 8 or 9 entries. Each horse has specific data such as name, driver and trainer. Race-specific information includes the gait, 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.

Models are then built depending upon the amount of race history to be tested. Once the system has been trained on the data provided, the results are tested along three dimensions of evaluation: accuracy, payout and efficiency. Accuracy is the number of winning bets divided by the number of bets made. Payout is the monetary gain or loss derived from the wager. Efficiency is the payout divided by the number of bets.

The Betting Engine examines seven different types of wagers: Win, Place, Show for the traditional wagers and Exacta, Quinella, Trifecta and Trifecta Box for the

ISSN: 1941-6679-On-line Copy

exotic wagers. If betting on a Win, the bettor receives a payout only if the selected horse comes in first place. If betting on Place, the bettor receives differing payouts if the selected horse comes in either first or second place. If betting on Show, the bettor receives differing payouts if the selected horse comes in first, second or third place. These differing payouts are dependent upon the odds of each finish. Wagers on Exacta mean that the bettor is predicting the first two horses to cross the finish line in order. Quinella is a similar two horse wager except the order of finish does not matter, as long as the predicted two horses finish within the top two positions. Trifecta is a three horse wager in order whereas Trifecta Box is picking the first three horses in any order.

EXPERIMENTAL DESIGN

To perform our experiment, we gathered data from Northfield Park; a USTA sanctioned harness track outside of Cleveland, Ohio. After data has been gathered, it is parsed for specific race variables before it is sent to S&C Racing for prediction.

For our collection we chose a study period of October 1, 2009 through December 31, 2010. The data was partitioned into a twelve month training set (October 1, 2009 – September 30, 2010) and a three month testing set (October 1, 2010 – December 31, 2010). Prior studies focused on only one racetrack, manually input their data, used 10-fold cross-validation rather than separate training/testing sets and had small datasets. Chen et. al. (1994) used 1600 training cases from Tucson Greyhound Park, Johansson and Sonstrod (2003) used 449 training cases from Gulf Greyhound Park in Texas and Schumaker and Johnson used 41,473 training cases from across the US. Our study is comparable with between 1,136 and 14,503 training cases depending upon the model and amount of race history that model requires.

From the data, we built 55 models, varying the primary and secondary race history variables between 1 (uses only the variables from the most recent race) to 10 (uses the most recent 10 races); hence race histories take the form of Time 1-1 for one-race primary and secondary history, to Time 10-10 for ten-race primary and secondary history respectively. Since the primary race history is always greater than or equal to the secondary race history, it dictates the amount of data used. The reason we chose the primary/secondary approach was to maintain consistency with prior studies. Table 1 illustrates the amount of training/testing races/cases for each of the ten primary race histories.

Table 1. Training/Testing Races/Cases for the ten primary race histories

	Time 1-x	Time 2-x	Time 3-x	Time 4
Training Races	1,744	1,223	914	
Training Cases	14,503	10,154	7,568	5,
Testing Races	519	370	263	

We chose to have differing training/testing cases between the primary race history models rather than a stable set. This is because our focus is on the betting engine to select which races to wager upon and reflects how the system would be implemented under real-world conditions – where a one-race history is easier to obtain than a ten-race history. This will be further explained in detail shortly.

For the Time 10-x models, we gathered 1,136 useable training cases for 138 races. The reason for so few usable races over a year's time is because we adopted a stringent requirement that *every horse* within a useable race needs to have that minimum amount of race history. In this model's case, we required a ten-race history for each horse. Since new entries would arrive in the Northfield market frequently and consequently would lack a ten-race history, only 138 races could meet this requirement.

Using the work of Chen et. al. (1994) as a guide, we built our models using the following eight variables: Fastest Time, Win Percentage, Place Percentage, Show Percentage, Break Average, Finish Average, Average Time of the Primary race history and Average Time of the Secondary race history. Two variables from Chen et. al. (1994) were not applicable to harness racing and not used; Grade Average and Upgrade. Both of these variables refer to the competitiveness of the race and have no equivalent in Harness racing.

Because of the system complexity with the different models, usage of training and testing data, sensitivity analysis and wager optimization, we present the following pseudo-code and describe the key aspects afterwards.

ISSN: 1941-6679-On-line Copy

For each Primary race history (x), iterate from 1 to 10
  For each Secondary race history (y), iterate from 1 to x
    Build $model_{Training,x,y}$ on Training set using only those races in which every horse has an x race history

    Run $model_{Training,x,y}$ through the SVR algorithm to construct a high dimension mathematical $equation_{x,y}$ to predict finish order

    Implement $equation_{x,y}$ against $model_{Training,x,y}$ in preparation for a Sensitivity Analysis

    For each $wager_{x,y}$ (Win, Place, Show, Exacta, Quiniela, Trifecta and Trifecta Box)
      For each metric (Accuracy, Payout, Betting Efficiency)
        Construct a Sensitivity Analysis on $model_{Training,x,y}$
        For Cutoff = 1.0 to 8.0, increment by 0.1
          Wager only on those races where the lowest Predicted Finish ≤ Cutoff
          Tabulate $\sum value_{x,y,wager,metric,Cutoff}$
        If NumRaces ≥ 30, $Cutoff_{x,y,wager,metric}$ = max($value_{x,y,wager,metric}$)

    Build $model_{Testing,x,y}$ on Testing set using only those races in which every horse has an x race history

    Implement $equation_{x,y}$ against $model_{Testing,x,y}$

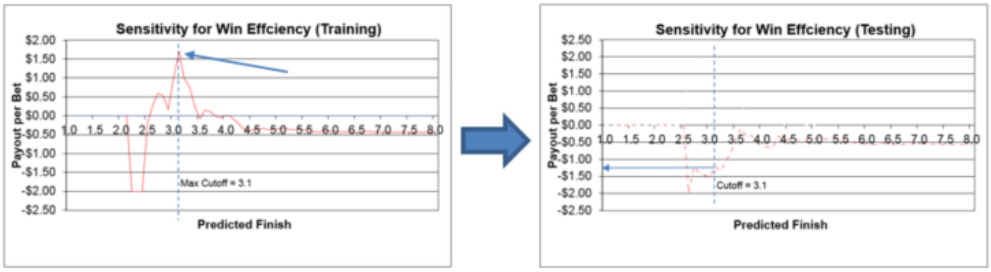
    For each $wager_{x,y}$ (Win, Place, Show, Exacta, Quiniela, Trifecta and Trifecta Box)
      For each metric (Accuracy, Payout, Betting Efficiency)
        For Cutoff = 1.0 to 8.0, increment by 0.1
          Tabulate $\sum value_{x,y,wager,metric,Cutoff}$
        Determine Testing $value_{x,y,wager,metric}$ using $Cutoff_{x,y,wager,metric}$ from earlier

While complex, the sensitivity analysis allows us to interrogate the data to optimize wagering, rather than treating it as a blackbox as previous studies have done. We increment a Cutoff value from 1 (wagering on races where the lowest predicted finish value is ≤ 1[3]) to 8 (wagering on races where the lowest predicted finish value is ≤ 8[4]), incrementing the Cutoff value by 0.1. Our first aim is to maximize the metrics in the Training data for each wager type. As an example using Win and betting efficiency, we identify the Cutoff associated with the maximized betting efficiency value. Then using that Cutoff we turn to the Testing set and look up the betting efficiency for that particular Cutoff value as shown in Figure 4.

[3] While unlikely, the possibility for this situation exists depending upon the inputs given to the system.

[4] In practicality this wagers on all races.

Figure 4. Finding and Using Cutoff values between Training and Testing sets



The premise is that if a hidden pattern is within the data, then a system should be able to identify it in the training set and also observe it in the testing set. In other words, the pattern discovered could be successfully arbitraged for the purposes of improved accuracy, abnormal positive payouts or increased betting efficiency.

As an example of how the system works, each horse in each race is given a predicted finish position by the SVR algorithm. Looking at *Miss HKB* for the Time 7-3 model, we compute the variables for the prior seven and three races as shown in Table 2, send them to S&C Racing's SVR algorithm and receive the predicted finish (2.93) from SVR.

Table 2. *Miss HKB* variable data

Fastest Time	119.56
Win Percentage	14.29%
Place Percentage	42.86%
Show Percentage	14.29%
Break Average	5.14
Finish Average	2.71
Time7 Average	120.19
Time3 Average	119.90
Predicted Finish	**2.93**

For this particular race, Race 4 on October 8, 2010, S&C Racing predicts that *Miss HKB* should finish 2.93 which is a good finish, but cannot be fully interpreted until compared with the predicted finishes of other horses in the race. The lower the predicted finish number, the stronger the horse is expected to be and the predicted finish value is independent of the other horses in the race. For context, Table 3 shows the race output for Northfield Park's Race 4 on October 8, 2010.

ISSN: 1941-6679-On-line Copy

Table 3. Predicted Values for Race 4 on October 8, 2010

Horse Name	Predicted Finish
Miss HKB	2.93
B B Big Girl	3.90
Friendly Kathy	4.30
ShadyPlace	5.27
St Jated Strike	5.72
Honey Creek Abby	5.78
Mad Cap	5.93
Pamela Lou	6.51
Winning Yankee	6.86

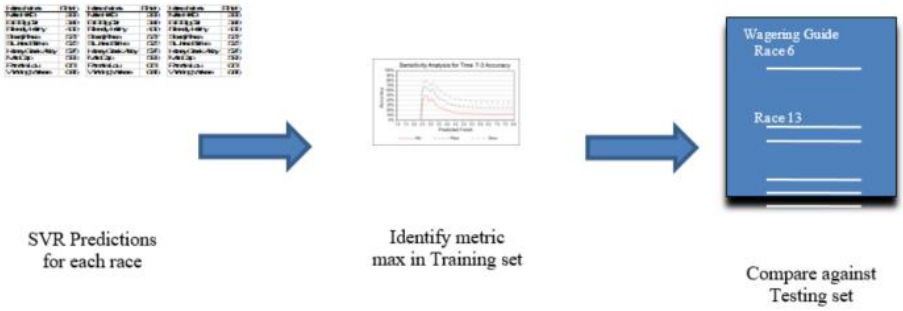
From this table, we can establish a rank-order of expected finishes. The information at this stage will give us wagering opportunities. However, S&C Racing goes a step further in isolating maximized metrics and then selectively wagering on the strongest races. This is where the knowledge component comes in with the sensitivity analysis derived cutoffs.

For each time combination and metric, we identify the maximum value in the training set and save the corresponding Cutoff value. In this case we take Cutoff value of 3.1 for Win,[5] 3.1 for Place and 3.1 for Show. We then take that Cutoff value to the testing set and retrieve the corresponding metric. Which in this case is 16.67% accuracy for Win, 58.33% for Place and 75.0% for Show.

So by identifying the Cutoffs in the training set representing the maximum metric across all time combinations, we can tune the S&C Racing system to focus on specific time combinations for each wager type. This means that not every race will be wagered upon. In essence, S&C Racing is given the ability to be selective in choosing races, as shown in Figure 5.

[5] While Cutoff 2.7 represents the maximum Win accuracy of 50.0%, it only represented 16 races; short of our 30 race limit.

Figure 5. S&C Racing chooses which races to wager upon



EXPERIMENTAL FINDINGS AND DISCUSSION

To answer our research questions we constructed the S&C Racing system. We first analyze the impact of race history on the seven wager types by analyzing the betting efficiency metric. Once we have narrowed down the best performing race history combinations we will constrain ourselves to a more detailed study of the impact it has on the different wager types, which answers our second research question. For the third research question we open up the constraints and look towards maximizing accuracy and wagering payout for specific time combinations.

We present Table 4 that looks at the averaged betting efficiencies of all seven wagers with respect to differing primary and secondary race histories.

Table 4. Averaged Betting Efficiencies across all Seven Wagers

Efficiency	Time1	Time2	Time3	Time4
Time1	$0.01	-$0.03	$0.01	$0.15
Time2		-$0.02	-$0.07	$0.02

From this table, Time 4-3 had the best return per wager at $1.27 followed by Time 8-5 at $0.58 and Time 8-7 at $0.27. Compared to the betting efficiency of Time 7-3 (which has been the defacto standard for race combination history in prior results)

at -$0.11, it would appear that different time combinations do work better with harness racing.

To determine the effect of race history on prediction, we performed a single factor Anova on both the Primary and Secondary race histories as shown in Table 5a and 5b.

Table 5a. Anova of Primary Race Histories
Table 5b. Anova of Secondary Race Histories

Anova: Single Factor

by Columns (Primary Time variable)

SUMMARY

Groups	Sum	Average	Variance

Anova: Single Factor

by Rows (Secondary Time variable)

SUMMARY

Groups	Sum	Average	Variance

From Table 5a, we have an f-measure of 9.94 and a between groups p-value < 0.001 indicating statistically significant differences between primary race histories. From the descriptive stats, a four race primary history performed best with an average of $8.70. Comparing this result to the second best primary race history of two ($5.80) using a t-test, we found a statistically significant difference (p-value < 0.001). By comparison, the seven race history, which has been the de facto standard in several other race studies, had a meager $1.61 return. While a seven race primary history may be important in other racing domains such as greyhounds, it did not capture the essence of predicted harness performance. This leads us to believe that too much time elapses where the history starts to become irrelevant and actually harms the predicted results.

Table 5b looks at manipulating the secondary race histories. From this Anova, we had an f-measure of 0.72 and found no statistical difference between the groups. This means that varying the second time variable by itself does not make much difference, statistically.

If we instead concentrate on Time 4 as the primary time variable, Table 6 looks at the accuracy, payout and betting efficiency of each time Time 4-x combination.

Table 6. Time 4-x Accuracy, Payout and Efficiency

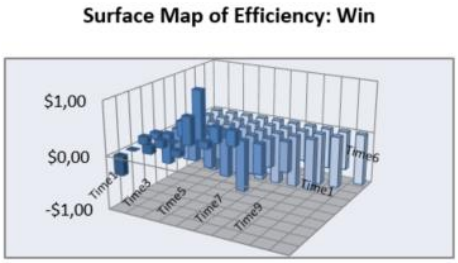
	Time4		
	Accuracy	Payout	Efficiency
Time1	50.81%	$508.17	$7.33
Time2	38.14%	$568.47	$8.44

From this table, Time 4-1 had the best accuracy at 50.81%, however, it was found to be statistically equivalent to Time 4-3 (50.44%). Comparing Time 4-3 to Time 4-2 in accuracy (50.44% to 38.14%) we did achieve statistical significance (p-value < 0.001). For Payout, Time 4-3 had the best Payout at $609.34 and was found to be statistically different from the second-best payout of Time 4-2 at $568.47 (p-value < 0.001). Efficiency of Time 4-3 was also the highest at $10.06 and was statistically different from its nearest competitor of Time 4-4 at $8.96 (p-value < 0.001).

From this analysis, it would appear that Time 4-3 maximizes the most metrics. For payout and efficiency Time 4-3 had the statistically superior values of all seven wager types. Time 7-3 did not fare so well with a 35.42% accuracy, $43.39 payout and $0.32 betting efficiency. While Time 7-3 may be appropriate to Greyhound racing, it is again clearly not optimal for harness. For accuracy, Time 4-3 and Time 4-1 were statistically equivalent in terms of average accuracy across all seven wager types. However, Time 4-1 payout and efficiency was not a good as that of Time 4-3.

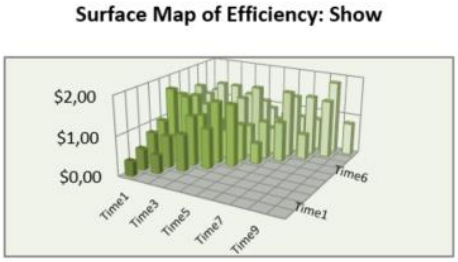
ISSN: 1941-6679-On-line Copy

Figure 6. Betting Efficiencies of Traditional Wagers



Panel 6a. Betting Efficiencies for Win

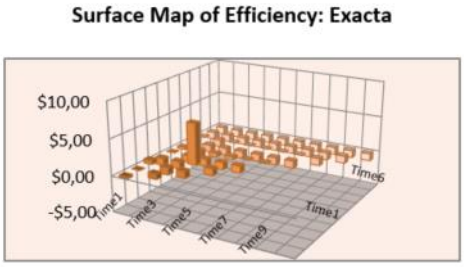
Panel 6b. Betting Efficiencies for Place



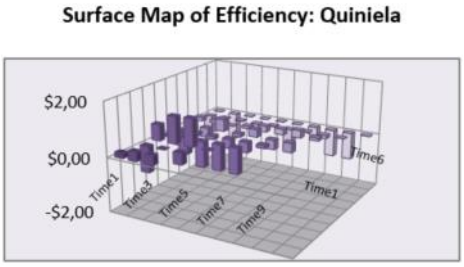
Panel 6c. Betting Efficiencies for Show

From this data, Win had a maximum betting efficiency of $0.95 at Time 5-3. From this maximum betting efficiency, the $0.95 represents the excess return per dollar wagered. Place had a maximum betting efficiency of $1.70 at Time 9-8. Place showed much more uniformity across the models as opposed to Win with an average return of -$0.57 and $0.50 standard deviation. Show had a maximum betting efficiency of $1.84 at Time 10-9 with an average excess return of $1.06 and $0.39 standard deviation. Given the much more uniform returns of the Show wager, it would seem that the amount of race history appears less important to Show wagers.

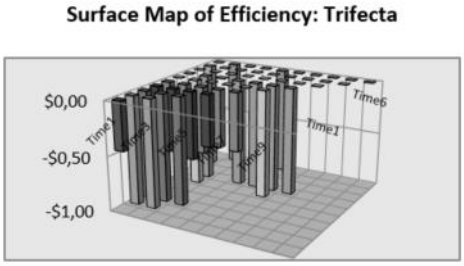
Figure 7. Betting Efficiencies of Exotic Wagers



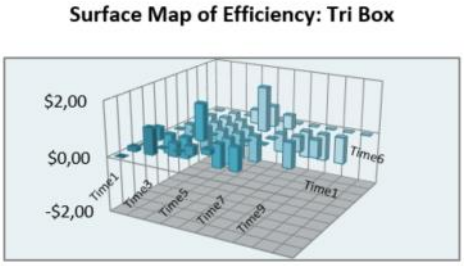
Panel 7a. Betting Efficiencies for Exacta



Panel 7b. Betting Efficiencies for Quiniela



Panel 7c. Betting Efficiencies for Trifecta



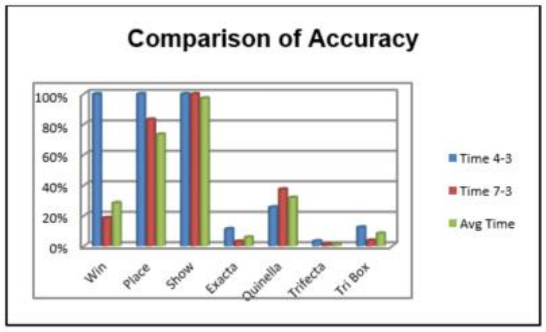
Panel 7d. Betting Efficiencies for Trifecta Box

Looking at the betting efficiencies of the exotic wagers in Figure 7, Exacta had a maximum betting efficiency of $5.85 at Time 4-3, an average excess return of -$0.61 with $0.99 standard deviation. Quiniela fared similarly well with $1.06 maximum betting efficiency at Time 4-3, and average excess return of -$0.16 with $0.47 standard deviation. The results for Exacta and Quiniela would indicate that a four race primary variable may be ideal for 2 horse wagering. Trifecta had no positive returns, average -$0.38 loss with $0.44 standard deviation. Trifecta Box maxed out at Time 8-5 with a $1.68 excess return, average -$0.31 with $0.65 standard deviation.

To answer our research question of *what wager combinations work best and why*, we compare Time 4-3 against the de facto standard of Time 7-3 and also an average time (which is the average of all 55 models) as shown in Figure 8.

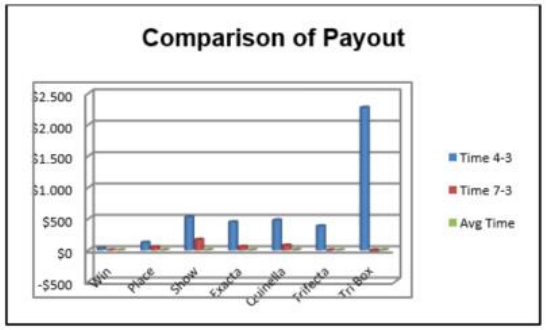
ISSN: 1941-6679-On-line Copy

Figure 8b. Comparing Payout across Wager Types and p-values



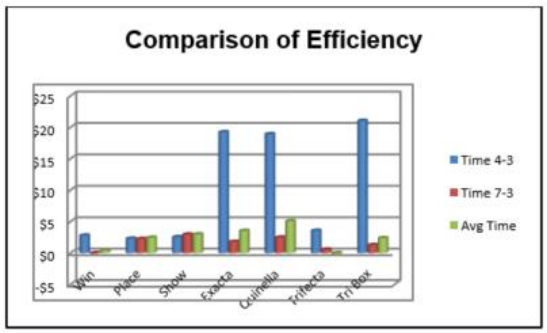
	4-3 vs 7-3	4-3 vs Avg	7-3 vs Avg
Win	0.283	0.303	**0.002**

Figure 8a. Comparing Accuracy across Wager Types and p-values



	4-3 vs 7-3	4-3 vs Avg	7-3 vs Avg
Win	<0.001	<0.001	0.153

Figure 8c. Comparing Efficiency across Wager Types and p-values



	4-3 vs 7-3	4-3 vs Avg	7-3 vs Avg
Win	0.106	0.126	<0.001

From Figure 8a, while it looks like Time 4-3 outperformed in many wager types, statistical significance could only be obtained in a few wager combinations. This had to do with the maximized accuracy coming at the expense of a lower number

of wagering opportunities. This lower number of wagering opportunities, with lower degrees of freedom, raises the threshold of statistical significance. In many cases, the comparisons were unable to exceed this threshold.

In Figure 8b, where the focus is on maximizing Payouts, a larger amount of wagering opportunities exist and from this we found that Time 4-3 had the highest Payouts in all seven wager types. It was also interesting to note that all comparisons had p-values < 0.001 except for the comparison of Time 7-3 and the Average Time with a p-value of 0.153.

From Figure 8c, all Time 4-3 wagers except Place and Show outperformed both Time 7-3 and the Average Time with statistical significance. For Place, the Average Time was highest at a $2.45 return compared to Time 4-3 at $2.30. For Show, the Average Time was the highest at a $2.98 return compared to Time 4-3 at $2.55.

Taking all of these results together, Time 4-3 was superior. Comparing it to both Time 7-3 and the Average Time, Time 4-3 appeared to have better accuracy results, but did not achieve a statistical significance in most cases due to a low number of wagering opportunities with the highest accuracy. Time 4-3 outperformed in Payouts in all seven wager types and outperformed in Betting Efficiency in five of the seven wagers. This indicates that a four race primary history coupled with a three race secondary history was best able to predict future finishes. Quantitatively, this could be considered the sweet spot of the amount of race history needed to optimize predictions.

CONCLUSIONS AND FUTURE DIRECTIONS

From our investigation we found that S&C Racing was able to predict harness races fairly well depending upon the wager desired and the amount of race history input to the system. When looking at the Betting Efficiency of wagers, both Exacta and Quinella were uniform in their returns whereas Trifecta and Trifecta Box were not. It would appear that picking the first two horses was easier for the S&C Racing system than picking the third. In looking at what amount of race history works best, Time 4-3 maximized the most metrics. In Payout and Betting Efficiency Time 4-3 was superior to Time 4-x. For Accuracy, Time 4-3 and Time 4-1 were both superior and statistically equivalent. When comparing Time 4-3 against the de facto standard of Time 7-3 and the Average Time, Time 4-3 again maximized most of the metrics. In Payout and Efficiency, Time 4-3 was superior in a majority of wagers.

ISSN: 1941-6679-On-line Copy

Future directions for this stream of research include an analysis of training frequency, fraud detection and expanding these techniques into other domains. For all of the prior race studies, this one included, a static model of prediction was built and tested. In the case of this paper, the model was built over one year worth of data and applied to a three month testing set. We feel that a potential research area includes determining how often the model should be refreshed. Would every day fresh be appropriate, coming from the information retrieval domain when dealing with critical data, or would the model still be valid for a certain period of time, thus decreasing the amount of computation necessary. Another potential research area is to build a fraud detection framework. Now that a prediction model can be built, we have the potential to look for a pattern of outlier data that may indicate either an undisclosed injury or race misconduct occurring. A third potential research area includes expanding these techniques to other domains such as thoroughbred, track and field and Nascar racing. The techniques used here could be a baseline for further investigation of other racing domains.

REFERENCES

Ackoff, R., (1989). From Data to Wisdom. *Journal of Applied Systems Analysis* 16: 3-9.

Allan, J., (1995). Relevance Feedback with too Much Data, *ACM SIGIR*, Seattle, WA.

Bball, (2008). Online Interactive Historical Sports Statics Databases from *http://www.bballsports.com/*. Retrieved Jan 30, 2008.

Burby, J., (2006). Data Smog: The Too-Much-Data Problem from *http://www.clickz.com/clickz/column/1691000/data-smog-the-too-much-data-problem*. Retrieved Oct 19, 2011.

Burns, E., R. Enns and D. Garrick, (2006). The Effect of Simulated Censored Data on Estimates of Heritability of Longevity in the Thoroughbred Racing Industry. *Genetic Molecular Research* 5(1): 7-15.

Chen, H. and M. Chau, (2004). Web Mining: Machine Learning for Web Applications. *Annual Review of Information Science and Technology (ARIST)* 38: 289-329.

Chen, H., P. Rinde, L. She, S. Sutjahjo, C. Sommer and D. Neely, (1994). Expert Prediction, Symbolic Learning, and Neural Networks: An Experiment on Greyhound Racing. *IEEE Expert* 9(6): 21-27.

DataSoftSystems, (2009). Data Mining - History and Influences from *http://www.datasoftsystem.com/articles/article-1380.html*. Retrieved Sept. 2, 2009.

Helland, P., (2011). If You Have Too Much Data, then "Good Enough" is Good Enough. *ACM Queue* 9(5).

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

Lazar, A., (2004). Income Prediction via Support Vector Machine, *International Conference on Machine Learning and Applications*, Louisville, KY.

Pincher, M., (2010). A Guide to Developing Taxonomies for Effective Data Management from *http://www.computerweekly.com/Articles/2010/04/06/240539/A-guide-to-developing-taxonomies-for-effective-data.htm*. Retrieved Oct 19, 2011.

Rogers, G., (2002). Death by Assessment: How Much Data Are Too Much? from *http://www.abet.org/Linked%20Documents-UPDATE/Assessment/Assessment%20Tips2.pdf*. Retrieved Oct 19, 2011.

Schumaker, R., O. Solieman and H. Chen, (2010). Sports Data Mining. New York, Springer.

Schumaker, R. P. and H. Chen, (2008). Evaluating a News-Aware Quantitative Trader: The Effects of Momentum and Contrarian Stock Selection Strategies. *Journal of the American Society for Information Science* 59(1): 1-9.

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

  ISSN: 1941-6679-On-line Copy

Schumaker, R. P., (2013). Data Mining the Harness Track and Predicting Outcomes, *Journal of International Technology and Information Management* 22(2):103-107.

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