

Using SVM Regression to Predict Harness Races: A One Year Study of Northfield Park

Robert P. Schumaker

Computer and Information Sciences Department
Cleveland State University, Cleveland, Ohio 44115, USA
rob.schumaker@gmail.com

Word Count: 4,106

Abstract

Can data mining tools be successfully applied to wagering-centric events like harness racing? We demonstrate the S&C Racing system that uses Support Vector Regression (SVR) to predict harness race finishes and tested it on one year of data from Northfield Park, evaluating accuracy, payout and betting efficiency. Depending upon the level of wagering risk taken, our system could make either high accuracy/low payout or low accuracy/high payout wagers. To put this in perspective, when set to risk averse, S&C Racing managed a 92% accuracy with a \$110.90 payout over an entire year. Conversely, to maximize payout, S&C Racing Win accuracy dropped to 57.5% with a \$1,437.20 return. While interesting, the implications of S&C Racing in this domain shows promise.

1. Introduction

Harness racing is a fast-paced sport where standard-bred horses pull a two-wheeled sulky with a driver. Races can either be trotting or pacing which determines the gait of the horse; trotting is the simultaneous leg movement of diagonal pairs whereas pacing refers to lateral simultaneous leg movement. North American harness racing is overseen by the United States Trotting Association which functions as the regulatory body.

Within this sport is the ability to wager on forecasted races, however, making accurate predictions has long been a problem. Even in situations where accurate forecasts are possible, it is entirely possible to focus on unimportant aspects. This 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).

Before making a wager, a bettor will typically read all the information on the race card and gather as much information about the horses as possible. They will also examine data concerning a horse's physical condition, how they have performed historically, their breeding and bloodlines, who is their trainer or owner, as well as odds of winning.

Automating this decision process using machine learning may yield equally predictable results as greyhound racing, which is considered to be the most consistent and predictable form of racing. Consistency lends itself well to machine learning algorithms that can learn patterns from historical data and apply itself to 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 create an arbitrage through its predictions. In particular, we will focus on closely examining the effect of longshots on racing prediction and payouts.

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.

2. Literature Review

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

For greyhound racing in particular, several studies have been successful in using machine learning algorithms in a limited fashion to make track predictions. While these studies used less accurate algorithms by today's standards, they laid the groundwork for racing prediction.

2.1 Predicting Outcomes

The science in prediction is not entirely concrete. However, there are methods that focus on informational discrepancies within a wagering environment. Methods such as market efficiency, mathematics and data mining can help us to better understand the motivations behind wagering activity.

Market Efficiency is all about the movement and use of information within a tradeable environment. This area of predictable science includes the use of statistical tests, conducting behavioral modeling of trading activity and forecasting models of outcomes to create rules from observation [1]. In statistical testing, it is assumed that sporting outcomes will closely mirror expectations, such that no betting strategy should win in excess of 52.4% [1]. This assumes that information is widely available to bettors and odds-makers alike. Deviations from these expectations could indicate the introduction of new or privately held information.

In Behavioral models, models of bettor biases are tested in order to determine any predictable decision-making tendencies. Perhaps the best known behavioral model is to select a wager according to the longshot bias where bettors will over-value horses with higher odds. Arrow-Pratt theory suggests that bettors will take on more risk in order to offset their losses [2]. In this approach it is argued that betting on favorites should be as profitable as betting on longshots. However, this is not the case which leads to a bias of longshot odds.

Forecasting models are a primitive use of historical data where seasonal averages, basic statistics and prior outcomes are used to build models and test against current data [1]. However, it was found that this approach had limitations and that the models were too simplistic and were poor predictors of future activity.

Mathematics is the area that focuses on fitting observed behavior into a numerically calculable model. It differs from Market Efficiency in that information within the market is not considered or tested. This area of predictive science includes the Harville formulas, the Dr. Z System and streaky player performance. Harville formulas are a collection of equations that establish a rank order of finish by using combinations of joint probabilities [3]. With Harville

formulas, it is believed that odds can be over-estimated on certain favorites which can lead to an arbitrage opportunity.

A little more sophisticated than the Harville formulas is the Dr. Z System. In this system, a potential gambler will wait until 2 minutes before the race, select those horses with low odds and bet Place (i.e., the horse will finish in 2nd place or better) on those with a win frequency to place frequency greater than or equal to 1.15. Subsequent studies later found that bettors were effectively arbitraging the tracks and that any opportunity to capitalize on this system was lost [4].

In streaky behavior, player performance is analyzed for the so-called “hot-hand” effect to see if recent player performance influences current performance [5]. While this phenomenon was studied in basketball and there was no evidence of extended streaky behavior [5], baseball academics remain unconvinced. In a study modeling baseball player performance, it was found that certain players exhibit significant streakiness, much more so than what probability would dictate [6].

While mathematics and statistics lay at the heart of data mining, the two are very different. Statistics are generally used to identify an interesting pattern from random noise and allow for testable hypotheses. The statistics themselves do not explain the relationship; that is the purpose of data mining [7]. Data Mining can be broken into three areas; Simulations, Artificial Intelligence and Machine Learning. Statistical simulations involve the imitation of new game data by using historical data as a reference. Once constructed, the simulated play can be compared against actual game play to determine the predictive accuracy of the simulation. Entire seasons can be constructed to test player substitution or the effect of player injury. The

drawback is that simulated data cannot address the complexities involved with large numbers of varying parameters.

Artificial Intelligence differs from other methods by applying a heuristic algorithm to the data. This approach attempts to balance out statistics by implementing codified educated guesses to the problem in the form of appropriate rules or cases. Heuristic solutions may not be perfect, however, the solutions generated are considered adequate [7].

The third branch, machine learning, uses algorithms to learn and induce knowledge from the data. Examples of these algorithms include both supervised and unsupervised learning techniques such as genetic algorithms, neural networks and Bayesian methods. These techniques can iteratively identify previously unknown patterns within the data and add to our understanding of the composition of the dataset [7]. These systems are better able to generalize the data into recognizable patterns.

2.2 Racing Prediction Studies

Predictive algorithms have been adopted successfully in non-traditional sports, such as greyhound and thoroughbred racing. These types of predictions generally involve machine learning techniques to train the system on the various data components, feed in new data and then extract predictions from it.

In a prior 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 domain experts, on 100 races at Tucson Greyhound Park [8]. From their work they made binary decisions as to whether or not each greyhound would finish first based on their historical race data. If a dog was predicted to finish first, their system would make a \$2 wager. Their ID3 decision tree was accurate 34% of the time with a \$69.20 payout while the BPNN was 20%

accurate with a \$124.80 payout. This disparity in decreased accuracy and increased payout is justified by arguing that the BPNN was selecting longshot winners. By doing so, accuracy would decrease but higher payouts could be gained because of the longer odds.

In a follow-up study that expanded the number of variables studied to 18, Johansson and Sonstrod also used a BPNN [9]. Their study on 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 as compared to Chen et. al. and would imply that either the additional variables or too few training cases (449 as compared to Chen's 1,600) hampered their ability to identify longshots.

In another study that focused on using discrete numeric prediction rather than binary assignment, Schumaker and Johnson used Support Vector Regression (SVR) on 10 performance-related variables [10]. Their study of 1,953 races across the US managed a 45.35% Win accuracy with a \$75.20 payout. To maximize payout, AZGreyhound had 23.00% Win accuracy with a \$1,248.40 payout. They found the same tradeoff between accuracy and payout.

In a study of thoroughbred racing, Williams and Li measure 8 race performance variables on 143 races and built a BPNN for each horse that raced [11]. This differed from other BPNN studies that created one network for all races. The system did manage 74% accuracy in selecting a winner.

2.3 Common Study Drawbacks

The drawbacks of the BPNN design used in previous studies is that win or loss categories are governed as a binary assignment as shown in Table 1.

Race #	Dog # - Name	Win/Loss
1	4 - Kma Baklava	1
1	2 - Coldwater Bravo	1
1	7 - Dollar Fa Dollar	1
1	3 - Stat U S Mystic	0
1	8 - Bf Oxbow Tiger	0
1	6 - Flyin Low	0
1	5 - Jr B-s Diesel	0
1	1 - Shining Dragon	0

Table 1. BPNN Binary Assignment

In Table 1, three dogs had strong enough historical data to warrant the system to pick them to finish first. This is because each dog’s chance of winning is evaluated independently of others. In this situation, you bet \$2 of each of the three to win, guaranteeing a minimum \$4 loss which adds overhead to the payout and in effect artificially depresses returns because only one dog can be declared the winner. This arrangement also fails to provide a rank ordering of finish. Is “Kma Baklava” more likely to finish in first place than “Coldwater Bravo?” The other problem with this type of a binary structure is how to decide bets on Place or Show.

From our investigation, we found several research gaps. The first of which is that not a lot of study has been done on discrete prediction. Most prior studies have relied on binary assignment, which is inefficient. The second gap is that none of these studies focus on harness racing. While harness racing enjoys some similarity between other racing types, it may be different enough that techniques from one racing type cannot be satisfactorily applied. The third gap is that many of the prior studies are black-box in nature and lack a fine-grain analysis. Neural networks generally provide one answer without any explanatory power regarding how the answer was derived. By using different learning techniques and exploring variations around answers, perhaps new knowledge can be gleaned.

3. Research Questions

From our analysis we propose the following research questions.

- ◆ *Can harness races be predicted better than random chance?*

We believe that using an SVR style approach may lead to better results than prior studies use of BPNN and ID3. SVR is capable of returning discrete values rather than strict classifications which can allow for a deeper inspection of accuracy results.

- ◆ *How profitable is the system?*

From prior research we know that accuracy and payout are inversely related. By increasing the amount of wagering risk to include longshots, we believe that there is a balancing point between them.

- ◆ *What wager combinations work best and why?*

Perhaps wagers on place and show may be more profitable. By looking at maximizing the return per wager, we can find the answer.

4. System Design

To address these research questions, we built the S&C Racing system shown in Figure 1.

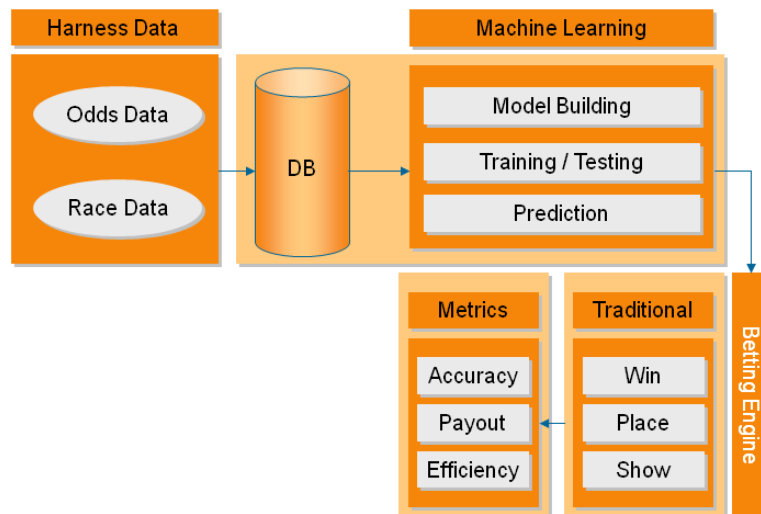


Figure 1. The S&C Racing System

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, 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.

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 three different types of wagers: Win, Place or Show. 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.

5. Experimental Design

To perform our experiment, we automatically gathered data from online web sources which consists of prior race results and wager payouts at Northfield Park; a USTA sanctioned harness track outside of Cleveland, Ohio. After data has been gathered, it is parsed for specific race variables before being sent to S&C Racing for prediction. For our collection we chose a study period of October 1, 2009 through September 30, 2010.

Prior studies focused on only one racetrack, manually input their data and had small datasets. Chen et. al. (1994) used 1600 training cases from Tucson Greyhound Park, Johansson and Sonstrod (2003) used 449 training case from Gulf Greyhound Park in Texas and Schumaker and Johnson used 41,473 training cases from across the US. Our study differs by automatically gathering race data from Northfield Park. In all, we gathered 1533 useable training cases covering 183 races. The reason for so few races was to maintain consistency with Chen et. al.'s approach which required a race history of the prior seven races. Since new entries would arrive in the Northfield market frequently, only 183 races could meet this stringent requirement.

Following Chen et. al. (1994), we limited ourselves to the following variables:

- ◆ Fastest Time – a scaled difference between the time of the race winner and the horse of interest, where slower horses experience larger positive values
- ◆ Win Percentage – the number of wins divided by the number of races
- ◆ Place Percentage – the number of places divided by the number of races
- ◆ Show Percentage – the number of shows divided by the number of races
- ◆ Break Average – the horse's average position out of the gate
- ◆ Finish Average – the horse's average finishing position
- ◆ Time7 Average – the average finishing time over the last 7 races, in scaled difference from the winners
- ◆ Time3 Average – the average finishing time over the last 3 races, in scaled difference from the winners

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 'Road Storm,' we compute the variables for the prior seven races as shown in Figure 2.

Horse	Race Date	Race #	Fastest Time	Win Percentage	Place Percentage	Show Percentage	Break Average	Finish Average	Time7 Average	Time3 Average	Predicted Finish
Road Storm	2/27/2010	11	0.0	6.67%	11.11%	0.00%	2.67	2.00	0.24	0.00	0.8846

Figure 2. Road Storm variable data

For this particular race, Race 11 on February 27, 2010, Road Storm’s Fastest Time is 0.0 implying that they were the winner of this race. If they were 1.0 seconds behind the winner, the Fastest Time would be 1.0. Road Storm has a Win percentage of 6.67%, a Place percentage of 11.11% and a Show percentage of 0.00% indicating how often they finish in each of those positions. The Break Average is 2.67 meaning that this horse is typically in the front of the pack coming out of the gate. The Finish Average is 2.00 indicating plenty of second place finishes. The Time7 Average shows that over the past seven races, Road Storm has been 0.27 seconds behind the leader. The impressive Time3 Average of 0.00 indicates first place finishes in the prior three races. S&C Racing predicts from its internal model that Road Storm should finish 0.8846 which is a strong finish, but cannot be fully interpreted until compared with the predicted finishes of other horses in a 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. As a visual aid, Table 2 shows the race output for Northfield Park’s Race 11 on February 27, 2010.

Horse # - Name	Finish	Predicted Finish
3 - Road Storm	1	0.8846
1 - Saucy Brown	2	3.6521
6 - Jeff the Builder	3	4.4880
8 - Just Catch Ya	4	5.5708
2 - Tessallation	7	5.6452
4 - Key Western	5	6.3366
5 - Auto Pilot	6	6.6634
7 - Look Ma No Pans	8	8.3325

Table 2. Predicted Values

From this table, we can wager on Road Storm to win, Saucy Brown to place and Jeff the Builder to show based on S&C Racing's predicted finish.

For the machine learning component we implemented Support Vector Regression (SVR) using the Sequential Minimal Optimization (SMO) function through Weka [12]. SVR allows for discrete numeric prediction instead of classification. We also selected a linear kernel and used ten-fold cross-validation.

6. Experimental Findings and Discussion

To answer our first research question of *can harness races be predicted better than random chance* we present Figure 3 that looks at the amount of risk versus accuracy for the three wager types.

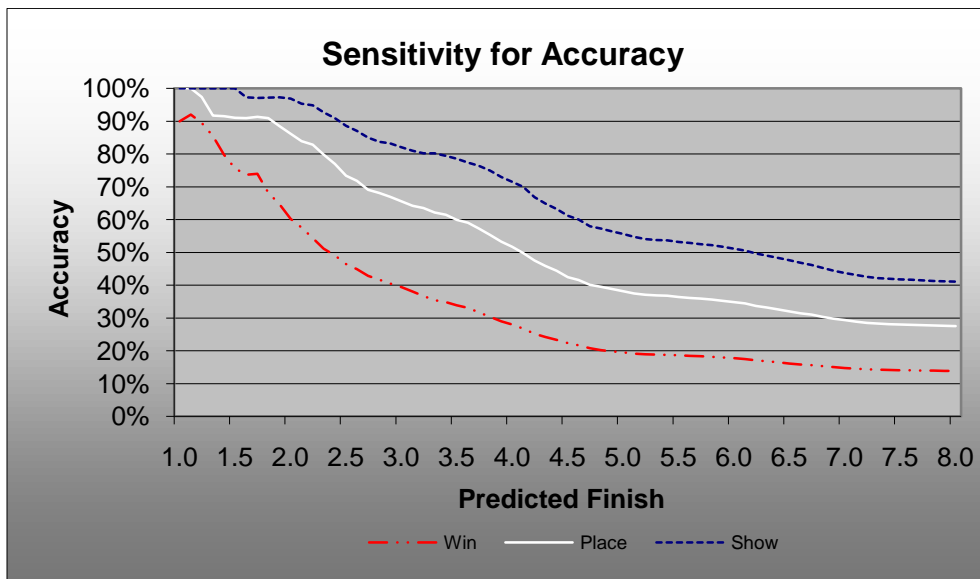


Figure 3. S&C Racing Accuracy

From our data, there was an average of 8.317 horses per race. The random chance of selecting a win wager is 12.02%, for place is 24.05% and show is 36.07%. To maximize accuracy, win had 92.0% accuracy by wagering on horses predicted to finish 1.1 or better (p-value < 0.01) by correctly selecting 23 of 25 wagers. This shows a tradeoff between high

accuracy and a low number of wagers; there were only 25 betting instances in a one year period where a horse was predicted to finish 1.1 or better. Place managed 100% accuracy on all 25 wagers for predicted finishes of 1.1 or better (p-value < 0.01). Show had 100% accuracy on 78 wagers for predicted finishes of 1.5 or better (p-value < 0.01). Some observations about accuracy was that as the strength of finish prediction decreased, so to did accuracy. This was expected as weaker horses are being selected. Accuracies approached random chance when wagering on every horse, which was expected.

To answer our second research question of *how profitable is the system* we analyze wager payouts and a function of prediction strength as shown in Figure 4.

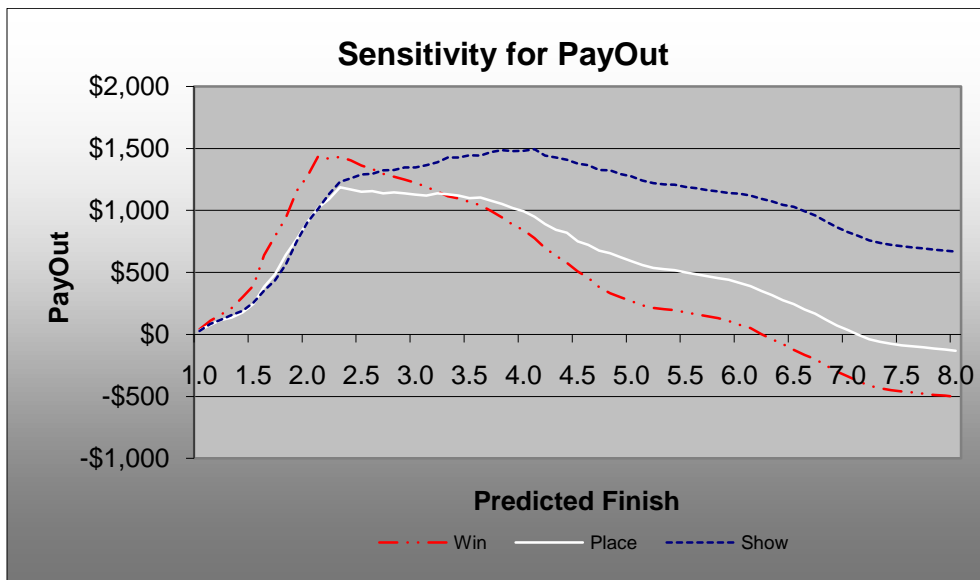


Figure 4. S&C Racing Payouts

In maximizing payouts, win had a payout of \$1,437.20 with 57.5% accuracy on 299 wagers (p-value < 0.01). Any horse wagered on with a predicted finish of 2.1 or better maximized win payout at the expense of accuracy. Keep in mind that the payout return is over a one year period. For Place, the maximum payout was \$1,168.30 with 76.9% accuracy on 373 wagers (p-value < 0.01). Any horse wagered on with a predicted finish of 2.4 or better

maximized place payout. For Show, the maximum payout was \$1,492.40 with 70.0% accuracy on 689 wagers (p-value < 0.01). Any horse wagered on with a predicted finish of 4.1 or better maximized show payout.

To answer our third research question of *what wager combinations work best and why* we analyze betting efficiency as a function of prediction strength as shown in Figure 5.

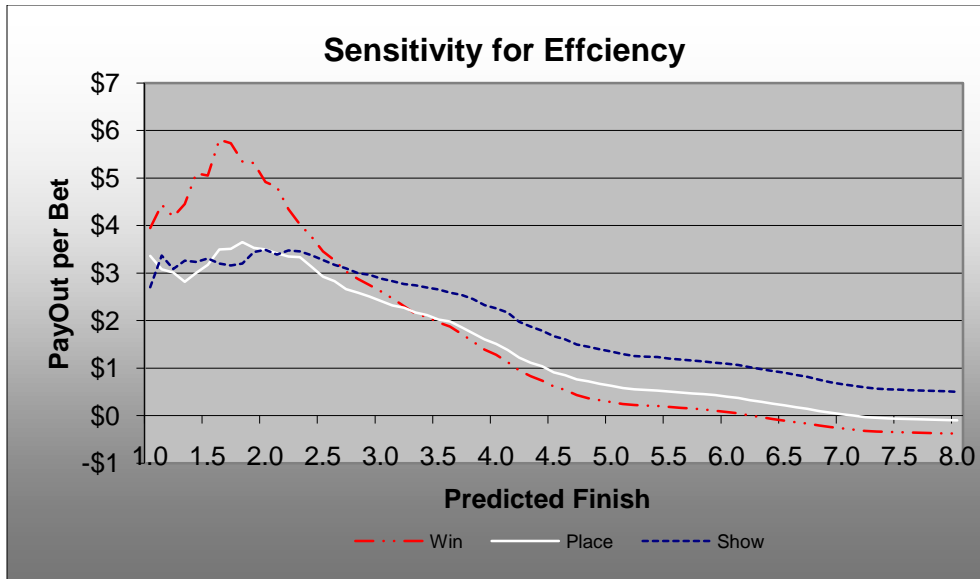


Figure 5. S&C Racing Betting Efficiency

In maximizing betting efficiency, the amount of return per wager, Win had a peak return efficiency of \$5.81 when wagering on horses predicted to finish 1.6 or better (p-value < 0.01). Place peaked at \$3.65 return when wagering on horses predicted to finish 1.8 or better (p-value < 0.01). Show peaked at \$3.49 return when wagering on horses predicted to finish 2.0 or better (p-value < 0.01). Win had 73.6% accuracy on 110 wagers. When wagering on horses predicted to finish 6.3 or worse, win had a negative return. Place had 90.9% accuracy on 176 wagers. When wagering on horses predicted to finish 7.1 or worse, place had a negative return. Show had 96.9% accuracy on 259 wagers and efficiency leveled out to twenty cents return when wagering

on all horses to show. While positive, a twenty cent return is somewhat unattractive to many bettors.

7. Conclusions and Future Directions

From our investigation we found that S&C Racing was able to predict harness races somewhat reasonably. When wagering on win, the system had 92% accuracy but had a low number of wagering opportunities. Place and Show both had 100% accuracy, but also a dearth of opportunity to use it. The Show wager was the most profitable with a \$1,492.40 payout. Accuracy decreased as the payout increased. The Win wager was the most efficient wager returning an excess of \$5.81 per wager. However, as the system bet on more horses, efficiency and accuracy declined.

Future directions for this stream of research include looking into more exotic wagers and examine their profitability as well as looking into thoroughbred racing and see if the same techniques developed for greyhound racing can be applied within that domain.

References

- [1] J. Dana, and M. Knetter, "Learning and Efficiency in a Gambling Market," *Management Science*, vol. 40, no. 10, pp. 1317-1328, 1994.
- [2] J. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, vol. 32, no. 1-2, pp. 122-136, 1964.
- [3] R. Sauer, "The Economics of Wagering Markets," *Journal of Economic Literature*, vol. 36, no. 4, pp. 2021-2064, 1998.
- [4] J. Ritter, "Racetrack Betting - An Example of a Market with Efficient Arbitrage," *Efficiency of Racetrack Betting Markets*, D. Hausch, V. Lo and W. Ziemba, eds., San Diego: Academic Press, 1994.
- [5] A. Tversky, and T. Gilovich, "The Cold Facts About the "Hot Hand" in Basketball," *Anthology of Statistics in Sports*, J. Albert, J. Bennett and J. Cochran, eds., Philadelphia: SIAM, 2005.
- [6] J. Albert, "Streaky Hitting in Baseball," *Journal of Quantitative Analysis in Sports*, vol. 4, no. 1, 2008.
- [7] R. Schumaker, O. Solieman, and H. Chen, *Sports Data Mining*, New York: Springer, 2010.
- [8] H. Chen, P. Rinde, L. She *et al.*, "Expert Prediction, Symbolic Learning, and Neural Networks: An Experiment on Greyhound Racing," *IEEE Expert*, vol. 9, no. 6, pp. 21-27, 1994.
- [9] U. Johansson, and C. Sonstrod, "Neural Networks Mine for Gold at the Greyhound Track," in International Joint Conference on Neural Networks, Portland, OR, 2003.
- [10] R. P. Schumaker, and J. W. Johnson, "Using SVM Regression to Predict Greyhound Races," in International Information Management Conference, San Diego, CA, 2008.
- [11] J. Williams, and Y. Li, "A Case Study Using Neural Network Algorithms: Horse Racing Predictions in Jamaica," in International Conference on Artificial Intelligence, Las Vegas, NV, 2008.
- [12] I. H. Witten, and F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*, San Francisco: Morgan Kaufmann, 2005.