Cluster Analysis Model for Selection of High Performing Greyhounds

Anil Gulati\textsuperscript{1} and William Bosworth\textsuperscript{2}

Abstract

Conventional economic and finance theory postulates that observed prices in a competitive market will reflect an accurate assessment of expected outcomes. In legal sports wagering the relative value of a risky asset (a wager) can be compared to its relative expected outcome to gain insight into the process of price discovery of the asset. At least 15 studies spanning 30 years and four continents provide evidence that participants in pari-mutuel betting make consistent, exploitable errors in their betting decisions. There are four explanations for this observed market inefficiency, including behavioral anomalies, asymmetric information, measurement error, and noisy signaling inherent in the bidding process. This study tests a betting rule based on the win history of racing greyhounds using data taken from six greyhound race tracks. Applying cluster analysis to racing greyhounds we show that readily available information is not efficiently exploited giving rise to a statistically and economically significant profitable betting rule. Our findings are consistent with behavioral anomalies and do not support the asymmetric information hypothesis and the measurement error explanation suggested in other studies.

Key Words: pari-mutuel betting, gambling, favorite longshot bias, greyhound racing.

1. Introduction and Review of Literature

Pari-mutuel betting, the kind of wagering that takes place at horse and greyhound racing venues in the U.S. and other nations, has provided a rich source of data with which to investigate the theory of market efficiency. According to this theory, an efficient market is one in which the price of an asset is an unbiased estimate of its expected yield. If the market for wagers in horse and greyhound racing is efficient, racing market participants will respond to imbalances between payout ratios and reward probabilities. Eventually the “market odds” or payout ratios should reflect the relative probabilities of winning. In general a participant should, on average, realize negative returns equal to the portion of the pool retained by the track operator. While many studies\textsuperscript{3} have detected betting patterns inconsistent with efficient markets, others\textsuperscript{4} have provided evidence that the methodologies linking relative probabilities and market odds are biased in favor of finding against efficient markets in pari-mutuel betting. This study uses a new technique to overcome some of the measurement errors involved in previous studies. We apply this technique to greyhound racing and find statistically and economically significant unexploited betting opportunities that are inconsistent with an efficient market.

Pari-mutuel betting is a form of betting in which all similar wagers are pooled, a certain portion is deducted from the pool by track operators for overhead, and the remainder is proportionally divided among the winners. As bets are cast participants are able to view the payout ratios on the Tote board (an electronic display) as the pools accumulate. The efficient markets hypothesis holds that payout ratios for the racing entries that have a low (high) probability of winning should be large (small). Any payout ratio that is large relative to the probability of winning will provide an opportunity for a bet with an expected profitable outcome. A bet with a potential payout that is small relative to the probability of winning has an expected loss. The former betting opportunity will attract more bets causing the pool to be divided among more potential winners thus reducing the payout ratio. At the same time the latter will be avoided and the payout ratio will increase as the potential pool is distributed over fewer winning bets. Griffith (1949) observed that the behavior of betters is consistent with agents who can detect the differences in the probability of different outcomes. But a seminal study by Ali (1977) uncovered an anomalous behavior. In his study of horse race betting he found that bettors tend to “under bet” horses with a low payout ratio (high probability of winning) and “over bet” horses with a high payout ratio (low probability of winning).

\textsuperscript{1}Corresponding Author, Western New England University, 1215 Wilbraham Road, Springfield MA. 01119, USA.
\textsuperscript{2}Western New England University, 1215 Wilbraham Road, Springfield MA. 01119, USA.
\textsuperscript{4}Walls and Busche (2003) and Busche and Walls (2001).
This anomaly, known as the favorite long shot bias (FLB), has since been detected in many other studies of horse racing including Snyder (1978), Ziemba and Hausch (1984, 1987), Thaler and Ziemba (1988), and Asche and Quandt (1990), and Winter and Kukuk (2006). While Gramm and Owens, (2006) detect FLB they also find evidence that increased access to betting through simulcasting is reducing inefficiency. The existence of the FLB is not without controversy. Snyder (1978), Ziemba and Hausch (1984, 1987) show that the win market is weakly efficient in the sense that there is no profitable betting strategy. Gramm et. al. (2012) argue that FLB is not economically significant. Busche and Walls (2003) and Walls and Busche (2003) offer an alternative explanation of the observed FLB. Pari-mutuel betting includes not only a track “take-out” but also a protocol of rounding payout ratios down to the nearest amount divisible by 20 cents in the case of horse racing and to the nearest 10 cents in the case of greyhound racing. That is, if a correctly calculated payout is $16.59 on a two dollar bet, the actual amount paid out is only $16.40. The track retains the 19 cents in addition to its percentage take-out of the total pool. Busche and Walls (2003) point out that this fee, called “breakage” is an average 10 cents per bet and is uniformly distributed across all payments. The favorites have low payout ratios so that the breakage represents a higher percentage of the winnings than when a long shot wins with a relatively high payout ratio. Walls and Busche (2003) using data from horse tracks in Hong Kong and Japan and Busche and Walls (2003) using Ali’s (1977) data arranged the reported racing results in order of the size of the realized breakage amounts. They found that FLB existed only in the high “breakage” races.

Gulati and Shetty (2007) applied Ali’s (1977) methodology to greyhound racing. Not only did they find evidence that rejects FLB, they found that high probability greyhounds were over bet and low probability greyhounds were under bet, that is, the opposite of FLB. They conjectured that the discrepancy reflects the manner in which racing fields are selected in greyhound racing. In greyhound racing entrants are entered into a particular race based on their recent success/failure history. Winning entrants are moved up to a higher/more competitive grade and vice versa. Winners of the most competitive races are bound by the lack of more competitive races to enter. Entries in a race of any grade, other than the highest, may be one of three types: those just moved up by winning a race in the next lower grade, those just moved down from a higher grade due to poor showing and those which maintaining the current grade based on above average performance. This composition of the field creates an illusion that favorite greyhounds are more likely to win in less competitive races and these are the greyhounds that are over bet.

The explanations for why FLB is observed in pari-mutuel betting fall into two categories, behavioral explanations and game theoretic explanations. Behavioral explanations include Griffith (1949) - miscalculation of the odds; Rosett (1965), Weitzman (1965) Quandi (1986) and Ali (1977) – risk seeking behavior; Kahneman and Tversky (1984) and Thaler (1985) – “mental accounting” or compartmentalization of wealth; and Canfield, Fauman, and Ziemba (1987) – the value of “bragging rights” from successful long shot bets. The game theoretic assumes that there are both privately informed and “noisy” bettors. Embedded in the subjective probabilities distribution of the noisy bettors are the objective probabilities known to the informed bettors. Therefore the posted odds influence the noisy bettors’ decisions. That is, noisy bettors observe posted odds and try to extract information from them. In Potters and Wit (1996), Feeney and King (2001), and Koessler and Ziegelmeier (2003), informed bettors are able to influence the final odds after the noisy bettors have cast their bets. Ottaviani and Sorensen (2003) show that the strength of belief of informed bettors is higher for high probability entrants. In the moments leading up to the race, noisy bettors update their subjective probabilities by observing posted odds. Therefore informed bettors wait until the last moment to bet because an early bet can provide a signal to noisy bettors that would dilute their potential winnings as bets already placed cannot be withdrawn. Koessler et. al. (2008) and Axelrod et. al. (2009) model similar betting externalities. These models predict the behavior observed by Asch, Malkiel, and Quandt (1982) and Gandar, Zuber, and Johnson (2001) in which the final odds of the ultimate winners decline over the betting period. In these studies the problem of unequal distribution of information is unresolved due to an incomplete tâtonnement process resulting in FLB.

All of these game theoretic models assume that there is asymmetric information. If, in fact, there is publicly available information that could lead to a profitable betting rule then the game theoretic could be rejected in favor of alternative explanations including behavioral anomalies. Brately (1973) attempted to use win history in the context of horse racing to form a betting rule but failed to do so. Bolton and Chapman (1986) use a multinomial Logit model to derive a successful horse racetrack betting rule using a list of explanatory variables that includes win percentage, average speed in previous races, weight, post position, whether the distance of the race is within the horse’s experience, the win percent of the jockey, and the number of wins of the jockey.
Goodwin (1996) argued that models based on an assumption of any particular distribution of expectational errors (e.g. normal, logistic, etc.) can lead to biased estimators of underlying parameters. Further, heteroscedasticity in greyhound racing due to the groupings of similar dogs (e.g. successful dogs, less successful dogs) in different races can lead to inconsistent parameter estimates. He built a model that uses various historical variables, handicapping and posted odds as predictors. His model significantly outperforms the predictions of professional handicappers. In contrast to both these studies, we directly test a simpler betting strategy based only on win history. We avoid parameterization by using cluster analysis which minimizes the distance functions among greyhounds based on their win history.

All of the empirical studies cited so far (except Griffith, 1949) have grouped the data and calculated the statistics to test the hypothesis of market efficiency in the manner introduced by Ali (1977). In this method the racing entrants are ranked based on their betting odds. That is, the entrant with the lowest (highest) payout ratio is ranked first (last). The gross payout ratio for the \( i^{th} \) entry is a function of the entry’s relative share of the betting pool.

\[
P_i = \frac{(1 - t) \times \sum_{i=1}^{G} X_i}{X_i},
\]

(1)

where,

- \( P_i \) = The gross return,
- \( t \) = The percentage “take-out” retained by the track operator,
- \( X_i \) = The amount of money bet on the \( i^{th} \) entry,
- \( G \) = The number of entries racing in the race,
- \( \sum X_i \) = The total amount in the betting pool across all entries.

The probability of winning implied by the pattern of wagers (IP) is,

\[
IP_i = \frac{X_i}{\sum_{i=1}^{G} X_i}.
\]

(2)

To adjust for the “breakage” the implied probability is calculated as,

\[
IP_i = \frac{X_i}{\left(\sum_{i=1}^{G} X_i + \frac{RD}{2}\right)}
\]

(3)

Where, RD is the maximum possible “breakage.”

Wall and Busche (2003) and Busche and Wall (2003) departed from this procedure by omitting the RD/2 term. Instead, they further grouped observations ranked by the amount of breakage actually realized. Thus, each entry was grouped in two ways, once by IP and once by the breakage that resulted in the final payout. When they did so they found statistically significant evidence that ignoring breakage costs biases the results in favor of FLB.

The present study departs from the method of Ali (1977) by using cluster analysis in lieu of the implied probability calculation of Equation 3. In greyhound racing the win history of every entry is available to all race track participants. Therefore, it is possible to calculate the actual probability of its win predicated on its history. If, using historical data instead of the posted payout ratios, any bias can still be detected then the efficient markets hypothesis is rejected by the data. Section II reports on the data collection and the methodology, Section III reports the results, and Section IV concludes.

2. Data and Methodology

We chose six greyhound racetracks for this study with the purpose of creating a geographically diverse sample. Collectively, the observation periods on these six tracks represent about 230,000 races. The track names and the observation periods are given below.

Insert Table 1 Here

Data Selection

To give our study a clear focus we narrowed our study by using information that is publicly available to bettors. On each track we only considered races that started with a full box – i.e. eight greyhounds.
Further, all greyhounds that had not competed in each of the top three grades in their racing careers were removed. Having defined our field of study we calculated the win ratio of each of the greyhounds. The result is a table of every greyhound that competed in the top three grades and their corresponding win ratio in each of the top three grades.

2.1 Data Analysis

Using historical performance as the defining characteristic we clustered the greyhounds on each track into two groups: “High” win frequency and “Low” win frequency. The clustering was performed using the k-means cluster analysis in SPSS. All groups converged within 20 iterations. Table II shows the number of greyhounds classified as “High” win frequency, and “Low” win frequency, for each of the six tracks.

Insert Table 2 Here

For the entire field, the number of “Low” win frequency greyhounds outnumbered the “High” win frequency counts by about a multiple of seven. The range was from 5.97 times (Gulf) to 7.04 times (Birmingham). The k-means analysis tool in SPSS was used to classify the greyhounds into the two clusters. The final centers of the “High” win frequency cluster are shown in Table III below. The win percentage is decomposed by grade (highest, second and third).

Insert Table 3 Here

For the highest grade, the win proportion of the “High” win frequency cluster is low relative to the other two grades. Greyhounds are moved up a grade with superior performance and are downgraded to the next lower grade due to sub-par performance. Superior performers in the top grade cannot move any higher. Better performers remain in the highest grade and sub-par performers are downgraded. Therefore, the field in a top grade race is composed of greyhounds maintaining that grade or moving up by winning the next lower grade race. Therefore, the field in the top grade is more competitive (evenly matched) which results in lower win percentage. The final centers of the “Low” win frequency greyhounds cluster are shown in Table IV, below. The win percentage is decomposed by grade, as before.

Insert Table 4 Here

As noted in the case of the “High” win frequency greyhounds, the “Low” win frequency greyhounds also won the least in the highest grade.

2.2 Methodology

Our hypothesis is that the greyhound racing market is dominated by rational racing participants. The null hypothesis is the weak form of the efficient market hypothesis: no bets should have positive expected values. We assume that racing participants have access to win histories of the greyhounds and are able to form expectations based on those histories. We do not attempt to model the information available from handicappers, odds posted by the track operators, or evolving posted odds. We assume that betting participants study the history of the greyhounds in a race and determine the probability of their winning based on their past performance, compare those probabilities to the posted odds before making betting decisions. The clusters therefore classify the greyhounds in two groups, i.e. the group with high winning potential and the other with low probability of winning the race. If bettors are able to assess these probabilities accurately and bet accordingly, there should be no opportunities for excess returns. Actually, the returns should be negative and equal the track take-out. To test our hypothesis, we placed two dollar hypothetical bets on each greyhound in the two clusters. We segregated the returns by cluster for Favorites and Longshots and the results are presented in the next section.

3. Results

Table V shows results of a hypothetical $2 bet to “win” wagered on the greyhound in the “High” cluster when the entry was the “favorite to win.” Similarly a hypothetical $2 bet to “win” was wagered on the favorites in the “Low” cluster, for the top three grades studied. These returns are calculated net of the track take-out.

Insert Table 5 Here

The returns in “High” cluster are positive for each of the six tracks, thus representing profit making opportunities. Conversely, the returns generated in “Low” cluster are negative across the board, with one exception of Derby Lane. Explanation of this exceptional behavior would require deeper analysis which is beyond the scope of the present study. For now, we attribute this exception to shorter observation period.
Table VI shows results of similar hypothetical $2 bet to “win” wagered on the greyhound in the “High” cluster when it was the Longshot. Similarly a hypothetical $2 bet to “win” was wagered on the Longshot in the “Low” cluster, for the top three grades studied. These returns are calculated net of the track takeout.

Insert Table 6 Here

These results generally follow the pattern that returns in the “High” cluster are far superior than the “Low” cluster. Also, we must underscore the positive returns experienced by the “Low” cluster greyhounds. We note that these results signal that the FLV bias is present in our data set.

3.1 Discussion of Results

Our results show that bettors are unable to differentiate between the “High” and “Low” potential (cluster) entries. Significant and positive returns in the “High” cluster result for these entries, collectively, attracting smaller percentage of the Win pool as compared to the true odds of winning. Therefore, we conclude that these entries are significantly under-bet. Conversely, the entries in “Low” cluster are either over-bet, or at the very least represent fair bets. Our method of analysis discriminates between entries with a potential to consistently win races and between a second group with more unpredictable outcomes. Second, we show that market participants are unable to discriminate between the two groups.

4. Conclusions

In this study we tested a simple model in which racing participants used historical data to bet on greyhound races. Our hypothesis is that racing participants efficiently exploit publicly available information. Our test consisted of exploiting the win history of the greyhounds in each race as the sole predictor of the win probability. The results of our study reject the hypothesis of an efficient market in pari-mutuel greyhound racing. Deviations from rational betting payouts were both statistically and economically significant for the six race tracks observed in this study. The results of this study do not support the asymmetrical explanation of systematic bias in pari-mutuel betting of Potters and Wit (1996), Feeney and King (2001), Koessler and Ziegel (2003), Ottaviani and Sorensen (2003), Koessler et. al. (2008), and Axelrod et. al. (2009). Moreover, this study overcomes the measurement problems of Walls and Busche (2003) and Busche and Walls (2003). We therefore conjecture that we are observing a behavioral phenomenon inconsistent with efficient markets.

References


<table>
<thead>
<tr>
<th>Track Name</th>
<th>Track Code</th>
<th>Racing results from</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Birmingham Race Course</td>
<td>Birmingham</td>
<td>Apr 2002 - Aug 2011</td>
</tr>
<tr>
<td>2 Bluffs Run Greyhound park</td>
<td>Bluffs</td>
<td>Jan 1997 - Aug 2011</td>
</tr>
<tr>
<td>3 Derby Lane</td>
<td>Derby</td>
<td>Mar 2008 - Aug 2011</td>
</tr>
<tr>
<td>5 Jacksonville Greyhound Racing</td>
<td>Jacksonville</td>
<td>May 2002 - Aug 2011</td>
</tr>
<tr>
<td>6 Palm Beach Kennel Club</td>
<td>Palm</td>
<td>Jun 2004 - Aug 2011</td>
</tr>
</tbody>
</table>
### Table II. Cluster Analysis Results, Number of Greyhounds Classified in the Two Clusters

<table>
<thead>
<tr>
<th>Track Code</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Birmingham</td>
<td>407</td>
</tr>
<tr>
<td>2</td>
<td>Bluffs</td>
<td>581</td>
</tr>
<tr>
<td>3</td>
<td>Derby</td>
<td>158</td>
</tr>
<tr>
<td>4</td>
<td>Gulf</td>
<td>219</td>
</tr>
<tr>
<td>5</td>
<td>Jacksonville</td>
<td>336</td>
</tr>
<tr>
<td>6</td>
<td>Palm</td>
<td>269</td>
</tr>
</tbody>
</table>

### Table III. Final Centers of the "High" Cluster (Percent of Races Won in Top Three Grades)

<table>
<thead>
<tr>
<th>Track Code</th>
<th>Highest</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Birmingham</td>
<td>14.64%</td>
<td>46.41%</td>
</tr>
<tr>
<td>2</td>
<td>Bluffs</td>
<td>14.41%</td>
<td>34.10%</td>
</tr>
<tr>
<td>3</td>
<td>Derby</td>
<td>15.01%</td>
<td>31.96%</td>
</tr>
<tr>
<td>4</td>
<td>Gulf</td>
<td>17.46%</td>
<td>48.08%</td>
</tr>
<tr>
<td>5</td>
<td>Jacksonville</td>
<td>13.21%</td>
<td>29.70%</td>
</tr>
<tr>
<td>6</td>
<td>Palm</td>
<td>15.39%</td>
<td>37.11%</td>
</tr>
</tbody>
</table>

### Table IV. Final Centers of the "Low" Cluster (Percent of Races Won in Top Three Grades)

<table>
<thead>
<tr>
<th>Track Code</th>
<th>Highest</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Birmingham</td>
<td>8.29%</td>
<td>12.59%</td>
</tr>
<tr>
<td>2</td>
<td>Bluffs</td>
<td>5.05%</td>
<td>13.61%</td>
</tr>
<tr>
<td>3</td>
<td>Derby</td>
<td>5.36%</td>
<td>15.42%</td>
</tr>
<tr>
<td>4</td>
<td>Gulf</td>
<td>6.09%</td>
<td>16.00%</td>
</tr>
<tr>
<td>5</td>
<td>Jacksonville</td>
<td>5.00%</td>
<td>15.23%</td>
</tr>
<tr>
<td>6</td>
<td>Palm</td>
<td>5.25%</td>
<td>16.08%</td>
</tr>
</tbody>
</table>

### Table V. Hypothetical Betting Results for "High" and "Low" Clusters - Rates of Return for $2 Bet to Win on Favorites.

<table>
<thead>
<tr>
<th>Track Code</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Birmingham</td>
<td>23.72%</td>
</tr>
<tr>
<td>2</td>
<td>Bluffs</td>
<td>7.92%</td>
</tr>
<tr>
<td>3</td>
<td>Derby</td>
<td>26.15%</td>
</tr>
<tr>
<td>4</td>
<td>Gulf</td>
<td>8.84%</td>
</tr>
<tr>
<td>5</td>
<td>Jacksonville</td>
<td>9.25%</td>
</tr>
<tr>
<td>6</td>
<td>Palm</td>
<td>1.61%</td>
</tr>
</tbody>
</table>

### Table VI. Hypothetical Betting Results for "High" and "Low" Clusters - Rates of Return for $2 Bet to Win on Longshots.

<table>
<thead>
<tr>
<th>Track Code</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Birmingham</td>
<td>179.34%</td>
</tr>
<tr>
<td>2</td>
<td>Bluffs</td>
<td>91.28%</td>
</tr>
<tr>
<td>3</td>
<td>Derby</td>
<td>28.85%</td>
</tr>
<tr>
<td>4</td>
<td>Gulf</td>
<td>49.79%</td>
</tr>
<tr>
<td>5</td>
<td>Jacksonville</td>
<td>42.20%</td>
</tr>
<tr>
<td>6</td>
<td>Palm</td>
<td>84.64%</td>
</tr>
</tbody>
</table>