EDA and Analysis of Factors on Horse Racing Data

Hita Juneja
Dept. of CSE
PES University
Bangalore, India
hitajuneja@gmail.com

Ishita Bharadwaj
Dept. of CSE
PES University
Bangalore, India
ishitabharadwaj04@gmail.com

***Abstract*—We choose to examine the many facets of derbies held throughout the world in our literature review and analyze those components that significantly influence final race positions. For a thorough understanding of this subject, we studied articles, academic papers, and watched stakes races. This report provides exploratory data analysis on the nature of derby racing using trajectory plots and data visualisation. We have concentrated on how jockey's strategies and drafting techniques influence a horse's speed.**

***Index Terms*—Derbies, Horse Racing, Jockey's strategies, Drafting Techniques**

I. Introduction

America's most prestigious and oldest horse races are held annually at Aqueduct, Belmont, Kentucky, Saratoga, Travers Stakes and Breeder's Cup. Horse racing in the United States is an extensive sport that spreads across the country and facilitates wide-ranging betting opportunities.

Each horse is tracked every 0.25 seconds, allowing for a substantial deduction of new attributes and reasoning. Analysis is also done on how a jockey's weight and drafting technique affect a horse's performance in a race. To win the race, there is a lot of planning and preparation that takes place in the backdrop. The jockey must take the pace of the horses into account while planning his drafting strategy because the distance varies from race to race. The pace of the horses is determined by their location on the course and their relative position to the competition. Purse refers to the total amount of money paid out to the owners of horses racing at a particular track. Percentages of a race's total purse are awarded to each of the top 5 or 6 finishers. Understanding the derby dataset helps to enhance equine welfare, performance, and rider decision-making as well as innovative methods of training and competing in horse racing.

Time series analysis is incorporated into this article, since data is continuously recorded after predetermined interval of 0.25 seconds. During a derby race, the horses' locations (trakus index, longitude, and latitude) are noted. These readings are collected chronologically and are then utilised to reveal patterns and changes over time.The more regular and repeatable patterns, the easier it is to forecast the series.The approximate entropy will be used to quantify the unpredictability. The higher the approximate entropy, the more difficult it is to forecast it [1].

In Derby racing, it is extremely important for the bettors to know the different types of derbies in order to make wise bets. Horses that have never won compete in maiden races. Horses remain maiden until they notch their first win. In handicap racing, horses that are thought to have an early edge carry additional weight to provide the other horses in the field with a handicap, effectively balancing the situation. In claiming races, horses are sold at races because bids as placed on them prior of the events. The winner offers the owner his winnings in exchange for the horse. Allowance races are the next level up from claiming races. In order to level the playing field for racing, the horses in this case are not for sale and must carry weight. They may include races for non-winners of a certain number. Stakes are the highest level of horse races, and naturally, pay the largest purse. The Kentucky Derby, whose data set we are using in this study, is one of the most prestigious Grade I stakes races.

The aim of this paper is to analyze horse racing tactics, drafting strategies, and path efficiency. Considering different venues, surfaces and race distances, optimal racing strategies will be inferred.

II. Related Work

In his article, D. S. Gardner has emphasized the parallel development of derby racing for horses and human marathon runners. A record for the Kentucky Derby was achieved in 1973 in 1.59.00 minutes, an improvement of 4% over the previous century, and an improvement of 11% over the Epsom Derby. The winning time in horse races is still influenced by a variety of racing strategies and additional elements including stall position, track circumstances, jockey skills, etc. This paper emphasizes the percentage gain in performance over time by comparing horse race timings and human winning times before and after the 1950s [2].

In addition, it analyses the slope and variance coefficients for every race held at Epsom Derby, Melbourne Cup, and Kentucky Derby. The pre-1950 era had a substantially higher value of slope indicating the rate of change in timings than the contemporary era, with an average reduction in time of 4.2%. This is because there was far more room for improvement in the pre-1950 era. On the other hand, men's performance has improved by an average of 10.4%, with a specific increase of 32% for women's marathons. This is significant when compared to the development in horse racing. The principles of exercise physiology, nutrition, and interval training can

therefore be used in horse racing to yield results that are comparable to those seen in man during the past 50 years.

Sameerchand Pudaruth's study examines how changes in horse racing odds affect the results of races. The odds differ between tracks and races; they indicate how much money can be made in relation to the wager. It serves as a predictor of the horse with a better chance of winning the race. The variance in odds leading up to the races was recorded and fed to a NeuroXL software neural network. Ranks were thresholded to a win/lose characteristic. 7.4% of predictions were correct, which was fewer than the 11.4% of predictions that were correct due to pure chance. Thus, it was discovered that the variance in odds has no discernible impact on the winner of a race [3].

III. Dataset Overview

Start, racing, and tracking data are part of the dataset, provided by The New York Racing Association (NYRA) and the New York Thoroughbred Horsemen's Association (NYTHA). World class thoroughbred races are held in the United States at either Aqueduct Racetrack, Belmont Park, or Saratoga Race Course. There are listed race dates, post times, race numbers, and race distances. There are various course types used for races, including hurdle surfaces, turf, and dirt. Different track conditions, such as muddy, sloppy, firm, yielding, and firm, are present on each type of course.

Trakus, a commonly used method of collecting the latitude/longitude of the horse in the race every 0.25 seconds, is used to track the horses' locations. Trakus is a real-time tracking system that pinpoints each horse's precise location throughout a race, allowing it to instantly collect and send information about each horse's relative position to the leader as well as its location, speed, and distance travelled. Trakus embeds transponders in each runner's saddlecloth; an antenna at the trackside picks up the signal from the chip and determines the horse's precise location. The race's winning chances were given as it's odds.

IV. Exploratory Data Analysis

a) Data Visualisation: About 41% of the races were conducted at Aqueduct, that accounts for 825 of the total races, followed by Belmont, with 772 races, or about 38% of the total, and the third most raced location was Saratoga, with 403 races, or 20% of the total.

Shown in Fig. 1, dirt(D) races were held 1,351 times, turf(T) and inner turf(I) races held more than 200 times, while outer turf(O) races were been held 67 times. Hadle(H) has been held 9 times, accounting for less than 0.5% of the total. The type of course definitely affects the horse's speed, as well as his jockey's drafting technique. Along similar lines, after course types, prevalent track conditions are visualised in Fig. 2.

Fast track(FT) is the dominant track condition, accounting for 72% of the total. Second in line is sloppy(SY), easily one-fifths of fast track: 15%. Good(GD) and muddy(MY) make up for the remaining 11%.

So, when a horse race is held, the ground is firm 455 times, good 161 times, yielding 30 times, and soft 3 times, just 0.4% of the total. From this, it can be fairly said that track conditions are firm, fast and are conducted on dirt tracks.

We assumed a horse race as just that, but the fact that there are different race types should also be considered, because different ranks of horses participate in different races, with their respective goals. Distribution of the race types were observed. 471 of the 2000 races were claiming races, wherein the purse is bartered for the horse. Maiden special weight closely followed with 406 races, maiden claiming races with 325. Maiden races tend to be comparatively less reliable, as none of the competing mounts have won a race before, betters have to gauge the winner with their purse. These 3 accounted for more than 55% of the races. Stakes, the most prestigious races were 10% of the total. In this, the horse's owner must pay either a nomination fee. The fees paid by the owners is added to the purse money.

It is important to know that 98.95% of the races have a distance of more than 12 furlongs. This would affect the horse's stamina and finish line racing tactics.

Fig. 3, 4 and 5 show scatter plots of each of the race tracks with respect to their geometric coordinates. All of them have a straight runway path and an oval shaped track.

The succeeding subsections explore how the jockey brings out a horse's ability for the course.

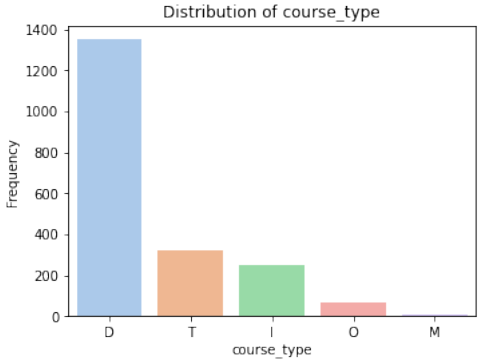

Fig. 1. Frequency of every course type.

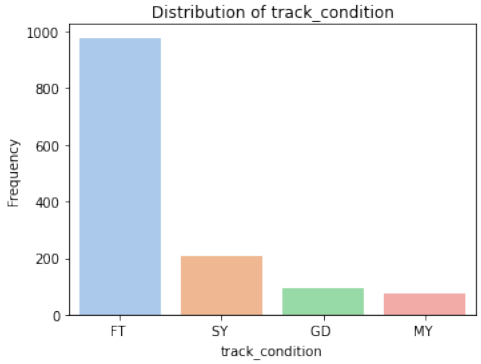

Fig. 2. Frequency of track condition.

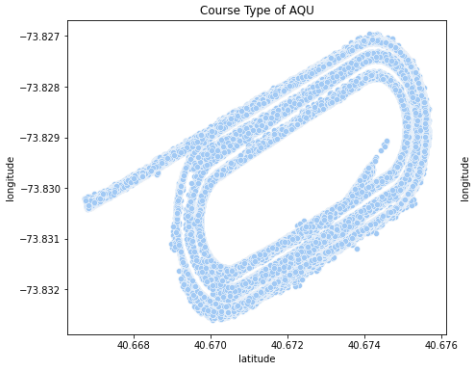

Fig. 3. Scatterplot of horse track at Aqueduct.

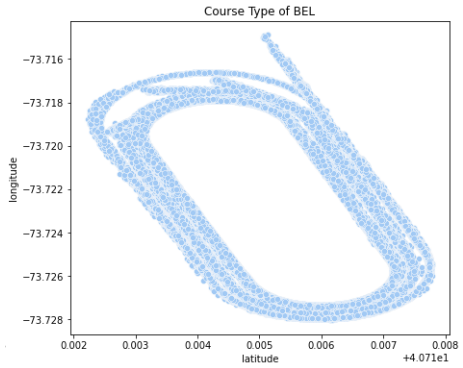

Fig. 4. Scatterplot of horse track at Belmont.

b) Analysing Horse Speed: Velocity and acceleration parameters can be computed from the geometric data. This, apart from the horses' kinetic energy, is largely governed by it's jockey. Could the rider make the horse run faster, or do the more successful jockeys ride, the better horses?

To arrive at an answer, first the geodetic coordinates specified by latitude, longitude, and elevation are transformed to geocentric Earth-Centered Earth-Fixed (ECEF) cartesian coordinates. Velocity and acceleration are obtained from the change in position information. Anomalous velocities are removed using the IQR metric. Then, one-second moving averages for the velocity and acceleration are computed. Finally, a summary statistic of the moving average of the horse's velocity and acceleration for each race is generated.

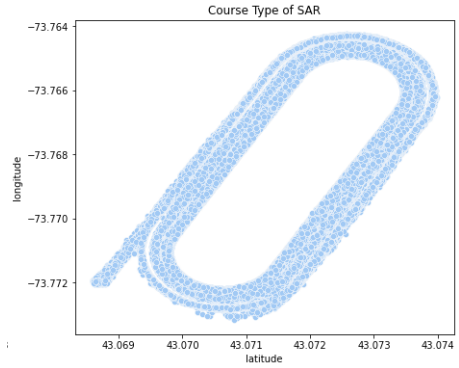

Fig. 5. Scatterplot of horse track at Saratoga.

The mean speed was found to vary at track locations. Aqueduct and Belmont had more horses with mean speed despite the fact that all three courses have the same mean speed. This may be due to the fact that more horses competed at these venues or that more horses were ridden by stronger jockeys, creating a fierce competition and a close finish line.

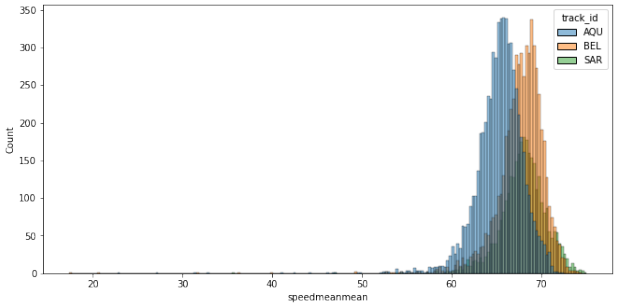

Fig. 6. Frequency distribution of mean velocity at Aqueduct, Belmont and Saratoga.

The unique movement of a rider on a horse's back "drives" a horse faster by creating kinetic energy. All racehorse jockeys ride similarly, but some jockeys are better than other jockeys at making their horse run more quickly. Monkey-crouch position, horse whip, correctness of pace, jockey's strengths are some factors that have proven to increase a horse's speed.

c) Drafting Strategies: Performance in a race is believed to be largely dependent on the pace strategy chosen and the advantages of aerodynamic drafting. Drafting is the act of following another competitor closely in order to minimise the amount of effort needed to overcome drag (pushing fluid, such as air or water, out of the path). The second environmental component, the requirement to overcome resistance when travelling through air or water, can be lessened via drafting.

An illustration of the trajectory of the winner and runner-up horses is provided in Fig. 7. The trajectory taken by the first horse is denoted by the black dots and that of the second horse is denoted by the peach dots.

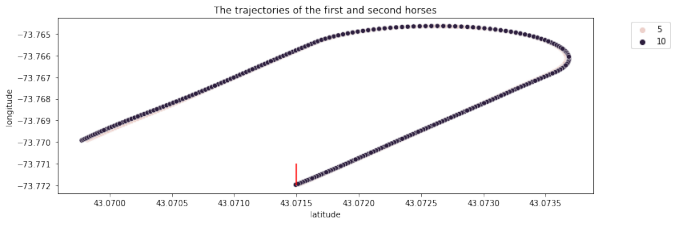

Fig. 7. Trajectory followed by the first and second horse.

A snapshot of the trajectory where the first-placed horse overtook the second-placed horse is shown is Fig. 8.

The winner horse (w1) overtook the initially leading horse (w2) by moving slightly to the outside of the track and

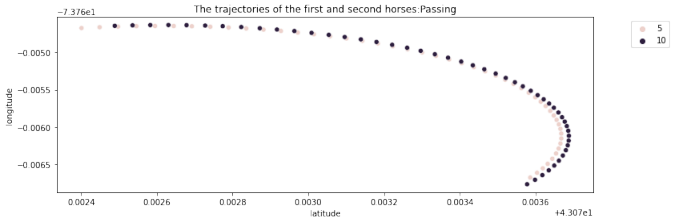

Fig. 8. Close up of first horse overtaking the second.

overtaking the w2 in second place from the outside. Fig. 9. depicts how the race ultimately ended. The red line is the finish line.

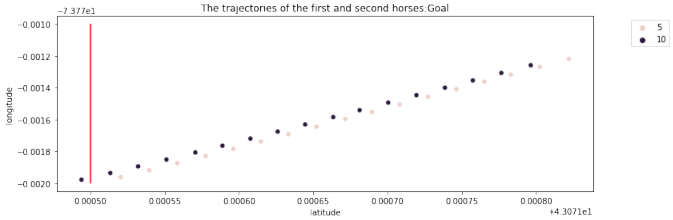

Fig. 9. Close up of first horse overtaking the second.

V. Plan for Future Weeks

We intend to put into practice a regression model that we researched in order to examine the relationship between important characteristics of the horses, their jockeys and it's effect on final positions. The correlation and statistical significance of attributes will also be studied.

Acknowledgment

We would like to acknowledge our Data Analytics Course Professor Dr. Gowri Srinivasa for providing feedback for choosing the title of our project. We would also like to acknowledge the teaching assistants who have been constantly providing resources to practice the learnt concepts.

References

[1] Little, Todd D., ed. The Oxford handbook of quantitative methods in psychology: Vol. 2: statistical analysis. Vol. 2. Oxford University Press, 2013.

[2] Gardner, David S. "Historical progression of racing performance in thoroughbreds and man." Equine veterinary journal 38, no. 6 (2006): 581.

[3] Pudaruth, Sameerchand, Nishchay Nemdharry, Trivartsingh Ramjeawon Harrykesh Ramma, and Ritesh Mungroo. "Impact of the variation of horse racing odds on the outcome of horse races at the Champ de Mars." In 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-4. IEEE, 2017.

[4] Spence, Andrew J., Andrew S. Thurman, Michael J. Maher, and Alan M. Wilson. "Speed, pacing strategy and aerodynamic drafting in Thoroughbred horse racing." Biology letters 8, no. 4 (2012): 678-681.