Analysis and Odds Prediction 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. Unsupervised clustering techniques(K Means Clustering and Agglomerative Clustering) are employed to undertstand the patterns in data, followed by modelling the odds of horse winning using multiple linear regression(MLR), and a neural network.**

***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.

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. Approximate entropy is used to quantify the unpredictability [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 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 explore the dataset, deduce reasoning for numerous distributions of the attributes, find patterns within it, and finally predict the odds of a horse winning.

The paper is organized into six sections. Following this introduction, Section 2 describes the related work, Section 3 provides an overview of the dataset. We put forth the explorartory data analysis of 3 tables - racing, track and start files in Section 4. Section 5 describes the methodology in preprocessing, and modelling. Contribution of our team is explained in Section 6. Peer review is contained in Section 7, followed by acknowledgements and a list of references.

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. A fast, dirt track is ideal for racing, while a muddy track has more moisture, while the track condition after rain is a sloppy track. Horses may race on the outer boundary of the track, called outer turf, or inside the boundary, called inner turf.

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

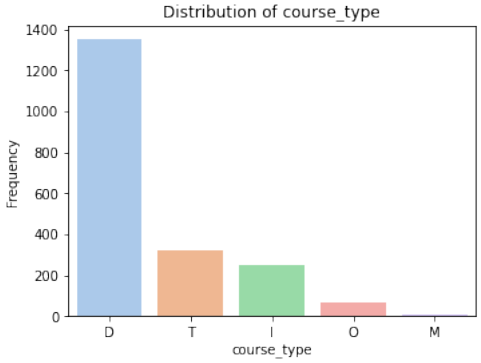

Fig. 1. Frequency of every course type.

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.

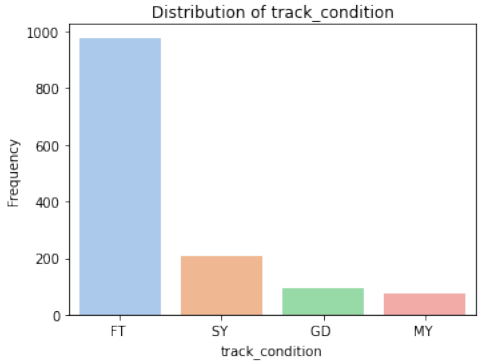

Fig. 2. Frequency of track condition.

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.

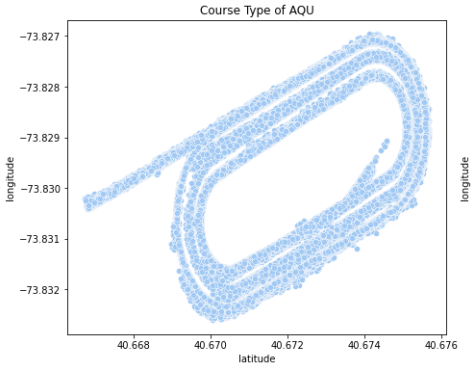

Fig. 3. Scatterplot of horse track at Aqueduct.

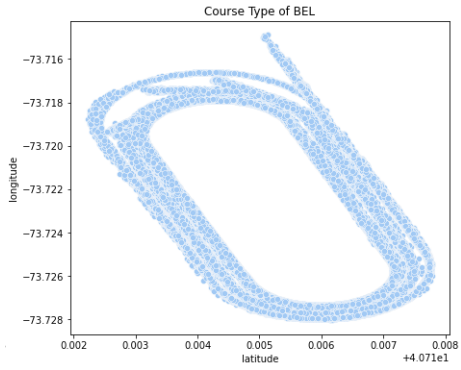

Fig. 4. Scatterplot of horse track at Belmont.

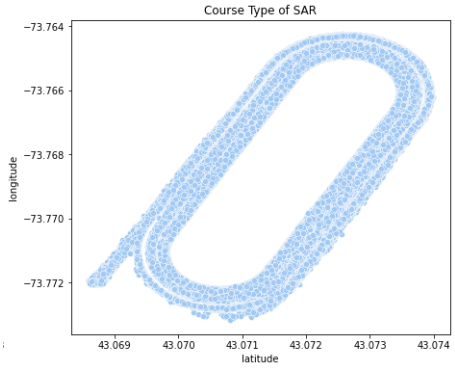

Fig. 5. Scatterplot of horse track at Saratoga.

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

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.
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.

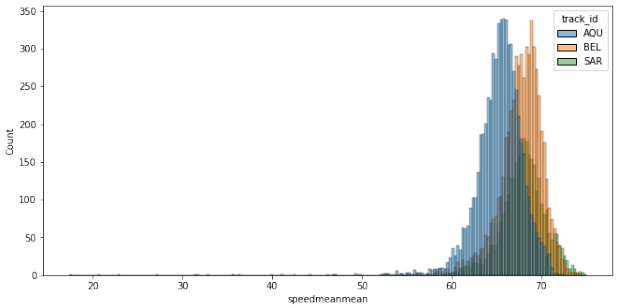

Fig. 6. Frequency distribution of mean velocity at Aqueduct, Belmont and Saratoga.

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.

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

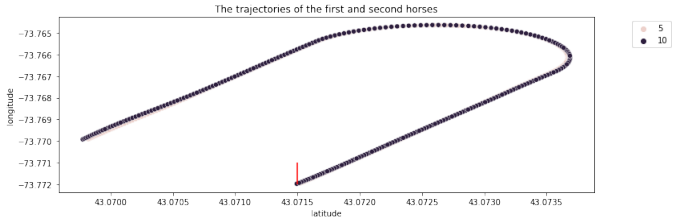

Fig. 7. Trajectory followed by the first and second horse.

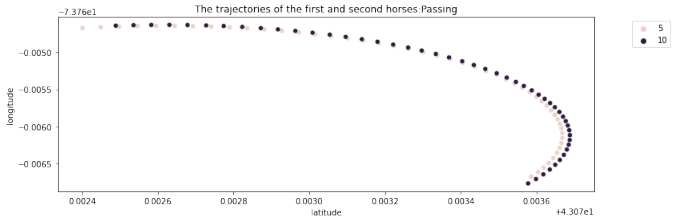

Fig. 8. Close up of first horse overtaking the second.

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

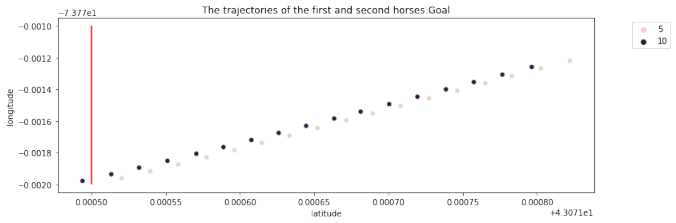

Fig. 9. Close up of first horse overtaking the second.

V. Methodology

Data Preprocessing

We discarded the duplicate rows and dropped some unnecessary columns from the dataset since they were not relevant to the target 'odds'. The category fields, like "race date," just provide timestamps and no other information. Similarly, the 'trakus index', 'latitude', 'longitude', and 'distance id' only offer us information on the horse's geographic location, making them less useful. Following data cleaning, categorical attributes like "track id," "program number," "course type," "track condition," "race type," and "jockey" were converted to numeric values for our models' convenience.

We used MinMaxScaler for categorical data and StandardScaler for continuous data to standardise the impact of each factor on the target variable (in our instance odds). The cleaned and standardized dataset can be seen in Fig 10.

Lastly, we created a correlation heatmap (Fig 11.) of all the features to assess the correlation between attributes and the target, and we observed that the features "course type" - "run up distance" and "purse" - "race type" had higher correlations than the others (0.43 and 0.42, respectively).

	track_id	race_number	program_number	course_type	track_condition	run_up_distance	race_type	purse	post_time	weight_carried	jockey	odds
3626025	0.5	0.750000	0.885714	1.00	0.333333	0.626613	1.000000	-0.373417	0.365775	0.153722	0.785311	-0.343734
2387002	0.0	0.416667	0.914286	0.00	0.333333	-0.401470	0.272727	-0.425871	-0.565106	0.436405	0.237288	-0.487177
3688499	0.0	0.250000	0.857143	0.75	0.333333	-0.506690	0.363636	-0.119889	-0.693377	-0.411644	0.435028	-0.415455
2682742	1.0	0.083333	0.914286	1.00	0.000000	-1.213116	0.181818	-0.277251	-1.133164	-0.128961	0.672316	1.751568
4367312	0.5	0.500000	0.742857	1.00	0.000000	1.113601	0.363636	-0.076177	-0.000714	-0.411644	0.785311	0.050736
2644227	0.5	0.583333	0.057143	0.25	0.000000	-0.022703	0.818182	-0.251024	0.109233	-0.694327	0.429379	-0.031231
5037433	0.0	0.166667	0.714286	1.00	0.333333	0.518394	0.727273	0.142382	-0.667723	-0.694327	0.672316	-0.625498
3212750	0.5	0.166667	0.828571	0.00	0.166667	-0.563799	0.000000	-0.058692	-0.026368	-0.128961	0.372881	-0.410332
1536606	1.0	0.500000	0.514286	0.25	0.000000	0.680723	0.090909	-0.006238	0.105568	0.436405	0.598870	1.559456
282806	0.0	0.166667	0.028571	0.00	0.333333	-0.590854	0.000000	-0.119889	-0.631074	0.436405	0.542373	-0.666482

Fig. 10. Cleaned data

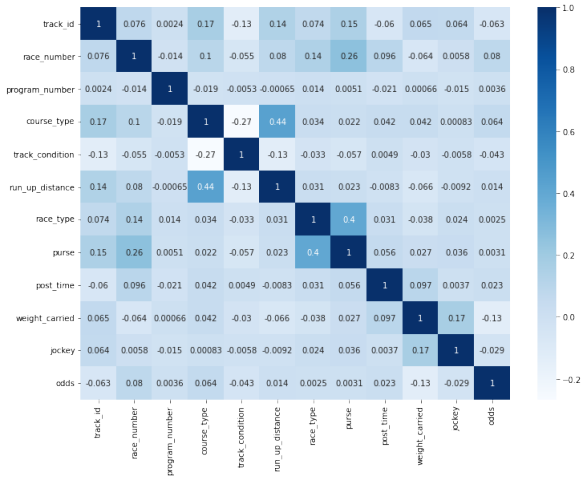

Fig. 11. Histogram showing correlation between features

Clustering

In order to find and understand any patterns in the data, unsupervised learning-clustering is implemented. To analyse which method clusters the data points the best, two methods—K-means and agglomerative clustering—were implemented.

Data values are clustered using K-means, an unsupervised learning technique, based on how similar they are to one another. Prior to the process, the number of clusters must be determined. This is not required in agglomerative clustering which uses a bottom-up approach, wherein each data point starts in its own cluster. Then, these clusters are greedily connected by merging the two clusters that are the most similar to one another. While the agglomerative graph in Fig. 30 had a good knee shape at k equal to 6, the k-means graph in Fig. 29 was a bit oscillatory until k equal to 5. Consequently, N=6 was chosen as the number of clusters in agglomerative clustering, and the cluster labels were generated for each data point.

Fig. 12. K-means elbow plot of score versus k ranging from 2 to 9.

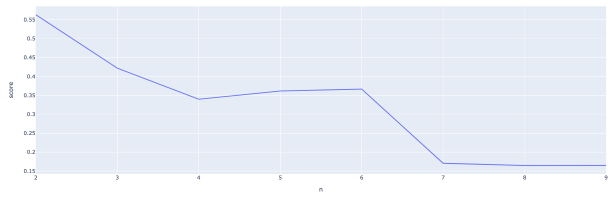

Fig. 13. Agglomerative knee plot of score versus k ranging from 2 to 9.

Metric	Value
RMSE	0.59646116
MAE	0.55489073

TABLE I
RMSE AND MAE VALUES FOR MLR MODEL

Pearson correlation coefficient between these labels and the target variable, odds, is equal to 0.501. It's interesting to note that this correlation is the highest between any two variables, which is why these labels are used as an additional attribute to the predict the odds.

Modeling

Before modeling, PCA is performed on the preprocessed data, as exactly 5 features explained 90% of variance in the dataset. By downsizing a huge amount of variables while retaining the majority of the information, PCA is used to reduce the dimensionality of datasets. 15081 rows in total are divided into training (11310 rows) and test (3771 rows).

Multiple Linear Regression: It enables the creation of a linear relationship between numerous independent factors and a single dependent variable. This method produced an experimental R^2 value of 0.371, RMSE value of 0.596 and MAE value of 0.5548 as seen in Table I. The comparison between the anticipated odds value and the true value is shown in Fig. 14. Other forms of representation did not produce a comprehensible pattern than the one shown in the figure.

Fig. 15 is a plot of the r-squared values for number of features leading up to 13. After five features, the values stabilized around 0.37, which is consistent with the PCA-based feature selection.

This model underwent 5-fold cross-validation; Table II shows the average and standard deviation for the 5-folds.

Metric	Mean	Standard Deviation
R^2	0.3685	0.0315
Negative MSE	-0.6303	0.053

TABLE II
5 FOLD CROSS VALIDATION SCORES - MLR

This implies that there isn't much variation in the outcomes between test sets. It confirms that the performance of the model is approximately 0.37 (R^2), which isn't very well, therefore we explore and choose a neural network.

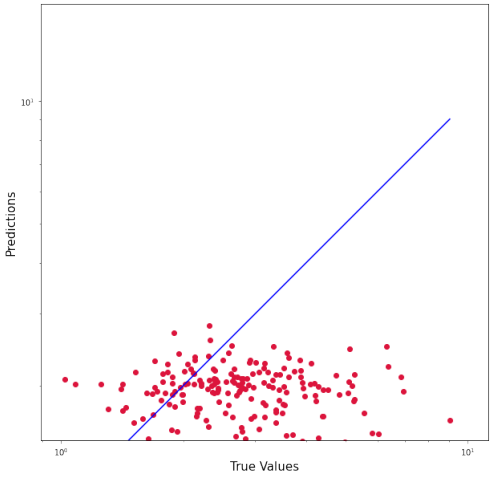

Fig. 14. Scatterplot of log(Predicted values) against log(True values)

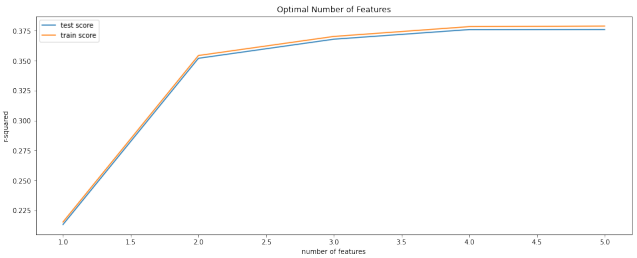

Fig. 15. R-squared for K-fold MLR

1) DL Neural Network: The regressive nature of this problem led to the implementation of a 5-layer deep learning neural network with linear and relu activation functions. The neural network's final layer contains one neuron that is a continuous numerical value, the odds. An Adam optimizer is used to iteratively tune the network weights based on training data during the model's 200-epoch, 200-batch, 0.0001 learning rate training.

As illustrated in Figs. 16, 17, and 18, the MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and MAE (mean absolute error) decrease continually, insinuating that the model is learning well. The average MSE loss, RMSE and MAE for deep learning model is inferred from Table III.

Metric	Value
Loss	0.56776154
RMSE	0.52132695
MAE	0.52132695

TABLE III
LOSS, MEAN, RMSE FOR NEURAL NETWORK

The neural network has lower RMSE and MAE values than MLR (0.0751 and 0.03356, respectively). Therefore, we can conclude that the deep neural network outperformed the MLR model by a small margin. This is may be because deep neural networks use activation functions to account for data non-linearity.

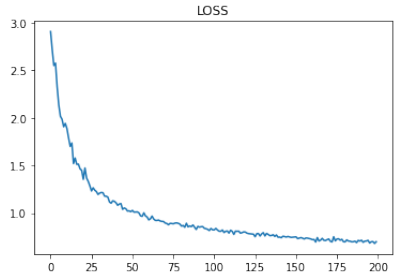

Fig. 16. Loss for Deep Neural Network

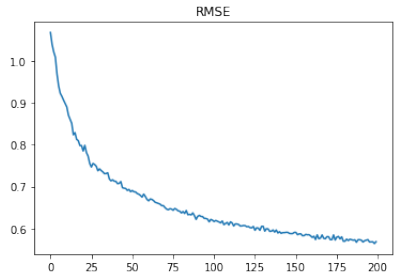

Fig. 17. RMSE for Deep Neural Network

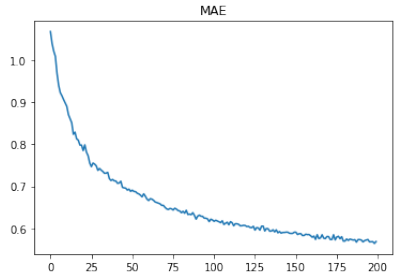

Fig. 18. MAE for Deep Neural Network

VI. Contribution of Team

The project work was done together on calls and in person. So, we contributed equally to it. We inferred the dataset, and provided an overview of it, visualised the 3 tables, with distributions of track locations, course type, track condition, weight carried and other factors. Scatterplots of racing tracks, average horse speed versus odds. Analysed the velocity and acceleration of horses at the three track locations. Visualised a glimpse of the course taken by a winning horse at it's final race moments. This lead to a thorough reading of the impact of a jockey to it's horse. The many facets of winning, more than just the speed of the horse - the determination of it's rider, drafting strategies and race types.

We performed preprocessing, clustering, and modelling using MLR, neural network in phase 2 of the project. Experimental results and inference. Feature selection, data scaling and standardisation, PCA. Implemented and inferred K Means Clustering an Agglomerative Clustering, elbow plots of both. We studied and summarized 4 papers as a literature survey, and drafted the report.

VII. Peer Review

Our peer review team suggested us to understand the math behind clustering. We played around by taking from-scratch clustering functions and feeding it the same preprocessed data. Luckily it was taught around the same time in class as part of Unit 4 so we were able to solve and see plots of simple clustering questions.

They also recommended that we justify our decision to use a certain prediction model. We had done clustering initially, around the time of this peer review, and then set allocated clusters as the target. KNN, Random Forest, and SVM were utilised to evaluate this target as a classification problem. We examined this and concluded that it wasn't really helpful because clustering horses based on the races they participated in doesn't produce any beneficial outcome. Predicting a horse's likelihood of winning, which is a continuous variable, would be more helpful. In order to transfer the problem statement to a regressive kind, odds must be predicted.

Since the clusters and odds were correlated, they were added as a feature. We utilised MLR and a simplistic deep learning neural network because we saw that there were many independent variables and one target variable as a result of them. Despite the fact that not all variability could be predicted by the model, the latter strategy was only marginally better than MLR.

Conclusion

This study examined the New York Racing Association's (Big Data Derby) dataset on horse racing with the goal of estimating the likelihood that a specific horse will win a race using multiple linear regression and a simplistic neural network. The latter performed slightly better in terms of having lower RMSE and MAE measures. Although the exact cause of this very slight difference remains uncertain, we hypothesized that it was because the neural network, as opposed to MLR, could take into consideration non-linear patterns in the data. To the best of our knowledge, there aren't many studies that discuss horse racing in the context of artificial intelligence and machine learning. One investigation looked at the impact of changing horse racing odds on outcomes of races [4].

Given that each horse's latitude and longitude have been provided at intervals of 0.25 seconds, there is a wide range of analysis that can be done on this dataset. When compared to the odds, this might be used to predict the horse's place in the final laps of a race and analyse variations between real and odds. The relative relevance of drafting can also be determined. In a very traditional business, this might lead to novel approaches to training and racing. One could contribute to bettering equine welfare, performance, and rider decision-making by making better use of horse tracking data.

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.