DS502
Project Proposal

Sidhanth Subhash Jain, Chandra Rachabathuni, Darshan Swami

Problem Description

Injury prevention is essential in modern athletics. Animal-related sports, such as horse racing, are no different than human sports. Typically, movement efficiency correlates to both improved performance and injury prevention. We will build a model to interpret one aspect of this new data in this project. We could use the data to analyze jockey decision making, compare race surfaces, or determine the relative importance of drafting. The project will assist racing horse owners, trainers, and veterinarians in better understanding the relationship between equine performance and welfare. Equine welfare could improve significantly with better data analysis.

Data Description

The data available is described in the following excerpt from the Kaggle page:
A wealth of data is now collected, including measures for heart rate, EKG, longitudinal movement, dorsal/ventral movement, medial/lateral deviation, total power and total landing vibration.
The data is stored in separate csv files before being merged into a single csv file, as shown below.

- nyra_start_table.csv - horse/jockey race data
- nyra_race_table.csv - racetrack race data
- nyra_tracking_table.csv - tracking data
- nyra_2019_complete.csv - combined table of three above files

nyra_start_table.csv

1. track_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
2. race_date - date the race took place. YYYY-MM-DD.
3. race_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
4. program_number - Program number of the horse in the race passed as 3 characters. Should remain 3 characters as it isn't limited to just numbers. Is essentially the unique identifier of the horse in the race.
5. weight_carried - An integer of the weight carried by the horse in the race.
6. jockey - Name of the jockey on the horse in the race. 50 character max.
7. odds - Odds to win the race passed as an integer. Divide by 100 to derive the odds to 1. Example - 1280 would be 12.8-1.
8. position_at_finish - An integer of the horse's finishing position. (added to the dataset 9/8/22)

nyra_race_table.csv

1. track_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
2. race_date - date the race took place. YYYY-MM-DD.

3. race_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
4. distance_id - Distance of the race in furlongs passed as an integer. Example - 600 would be 6 furlongs.
5. course_type - The course the race was run over passed as one character. M - Hurdle, D - Dirt, O - Outer turf, I - Inner turf, T - turf.
6. track_condition - The condition of the course the race was run on passed as three characters. YL - Yielding, FM - Firm, SY - Sloppy, GD - Good, FT - Fast, MY - Muddy, SF - Soft.
7. run_up_distance - Distance in feet of the gate to the start of the race passed as an integer.
8. race_type - The classification of the race passed as as five characters. STK - Stakes, WCL - Waiver Claiming, WMC - Waiver Maiden Claiming, SST - Starter Stakes, SHP - Starter Handicap, CLM - Claiming, STR - Starter Allowance, AOC - Allowance Optionl Claimer, SOC - Starter Optional Claimer, MCL - Maiden Claiming, ALW - Allowance, MSW - Maiden Special Weight.
9. purse - Purse in US dollars of the race passed as an money with two decimal places.
10. post_time - Time of day the race began passed as 5 character. Example - 01220 would be 12:20.

nyra_tracking_table.csv

1. track_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
2. race_date - date the race took place. YYYY-MM-DD.
3. race_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
4. program_number - Program number of the horse in the race passed as 3 characters. Should remain 3 characters as it isn't limited to just numbers. Is essentially the unique identifier of the horse in the race.
5. trakus_index - The common collection of point of the lat / long of the horse in the race passed as an integer. From what we can tell, it's collected every 0.25 seconds.
6. latitude - The latitude of the horse in the race passed as a float.
7. longitude - The longitude of the horse in the race passed as a float.

nyra_2019_complete.csv - This file is the combined 3 files into one table. The keys to join them trakus with race - track_id, race_date, race_number. To join trakus with start - track_id, race_date, race_number, program_number.

1. track_id - char(3)
2. race_date - date
3. race_number - char(3)
4. program_number - char(3)
5. trakus_index - int
6. latitude - float

7. longitude - float
8. distance_id - int
9. course_type - char(1)
10. track_condition - char(3)
11. run_up_distance - int
12. race_type - char(5)
13. post_time - char(5)
14. weight_carried - int
15. jockey - char(50)
16. odds - int
17. position_at_finish - An integer of the horse's finishing position.

Prediction type

On this data, two types of predictions can be made: regression and classification.
Classification can be used to predict whether or not a jockey won a race by engineering a binary target based on position at finish (an integer representing the horse's finishing position). Clustering is used to group specific horses together and perform classification to predict whether there will be horse injury, which is graded from "0" to "5".
Regression can be used to forecast a horse's odds (Odds are the return you can expect to get if the horse you bet on is successful)

Methods

Before we start building the model, we'll try to convert the categorical attributes to ordinal attributes. For example, the jockey name must be one-hot encoded because it represents the horse's name, and there are other variables that must be encoded.
We would also do feature engineering and feature selection before feeding into the model to simplify the model. We would attempt to do these feature selections:

- **Feature selection with correlation** - find out which features are correlated and then drop all except one.
- **Univariate feature selection** - Univariate feature selection works by selecting the best features based on univariate statistical tests.
- **Recursive feature elimination** - Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.

We will conduct preliminary statistical analysis on the data to gain insights into various features and to try to answer some questions posed by the data. Following the preliminary analysis, we will perform some data preprocessing and attempt to reduce dimension using PCA. In addition, we will classify the chances of horse injury by clustering on the pre-processed data and using the cluster centroids to predict the chances of horse injury.
We will then train various classification models to see which model provides the best accuracy for the data.
We can also use the combined data to predict the odds of a horse winning by applying a regression model on top of it; this would allow bookmakers to take calculated risks on their bets.

Algorithms and Error Metrics

For classification:

1. SVM (Support Vector Machine)
2. Logistic Regression
3. KNN
4. XGBoost
5. Random Forests

For Clustering:

1. KMeans

For Regression:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression

Error Metrics:

1. RMSE for regression
2. ROC curve and F1 score for multi-class classification

Comments and Concerns

The Big Data Derby 2022 data has a unique structure that provides us with unique insights and challenges. Rather than each row representing an observation, collections of rows are distinguished by a 'Track ID.' represent the racetrack. There are numerous intriguing preprocessing techniques for transforming this data into something that can be modeled. There are numerous exploration topics in the data such as horse injury and odds predictions that can motivate our models and insights about this unique data, such as horses being affected by weather conditions, which is not present in the current dataset.
The data type of the program number feature is incorrect. There are also spaces. This is how the data appears as '6 '.
The current data set contains no information on:
a) the horse's unique identity (only its program number in an individual race)
b) its final position