|  |
| --- |
| Grandview University |
| Big Data Derby: Project Proposal |
| Christian Webster  September 9, 2022 |
| **ABSTRACT**  One of the most important aspects of sports racing is strategy. With a big enough collection of data, it would be possible to create analytical models that give new insights in tactics, drafting strategies, and path efficiency. The goal of this project will be to analyze different horse racing strategies, which may be done through analyzing the latitude and longitude of horses during races. These conclusions not only allow the creation of new racing strategies, but also may help horse veterinarians improve horse performance and welfare. |
|  |

|  |
| --- |
|  |

**TYPE OF PROJECT**

This project will attempt to construct a model that can help analyze different horse racing strategies. This may also help improve care methods by veterinarians in order to improve horse performance and welfare.

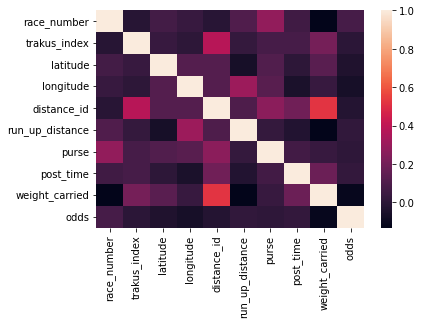
**DATA DIMENSIONALITY**

This dataset consists of 17 variables and 5,228,430 observations. The target variable would be “odds,” which is the odds to win the race (expressed as an integer). To get the odds-to-one, you would divide by 100, so a table value of 1300 would be odds of 13-to-1. Below are our initial variables and their descriptions:

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Description |
| Track\_id | Char | 3-character id for the track the race took place at  AQU – Aqueduct, BEL – Belmont, SAR - Saratoga |
| Race\_date | Date | Date the race took place. YYYY-MM-DD |
| Race\_number | Char | Number of the race |
| Program\_number | Char | Program number of the horse in the race |
| Trakus\_index | Int | Common collection point of lat/long of the horse |
| Latitude | Float | Latitude of the horse |
| Longitude | Float | Longitude of the horse |
| Distance\_id | Int | Distance of the race in furlongs (1/8th mile) |
| Course\_type | Char | Type of course the race was run  M – Hurdle, D – Dirt, O – Outer Turf, I – Inner Turf, T - Turf |
| Track\_condition | Char | Condition of the course the race was run on |
| Run\_up\_distance | Int | Distance in feet to the gate at the start of a race |
| Race\_type | Char | Classification of the race |
| Post\_time | Char | Time of day the race began |
| Weight\_carried | Int | Amount of weight carried by the horse |
| Jockey | Char | Name of the jockey in the race |
| Position\_at\_finish | Int | Horses finishing position |
| Odds | int | Odds to win the race (target variable) |

A picture containing text

Description automatically generated**EXPLORATORY ANALYSIS**

***Figure 1****.* Checking for null values in our dataset

***Figure 2****.* Correlation heatmap of our variables. We can see that the highest correlation with odds is weight carried by far. This makes sense because if the horse is carrying less weight, they can run faster and expend less energy.

**Chart, bar chart

Description automatically generated*Figure 3.*** General bar plot of our track\_id. We can see that the majority of our races take place on AQU, or Aqueduct.

Chart, bar chart

Description automatically generated***Figure 4.*** General bar plot of our track\_condition, We can see that the majority of our race take place when the condition is FT, or Fast.

**DATA SOURCE**

All data was acquired from *Kaggle.com,* provided by the currently running competition “Big Data Derby.”

**IMPORTANCE OF THIS PROJECT**

As strategy is the most important aspect of any sport, determining new strategies based on existing data for horse racing is of vital importance to stakeholders of horse racing teams. Through the analysis of this data, new racing strategies may be born, along with new horse care methods.

**RESEARCH PLAN**

|  |  |  |
| --- | --- | --- |
| Date | Objective | Person(s) in charge |
| Sept 2  Sept 9  Sept 9 – Oct 30  Nov 1-11  Nov 14 – 21  Nov 21 - 28  Dec. 5 | Initial Meeting with Oscar  Creation of Project Proposal  Analysis of Data  Creation of Preliminary Project Report  Finalization of Code and Presentation  Finalization of Project Report and Presentation  Submission of Code and Report | Christian  Christian  Christian  Christian  Christian  Christian  Christian |

**REFERENCES**

https://www.kaggle.com/competitions/big-data-derby-2022/overview