Big Data Derby 2022

Analysis of Horse Racing Data and Understanding Factors that Affect Position

Applied Predictive Analytics – George Mason University

Syed Abdul Hadi

Introduction and Objective:

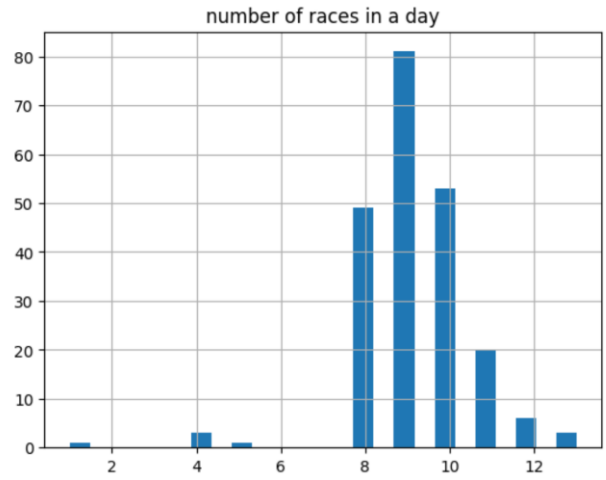
Strategy and decision making based on game environment is a critical component in modern athletics. Sports that involve animals, such as horse racing, are no different than human sport. Typically, efficiency in movement correlates to both improvements in performance and injury prevention. A wealth of 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 motivation of this project is to decipher what makes the most positive impact. The goal is to create a model that makes use of race data, field conditions and jockey information to identify what impacts the final position of a horse in a race and how successfully can this outcome be modelled.

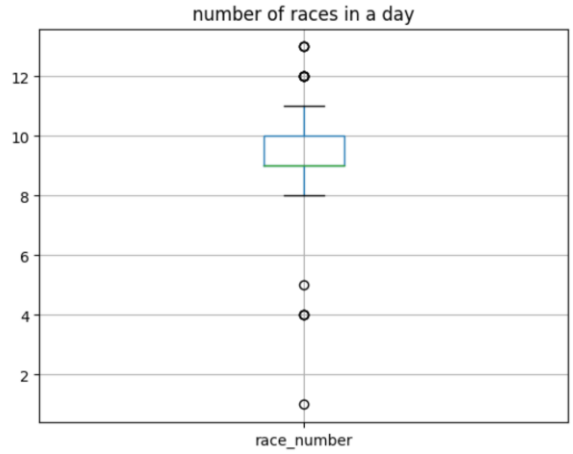
The dataset being used has recently been made available by the New York Racing Association (NYRA) and the New York Thoroughbred Horsemen's Association (NYTHA). It is the first to provide researchers with X/Y coordinate mapping of horses during races. The dataset provides information that can be used to analyze jockey decision making, compare race surfaces, or measure the relative importance of drafting. With considerable data, it is possible to explore new ways of racing and training in a highly traditional industry. With improved use of horse tracking data, this project aims to improve equine welfare, performance and rider decision making.

Exploratory Analysis:

Races Per Day:

There are between 1 and 14 races on any of the race days. This does not include days on which there were no races.



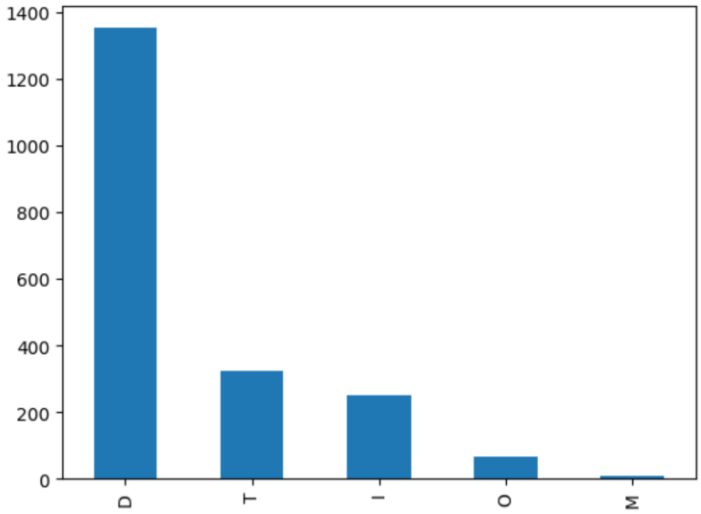


Outliers were not removed from this column because races and outcomes are independent of each other.

Course Type Analysis:

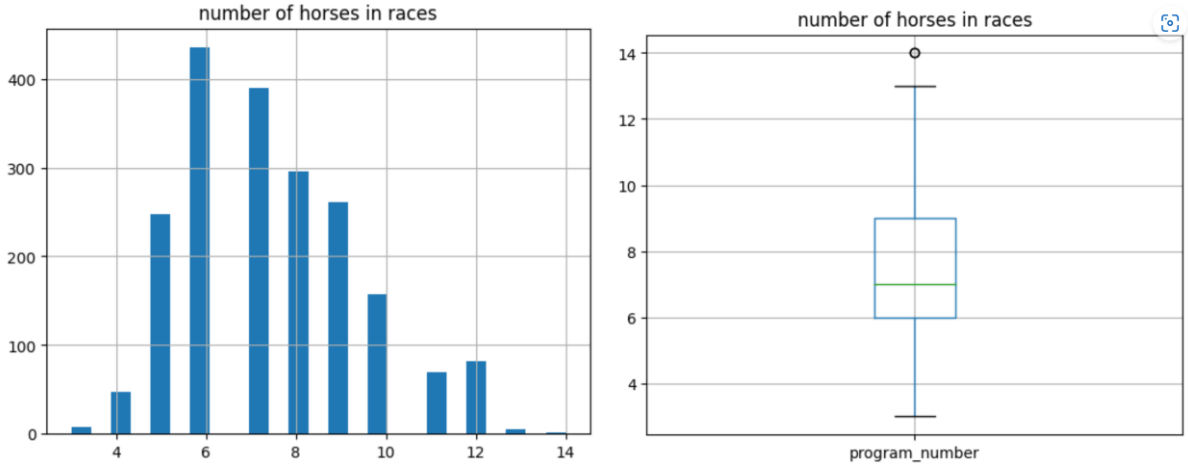
There are the following four course types in the dataset.

M - Hurdle, D - Dirt, O - Outer turf, I - Inner turf, T - turf.



Course type ‘M’ (Hurdle) has been removed from the dataset because it is a fundamentally different category of courses and it won’t be representative of performance in other courses.

Horse Analysis:

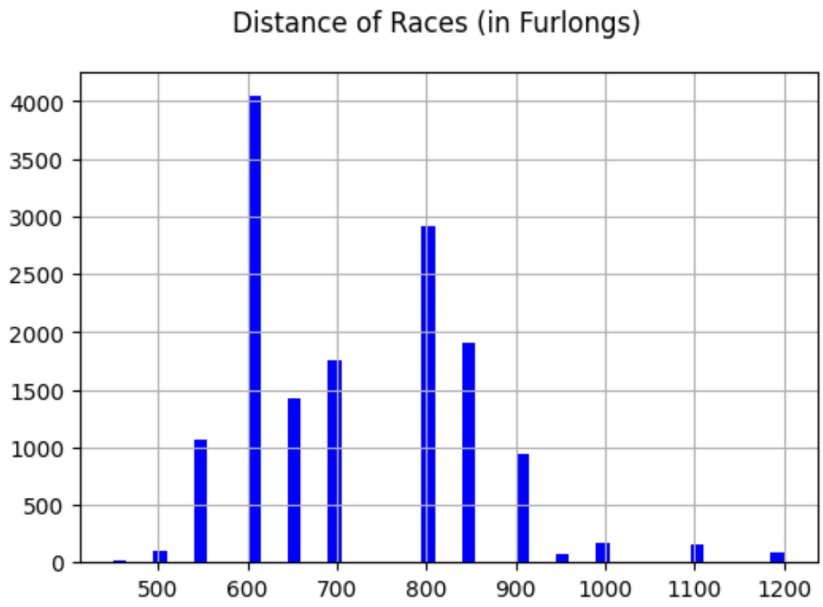


There are between 3 and 14 horses per race, races with less than 5 and more than 12 horses have been removed because the final position of any horse is determined by other horses in the race as well, too many or too few will inflate or discount the performance of individual horses.

Race Distance Analysis:

Race distance is measured in furlongs. Distributed as follows:

1 furlong = 220 yards

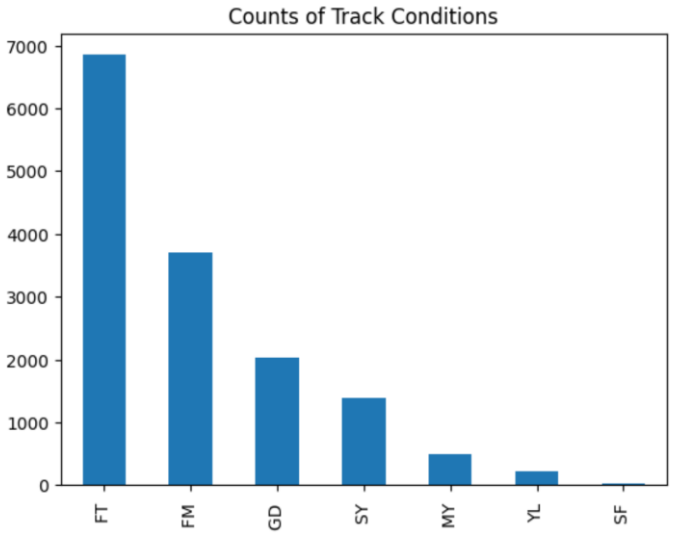


Races with distance greater than 1200 furlongs have been removed because we do not have enough data on whether performance diminishes as race distance increases.

Track Conditions Analysis:

The track conditions are distributed as follows. Encoded as:

YL - Yielding, FM - Firm, SY - Sloppy, GD - Good, FT - Fast, MY - Muddy, SF - Soft.

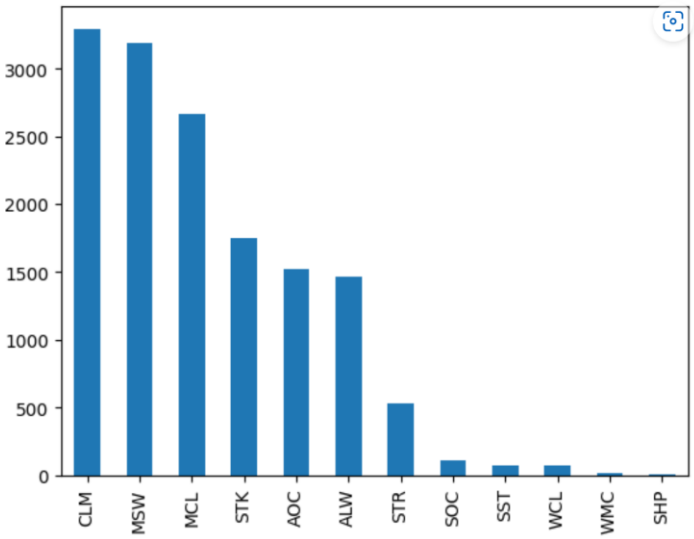


Tracks with 'soft' condition (coded as SF) have been removed because we have very little data on performance on the soft track.

Race Type Analysis:

The following race types are found within the data, encoded as:

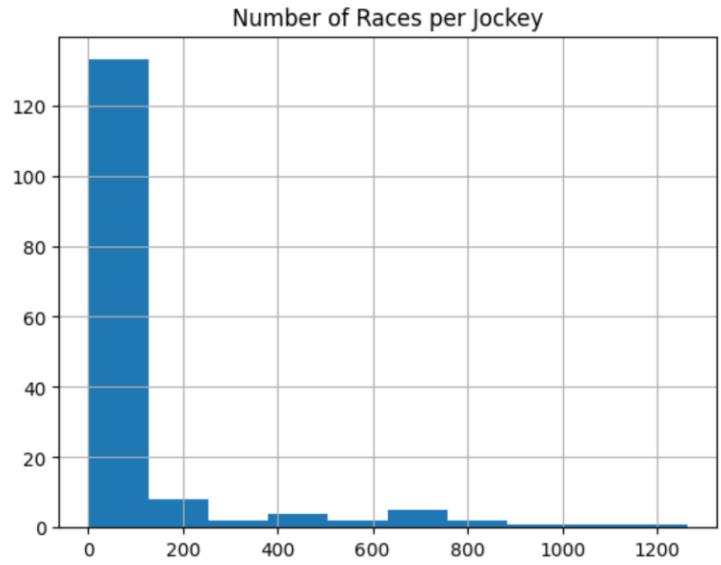
STK - Stakes, WCL - Waiver Claiming, WMC - Waiver Maiden Claiming, SST - Starter Stakes, SHP - Starter Handicap, CLM - Claiming, STR - Starter Allowance, AOC - Allowance Optionl Claimer, SOC - Starter Optional Claimer, MCL - Maiden Claiming, ALW - Allowance, MSW - Maiden Special Weight.



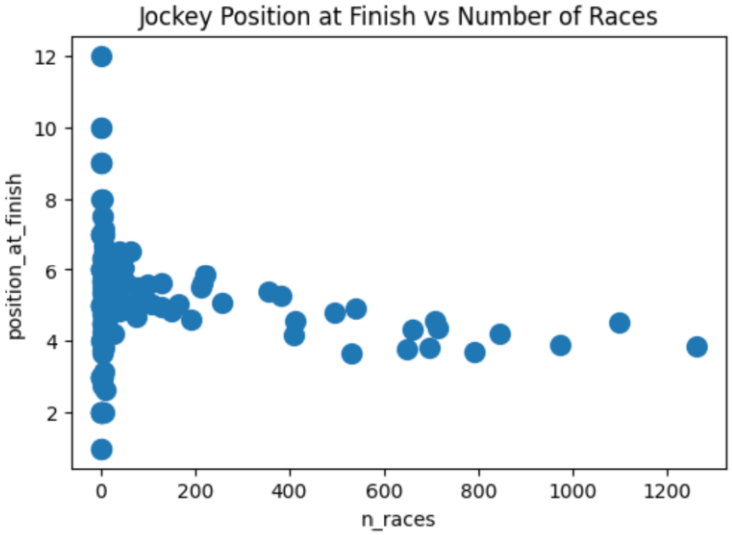
SHP and WMC are removed these are highly different race types and there isn't enough data to model horse performance in these races.

Jockey Analysis:

There are 159 jockeys in the dataset and they have between 1 and 1263 races recorded. It has been observed that 39 jockeys have only 1 race, 73 jockeys have up to 3 races and 44 jockeys have more than 20 races.



The performance of jockeys relative to their number of races can be observed below:



It can be observed that there is a downward trend i.e., increase in performance as jockey experience on a track goes up. Number of races up until a race date can be used to obtain a better position at finish when modeling.

Feature Engineering:

Places refers to holding a place within top 3 in a race. The following 6 features have been created using the data available:

1. Cumulative Jockey Races
2. Cumulative Jockey Wins
3. Cumulative Jockey Places
4. Cumulative Horse Races
5. Cumulative Horse Wins
6. Cumulative Horse Places

Each of these figures, once calculated, are shifted down by 1 index in a group wise manner (grouped over jockey/horse by race). This is done to ensure that if a horse (or jockey) won a race, their cumulative statistics do not represent this win in the same row that may be used to train the model.

Outcome Engineering:

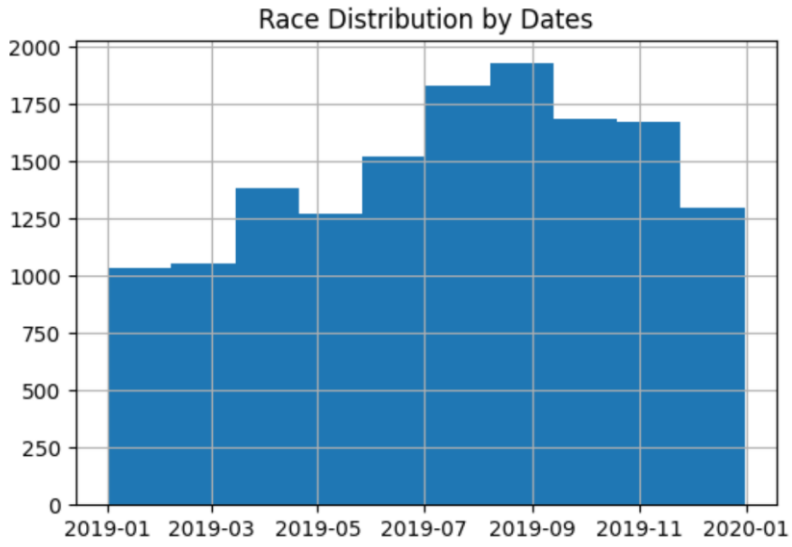
The target variable (label), has been engineered as follows:

1 : if position is 1^{st}.

0 : if position is not 1^{st}

Train Test Split:

Data is split into train and test. This is based on dates – races before '2019-09-01' are part of training data, races after this are part of test data.



The outcome variables has the following distribution in the training and test data:

Train:

0 0.862994
1 0.137006

Test:

0 0.870642
1 0.129358

Logistic Regression Results:

Optimization terminated successfully.
Current function value: 0.345724
Iterations 9

Results: Logit

Model:	Logit			Pseudo R-squared:	0.124	
Dependent Variable:	label			AIC:	10210.1618	
Date:	2022-11-28 13:27			BIC:	10468.3382	
No. Observations:	14668			Log-Likelihood:	-5071.1	
Df Model:	33			LL-Null:	-5788.3	
Df Residuals:	14634			LLR p-value:	1.2332e-280	
Converged:	1.0000			Scale:	1.0000	
No. Iterations:	9.0000					

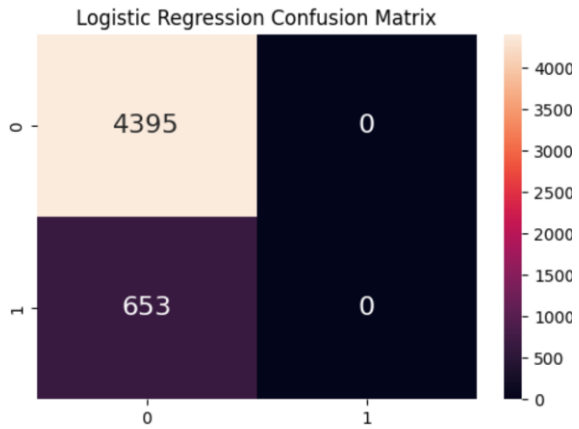
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
weight_carried	0.0192	0.0103	1.8559	0.0635	-0.0011	0.0394
odds	-0.0014	0.0001	-22.7521	0.0000	-0.0015	-0.0013
distance_id	-0.0000	0.0002	-0.0017	0.9986	-0.0004	0.0004
run_up_distance	-0.0003	0.0009	-0.3597	0.7190	-0.0021	0.0014
purse	-0.0000	0.0000	-0.2004	0.8411	-0.0000	0.0000
post_time	-0.0000	0.0001	-0.0168	0.9866	-0.0002	0.0002
horse_id	0.0000	0.0000	1.0732	0.2832	-0.0000	0.0001
track_id=AQU	-1.2202	2565631.9532	-0.0000	1.0000	-5028547.4460	5028545.0056
track_id=BEL	-1.2217	2560091.7862	-0.0000	1.0000	-5017688.9197	5017686.4764
track_id=SAR	-1.2729	2557334.8916	-0.0000	1.0000	-5012285.5569	5012283.0110
course_type=D	-0.8635	1461613.1361	-0.0000	1.0000	-2864709.9696	2864708.2426
course_type=I	-0.9344	1463818.1711	-0.0000	1.0000	-2869031.8298	2869029.9609
course_type=O	-0.9776	1448050.4083	-0.0000	1.0000	-2838127.6256	2838125.6704
course_type=T	-0.9393	1431313.9493	-0.0000	1.0000	-2805324.7306	2805322.8519
track_condition=FM	-0.6154	1046549.0372	-0.0000	1.0000	-2051199.0364	2051197.8057
track_condition=FT	-0.6223	1037156.7002	-0.0000	1.0000	-2032790.4010	2032789.1564
track_condition=GD	-0.6223	1088470.6198	-0.0000	1.0000	-2133363.8354	2133362.5909

track_condition=MY	-0.6157	1023176.2314	-0.0000	1.0000	-2005389.1790	2005387.9476
track_condition=SY	-0.6101	1058058.8301	-0.0000	1.0000	-2073757.8106	2073756.5904
track_condition=YL	-0.6291	1055407.4023	-0.0000	1.0000	-2068561.1267	2068559.8684
race_type=ALW	-0.4101	2103101.9046	-0.0000	1.0000	-4122004.3990	4122003.5788
race_type=AOC	-0.4058	2103101.9046	-0.0000	1.0000	-4122004.3948	4122003.5831
race_type=CLM	-0.3904	2103101.9046	-0.0000	1.0000	-4122004.3793	4122003.5986
race_type=MCL	-0.3269	2103101.9046	-0.0000	1.0000	-4122004.3158	4122003.6620
race_type=MSW	-0.3306	2103101.9046	-0.0000	1.0000	-4122004.3195	4122003.6583
race_type=SOC	-0.3509	2103101.9046	-0.0000	1.0000	-4122004.3398	4122003.6380
race_type=SST	-0.4041	2103101.9046	-0.0000	1.0000	-4122004.3930	4122003.5849
race_type=STK	-0.3196	2103101.9046	-0.0000	1.0000	-4122004.3085	4122003.6693
race_type=STR	-0.4280	2103101.9046	-0.0000	1.0000	-4122004.4169	4122003.5610
race_type=WCL	-0.3486	2103101.9046	-0.0000	1.0000	-4122004.3375	4122003.6404
jockey_id	-0.0012	0.0009	-1.3130	0.1892	-0.0029	0.0006
cumulative_horse_races	0.0130	0.0222	0.5846	0.5588	-0.0305	0.0564
cumulative_horse_wins	0.0335	0.0520	0.6447	0.5191	-0.0684	0.1355
cumulative_horse_hold3	-0.0194	0.0370	-0.5244	0.6000	-0.0919	0.0531
cumulative_jockey_races	-0.0006	0.0006	-0.9780	0.3281	-0.0017	0.0006
cumulative_jockey_wins	0.0020	0.0029	0.6917	0.4891	-0.0036	0.0076
cumulative_jockey_hold3	0.0003	0.0018	0.1511	0.8799	-0.0032	0.0037

Test Data Performance:

Accuracy of logistic regression classifier on test set: 0.87

	precision	recall	f1-score	support
0	0.87	1.00	0.93	4395
1	0.00	0.00	0.00	653
accuracy			0.87	5048
macro avg	0.44	0.50	0.47	5048
weighted avg	0.76	0.87	0.81	5048



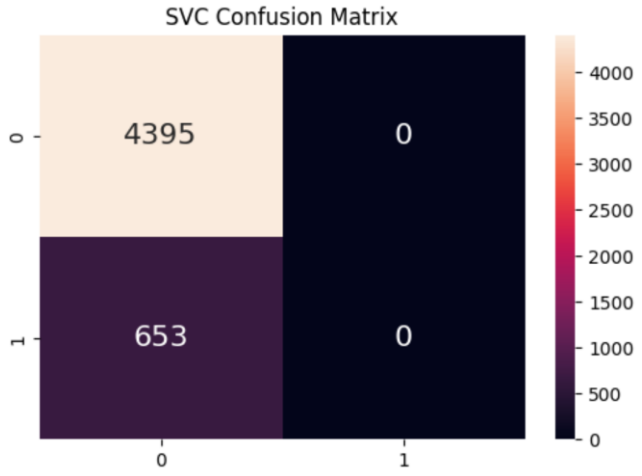
A Box-Tidwell test was not performed to validate the assumptions of logistic regression because there aren't many continuous variables within the dataset. But it is an empirical understanding that the variables do not have a reasonable linearity with the outcome of the logit. Therefore, a logistic regression model is not successful. Rather we need a model that can work with the non-linearity and the categorical (one-hot-encoded) terms in the data.

Support Vector Classifier:

```
Model accuracy score with default hyperparameters: 0.8706
              precision    recall  f1-score   support

           0       0.87      1.00      0.93      4395
           1       0.00      0.00      0.00       653

    accuracy                           0.87      5048
   macro avg       0.44      0.50      0.47      5048
weighted avg       0.76      0.87      0.81      5048
```

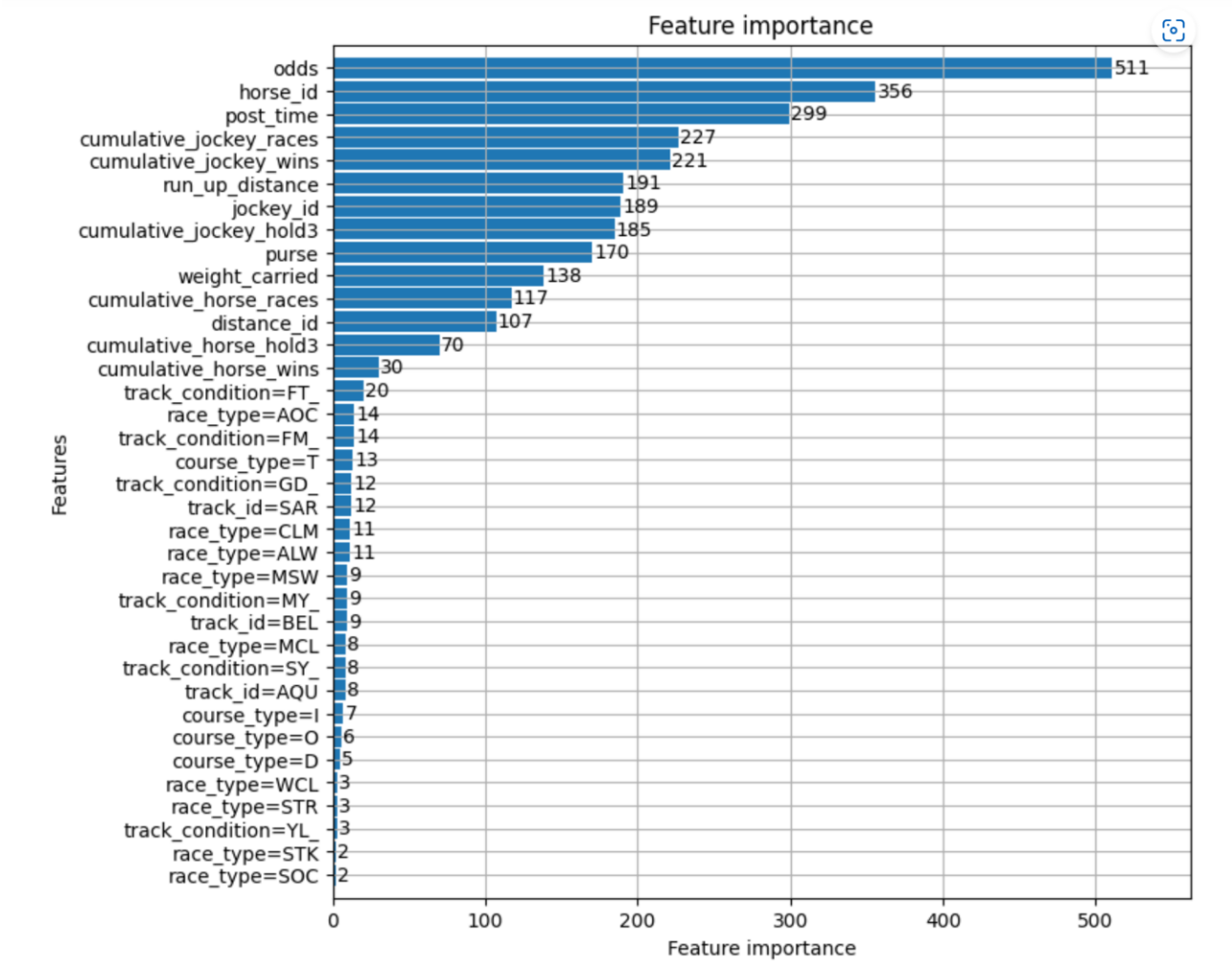


As expected, we obtain similar results. Class '0' dominates the outcomes, indicating towards and imbalance within the dataset. It may not be useful to balance the dataset by oversampling or underdamping because a race may only have one winner, and artificially changing this ratio may negatively affect model performance. We need to better model upon the priors as found within the dataset. We'll first test a support vector classifier before implementing a model to fit the non-linearity and the class imbalance.

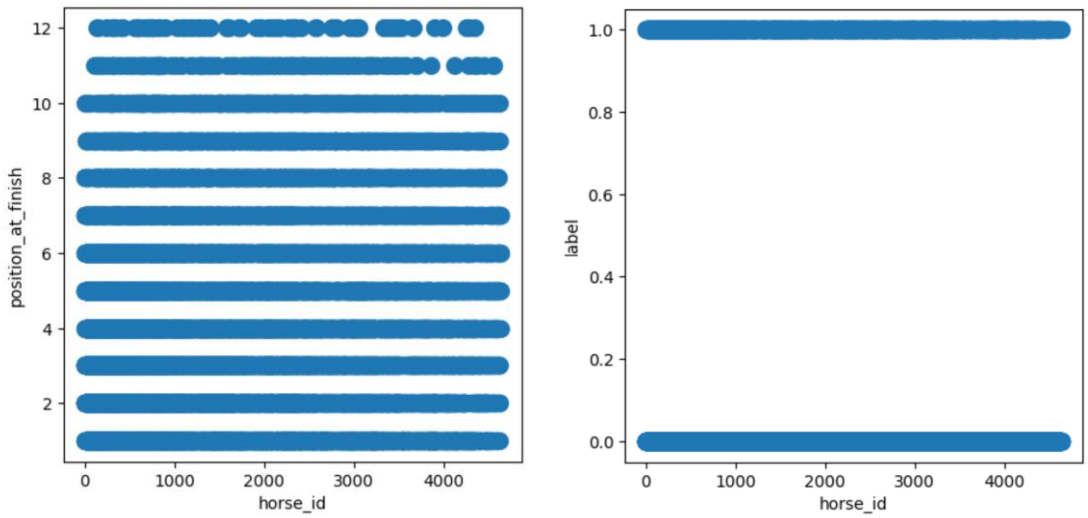
LightGBM:

LightGBM, short for light gradient-boosting machine, is a free and open-source distributed gradient-boosting framework for machine learning. We can use this to extract feature importance.

Feature Importance:

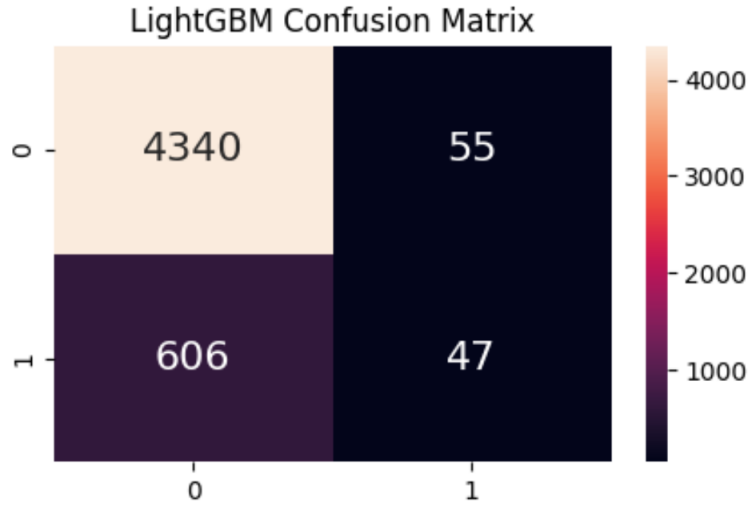


Horse ID gives a high importance, check to see if raw Horse_ID values show any linearity or trend with the outcome variable



Light GBM Classifier Results:

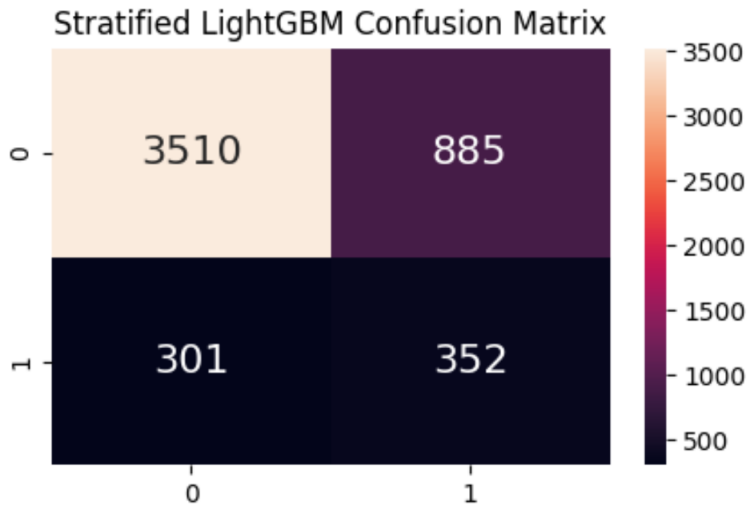
	precision	recall	f1-score	support
0	0.88	0.99	0.93	4395
1	0.46	0.07	0.12	653
accuracy			0.87	5048
macro avg	0.67	0.53	0.53	5048
weighted avg	0.82	0.87	0.83	5048



LightGBM Stratified:

This model generates stratified folds of the training data which maintain the balance of the outcome class in each fold.

	precision	recall	f1-score	support
0	0.92	0.80	0.86	4395
1	0.28	0.54	0.37	653
accuracy			0.77	5048
macro avg	0.60	0.67	0.61	5048
weighted avg	0.84	0.77	0.79	5048



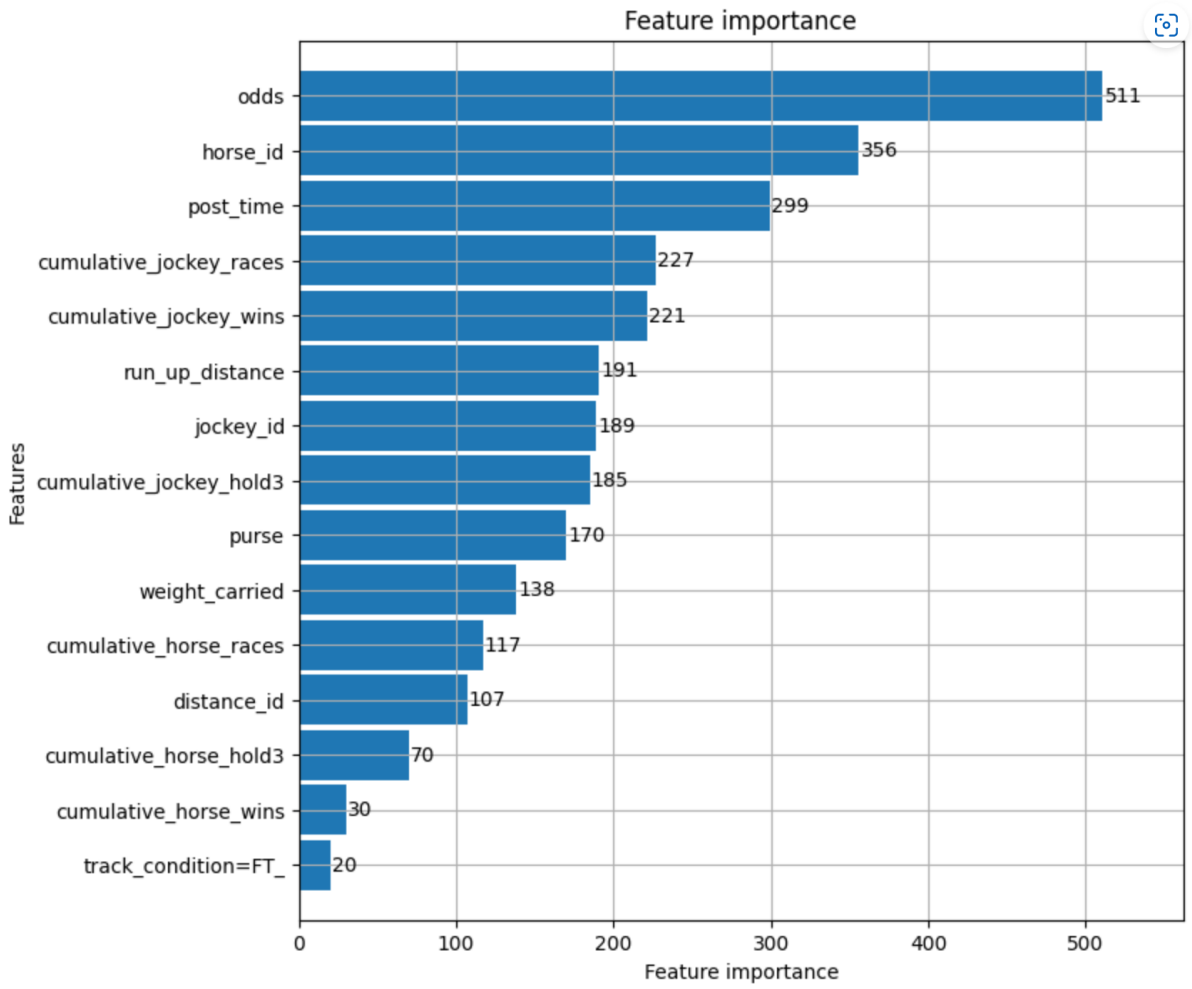
Performance seems to increase significantly by using the stratified LightGBM model.

Handling Class Imbalance with SMOTE:

Synthetic Minority Oversampling Technique (SMOTE) generates artificial data to balance the outcome variable.

SMOTE first selects a minority class instance a at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.[1]

15 of the most important features are chosen for SMOTE resampling.

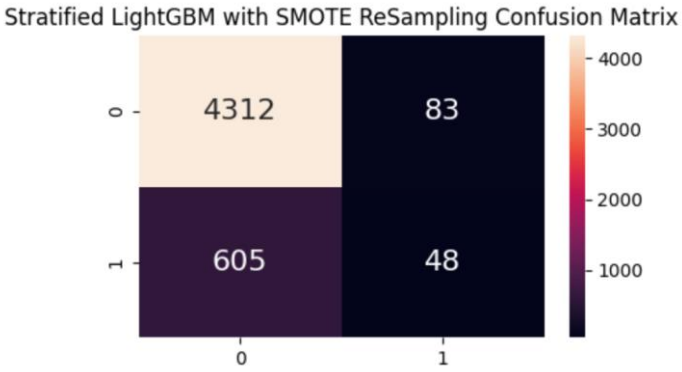


The outcome variable is balanced to 50% through artificial points:

Original Balance = 0.137006237006237
SMOTE Oversampling Balance = 0.5

LightGBM is then trained on the SMOTE resample data. The results of the model are:

	precision	recall	f1-score	support
0	0.88	0.98	0.93	4395
1	0.37	0.07	0.12	653
accuracy			0.86	5048
macro avg	0.62	0.53	0.52	5048
weighted avg	0.81	0.86	0.82	5048



This does not help with model performance, this is likely because there can only be one winner in a race and artificially introducing more winners does not positively impact model predictability.

Re-Engineering the Target Variable:

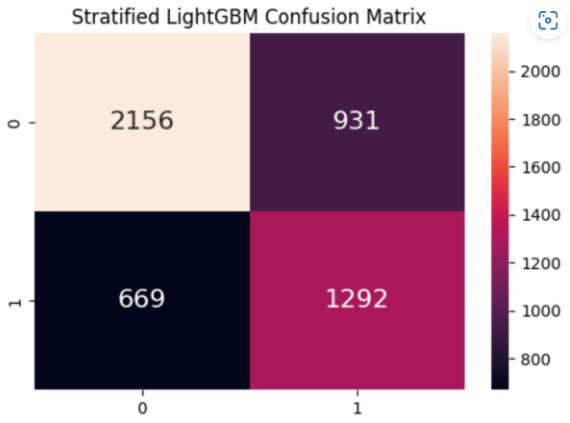
The target variable is redefined as follows:

1: if position at finish is 1, 2 or 3.

0: if position at finish is not among top 3.

Stratified LightGBM is then trained again with the new target variable.

	precision	recall	f1-score	support
0	0.76	0.70	0.73	3087
1	0.58	0.66	0.62	1961
accuracy			0.68	5048
macro avg	0.67	0.68	0.67	5048
weighted avg	0.69	0.68	0.69	5048



Performance has improved significantly and we're able to predict whether a place was held in the races with a much better recall.

Conclusion and Future Expansion:

The project aims at identifying factors that affect the final position of a horse in a race, followed by understanding whether we can use these factors to predict the final position of a horse a horse in a race. Feature importance as calculated by LightGBM has been reported and LightGBM with stratified folds has shown to give a recall of 66% for cases when a horse holds a place in top 3, and 70% for cases when a horse does not hold a place in top 3. Moreover, resampling for handling class imbalance has shown to be harmful in such modeling because a race only constitutes of one winner, balancing this ratio incorrectly trains a model.

Although we have a decent understanding of race factors that affect position, it is important to analyze positional data to understand performance as well, doing so would have the added advantage of accounting for the relative nature of racing i.e., the final position of a horse not only depends on its own performance but also the performance of other horse taking part in the race. This can be a good direction to expand this project in.

References:

[1] He, H., & Ma, Y. (Eds.). (2013). Imbalanced Learning. Wiley. https://doi.org/10.1002/9781118646106