Big Derby Data

Darshan Swami, Chandra Harsha, Sidhanth Subhash Jain

INTRODUCTION

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 analyse 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

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.

Objective 1: Data Insights. Explored the data before beginning any kind of statistical testing, by doing some Exploratory Data Analysis and came up with great insights about the data.

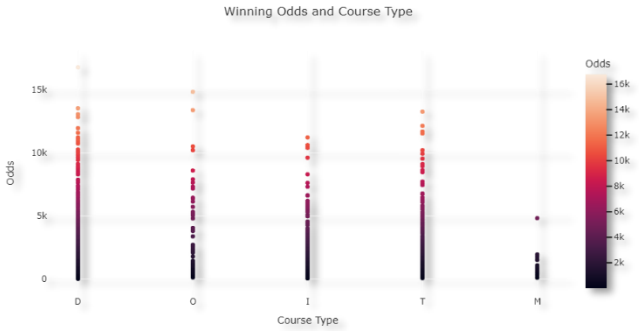
Objective 2: Injury Analysis. If there is an injury at the start of the career for a horse it would be fatal for its future so it's important to analyse the starting races of the horses to prevent long term injuries.

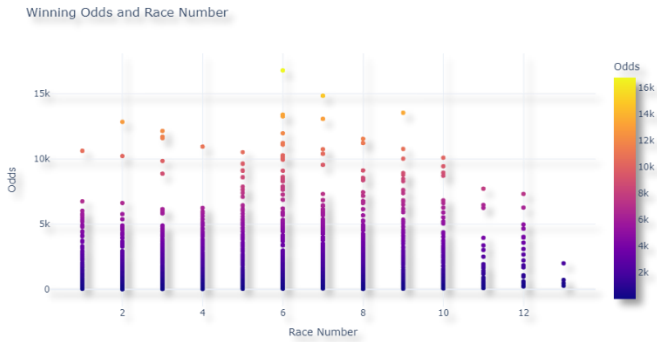
Objective 3: Data Modelling and Feature Engineering. Develop a prototype of Yelp Predictor by acquiring a sample data set and implementing the previously designed math solution.

OBJECTIVE 1: Data Insights

Winning Odds and Course Type.

Low odds indicate that a lot of money has been wagered on the horse to win (a favorite). High odds indicate that less money has been wagered on him to win (a longshot or underdog). The plot shows that Course Types D and O have the best odds of winning more money in a race. If the odds are higher, you will make more money.



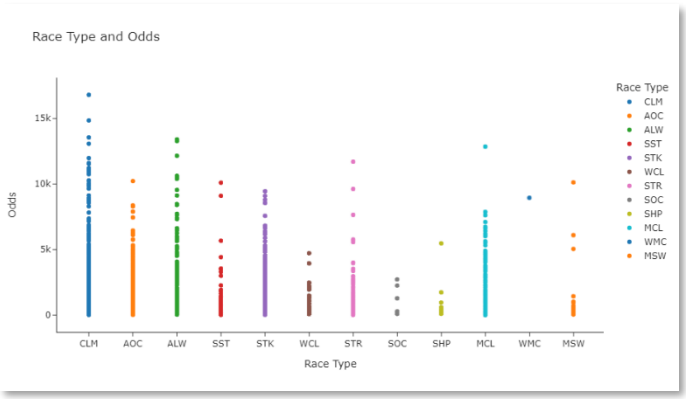


Winning Odds and Race Number.

The plot shows the relationship between the race number and winning odds. In a day there are multiple races conducted and are ordered ascendingly from 1 to 13. Races 6 and 7 have the best chances of winning more money. You will make more money if the odds are higher.

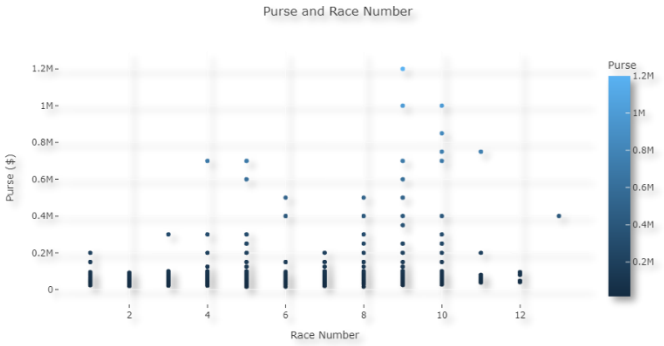
Race Type and Odds.

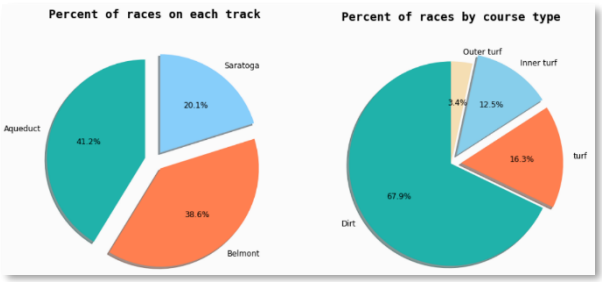
The odds of winning more money are for race type CLM i.e., Claiming. The odd of winning more money is lower for type SOC i.e., Starter Optional Claimer

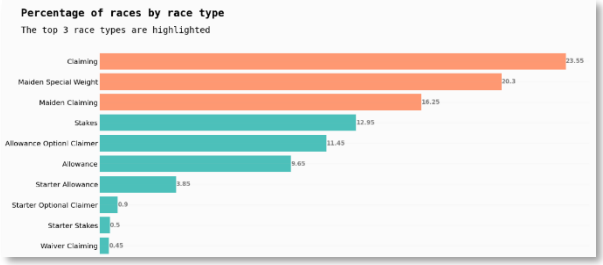


Purse and Race Number.

There are various types of races based on the size of the purse and the horse's experience. The race types are as follows, and we attempt to determine which race will provide us with the best odds of winning the bet. Race 9 has the highest prize money. So, in order to earn more money, the jockey should compete in Race 9.





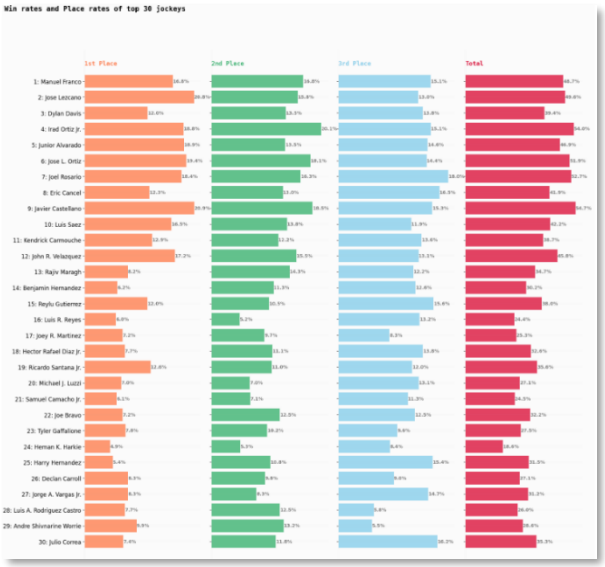


Racing Insights.

Claiming races accounted for more than 23% of races. Maiden Special Weight races accounted for around 20% of races. Maiden Claiming races are for horses that have never won a race and are eligible to be claimed. This type accounted for 16% of races. Stakes race is a horse race in which the prize offered is made up at least in part of money (such as entry fees) put up by the owners of the horses entered. This type of race accounted for about 13% of races.

Jockey Insights.

Win Rate as percentage of races a jockey finished in 1st position out of all the races, he participated in. And Place Rate as percentage of races a jockey finished in 1st, 2nd, or 3rd position out of all the races he participated in. Javier Castellano had the highest place rate and win rate. Both Javier Castellano and Jose Lezcano had a win rate over 21%.



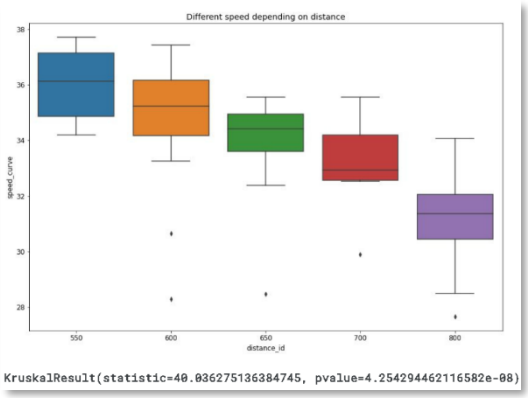
OBJECTIVE 2: Injury Analysis

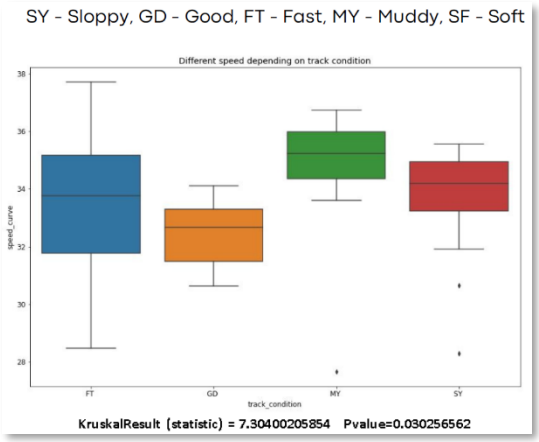
In order to prevent injuries, the idea is to analyse the average speed in a corner during the race. The Maiden Claiming Races during the winter season will be analysed. In order to maximize their chance of victory, it is important to know the race conditions. These race conditions, especially the average speed in the curve, allows the trainer to:

1. train horse to minimize the risk of injury
2. if the horse can follow this rhythm.
3. maximize chances of victories by selecting the most suitable race profile of the horse.
4.

Does the distance of the course influence the speed in the curve?

It seems that depending on the distance of the race the speed would be different. The p-value is smaller than 0.05%. We conclude that there is a statistically significant association between the speed in the curve and the distance. So, we can reject the null hypothesis



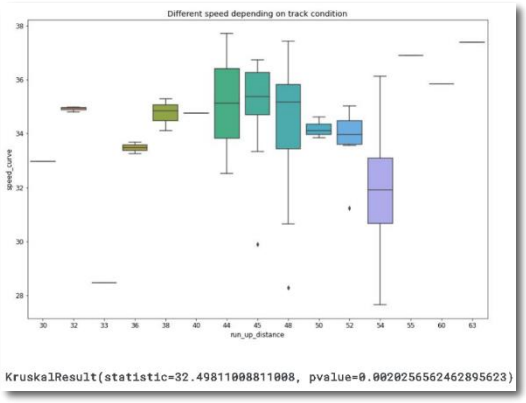


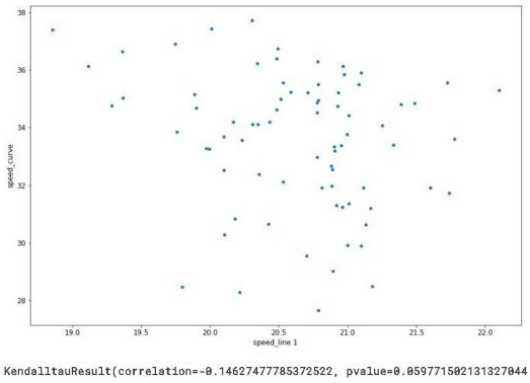
Does the track condition influence the speed in the curve?

Depending on the track condition the speed would be different. Null hypothesis: The speed is not depending on the track condition. The p-value is lower than 0.05%. We can reject the null hypothesis.

Does the run-up-distance influence the speed in the curve?

Depending on the run-up-distance the speed would be different. Null hypothesis: The speed is not depending on run-up-distance. The p-value is smaller than 0.05%. There is a statistically significant association between the speed in the curve and the run-up-distance. We can reject the null hypothesis

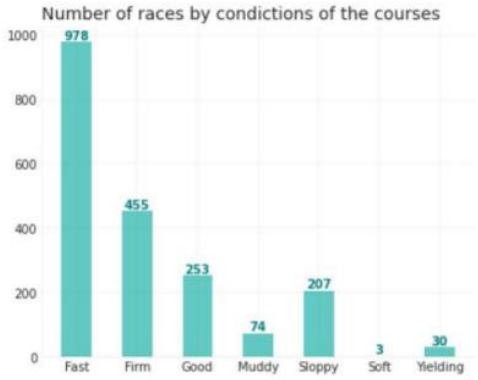




Does the speed in first line influence the speed in the curve?

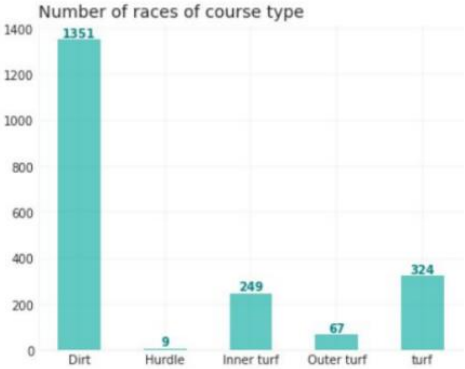
Null hypothesis: The speed on the curve is not depending on Speed in the first line .We can answer that the p-value is slightly higher than 0.05%. So, We should accept the null hypothesis. The speed in the first line doesn't influence the speed in the curve.

OBJECTIVE 3: (A) Feature Engineering



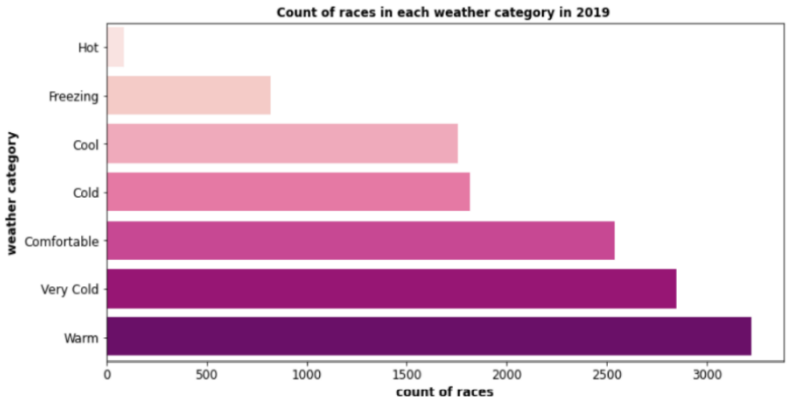
Understand the Indicating Factors of a Win

- Remove dates (race_date) on which there were less than 8 races
- Remove races in which there were less than 5 horses or more than 12 horses
- Remove 'Hurdle' races remove races with track condition 'Soft'
- Remove race_type in Waiver Maiden Claiming, Starter Handicap, Waiver Claiming



The descriptive statistical features of jockey's performances

- win rate before current race_date
- place rate before current race_date
- race count before current
- race_date
- win rate/place rate of current
- Course type/track



Also, the feature "weather" including all these categories (Freezing, Very Cold, Cold, Cool, Comfortable, Warm, Hot).

OBJECTIVE 3: (B) Training Machine Learning Model

To find the best hyper-parameters forLightGBM

Divide the whole dataset into 2 subsets, namely the training dataset and holdout dataset. The holdout data will not be used to train the models. Then, we create 3 expanding windows (as show in the following picture) using the training data. Thirdly, we run the hyperparameter tuning on the 3 windows (i.e. 3-folds) and find the best set of hyperparameters using OOF (out of the fold) performance.

```
2022-11-27 23:30:42.215931 0
100%|██████████| 100/100 [05:27<00:00,  3.28s/trial, best loss: -0.9128025975131014]
{'boosting_type': 'gbdt', 'colsample_bytree': 0.65, 'learning_rate': 0.08, 'max_bin': 40, 'max_depth': 1, 'metric': 'auc', 'min_chi
ld_samples': 45, 'min_data_in_bin': 75, 'n_estimators': 661, 'n_jobs': 4, 'num_leaves': 201, 'objective': 'binary', 'random_state':
1234, 'reg_alpha': 5, 'reg_lambda': 0.001, 'scale_pos_weight': 8, 'subsample': 0.95, 'subsample_freq': 12}
2022-11-27 23:36:10.154588 1
100%|██████████| 100/100 [05:26<00:00,  3.26s/trial, best loss: -0.9130168439080241]
{'boosting_type': 'gbdt', 'colsample_bytree': 0.75, 'learning_rate': 0.05, 'max_bin': 45, 'max_depth': 1, 'metric': 'auc', 'min_chi
ld_samples': 30, 'min_data_in_bin': 90, 'n_estimators': 939, 'n_jobs': 4, 'num_leaves': 126, 'objective': 'binary', 'random_state':
1234, 'reg_alpha': 0.1, 'reg_lambda': 15, 'scale_pos_weight': 8, 'subsample': 0.9, 'subsample_freq': 14}
2022-11-27 23:41:36.432007 2
100%|██████████| 100/100 [05:59<00:00,  3.60s/trial, best loss: -0.911592443187408]
{'boosting_type': 'gbdt', 'colsample_bytree': 0.4, 'learning_rate': 0.02, 'max_bin': 80, 'max_depth': 2, 'metric': 'auc', 'min_chil
d_samples': 20, 'min_data_in_bin': 95, 'n_estimators': 1043, 'n_jobs': 4, 'num_leaves': 41, 'objective': 'binary', 'random_state':
1234, 'reg_alpha': 15, 'reg_lambda': 10, 'scale_pos_weight': 8, 'subsample': 0.95, 'subsample_freq': 15}
```

Removing collinearity

Before we fit the whole training data into final hyperparameters, we used VIF to remove co-linear features. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

```
ReduceVIF fit
ReduceVIF transform
Dropping pre31_win_sum with vif=139.7280478393303
Dropping pre31_plc_rate with vif=67.63434828499291
Dropping pre31_plc_sum with vif=67.01383966869585
Dropping month with vif=64.89766358289387
Dropping distance_id with vif=60.438808641724144
Dropping weight_carried with vif=36.17914074789011
Dropping pre_place_rate with vif=24.672851580866784
```

Results

- To evaluate the performance of the models on test data, we consider the precision and recall.
- For lightgbm, the precision and recall are 0.68 and 0.28 respectively.
- The reason for a bit high precision and low recall is highly unbalanced data.
- In this case Actual Win is of only 12% and No Win constitutes to 88% of the whole data.
- To address this issue, we used SMOTE which is an up-sampling techniques.
- After using smote, the precision and recall are 0.590 and 0.449.
- Indicating factors of Winning.

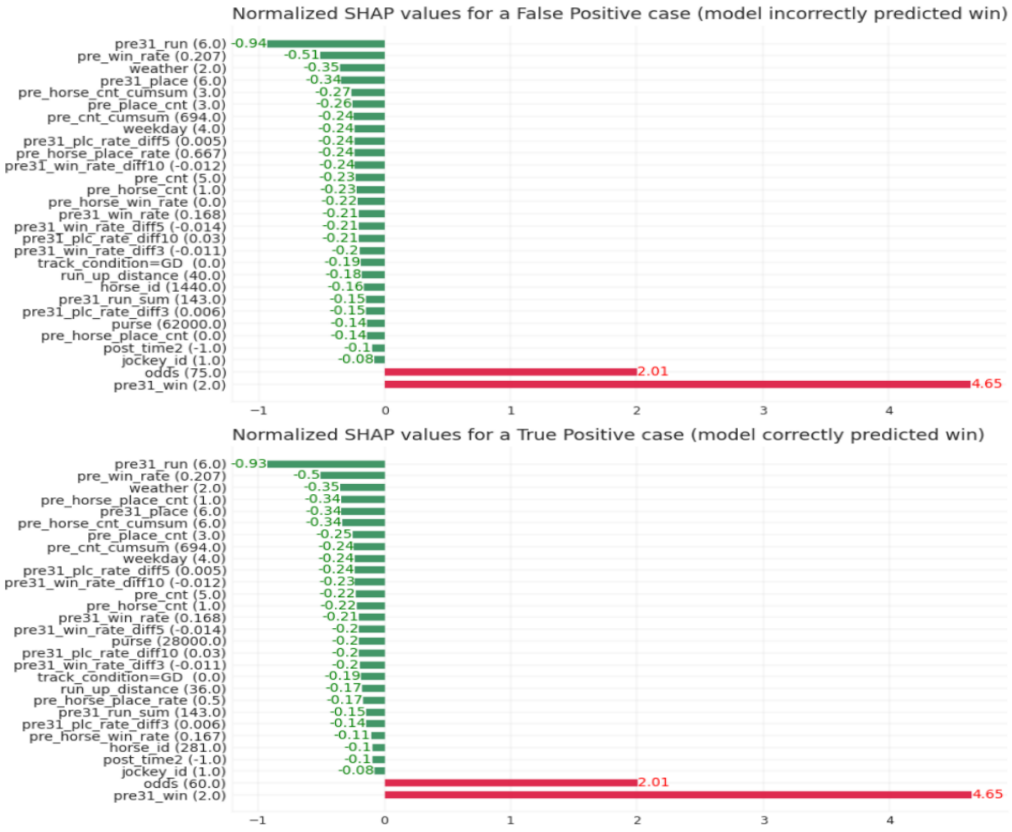
	Actual: Win	Actual: No Win
Predict: Win	97	34
Predict: No Win	250	2547

Confusion matrix without Smote

	Actual: Win	**Actual: No Win**
Predict: Win	156	108
Predict: No Win	191	2473

Confusion matrix after using Smote

For lightgbm, odds, jockey's previous (before current race day) wining counts and rate are the 4 most critical indicator of a possible win.



CONCLUSIONS

Data Insights:

- Features like course type, purse, race type and race number helps decide the winning odds of a jockey.
- Jockey selection based on winning percentage can be an important factor in deciding who to bet on.

Injury Analysis:

- Distance_id, run_up_distance and track_conditions are the most vital features to predict speed in the curve that determines the chances of horse being injured

Predict Winning:

- The developed model didn't perform as aspected because of lack of data.