Data Pirates

DS 502 Statistical Methods for Data Science

Big Data Derby

Darshan Swami

dswami@wpi.edu

Sidhanth Jain

sjain3@wpi.edu

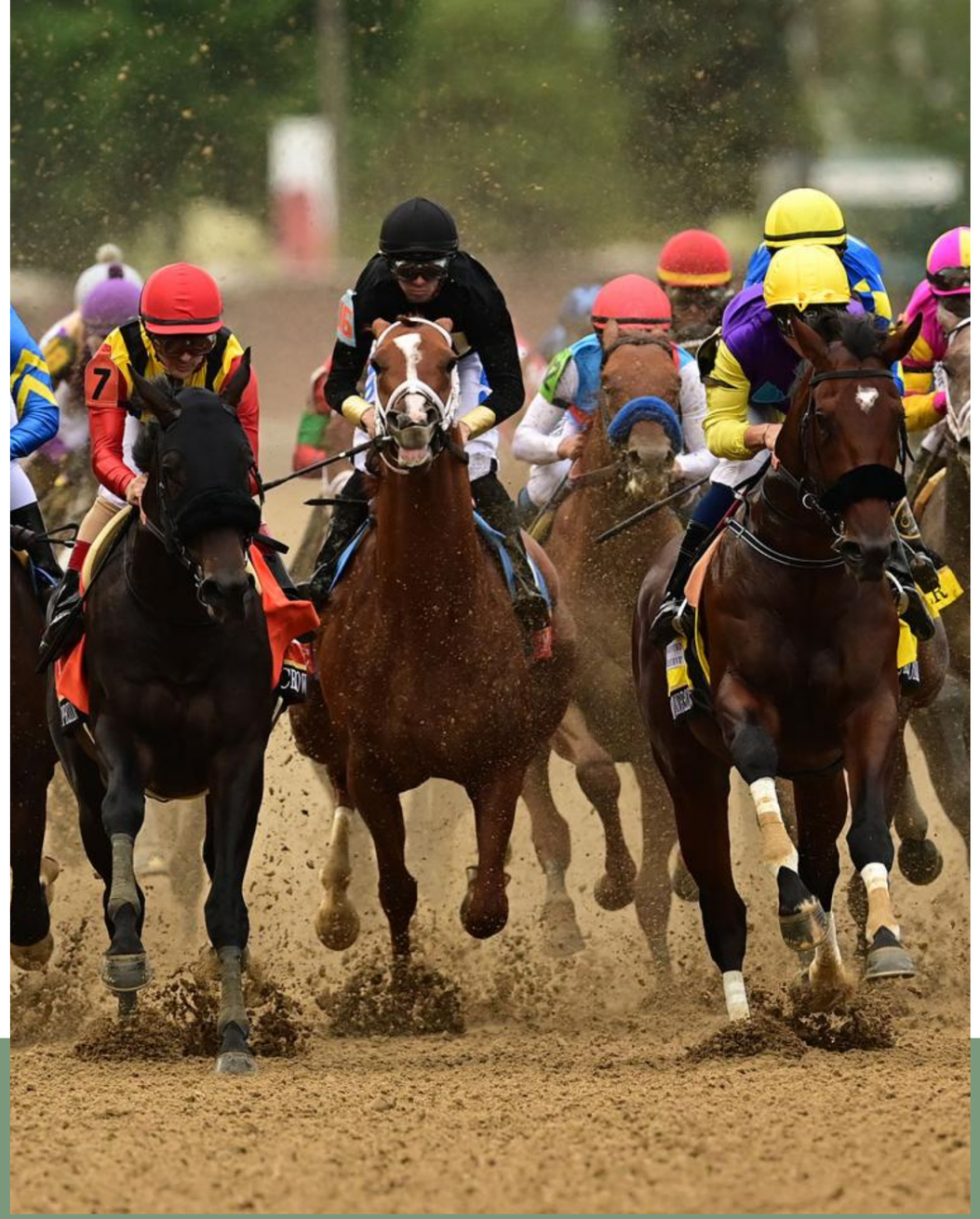
Chandra Harsha Rachabathuni

chrachabathuni@wpi.edu

Data Pirates

Agenda

- Introduction
- Data
- Problem Statement
- Data Insights
- Statistical Tests
- Data Modeling



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 analyze jockey decision making, and compare race surfaces. The project will assist racing horse trainers, and how efficiently one should bet on any race.

Data :

The data available is described in the following excerpt from the Kaggle page.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 - horse/jockey race data
- nyra_tracking_table.csv - tracking data
- nyra_2019_complete.csv - combined table of three above files
- nyc_weather_2019.csv - Weather data for the year 2019.

	track_id	race_date	race_number	program_number	trakus_index	latitude	longitude	distance_id
0	AQU	2019-01-01	9	6	72	40.672902	-73.827607	600
1	AQU	2019-01-01	9	6	73	40.672946	-73.827587	600
2	AQU	2019-01-01	9	6	74	40.672990	-73.827568	600
3	AQU	2019-01-01	9	6	63	40.672510	-73.827781	600
4	AQU	2019-01-01	9	6	64	40.672553	-73.827762	600

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

Problem Statement

01

Injury Analysis

Analysing injury affecting coditions in maiden claiming races.

02

Feature Selection

Indicating important features using VIF and creating new features to help in modelling.

03

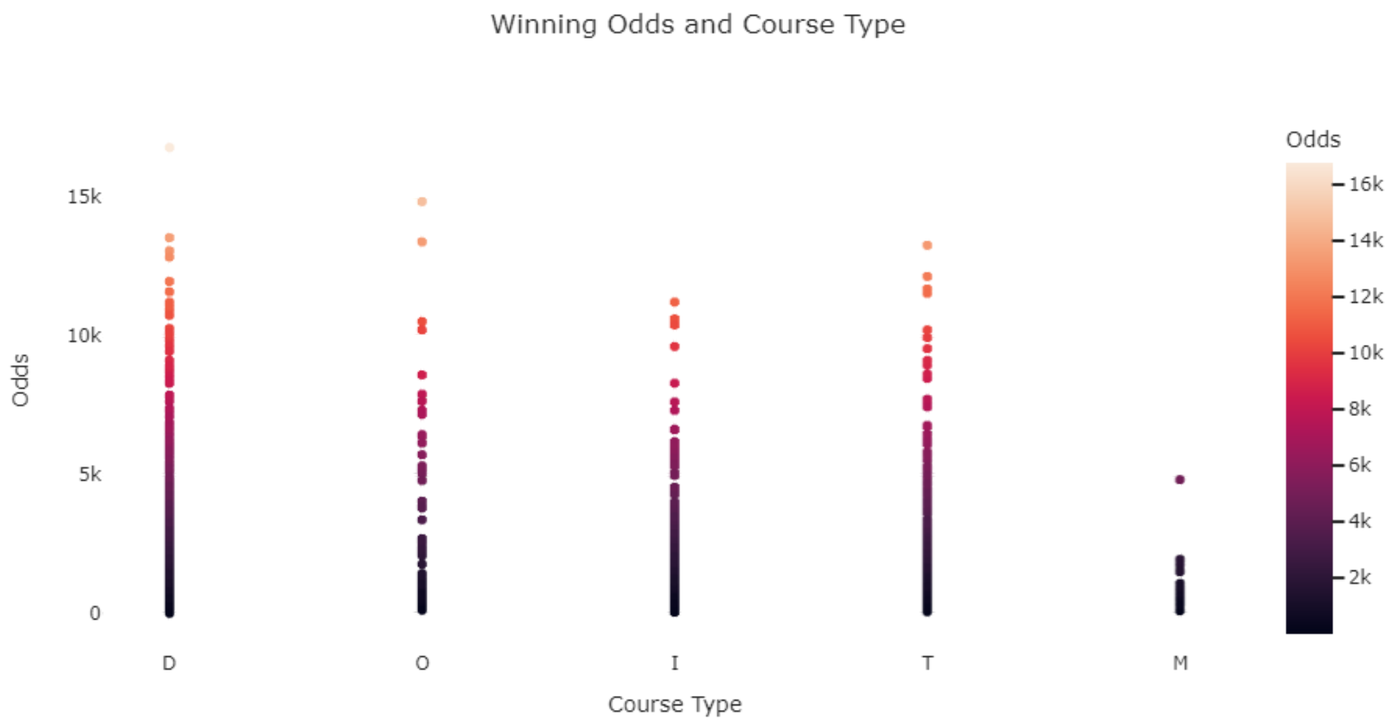
Applying Regression and Classification

Determining probability of horse winning using different models

Data Insights

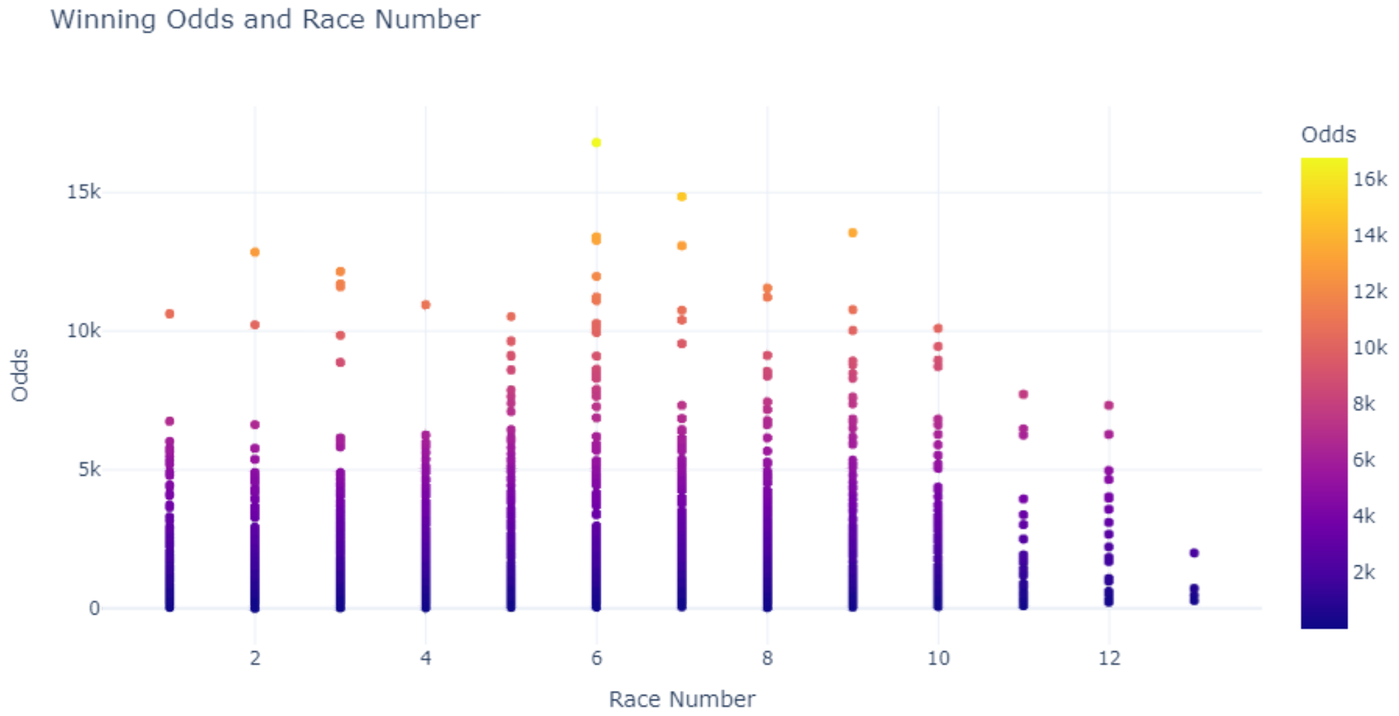
Low odds indicate that a lot of money has been wagered on the horse to win (a favourite)
High odds indicate that less money has been wagered on him to win (a longshot or underdog)

The plot show 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.



Data Insights

Races 6 and 7 have the best chances of winning more money. You will make more money if the odds are higher.

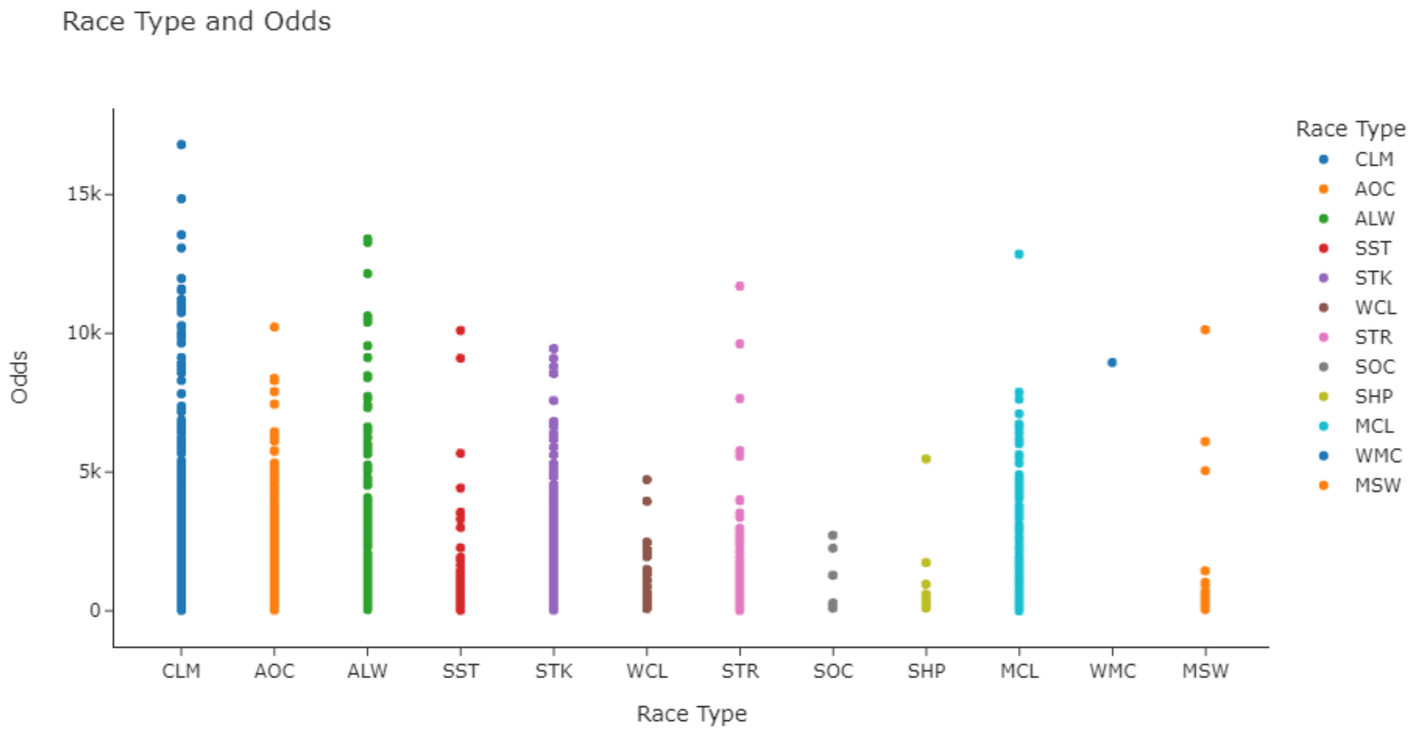


Data Insights

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.

The odds of winning more money is for race type CLM i.e. Claiming.

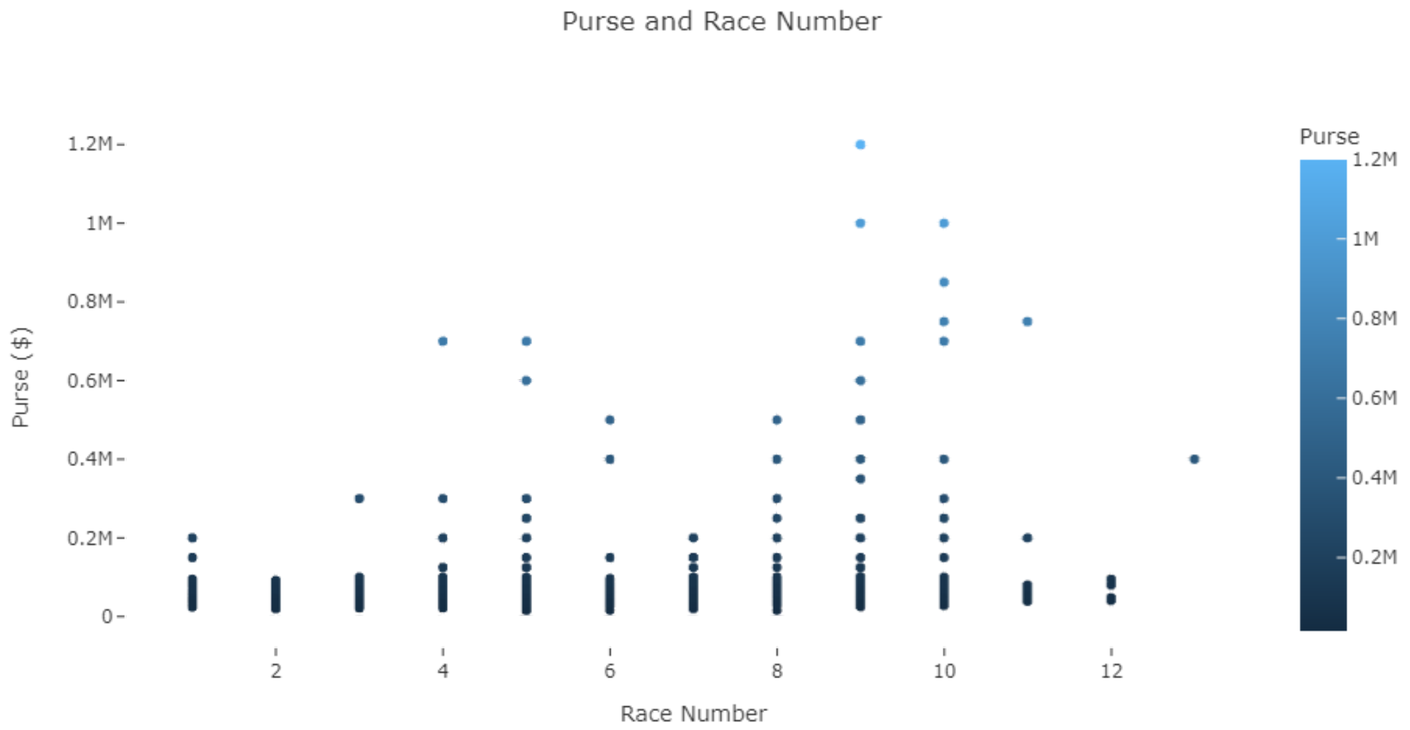
The odds of winning more money is lower for type SOC i.e. Starter Optional Claimer



Data Insights

Purse refers to the total amount of money paid out to horse owners at a specific track over a given time period, or to the percentages of a race's total purse that are awarded to each of the top finishers.

Race 9 has the highest prize money. So, in order to earn more money, the jockey should compete in Race 9

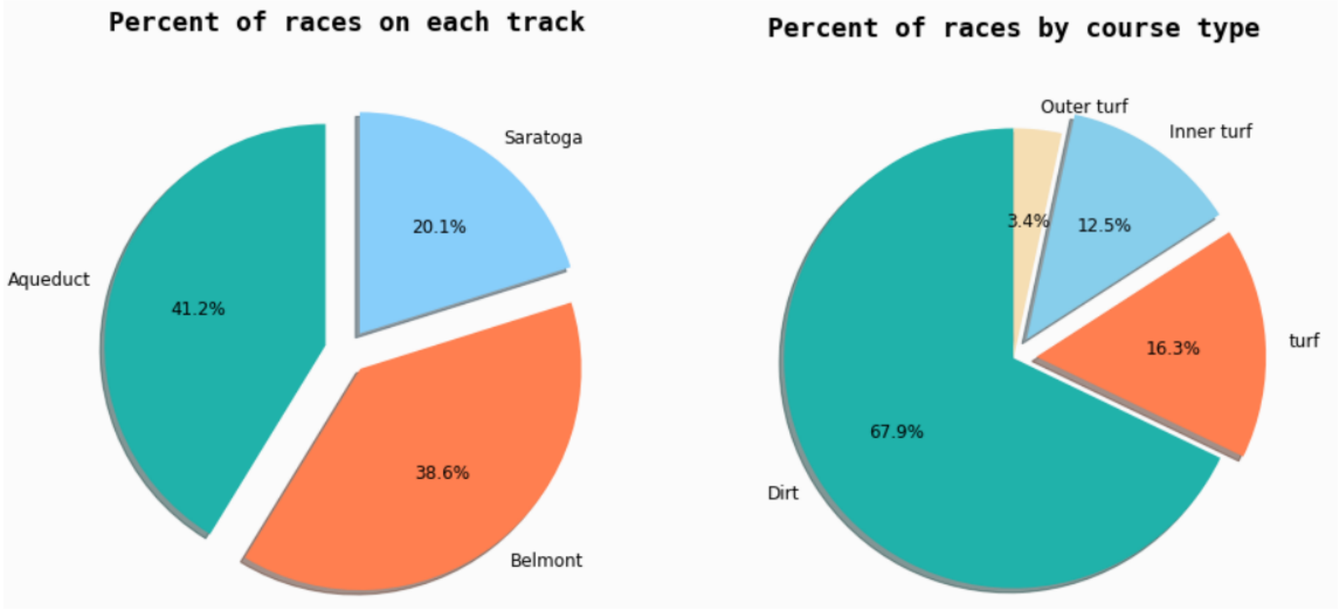


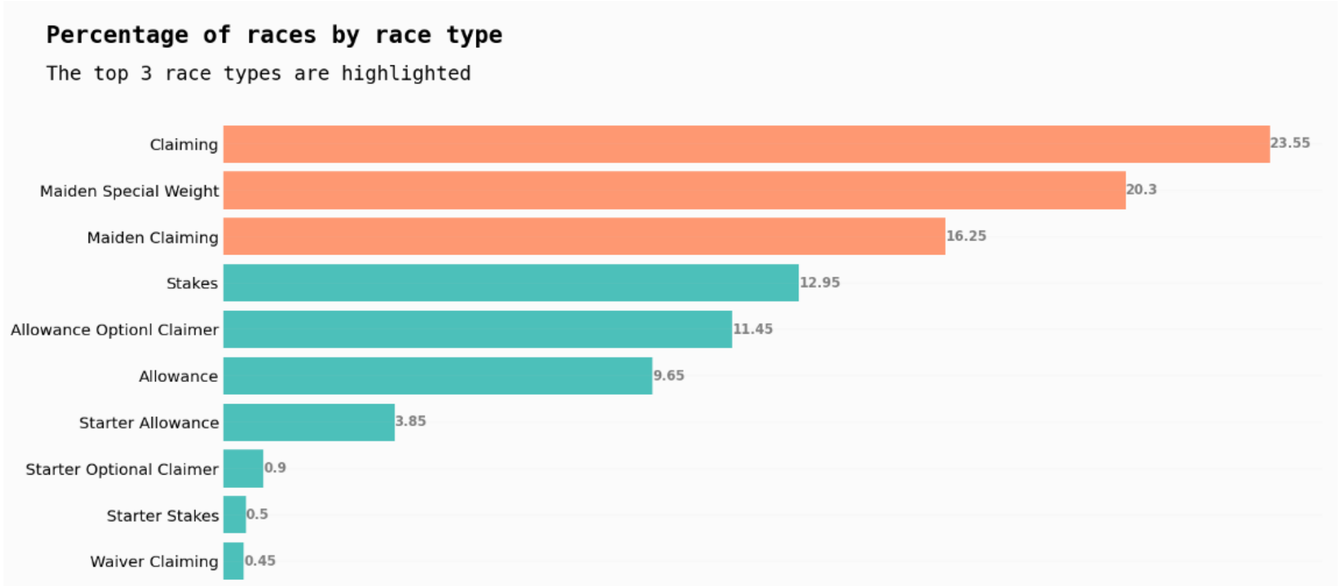
Data Insights

Claiming races accounted for more than 23% races. Maiden Special Weight races accounted for around 20% races.

Maiden Claiming races are for horses that have never won a race and are eligible to be claimed. This type accounted for 16% 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 races accounted forabout 13% races.

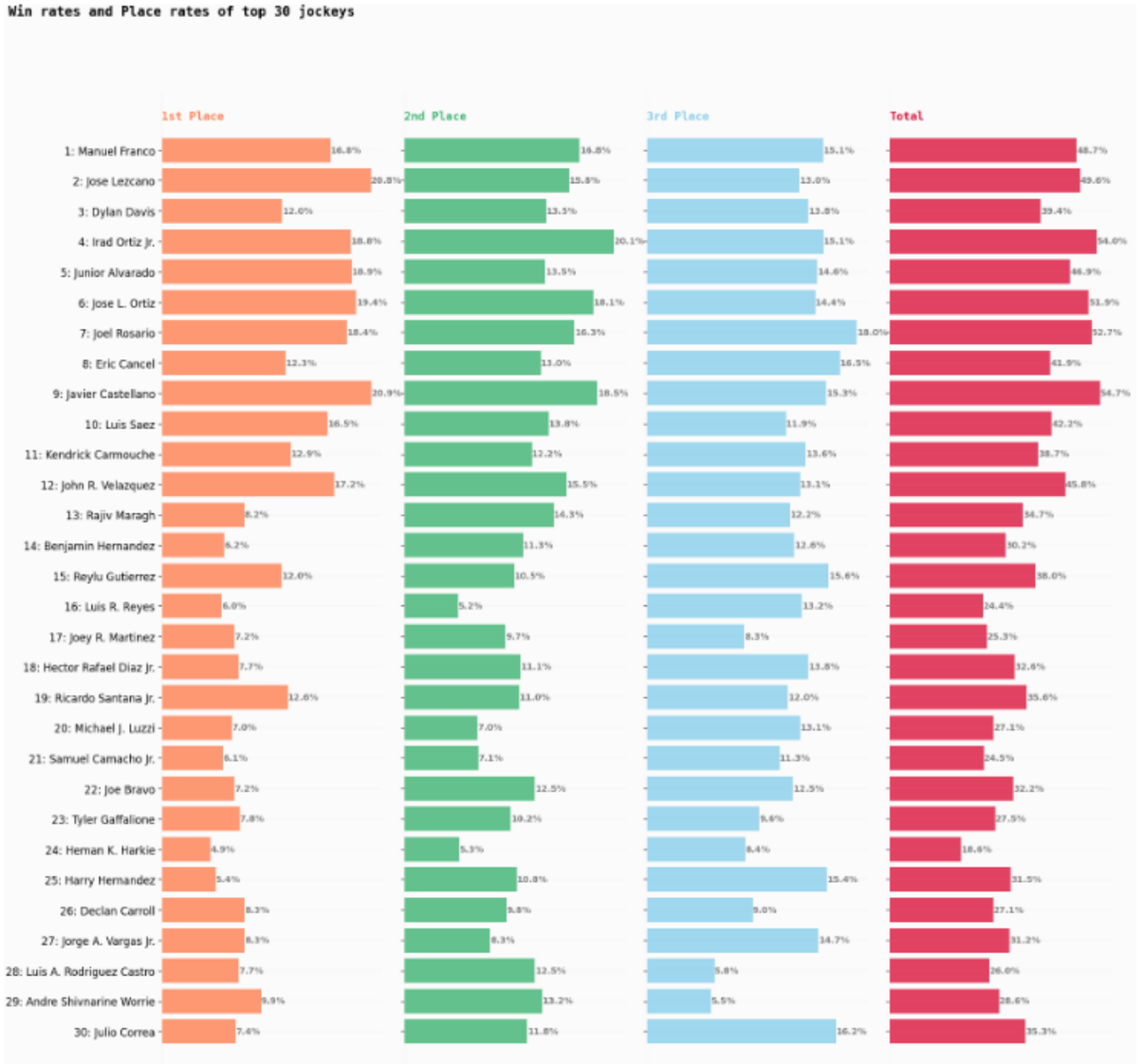




Data Insights

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%.



Injury Analysis

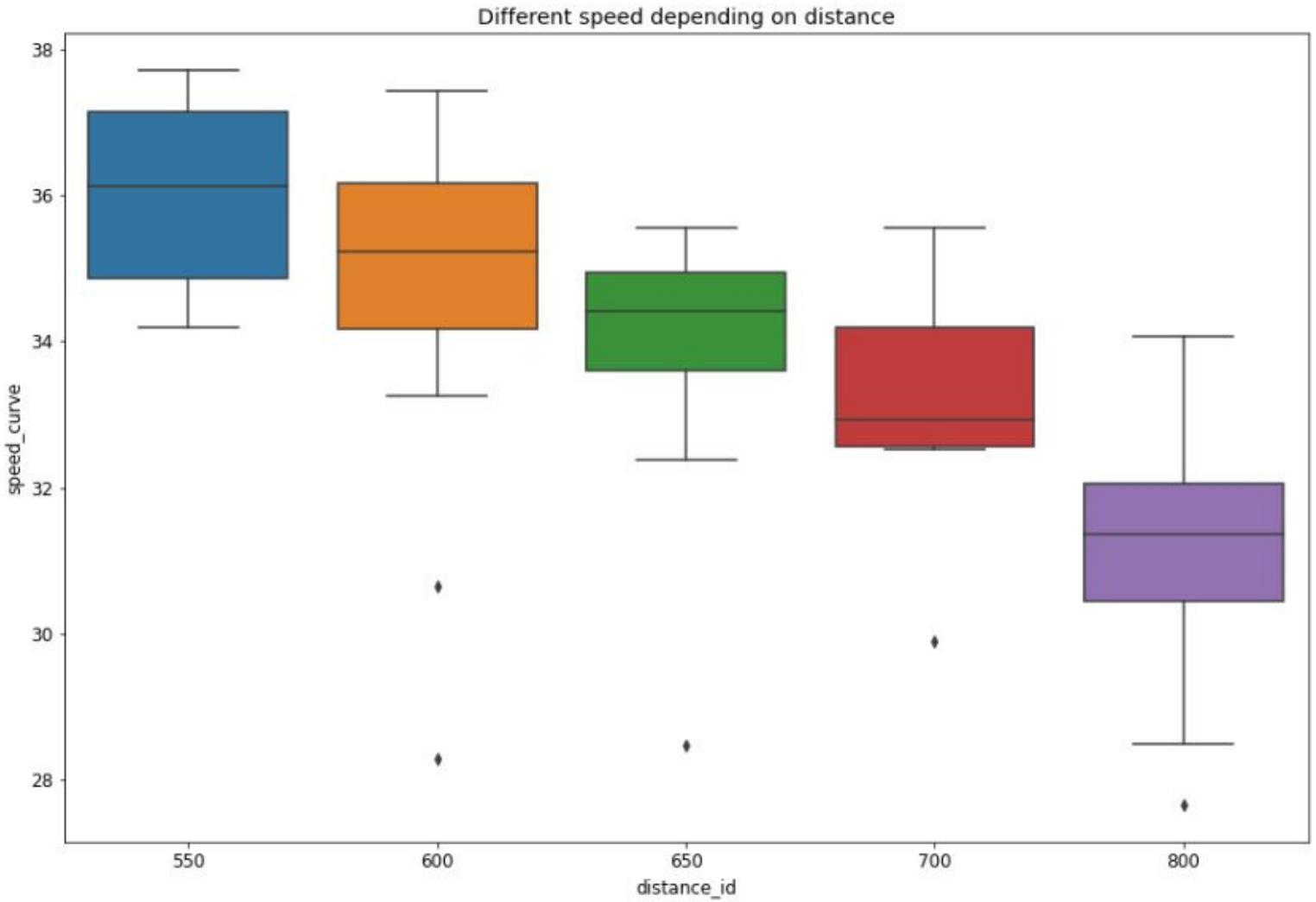
- In order to prevent injuries, the idea is to analyze the average speed in a corner during the race.
- The Maiden Claiming Races during the winter season will be analyzed.
- In order to maximize their chance of victory, it is important to know the race conditions.
- These race conditions, especially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

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

Injury Analysis

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.



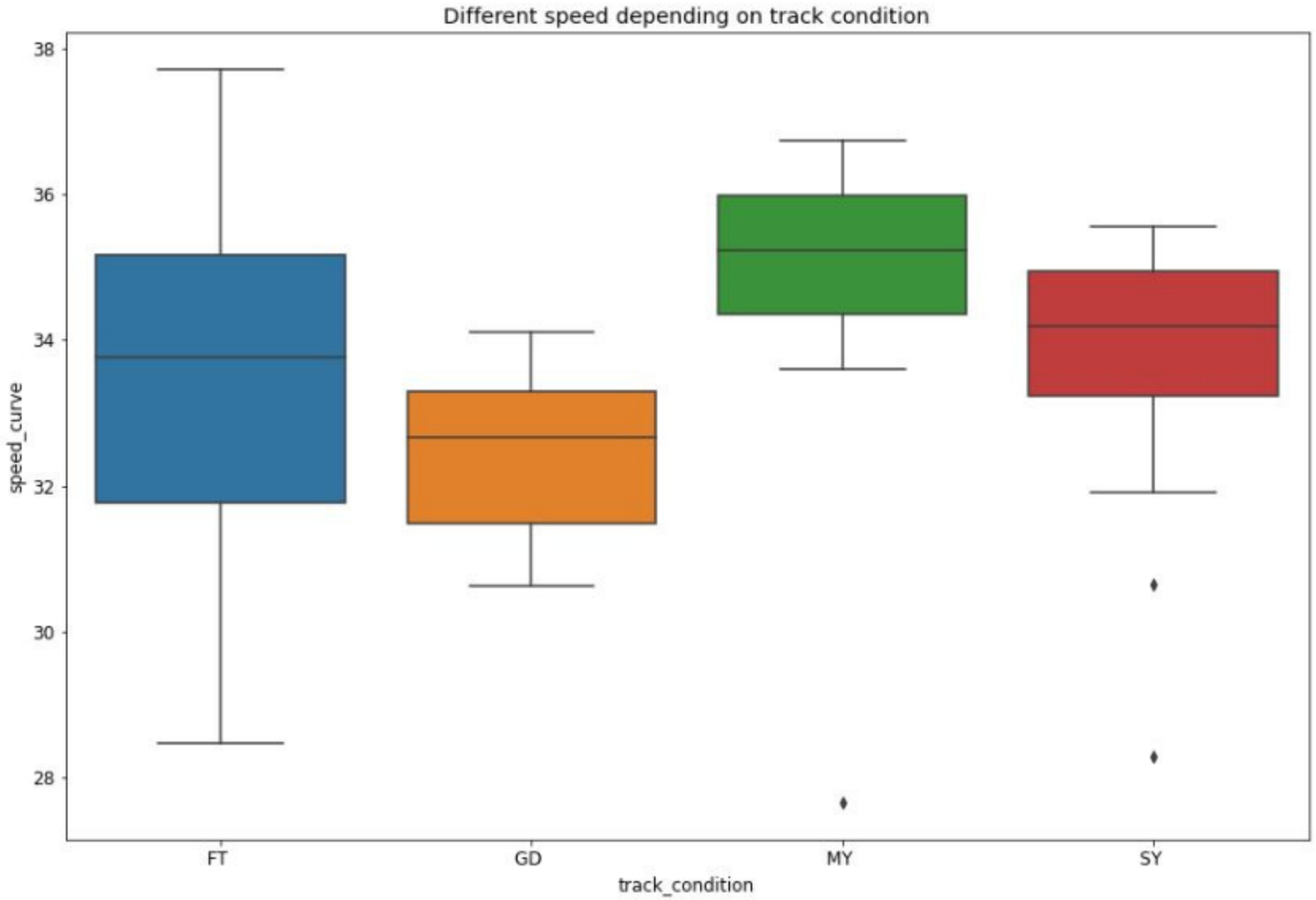

```
KruskalResult(statistic=40.036275136384745, pvalue=4.254294462116582e-08)
```

Injury Analysis

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.

SY - Sloppy, GD - Good, FT - Fast, MY - Muddy, SF - Soft

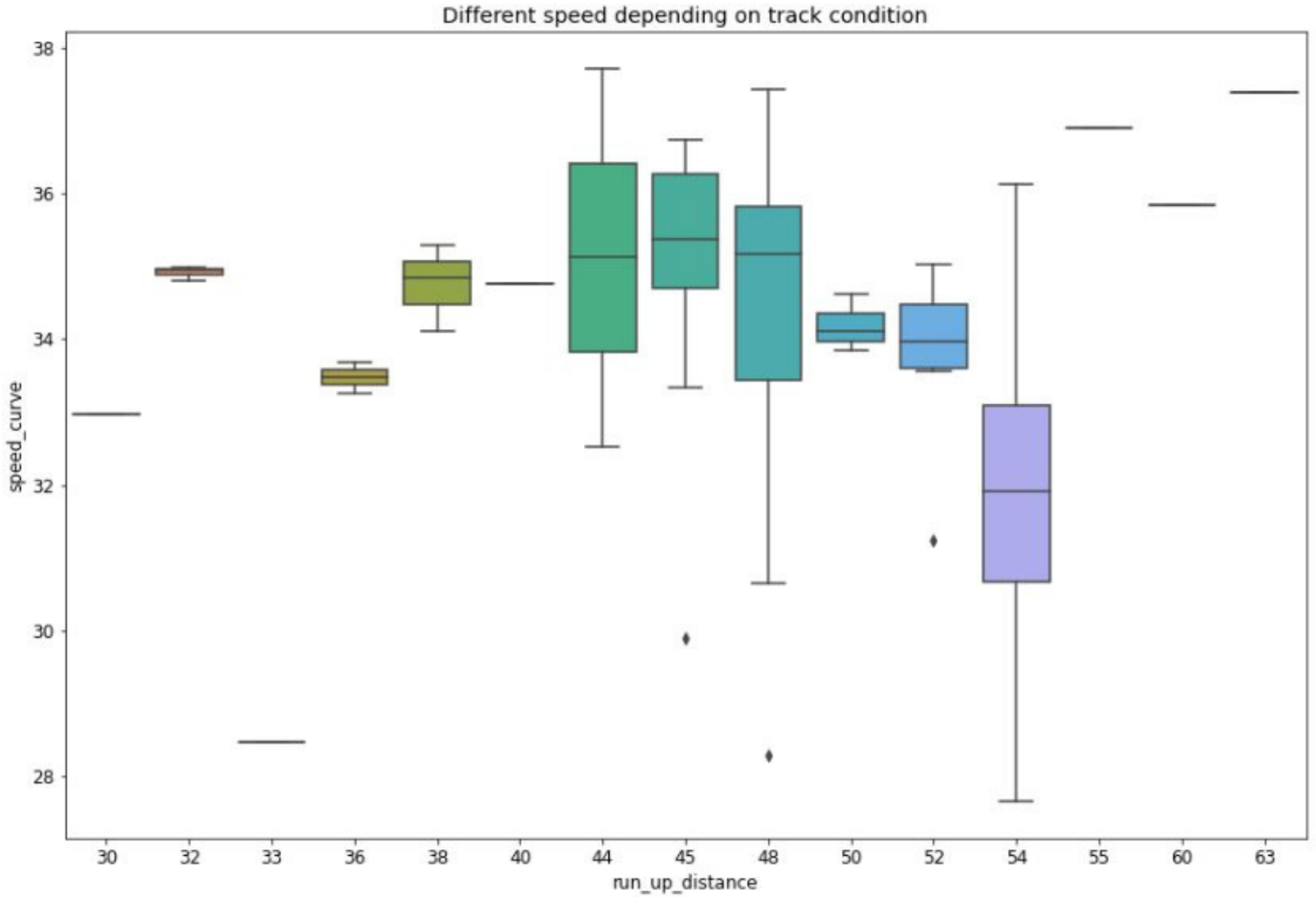


KruskalResult (statistic) = 7.30400205854 Pvalue=0.030256562

Injury Analysis

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.

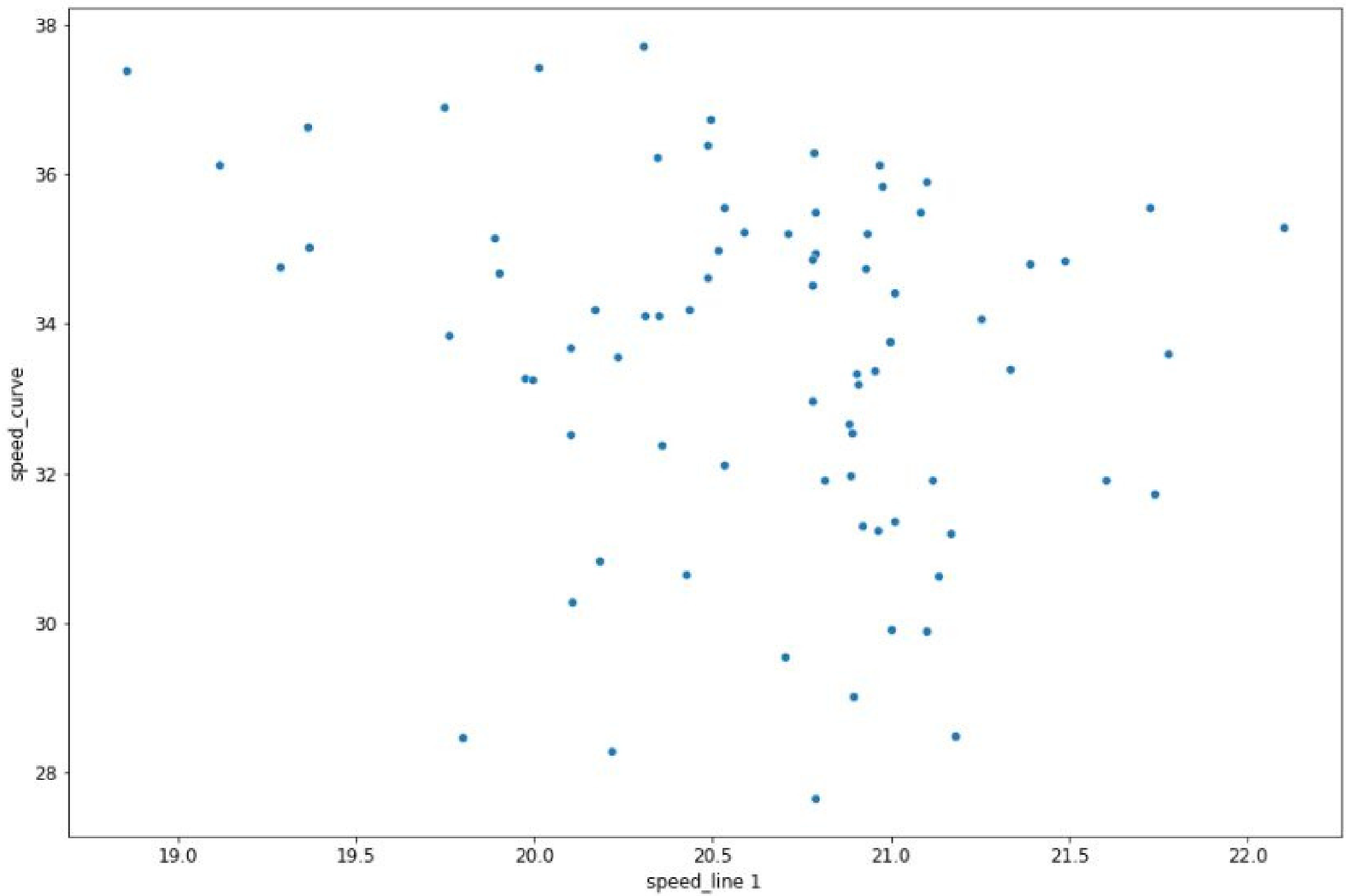



```
KruskalResult(statistic=32.498110088110008, pvalue=0.0020256562462895623)
```

Injury Analysis

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




```
KendalltauResult(correlation=-0.14627477785372522, pvalue=0.059771502131327044)
```

Injury Analysis

Predict the speed according to run up distance ,distance and track condition.

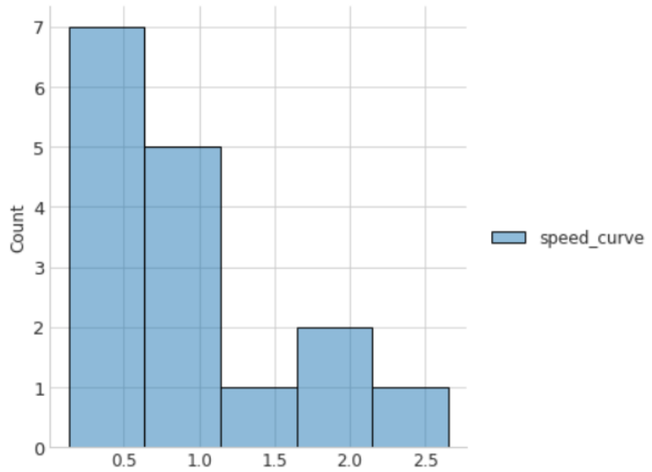
Selection of 4 regression models with different hyperparameters to be tested

- Ordinary least squares
- Ridge
- Lasso
- XGBoost

The best model to be used is XGBoost, since the test RMSE for XGBoost is 1.07 which is the lowest among other models.

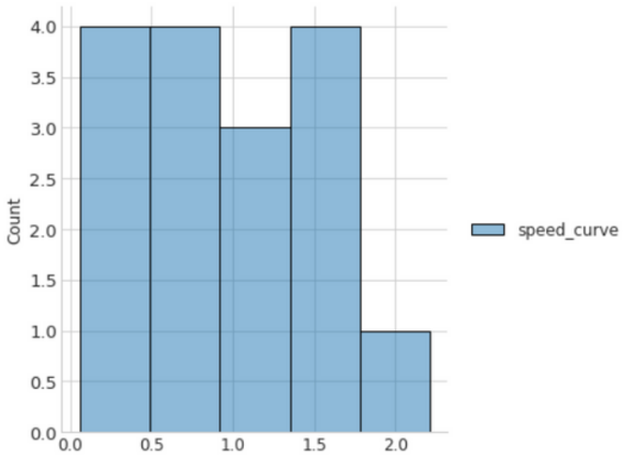
RMSE : On average, the prediction error is 1 mile per hour.

MODEL TESTED :OLS
<Figure size 360x360 with 0 Axes>



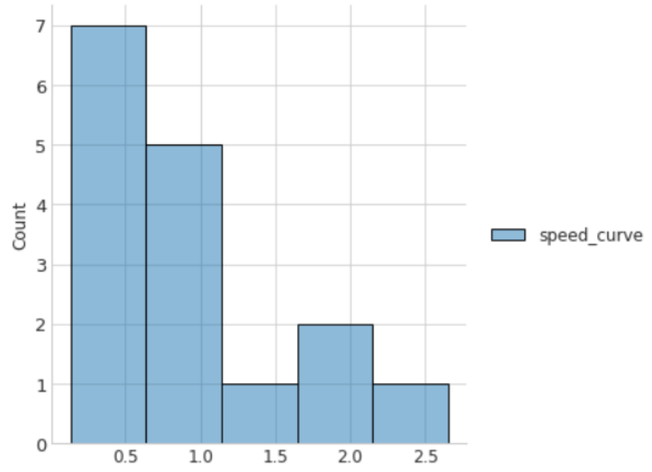
RMSE : 1.134154e+00

MODEL TESTED :LASSO
<Figure size 360x360 with 0 Axes>



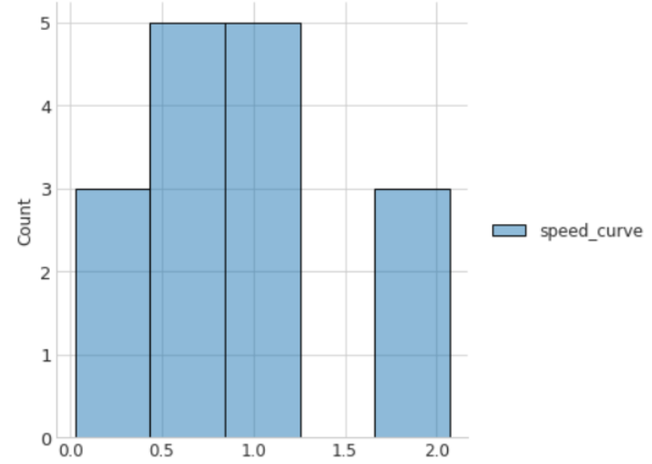
RMSE : 1.141511e+00

MODEL TESTED :RIDGE
<Figure size 360x360 with 0 Axes>



RMSE : 1.134136e+00

MODEL TESTED :XGB
<Figure size 360x360 with 0 Axes>

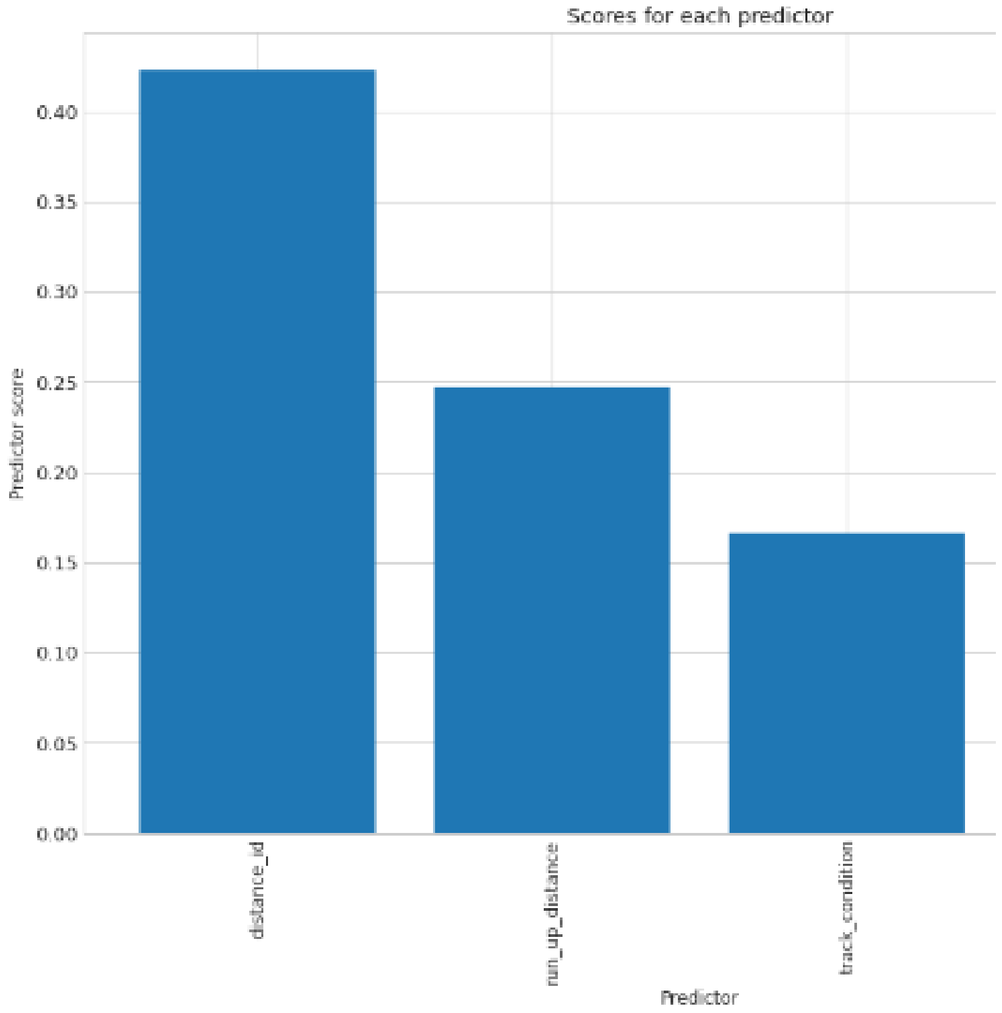


RMSE : 1.074866e+00

Injury Analysis

Our Viewpoint

- Predict speed in the curve in order to train horses and select the suitable race
- Distances and run-up-distance are the most useful features to predict speed.
- The model suffers from a lack of data.
- Extend the study to all races (racetrack, season, type)

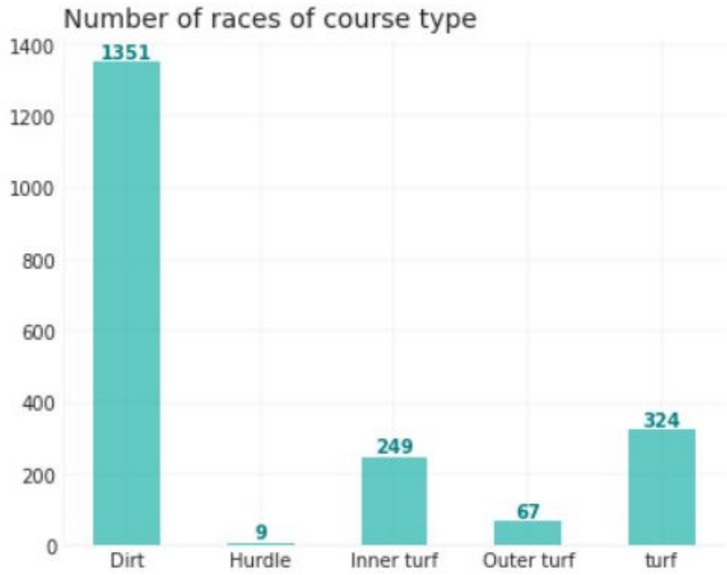


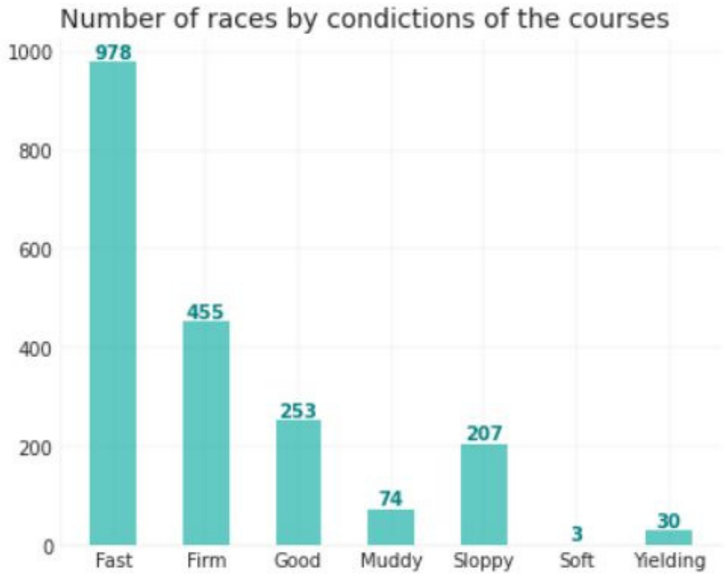
Data Modelling

Understand the Indicating Factors of a Win

Based on the EDA on the data, the following data processing is done:

- 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





Feature Engineering

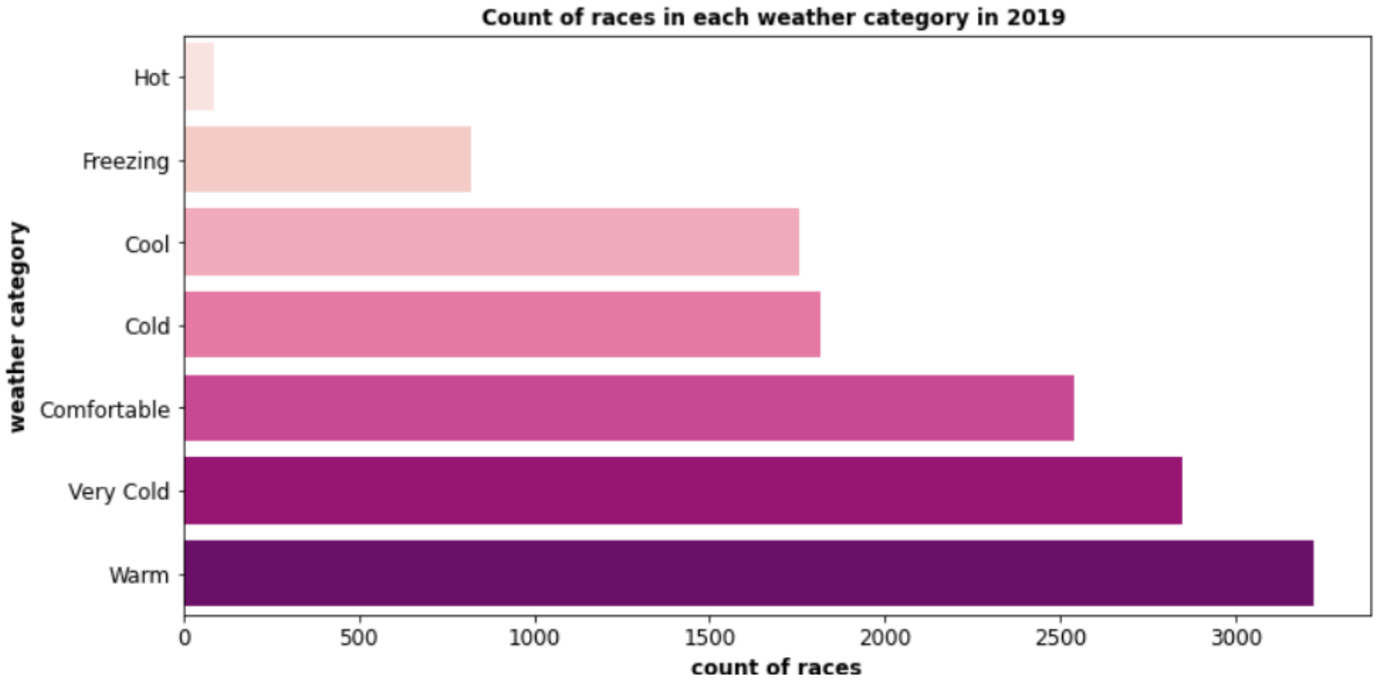
The descriptive statistical features of jockey's performances

- 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).

Why this feature?

Indeed, it is known that among Thoroughbreds, horses seem to prefer cold weather to warm, and their speed rises when the temperature sinks.



Training Machine Learning Model

To find the best hyper-parameters for LightGBM

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_child
_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 lightbgm, 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

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

Confusion matrix without Smote

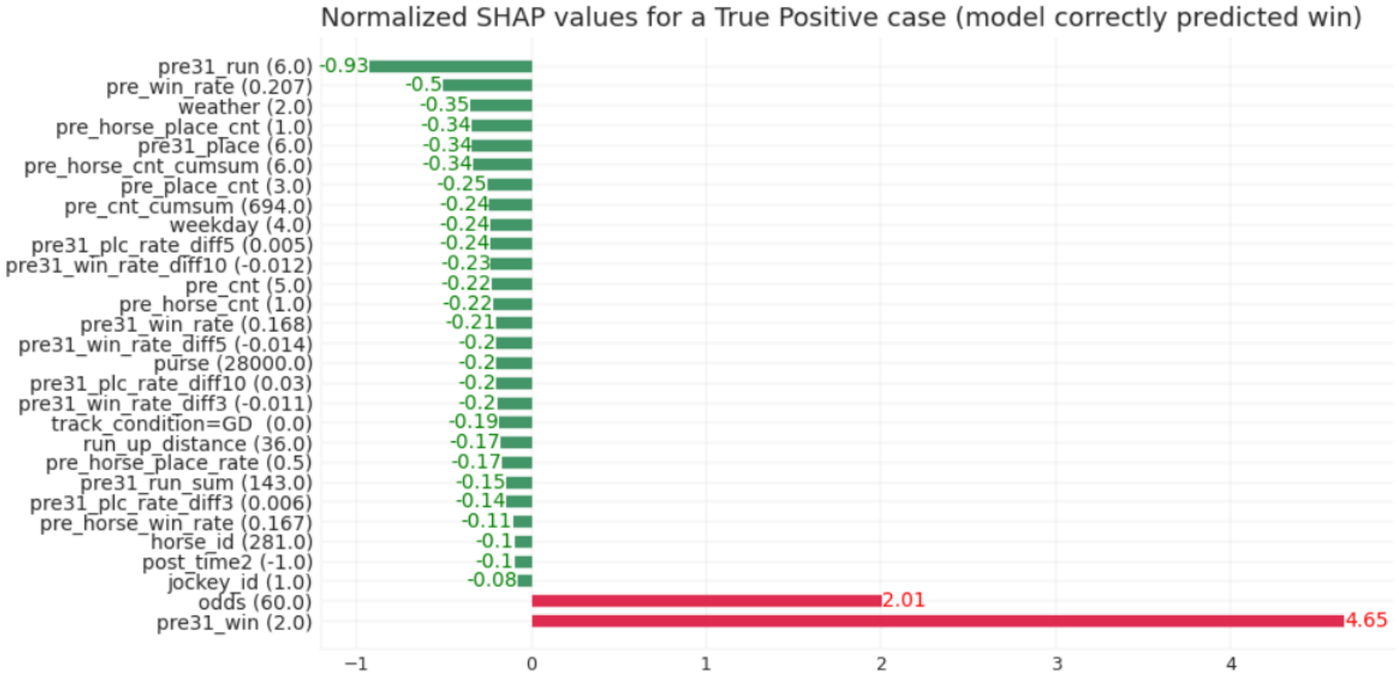
	Actual: Win	**Actual: No Win**
Predict: Win	156	108
Predict: No Win	191	2473

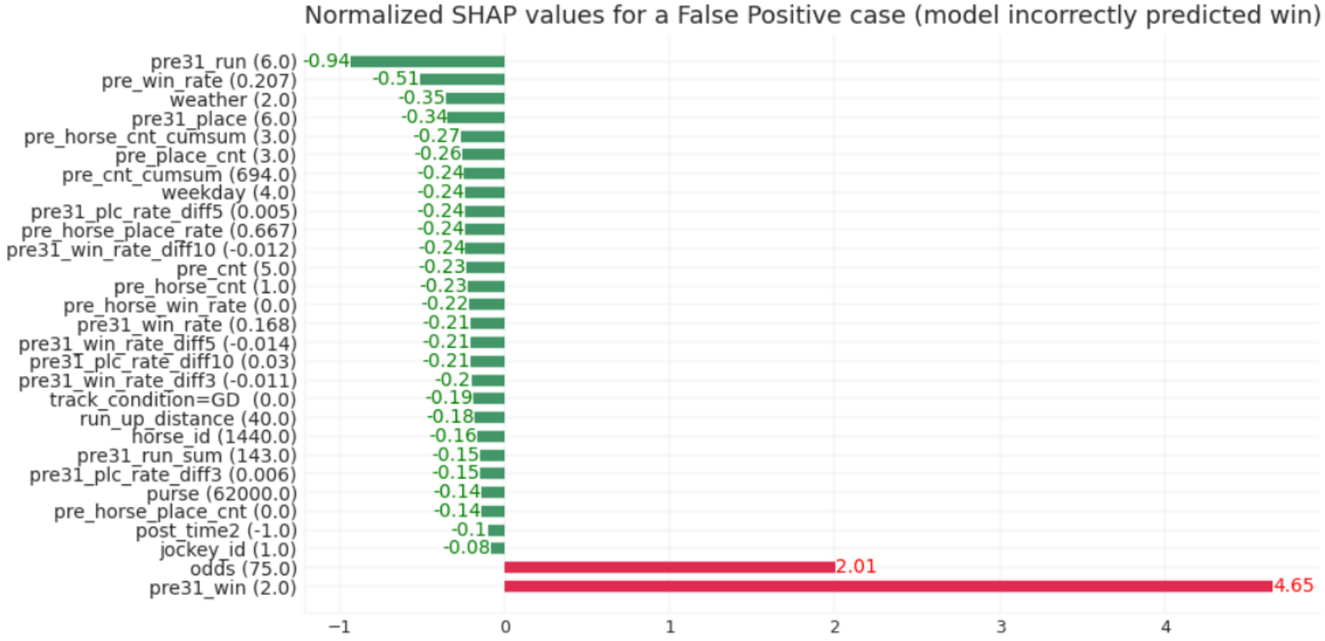
Confusion matrix after using Smote

Results

Indicating factors of Winning

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





Conclusion

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.

Next Step

- We can extend the injury analyses for other tracks and races.
- Study can be extended with the help of better and precise data about horse injuries.
- We can use robust feature extraction to get insights about drafting.

Thank You