Analysis of Horse Racing

Analyzing horse racing and building model to predict winners

Jimmy Chan

June 18, 2018

Abstract

The project analyzed the 2014-2016 Hong Kong racing season data from Hong Kong Jockey Club using Python and extracted features from the data. The purpose of this project is to examine the possibility of yielding a positive return on Hong Kong horse racing by statistical and machine learning approaches.

1 Introduction

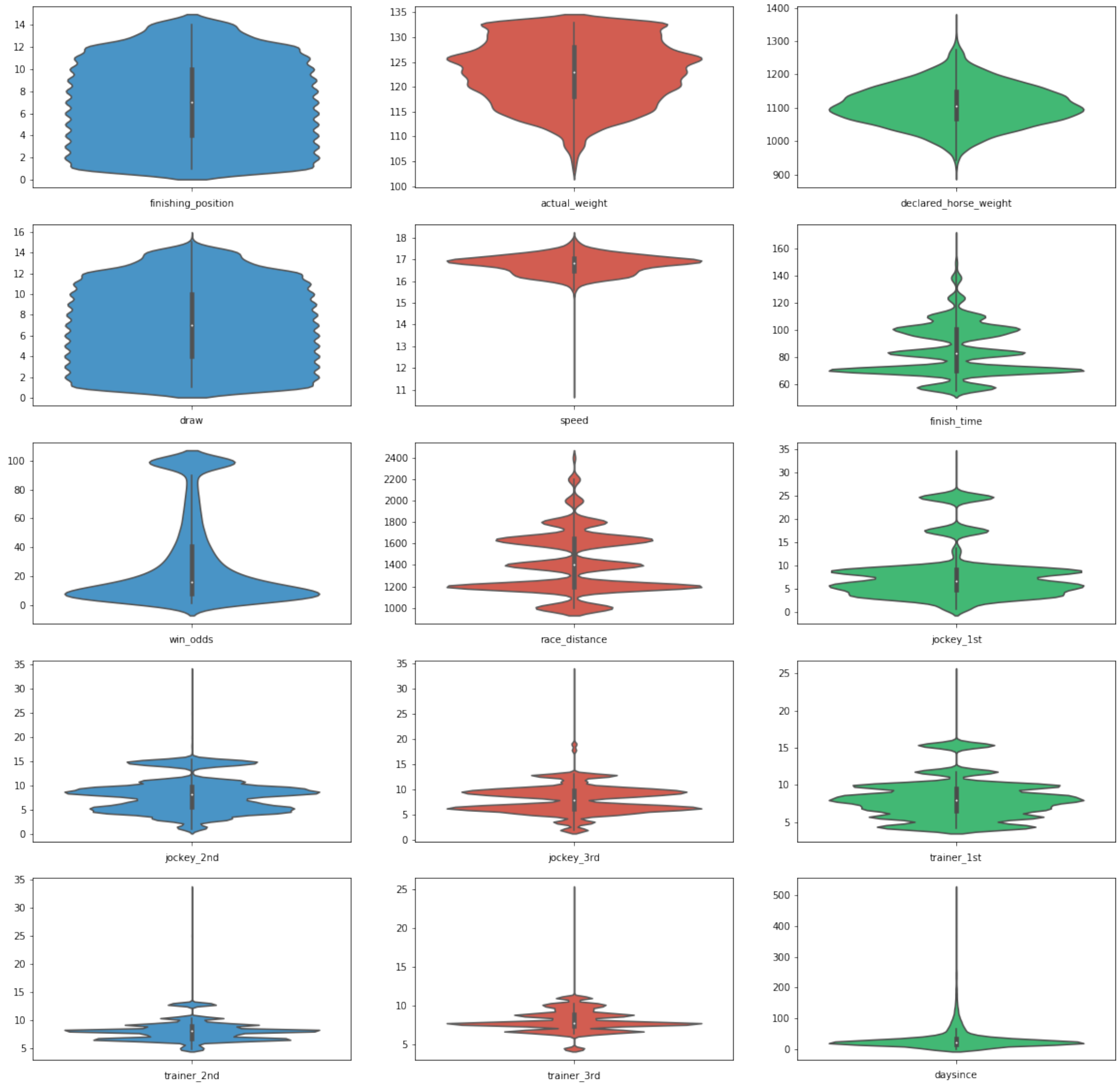
In this report, I will do the exploratory data analysis of the cleaned dataset that we have from the data wrangling notebook. We are trying to gain insight into the data set and identify the most important variables. Then, I will build a logistic regression model to predict the winner.

2 Exploratory data analysis

2.1 Predictor Variables Description

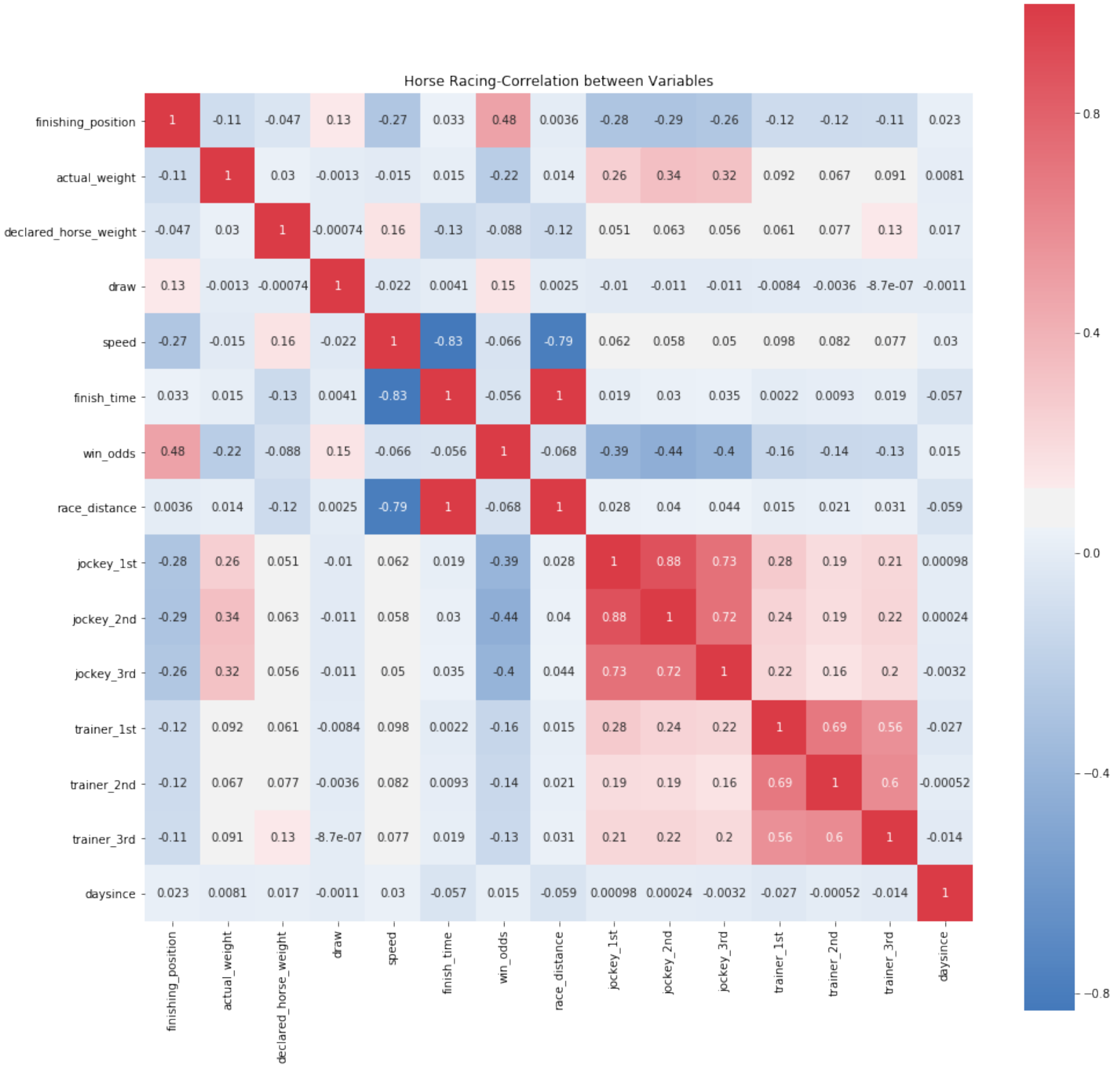
Finishing_position is the dependent variable which is the result of a horse race we want to predict. **Actual_weight, declared_horse_weight, draw, finish_time, win_odds, race_class, race_distance, jockey_1st, jockey_2nd, jockey_3rd, trainer_1st, trainer_2nd, trainer_3rd, daysince** are the independent variables for the prediction model which predict or forecast the values of the dependent variable in the model.

2.2 Basic Descriptive Statistics Plots



2.3 Correlation Map - correlation between Variables

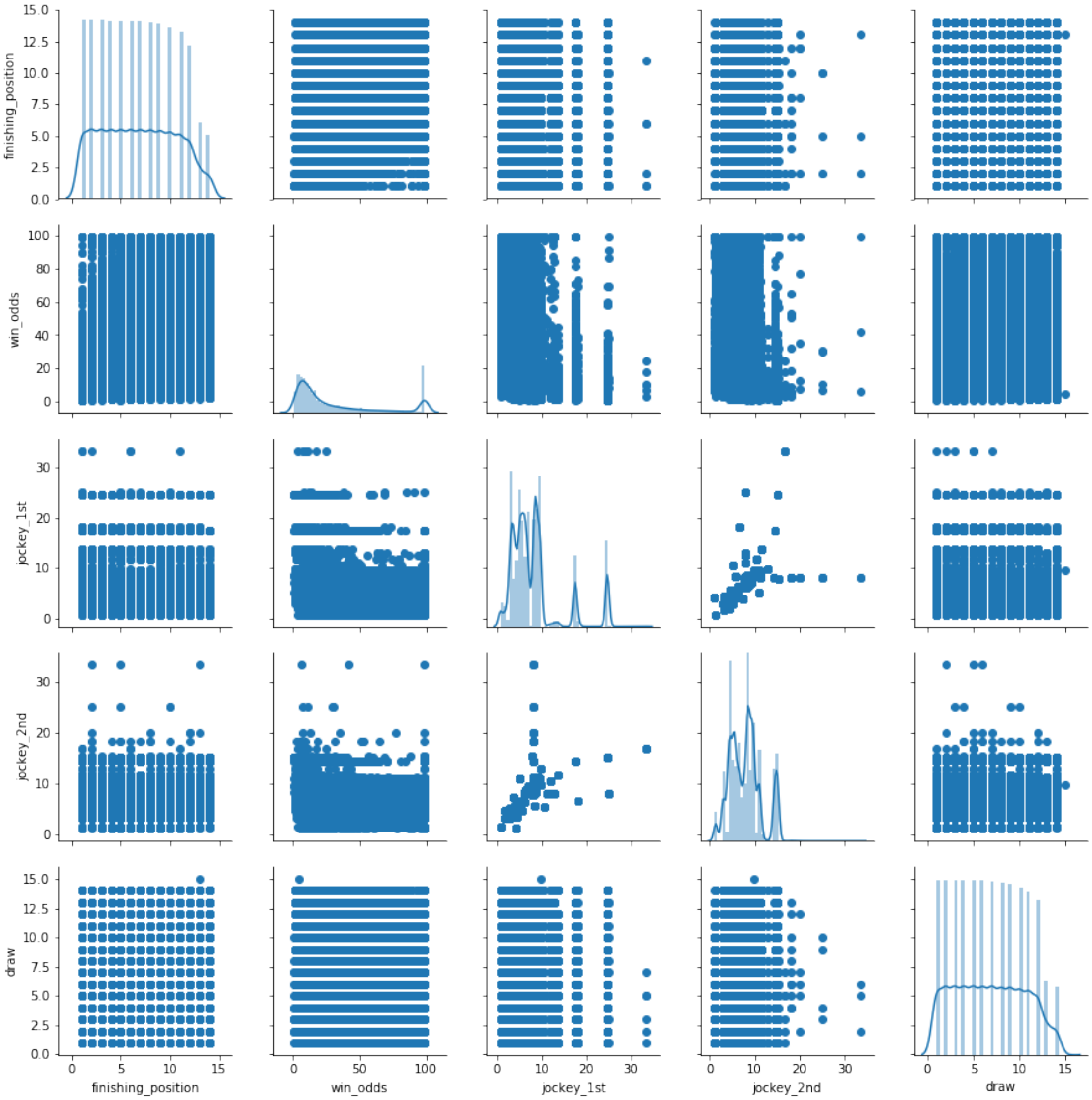
We can find the relationship between variables by plotting the correlation map.



We found that the correlation coefficient between race distance and finish time is 1. When the correlation coefficient is 1, it implies that race distance and finish time have a perfect positive relationship. It makes sense because it takes more time for a longer distance, and vice versa.

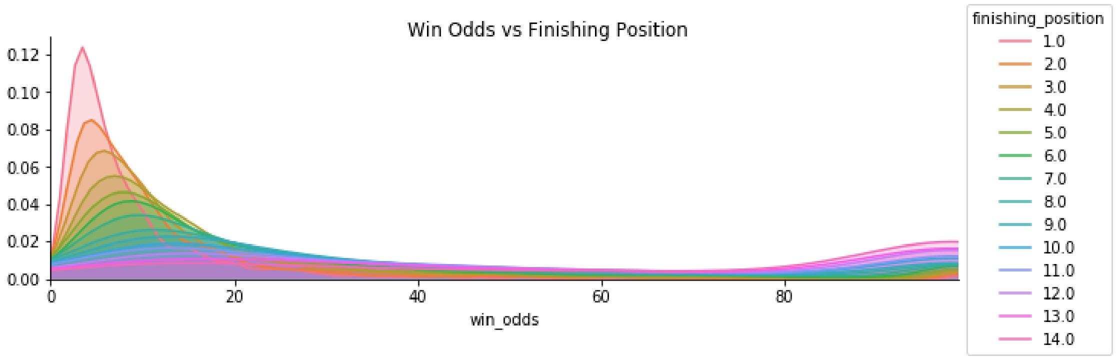
As we can see the win_odds, jockey_1st, speed and draw have higher negative/positive correlations(0.48, -0.28, -0.27, 0.13, 0.096) with the finishing_position. We can plot the plotting pairwise relationships between these variables.

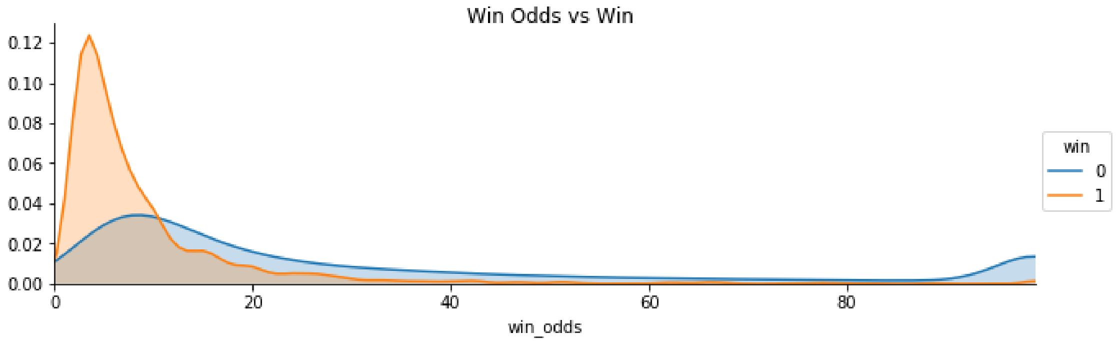
2.4 Pairwise Relationships Between Variables



2.5 Separate Plots and Analysis of High Correlations Variables vs Finishing Position and Win (Finishing Position =1)

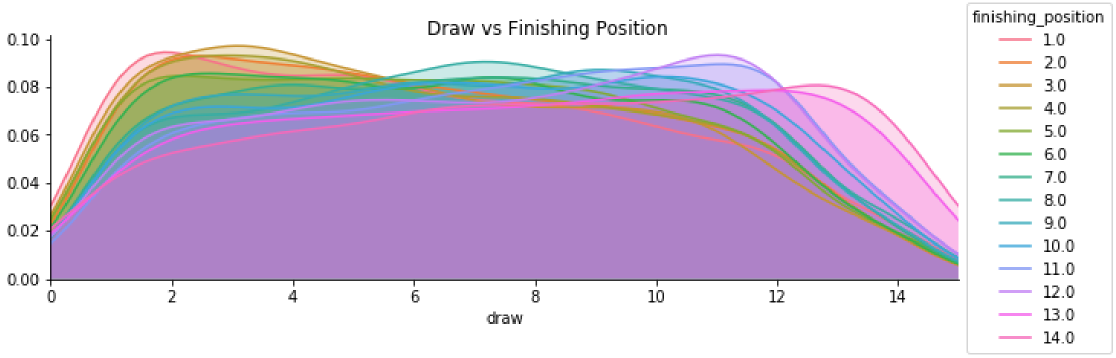
Win Odds vs Finishing Position (Win)

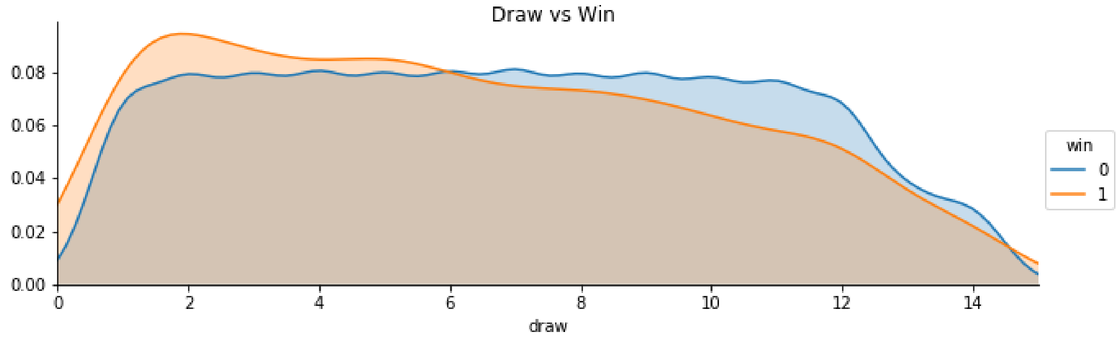




As we can see from the graphs above, it shows that win odds less than 15 have higher rates of a good finishing position, especially when finishing positions 1. It implies the lower the win odds, the higher the chance of winning.

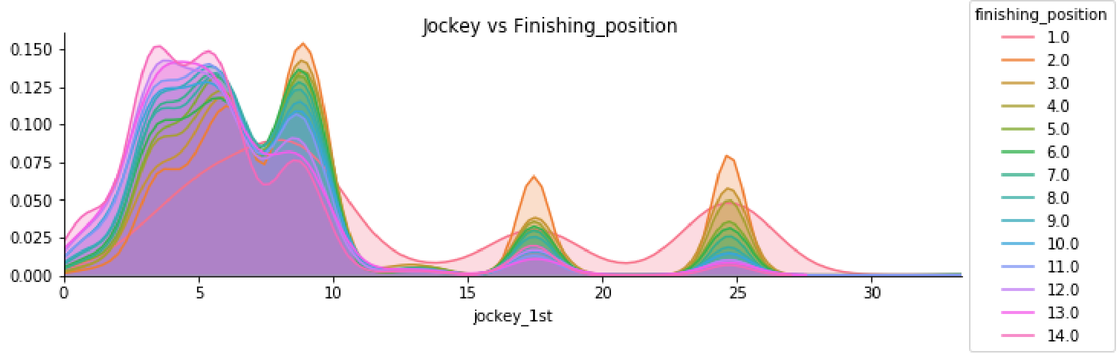
Draw vs Finishing Position (Win) Draw of a horse decides in which individual stall a horse is placed. The smaller the draw number, the closer the horse is to the inside rail, it means smaller draw numbers has a slight advantage over larger draw numbers since a shorter distance to be covered at the turns.

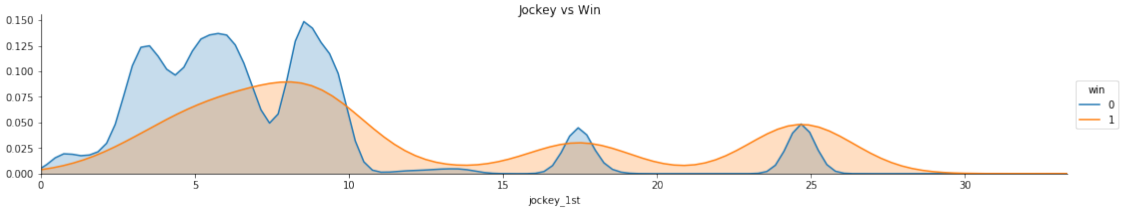




As we can see from the graphs above, it is proved that draw numbers smaller than 6 has an advantage over draw numbers larger than 6. And the smaller the draw number, the larger the advantage.

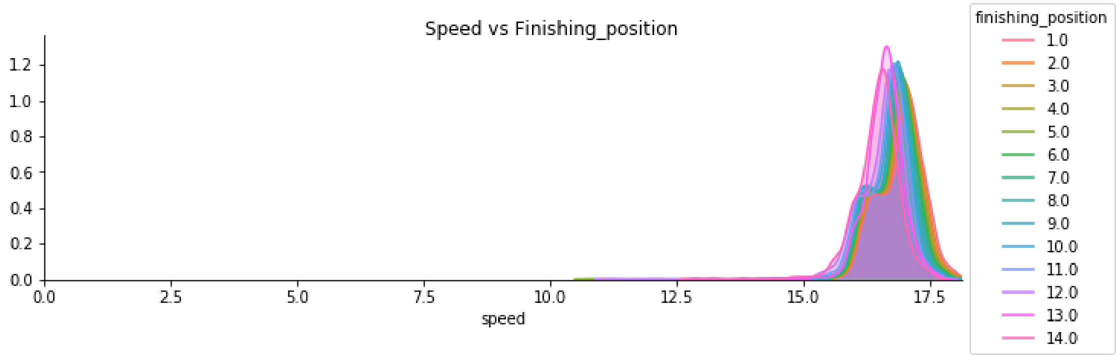
Winning Percentage of Jockey vs Finishing Position We can say that there is a relationship between jockey performance and finishing position based on common sense. But we don't know the impact of a jockey on horse racing by guessing. However, somebody believes a good horse will win despite the jockey.

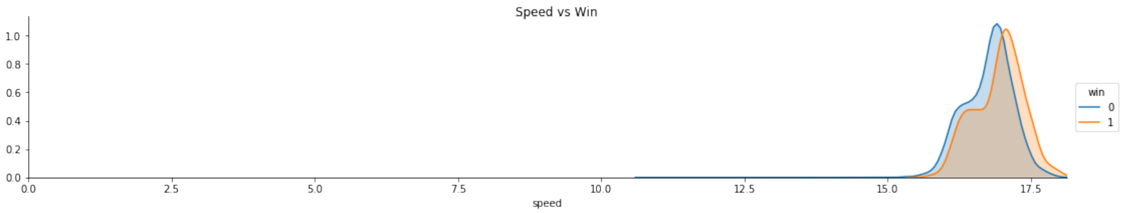




As we can see from the graphs above, it shows that the winning percentage of the jockey is less than 10 have much higher rates of not winning races based on our data set. It is proved that there is a strong relationship between jockey performance and finishing position.

Speed vs Finishing Position Speed should be one of the most important variables when we are picking the winner of a race. In the data wrangling, we divided the race distance(meter) by finish time(second) to get the speed. Therefore the speed is measured by meters per second.





As we can see from the graphs above, it shows that the speed faster than 17 meters per second has a better chance of winning. It is proved that there is a relationship between speed and finishing position but not a strong relationship.

3 Logistic Regression Model

```
"""
                                Results: Logit
==============================================================================
Model:                   Logit                  No. Iterations:     9.0000
Dependent Variable:      win                    Pseudo R-squared:   0.148
Date:                    2018-06-18 18:32       AIC:                9889.6925
No. Observations:        20554                  BIC:                10111.7552
Df Model:                27                     Log-Likelihood:     -4916.8
Df Residuals:            20526                  LL-Null:            -5768.0
Converged:               1.0000                 Scale:              1.0000
------------------------------------------------------------------------------
                             Coef.  Std.Err.    z     P>|z|    [0.025   0.975]
------------------------------------------------------------------------------
actual_weight               -0.0310   0.0046  -6.7626 0.0000 -0.0399 -0.0220
```

```
declared_horse_weight       -0.0008   0.0005  -1.8294 0.0673 -0.0017  0.0001
draw                        -0.0107   0.0074  -1.4498 0.1471 -0.0253  0.0038
speed                        0.2190   0.0401   5.4642 0.0000  0.1404  0.2975
win_odds                    -0.0692   0.0036 -18.9835 0.0000 -0.0763 -0.0620
race_distance               -0.0000   0.0001  -0.2272 0.8203 -0.0002  0.0002
jockey_1st                   0.0489   0.0099   4.9365 0.0000  0.0295  0.0683
jockey_2nd                  -0.0035   0.0178  -0.1991 0.8422 -0.0385  0.0314
jockey_3rd                  -0.0186   0.0157  -1.1836 0.2366 -0.0493  0.0122
trainer_1st                  0.0756   0.0152   4.9771 0.0000  0.0458  0.1053
trainer_2nd                 -0.0520   0.0219  -2.3733 0.0176 -0.0949 -0.0091
trainer_3rd                 -0.0690   0.0262  -2.6304 0.0085 -0.1204 -0.0176
daysince                     0.0006   0.0009   0.7123 0.4763 -0.0011  0.0023
Class 2                     -0.1746   0.1891  -0.9233 0.3559 -0.5453  0.1961
Class 3                     -0.1615   0.1750  -0.9229 0.3561 -0.5044  0.1815
Class 3 (Special Condition)  0.5318   0.8110   0.6557 0.5120 -1.0577  2.1214
Class 4                     -0.1589   0.1752  -0.9073 0.3642 -0.5023  0.1844
Class 4 (Restricted)        -0.0951   0.5172  -0.1839 0.8541 -1.1088  0.9186
Class 4 (Special Condition) -0.3456   0.6370  -0.5425 0.5875 -1.5942  0.9030
Class 5                     -0.1713   0.1852  -0.9248 0.3550 -0.5342  0.1917
Griffin Race                 0.1612   0.3693   0.4366 0.6624 -0.5626  0.8851
Group One                   -0.0161   0.2976  -0.0543 0.9567 -0.5995  0.5672
Group Three                 -0.0643   0.4116  -0.1563 0.8758 -0.8711  0.7424
Group Two                    0.1362   0.3981   0.3423 0.7321 -0.6439  0.9164
Hong Kong Group One          0.4970   0.4400   1.1296 0.2586 -0.3654  1.3594
Hong Kong Group Three       -0.2870   0.3166  -0.9066 0.3646 -0.9075  0.3335
Hong Kong Group Two         -0.2074   0.5213  -0.3979 0.6907 -1.2291  0.8142
Restricted Race              0.2967   0.5421   0.5472 0.5842 -0.7659  1.3592
============================================================================

"""
```

As we can see from the logistic regression model summary, most variables have small p-values which mean they are statistically significant. However, some variables have a large P-value which can be removed from the model, e.g., race_distance, jockey_2nd. Meanwhile, some variables have a p-value between 0.2 to 0.5 which is not statistically significant, but we cannot determine whether we should remove it or not unless we do the cross-validation.