QF 206 - Quantitative Trading Strategies (G1)
AY2023/24 Term 2

Written By: Group 3

Ian Chia Chern Yi	01415127
Juan Sebastian	01409757
Justin Ho Rui Hwa	01388166
Tham Sheng Le	01370375

Table of Contents

1.0 Executive Summary

The research aimed to explore potential opportunities within the Hong Kong Horse Betting Market. However, initial findings indicated that conventional betting strategies did not consistently yield positive returns, resulting in a depletion of portfolio wealth. To address this, quantitative trading strategies and machine learning techniques were employed to enhance existing strategies and assess the feasibility of generating positive returns given available data and payouts.

Utilizing data sourced from the Hong Kong Jockey Club (HKJC), various benchmarking and betting strategies were developed to exploit market factors and potentially uncover alpha. The Kelly Criterion was employed to appropriately size bets, maximizing the return-to-risk ratio and optimizing long-term betting placements.

Upon initial analysis of wealth curves and strategy performance, it was observed that incorporating the Exponential Moving Average (EMA) Unbiased Average Speed variable improved predictive ability. Subsequently, machine learning techniques, including regression and classification models, were utilized to predict horse performance metrics such as average race speed, finish time, and probability of finishing in the top three positions, aiding in betting decisions.

Regression models for race speed and finish time exhibited profitability within a limited timeframe, from October 12, 2016, to July 16, 2017, with prediction accuracies of 57.3% and 55.1%, respectively. Conversely, classification models, particularly for finish position, displayed lower prediction accuracies (53.9%) and diminished returns compared to regression models. This discrepancy could be attributed to classification models predominantly identifying bets with lower win odds, thereby restricting potential profits.

When implementing the Kelly Criterion for bet allocation, absolute profits are reduced. However normalised average returns per bet consistently outperformed flat betting across all methods and target variables, highlighting the effectiveness of risk-adjusted Kelly betting strategies in maximizing returns.

Best Performance from each Solutions:

Performance Statistic	Benchmark	Naive Mean Variance		EMA Mean Variance	
Target Variable	NA	Speed		Time	
Bet Type	**Fixed Bet**	**Fixed Bet**	**Kelly**	**Fixed Bet**	**Kelly**
Absolute Returns	-$32,375.00	$15,650.00	$35,226.72	-$43,470.00	$5,388.03
Mean Returns (%)	-13.68%	6.58%	1.02%	-18.19%	7.02%
Standard Deviation (%)	135.30%	517.17%	497.00%	419.34%	497.74%
Max Drawdown	-$32,800.00	-$25,190.00	-$5,610.39	-$46,020.00	-$2,533.67
Sharpe Ratio	-0.1011	0.0127	0.0021	-0.0434	0.0141

Performance Statistic	Benchmark	Regression		Classification	
Target Variable	NA	Speed		Rankings	
Bet Type	**Fixed Bet**	**Fixed Bet**	**Kelly**	**Fixed Bet**	**Kelly**
Absolute Returns	-$32,375.00	$41,250.00	$30,360.08	$8,430.00	$4,914.48
Mean Returns (%)	-13.68%	4.57%	16.50%	1.59%	4.97%
Standard Deviation (%)	135.30%	42.21%	140.59%	45.39%	128.53%
Max Drawdown	-$32,800.00	-$5,010.00	$2,980.16	-$6,230.00	-$2,820.93
Sharpe Ratio	-0.1011	0.1081	0.1174	0.0350	0.0387

2.0 Introduction to the Business Problem

The project employs techniques commonly used in Quantitative Trading to investigate potential opportunities within the Hong Kong Horse Betting Market, utilizing data from the Hong Kong Jockey Club (HKJC). Following an Exploratory Data Analysis (EDA) to understand market characteristics, the analysis focuses on win-odds behaviour, regression analysis of various influencing factors, and devising a benchmarking strategy for evaluating performance. The goal is not to guarantee profits but to apply quantitative finance methods to enhance betting performance, emphasizing systematic and data-driven decision-making to potentially uncover market edges.

2.1 Introduction to Horse Betting

Central to the horse betting experience in Hong Kong is the pari-mutuel betting system, wherein all bets are pooled together into a single betting pool. The odds for each horse are dynamically determined based on the total amount wagered on that horse in relation to the overall pool. This fluidity ensures that odds reflect real-time betting activity, with higher odds indicating lower levels of betting on a particular horse and vice versa. Payouts for pari-mutuel betting are detailed in Appendix 6.4 Figure 1, offering transparency and fairness to participants.

In this system, bettors compete against each other rather than against the house or bookmaker, fostering a dynamic and competitive betting environment. The fluctuating odds represent the collective insights and sentiments of the betting public. Following the race's conclusion and outcome determination, the total betting pool is distributed among winning bettors, with deductions for taxes, operational costs, and charitable contributions by the HKJC. Events offering pari-mutuel betting are outlined in Appendix 6.4 Figure 2, showcasing the breadth of opportunities available for participation in the thrilling world of horse racing.

2.2 The Data Set

The data is taken from Kaggle (Camara & Cheng, 2017) which is pulled from the HKJC (The Hong Kong Jockey Club*,* 2024). The data consist of 2 parts: the horses race data, and the racetrack data. The horses race data contains details such as the weights of the horse, the lane the horse, win odds, etc. The racetrack data contains details such as the weather condition, the distance of the track, etc.

2.3 Target Measures

The performance of the trading strategies will be compared to the benchmark. The primary comparison will be based on the mean returns (%), standard deviations (%), max drawdown, and Sharpe ratio. Mean returns, standard deviations, and Sharpe ratios are used as comparison as they normalize the values between different betting strategies. Max drawdown is used as a risk measure as it allows us to anticipate the minimum amount to have on hand to avoid a wipeout.

3.0 Mean Variance Solution

Target Variable 1: Average Time

Idea

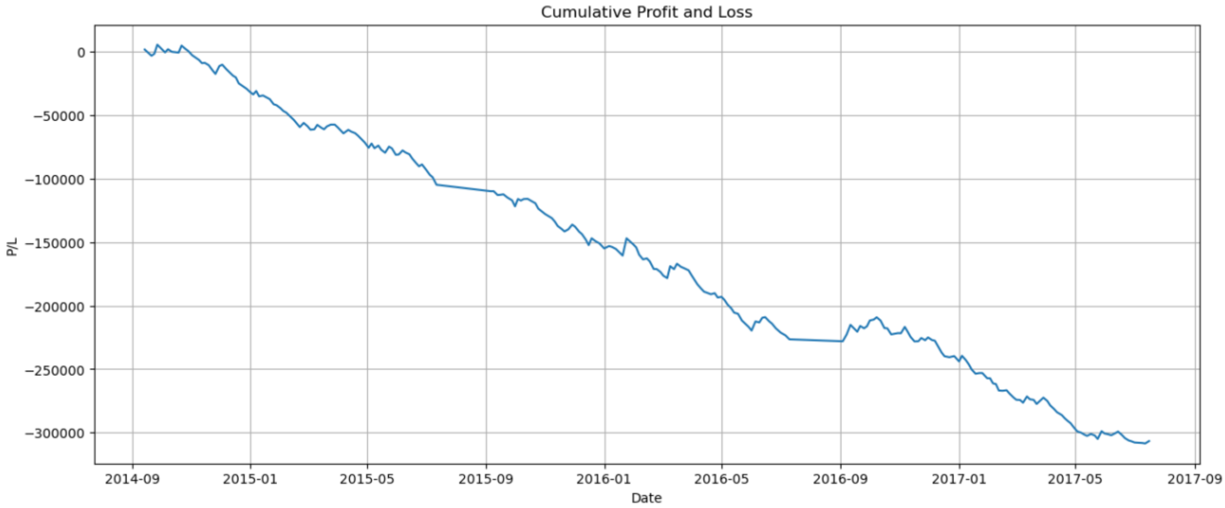
A straightforward approach suggests that horses with shorter average completion times for races are more likely to win.

Method

To estimate the mean and variance of a particular horse's track time, we group the data by the horses' ID and compute the sample mean and sample variance. Assuming a normal distribution for the horses' running statistic, we then conduct Monte Carlo simulations to calculate the probability of winning. This approach is adopted because exact computation of these values may be complex, especially when dealing with multiple variables. Instead, we opt for Monte Carlo simulations to simulate runs and determine the probability of a horse winning a race.

Betting Strategy 1: Betting when empirical > implied win probability

A straightforward betting strategy involves placing a $100 bet on a horse if our estimated win probability from the Monte Carlo simulation exceeds the implied win probability of the race. The implied win probability is determined by taking the reciprocal of the win odds assigned to the horse in that race. Theoretically, this strategy should yield more wins than losses over the long run if our estimation is accurate. Additionally, to bet on horses that we are confident in, we only bet on the horses that has ran at least 10 races.



Win Rate: 1.18 %	Total Pct Return: -30.11 %	Median: -100.0 %
Total Bets: $1,018,100	Mean: -30.11 %	Max Drawdown: -$315,970.0
Total Return: -$306,580.0	Standard Deviation: 498.73 %	Sharpe Ratio: -0.0604

However, we lost money. Clearly the model did not work. However, it might be possible that, the betting strategy is also at fault as we are not betting optimally. Thus, this led us to betting strategy 2.

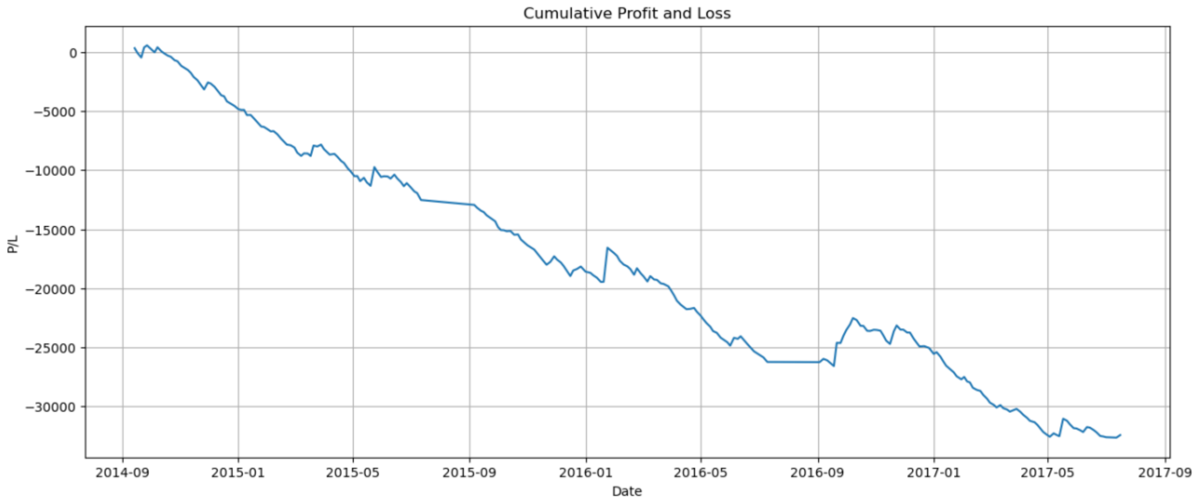
Betting Strategy 2: Fixed Max Bet Kelly Criterion

The Kelly Criterion (Appendix 6.1) is helps to calculate what percentage of capital should be risked maximizing returns in the long run.

In Strategy 2, we implement a fixed maximum bet size for each horse. The Kelly criterion is then applied against this fixed maximum bet size to determine the appropriate betting amount. For instance, if the fixed max bet size is $100 and the Kelly criterion suggests allocating 10% of capital for horse X and 20% for horse Y, the bet amounts would be $10 for horse X and $20 for horse Y.

Importantly, in subsequent races, regardless of the outcome of the previous race, the fixed maximum bet size remains unchanged at $100.

This strategy aims to optimize betting amounts based on the Kelly criterion while mitigating the risk of excessive loss or capital depletion by bringing fresh capital to capture winnings create by long shot horses.



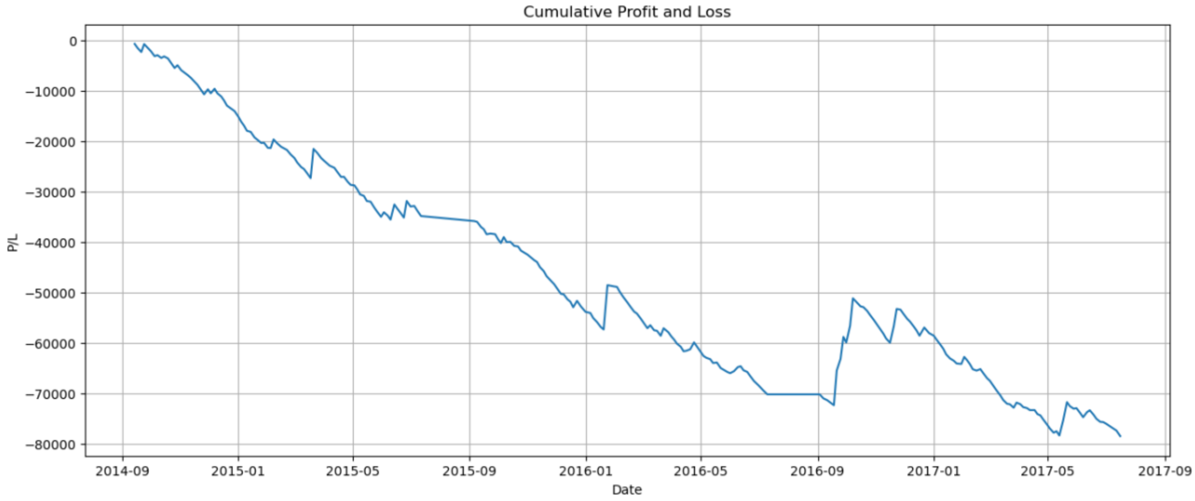
Win Rate: 1.18 %	Total Pct Return: -34.7 %	Median: -100.0 %
Total Bets: $93,512.92	Mean: -30.14 %	Max Drawdown: -$33,572.63
Total Return: -$32,449.47	Standard Deviation: 498.63 %	Sharpe Ratio: -0.0604

However, the performance did not improve over strategy 1. A potential issue identified could be due to the multiple bets placed on multiple horses in a single race. This was a potential problem as there can only be 1 actual winner. This leads us to betting strategy 3.

Betting Strategy 3: Betting on Horse with Highest Win Probability

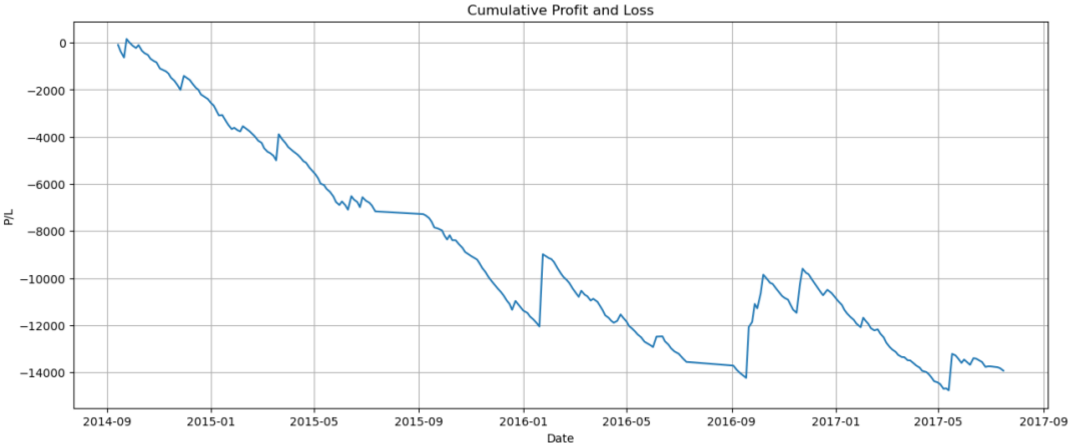
This strategy hypothesis that the reason for no improvement seen between strategy 1 and 3 is caused by multiple bets on multiple horses in a single race. Such bets require the assumptions that each bet placed on each horse in a race should be independent of one another. However, they are not independent of each other as there can only be 1 winner. Thus, the new approach is to only bet on the horse with highest empirical win probability in a race. This also gives us the advantage of having, theoretically, lower variance in our returns.

Without Kelly:



Win Rate: 0.55 %	Total Pct Return: -33.0 %	Median: -100.0 %
Total Bets: $238,000	Mean: -32.98 %	Max Drawdown: -$78,710.0
Total Return: -$78,530.0	Standard Deviation: 431.13 %	Sharpe Ratio: -0.0765

With Kelly:



Win Rate: 0.28 %	Total Pct Return: -34.4 %	Median: -100.0 %
Total Bets: $40,496.1	Mean: -27.89 %	Max Drawdown: -$15,059.38
Total Return: -$13,932.04	Standard Deviation: 486.98 %	Sharpe Ratio: -0.0573

This betting strategy seems to have slightly better performance than strategy 1 and 2 based on its Sharpe ratio.

Target Variable 2: Average Speed

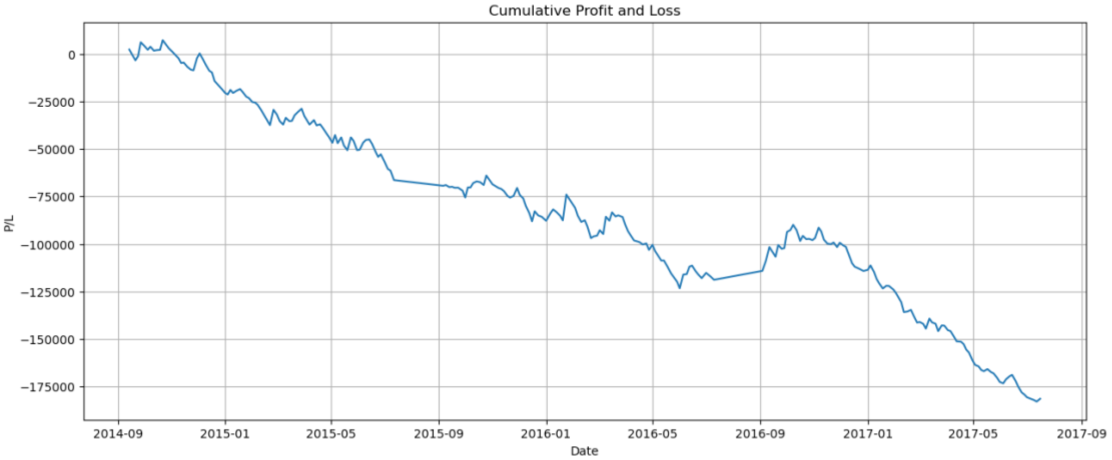
Idea

Upon further examination of the data, it becomes apparent that different racetracks feature varying lengths for the races, which directly impacts the performance of the horses. Considering this observation, a potential avenue for improvement lies in adjusting the target variable to incorporate the speed of the horse, thus accounting for the differences in racetrack lengths.

Method

The same method is applied to how we processed the previous time variable. To estimate the mean and variance of a particular horse's speed, we group the data by the horse ID and compute the sample mean and sample variance. Assuming a normal distribution, we then conduct Monte Carlo simulations to calculate the probability of winning.

Betting Strategy 1: Betting when empirical > implied win probability

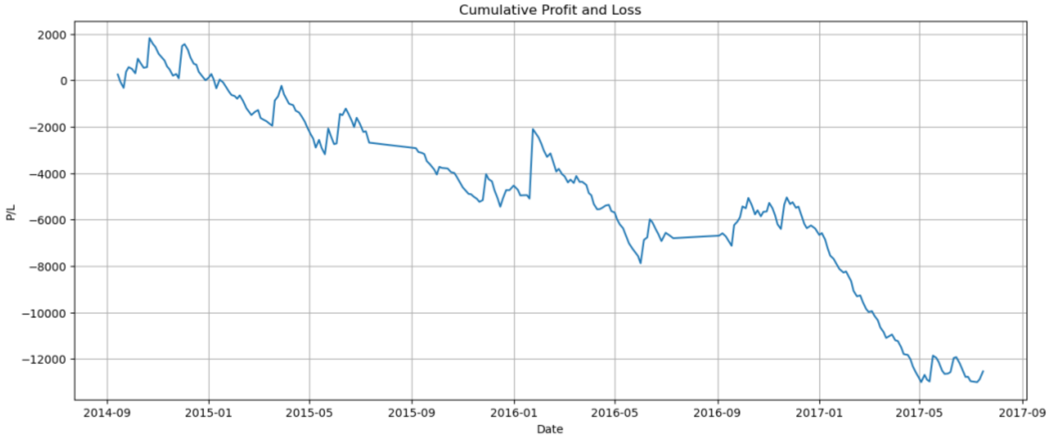


Win Rate: 1.62 %
Total Bets: $1,040,300
Total Return: -$181,500.0

Total Pct Return: -17.45 %
Mean: -17.45 %
Standard Deviation: 520.13 %

Median: -100.0 %
Max Drawdown: -$193,270.0
Sharpe Ratio: -0.0335

Betting Strategy 2: Fixed Max Bet Kelly Criterion



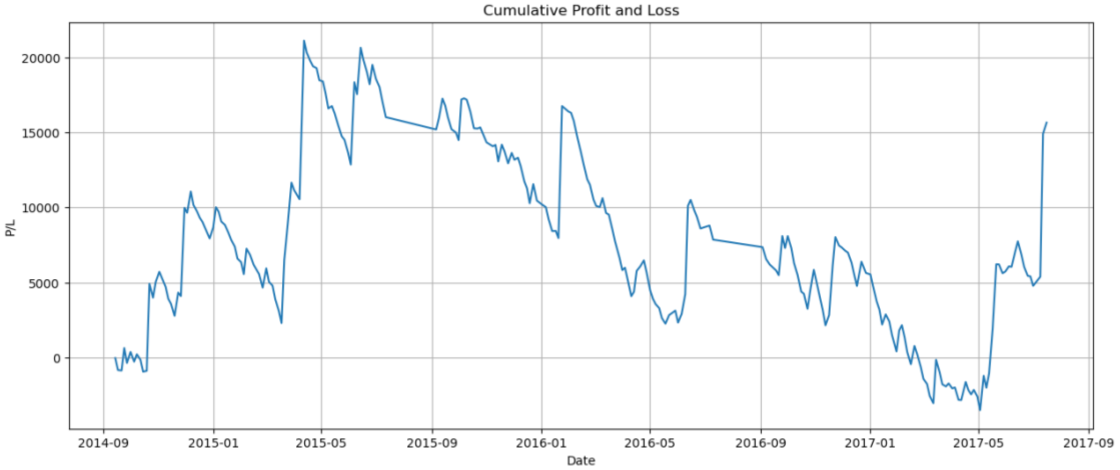
Win Rate: 1.62 %
Total Bets: $83,656.24
Total Return: -$12,534.09

Total Pct Return: -14.98 %
Mean: -17.47 %
Standard Deviation: 520.06 %

Median: -100.0 %
Max Drawdown: -$15,222.38
Sharpe Ratio: -0.0336

Betting Strategy 3: Betting on Horse with Highest Win Probability

Without Kelly Criterion:

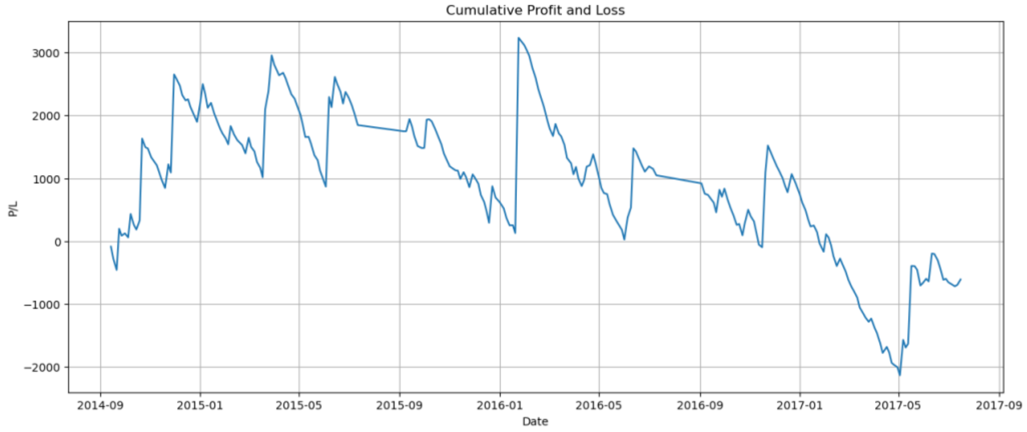


Win Rate: 1.04 %
Total Bets: $237,900
Total Return: $15,650.0

Total Pct Return: 6.58 %
Mean: 6.58 %
Standard Deviation: 517.17 %

Median: -100.0 %
Max Drawdown: -$25,190.0
Sharpe Ratio: 0.0127

With Kelly Criterion:



Win Rate: 0.5 %	Total Pct Return: -1.73 %	Median: -100.0 %
Total Bets: $35,226.72	Mean: 1.02 %	Max Drawdown: -$5,610.39
Total Return: -$609.82	Standard Deviation: 497.0 %	Sharpe Ratio: 0.0021

Using speed as the target variable yields improved results across the board. This is particularly evident with betting strategy 3, where the strategy starts to make profit and Sharpe ratio starts to enter the positive region. Moving forward, we will maintain our focus on betting strategy 3.

Target Variable 3: Unbiased Average Speed

Idea

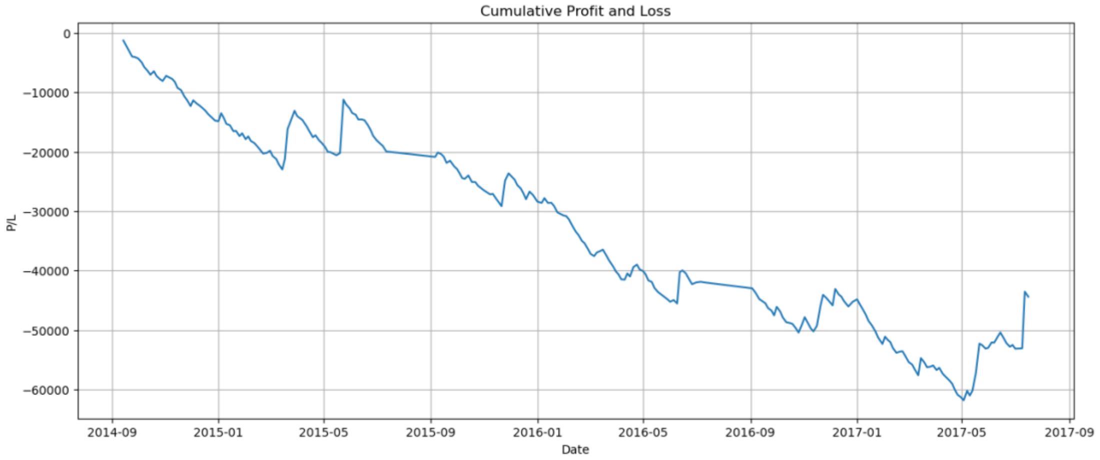
Analysing target variable 1 and 2, it is possible that we might have given ourselves forward bias as we calculated the mean and variance of the horse across all its race ahead of betting.

Method

To address this issue, we implement an expanding window approach, considering only the races that have already occurred to calculate the average speed of the horse. Subsequently, we employ the same methodology used for the previous two target variables through Monte Carlo simulation.

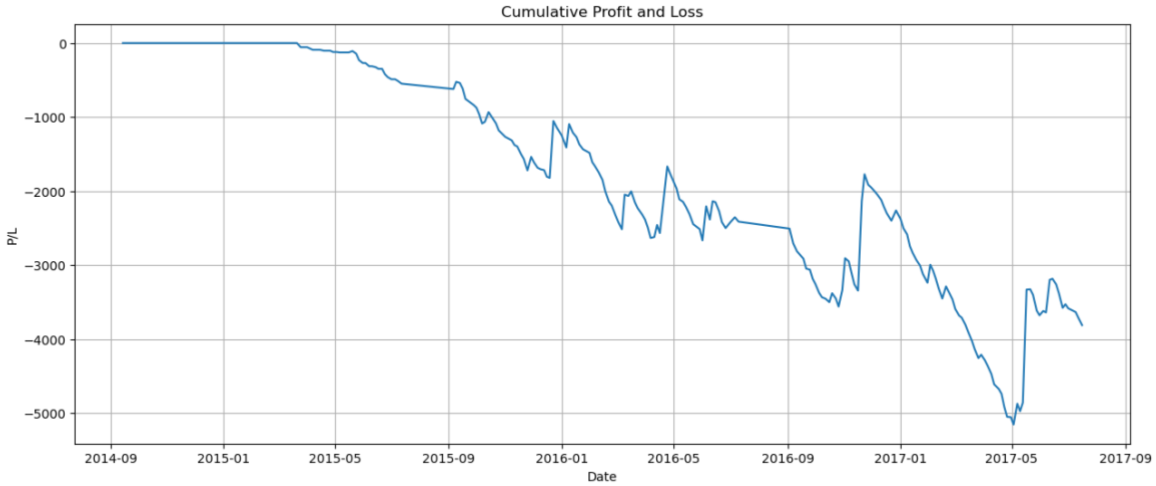
Betting Strategy 3: Betting on Horse with Highest Win Probability

Without Kelly:



Win Rate: 0.86 %	Total Pct Return: -18.63 %	Median: -100.0 %
Total Bets: $238,200	Mean: -18.63 %	Max Drawdown: -$62,400.0
Total Return: -$44,380.0	Standard Deviation: 434.72 %	Sharpe Ratio: -0.0429

With Kelly:



Win Rate: 0.18 %	Total Pct Return: -22.25 %	Median: -100.0 %
Total Bets: $17,141.24	Mean: -24.26 %	Max Drawdown: -$5,266.94
Total Return: -$3,813.26	Standard Deviation: 352.53 %	Sharpe Ratio: -0.0688

The model performs worse as expected as forward bias has been removed.

Target Variable 4: EMA Unbiased Average Time & Speed

Idea

A horse more recent performance is more relevant than older performance. This can be due to gain in experiences, muscle mass, age, etc. Thus, this presents a potential alpha based on momentum strategy.

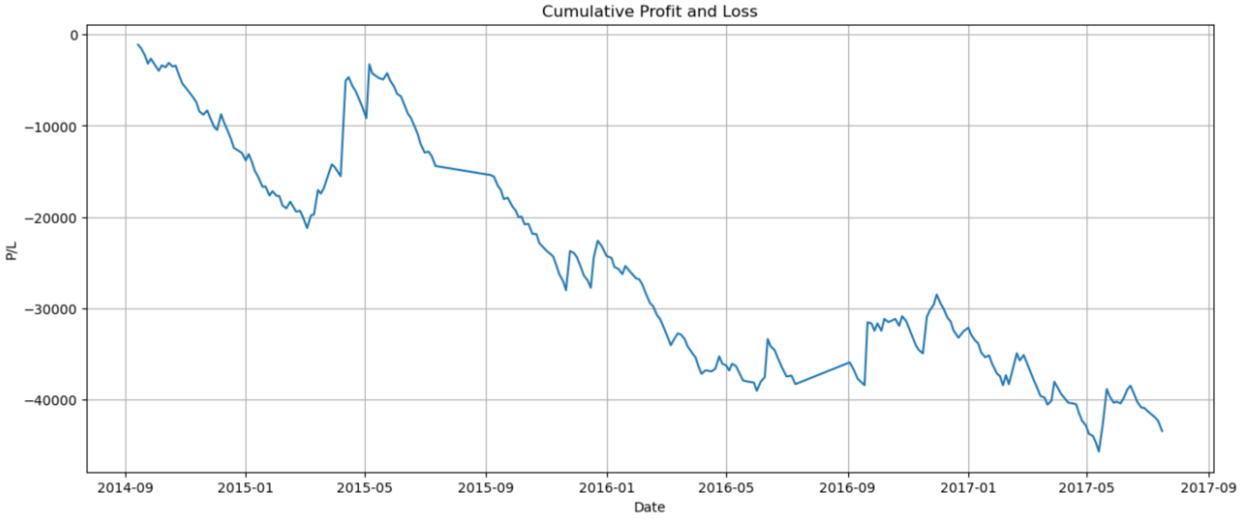
Method

Setting $\alpha = 0.2$

Presuming that horses' momentum influences their performance, we employed Exponential Moving Average (Appendix 6.2) on both the time and speed variables. Subsequently, we applied the same Monte Carlo simulation methodology used previously. We set $\alpha = 0.2$ which is equal to an 11-data point Moving Average.

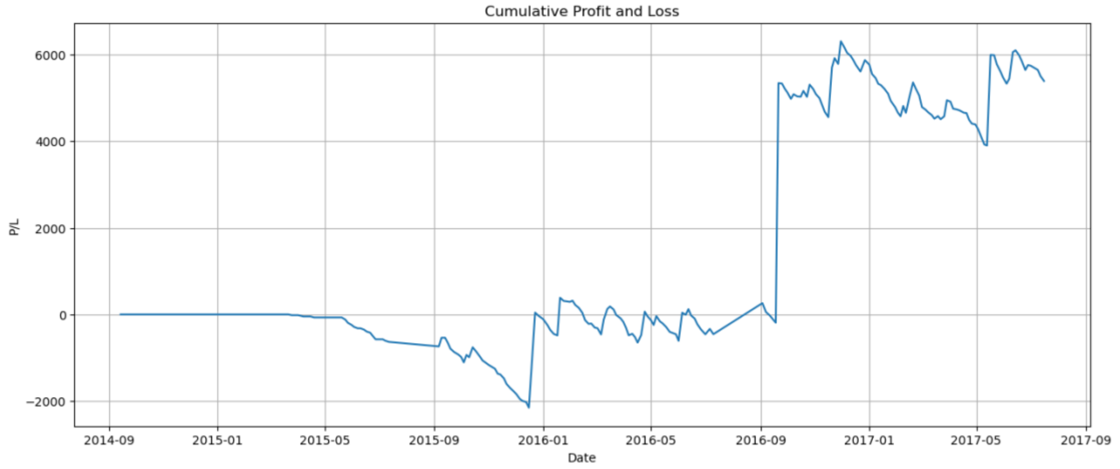
Betting Strategy 3: Betting on Horse with Highest Win Probability

Using Time as target variable (Without Kelly):



Win Rate: 0.74 %	Total Pct Return: -18.19 %	Median: -100.0 %
Total Bets: $239,000	Mean: -18.19 %	Max Drawdown: -$46,020.0
Total Return: -$43,470.0	Standard Deviation: 419.34 %	Sharpe Ratio: -0.0434

Using Time as target variable (With Kelly):

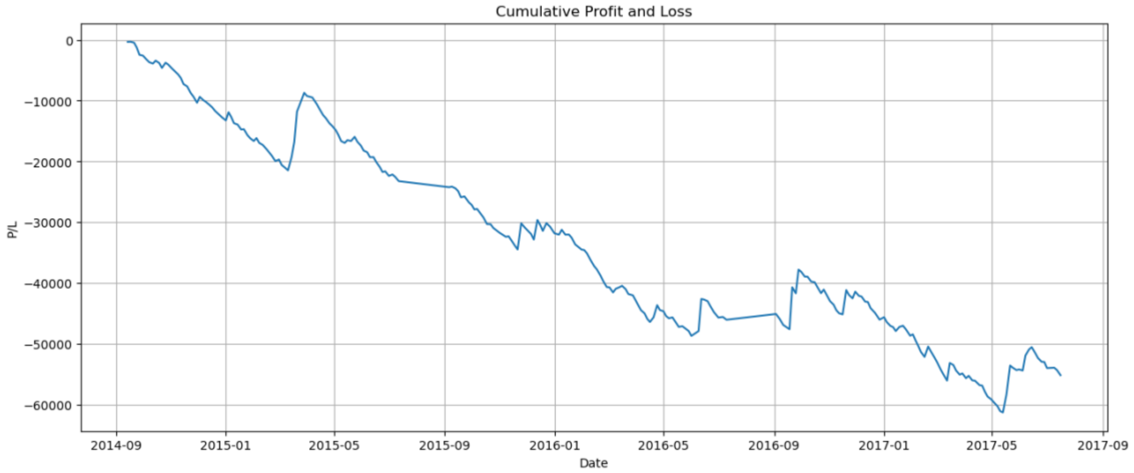


Win Rate: 0.19 %
Total Bets: $18,852.71
Total Return: $5,388.03

Total Pct Return: 28.58 %
Mean: 7.02 %
Standard Deviation: 497.74 %

Median: -100.0 %
Max Drawdown: -$2,533.67
Sharpe Ratio: 0.0141

Using Speed as target variable (Without Kelly):

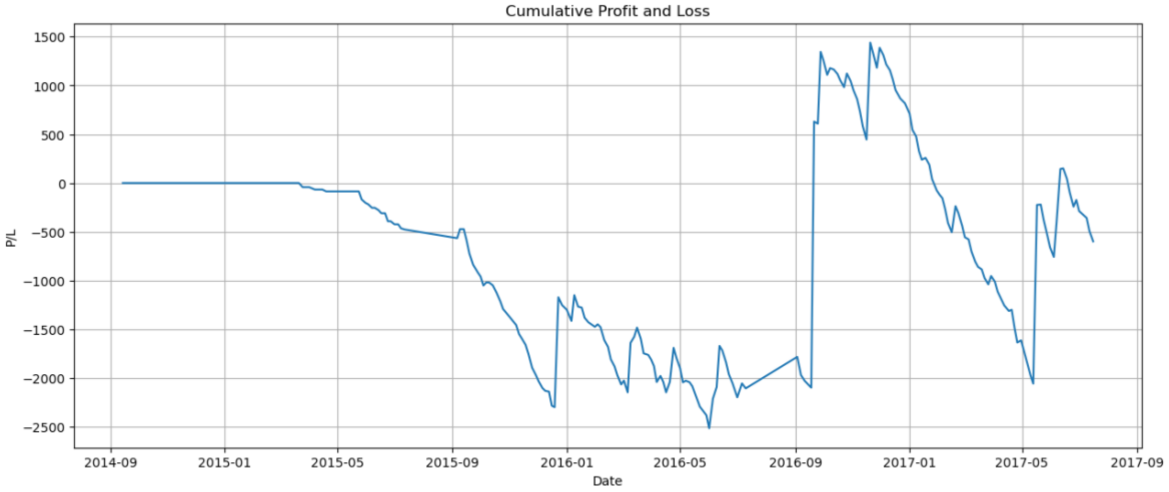


Win Rate: 0.85 %
Total Bets: $239,100
Total Return: -$55,110.0

Total Pct Return: -23.05 %
Mean: -23.05 %
Standard Deviation: 375.68 %

Median: -100.0 %
Max Drawdown: -$62,370.0
Sharpe Ratio: -0.0614

Using Speed as target variable (With Kelly):



Win Rate: 0.15 %	Total Pct Return: -3.58 %	Median: -100.0 %
Total Bets: $16,738.64	Mean: -15.29 %	Max Drawdown: -$3,583.28
Total Return: -$598.48	Standard Deviation: 470.74 %	Sharpe Ratio: -0.0325

It is observed that Kelly played a significant role for the profitability of this strategy. This is possibly due to Kelly betting strategy only works effectively when your probability estimates are accurate.

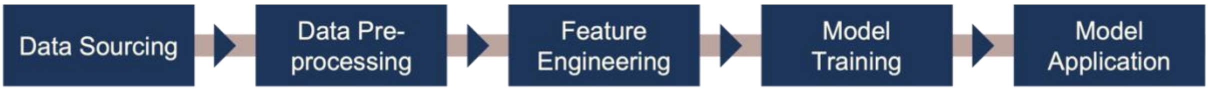
A significant improvement is evident in both Time and Speed based target variable. It is especially so when using time as the target variable, to the extent that we generated significant profits by the end of it.

Another observation is the better performance seen in Time compared to Speed based target variable. This is unexpected as all previous data points suggested that the Speed based target variable should perform better. Observing the graph, the difference is caused by lower bets placed on winning horses. Such as during the large jump at 2016-09. Time based profited about $5000 while Speed based profited only about $3000. This is caused by Kelly criterion which is affected by the estimated win probability. This means that the EMA model is better at modelling the race time compared to the speed of the horse.

4.0 Applying Machine Learning to predict horse performance metrics

4.1 Machine Learning Methodology

The end-to-end process of training a machine learning model to predict horse performance metrics and subsequently applying it to make probable betting decisions comprises of 5 steps.



4.2 Data Sourcing

Horse racing happens in an open environment making it highly susceptible to a wide range of different factors. Hence, the first step of Data Sourcing is extremely essential in capturing as much of these various feature data that can potentially explain the variance in the horse performance.

On top of using the main race dataset, we leveraged web scraping to obtain individual horse data, such as the horse metadata and their past race performances from Hong Kong Jockey Club. We also obtained the historical average daily weather data for each day of the race we have from OpenMeteo.

4.3 Data Pre-processing

After sourcing for a wide variety of data, we merged the columns into a single dataset for consolidated use. In the second step of Data Pre-processing, we removed invalid rows such as when a horse did not finish a race and hence did not have a valid finishing position.

4.4 Feature Engineering

In the third step of Feature Engineering, we worked on reducing the number of dimensions from an initial total of 58 features to 18 features, with the purpose of engineering impactful features.

4.4.1 Breaking down Race Features

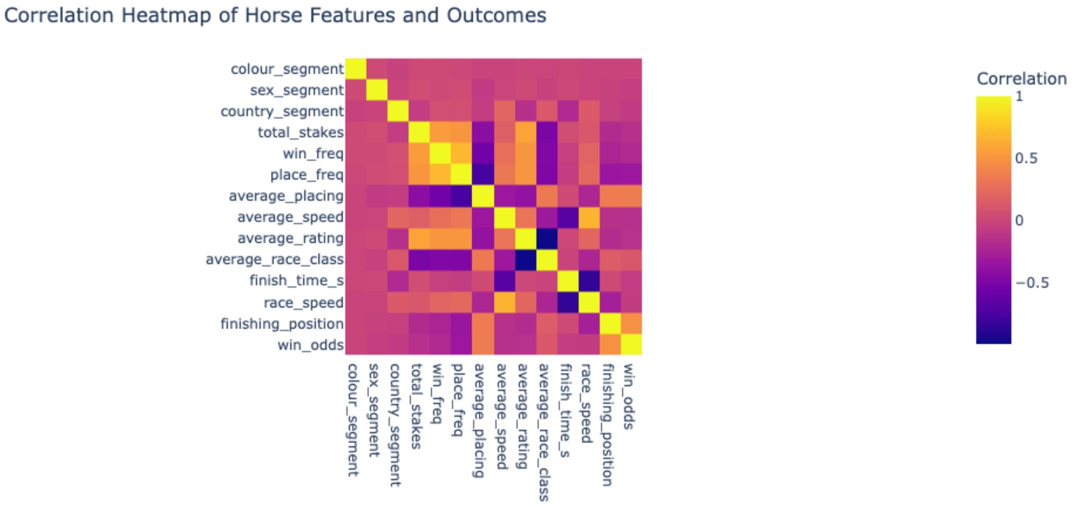
When managing race features that were categorical in nature, we had to distinguish whether these features were nominal or ordinal in nature.

For nominal features, such as race course, these categorical values had no measure of rank between them, hence we chose to encode them for easy interpretation by the model subsequently. For instance, it is hard to argue if a “Sha Tin” race course is better or worse than a “Happy Valley” race course.

For nominal features which has a large number of different categories, we chose to recategorise and group those that had similar outcomes so as to reduce the complexity of the data and highlight patterns more effectively.

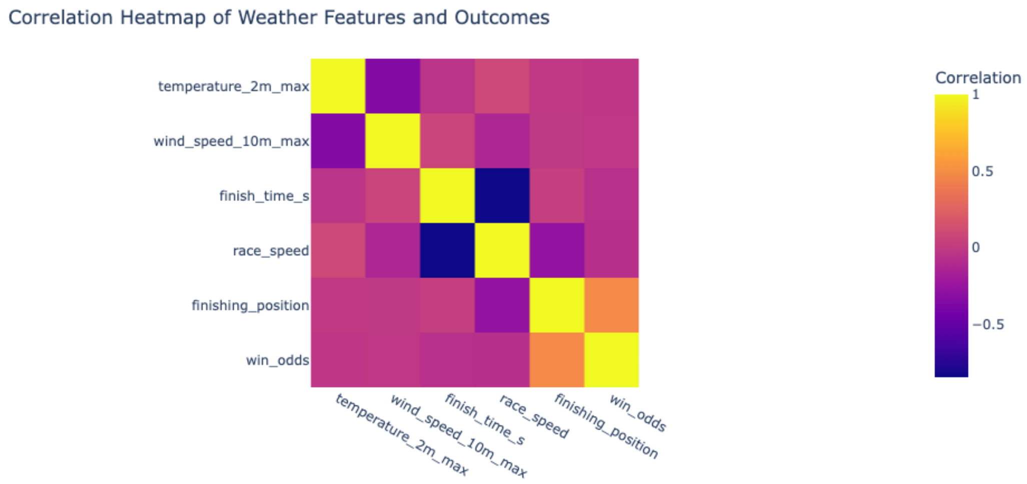
For ordinal features, such as track condition, these categorical values had a measure of rank between them, where one value could be taken to better or worse than another. For instance, a turf track condition that had “GOOD TO FIRM” value would have a lower penetrometer value than one that had “GOOD” value, hence we chose to map these track condition categories to discrete penetrometer values (Hong Kong Turf, 2018).

In summary, these were the race features we engineered, refer to appendix 6.5.1 to view individual feature engineering.



4.4.2 Breaking down Weather Features

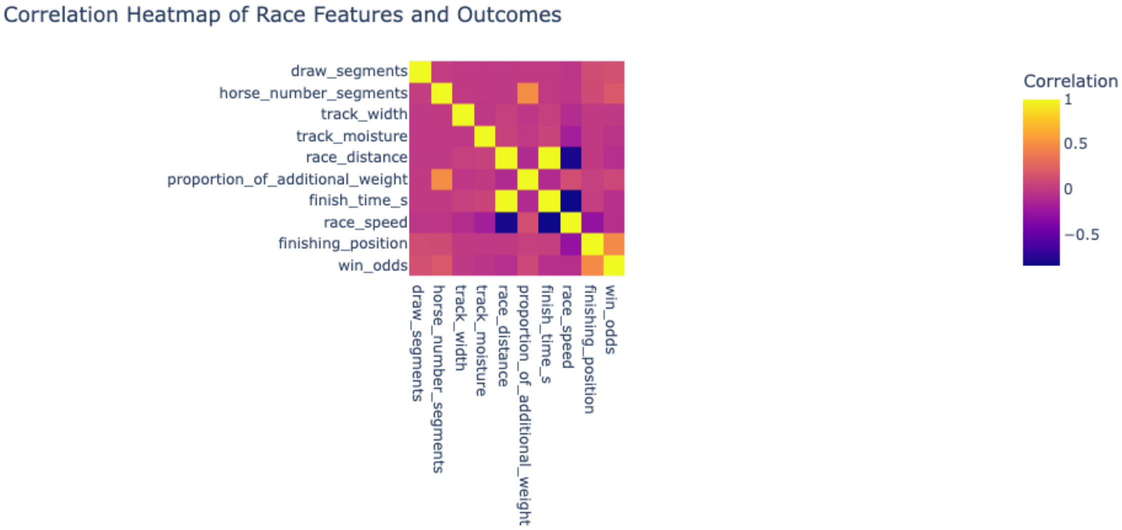
From the weather data acquired from OpenMeteo, we noticed that there existed a high degree of correlation between weather features. As such, we picked temperature_2m_max and wind_speed_10m_max as they had the highest correlation with our dependent variables and relatively correlation with each other.



4.4.3 Breaking down Horse Features

The horse data scraped from Hong Kong Jockey Club consisted of 2 main components – the horse metadata and the horse historical race performance. Given the historical race performance of a horse, we aggregated these metrics to summarise large volumes of race data and smooth out noise.

In summary, these were the horse features we engineered, refer to appendix 6.5.2 to view individual feature engineering.



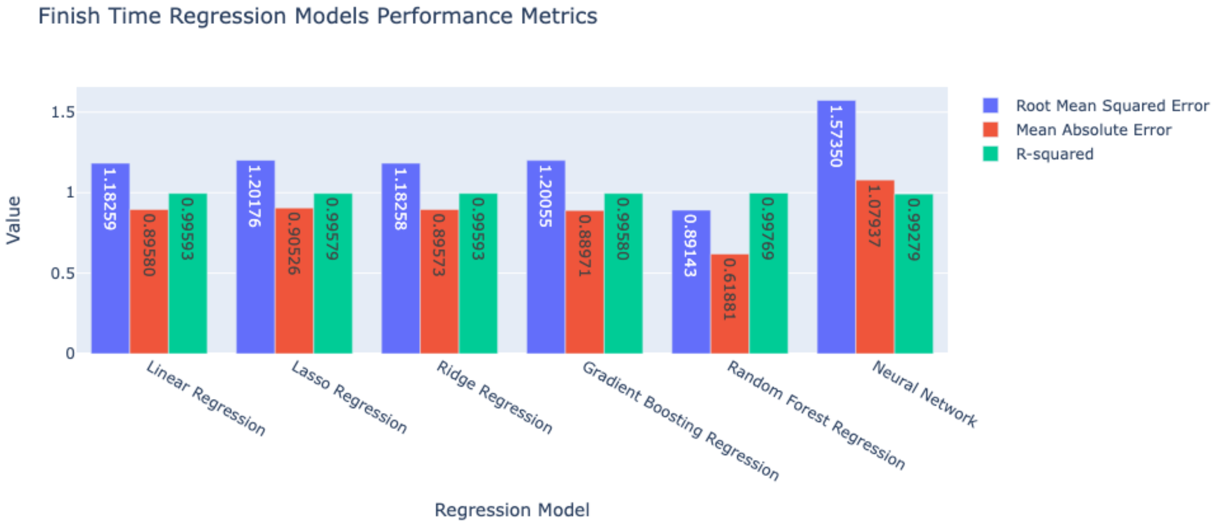
4.5 Model Training

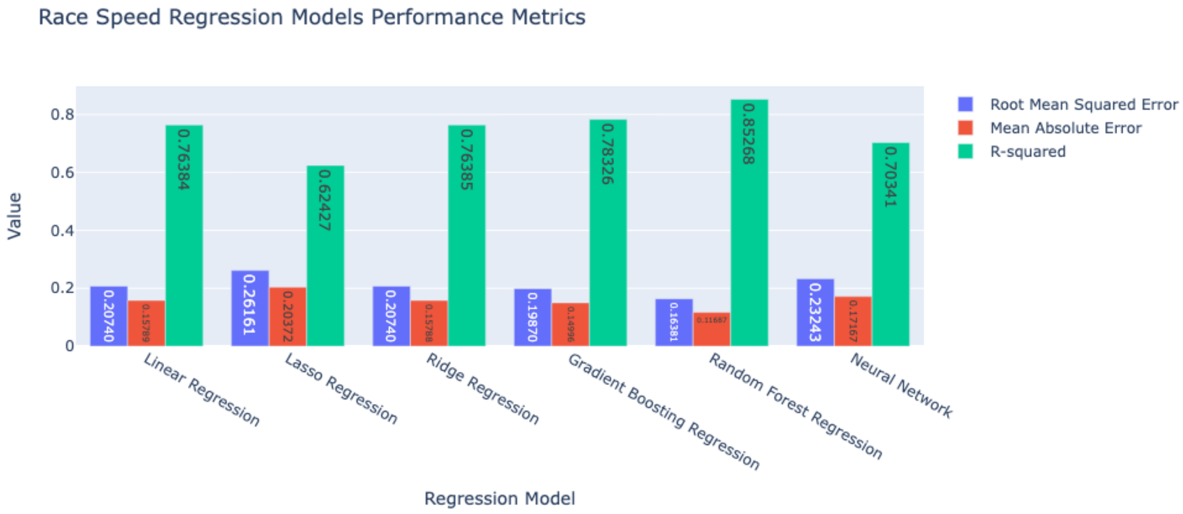
To make predictions on the winner of a race, we leveraged various regression and classification models to predict 3 different horse performance metrics – race speed, finish time and probability of finishing in the top 3 ranks. This would be used to estimate the probability of a horse winning a race and make subsequent betting decisions with.

To ensure there was no data snooping, we segmented our final dataset such that all data before 12th Oct 2016 would be used exclusively for training our models, and all data from 12th Oct 2016 would be used exclusively for testing our models and making betting decisions.

<u>Predicting Horse Race Speed and Finish Time with Regression Models</u>

We experimented with 6 different Regression models and concluded that Lasso Regression, Ridge Regression, Gradient Boosting Regression and Random Forest Regression had the best regression performances based on metrics like Root Mean Squared Error, Mean Absolute Error and R-squared.

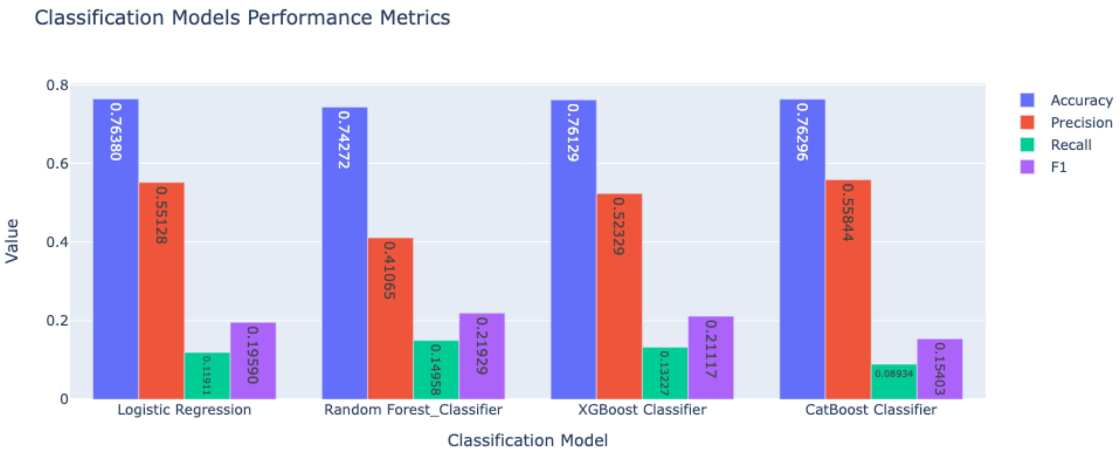




Predicting Finishing Position Class with Classification Models

From the finishing position column data, we created a new finishing position class where positions 1st to 3rd would be Class 1, and positions 4th onwards would be Class 0. This was done so as to simplify the problem compared to have to predict multiple classes of the exact position. In such cases, the model may also not be able to distinguish the differences between the multiple classes that it has to predict. Using Classification models such as Logistic Regression, we would be able to make predictions on the finishing position class for each horse in a race, and the probability of the horse getting classified as said finishing position class.

We experimented with 4 different Classification models, and concluded that Random Forest Classifier, XGBoost Classifier and Cat Boost Classifier had the better classification performances based on metrics like accuracy, precision, recall and F1 score.



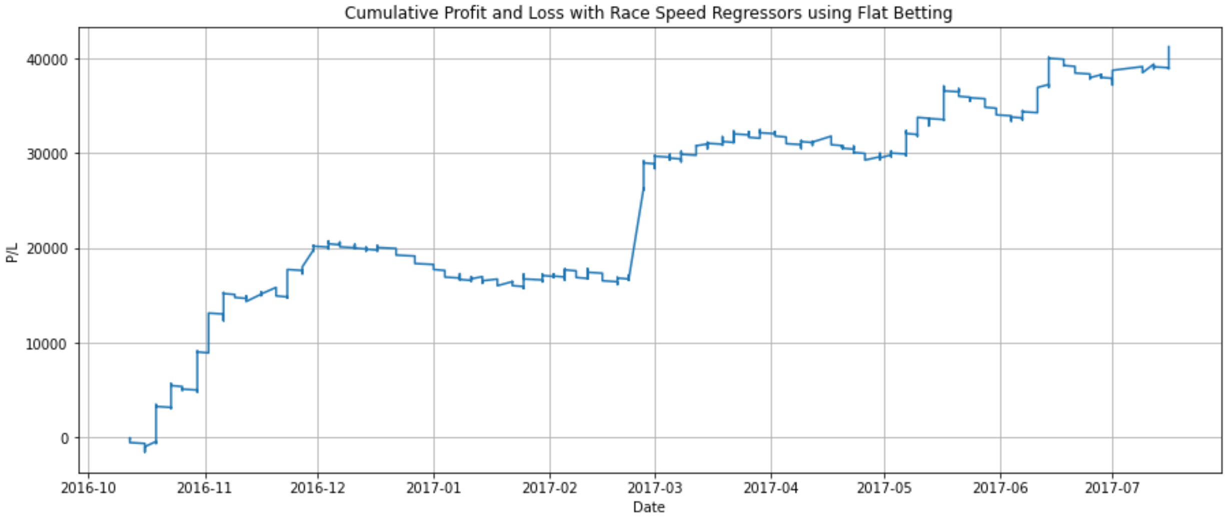
4.6 Model Application

Betting decisions with Average Horse Speed and Horse Finish Time Regressors

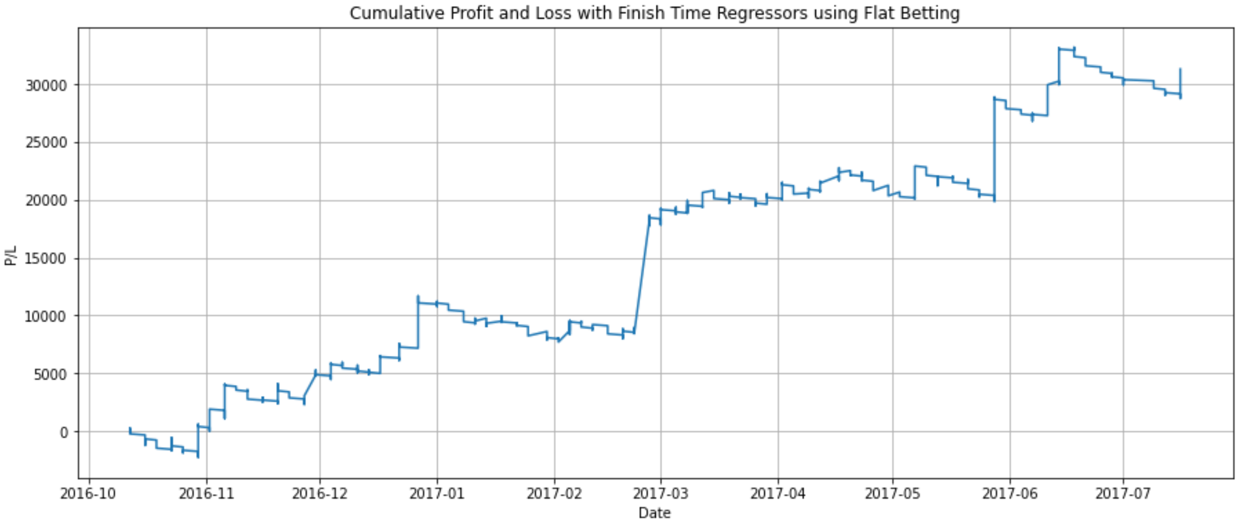
Using the 4 best-performing Regression Models, each model was used to predict the Average Horse Speed or Horse Finish Time. From these predictions, we derived a predicted finishing position for a horse in a race, by ranking the horse in descending order of average speed or ascending order of finish time. The probability of a horse winning a race would be taken as the average of the inverse predicted finishing position (Appendix 6.3).

This formula averages the inverses of the predicted finishing positions, providing a measure of the horse's likelihood to win based on its expected performance across the races (Appendix 6.3)

By making flat bets, where we used fixed size bets of $100 regardless of the probability of a horse winning, on the most likely horse to win a race, we achieved significant returns where our race speed regression models made a profit of $41,250, and our finish time regression models made a profit of $31,290 for less than a year of betting from 12th Oct 2016 to 16th Jul 2017. Out of all the races from this period, the race speed regression models and the finish time regression had a prediction accuracy of 57.3%, of 55.1% respectively.

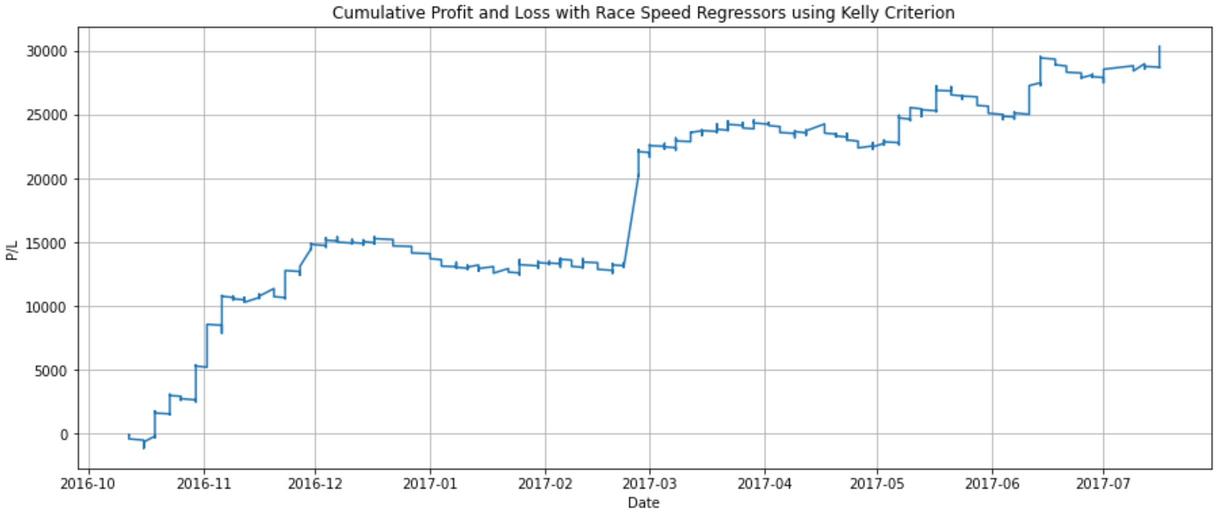


Win Rate: 21.04%	Total Pct Return: 4.79%	Median: -7.69%
Total Bets: $862,000	Mean: 4.57%	Max Drawdown: -$5,010
Total Return: $41,250	Standard Deviation: 42.21%	Sharpe Ratio: 0.1081

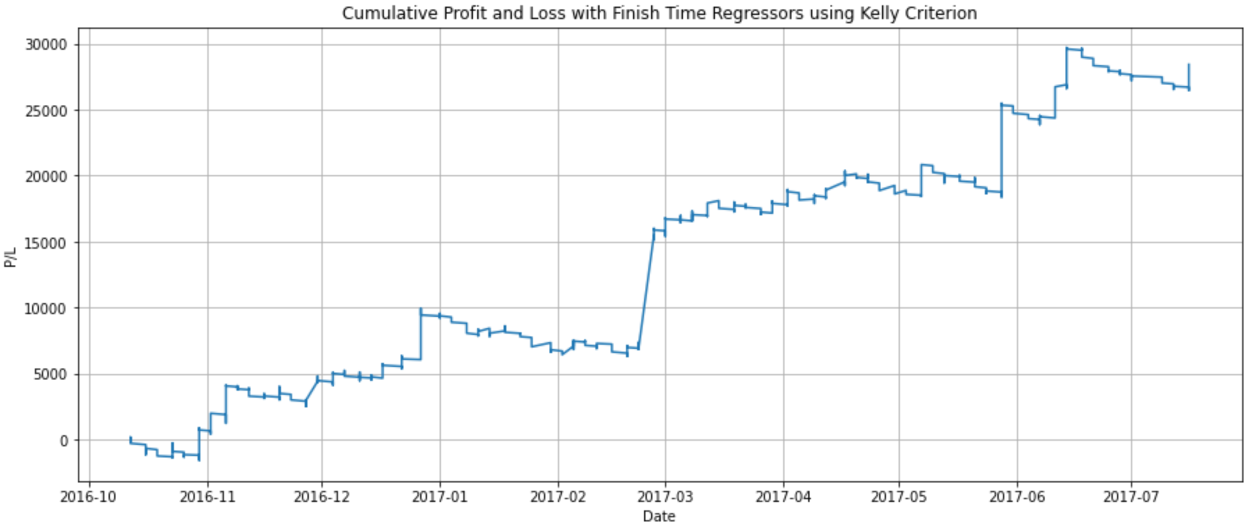


Win Rate: 20.2%	Total Pct Return: 3.63%	Median: - 8.33%
Total Bets: $862,900	Mean: 4.76%	Max Drawdown: -$4,420
Total Return: $31,290	Standard Deviation: 65.69%	Sharpe Ratio: 0.0726

When using Kelly Criterion to allocate bet amounts based on the probability of a horse winning and its win odds, we found that there was a slight decrease in absolute returns where the race speed regression models and finish time regression models made a profit of $30,360.08 and $28,419.82 respectively. This can be attributed to the reduced sized winnings from betting adjusted amounts. However, looking at the normalised returns using Mean and Sharpe Ratio, we observe that Kelly Criterion performs better than flat betting.



Win Rate: 20.99%	Total Pct Return: 17.93%	Median: - 27.63%
Total Bets: $169,322.30	Mean: 16.5%	Max Drawdown: -$2,980.16
Total Return: $30,360.08	Standard Deviation: 140.59%	Sharpe Ratio: 0.1174



Win Rate: 20.2% Total Pct Return: 14.93% Median: -28.45%

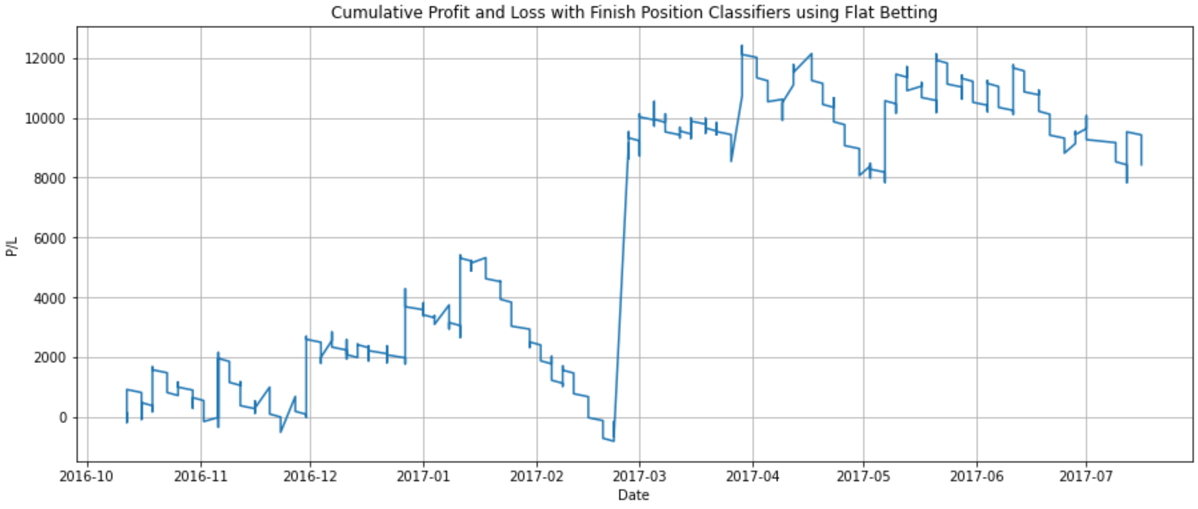
Total Bets: $190,352.67 Mean: 14.69% Max Drawdown: -$3,644.38

Total Return: $28,419.82 Standard Deviation: 181.17% Sharpe Ratio: 0.0811

Betting decisions with Finishing Position Class Classifiers

Using the 3 best-performing Classification Models, each model was used to predict the probability of a horse being classified as class 1, indicating the horse finishing within the top 3 positions. The horse that had the best average probability from all 3 models would then be our predicted horse to win.

By making flat bets, our finish position classification models had a prediction accuracy of 53.9%, but made a profit of $8,430, a significant decrease from the regression models. This could be due to the classification models getting only the bets with lower win odds right, thereby limiting the amount of profits that could have been made.

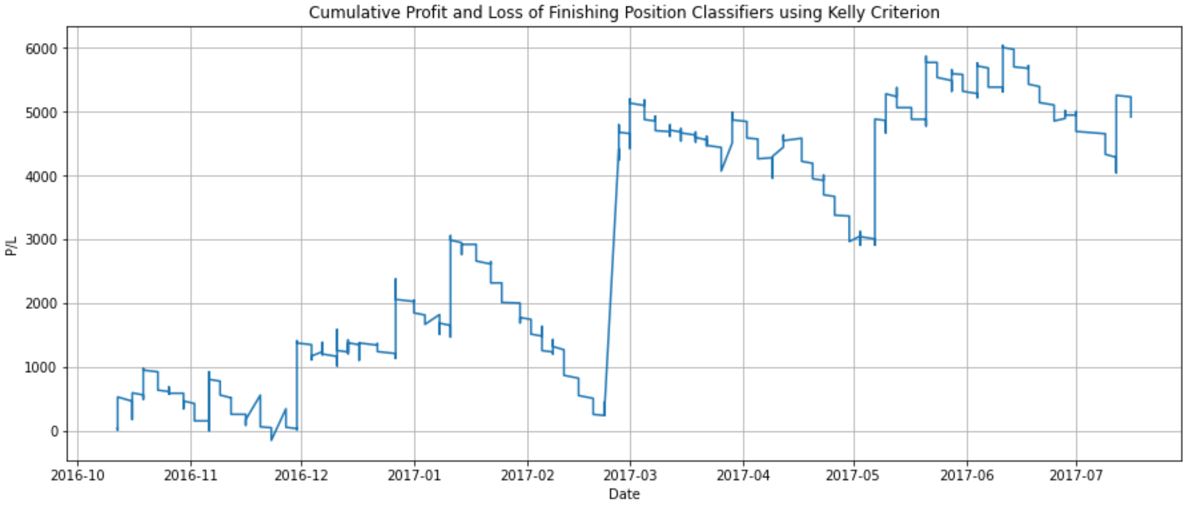


Win Rate: 17.39% Total Pct Return: 0.99% Median: - 8.33%

Total Bets: $852,200 Mean: 1.59% Max Drawdown: -$6,230

Total Return: $8,430.0 Standard Deviation: 45.39% Sharpe Ratio: 0.035

When using Kelly Criterion to allocate our bet amounts, this reduced our nominal profits as expected to $4,914.18. Similarly, looking at the normalised performance using mean returns and Sharpe ratio, we observe that Kelly Criterion performs better than flat betting.



Win Rate: 16.52%	Total Pct Return: 3.43%	Median: -17.93%
Total Bets: $143,148.79	Mean: 4.97%	Max Drawdown: -$2,820.93
Total Return: $4,914.48	Standard Deviation: 128.53%	Sharpe Ratio: 0.0387

5.0 Betting Strategy Creation and Benchmarking

5.1 Benchmarking Strategies

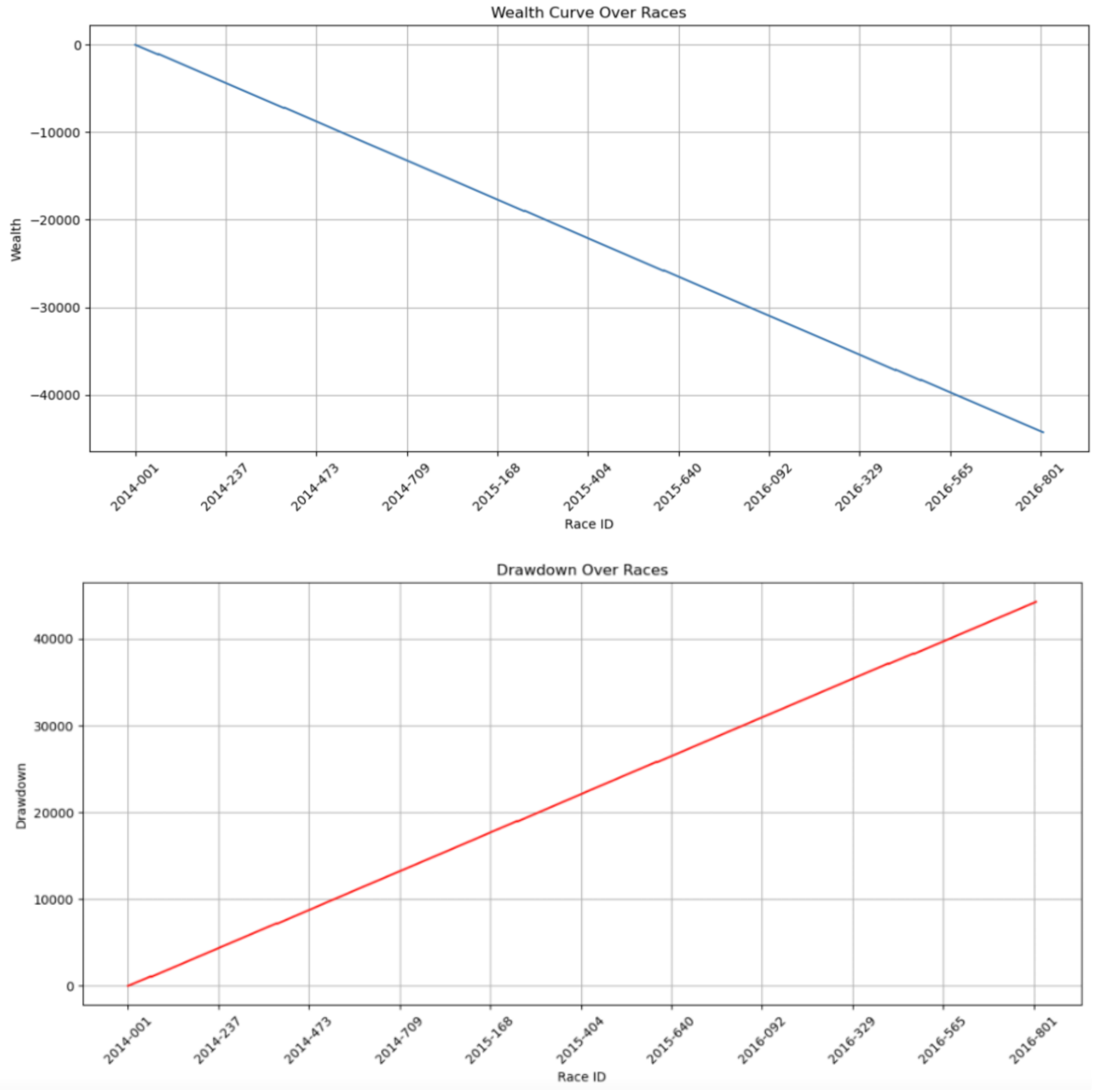
The following strategies are used as a benchmark to evaluate the effectiveness and profitability of the betting strategy discussed in the earlier section.

5.1.1 Betting Proportionally to Win Odds

This strategy is adopted by bettors for risk management purposes and to increase the possibility of generating returns from betting on multiple horses. Typically, the implementation of this strategy involves the bettor identifying horses with win odds that are perceived to be undervalued and allocating the bet amount proportionally to the win odds. For simplicity, the benchmarking process will assume that the bettor places a bet on every horse with the bet weight corresponding to the normalised implied winning probability.

The implied winning probability can be calculated by taking the reciprocal of the win odds. The higher the win odds, the lower the chances of winning. However, the sum of implied winning probabilities for each horse always adds up to greater than 100%. To allow bet weights to be assigned according to the winning probability, the implied winning probabilities must first be normalised to 100%. This is achieved by dividing the implied winning probability of each horse by the sum of implied winning probabilities of all horses in a race. This normalised winning probability becomes the bet weight of the horse, summing up to 1.

To calculate the total returns of this strategy, the first step is to calculate the returns per horse. If the horse is a winner, the returns are equal to the bet weight multiplied by the win odds and subtracting the bet weight. If the horse loses, the returns are equal to negative of the bet weight. Adding up the returns for each horse in a race provides the betting returns for each race. This is then used to plot the wealth curve, assuming that $100 is bet for each race.



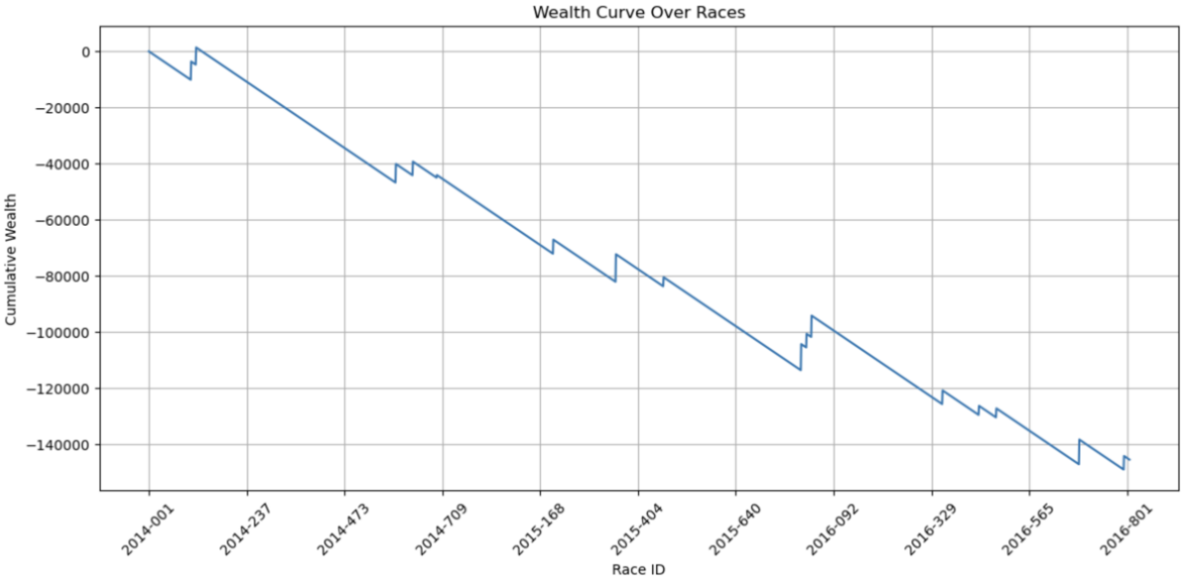
Win Rate: 7.86%	Total Pct Return: -18.71%	Median: -100%
Total Bets: $236,621.11	Mean: -26.9%	Max Drawdown: -$44,336.45
Total Return: -$44,275.34	Standard Deviation: 406.11%	Sharpe Ratio: -0.0662

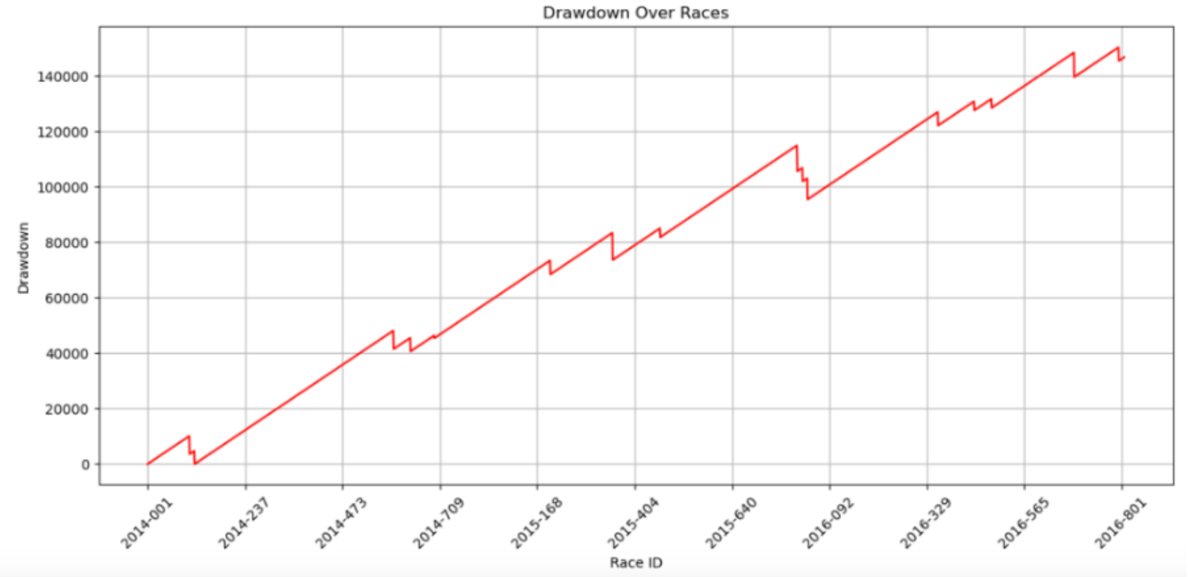
Betting on every horse is a suboptimal strategy due to the existence of the bookmaker’s margin. Since the sum of all winning chances always add up to above 100%, the bookmaker receives more money from losing bets than it pays out to winning bets. This guarantees that anyone who bets on all horses will lose money, which is reflected in this strategy’s mean returns of -26.9%.

5.1.2 Long Shot Betting

This strategy is adopted by bettors who accept lower chances of winning in exchange for the potential for greater profits. This involves betting on the horse with the lowest probability of winning (highest win odds). For cases where there are multiple horses that share the highest win odds, the bet weight is normalised such that it is never greater than 1. This means that if there are 2 horses with the highest win odds, the bettor will split the bet amount equally between the 2 horses, with the total bet weight summing up to 1. The other horses will have a bet weight of 0.

The profit per horse can be calculated by taking bet weight multiplied by the win odds and subtracting the bet weight for horses which finish in first. For the horses that did not win, the profit is simply the negative value of the bet weight. The profit per race is then calculated by adding up the profits for all horses in the race.





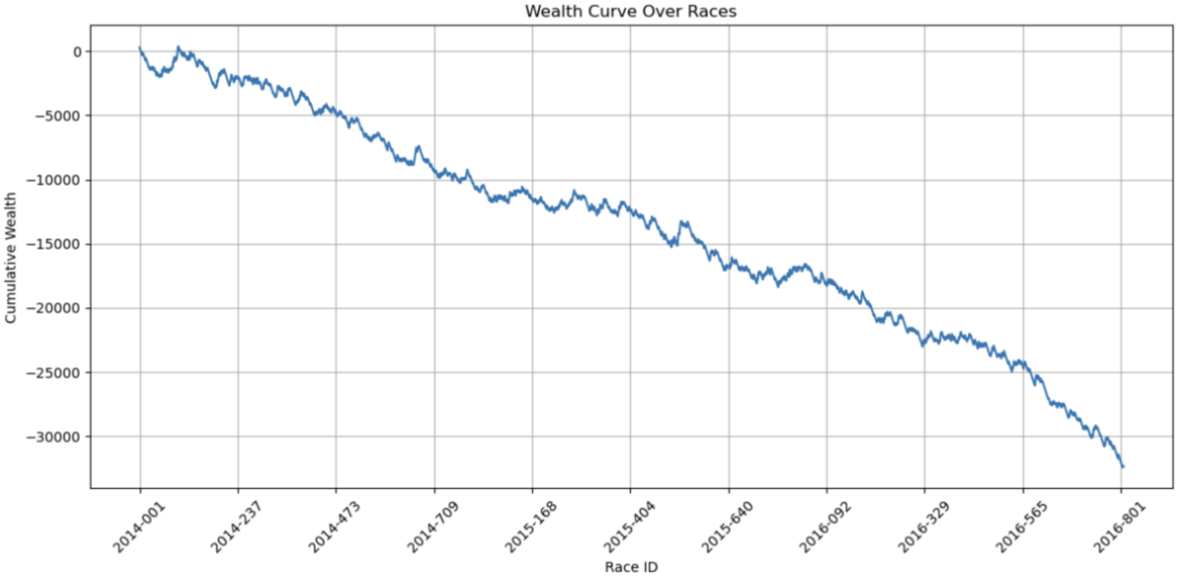
Win Rate: 0.68%	Total Pct Return: -61.47 %	Median: -100.0 %
Total Bets: $236,700	Mean: -61.47%	Max Drawdown: -$150,350
Total Return: -$145,500	Standard Deviation: 506.06%	Sharpe Ratio: -0.1215

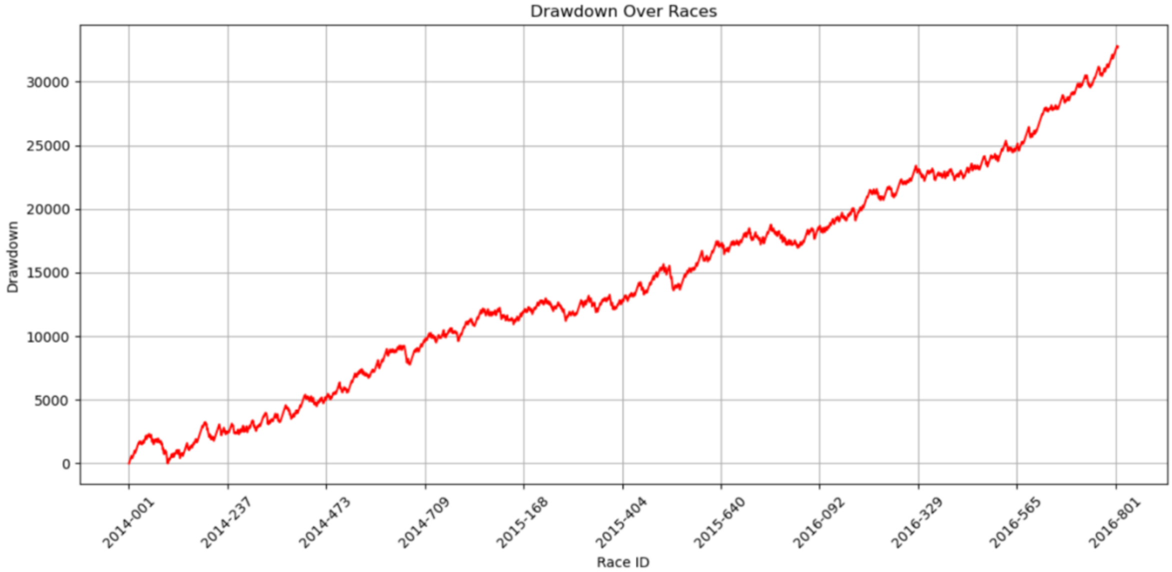
Betting on the long shot has the potential for very high returns given that the win odds are the highest. However, the probability of winning is incredibly low, with a win rate of 0.68%. This makes the long shot strategy highly unprofitable in the long run with a mean return of -61.47%.

5.1.3 Favourite Betting

This strategy is adopted by bettors who accept lower profits in exchange for a higher chance of winning. This involves betting on the horse with the highest probability of winning (lowest win odds). For cases where there are multiple horses that share the lowest win odds, the bet weight is normalised such that it is never greater than 1. This means that if there are 2 horses with the lowest win odds, the bettor will split the bet amount equally between the 2 horses. For horses with the lowest win odds, the bet weight will sum up to 1 while the other horses will have a bet weight of 0.

The profit per horse can be calculated by taking bet weight multiplied by the win odds and subtracting the bet weight for horses which finish in first. For the horses that did not win, the profit is simply the negative value of the bet weight. The profit per race is then calculated by adding up the profits for all horses in the race.





Win Rate: 31.61%	Total Pct Return: -13.68%	Median: -100.0 %
Total Bets: $236,600	Mean: -13.68%	Max Drawdown: -$32,800
Total Return: -$32,375	Standard Deviation: 135.30%	Sharpe Ratio: -0.1011

Betting on the favourite while unprofitable, performs the best relative to the other benchmarking strategies, with a mean return of -13.68%. This suggests that the implied winning odds are not accurate at predicting the actual winner of each race as the win rate is only 31.61%. Furthermore, given that the win odds for the favourite are the lowest, the returns from winning are insufficient to cover the losses in other races.

6.0 Appendix

6.1 Kelly Criterion

Kelly formula:

$$f = (bp - q) / b$$

where:

$$f = the\ fraction\ of\ the\ current\ capital\ to\ wager$$
$$b = the\ odds\ received\ on\ the\ bet$$
$$p = the\ probability\ of\ horse\ winning$$
$$q = the\ probability\ of\ horse\ losing, which\ is\ 1 - p$$

6.2 Exponential Moving Average

EMA formula:

$$EMA_{Current} = \alpha * \mathrm{x} + (1 - \alpha) * EMA_{Previous}$$

Where:

$$\alpha = \text{Smoothing factor}$$
$$\mathrm{x} = \text{Most recent data}$$

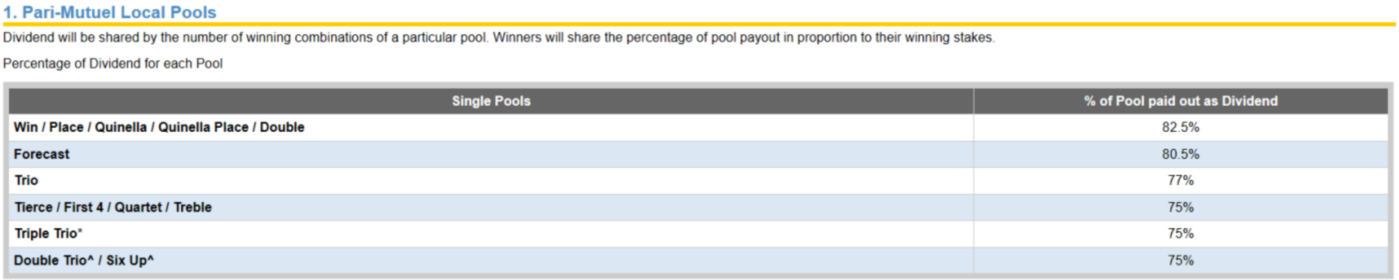
6.3 Average Probability of a Horse Winning a Race

$$P_{win} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}$$

where:

$$P_{win} = probability\ of\ the\ horse\ winning,$$
$$n = number\ of\ races,$$
$$P_i = predicted\ finishing\ position\ of\ the\ horse\ in\ the\ i^{th}\ race$$

6.4 Other tables and charts used



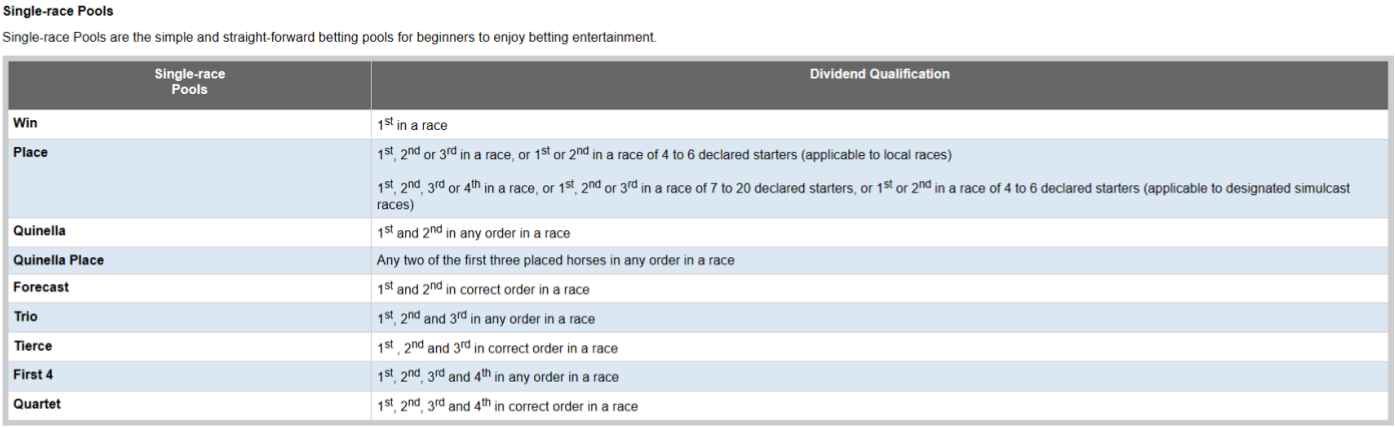
1. Pari-Mutuel Local Pools

Dividend will be shared by the number of winning combinations of a particular pool. Winners will share the percentage of pool payout in proportion to their winning stakes.

Percentage of Dividend for each Pool

Single Pools	% of Pool paid out as Dividend
Win / Place / Quinella / Quinella Place / Double	82.5%
Forecast	80.5%
Trio	77%
Tierce / First 4 / Quartet / Treble	75%
Triple Trio*	75%
Double Trio^ / Six Up^	75%

Figure 1: Pay outs for Pari-Mutuel Betting, taken from HKJC



Single-race Pools

Single-race Pools are the simple and straight-forward betting pools for beginners to enjoy betting entertainment.

Single-race Pools	Dividend Qualification
Win	1st in a race
Place	1st, 2nd or 3rd in a race, or 1st or 2nd in a race of 4 to 6 declared starters (applicable to local races) 1st, 2nd, 3rd or 4th in a race, or 1st, 2nd or 3rd in a race of 7 to 20 declared starters, or 1st or 2nd in a race of 4 to 6 declared starters (applicable to designated simulcast races)
Quinella	1st and 2nd in any order in a race
Quinella Place	Any two of the first three placed horses in any order in a race
Forecast	1st and 2nd in correct order in a race
Trio	1st, 2nd and 3rd in any order in a race
Tierce	1st , 2nd and 3rd in correct order in a race
First 4	1st, 2nd, 3rd and 4th in any order in a race
Quartet	1st, 2nd, 3rd and 4th in correct order in a race

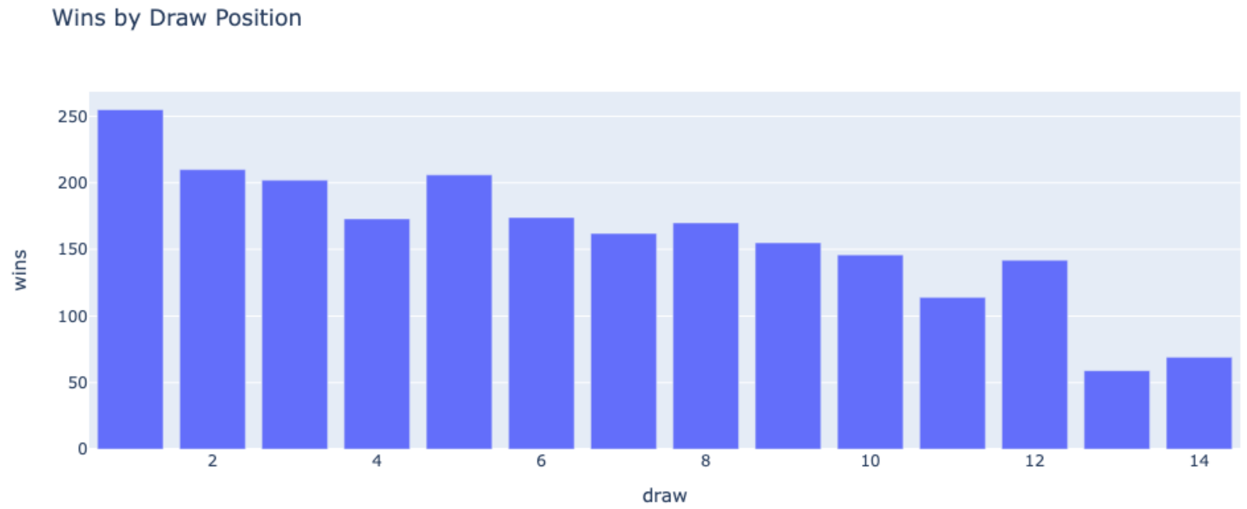
Figure 2: Race results qualification explained, taken from HKJC

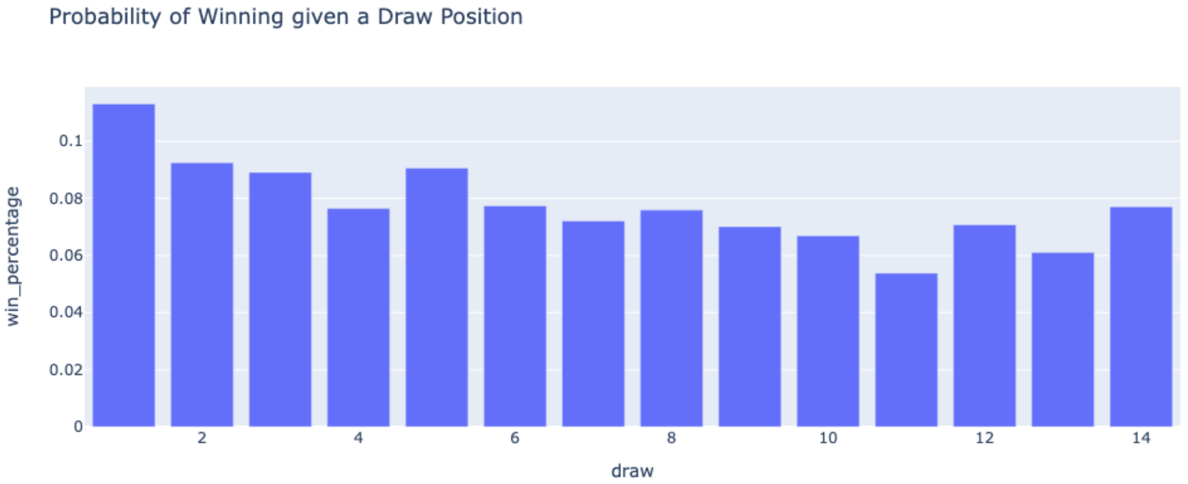
6.5 Feature Engineering

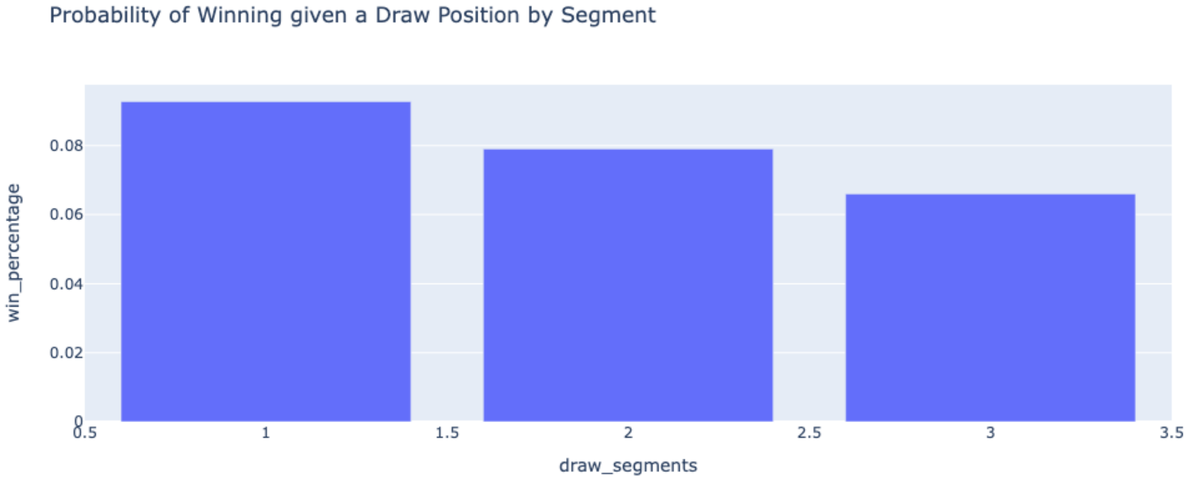
6.5.1 Race Features Engineering

Draw Position

Draw position refers to the lane which the horse would be running in, which is randomly assigned to a horse a few days before a race. For tracks which are circular in nature, a low draw number is advantageous to a horse as it would be running on the inner lanes which have a lower race distance to cover than the outer lanes. We decided to group and segment the draw based on those with similar probabilities to reduce the number of dimensions in the values.

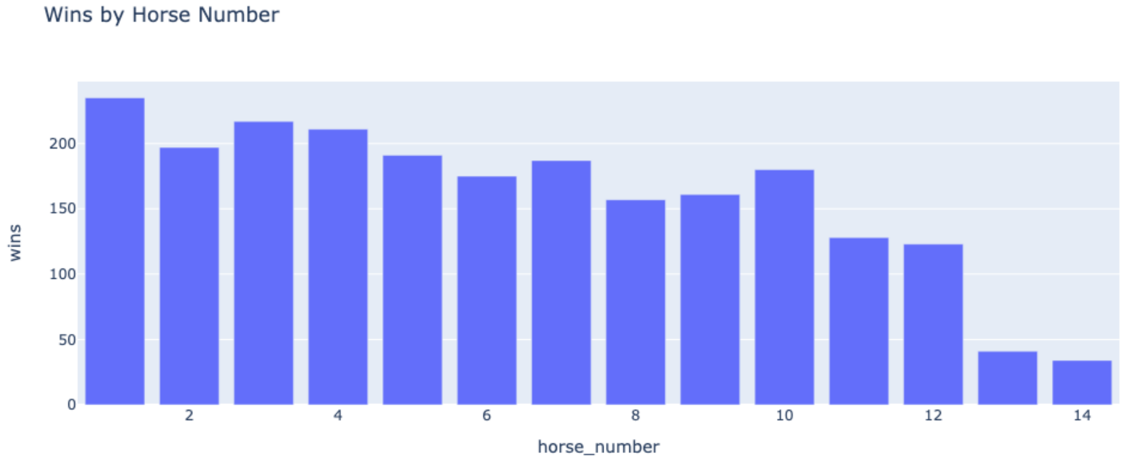


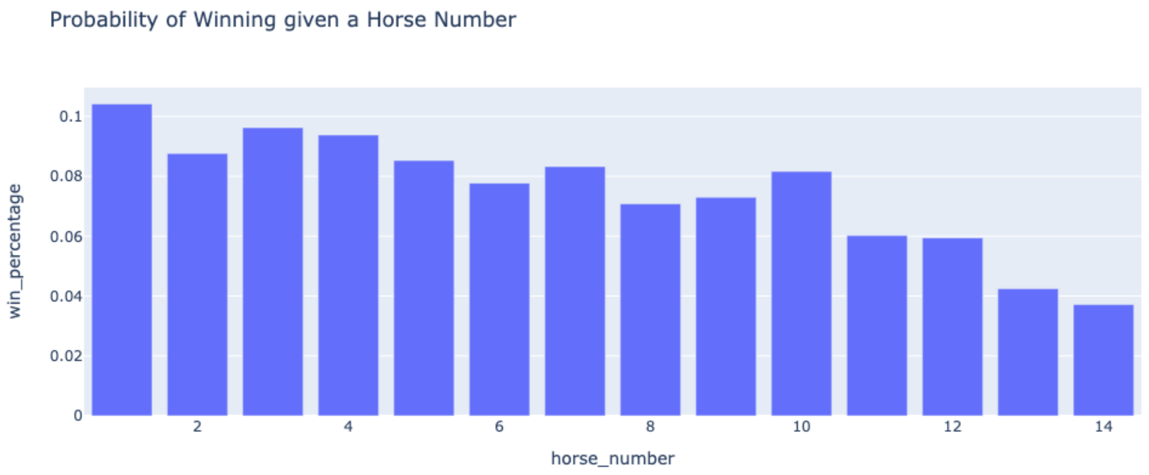


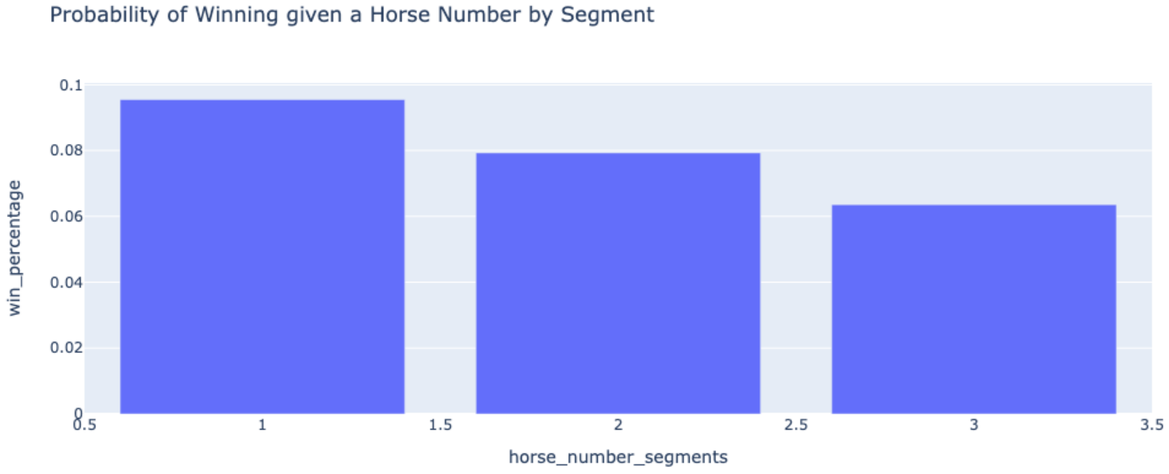


Horse Number

Horse numbers are the relative ratings of the horses in a race, where horse number 1 refers to the highest rated horse in a particular race and horse number 14 refers to the lowest rated horse in a particular race. We decided to group and segment the horse number based on those with similar probabilities to reduce the number of dimensions in the values.



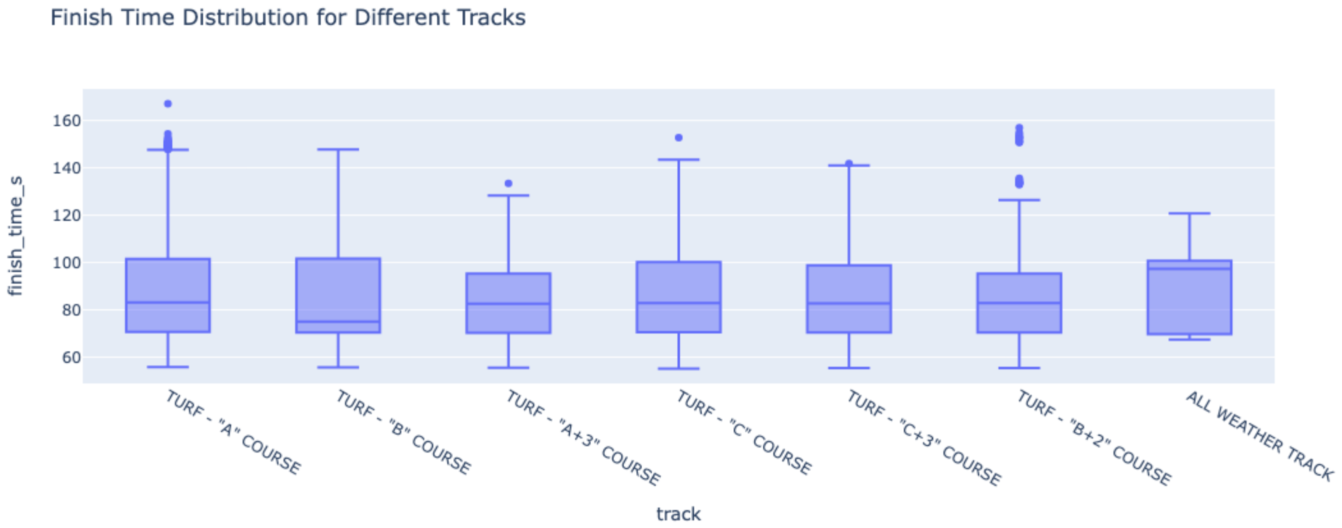




Track Course

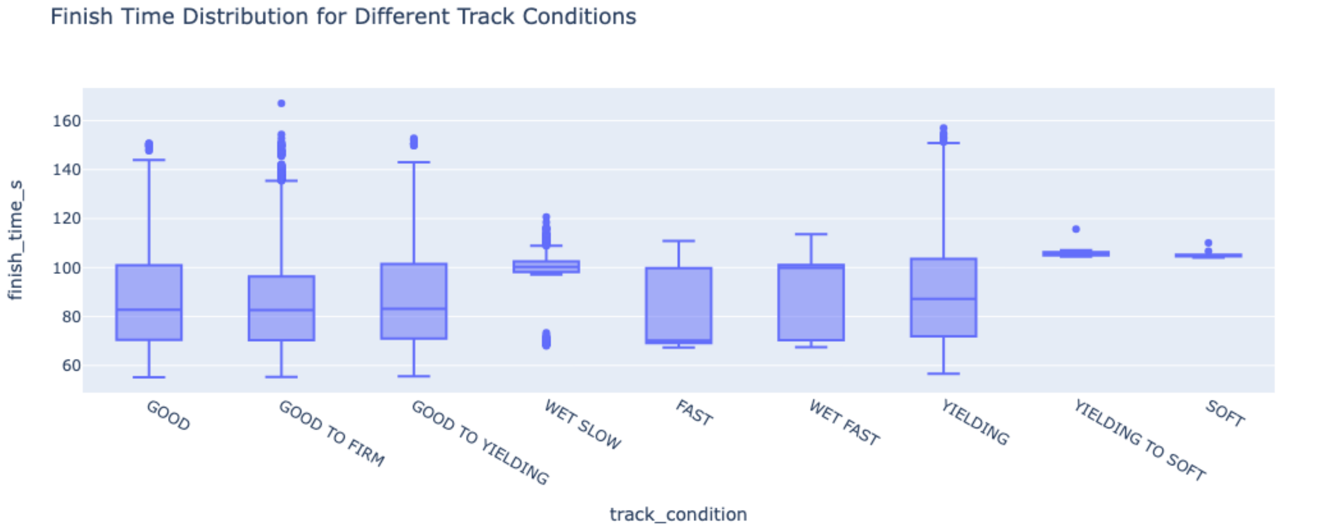
There are 2 broad categories of track course, turf and all weather track. Among turfs, there are a wide range of different turf types such as “A”, “B” and “C”. The main difference between these different types lies in the track width, where across different types, they may have a longer or shorter track width. Between subtypes such as “A” and “A+3”, track widths would shrink by that constant, for instance “A+3” would be 3 meters narrower than “A”. Hence we translated these track course categorical values into a track width numerical value.

Track widths are extremely crucial race information as they affect the ease of horses overtaking as a narrower track makes it harder as compared to a wider track.



<u>Track Condition</u>

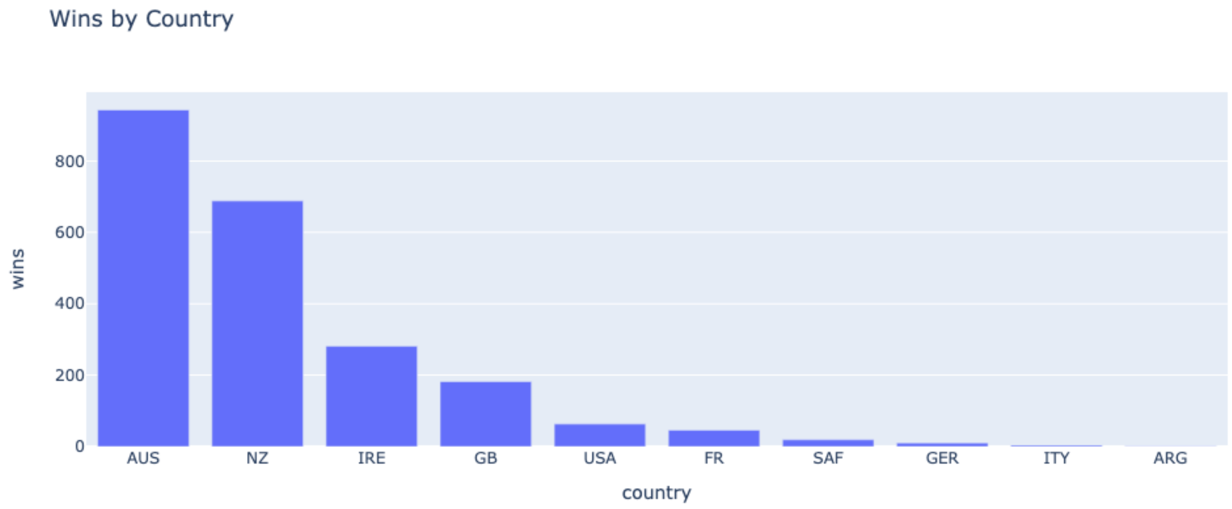
Track condition refers to the amount of moisture or dryness of the track. If the track is too moist, horses may sink into the track or slip while running causing them to run slower as we can see from a lower and more condensed finish time distribution for “WET SLOW”, “WET FAST”, “SOFT” and “YIELDING TO SOFT” track conditions. Hong Kong Jockey Club classifies these track conditions based on a quantitative penetrometer reading of the track, hence we re-translated these categorical values back into their original numerical penetrometer reading for ease of interpretation by the model.

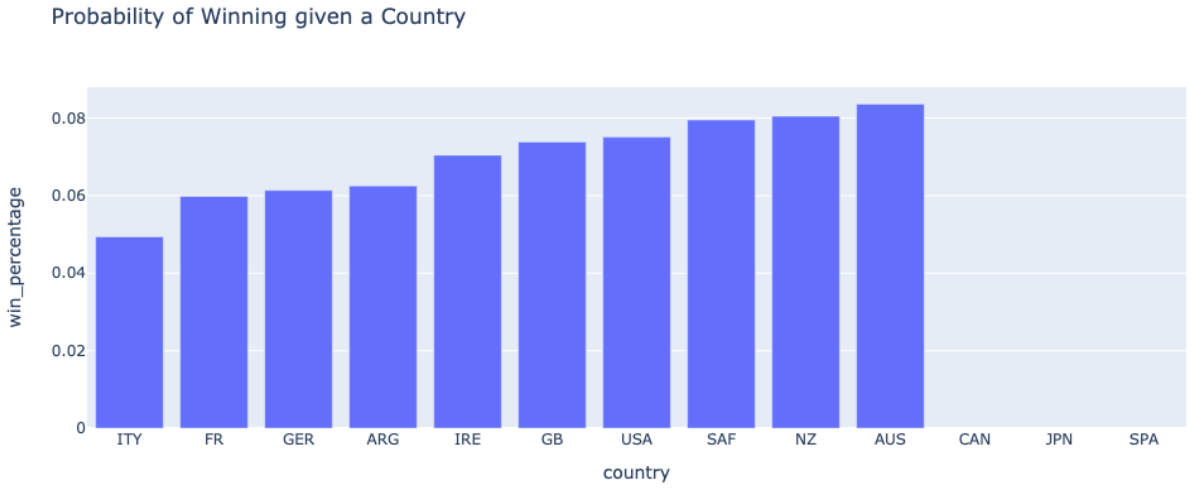


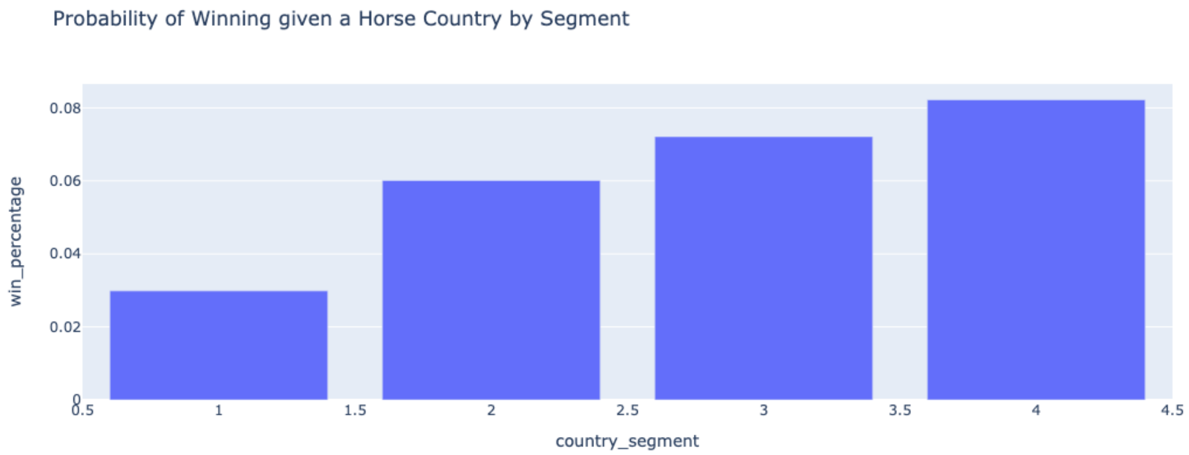
6.5.2 Horse Features Engineering

<u>Country</u>

Based on the breakdown of wins by country, we see that horses from Australia win a significant amount of races more than others. However, when adjusted based on the amount of data we have from each country, we see that they are relatively similar now. This might be due to a disproportionate amount of Australian horses running in races. We decided to group and segment the country based on those with similar probabilities to reduce the number of dimensions in the values.

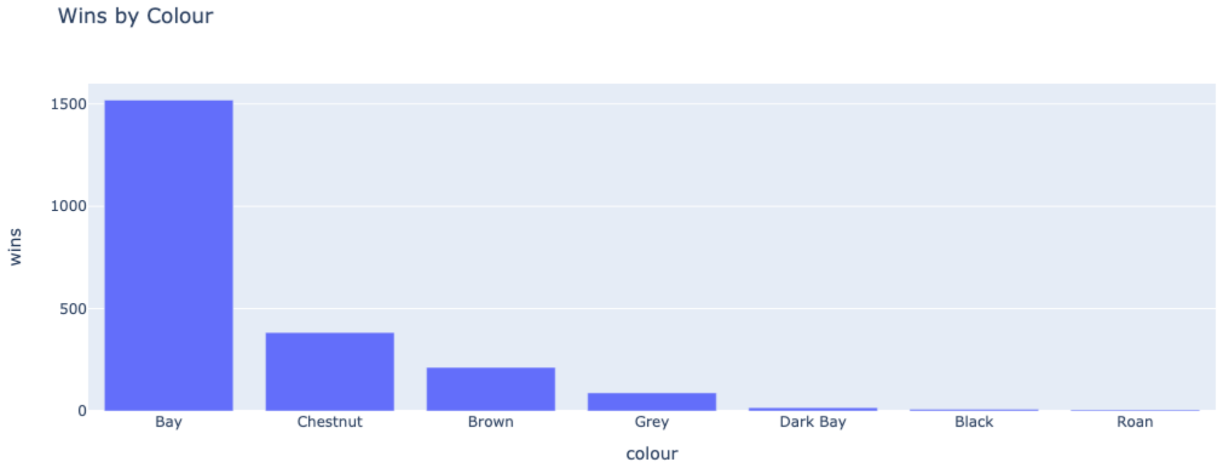


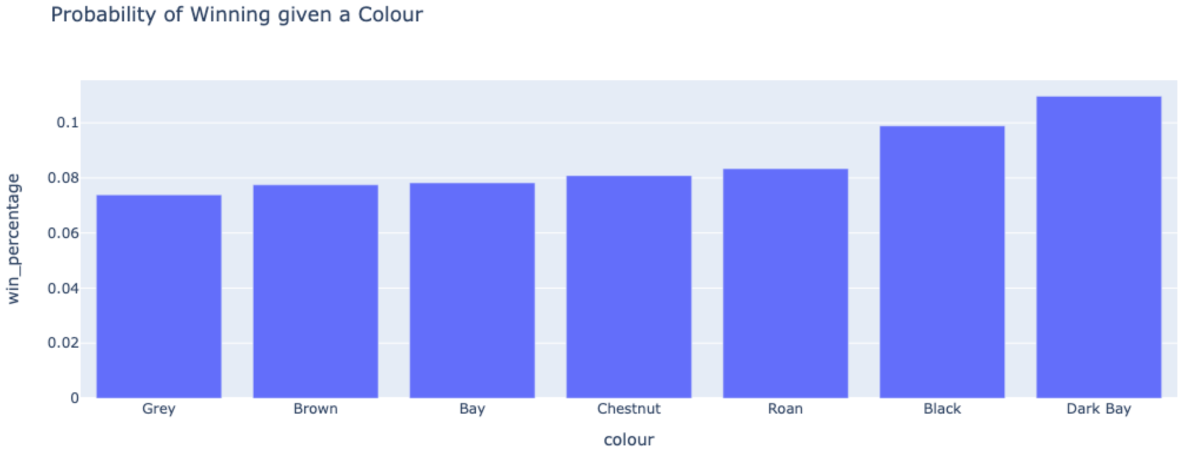


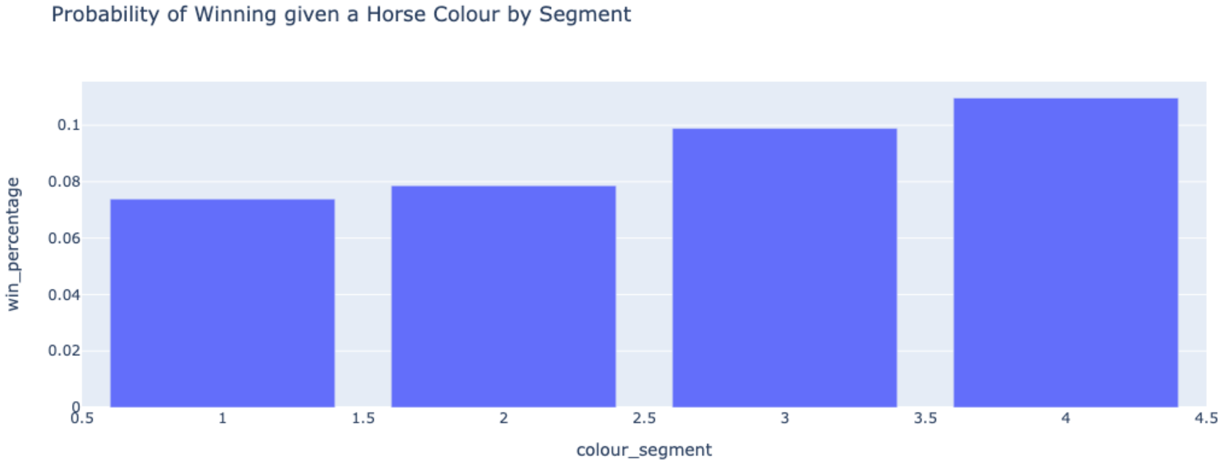


Colour

Based on the breakdown of wins by colour, we see that horses of Bay colour win a significant amount of races more than others. However, when adjusted based on the amount of data we have from each colour, we see that Dark Bay colour has the highest probability of winning now. This might be due to a disproportionate amount of Bay gender horses running in races. We decided to group and segment the country based on those with similar probabilities to reduce the number of dimensions in the values.

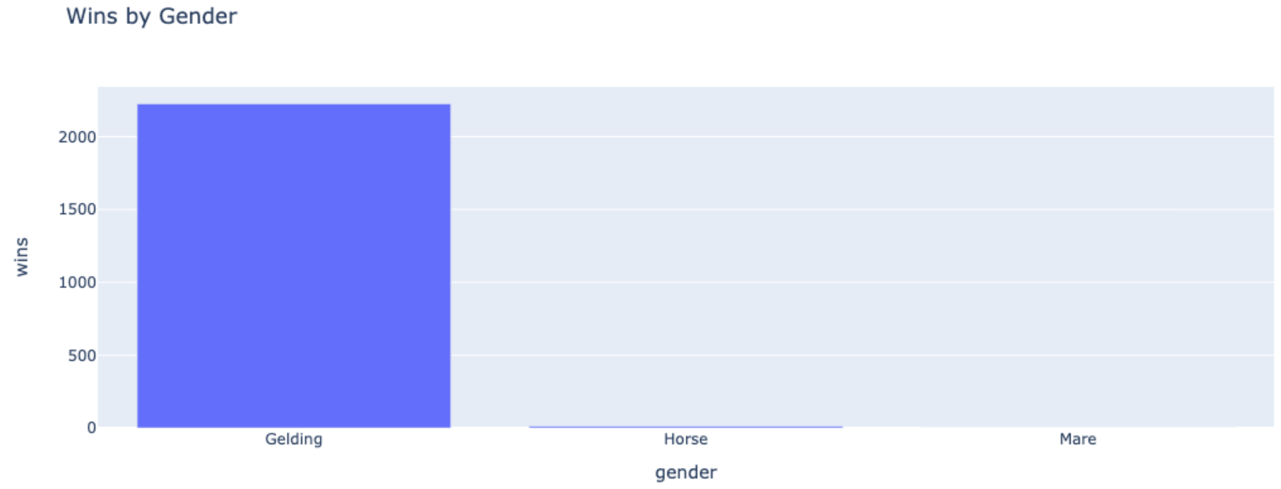


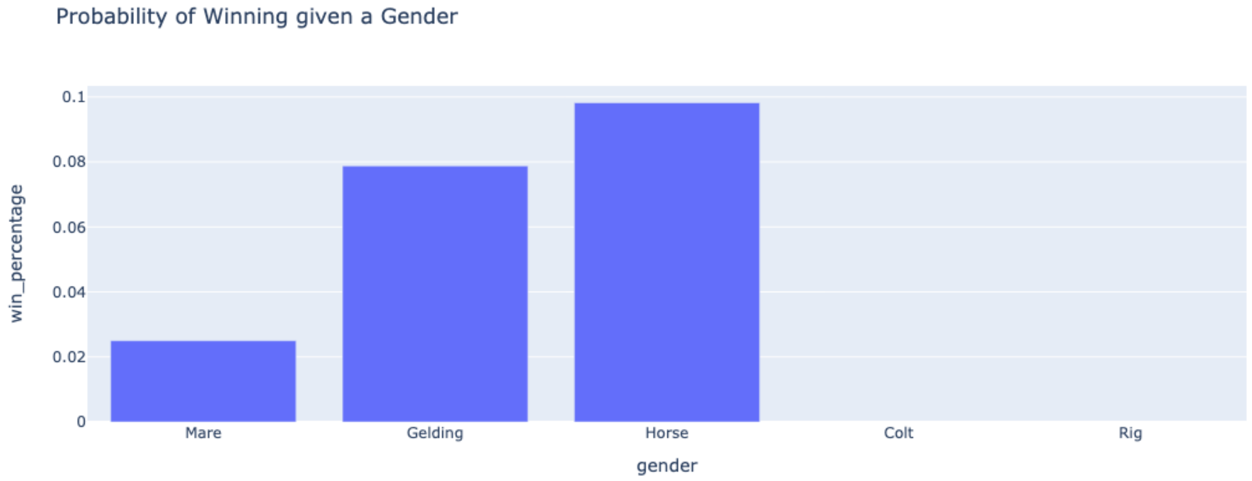


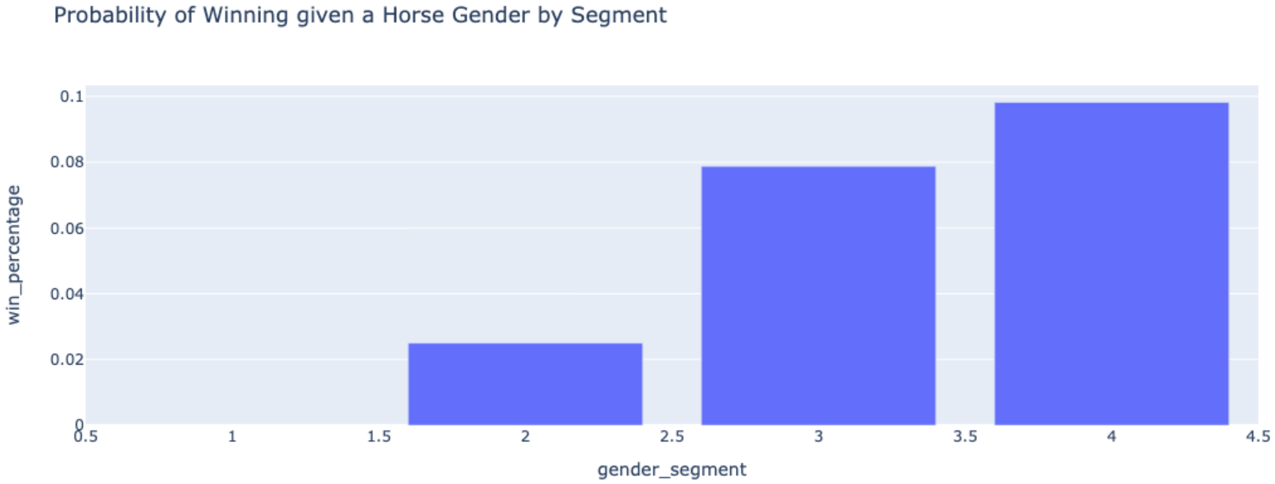


Gender

Based on the breakdown of wins by gender, we see that horses of Gelding gender win a significant amount of races more than others. However, when adjusted based on the amount of data we have from each country, we see that horses of Horse gender have the highest probability of winning. This might be due to a disproportionate amount of horses of Gelding gender running in races. We decided to group and segment the country based on those with similar probabilities to reduce the number of dimensions in the values.







7.0 Reference

Adams, B. R., Rusco, F. W., & Walls, W. D. (2002). Professional bettors, odds-arbitrage competition, and betting market equilibrium. *The Singapore Economic Review*, *47*(01), 111-127. https://doi.org/10.1142/S021759080200033X

Agars, S. (2024, March 18). *HK Racing's softer tracks and slow times: Overwatering, jockeys or seasonal?* South China Morning Post. https://www.scmp.com/sport/racing/article/3255833/hong-kong-racings-softer-tracks-and-slow-times-overwatering-jockeys-or-simply-seasonal

Camara, L., & Cheng, A. (2017, August 18). *Hong kong horse racing results 2014-17 seasons*. Kaggle. https://www.kaggle.com/datasets/lantanacamara/hong-kong-horse-racing

Hausch, D. B., Lo, V. S. Y., & Ziemba, W. T. (2008). *Efficiency of racetrack betting markets* (1st ed., Vol. 2). World Scientific Publishing Co. Pte. Ltd. https://doi.org/10.1142/6910

HKSpeedKing. (2023, September 6). *Top tips for betting on Hong Kong racing*. https://www.hkspeedking.com/post/guide-to-betting-on-hong-kong-racing

Hong Kong Turf. (2018, May 11). *Understanding different race track conditions at Hong Kong racing*. Understanding Different Race Track Conditions At Hong Kong Racing. https://hongkongturf.blogspot.com/2018/05/understanding-different-race-track.html

Powell, D. (2022, September 10). *Hong Kong Racing Study Guide: The horse class system, Expalined* Paulick Report. https://paulickreport.com/news/horseplayers-category/hong-kong-racing-study-guide-the-horse-class-system-explained

The Hong Kong Jockey Club. (n.d.-a). *Racing Information*. Course information - reference information - horse racing. https://racing.hkjc.com/racing/english/racing-info/racing_course.aspx

The Hong Kong Jockey Club. (n.d.). https://www.hkjc.com/home/english/index.aspx