AGAINST THE ODDS

Using Statistics and Machine Learning to Predict the Winning Horses

MICHAEL HOTALING | MJHOTALING@MY365.BELLEVUE.EDU

SUMMARY

Often referred to as "games of chance", gambling can often be broken down into relatively basic mathematics. All outcomes in gambling games have defined probabilities that depend on sample spaces, or the total number of possible outcomes. In poker, the odds of having a winning hand can be calculated based on the probabilities of the cards drawn, as well as some underlying psychology.

Horse racing and other sports events can also be broken down in a similar way. By analyzing physical attributes of each of the horses such as average speed, number historical wins, terrain conditions, weather, and dozens of other data points, we can create a model that can return probabilities of certain horses winning.

Hong Kong has the largest industry for betting on horses with an average pool size of $12.7 million per race, followed by France with an average pool size of $2 million.

BUSINESS UNDERSTANDING

PROBABILITY

Probability is at the center of all gambling. For simple card games such as blackjack or poker, the probability of winning can be easily understood based on which cards a user has in their possession as well as their opponents' cards or the cards in the flop. Skilled players recognize the sample space of the game and the probabilities associated with each of their hands. They are able to make calculated bets based on the likelihood of winning. If their odds of winning are low, they may bet very little or fold.

EXPECTED VALUE

It's not enough to know the probabilities of winning a hand are. We are interested in how much money we can make from each game. This is called the expected value and is calculated based on the probabilities multiplied by their associated gain or loss. For example, casinos have a negative expected value due to the house advantage in all games.

ACTUAL VALUE

More often than not, gamblers overleverage their bets due to superstitions, logical fallacies, and a general lack of understanding for the underlying mathematics of gambling, making betting against them much more lucrative. If the calculated risk from our expected value is much less than the implied risk of the actual value of the bet, it may be financially beneficial to take advantage of that arbitrage. For example, if we calculate the odds of a horse winning placing first in a race is 60% likely, and the betting odds for that horse are +100 (or $\frac{2}{1}$ or 50%), we create a situation which we are able to double our money on a 60% chance. By utilizing the Kelly Criterion, we can wager an appropriate level of money for each bet, ensuring that we don't overextend our positions

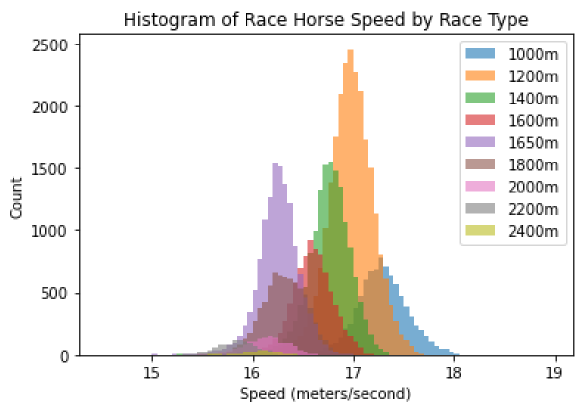
"Gambling is not about how well you play the games, it's really about how well you handle your money."
V. P. Pappy

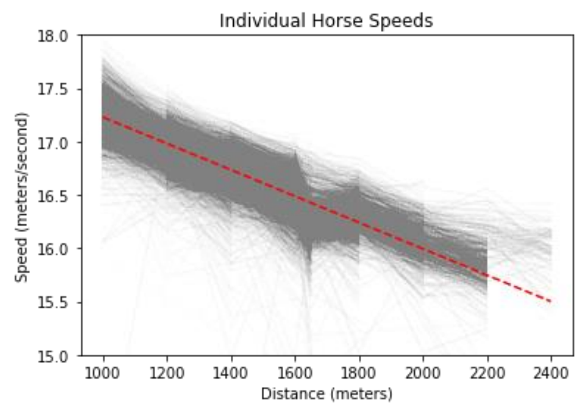
UNDERSTANDING THE DATA

We have access to several years of historical racing data from two famous tracks in Hong Kong containing the date, track conditions, track surface type, and dozens of other useful attributes. We can generate profiles for each horse that participated and build a statistical model to help us evaluate the likelihood of each horse placing in the race.

Horse racing dates back as far as 4500BCE when Nomadic tribesmen raced horses in Central Asia

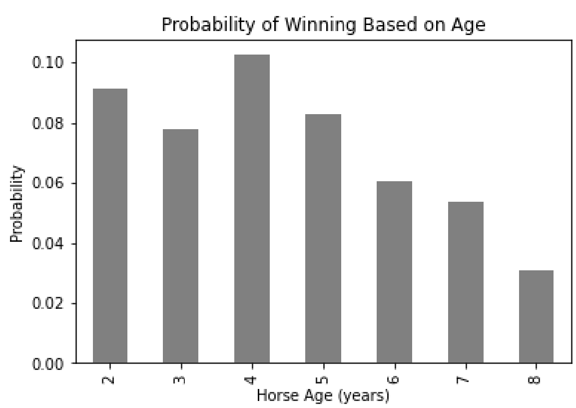
Using the distance of the track as well as the finishing times for each horse, we can estimate the average speed of each horse. For longer distance races, we can see that the jockeys slow down their horses to prevent exhaustion. Any horses with statistically abnormal speeds for their assigned race can be identified and investigated.

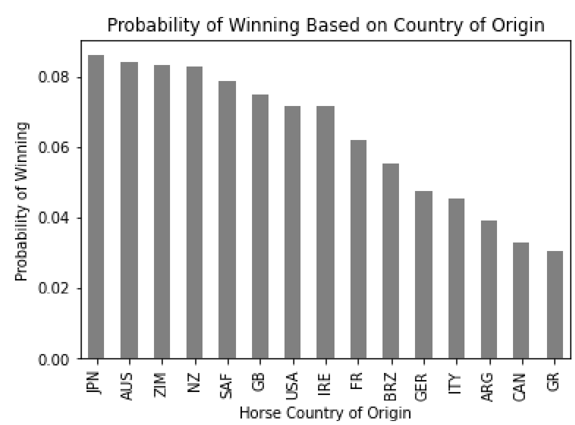




We also have information about the age of our horses. From preliminary exploratory data analysis, we can see that the older a horse gets, the less likely it will win. We can also see that Japanese, Australian, and New Zealander horses are at a slight advantage over horses from other parts of the world, such as Italy, Canada, and Brazil. Perhaps horses experience jetlag.

Betting favorites in horse races prevail in about a third of races, meaning that more than two thirds of races are upsets





DATA PREPROCESSING

We will need to wangle our data quite a bit to build a proper machine learning model. Firstly, we will need to start with creating a dataframe containing statistics for each horse.

Since speed is such an important attribute in racing, we can first start by creating a dataframe containing the average speed for each horse for each different type of race. Since not all horses compete in all races, we may want to create a model for each different type of race since many NA values will be present, or we may be able to estimate the speed of our horses missing statistics using linear extrapolation from horses in our sample set.

distance	1000	1200	1400	1600	1650	1800	2000	2200	2400
horse_id									
3038	NaN	NaN	NaN	16.241464	16.291279	16.097318	NaN	NaN	NaN
1239	17.525412	16.959381	16.699561	16.797082	16.256268	16.012810	NaN	NaN	NaN
4142	17.192433	16.970652	16.778523	NaN	NaN	NaN	NaN	NaN	NaN
323	17.396623	17.040541	16.840838	16.765147	NaN	NaN	NaN	NaN	NaN
504	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	16.081731
3975	17.099863	17.038194	16.780635	16.578090	16.220512	16.415332	16.265319	16.061031	NaN
1653	NaN	NaN	NaN	16.335092	16.155711	16.181971	NaN	15.669516	NaN
2477	NaN	16.733715	16.653688	NaN	16.419544	NaN	NaN	NaN	NaN
1075	NaN	16.955667	16.705947	16.530633	16.241359	NaN	NaN	NaN	NaN
1297	NaN	16.801089	16.785150	16.648979	16.289964	16.261131	NaN	NaN	NaN

Some horses perform better on softer or firm ground than others. We can also generate speed statistics based on the track conditions for both turf and dirt tracks, imputing missing values based on linear extrapolation from other horses.

surface	Dirt						Turf				
going	GOOD	GOOD TO FIRM	GOOD TO YIELDING	SOFT	YIELDING	YIELDING TO SOFT	FAST	GOOD	SLOW	WET FAST	WET SLOW
horse_id											
1867	16.852310	17.049474	17.078621	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3638	16.944197	16.984405	17.196905	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1599	16.687641	16.728286	16.293887	NaN	NaN	NaN	NaN	16.837083	NaN	NaN	16.488458
3088	16.780971	16.722862	16.830295	NaN	17.132088	16.027476	NaN	16.945208	16.999575	NaN	16.652789
2666	16.474484	16.522623	16.031918	NaN	NaN	16.000000	16.609763	16.398674	NaN	NaN	17.133067
3942	16.854954	16.464519	NaN	NaN	17.102788	NaN	NaN	NaN	NaN	NaN	NaN
681	17.050984	17.141063	NaN	NaN	NaN	NaN	NaN	17.178755	NaN	NaN	16.583748
694	17.081228	17.292063	16.666667	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2652	16.908328	17.375508	15.853190	NaN	NaN	NaN	NaN	17.111723	NaN	NaN	NaN
1950	16.694491	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

MODEL DEVELOPMENT

Once our data is formatted to the appropriate structure with all of our desired attributes and statistics in place, we can start training a machine learning model. Our input data will be the data we generated about each of the horse and our output should be the calculated odds of the horse placing in a certain position. For simplicity's sake, we will only assume that we are interested in the winning horses.

Our neural network will output scores based on whether it predicts our horse may win or lose from zero to one. We can then scale our data based on the total sum of our predictions to establish the predicted finishing order.

Here's an example from one of our models. We were able to accuracy predict that horse number 2998 placed in the top 3.

	race_id	horse_id	horse_age	avg_speed	actual_weight	wins	nn_pred	real_loss	real_win	nn_pred_win
6	0	911	3	16.636958	121.958333	3.0	0.074937	1	0	1
9	0	2998	3	15.927509	116.600000	0.0	0.074929	0	1	0
4	0	2796	3	16.758655	125.916667	0.0	0.074786	1	0	0
3	0	1853	3	16.682010	126.250000	0.0	0.074611	1	0	0
8	0	1730	3	16.627078	121.181818	1.0	0.074569	1	0	0
11	0	2617	3	16.706089	123.000000	0.0	0.074335	1	0	0
13	0	306	3	16.615042	112.428571	0.0	0.072924	1	0	0
1	0	2157	3	16.754428	122.468750	2.0	0.072413	1	0	0
7	0	2170	3	16.876863	121.000000	2.0	0.071047	1	0	0
0	0	3917	3	16.619423	120.500000	2.0	0.070808	1	0	0
5	0	3296	3	16.706621	119.625000	1.0	0.068179	1	0	0
10	0	1733	3	16.847438	124.608696	3.0	0.068165	1	0	0
2	0	858	3	16.827020	118.863636	2.0	0.064900	1	0	0
12	0	727	3	16.766933	124.391304	3.0	0.063397	1	0	0

Now that we have established like likelihood of each horse placing in first, we can compare the house odds of each horse winning. If the house odds imply more volatility than we calculate, we can place a bid using the Kelly Criterion. Even if a horse has a very low chance of winning, if we calculate the odds being slightly higher than the house, we can place a bid based on our statistics and get a large return if we are proven correct.

MODEL DEPLOYMENT

Now that we have established like likelihood of each horse placing in first, we can compare the house odds of each horse winning. If the house odds imply more volatility than we calculate, we can place a bid using the Kelly Criterion. Even if a horse has a very low chance of winning, if we calculate the odds being slightly higher than the house, we can place a bid based on our statistics and get a large return if we are proven correct.

We can set up a data pipeline to collect live statistics on horse races all over the world and use online betting website to place our bets. Useful statistics can include historical data such as horse speeds per terrain status, a rolling number of wins per horse as well as environmental attributes the day of the race like weather, humidity, and temperature.

We can also track jockey and trainer performance such as how well their horses perform on turf, dirt, or in a compromised going such as mud. These indicators can stack up to create a very robust algorithm

Once the model is deployed, it may be advantageous to loop in return on investment (ROI) per horse as a feedback attribute for the model. If horses tend to perform poorly, the algorithm will pick them less.

Depending on the performance of our model, we may be able to allow it to run without human intervention. This may be risky due to the fact that the algorithm will be handling real money.