

AGAINST THE ODDS

Using Statistics and Machine Learning to Predict the Winning Horses

MICHAEL HOTALING | MJHOTALING@MY365.BELLEVUE.EDU

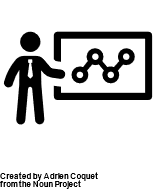
# SUMMARY

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

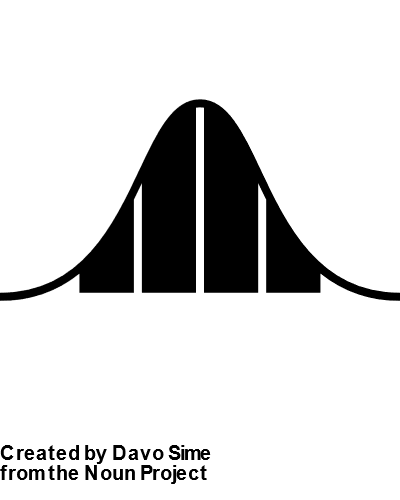
Probability is at the center of all gambling. For simple card games such as blackjack or poker, the probability of winning can be easily understood based on which cards a user has in their possession as well as their opponents’ cards or the cards in the flop. Skilled players recognize the sample space 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.

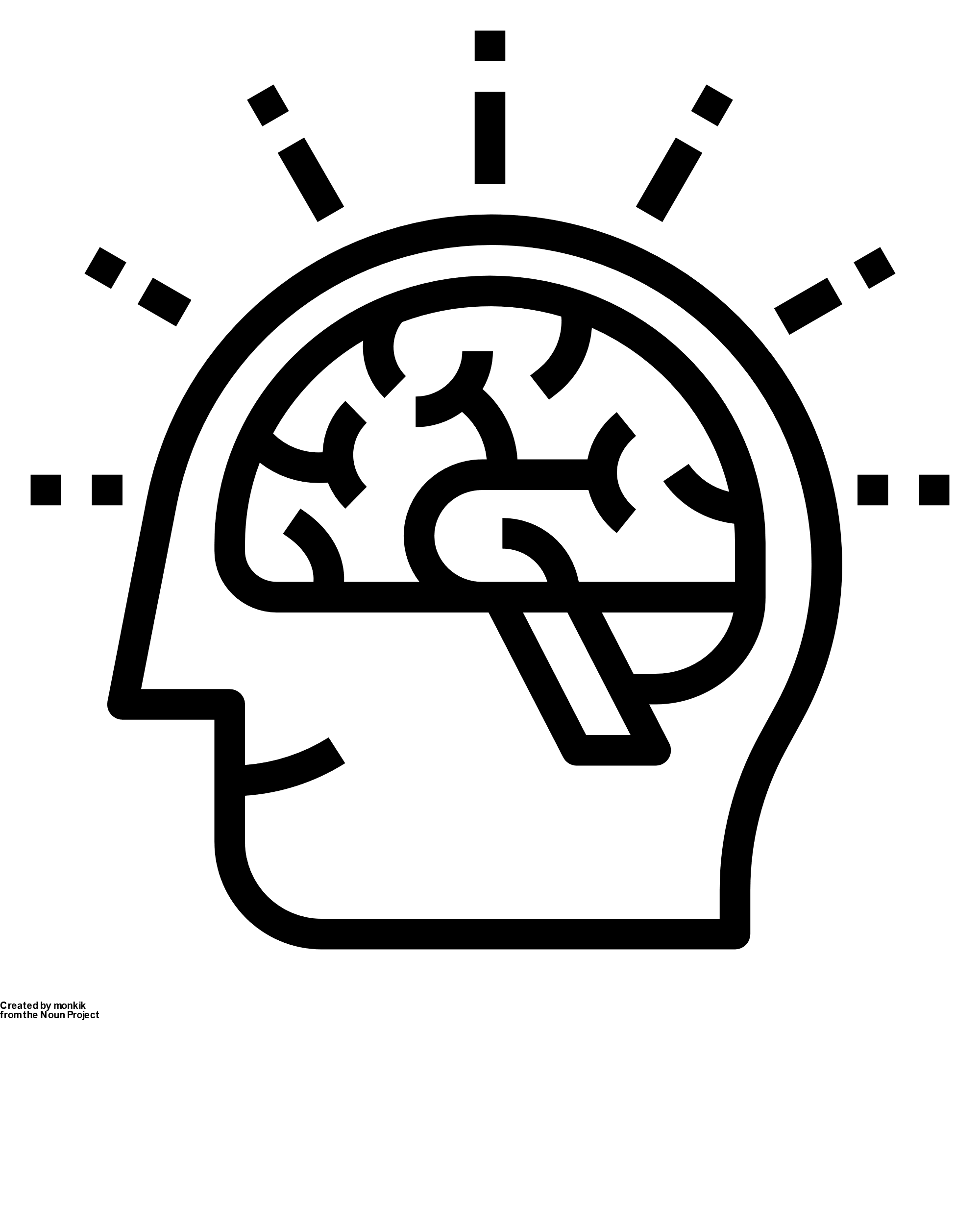
## ACTUAL VALUE

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 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





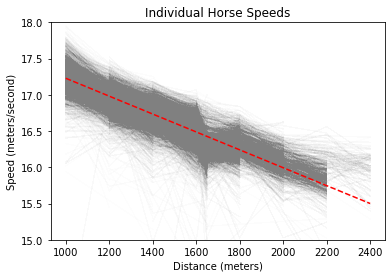
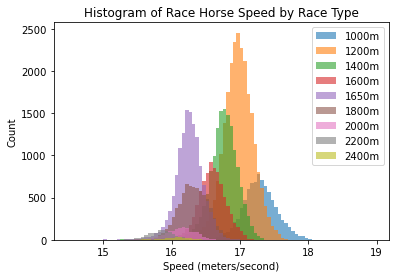
*“G*ambling is not about how well you play the games, it’s really about how well you handle your money.”  
V. P. Pappy



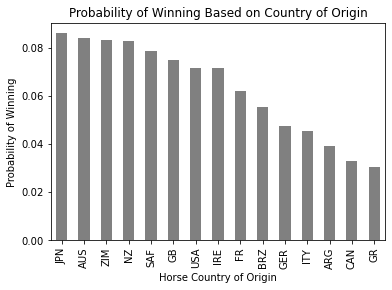
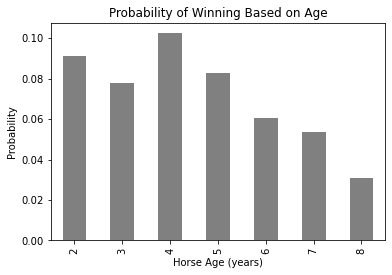
# UNDERSTANDING THE DATA

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.

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.

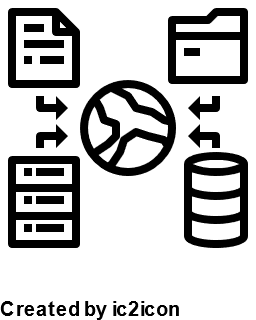


We also have information about the age of our horses. From preliminary exploratory data analysis, we can see that the older a horse gets, the less likely it will win. We can also see that Japanese, Australian, and New Zealander horses are at a slight advantage over horses from other parts of the world,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



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

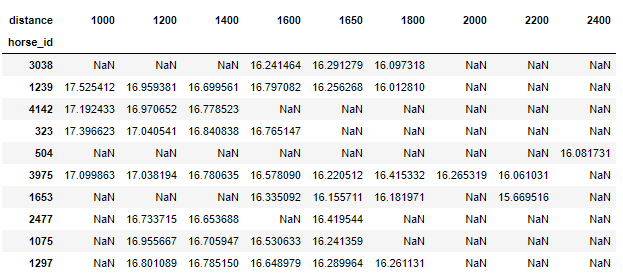
Betting favorites in horse races prevail in about a third of races, meaning that more than two thirds of races are upsets



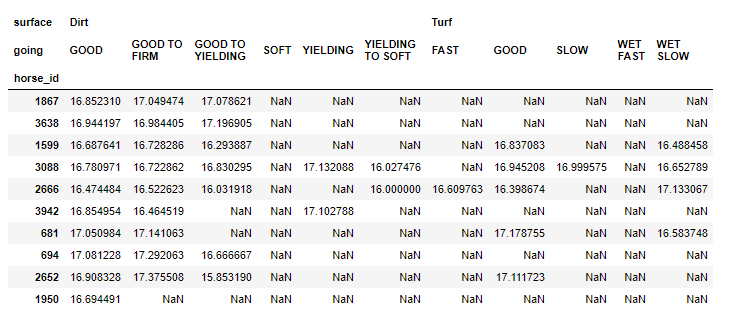
# DATA PREPROCESSING

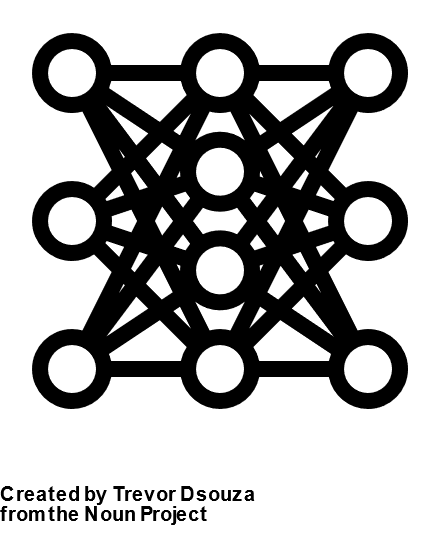
We will need to wangle our data quite a bit to build a proper machine learning model. Firstly, we will need to start with creating a dataframe containing statistics for each horse.

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.



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.



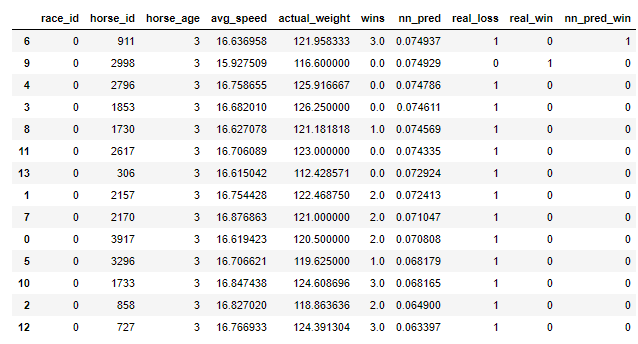


# MODEL DEVELOPMENT

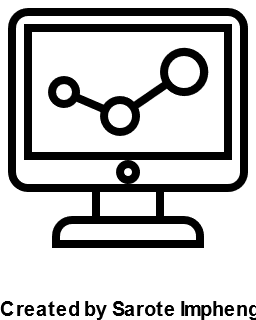
Once our data is formatted to the appropriate structure with all of our desired attributes and statistics in place, we can start training a machine learning model. Our input data will be the data we generated about each of the horse and our output should be the calculated odds of the horse placing in 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.



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.

