Shaefer Drew

SI 330 Final Report

Horse Racing

Motivation

I've always had a passion for predicting the future. I exercise this passion in many different forms: Stock trading, fantasy football, march madness, etc. Basically, I enjoy gambling and using data to provide valuable insights that will make my bets less risky and generate a higher return on investment. I came across horse racing data for multiple reasons: I love betting on sports and I'll be living within driving distance of the Kentucky 500 this summer.

Goal

My goal was to analyze the data sources and make a calculated guess based on my findings. I knew very little about horse racing, much less how the bets worked, going into this project. In order to get a basic understanding of my data sources and all its components, I first familiarized myself with the different types of bets and the odds system. After gaining a basic understanding of how horse race betting worked, I wanted to use the data to develop a betting strategy, as well as more of an in depth understanding of horse race history.

Questions I wanted to test were:

1. Are odds an accurate predictor of winners and losers?
    a. Are the odds of winning horses significantly smaller than the odds of losing horses?
    b. Are winning odds and finishing position correlated?
2. Based on their track record, which tipster is the most reliable?
3. Which horses, trainers, and jockeys are the top performers?
4. Are there groups of horses that carry similar traits and will those traits help us determine a winning horse?
5. Which factors are most highly correlated to finishing position?

By answering these questions, I could determine the most reliable combinations of horses, jockeys, tipsters, and other variables.

Data Sources

1. Tip Data

This dataset is a 2.7 MB CSV file from Kaggle, which I read into python using a pandas dataframe. It consists of 39,000 horse racing bets from 31 tipsters throughout various locations.

2. Race Data

This dataset is a 2.8 MB CSV file from Kaggle, which I again read into python using a pandas dataframe. It includes Hong Kong horse racing results from 2014-2017.

Data Manipulation Methods, Analysis, and Visualization

Data Cleaning

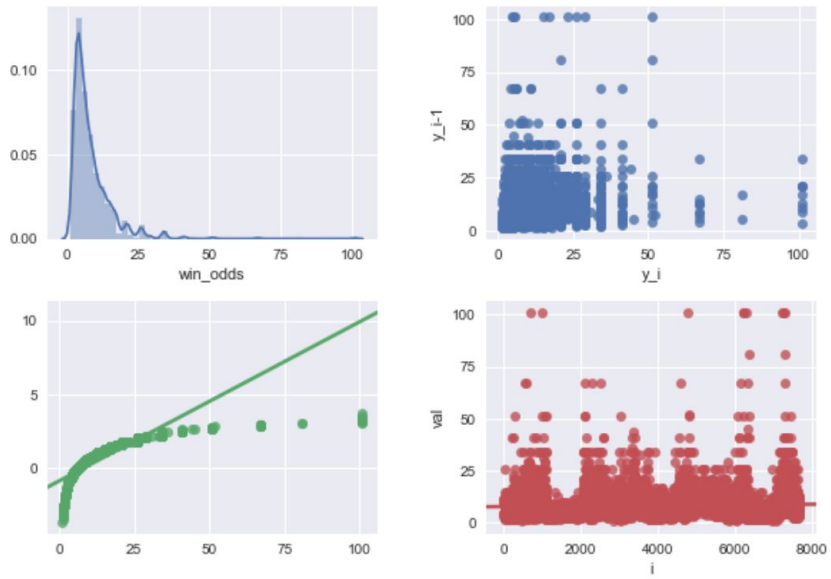
I began by reading the two datasets into pandas as independent dataframes, df_tips and df_races. I inspected the data, removing all of the null values, dropping and replacing certain string values and unknown terms, and converting data types to accurately represent different columns and make it as convenient as possible for their analysis. I approached the challenge of merging the two dataframes by first renaming columns with the same variables but different column names. I decided that I wanted to merge on the columns "horse_name, win_odds, and Result" in order to create an even larger dataset to analyze those relationships from. I ended up using pandas' concat function to concatenate the two datasets.

Question 1: Are odds an accurate predictor of winners and losers?

Some background: Odds are based primarily on payoff. Therefore, if you bet on a horse that has high odds and it wins, you will make more money. Horses with higher odds are said to be "underdogs" while horses with lower odds are the favorites. I wanted to determine if these odds were an accurate predictor of the actual winners and losers.

A. Are the odds of winning horses significantly smaller than the odds of losing horses?

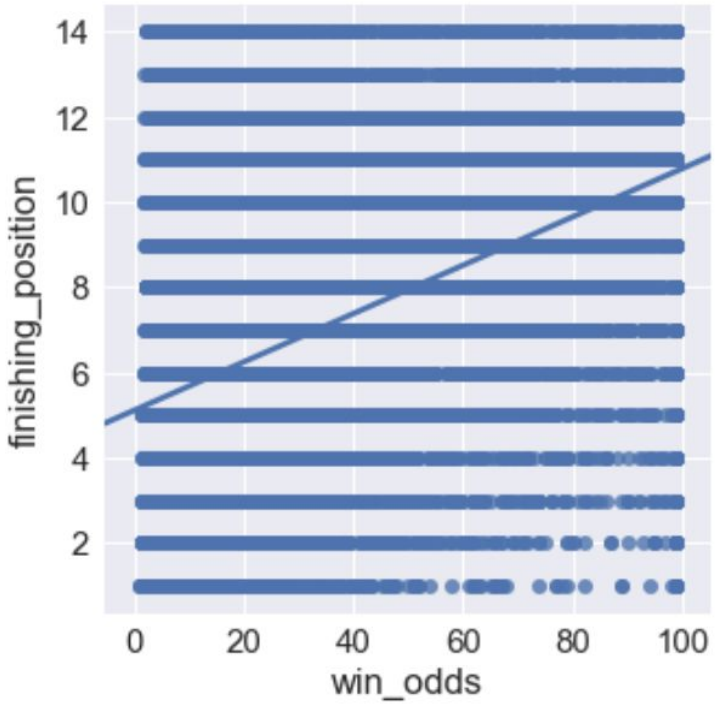
In order to answer this question I took the merged data frame and split it into 2 data frames, consisting of losing horse races and winning horse races. Next, I used a function I found from SI 370 and plotted the histogram, QQ plot, lag plot, and run sequence plot primarily for the purpose of determining if the data was normally distributed.



Since the data did not appear to be normally distributed, I used a kruskal-wallis test to determine if the difference in odds for winners and losers was significant. The test yielded a significant result, allowing a rejection of the null hypothesis that the median distribution of odds between the groups was the same. The odds of winning horses turned out to be significantly lower than the odds of losing horses, with an absolute mean difference of 12.42 and a median difference of 5.

B. Are winning odds and finishing position correlated?

As seen in the linear plot and using bivariate linear regression, we can conclude that winning odds and finishing position are positively correlated. Although, it is hard to predict a horse's position based on odds alone due to the many outliers.



Question 2: Based on their track record, which tipster is the most reliable?

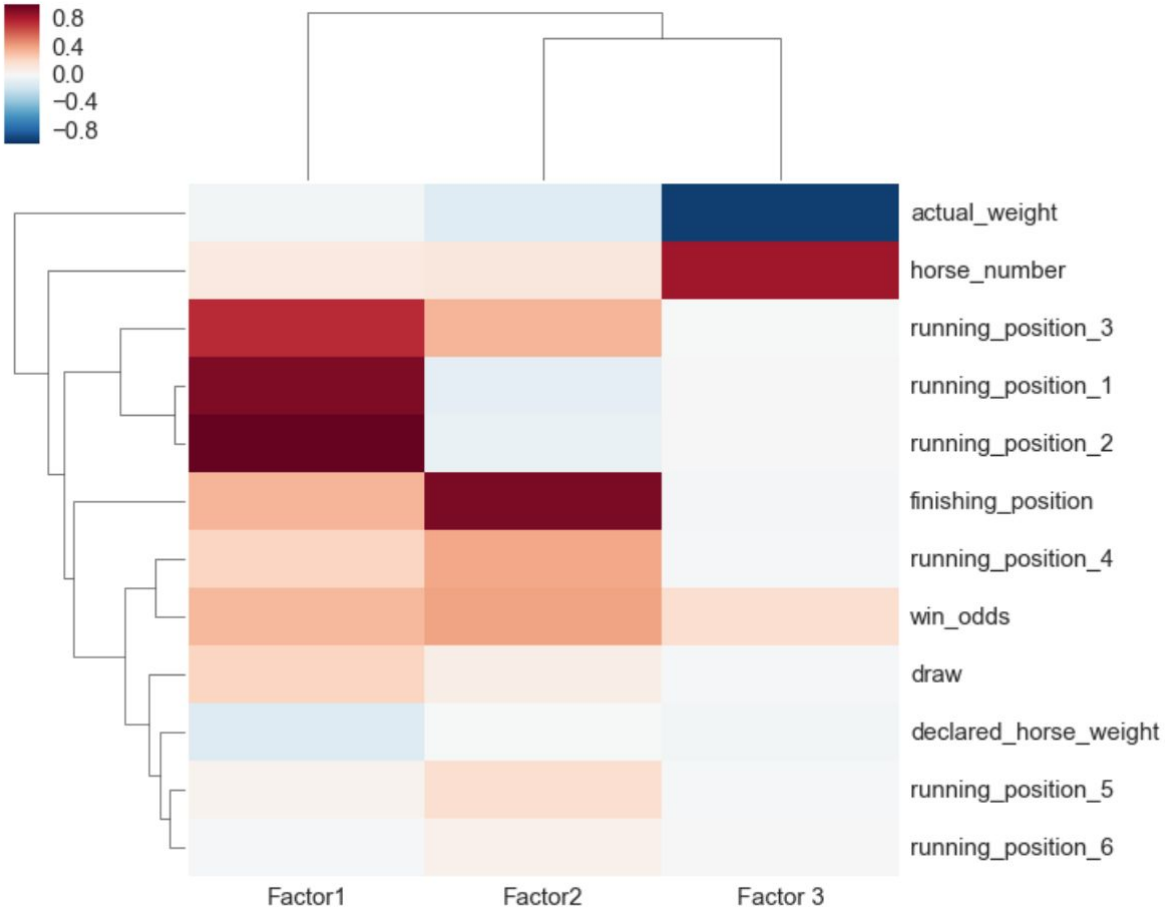
I used a pivot table to group different tipsters by the total number of wins and losses they had and calculated their success rate. From there, I sorted from highest to lowest success rate and determined who the most reliable tipsters were.

Question 3: Which horses, trainers, and jockeys are the top performers?

I created pivot tables for each category, calculating the average finishing position and winning odds and sorting by finishing position from least to greatest in order to see the top performers.

Question 4: Are there groups of horses that carry similar traits and will those traits help us determine a winning horse?

I decided to perform PCA factorial analysis to see if different groups of horses could belong to different clusters based on a number of factors.



The result wasn't too insightful, primarily because of the lack of characteristics and traits in the data; however, it makes logical sense. Horses who were frequently in low positions tended to have lower winning odds and by far the lowest finishing position. This also provided some insight into the type of analysis I needed to perform next.

Question 5: Which factors are most highly correlated to finishing position?

In order to determine correlation and accurately predict a horse's finishing position using the given factors, I built a multivariate linear model.

OLS Regression Results

Dep. Variable:	finishing_position	**R-squared:**	0.547
Model:	OLS	**Adj. R-squared:**	0.547
Method:	Least Squares	**F-statistic:**	3934.
Date:	Mon, 16 Apr 2018	**Prob (F-statistic):**	0.00
Time:	20:06:31	**Log-Likelihood:**	-68670.
No. Observations:	29364	**AIC:**	1.374e+05
Df Residuals:	29354	**BIC:**	1.374e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.3145	0.309	-1.017	0.309	-0.921	0.292
win_odds	0.0359	0.001	70.354	0.000	0.035	0.037
total_weight	0.0024	0.000	9.960	0.000	0.002	0.003
horse_number	0.0036	0.004	0.906	0.365	-0.004	0.012
running_position_1	-0.0837	0.011	-7.658	0.000	-0.105	-0.062
running_position_2	-0.2329	0.012	-19.086	0.000	-0.257	-0.209
running_position_3	0.6206	0.006	104.027	0.000	0.609	0.632
running_position_4	0.2240	0.004	63.529	0.000	0.217	0.231
running_position_5	0.1083	0.006	17.944	0.000	0.096	0.120
running_position_6	0.0569	0.016	3.635	0.000	0.026	0.088

Omnibus:	149.075	**Durbin-Watson:**	1.220
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	151.821
Skew:	-0.170	**Prob(JB):**	1.08e-33
Kurtosis:	3.096	**Cond. No.**	2.61e+04

Running positions and winning odds had the highest correlation and influence on the model. With an R-squared value of .547, this model is relatively reliable, given the nature of horse racing, in predicted the finishing position.

Conclusion

Using data analytics, I was able to expand my knowledge of horse racing and build an actual strategy around a sport that is said to be based on luck. By analyzing data and learning who the best tipsters, horses, and jockeys were, then learning how to predict placement based on the other given factors, I was able to optimize my betting strategy.

Sources and Data

1. https://www.kaggle.com/gunner38/horseracing/data
2. https://www.kaggle.com/lantanacamara/hong-kong-horse-racing/data
3. SI 370 course material