CS 3250
Team KAM
https://kevinalig.github.io/CS3250/

Kevin Alig
Andrew Sjoberg
Matthew Burks

Wagering on Races: A Study on How to Increase Payouts

Betting payout odds in horse racing are based on a parimutuel betting strategy. The pool of money wagered, minus contestant prizes and "house" cut of about 15%, is divided so all winning combinations share the winnings. This is similar to a lottery payout, except the number of wagers on each horse is known, while a lottery doesn't know how many wagers are on each number combination until after the numbers are drawn. Because the odds change based on wager amounts placed, the higher odds will have a lower payout per wager. Hence a 'long-shot' horse (a horse with few wagers) will share the prize with fewer betters and will receive a higher payout. Our goal with this research project is to increase the payout for an individual, and the following hypothesis aims to create data analysis experiments on the nature of the betting methods provided by institutions.

Our first hypothesis postulates that a favored horse is not the most likely to win; therefore, betting against the favored horse should provide a bigger payout. According to some analysts, the favored horse is simply a popularity contest (USAPlayers.com), so a bet on a favored horse simply leads to a smaller payout divided amongst many participants. We will look into our data set to show that the favorite horse is not necessarily the winner.

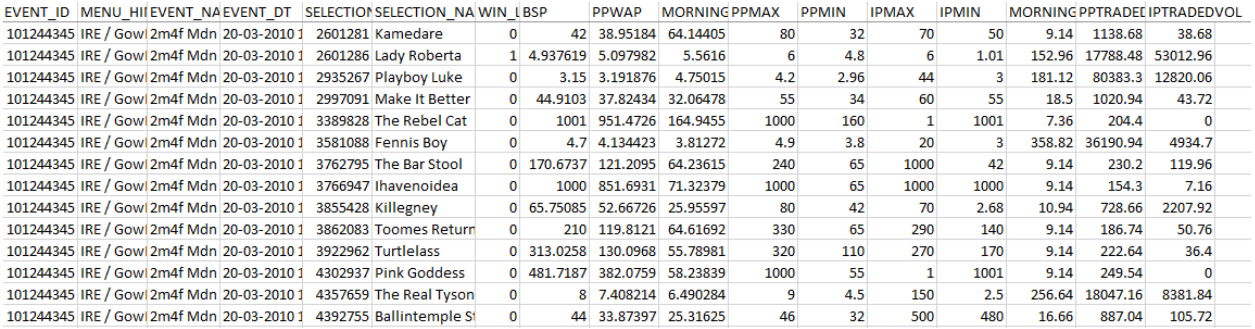
This first hypothesis brings us to the next approach in our analysis: Betting on 'lower odds' (less favored) horses will provide a higher payout. We propose that over the long run, placing wagers on horses with greater payouts could be a smarter bet given an enough length of time. Since the payout of 'lower-odds' bets is higher than bets placed on favored horses, bets of this nature afford a higher margin of losses to increase payout over time.

An alternative to picking individual horses is to create wagers that fall into a 'window of odds'. Our data should show that when the odds vs the winner is averaged together, a grouping will emerge indicating where smart bets should be placed. By comparing odds to actual winners, we believe the data analysis will show that inside of an odds window, say between 4-1 and 8-1, will provide the best chance of payout over time.

Datasets for this study were obtained from the biggest betting exchange worldwide: Betfair (TheGuardian.com). Betfair started its operations in 2000, with a steady growth that allowed the purchasing of England's biggest horseracing publishing company, Timeform, in 2006. In 2009, Betfair purchased the TVG Network in the United States, which allowed for Betfair to offer wagering services in 31 of the States. Betfair also reached a deal in 2009 with the New York Racing Association, which allowed for Betfair to offer wagering and exchanges on Aqueduct's thoroughbreeding races (ThoroughbredTimes.com). More recently in 2015, Betfair merged with the biggest bookmaker in Ireland, Paddy Power, which cemented Betfair as the biggest betting exchange operating across the globe. Although public access to archived data only extends back to 2007, the dataset Betfair holds on sport gamblings far surpass those offered by any other research firm or corporate entity such as ESPN. Therefore, our datasets for this research topic are exclusively gathered from Betfair.

Betfair offers an up-to-the-minute database via their API. However, this requires an active account. For our purposes, we require historical data rather than the latest data. Betfair provides previous days' racing data freely through its public website at the following link: https://promo.betfair.com/betfairsp/prices/index.php. These .csv files date back to 2007 with data from horse racing in multiple countries, to include the United Kingdom, Ireland, Australia, the United States of America, and South Africa. Collecting the off of their server was a trivial task with web-browser plug-in tools such as "DownloadThemAll", which scrapes a server and downloads all dependent files from a public site.

Each .csv file provided by Betfair has a patterned name of their 'exchange-price-location-winnings-date'. The files are formatted the same way internally. Each file represents the days' worth of races (times based out of Gibraltar). Attributes in these files include the 'Event ID', 'Name of Runner (horse)', 'Win/Lose (or 'Place'; effectively 1st place, 2nd place, or loss), 'Starting Price' on wagers, and amount that each wager sways up to the time of the race as well as the amount of trading through the Betfair exchange.



EVENT_ID	MENU_HI	EVENT_NA	EVENT_DT	SELECTIOI	SELECTION_NA	WIN_L	BSP	PPWAP	MORNING	PPMAX	PPMIN	IPMAX	IPMIN	MORNING	PPTRADEI	IPTRADEDVOL
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	2601281	Kamedare	0	42	38.95184	64.14405	80	32	70	50	9.14	1138.68	38.68
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	2601286	Lady Roberta	1	4.937619	5.097982	5.5616	6	4.8	6	1.01	152.96	17788.48	53012.96
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	2935267	Playboy Luke	0	3.15	3.191876	4.75015	4.2	2.96	44	3	181.12	80383.3	12820.06
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	2997091	Make It Better	0	44.9103	37.82434	32.06478	55	34	60	55	18.5	1020.94	43.72
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	3389828	The Rebel Cat	0	1001	951.4726	164.9455	1000	160	1	1001	7.36	204.4	0
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	3581088	Fennis Boy	0	4.7	4.134423	3.81272	4.9	3.8	20	3	358.82	36190.94	4934.7
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	3762795	The Bar Stool	0	170.6737	121.2095	64.23615	240	65	1000	42	9.14	230.2	119.96
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	3766947	Ihavenoidea	0	1000	851.6931	71.32379	1000	65	1000	1000	9.14	154.3	7.16
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	3855428	Killegney	0	65.75085	52.66726	25.95597	80	42	70	2.68	10.94	728.66	2207.92
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	3862083	Toomes Returr	0	210	119.8121	64.61692	330	65	290	140	9.14	186.74	50.76
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	3922962	Turtlelass	0	313.0258	130.0968	55.78981	320	110	270	170	9.14	222.64	36.4
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	4302937	Pink Goddess	0	481.7187	382.0759	58.23839	1000	55	1	1001	9.14	249.54	0
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	4357659	The Real Tyson	0	8	7.408214	6.490284	9	4.5	150	2.5	256.64	18047.16	8381.84
101244345	IRE / Gowl	2m4f Mdn	20-03-2010 1	4392755	Ballintemple St	0	44	33.87397	25.31625	46	32	500	480	16.66	887.04	105.72

Figure 1 - Bitfair CSV File for a sample event.

In thanks to the same attributes being present across all of the files, merging the files is also a trivial process. In Windows, the command line shell allows for a simple argument of "[directory] copy *.csv mergedFileName.csv" (AnalystCave.com). A Python script would require a File I/O function which itself would still be trivial (fout, for loop, write).

We have over 22,500 data sets gathered from Betfair, with an average of 120 rows of data per set. The datasets offered by Betfair gives us two minor hurdles. Their horseracing data mixes in their dog racing data. Fortunately, greyhound race results are aptly named 'greyhound' as part of their filename and can easily be sorted out. The second issue comes in the form of poorly kept records; some races do not have betting data included. Without the starting wagers, we do not have a way in which to calculate odds and favors. Some datasets also exclude the list of winners and losers, which also hinders us from calculating final payouts based on wins. Therefore, whenever we complete our data analysis in R, we will need to exclude entire 'Event ID' sets that do not have either a complete set of betting data or winners.

In order to answer our hypotheses regarding an overall favored horse's likelihood to win its given race, we should have enough data points to provide a good probability model fitting a binomial distribution, since we are just measuring the win/loss (two-outcome) event per race. With over 22,500 datasets based on daily events, each file contains an average of four events, meaning that our dataset for this probability model will have about 90,000 data points. Although we have a large number of data points, a simple binomial distribution curve may be the best way to visualize this data. Since our hypothesis is meant to provide insight into our big question of 'how to increase payouts from horse racing bets', the normalized probability distribution should be directly proportional to the payout odds. This same methodology would address our additional hypothesis in which we wish to see if we receive a better payout by betting on a horse favored in second to the top favored horse.

In order to analyze the data in regards to our first hypothesis, we want to focus on the favored horses. We will record how many times a horse with the highest total wager placed won or lost, and to compare those two results. This would not only give us an analysis on win/loss ratio, but it also shows how often a favored horse performs well on the racetrack. That in turn determines if that popular favorite should truly be voted on or not in order to maximize a good betting return.

We also want to know if placing a wager on a lower ranked horse increases our payout chances. We postulate that earnings from a lower-ranked wager can overcome money sunk into a loss. A horse is lower ranked than the highest if it receives a lower average wager, which is an attribute provided in our datasets. The dataset also has information regarding the final amount that was wagered before the race began. To make an analysis, we just need to compare the amount that was won or lost, and sum up all of these up over a specific time interval. If the result is positive, then all of the wagers were worth paying for at the end of the time frame.

To address our third hypothesis on a 'window of odds', we will need to create an artificial bracket of odds. For example, a horse favored 1-1 to 5-1 are all grouped in the same bracket, then 6-1 to 10-1 are grouped in a different bracket, and so on and so forth. This creates a sliding bracket (window) of odds. Then in R, we can compare how many horses won per each window, with an accompanying bar graph for each betting window. We will then create a correlated graph showing the payout for each bracket according to their winning probabilities. This can be a dynamic graph that is generated in Tableau so that the user can switch between each graph (Tableau.com), overlay the graphs, or generate their own brackets and graphs using our data (as opposed to viewing our pre-made graphs).

Works Cited

"Betfair NYRA Reach Wagering Agreement." Thoroughbred Times. N.p., 02 Nov. 2009. Web. 26 Oct. 2016.

"Display Measures Dynamically." Tableau Software. N.p., 16 Aug. 2016. Web. 26 Oct. 2016.

"Horse Betting Guide." U.S. Players. N.p., n.d. Web. 26 Oct. 2016.

"Merge CSV Files or TXT Files in a Folder - Using Excel or CMD." The Analyst Cave Excel VBA Programming and More. N.p., Mar.-Apr. 2016. Web. 26 Oct. 2016.

"Will Betfair Become a Player in the U.S.? - Horse Racing News | Paulick Report." Horse Racing News Paulick Report. N.p., 2010. Web. 26 Oct. 2016.

Wood, Greg. "World's Biggest Betting Exchange Betfair and Paddy Power to Merge." The Guardian. Guardian News and Media, 2015. Web. 26 Oct. 2016.