Bettor’s Bet: How Las Vegas sportsbooks determine lines

MICHIGAN PROBLEM GAMBLING HELPLINE
1-800-270-7117
TRAINED COUNSELORS AVAILABLE 24/7

Introduction

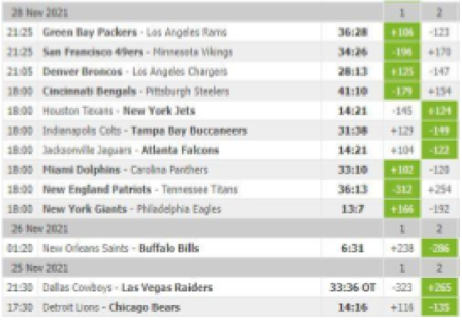
The continued rise of sports betting in the U.S. had me pondering if this type of gambling is any ‘safer’ than gambling at a casino. After all, making an educated guess on a game’s outcome takes at least some skill as opposed to leaving it all to chance.

Having some knowledge of a game’s possible outcome means bookies hedge their risk by setting odds; returning less money if an outcome is likely, and more money if it’s less likely. But how exactly do they set these lines?

Is it possible to derive your own odds to then compare with bookie odds to find the best bets to make? Is it all luck after all?

Why do ‘favorites’ lose?

-### = favorite   +### = underdog



28 Nov 2021			1	2
21:25	**Green Bay Packers** - Los Angeles Rams	36:28	+106	-123
21:25	**San Francisco 49ers** - Minnesota Vikings	34:26	-196	+170
21:05	**Denver Broncos** - Los Angeles Chargers	28:13	+125	-147
18:00	**Cincinnati Bengals** - Pittsburgh Steelers	41:10	-179	+154
18:00	Houston Texans - **New York Jets**	14:21	-145	+124
18:00	Indianapolis Colts - **Tampa Bay Buccaneers**	31:38	+129	-149
18:00	Jacksonville Jaguars - **Atlanta Falcons**	14:21	+104	-122
18:00	**Miami Dolphins** - Carolina Panthers	33:10	+102	-120
18:00	**New England Patriots** - Tennessee Titans	36:13	-312	+254
18:00	**New York Giants** - Philadelphia Eagles	13:7	+166	-192
26 Nov 2021			1	2
01:20	New Orleans Saints - **Buffalo Bills**	6:31	+238	-286
25 Nov 2021			1	2
21:30	Dallas Cowboys - **Las Vegas Raiders**	33:36 OT	-323	+265
17:30	Detroit Lions - **Chicago Bears**	14:16	+116	-135

Source: https://www.oddsportal.com/american-football/usa/nfl/results/

Data

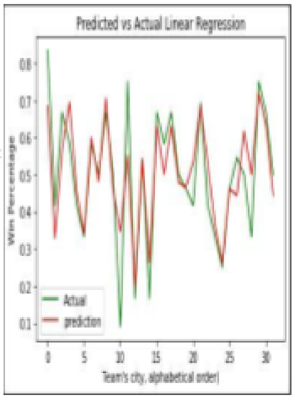
Algorithms used by bookies to set lines are proprietary, and not available to the public. Therefore, I selected data points from different sources that *might* be used to create lines.

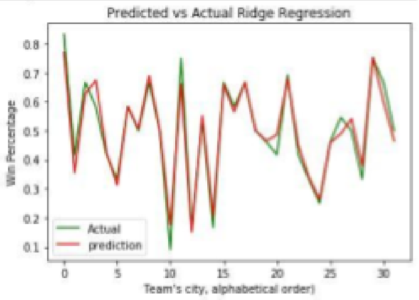
- Points for/against/differential, margin of victory, strength of schedule, offensive/defensive Simple Rating (PFR metric)
  - https://www.pro-football-reference.com/years/2021/
- Average points team covers the spread by
  - https://www.teamrankings.com/nfl/trends/win_trends/
- Yards - total/per game, pass total/per game, run total/per game, points per game
  - https://www.footballdb.com/stats/teamstat.html?group=O&cat=T&yr=2021&lg=NFL
- Odds calculations
  - https://www.legitgamblingsites.com/online-betting/calculating-odds/

Results

Results vary to say the least. One model came up with a high accuracy score, inferring possible overfitting. Other models return accuracy scores similar to random guessing.

I initially thought my models weren’t very good (at all), until I realized they were inline with weekly odds results like the image shown to the left. Manually counting the number of favorite and underdog wins reveals it’s almost 50/50. The football week shown on the left was chosen at semi-random. The ‘bias’ was in knowing my favorite team won that week.





Possible overfitting from ridge regression model

Further Analysis

In attempting to figure out the ‘secret sauce’ bookies use to determine game lines, I was caught in a cycle between determining my own lines and trying to reverse engineer bookie lines. Learning of the ‘moneyline guessing game’ inclines me to further develop my own lines, and use those to exploit mispriced odds.

Conclusion

It’s very possible gambling on a game’s outcome is a game of chance after all, or it could be what the bookies want you to think. The job of the oddsmakers is to generate revenue for the casino or sportsbook. I think it is very possible oddsmakers intentionally mislead the public with ‘favorites’ and ‘underdogs’ in order to draw bets from both sides. The more we guess the more they win.

Bettor's Bet:
How Las Vegas Sportsbooks Determine Lines

Chris Peña
School of Information
University of Michigan
chpena@umich.edu

December 13, 2021

Abstract

Sports betting is becoming legal in more states each year, and it's popularity continues to increase. Bettors wager on the outcome of sporting events, and winning bets receive a payout that is determined by a pre-established 'line'. Las Vegas casinos and sportsbooks have been known to be the gold standard in setting lines, and this project explores the science behind those lines. Different machine learning models were built and ran in attempts to reverse engineer the lines, while others attempted to predict outcomes of NFL games. Results were mixed, but they reveal the profit-driven business model of sportsbooks. Determining independent lines would take more time and research than this project timeline allowed.

Introduction:

Sports betting is rapidly becoming one of the biggest online markets in the country. Its growing popularity is encouraging citizens to petition or vote for public office candidates that support online sports betting, and we see more and more states legalizing sports betting each year as a result. Its rise is met with an influx of competing sportsbooks, or bookies, trying to tap into the growing market. Each sportsbook offers bets for multiple sporting events, and bets can be made based on the bookie’s ‘betting lines’. The betting line is set based on likelihood odds, with ‘more likely’ outcome odds paying less than ‘less likely’ outcomes. Each sportsbook offers their own lines, meaning the same bet could pay more using one bookie over another. With its long history of profiting from gambling, Las Vegas sportsbooks are generally considered to be the consensus, or gold standard, in setting betting lines. But how do they develop these lines with such accuracy that it often seems like they know the outcome before the event even takes place?

The premise of this project was founded on that question, but it quickly became a rabbithole of research and data pre-processing. Unlike most assignments or projects I’ve done where new insight was the goal of data collection and analysis, this project was more in the realm of reverse engineering. Finding the betting lines and results is a simple task. Trying to dissect the black box of Las Vegas oddsmakers was far more complicated than I had originally imagined.

I focused on National Football League games for this project as I possess good domain knowledge of the sport, and I figured I’d have a solid understanding of what data oddsmakers might use to develop their lines. The process seemed straightforward at first with the path of gathering data (various game stats), gathering outcomes (the final line), and training one or multiple machine learning models to develop the lines set by Las Vegas.

Data and Methods:

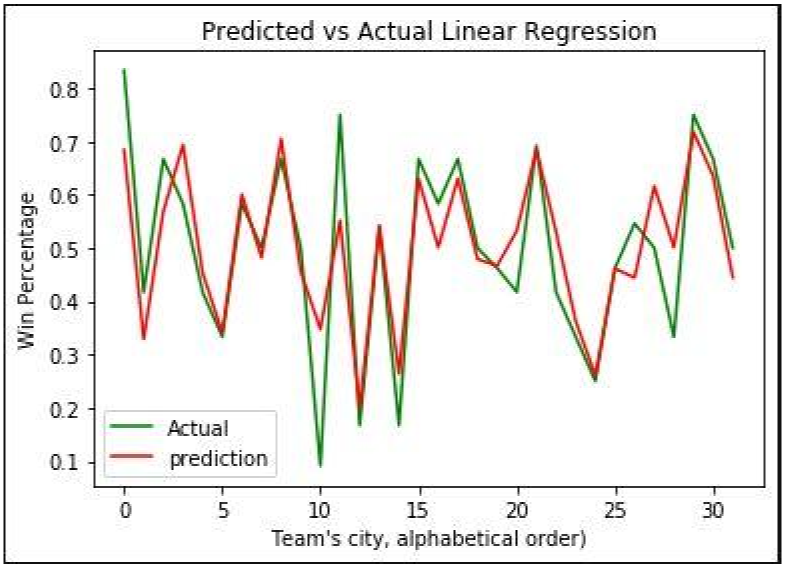
As mentioned, I began data collection by combining several different game statistics for each team. I used multiple sources for the game stats, selected the ones I felt were most relevant, and created a single dataframe for training and testing. Pro Football Reference is a well-known source of NFL statistics, and I was able to collect a large portion of the dataset from their current year page. Another large portion of the data was found on a newly-discovered site, The Football Database. Their data is a great breakdown of the overview data found on sites such as Pro Football Reference. The breakdown could be extremely helpful as it explains, for example, the distribution of rushing and passing yards to the total yards and total yards per game. This type of data could be a predictor of game outcomes if it, say, compares Team A’s total yards allowed and total yards per game allowed to Team B’s total yards and total yards per game. If Team B amassed most of its total yards from rushing, but Team A’s yards allowed per game is primarily from passing, then it could be a beneficial signal for Team A even if Team B is considered the better team. The idea here is if Team A's defensive strength is Team B's offensive weakness, the model could potentially learn this pattern to create appropriate betting lines. Team Rankings was another source found that offered potentially useful information, and some data was used from their site. The betting lines, both upcoming and historical for the year, were conveniently found on the Odds Portal website. It could have been by mere chance, but their website design made a possible pattern pop out. Considering it's a sports betting site it is very possible that the design was intentional. I didn't think much of the possible pattern at first, but it revealed something interesting later on.

Regression analysis seemed like the clear way to go for model building. A standard linear regression model was used, followed by a ridge regression. The standard linear regression would be used as a baseline model. The ridge regression model could then be used to apply weights to the model in hopes of better prediction results. I wanted to save a section of the project to win / loss predictions for teams, but I didn't think I'd spend too much time on it as it wasn't the intended goal of the project. I would build a linear SVC and a logistic regression model for this purpose.

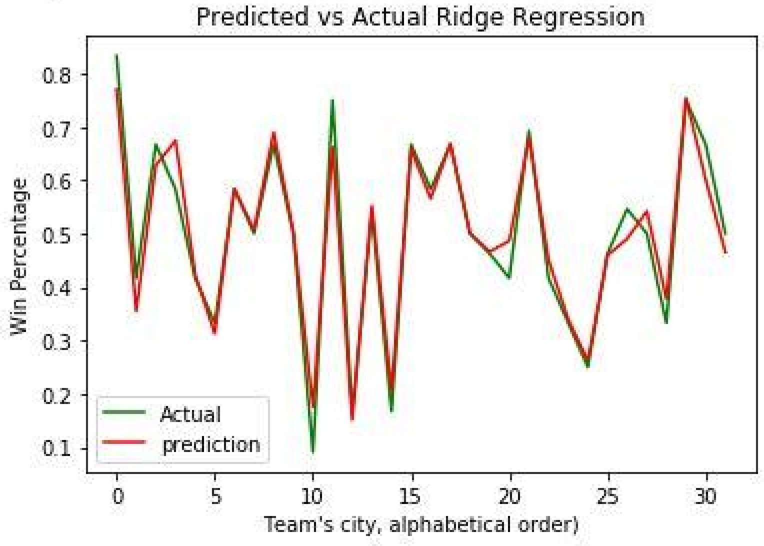
General knowledge of sports betting and creating outcome odds along with lines was sourced from multiple sites. The ambiguously named site Legit Gambling Sites provided a simple to follow guide of odds and lines conversion, along with a basic explanation of odds making. Just Into Data laid out a simple but effective description of betting line evolution, including various methods used currently and in the past.

Results:

I started with a standard linear regression model, and the accuracy score was disappointing at 53%. I thought maybe tweaking the model would improve the accuracy, but no matter what hyperparameter adjustments I made the accuracy remained low. It's likely the model wasn't applying appropriate weights (such as the Team A / Team B example above) as an explanation, or maybe I'm terrible at making prediction models. Either way I decided to follow the standard linear regression model with a ridge regression model. It occurred to me during this transition that there may be an angle I'm not considering here, and that is betting line deception. More on that later.

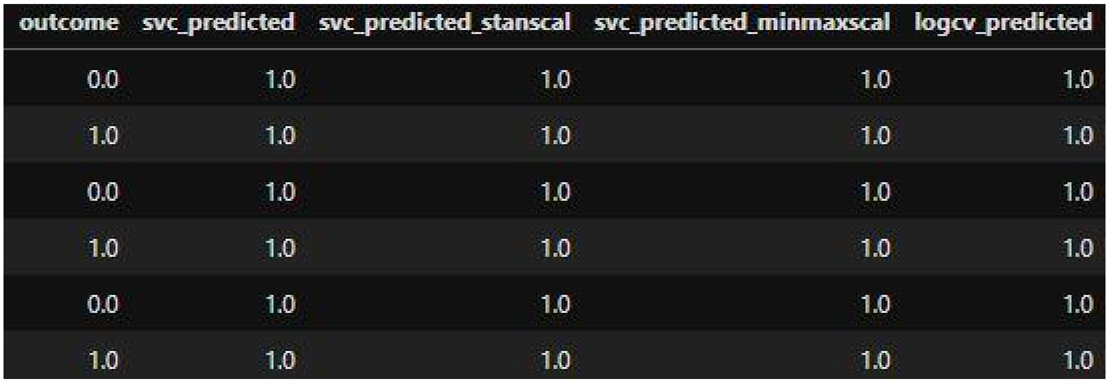


The weighted model proved to be substantially more accurate when it came to scoring with a 94%, and I began considering these results as the basis for developing outcome odds and betting lines. Was this the right way to go? Something about that initial regression model wasn't sitting right with me, but I needed to continue on with the project. With some confidence restored after the (almost too) high accuracy score from the ridge regression it was a good time to focus some attention on game outcome prediction without the lines.



Predicting win / loss outcomes was quite the roller coaster ride. I went into the model build thinking the accuracy would be lower than what is normally considered good; maybe in the 70 - 75% range due to game upsets that occur rather regularly in the NFL. The first model was a linear SVC model, which returned an outcome prediction of only five out of 22 teams winning. Satisfyingly enough all five of those teams did in fact win their previous game. 22 teams on a given football day of course means there should be 11 winners, barring a tie which is very rare. I began to tinker with hyperparameters, and increased the accuracy score to 83%. This was higher than anticipated, which made me very excited. It wasn't until I attached the predictions to the dataframe for comparison that I realized something was very wrong here. Accuracy of the test data was high, but the real data run returned either all wins or all losses. This obviously can't happen, and it would also mean an accuracy of 50% if half of the teams either win or lose.

I moved on to a logistic regression CV to see if I can get a better idea of what was going on. While reading the documentation for logistic regression CV I noticed that the 'solver' parameter could be adjusted, and it even recommended an algorithm for smaller datasets such as the one I was working with. I set the solver to the recommended 'liblinear' algorithm, and the result was an abysmal 16% accuracy score. I returned the parameters to default, and the accuracy score soared up to, you guessed it, 83%. I was naively hopeful this time the predictions wouldn't be all win or lose, but that was unfortunately the case after I compared the predictions with the real data.



outcome	svc_predicted	svc_predicted_stanscal	svc_predicted_minmaxscal	logcv_predicted
0.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0
0.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0
0.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0

It was around this time that I was looking at my models trying to make sense of what exactly was going on. Then it clicked. The nagging feeling that there was something to that 53% accuracy of the linear regression model, even though it makes sense with the model not being weighted, and that possibly interesting pattern noticed on the Odds Portal website. I went back to Odds Portal, and looked at betting line results for a semi-random week. I say semi-random because it was a week that I knew my favorite team, the (coincidentally) Las Vegas Raiders, won their matchup. The interesting pattern that was noticed was a nice mix of highlighted, meaning winner, cells with both plus and minus signs in them. As mentioned before bookies set their lines based on probability of an outcome occurring. The more likely an outcome is to happen the smaller the payout relative to your wager (minus money), and the less likely an outcome is to happen the larger the payout relative to your wager (plus money). A quick manual count of winners categorized by favorites (minus money) or underdogs (plus money) revealed an unexpected split - nearly 50 / 50. I repeated the count for other weeks, and the near 50 / 50 split was reoccurring.

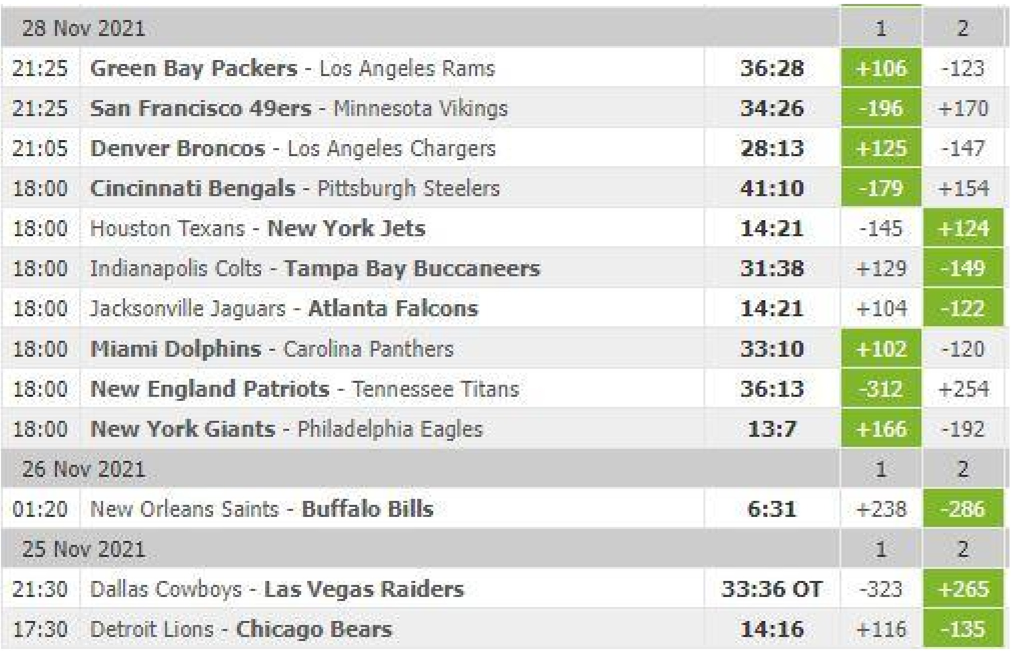
Findings:

What was going on here? How could the expert Las Vegas oddsmakers be roughly 50% accurate on who the favorite is in a NFL matchup? I decided to do some more research to find examples or any type of insight into how oddsmakers actually make their lines instead of trying to decipher it myself, and the old adage quickly came to mind - “The house always wins”.

The idea of ‘betting line deception’ came back around, and I started to wonder if oddsmakers alter their real projections before revealing them to the public as their betting lines. But how and why, when so many of their lines are very accurate? When I thought about ‘the house always winning’ I figured it was because casino games are set up such that your odds of winning are always significantly less than 50%. That being the case, on a long enough timeline, you’d eventually lose all your gambling money. Perhaps Las Vegas oddsmakers found a way to implement that strategy into sports betting.

Here’s how they would do it. Let’s suppose there are five matchups, and oddsmakers calculate their projections for all games. For this example we’ll say they project three teams with a general population consensus as favorites to win their matchup. They set those lines appropriately so the general population favorite is the actual favorite to win (minus money odds) and the underdog is projected to lose (plus money odds). The next matchup is projected to be an upset with the general population consensus favorite projected to lose. Instead of setting a line so the consensus favorite is projected to lose they set the line as if they’re projected to win. The odds are set fairly close to suggest the favorites won’t win by much, but will win. This mismatched pricing, or deception, is set so bettors feel even more confident that the consensus favorites will win, making them more likely to place a bet believing they’re getting great odds on the bet. The last matchup is projected to be very close, so which team will be the favorite in this instance? Both. The lines are set so both teams are technically favorites, meaning both bets will be minus money.

After all games are played and the finals are as they actually projected, what’s the financial outcome? The first three lines are right on, giving credibility to their system while collecting the money of those rooting for the underdog plus the ‘tax’ or ‘rake’ of a minus money bet for winning wagers. The house wins. The fourth matchup is an upset, and they collect on the (likely high volume) bets that were for the betting line favorite. The amount of bets for the underdog are so low that even at plus money they can pay those bets out with the losing bets collected and still make a profit. The house wins. The fifth matchup was indeed a close match, but it doesn’t matter which team wins. Both betting lines were minus money, so they collected a ‘tax’ on all the winning bets on top of gaining profits from all the losing bets. The house wins. Tiago Filipe Mendes Neves sums it up perfectly in his work “A data mining approach to predict probabilities of football matches” by saying, “Their business relies on you constantly placing bets and by complying with their system you will lose money in the long term.”



28 Nov 2021			1	2
21:25	**Green Bay Packers** - Los Angeles Rams	36:28	+106	-123
21:25	**San Francisco 49ers** - Minnesota Vikings	34:26	-196	+170
21:05	**Denver Broncos** - Los Angeles Chargers	28:13	+125	-147
18:00	**Cincinnati Bengals** - Pittsburgh Steelers	41:10	-179	+154
18:00	Houston Texans - **New York Jets**	14:21	-145	+124
18:00	Indianapolis Colts - **Tampa Bay Buccaneers**	31:38	+129	-149
18:00	Jacksonville Jaguars - **Atlanta Falcons**	14:21	+104	-122
18:00	**Miami Dolphins** - Carolina Panthers	33:10	+102	-120
18:00	**New England Patriots** - Tennessee Titans	36:13	-312	+254
18:00	**New York Giants** - Philadelphia Eagles	13:7	+166	-192
26 Nov 2021			1	2
01:20	New Orleans Saints - **Buffalo Bills**	6:31	+238	-286
25 Nov 2021			1	2
21:30	Dallas Cowboys - **Las Vegas Raiders**	33:36 OT	-323	+265
17:30	Detroit Lions - **Chicago Bears**	14:16	+116	-135

The example I just gave is only one possible scenario, and the oddsmakers can mix and combine these and other possible scenarios to deceive the bettors into making bad bets. I realize this all may sound like a big conspiracy theory, but we also have to accept the truth that casinos and oddsmakers are in the business of making money.

Related and Future Work:

There were two studies I found that touched on topics discussed in this project report. The first was already mentioned in the work of Tiago Filipe Mendes Neves. Though his work is on international football, what we Americans call soccer, the betting concepts are the same. His report also discusses betting mathematics, methodologies, and a guide on how he developed his own probability models to determine odds which he compared with real lines to find mispriced bets.

The second was the work of Lisandro Kaunitz et al. in “Beating the bookies with their own numbers - and how the online sports betting market is rigged”. The works are very similar with two key differences. Neves’ work provides more in depth machine learning techniques and algorithms that he used to develop his models, while Kaunitz et al. focus more on the statistical side of betting. The latter work also gives support to the notion I have dubbed betting line deception.

As for my own prediction models, there are some refinements I would like to implement. The first being following the lead of Neves, and expanding my projection models to include not only more data but other types of models as well; such as decision trees and neural networks. For my future projection model(s) I would also like to develop a feature or line of features that encompass a deeper understanding of yardage data. I would refer to this as the quality of yards per game, and it would combine the data collected for all yardage types with game scripts. The purpose would be to determine which yards can be produced more consistently, and which are generated by fluke, broken plays or ‘irrelevant’ plays that occur during a blowout game; often referred to as ‘garbage time yards’.

Conclusion:

The final message I want to deliver is sportsbooks, and all gambling for that matter, is like every other business in a capitalist society. Its business is to make money. If it wasn’t, could the Las Vegas strip have survived in its current state for so long? Do casinos offer compensated rooms, meals, and event tickets simply because they are nice people? Of course not. They rely on people to gamble so much that they’re willing to front a cost in order to get people on their gaming floors. Again, this may sound like some sort of conspiracy to some people, but after working on this project I'm of the opinion that betting lines are purposefully misleading in order to generate profits.

With that being said, it is a bit upsetting to admit that my initial work on this project - attempting to re-produce betting lines using real data - was not the ideal strategy. Even attempting to classify matchup winners and losers alone was not a great strategy. The correct approach should be to gain expert level knowledge on a sport, then gather the correct data points and build the predictive models. Implementing something like the aforementioned quality of yards to a model could have major positive insights into a matchup, but defining 'quality yards' cannot be achieved without a solid understanding of the sport. The same applies for recognizing situations such as broken plays resulting in unsustainably high yardage or scoring totals, or a game narrative where a coach will push for a player to break a record or have a 'revenge game' against their former team. Once you can build quality models and have a firm grasp of a sport to correctly interpret the results, then you can accurately identify and exploit mispriced or deceiving betting lines.

References:

"2021 Standings and Team". Pro Football Reference.
https://www.pro-football-reference.com/years/2021/

"2021 Team Offensive Stats". The Football Database.
https://www.footballdb.com/stats/teamstat.html?group=O&cat=T&yr=2021&lg=NFL

"Calculating Betting Odds". Legit Gambling Sites.
https://www.legitgamblingsites.com/online-betting/calculating-odds/

Devadiga, Thilakraj. "Linear Regression". BlobCity. Linear Regression chart code.
https://cloud.blobcity.com/code/explore/Regression/Linear%20Models/LinearRegression

Kaunitz, L., Zhong, S., Kreiner, J. "Beating the bookies with their own numbers - and how the online sports betting market is rigged." Research Center for Advanced Science and Technology, The University of Tokyo.
https://arxiv.org/ftp/arxiv/papers/1710/1710.02824.pdf

Lianne and Justin. “How to Improve Sports Betting Odds — Step by Step Guide in Python”. Just Into Data. February 28, 2021. https://www.justintodata.com/improve-sports-betting-odds-guide-in-python/

“Machine Learning in Python”. Scikit-learn. Multiple documentation pages. https://scikit-learn.org/stable/index.html

Mendes Neves, Tiago Filipe. “A data mining approach to predict probabilities of football matches”. Faculdade De Engenharia, Universidade Do Porto. July 14, 2019. https://repositorio-aberto.up.pt/bitstream/10216/121217/2/343145.pdf

“NFL Results and Historical Odds”. Odds Portal. https://www.oddsportal.com/american-football/usa/nfl/results/

“NFL Team Win Trends - All Games, 2021”. Team Rankings. https://www.teamrankings.com/nfl/trends/win_trends/