


Predicting NFL Point Spreads via Machine Learning

Daniel Brandon, Christian Brothers University, USA*

 <https://orcid.org/0000-0002-7502-7975>

ABSTRACT

This paper describes sports quantitative analysis research which investigates the use of statistics and modern machine learning methods applied to the problem of predicting the point spreads for United States (US) National Football League (NFL) football games. Insights and results are presented for several modern machine learning techniques for both exploratory analysis and predictive analysis. The case study presented here and the results thereof may be quite useful for those involved in the huge global sports betting arena both the gaming industry and the bettors therein. NFL game statistics also provides a rich source of relevant real-world data for the deployment of several modern data science methodologies and is thus a great teaching tool for the university classroom. Since sports gambling has now made its way onto college campuses with a growing number of schools signing million dollar deals with sports books and casinos, the topic of this article is of even more current relevance.

KEYWORDS

Data Science, Football Prediction, Machine Learning, NFL, Point Spread, Sports Statistics

Sports betting has been around for as long as sports have been played, and accurately predicting results is of vital importance for both those who place bets and those who take bets (sports books). Historically, predicting sports results was based on qualitative methods, including human judgments, intuition, feelings, inside information, etc. But in today's world, sophisticated quantitative data-analytics methods are used by most sports books and also many people or organizations who make bets. However, sports books do have ways to identify and discourage individuals that use sophisticated data analytics, people they label as "algorithmic bettors" (The Economist, 2022).

This paper describes research using real-world data to investigate the use of statistical and modern machine-learning methods applied to the problem of predicting the point spreads for United States (US) National Football League (NFL) football games. For this article we presume that the reader has a basic familiarity with statistics and data-analytics methods and tools and in particular predictive analysis and machine-learning technology.

DOI: 10.4018/IJDA.342851

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

This paper will attempt to answer several key research questions:

- Do traditional prediction tools such as multiple regression provide the accuracy needed to beat the official NFL games odds?
- Are some modern machine-learning methods better than others for NFL point-spread prediction?
- Are modern machine-learning methods able to beat the official NFL game odds when the standard 10% vigorish is included?
- Can today's best modern machine-learning methods produce competitive results for NFL point-spread prediction when the standard 10% vigorish is included and the early line is used?

The significance of this research goes beyond the obvious interest of data scientists, bettors, and sport books now that sports gambling is growing so rapidly around the world. The societal costs of gambling addiction are quickly growing. One study noted that most US high-school students are now betting online, and about 5% of high-school students have a gambling problem (The Kiplinger Letter, 2023). Online betting has now made its way onto college campuses, with a growing number of schools signing million-dollar deals with sports books and casinos. There are growing ethical and social issues with strong student exposure to sports gambling on their campus and sponsored by their university. Student education into the difficulty of winning bets and the high likelihood of growing losses is needed.

BACKGROUND

In 2018, the US Supreme Court allowed widespread sports gambling beyond the state of Nevada. The Supreme Court ruled that each state should have the right to regulate sports betting. This led to a gold rush of states and companies into this huge lucrative market. Most US states now permit some types of sports betting. Some states are in-person only, but most now also allow online betting. As a result, US sports betting has exploded in recent years, as shown in Fig. 1 (The Economist, 2022).

Online sports betting is eclipsing in-person betting. For the 12 months ended December 2021, DraftKings reported \$1.296 billion in revenue; it was more than double the \$614 million it generated in 2020 (Tatevosian, 2022). States in the US are benefitting on the order of hundreds of millions of dollars each year from sports-betting taxes. Football is the most popular US sport for betting, and the NFL earns billions of dollars indirectly from sports betting (AGA, 2018). According to Macquarie Research (Schafer, 2022), about \$1 billion of bets are placed weekly during the NFL season. The

Figure 1. US Sports Betting (Billions of Dollars) (Source: Image from The Economist (2022))

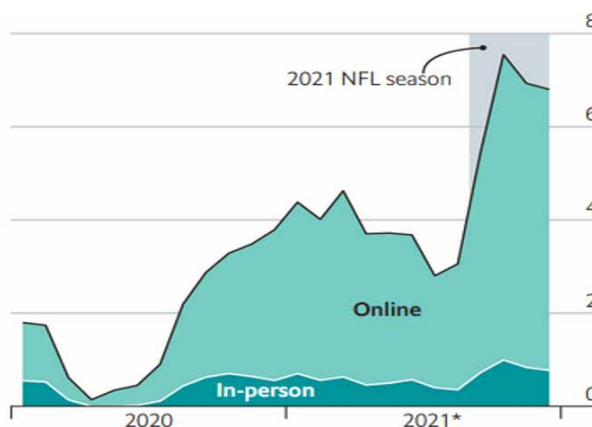


Table 1. Odds Display for One Game

Bet #	Team	Line	Over/Under	Money Line
134	NO Saints		40	+160
135	ATL Falcons	-6		-180

final quarter of the year during which most NFL games take place accounts for about 35% of yearly revenue for gambling companies. FanDuel currently leads the online market with about 50% market share in states where it’s operational, but several competitors are close behind, including DraftKings, Caesars Entertainment, and BetMGM.

Today, online sports gambling has even made its way onto college campuses. For example, as reported by *The New York Times*, Caesars Sportsbook struck a huge \$8.4 million deal with Michigan State University last year (Louis, 2023). As further reported by *The New York Times*, and what is considered very troubling to some, is that so far eight universities have partnered with online sports books and about another twelve athletic departments and booster clubs have signed agreements with traditional casinos. As stated by Dave Ramsey on *The Ramsey Show*, universities are “selling out your own students who you’re supposed to be caring for.” Gambling is the second largest addiction in North America today, and this rapidly growing addiction can start with sports betting as a gateway (Louis, 2023).

There are many types of NFL team bets, and one can bet in person (such as a physical casino) or online. The most common and simple team bets are straight bets, including point spread, over/under, and money line. Other, more complicated bets include parlays, prop bets, teasers, and futures bets. One can also bet on players and other NFL-related matters. An odds display on your smartphone, on your computer screen, or in some type of print media would appear similar to that shown in Table 1; the bottom team is the home team, and the top team is the away or visiting team.

The point spread is next to the team that is favored; here the ATL (Atlanta) Falcons are favored by 6 points. Point spreads vary during the week, by sports books, by location, and for other reasons. The favored team’s expected margin of victory is shown as a negative number. Sometimes a positive number (+6) would also be shown next to the expected losing team (underdog). Since ATL is a 6-point favorite here, it has to win by more than 6 points to cover a spread bet on it (Bet #135). If NO (New Orleans) wins the game or if NO loses by fewer than 6 points, then Bet #134 would be a winning bet. If NO loses by 6 points (ATL wins by 6 points), then either bet (#134 or #135) is a tie (push) and money is returned. To avoid ties, bookmakers will often add a hook of 0.5 to the point spread. To make a point-spread bet, one enters (or tells the attendant) the bet number, selects “point spread,” and enters the amount of the wager. If one wins a bet, the house (sports book or casino) will take a cut of your profit called the vigorish (also called the vig or juice). For point-spread betting, this is often shown next to the odds as:

ATL -6 (-110)
 NO 6 (-110)

110 indicates that you pay \$110 for a \$100 win (or about 10%). The vig has historically and typically been 10%, but it can vary by sports book, sport, bet type, and individual game. The vig is also adjusted when betting is moving heavily toward one team—both adjusting vig and adjusting point spreads are used to even out the betting on both sides. With the traditional 10% vig, a bettor would need to win 52.381% of their bets to break even.

With an over/under (total) bet, one bets that the total points scored by both teams will be more or less than the total shown in the over/under column. Here one can enter bet number, whether one

wants over or under, and the amount of the wager. With a money-line bet (or “to win”), one is betting which team will win the game. In the example here, ATL odds are -180, which means that a \$180 dollar bet on ATL would win \$100 for a return of \$280 if ATL won. A \$100 bet on NO to win would return \$260 if they did win (\$160 profit). Here one enters the bet number, selects “money line,” and enters the amount of the wager. However, accurately predicting the point spread is key to winning any bet type.

In addition to the US, there are currently about 20 countries around the world offering legal sports betting. The magnitude of international sports betting is enormous but hard to estimate due to vague and inconsistent global regulations and accounting practices. Estimates place global sports betting at about \$200 billion in 2021, with licensed online sports books accounting for roughly \$40 billion of that revenue (Statista, 2022). Offshore online sports books such as BetOnline, SportsBetting, and Legal Sports Betting have widespread global usage. In addition to the legally permitted US states, DraftKings currently operates in about seven other countries, including the UK and Canada (excluding Ontario). Note that the way odds are displayed varies around the world. Many countries do not use the “American” or money-line style but express odds in either decimal or fractional formats such as 10/1 (10 to 1). Most countries have income tax on sports-gambling wins, as does the US, but some do not, such as Canada. Thus, in determining gambling wins or losses, one also has to include taxes as well as the vigorish.

PREVIOUS RESEARCH

US football started about 1872 for college football and 1920 for the NFL. Standardized and formal statistics for football started in the late 1930s (Brown, 2021). Using computers and finding ways to predict football-game outcomes started in the 1960s (Lyons, 2020). In those early days of football digital analytics, linear regression and other linear methods would typically be programmed in FORTRAN. ASA-SIAM’s *Anthology of Statistics in Sports* (ASA-ISAM, 2005) noted many of these historical aspects, and Harville’s (1980) article in the *Journal of the American Statistical Association* is a classic example. Today many very sophisticated statistical and machine-learning methods are available, typically programmed in more powerful programming languages such as R and Python.

Past NFL research has largely concerned predicting game winners, game point spread, and wins against the spread. Bosch (2018) compared neural networks against classical machine-learning methods in predicting game winners. He used data from the 2009 to 2016 seasons and utilized mostly team stats but also average player age, weight, and height. Three types of neural networks were included in his study: traditional ANN (artificial neural networks), LSTM (long short-term memory), and RNN (recurrent neural network). The LSTM neural network gave the greatest accuracy of the neural net methods at 63.1%, but logistic regression was slightly better at 63.33% accuracy, as was support vector machines at 63.25% accuracy; random forests gave 62.26% accuracy.

Anyama and Igiri (2015) claimed 90.32% accuracy in predicting NFL game winners for 31 games in one season (2003) using machine learning by including independent variables for the Vegas spread and players’ performance ratings. Beal et al. (2020) compared several machine-learning methods for predicting the winning team for NFL games. They compared nine methods for 1,280 games over five seasons using 42 independent variables that were all team statistics for the current season plus an average for the past season. The results are shown in Table 2, with the best methods being naïve Bayes at 67.53%, followed by AdaBoost (a decision-forests method) at 66.35%; note that the Vegas game prediction closing line accuracy for that time period was 65.8%.

Much of the past research work for NFL point-spread bets involves classification: beating the spread or not (picking the winning bet given the point spread). Most of the point-spread research work involves using the closing Vegas spread as an independent variable to predict the actual game spreads and bets against the spread. Gimpel (2018) performed a study using NFL data for the 1992 through 1999 seasons with only team predictor variables plus Vegas point spread-based variables.

Table 2. NFL Winning Team Prediction Accuracy

Method	Accuracy
Support Vectors	0.5537
Nearest Neighbors	0.5748
Gaussian Process	0.4464
Decision Tree	0.6352
Random Forests	0.6341
AdaBoost	0.6635
Naïve Bayes	0.6753
QDA	0.5451
Neural Network	0.6071

He used both logistic regression and support vector machines to classify winning and losing against the spread and found the support vector machine was better for the 2000 season prediction (52.04% accurate) and logistic regression was better for the 2001 season (56.97% accurate). Wadsworth (2016) did an analysis with dependent variables of win streak, NFL ranking from previous season, current-season power ranking, and the Vegas closing spread. She used several machine-learning methods and found that the AdaBoosted forests was best and beat Vegas lines for spread bets 55% to 60% of the time, as shown in Fig. 2.

Szalkowski and Nelson (2012) used both the opening and closing Vegas lines to achieve 75% accuracy for NFL divisional winners. Seal (2018) reported that his MathBox decision-tree machine-learning algorithm had predicted results with accuracy very close to the Vegas closing accuracy using just early line information. Warner (2010) did a study using Gaussian-process machine learning based upon team data from the NFL 2000 through 2007 seasons as training data for 2008 and 2009 predictions. He used team statistics as independent variables but also included team computed strength and temperature difference between the two teams' cities. He achieved a win rate of 50.90% against the spread bets, which was below the threshold of 52.4% needed to come out ahead with a 10% vigorish. He stated: "In the end, this study confirms what many already know: the Vegas line-makers are very good at what they do." Currently, Lineups.com tracks their NFL prediction results for its preferred methods, random forests and logistic regression (the two categories are covering the spread or not). For 2017 (the year of our case study), their models on average won 52% of their point-spread bets (Lineups.com, 2017).

Figure 2. AdaBoosted Forests Versus Vegas Spread (Source: Image from Wadsworth (2016))

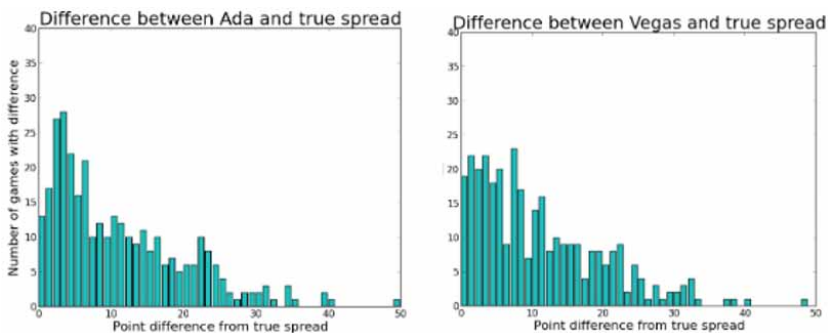
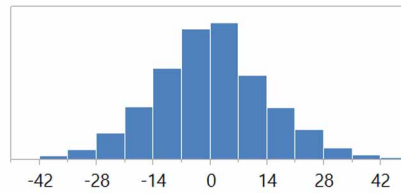


Figure 3. Vegas Spread Accuracy Distribution (Source: Image from RotoDoc (2016))



VEGAS LINES

Historically, sports odds and point spreads were initially set by individuals such as the notorious “Jimmy the Greek,” then by a consensus of experts. But today most sports odds are no longer set by humans. Casinos and online books employ data scientists and use sophisticated computer tools to set opening lines. While the total points scored by both teams have increased over the years as today’s high-powered passing offenses generate more points than the conservative running games of early NFL years, the average actual season point spread has remained between 10 and 12 for the last 50 years. As a result, the point spread as a percentage of total points continues to decline. In addition, the average of the point spreads for all games in a season, the home-field advantage, has decreased over the years from about 3 in the 1960s to about 2.5 in the 1980s and about 2 in recent seasons.

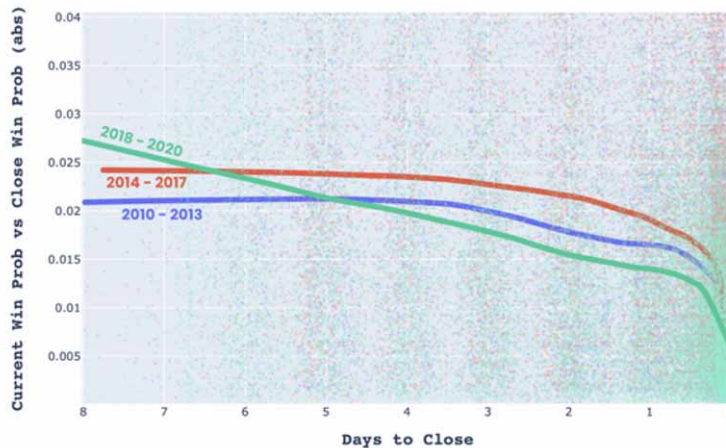
NFL point spreads set by the sports books generally come out on Tuesday for games later that week; these are the opening line or early line. Point spreads change during the week and are adjusted for a number of reasons such as the amount bet on each team and updated information such as player injuries. The final odds just before the start of the game are the closing line or public line. Therefore, the closing spreads are somewhat a function of public opinion. The point spreads of some sports books also vary by location to compensate for more money being bet on regional (home) team favorites.

Traditional sports books try to adjust the odds so that an even amount of money is bet on each team; this minimizes their risk. The odds that are in effect for one’s bet are the ones in place at the moment of the bet. Several studies have considered the accuracy of the closing Vegas line and whether the process of moving from the opening line to the closing line represents an efficient market hypothesis, whereby the price of a product over time is an unbiased predictor of the product’s value (Fama, 1970). Spinosa (2014) studied 21 NFL regular seasons from 1992 to 2012 and found through statistical analysis that the NFL Vegas closing line is efficient in the aggregate sense and does reflect the public’s betting preferences.

NFL Pickles (2007) did a study and found that from 1992 through 2006 the Vegas closing point spread had accuracy (standard deviation) ranging from 12.284 to 14.412 with a mean ranging from -1.29 to +1.96. In another study, RotoDoc (2016) examined the Vegas closing spread versus the actual spread for the seasons between 2005 and 2015 and found the variation was close to a normal distribution with a median difference of near zero and standard deviation of 13.7 points; see Fig. 3. Note that some studies use standard deviation and some use root mean square error (RMSE), but since the mean is about zero, we can directly compare the two measures. Boyd (2017) reported that the Vegas closing-line error does not vary much by the magnitude of the scores.

Greer (2021) did a study with NFL data for the 2007 to 2020 seasons and found Vegas closing spreads predicted the correct winner 65.9% of the time, while Vegas opening spreads predicted the correct winner only 63.5% of the time. Much of the change in the spread is at the end of the week closer to game time; see Fig. 4. In addition, for recent years the closing line was more accurate than past years and there was more difference between opening and closing lines. Therefore, having an accurate predictive tool that uses only data available early in the week is quite advantageous.

Figure 4. Vegas Lines Versus Days Until Game (Source: Image from Greer (2021))



CASE STUDY

As previous research has shown, it's very hard to beat the Vegas closing line given a 10% vigorish even with today's sophisticated machine-learning algorithms. There may be more betting opportunities early in the week against the opening Vegas line, so this study uses just team data available early in the week before injury information and consensus betting becomes available. In addition, this case study does not use either the opening or the closing Vegas line as a predictor variable. Only current-season team statistics were used, and other information such as player stats, the Vegas lines, and derivatives such as power ratings was not used as a predictor. Some other studies use other data such as individual player stats, coaching stats, city/stadium stats, closing Vegas lines, and even Twitter tweets. NFL team stats are available early in the week (Tuesday) at many sites such as pro-football-reference.com and NFL.com. Our stats were downloaded from NFL.com into Microsoft Excel; the number of team stats available on NFL.com has increased throughout the years.

The data gathering and numerical work were done in 2018 and 2019 for the 2017 NFL season, which was a typical NFL season (before COVID). The open-source R programming language and RStudio (now Posit) were used for all machine-learning methods. COVID did delay the completion of our analysis and writeup until now. For the 2017 regular-season games, the favorite won the game 179 times with 74 losses, and the favorite covered the spread (won a point-spread bet) 132 times (53.9%) with 113 spread losses and 8 ties (Sports Odds History, 2018). In 2017, NFL.com provided 19 offensive-team variables and 17 defensive-team variables. Not all the available NFL.com team stats were used. Our dependent variable is the actual game point spread, the difference between the home-team points scored and the visiting-team points scored. Twelve independent variables were used. For offensive, they were points per game, yards per game, third-down conversion percentage, yards penalized, fumbles, fumbles lost, net turnovers, and time of possession. For defensive, they were points allowed per game, yards allowed per game, third downs allowed percentage, and fumbles caused. The spreadsheet in Fig. 5 shows a partial view (2 of 17 weeks) for the NFL 2017 regular season.

EXPLORATORY DATA ANALYSIS

Pearson's r values were calculated for all pairs of variables in Excel as shown in Fig. 6. The strongest correlations for the point spread-dependent variable were for points/game, points allowed per game,

Figure 5. NFL Regular-Season 2017 Dataset

NFL 2017 Games by Week																	
Week #	Visiting Team	Home Team	Visiting Points	Home Points	Offensive Ind Variables				Defensive Ind Variables								
					Spread (H-V)	Points.G	Yds.G	3rdPct	Pen.Yds	Fumble	F.Lost	TO	ToFP	PA.G	YA.G	3rdPctAll	F.Caused
1	Arizona Cardinals	Detroit Lions	23	35	12	7.1	293.7	4	88	2	4	14	1	0.9	44.3	2	0
1	Atlanta Falcons	Chicago Bears	23	17	-6	-5.6	-77.4	-10	25	7	4	2	-1	0.3	0.7	1	0
1	Baltimore Ravens	Cincinnati Bengals	20	0	-20	-6.6	-24.9	0	278	3	7	26	-4	2.9	14	4	-15
1	Carolina Panthers	San Francisco 49ers	23	3	-20	-3	25.5	-3	306	-5	2	-2	-3	1.8	34.3	3	6
1	Indianapolis Colts	Los Angeles Rams	9	46	37	13.5	76.9	3	143	9	8	2	0	-4.6	-27.6	-7	0
1	Jacksonville Jaguars	Houston Texans	29	7	-22	-5	-45.9	-1	114	1	1	-22	-2	10.4	60.5	1	-11
1	Kansas City Chiefs	New England Patriots	42	27	-15	2.7	18.8	2	-209	-1	1	-9	0	-2.7	0.8	-1	0
1	Los Angeles Chargers	Denver Broncos	21	24	3	-4.1	-53.5	-1	-42	9	8	-29	1	6.9	-38.4	-7	-8
1	New Orleans Saints	Minnesota Vikings	19	29	10	-4.1	-34.3	6	-13	-8	-4	-2	1	-4.6	-60.6	-16	-4
1	New York Giants	Dallas Cowboys	3	19	16	6.7	17.7	10	44	-9	1	2	1	-3.4	-55.1	3	8
1	New York Jets	Buffalo Bills	12	21	9	0.3	-2.6	6	183	-6	-6	13	-2	-1.5	2.9	1	6
1	Oakland Raiders	Tennessee Titans	26	16	-10	2.1	-10.1	-5	-224	-23	-6	10	1	-1.1	-22.1	-5	3
1	Philadelphia Eagles	Washington Redskins	30	17	-13	-7.2	-40.9	-10	-230	3	15	-3	5.8	41.4	5	2	
1	Pittsburgh Steelers	Cleveland Browns	21	18	-3	-10.8	-49	-10	28	10	8	-30	-4	6.4	21.2	5	2
1	Seattle Seahawks	Green Bay Packers	9	17	8	-2.9	-24.7	2	-553	-4	2	-11	0	3.2	25.7	5	-2
1	Arizona Cardinals	Indianapolis Colts	16	13	-3	-2	-29.5	3	-101	1	-1	9	-1	2.6	56.2	9	3
1	Buffalo Bills	Carolina Panthers	3	9	6	3.8	21.1	0	-190	2	0	-10	4	-2	-38	0	-6
1	Chicago Bears	Tampa Bay Buccaneers	7	29	22	4.6	76.1	6	-148	3	3	-1	1	3.9	59	9	7
1	Cleveland Browns	Baltimore Ravens	10	24	14	10.1	-3.5	0	-175	-6	-9	45	3	-6.7	-3	-4	1
1	Dallas Cowboys	Denver Broncos	17	42	25	-4	-7.8	-4	-54	15	3	-16	1	3.1	-28.1	-11	-13
1	Detroit Lions	New York Giants	24	10	-14	-10.2	-23.6	-6	-73	-1	-3	-13	-1	0.7	17.4	2	-2
1	Green Bay Packers	Atlanta Falcons	23	34	11	2.1	59.1	6	129	3	-1	1	1	-4.3	-30.5	-5	-1
1	Houston Texans	Cincinnati Bengals	13	9	-4	-3	-39.5	-2	-23	-3	0	3	-3	-5.4	-7.5	6	-7
1	Miami Dolphins	Los Angeles Chargers	19	17	-2	4.6	68.9	8	-227	-6	-4	26	1	-7.6	-7.3	2	0
1	Minnesota Vikings	Pittsburgh Steelers	9	26	17	1.5	21	0	-53	4	-1	-3	0	1.4	31	11	3
1	New England Patriots	New Orleans Saints	36	20	-16	-0.6	-3	-3	127	6	6	1	0	1.9	-29.5	2	1
1	New York Jets	Oakland Raiders	20	45	25	0.2	18.9	4	-26	8	2	-10	-1	-0.6	-2.1	4	1
1	Philadelphia Eagles	Kansas City Chiefs	20	27	7	-2.7	9.6	-3	82	-11	-6	4	-2	3.8	58.6	8	0
1	San Francisco 49ers	Seattle Seahawks	9	12	3	2.2	-18.8	-2	354	4	-3	11	0	-3.1	-28.4	-5	-3
1	Tennessee Titans	Jacksonville Jaguars	37	16	-21	5.2	51.9	2	151	14	2	14	2	-5.4	-41.9	-2	2
1	Washington Redskins	Los Angeles Rams	27	20	-7	8.5	36.6	9	190	1	0	11	0	-3.6	-8.4	1	2

Figure 6. Pearson's r Correlation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Spread (H-V)	Points.G	Yds.G	3rdPct	Pen.Yds	Fumble	F.Lost	TO	ToFP	PA.G	YA.G	3rdPctAll	F.Caused	
2	Spread (H-V)	1												
3	Points.G	0.532227247	1											
4	Yds.G	0.446087329	0.785694	1										
5	3rdPct	0.297832703	0.469658	0.570287	1									
6	Pen.Yds	-0.006581847	-0.00379	0.049706	-0.10655	1								
7	Fumble	-0.15298681	-0.17553	-0.2302	-0.27652	0.092791	1							
8	F.Lost	-0.169431908	-0.17456	-0.19634	-0.20391	0.045398	0.793002	1						
9	TO	0.404427615	0.672681	0.430589	0.313368	-0.07902	-0.37158	-0.54706	1					
10	ToFP	0.377369865	0.628186	0.570432	0.412288	-0.26548	-0.16773	-0.24987	0.438178	1				
11	PA.G	-0.482753213	-0.62386	-0.56789	-0.43882	0.019852	0.376084	0.465923	-0.65085	-0.55513	1			
12	YA.G	-0.207667055	-0.1716	-0.17374	-0.22853	-0.08547	0.109145	0.089573	0.023755	-0.51712	0.563397	1		
13	3rdPctAll	-0.203036632	-0.25944	-0.15134	-0.08438	-0.11837	0.074903	0.091951	-0.08688	-0.5386	0.42835	0.738165	1	
14	F.Caused	0.066032763	0.134899	0.173286	0.200511	-0.22385	-0.13923	-0.09347	0.312064	0.046825	-0.07185	0.163743	0.355619976	1

yards/game, and net turnovers. There were also strong correlations between independent variables such as between yards per game and points per game or between fumbles and fumbles lost.

For a visual analysis, one can perform a scatter plot of paired variables such as the spread (home points minus visitor points) versus Points.G (average points per game for the home team minus average points per game for the visiting team) as shown in Fig. 7. We see wide dispersion about an imaginary midline.

There are many methods and algorithms for machine learning, as illustrated in Fig. 8, which shows the major methods. These are typically divided into supervised and unsupervised learning, but some methods can be used for either case. Classification determines categories typically for nominal and ordinal data types, and regression is for numeric data types. In our study, several popular methods were utilized both for exploratory data analysis and later for point-spread prediction.

K-nearest neighbors (KNN) is a supervised method in which one tries to classify data points into specific categories (Cover & Hart, 1967). KNN uses data from training cases to classify test cases based on a similarity. For classification, test points get classified in a certain category on the basis of voting from nearest neighbors, and for regression test, data get classified based on the averages of nearest neighbors; default metric for distance is typically Euclidean for continuous variables. The

Figure 7. Spread Versus Points Scored/Game Differential

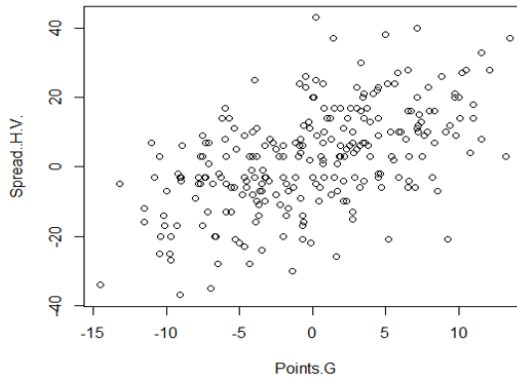
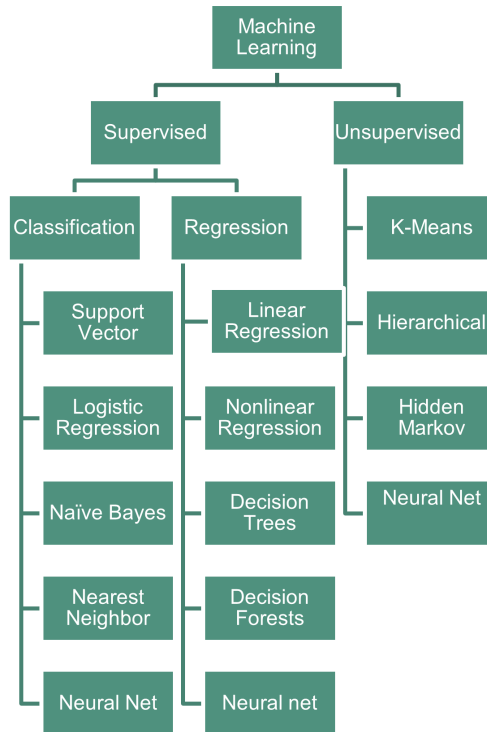
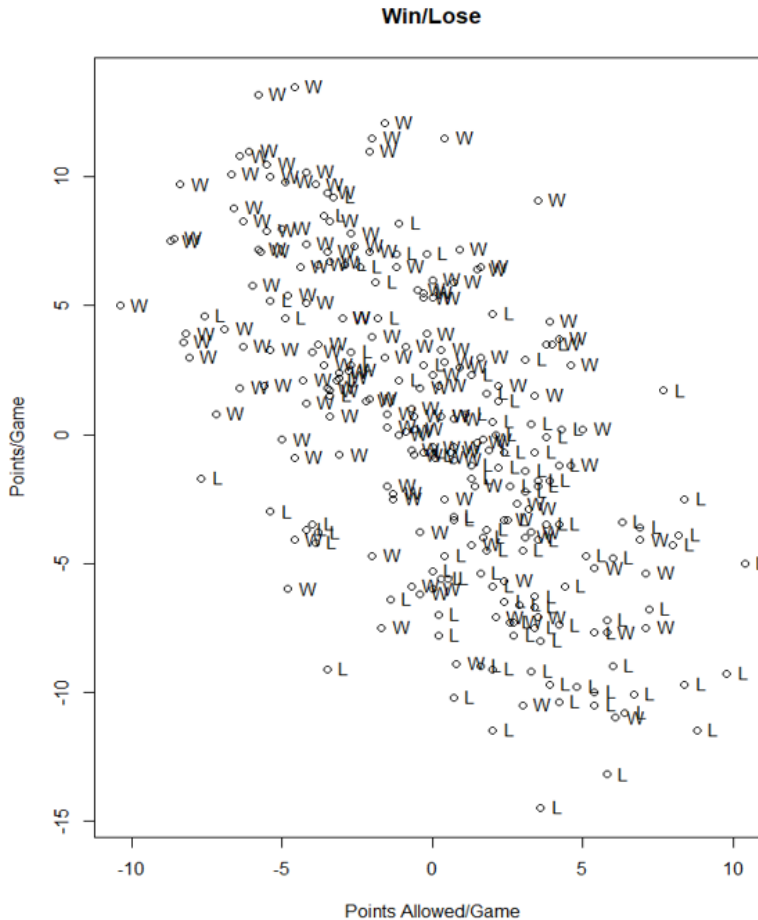


Figure 8. Machine-Learning Methods



choice of the parameter K is important; usually K gets initially set to the square root of the number of data points. A larger K value might reduce the variance with noisy data; however, a bias may be introduced since smaller patterns may have useful insights. It is customary to normalize all the variables to the same scale (i.e., 0 to 1), or else more weight would be given to the higher value data. A win/loss calculated column was added to the NFL regular-season dataset shown previously. A win/loss scatter plot of points per game versus points allowed per game is shown in Fig. 9. One would expect most losses near the bottom right and expect most wins at the top left. However, most of the area includes overlapping wins and losses, indicating the complexity of the decision surfaces.

Figure 9. Win/Loss Scatter Diagram

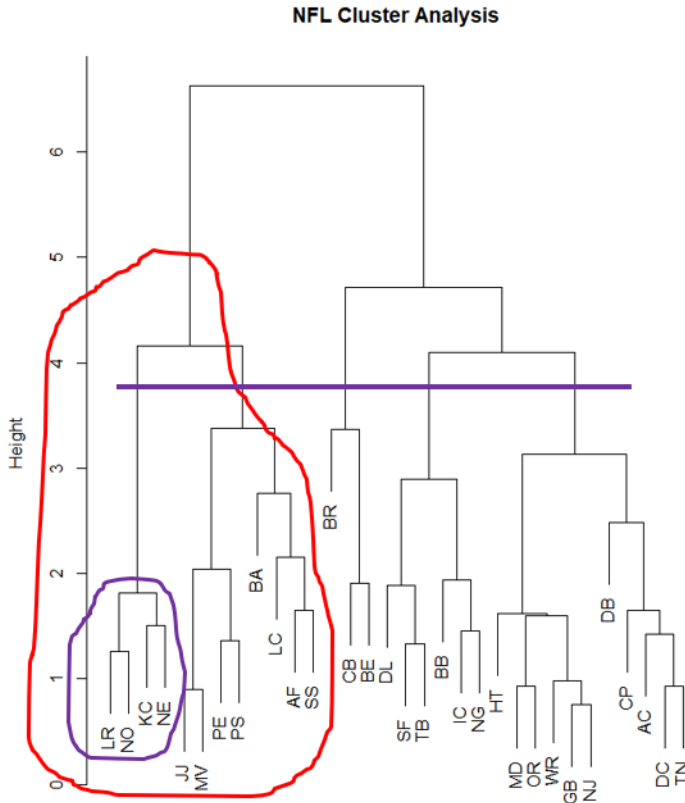


Next, a KNN analysis was performed with two features (points scored and points allowed) in R. The correct classification (win or loss) was obtained 69% of the time (14 out of 22 losses correct and 22 out of 30 wins correct). When more features (independent variables) were included, the results were less accurate.

One can also do a similar exploratory analysis with support vector machines (SVMs), which is a supervised method that analyzes data and is used primarily for classification but also regression (Suykens and Vandewalle, 1999). A SVM uses training data to fit hyperplanes in the independent variables space to create decision boundaries between categories. The hyperplane is placed to maximize the gap between the categories. However, instead of using the Euclidian notion of distance, a kernel trick is used to allow for nonlinear hyperplane surfaces. Using the same NFL data as for the previous KNN analysis and separating the data sets into a win set and a loss set, we found the accuracy was 69% the same as with KNN.

Cluster analysis is an unsupervised technique for classifying objects into groups. It develops a model that can be used to classify a case into a group based on the values of its variables. K-means and hierarchical are popular types of clustering analysis. A hierarchical clustering was performed to find the better teams in the NFL 2017 regular season using the following subset of stats: points per

Figure 10. Dendrogram of NFL Teams



game, yards per game, turnovers, time of possession per game, points allowed per game, and yards allowed per game.

The dendrogram is shown in Fig. 10; the better teams are shown in one of the two higher clusters, and the best teams are shown in the bottom left cluster (one of five clusters at that cut point). In comparison to the dendrogram, the actual NFL 2017 playoff teams are shown in Fig. 11.

Figure 11. NFL 2017 Playoff Teams

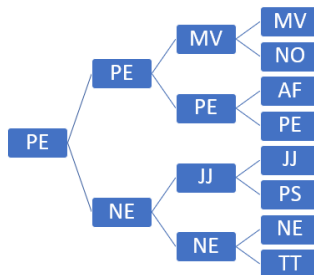
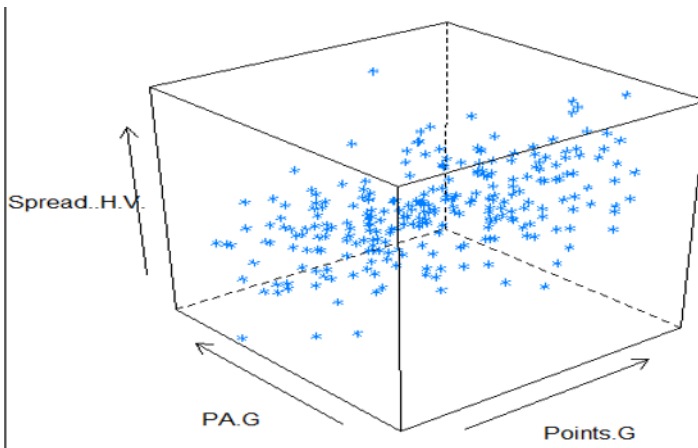


Figure 12. Cloud for Spread Versus Points/Game and Points Allowed/Game



PREDICTIVE ANALYSIS

One of the oldest forms of machine learning is simple linear regression, which is used to quantify the relationship between two numeric variables. The dependent variable (often called the result variable or y and shown on the vertical axis) and the independent variable (often called the decision variable or x and shown on the horizontal axis). In multiple linear regression (MLR), there is more than one independent (decision or predictor) variable. Stepwise MLR determines the most significant independent variables by adding and removing independent variables via measuring the adjusted r squared or AIC (Akaike information criterion) value. There are many forms of regression, and logistics regression is often used when the dependent variable is categorical, particularly when there are two categories (binary logistics regression). In sports betting, it is used to predict a win or loss situation for a single game, and it is also commonly used to predict whether a team will cover the spread for a single game with a particular given point spread. First, a standard MLR was performed using all our independent variables in R (lm function) for the NFL 2017 regular season. The adjusted r -squared was only .2963. The most significant predictors (lowest p -values) are points scored per game (Points.G) and points allowed per game (PA.G); the RMSE is 17.09.

Next a stepwise regression was done in R (stepAIC function). The best AIC was to use just points per game and points allowed per game. For visualization, a 3D scatter plot was obtained via the R cloud function, as shown in Fig. 12. Again, we see considerable variance about an imaginary mid-hyperplane.

Decision trees successively use questions or tests to divide training data until each division consists only of points from one category or one range. This is often called classification and regression trees. Trees are built (or grown) with questions using the independent variables, which leads to results or categories of the dependent variable at the bottom of the tree (Breiman et al., 1984). Random forests are a type of ensemble learning that generates multiple decision-tree models and then obtains a consensus. Random forests generally improve upon the accuracy of individual trees by creating a large number of smaller trees, a so-called decision forest (Breiman, 2001). A predicted outcome is obtained by compiling results for all of the trees (an average in regression, a majority vote in classification). A group of related methods called boosting methods utilizes a different method of ensemble formation. Random forests build an ensemble of deeper independent trees, but boosting builds an ensemble of shallower and weaker trees, with each tree improving on the previous (Freund & Schapire, 1999). Boosting usually outperforms random trees, and it can be employed for both classification and regression. Commonly used boosting algorithms today are AdaBoost (adaptive boosting), gradient

Table 3. Comparing Random Forests and MLR

Method	RMSE
AdaBoost (R gbm)	13.39
Gradient Boosting (R XGBoost)	13.29
MLR (R lm)	17.09

boosting, and XGBoost. In our case study, two popular decision-forest boosting algorithms in R were used. The first was the R gbm package, which is an implementation of extensions to Freund and Schapire’s (1996) AdaBoost. The other was the XGBoost (eXtreme Gradient Boosting) R package, which is an implementation of gradient boosting framework by Friedman (2001).

The results of random-forest methods versus MLR are shown in Table 3. From the results, one can see that the boosted-forests methods are much better than traditional multiple linear regression. For all methods the training data was 80% of the total dataset.

Neural networks have become very popular for many machine-learning applications, as this type of method has seen extensive improvements in recent years. They are at the forefront of AI research today. Neural nets (ANNs) emulate the brain process found in animal brains by having interconnected neurons (nodes) that pass weighted signals through the network. Can we improve upon our NFL point-spread predictions with modern ANN?

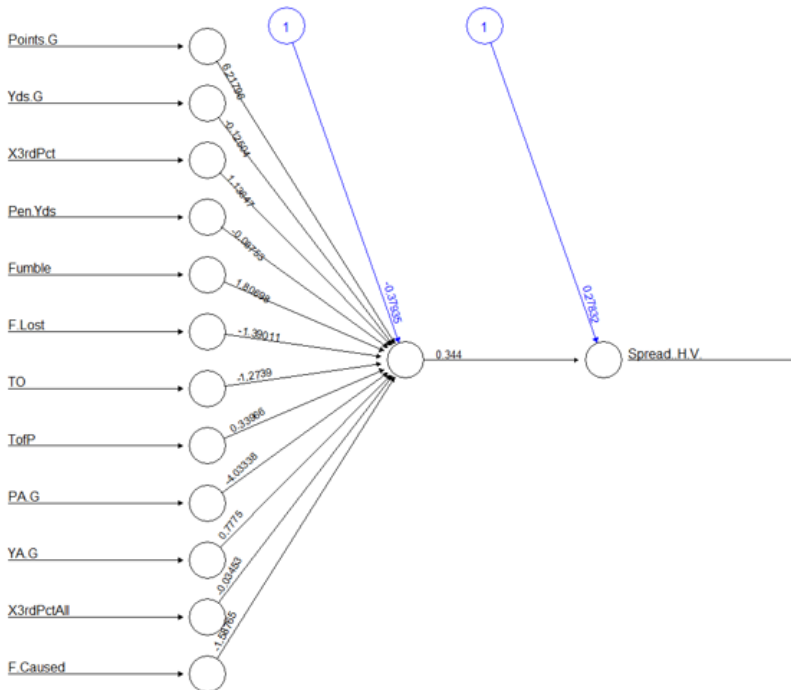
ANNs commonly have an input layer of nodes, an output layer of nodes, and one or more hidden (internal) layers. Deep-learning networks have multiple hidden layers. ANNs that have more layers and more neurons can learn more, but it takes more time and data to train them. Training involves exposing the ANN to cases of independent variable values with matching dependent variable values and then adjusting the node-to-node path weights to maximize accuracy of prediction. Specifically, ANN learning is accomplished by computing a loss function that is a measure of goodness of fit, and then the derivative of the loss is used to adjust the learning-machine path weights. Deep learning requires a huge amount of data. An analogy can be made with a puppy and a human baby. The baby can learn more than the puppy, but it takes more input data and much longer to train it.

For the first ANN model there were 12 input nodes corresponding to the 12 independent variables, with one output node for the point spread. But one does not know how many hidden layers nor how many nodes there should be in each for optimal results, although there are some guidelines in the literature based upon the size of the training dataset and other factors. A very simple model was first used with one hidden layer and one node in it. Later the complexity of the model can be increased if doing so improves upon the accuracy. The simple ANN model would be very similar to MLR if one used a linear activation function in the node. A number of different types of activation functions are employed in ANNs; the most commonly used are sigmoid (logistic), hyperbolic tangent, ReLU (rectifier linear unit), and Softplus.

The R neuralnet package was used with the default logistic activation function, and input data was normalized. R neuralnet provides both logistic and hyperbolic activation functions, or the user can supply their own function. The hyperbolic activation function was also tried, but the accuracy was less. The simple ANN model is shown in Fig. 13 (after training). The model sum of squares of errors (SSE) is 2.37 on the training data, but the correlation with test data is poor (0.5678) and the RSME was 97.89, much worse than our other learning models.

Next, more complex models were tried with up to four nodes in the hidden layer, as shown in Fig. 14. For four hidden nodes, the SSE on the training data was better at 1.467, but the correlation with the test data was worse at 0.272. The more complex models with more numbers of hidden nodes gave less accuracy than the simple ANN model. Therefore, more complex ANN models perform worse than the simple model on the NFL dataset here. The more complex the ANN model, the greater the

Figure 13. One-Node Hidden-Layer ANN



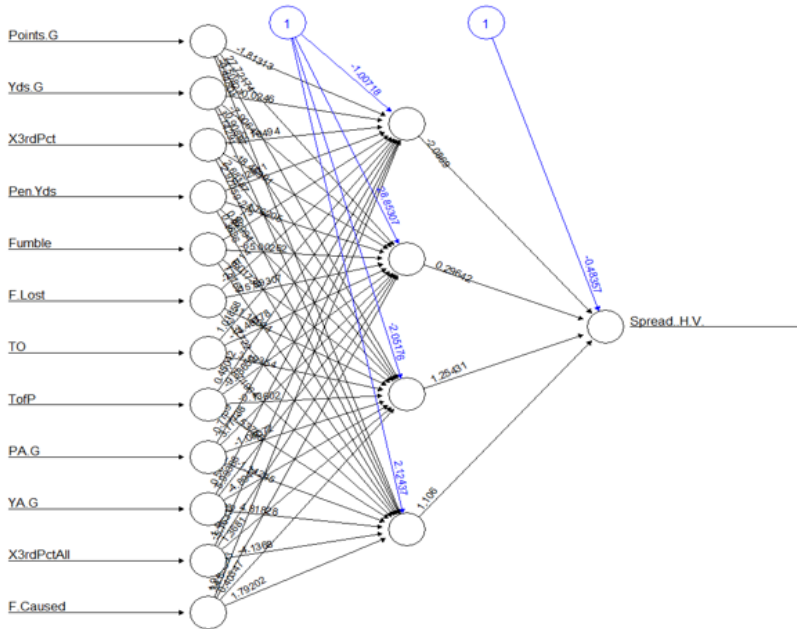
overfitting, which results in lower accuracy on the test dataset. In general, ANN models need many data points (thousands) to be good predictors. One might suggest using data points for multiple NFL seasons; however, the nature of the problem changes as the game changes, and the team makeup changes every year with new players (draft choices), retiring players, trades, new coaches, new rules, team relocations, etc.

For all the learning methods used herein, the training data was 80% of the total dataset and a single run was used. Accuracy may have been improved with other ratios of training to test data. Use of k-fold cross-validation may have also improved comparison, particularly with the ANN with its high overfitting. For example, in making 10 runs of the ANN four-node model where a different 80% training set is chosen each time, the RSME error ranged from 96.71 to 156.59, with a mean of 127.14. For each learning method used here, a separate R program was created using RStudio with hyperparameters set by user experience and industry guidelines. The world of machine learning continues to advance rapidly, and newer integrated development environments are now available to automatically deploy and tune multiple methods in one program (Patel et al., 2020). Hyperparameter optimization is also now commonly used (Shahul & Bajaj, 2023), and one such R package is EZtune (Lundell, 2023).

CONCLUSION

Football statistics have proven to be a rich real-world data source for teaching and illustrating data analytics and machine-learning techniques for both exploratory analysis and predictive analysis. This study has illustrated the usage of machine learning for football predictive analysis and compared several of the most modern techniques for such. The research here

Figure 14. Four-Node Hidden-Layer ANN



has gone beyond past research, which has focused on using the closing Vegas line and other factors beyond team statistics such as player status. Our research here uses only team statistics available at the start of each week of play.

As reported earlier, sports gambling is a growing problem in colleges, especially as our colleges are completing multimillion-dollar deals with sports books and casinos. As well as being a tool for teaching data analytics, this study should hopefully educate students about how difficult it is to beat the Vegas point spreads without using very sophisticated analytical tools.

The key research questions posed earlier can now be answered:

- Traditional prediction tools such as multiple regression do not provide the accuracy needed to beat the official NFL games odds, being about 4 points weaker.
- Some modern machine-learning methods are better than others for NFL point-spread prediction, notably random forests.
- Even modern machine-learning methods are not able to beat the official NFL game odds when the standard 10% vigorish is applied and the closing odds are used.
- However, today’s best modern machine-learning methods can produce competitive results for NFL point-spread prediction even with the standard 10% vigorish included if the early line is used.

For the particular football dataset here of the NFL 2017 regular-season games, the results are shown in Table 4. The random-forests method is much superior to both multiple linear regression and neural networks for this type of data. Note that the random forests’ accuracy using only team statistics available at the start of each week is comparable to the closing Vegas line historical accuracy of about 13 to 14 points, so betting opportunities are available against the early-week lines.

Table 4. Point-Spread Prediction Results

Method	RMSE
Neural Network (R neuralnet)	97.89
Random Forests (R XGBoost)	13.29
MLR (R lm)	17.09

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING STATEMENT

No funding was received for this work.

REFERENCES

- AGA (American Gambling Association). (2018, September 5). *How much does the NFL stand to gain from legal sports betting?* <https://www.americangaming.org/resources/how-much-does-the-nfl-stand-to-gain-from-legal-sports-betting/>
- Anyama, O. U., & Igiri, C. P. (2015). An application of linear regression & artificial neural network model in the NFL result prediction. *International Journal of Engineering Research & Technology*, 4(1), (Authors). (2005). (Title of chapter.). In J. Albert, J. Bennett, & J. Cochran (Eds.), *Anthology of statistics in sports* (pp. x–x). ASA-SIAM.
- Beal, R., Norman, T. J., & Ramchurn, S. D. (2020). A critical comparison of machine learning classifiers to predict match outcomes in the NFL. *International Journal of Computer Science in Sport*, 19(2), 36–50. doi:10.2478/ijcss-2020-0009
- Bosch, P. (2018). Predicting the winner of NFL games using machine and deep learning. <https://www.semanticscholar.org/paper/Predicting-the-winner-of-NFL-games-using-Machine-Bosch-Bhulai/bcd94e514ac1ed34622810faea2914669071f641>
- Boyd, J. (2017, May 19). *Vegas odds makers accuracy: Standard deviations by point spread*. BoydsBets. <https://www.boydsbets.com/ats-margin-standard-deviations-by-point-spread/>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. doi:10.1023/A:1010933404324
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Routledge Press., doi:10.1201/9781315139470
- Brown, T. (2021, September 7). *The emergence of football's game stats*. Football Archaeology. <https://www.footballarchaeology.com/p/the-emergence-of-game-stats-in-football>
- Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. doi:10.1109/TIT.1967.1053964
- Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. doi:10.2307/2325486
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In L. Saitta (Ed.), *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 148–156). Morgan Kaufmann.
- Freund, Y., & Schapire, R. E. (1999). A short introduction to boosting. *Jinkō Chinō Gakkaiishi*, 14(5), 771–780.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. doi:10.1214/aos/1013203451
- Gimpel, K. (2018). *Beating the NFL football point spread*. Carnegie Mellon University School of Computer Science. <https://www.cs.cmu.edu/~epxing/Class/10701-06f/project-reports/gimpel.pdf>
- Greer, R. (2021, September 15). *Why betting early is critical to beating NFL markets*. PFF. <https://www.pff.com/news/bet-why-betting-early-critical-beating-nfl-markets>
- Harville, D. (1980). Predictions for National Football League games via linear-model methodology. *Journal of the American Statistical Association*, 75(371), 516–524. doi:10.1080/01621459.1980.10477504
- Louis, S. (2023, January 16). *'You freakin' idiots': Dave Ramsey blasts US colleges for pitching online gambling to students—while making millions. Why young people are the perfect prey*. Yahoo Finance. <https://finance.yahoo.com/news/freakin-idiots-dave-ramsey-just-140000783.html>
- Lyons, K. (2020, January 5). *Computers in sport*. Keith Lyons Clyde Street. <https://keithlyons.me/2020/01/05/computers-in-sport/>
- Patel, D., Shrivastava, S., Gifford, W., Siegel, S., Kalagnanam, J., & Reddy, C. (2020). Smart-ML: A system for machine learning model exploration using pipeline graph. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 1604–1613). IEEE. doi:10.1109/BigData50022.2020.9378082

Pickles, N. F. L. (2007, September 7). *Vegas' accuracy of predicting the point spread*. <http://www.nflpickles.com/2007/09/vegas-accuracy-of-predicting-point.html>

RotoDoc. (2016). *Exploiting Vegas's lines in order to find NFL DFS value*. Rotogrinders. Retrieved October 10, 2022. <https://rotogrinders.com/lessons/introduction-808877>

Schafer, J. (2022, September 8). *NFL season brings 'the no. 1 acquisition moment' for sportsbooks*. Yahoo Finance. <https://finance.yahoo.com/news/nfl-season-brings-the-no-1-acquisition-moment-for-sportsbooks-180230238.html>

Seal, C. (2018, December 6). *We tied Vegas in our first attempt at predicting NFL game winners using machine learning*. Medium. <https://medium.com/fantasy-outliers/we-tied-vegas-in-our-first-attempt-at-predicting-nfl-game-winners-with-machine-learning-24a805ab3126>

Shahul, E. S., & Bajaj, A. (2023, August 30). *Hyperparameter tuning in Python: A complete guide*. Neptune AI. <https://neptune.ai/blog/hyperparameter-tuning-in-python-complete-guide>

Spinosa, C. L. (2014). *Testing the efficiency of the NFL point spread betting market* [senior thesis, Claremont McKenna College, Paper 986]. https://scholarship.claremont.edu/cmc_theses/986

Sports Odds History. (2018). *Historical NFL game odds*. Retrieved October 10, 2022. www.sportsoddshistory.com/nfl-game-odds

Statista. (2022). *Sports betting worldwide—Statistics and facts*. <https://www.statista.com/topics/1740/sports-betting/>

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293–300. doi:10.1023/A:1018628609742

Tatevosian, P. (2022, February 23). *DraftKings doubles revenue*. The Motley Fool. www.nasdaq.com/articles/draftkings-doubles-revenue-in-another-year-of-accelerating-growth

The Economist. (2022, February 10). *Sports betting in America is exploding*. <https://www.economist.com/graphic-detail/2022/02/10/sports-betting-in-america-is-exploding>

Wadsworth, C., & Vera, F. (2016). *Predicting point spread in NFL games*. Stanford University. <https://cs229.stanford.edu/proj2016/report/WadsworthVera-PredictingPointSpreadinNFLGames-report.pdf>

Warner, J. (2010). *Predicting margin of victory in NFL games: Machine learning vs. the Las Vegas line*. Cornell. https://www.cs.cornell.edu/courses/cs6780/2010fa/projects/warner_cs6780.pdf