Access*: Interdisciplinary Journal of Student Research and Scholarship

Volume 4 | Issue 1 Article 4

2020

Applying the Data: Predictive Analytics in Sport

Anthony Teeter
University of Washington, Tacoma, ateeter9@uw.edu

Margo Bergman
Uni of Washington - Tacoma, mwb4@uw.edu

Applying the Data: Predictive Analytics in Sport

Cover Page Footnote

If not for Dr. Margo Bergman this paper would not have been possible.

Abstract

The history of wagering predictions and their impact on wide-reaching disciplines such as statistics and economics dates to at least the 1700's, if not before. Predicting the outcomes of sports is a multibillion-dollar business that capitalizes on these tools but is in constant development with the addition of big data analytics methods. Sportsline.com, a popular website for fantasy sports leagues, provides odds predictions in multiple sports, produces proprietary computer models of both winning and losing teams, and provides specific point estimates. To test likely candidates for inclusion in these prediction algorithms, the authors developed a computer model and tested its accuracy compared to the professional black box model. The result, named the Turnover Model, has consistent performance with Sportsline.com in two cases: the moneyline and over/under wagers, and a superior performance in the against the spread wager. Logistic regression analysis was then employed to determine what factors contributed to this superior performance. Further work will refine this model to incorporate additional variables.

Applying the Data: Predictive Analytics in Sport

Conscious or not, people throughout the world conduct a form of future prediction from the time they awake to the time they go to sleep. Whether it is traffic patterns, mentally calculating wages versus the household budget, or the actual cost of that triple latte, people are performing calculated analyses in one form or another every day. This is also true of professional forms of analysis using analytics such as the stock market, scientific testing, medical research, or sports analysis. Every field has some sort of model that is used to predict a perfect outcome. These models are developed by the experts in these fields, who understand the details of the business models they are analyzing. Computer simulations are becoming more popular in the field of sports analytics, especially football (SportsLine, 2019). Sports betting is a $100 billion dollar market (Zion Market Research, 2019), and as such, the algorithms that are used will likely be top-notch.

The original motivation for this project was a curiosity about how the Las Vegas gambling industry developed their betting odds algorithms. I researched the development and background of the methods used to create the odds presented to bettors in Las Vegas and was unsuccessful in finding details, most likely due to their proprietary nature. This inspired me to create my own algorithm to estimate odds for certain types of bets, to be explained in the next section, and compare my results to the published outcomes. In this way, I could analyze what factors were likely included in the professional algorithms. The historical precedent for this is quite large. Some of the greatest concepts in the history of statistics and economics, including the discovery of the utility function, expected utility theory, and the diminishing marginal utility of money, were the direct result of thinking about a problem involving wagers—the Bernoulli Paradox (Samuelson, 1977). While we hope for more modest results, the long-term developments from the concepts in this paper can be far more wide-reaching than simply a new way to gamble.

In this paper, I will be analyzing a predictive model of football scores developed by myself, in conjunction with my faculty sponsor, Dr. Margo Bergman, based on my lifetime of experience playing, watching, and coaching the sport. Some of the factors included in the model were traditional, such as home versus away team designation, current weekly power rankings, and the game scores. The remaining factors are proprietary, but one important feature that arose from the analysis gives the model its name: “The Turnover Model”.

This analysis tests The Turnover Model of predicting football scores using chi-square and logistic regression analysis against a well-known leader in the industry of sports analytics and predictions: the predictive analysis website Sportsline.com. SportsLine (2020) is a website that provides sports fans with projections using a combination of advanced data models, the latest news and information, and the opinions of industry experts. SportsLine’s primary picks are based on a proprietary simulation model developed by Stephen Oh, a sports data science expert. Section 2 presents the methods of this study, Section 3 includes the results, and Section 4 provides the conclusion.

Methods

Definitions

Although there are many types of wagers one could place on a sporting event, we focused on the main three types. They are 1) the *moneyline* (ML) wager or "betting on a specific team to win a game" (Sports Interaction, 2020), where to win, one needs to pick the winner of the contest, either team A or Team B; 2) the *against-the-spread* (ATS) wager defined as "When you bet 'against the spread', it's not enough for the favorite to win the game; now, they have to win by more than a specified amount (the spread) in order for them to 'cover the spread;'" and 3) the *over-under* (Ov/Uv) wager" (Top 10 Sports Betting Sites, 2020).[1] Over/Under betting is also called a totals bet. The total in any given sporting event is a combined score of both teams. The total for these games is an amount that is set by oddsmakers based on how they envision a game will unfold from a scoring perspective. As a bettor, one would need to select if the total number of points scored by both teams will be OVER or UNDER the set total." (MLB Odds 2020, 2020). Here is an example of all three wagers:

Moneyline. Team A is favored to win in the contest with a ML value of -500. Team B, therefore, is the underdog to win with a ML value of +300. This means that if you want to choose Team A, you must wager $500 to win $100 if Team A wins, collecting $600; if Team A loses, you lose $500. If you believe Team B will win, then you wager $100, and you will win $300 if Team B wins, collecting $400; if Team B loses, you lose $100. The favorite always costs more to win less.

[1] Conversely, the underdog team cannot lose by the amount of points set on the "against-the-spread".

Against-the-Spread. In an ATS bet, you win the same amount that you wager. This is mostly true; however, one must factor in the 10% commission, or *juice*, that is taken into the sportsbook that creates the line and wager. So, if you bet $110, you win $100, collecting $210. This can almost be looked at as an insurance tax. If you are successful on your wager you receive the 10% wager amount back on top of your winnings. If you are unsuccessful in your wager, the sportsbook keeps the full amount of the $100 wager and the added in $10, or 10% *juice*. What you are wagering on is the number of points by which the team will win the game. If Team A is favored by 6.5 points, and you choose Team A, then Team A has to win the contest by 7 or more points for you to win your wager. If you select Team B, then you win the wager if Team B wins, or if Team B loses by no more than 6 points. Team B can *lose* the contest and you can still win if they lose by 6 points or fewer.

Over-Under. An over-under (Ov/Un) wager is a set total of points that the two teams will score in combination at the end of the contest. For example, if the Ov/Un line is set at 44.5 points, and you choose "Over", then both teams combined points must be higher than 45 for you to win. If you choose "Under", then the combined total must be 44 or under. This is also an even-money bet, like the ATS, so if you bet $110, you win $100, collecting $210[2].

Power Rankings

There is not a consensus when it comes to National Football League (NFL) power rankings for calculating wagers. In general, it is a ranking of the "best" to the

[2] Because the bet-offering entity (typically a casino) wishes to make money regardless of the outcome of the contest, the odds offered for these "even money" bets is typically -110, where one wins $10 for every $11 wagered.

“worst” of the teams, in any given week. The exact ordering might differ, within small blocks, such as the top four, but the subblocks of ranks Top 5, Top 10, Top 20, are generally the same. They are generated by industry experts and there is not a canonical method for creating NFL power rankings (2020 NFL Power Rankings, 2020).

Turnover Margins

A Turnover Margin is calculated by subtracting the total number of “giveaways”—interceptions & fumbles lost—from the total number of “takeaways”—interceptions & opponent fumble recoveries The Football Database, 2020).

Weather

In NFL Weather reports, wind and temperature are arguably the most important factors since they can contribute to the in-game decisions of the coach (play-calling). If the wind speed is high, it is very likely that a team will struggle throwing the football. Also, the wind plays a major factor in the kicking game and those easy field goals now become difficult. Low temperatures can make the football harder to catch, resulting in a more likely running game scenario (Vegas Insider, 2020).

Predictive Analysis Score Prediction Algorithm

This analysis compares ten weeks of predictions (Turnover Model versus SportsLine Model control) starting in week 4 of the NFL football season (SportsLine, 2020). Weeks 1, 2, and 3 were shown to have little standalone predictive value for the model in initial analysis of this iteration of the Turnover Model, given the particular set of factors that are included. By aggregating the information starting in week 4, this allows the previous 3 weeks of data collection to seed the predictive model process. Previous seasons' data cannot be used, because of excessive heterogeneity between teams due to coach and player changes, the NFL draft, and free agency.

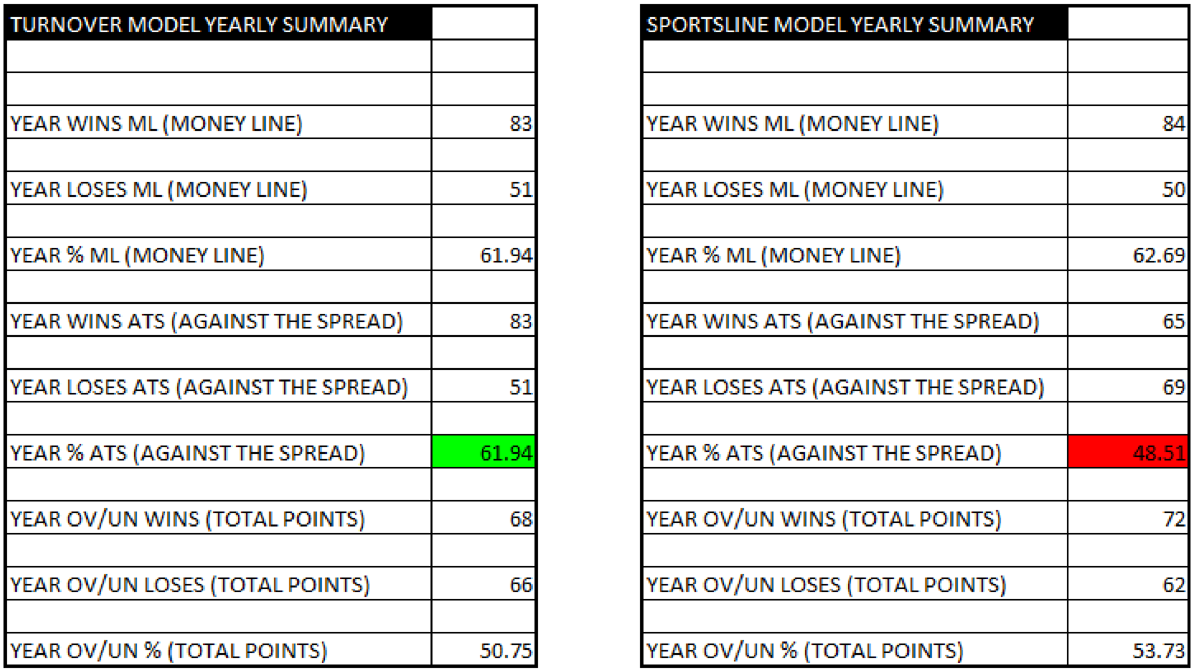
The predictions of the model are: moneyline wager, against-the-spread wager, and over-under wager. The Oakland Raiders were excluded from analysis[3].
The data is collected through a proprietary Excel spreadsheet, using publicly available variables, including, but not limited to, power rankings, turnover margins, and weather. No insider information, such as knowledge of contract negotiations or unreleased injuries, was obtained and or used. The variables are used to create a predictive set of scores for the previously mentioned outcomes. The data is entered into the Excel spreadsheet, which is pre-prepared with a series of formulas. The results from these formulas were then post-processed to consider any situational variables, such as current week injuries, to develop the final score prediction. These final predicted scores were collected into a table that shows all the games for easy reference and tally at the end of each week.

Results

The outcomes are compared for 10 weeks between the Turnover Model predictions vs. Sportsline.com predictions for the 2019 / 2020 season using chi-square analysis and logistic regression. The final outcomes after the completion of 10 weeks and N=134 games are shown in Table 1. In this table a "win" for a ML wager indicates that the team chosen won the contest. An against the spread "win" means that the team

[3] The Turnover Model algorithm includes some elements of human perception in its calculations. Therefore, since I am an Oakland Raiders fan, to eliminate bias, I do not perform any calculated analysis on their games, nor did I include SportsLine's predictions on their games for comparison equality.

chosen to win against the spread won against the spread. An over/under “win” means that the over/under wager chosen was met. The % wins are the total wins/N.



TURNOVER MODEL YEARLY SUMMARY	
YEAR WINS ML (MONEY LINE)	83
YEAR LOSES ML (MONEY LINE)	51
YEAR % ML (MONEY LINE)	61.94
YEAR WINS ATS (AGAINST THE SPREAD)	83
YEAR LOSES ATS (AGAINST THE SPREAD)	51
YEAR % ATS (AGAINST THE SPREAD)	61.94
YEAR OV/UN WINS (TOTAL POINTS)	68
YEAR OV/UN LOSES (TOTAL POINTS)	66
YEAR OV/UN % (TOTAL POINTS)	50.75

SPORTSLINE MODEL YEARLY SUMMARY	
YEAR WINS ML (MONEY LINE)	84
YEAR LOSES ML (MONEY LINE)	50
YEAR % ML (MONEY LINE)	62.69
YEAR WINS ATS (AGAINST THE SPREAD)	65
YEAR LOSES ATS (AGAINST THE SPREAD)	69
YEAR % ATS (AGAINST THE SPREAD)	48.51
YEAR OV/UN WINS (TOTAL POINTS)	72
YEAR OV/UN LOSES (TOTAL POINTS)	62
YEAR OV/UN % (TOTAL POINTS)	53.73

Table 1: Final Outcomes T.O. Model versus S.L. Model

Chi-Square Analysis

A Chi-square test was performed on the three predicted outcomes: moneyline wager, against the spread wager, and over/under wager. The test measures how well the observed distribution of data fits with the expected distribution if the variables are independent of the outcome of the test. Without any knowledge or analysis, the potential winner of a binary contest can be determined by a coin flip. Therefore, in our comparison, we assumed the observed distribution would be 50/50 for all three wagers. For the win/loss $\chi 2$ (1, N = 134) = 0.0159, $p > .05$, and over/under $\chi 2$ (1, N = 134) = 0.2393, $p > .05$, so there was no significant difference in results. In the case of the against the spread $\chi 2$ (1, N = 134) = 4.8892, $p < .05$, there was a significant difference.

Regression analysis

Following the chi-square analysis, I wanted to determine what factors, if any, were identifiable as predictive of the significant difference in the ATS percentage between myself and Sportsline.com. Regression analysis is used when you want to predict a dependent variable from any number of independent variables. If the dependent variable is dichotomous, as is the case in this situation, then logistic regression is the proper choice of regression method. The independent variables used in regression can be either continuous or dichotomous. One point to keep in mind with regression analysis is that causal relationships among the variables cannot be determined. While the terminology is such that we say that X "predicts" Y, we cannot say that X "causes" Y.

The variables chosen for analysis in the logistic regression were:

1. Predicted Against the Spread winner from the Turnover model (dependent variable) (0/1) 1 if Win
2. Predicted Score Team 1
3. Predicted Score Team 2
4. Final Score
5. Predicted winner from the Turnover Model (0/1) 1 if Home Team
6. Actual game winner (0/1) 1 if Home Team

Matching Variables:

7. Matched prediction between the Turnover Model and a personal prediction (0/1) 1 if Home Team

8. Matched prediction between the Turnover Model and Las Vegas prediction (0/1) 1 if Home Team

Interaction Variables:

9. Matched prediction between the Turnover Model and actual winner (0/1) 1 if Home Team

The purpose of matching the variables was to see if there was any predictive value for my personal predictions versus the computer predictions in understanding the capability of the Turnover Model to do so well for ATS wins.

Logistic Regression Results

Dependent Variable:			
1	Odds Ratio	2.5%	97.5%
(Intercept)	2.022	.0159	32.374
2	1.0546	0.9500	1.1736
3	1.0284	.09270	1.1289
6	0.3044	0.0791	1.1209
4	0.5341	0.1675	1.5615
7	0.4184	0.1588	1.0431
8	0.4171	0.1313	1.2302
5	0.0423	0.0089	0.1580
9	76.075	13.541	519.68

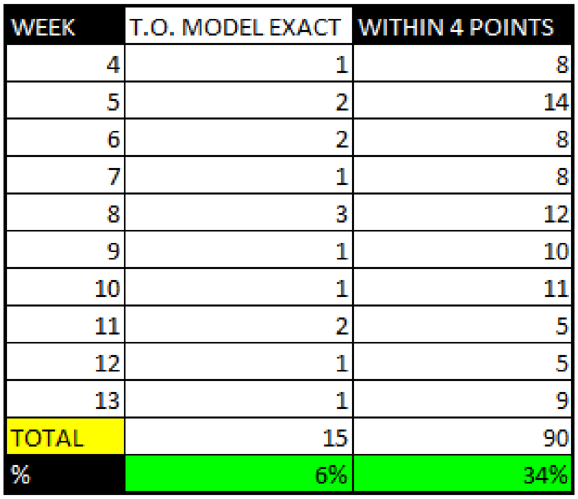
Table 2: Odds Ratios of Logistic Regression

An odds ratio is the relative probability of success to failure. Typically, in regression, a constant effect of the independent variable on the dependent variable is the measure of interest (Martin, 2020). However, in a logistic regression, the probability effect is not constant. Therefore, we must take a monotonic transformation of the data, such as the logarithm, and from this, probabilities can be calculated in percentage terms, rather than absolute. Instead of p-values, 2.5% - 97.5% confidence intervals are presented. If the confidence interval crosses 1, then that variable is not considered to be statistically significant. For example:

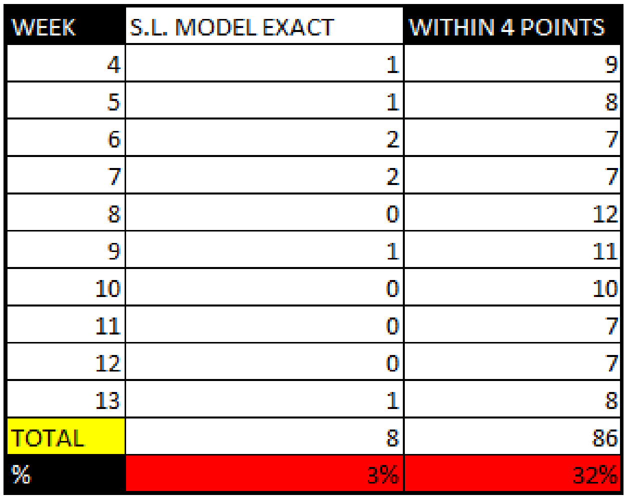
> “Let’s say that the probability of success of some event is .8. Then the probability of failure is 1 – .8 = .2. The odds of success are defined as the ratio of the probability of success over the probability of failure. In our example, the odds of success are .8/.2 = 4. That is to say that the odds of success are 4 to 1. If the probability of success is .5, i.e., 50-50 percent chance, then the odds of success is 1 to 1.” (Introduction to SAS, 2016)

Here we see that the Turnover model would have been 76 times MORE likely to win ATS prediction when the Turnover model prediction matched the real game winner. In more concrete terms, when the Turnover model was able to predict the winner of the game, it was also able to more accurately predict the score of the game than the competitor’s website, leading to increased ATS win percentages. Additionally, when the Turnover model predicted a Home team winner but the actual winner was the Away team, the Turnover model is 24 times LESS likely to win an ATS prediction. This shows a clear bias of some kind within the Turnover model, for the Home team. The exact nature of this bias will be analyzed in future research.

With a total of 134 games and 268 individual team score predictions, the Turnover model was able to correctly predict the actual score of 15 individual team scores with an exact prediction rate of 6%. Compared to 8 games that of SportsLine with an exact prediction rate of 3%. The Turnover Model was also within 4 points of 90 individual team scores with a 34% rate of games predicted within 4 points compared to games that of SportsLine with 86 games predicted with a 32% rate of games predicted within 4 points of the actual total.



WEEK	T.O. MODEL EXACT	WITHIN 4 POINTS
4	1	8
5	2	14
6	2	8
7	1	8
8	3	12
9	1	10
10	1	11
11	2	5
12	1	5
13	1	9
TOTAL	15	90
%	6%	34%



WEEK	S.L. MODEL EXACT	WITHIN 4 POINTS
4	1	9
5	1	8
6	2	7
7	2	7
8	0	12
9	1	11
10	0	10
11	0	7
12	0	7
13	1	8
TOTAL	8	86
%	3%	32%

Discussion

The outcome of the Chi-square test indicates that the Turnover model performs significantly better for the against the spread wager predictions, but the same for the moneyline or over/under, as compared to SportsLine. Therefore, it seems likely that the algorithm in use here is similar to SportsLine for the two wager predictions with non-significant algorithms. For the ATS predictions, given that the Turnover Model predicts a higher positive outcome, it would seem this model is superior to the algorithm in use by the professional Sportline algorithm.

In analyzing the potential reasons that the Turnover Model algorithm performs better for ATS, we determined that it has strong predictive power when the algorithm accurately predicts the game winner. However, when the real winner is the Away team, the Turnover Model suffers. The model, therefore, clearly includes many factors that are important for predicting a Home Team win.

Conclusion and Future Research

Over the course of ten weeks, the model described in this paper predicted the winner against the spread better than the SportsLine model predictions. It also seems to predict with equal success for the two other wager categories. Future research will include improving the model's ability to predict Away team wins, analyzing specific aspects of the Turnover Model's point prediction ability, and automating injury and weather information. The source of bias of the Home versus Away predictions is an important goal for future research.

* This paper is in no way condoning or suggesting illegal gambling. This study is the sole purpose of a statistical analysis of sports and numbers using mathematical scientific studies and statistical testing for the prediction of final outcomes of a sporting event.

References

UCLA: Statistical Consulting Group. (n.d.). Introduction to SAS. Retrieved from https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/ (accessed 9/9/2020)

Zion Market Research. (July 2019) “Research: US $155.49 bn for sports betting market size 2019 growing at 8.83% CAGR through 2024.” Retrieved from https://www.globenewswire.com/news-release/2019/07/26/1892289/0/en/Research-US-155-49-Bn-for-Sports-Betting-Market-Size-2019-Growing-at-8-83-CAGR-Through-2024.html

Grace-Martin, K. (n.d.). “Why use odds ratios in logistic regression”. Retrieved from https://www.theanalysisfactor.com/why-use-odds-ratios/

Odds Shark. (July 2020). MLB odds 2020 - Best baseball odds & lines for MLB. Retrieved from https://www.oddsshark.com/

Safest Betting Sites. (2020). “2020 NFL power rankings - NFL betting strategy guide.” Retrieved from https://www.safestbettingsites.com/nfl-betting/nfl-power-rankings

Samuelson, P. (March 1977). St. Petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature, 15* (1), 24–55.

SBD. (June 2019). Top 10 sports betting sites, betting odds, news & guides: SBD. Retrieved from https://www.sportsbettingdime.com/

Sportsinteraction.com (December 2019). Sports Betting at Sports Interaction (SIA), Canada's Online Sportsbook. Retrieved from https://www.sportsinteraction.com/

Sportsline.com. (August 2019). 2019 NFL win totals: Proven computer model releases best bets.” Retrieved from https://www.sportsline.com/insiders/2019-nfl-win-totals-proven-computer-model-releases-five-best-bets/

Sportsline.com (January 2019). SportsLine FAQ. https://www.sportsline.com/insiders/sportsline-faq/ accessed 12/13/2019

The Football Database. (July 2020). 2019 NFL turnover differential. Retrieved from https://www.footballdb.com/stats/turnovers.html

Vegasinsider.com. (September 2020) How does the weather affect NFL football. Retrieved from https://www.vegasinsider.com/nfl/weather/