Predicting The Regular Season with Machine Learning ~ Who Comes Out On Top?

STAT 440

Final Project

Xingyuan Fang

Eli Reber

Shawn Rodricks

Link to Presentation: shorturl.at/qsyE4

Table of Contents

Introducing the Problem

Predicting the outcome of sports games has always been an in-depth, attractive but challenging venture attracting a large audience as of late. Sports prediction now applies to an array of fields such as sports betting, sports business operations, player salaries, player development, etc. and this is only possible by the recent explosion of data analytics in sports. Ever since Billy Beane used Bill James' Moneyball theory that one can use math instead of money to operate his Oakland Athletics Baseball team, every sports owner has turned to data to find hidden knowledge to drive decisions and their sports organizations on a day to day basis.

Our project aims to study an application of statistics around machine learning and sports. With Covid-19 spreading around rapidly in the United States right now, all business industries have been hit in a multitude of ways. A huge industry hit is the entertainment industry and sports fans are reeling around the world. The sports industry in the US includes 4 major sport leagues: National Football League, National Basketball Association, Major League Baseball and National Hockey League.

With individual team season totals data, we aim to find the population mean winning percentage of all 4 US sports leagues champions (NFL, NHL, NBA, and MLB)… and investigate which league required the most amount of regular season dominance to win a championship. Then we will run machine learning algorithms to predict a teams total wins for 2020 (since this season has been suspended indefinitely) based on the previous year's data and historical data to find a hypothetical champion for each league in the suspended 2020 season and how likely they are to win the championship.

Introducing the Data

We needed two types of datasets. One where we can see every champion teams' winning percentage to calculate what it takes to become a consistent threat to winning a championship each year. We also needed a dataset with every team's scoring numbers (runs allowed vs. runs scored or touchdowns allowed vs. touchdowns scored etc.) as well some other advanced statistics tracked throughout the course of a season to run machine learning models and determine the pythagorean expected win totals for 2020.

Our initial data set (figure 1) includes each champion from each league dating back to each league's inception or post-merger. To explain, our MLB data begins in 1903, with the first ever World Series (finals) played. Our NFL data begins in 1966, when the first Super Bowl was played between the NFL and AFL champions. Our NHL data begins in 1967, after the major expansion occurred, when the league expanded from 6 to 12 teams. Our NBA data begins in 1977, after the merger happened between the ABA and NBA.

We collected all of this data by scrapping from various websites. We used baseball reference and this took the most time to scrap as there were a lot of features we needed to predict win totals for baseball and this sport is the oldest in the US. We also used pro football reference to scrap for football. NHL.com and NBA.com provided sufficient statistics with win/loss records for all teams so we stuck with those for hockey and basketball data. We scrapped in depth also to get additional data sets per each sport for the past decade, dating back to include the 09 season so we could get accurate statistics to run regression models as well.

Stratifying and Finding Champion Win %

Our strata are the representatives of the four leagues. Every league has a different number of regular season games played, so we felt that this was the most appropriate way to break up our clusters. The NBA & NHL have 82 games played in a season compared to MLB's 162 games played and NFL's 16 games played. First, we had to calculate the population variance for each of the league champions' winning percentages. The NBA and the NHL had a variance of about 0.005 which makes sense because both leagues played an 82 game regular season. The NFL had the highest variance of 0.008 which is a result of the 16 game season. For example, the 2011 NFL champion New York Giants had a regular season winning percentage of 0.563 which is very low while the 2016 New England Patriots had a winning percentage of 0.875. That range is simply due to the shorter season whereas the MLB's population variance was 0.0018. Playing the longest regular season of the "Big 4", having the most consistent championship format and having the least variability in teams in the league. In total, there were 43 NBA champions, 53 NFL champions, 52 NHL champions, and 115 MLB champions. These were all of our Ni values for each league. We used a bound of .03. Given we wanted equal strata sizes, our ai's for each strata was ¼. The sample size that we need from each sample is 18/4 = 4.5, which we then round up to ni = 5.

We created a box plot (figure 2) to verify if the variances are constant within the groups and means were different across groups. We believe the variances are close enough with the only issue possibly coming from the NFL; we concluded that this is due to the much shorter season of NFL (16 games) compared to the other leagues (82 games & 162 games). The shorter the season, the more variability, hence the outliers in the NFL and NBA. The stratified means are different as well, showing us that stratified is the correct method. Finally, because the leagues are independent, resulting in non-overlapping groups. Using a random number generator from Google Sheets, we randomly selected 5 championship teams' winning percentages for each league and found the yi bar's and sample variances with the randomly selected data. Our sample means and sample variances are as follows:

NFL: Y Bar = .820, Si^2 = .006	NHL: Y Bar = .602, Si^2 = .008
NBA: Y Bar = .771, Si^2 = .003	MLB: Y Bar = .586, Si^2 = .0002

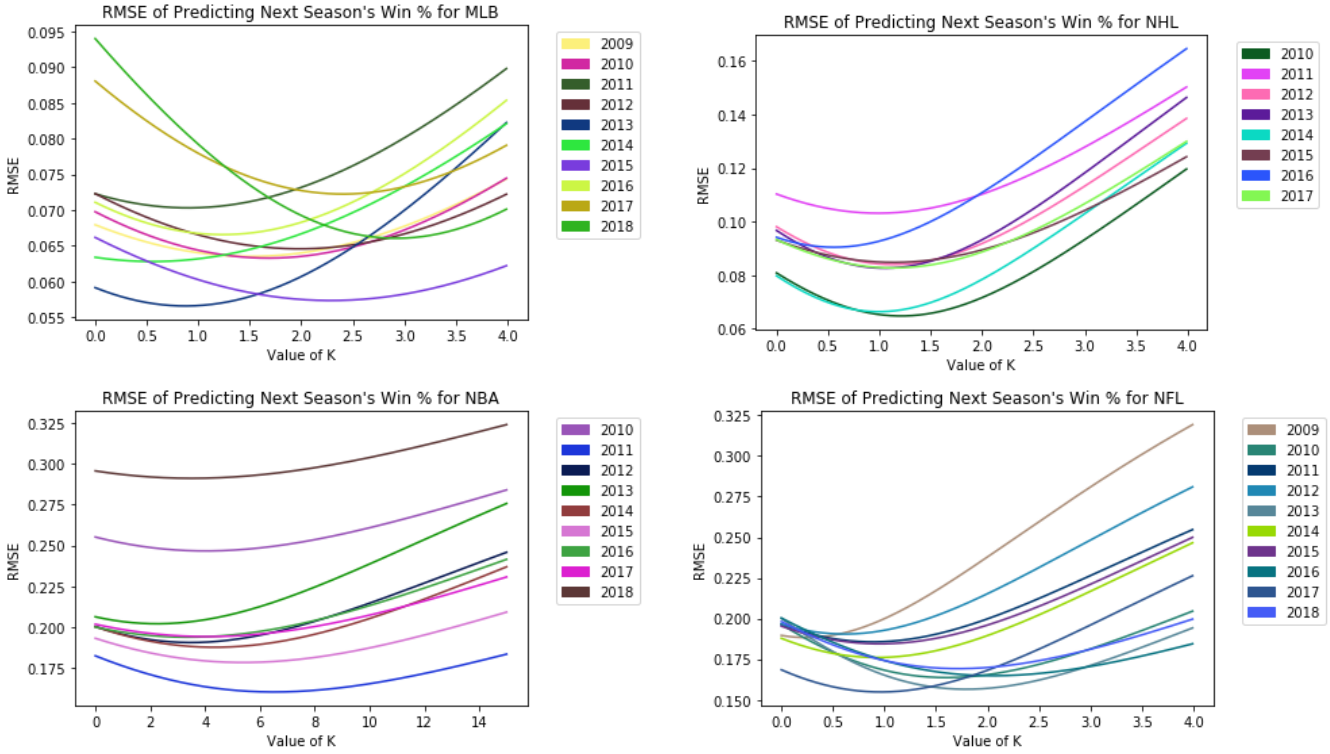
Using these numbers and all of the N, ni, and Ni values previously mentioned, we used the following formulas (figure 3) to calculate the stratified sample mean, variance, and bound. Using the stratified sampling system, our estimated population mean for championship team winning percentages is .667 ± .022.

Machine Learning Techniques

When it comes to using machine learning in the field of sports, one might ask how that really brings any value to the world. Machine learning models outperform Vegas bookmakers by 2-10%. Any skilled bettor could turn that difference into retirement money. Teams can use their own historical data to find hidden insights into where an aspect of their game may be lacking and what areas to improve. The widely available data that has been tracked on a game to game basis (as much as 14 baseball games played a day during the season) has made machine learning models very accurate in predicting team insights. This is what we are doing as well. Teams have in turn used statistics to value players and negotiate contracts so the whole structure of an organization can be traced to how well they can analyze statistics. We will utilize a neural network to predict the win totals of teams. Ideally, this information would be used to inform teams on where they need to improve to secure more wins.

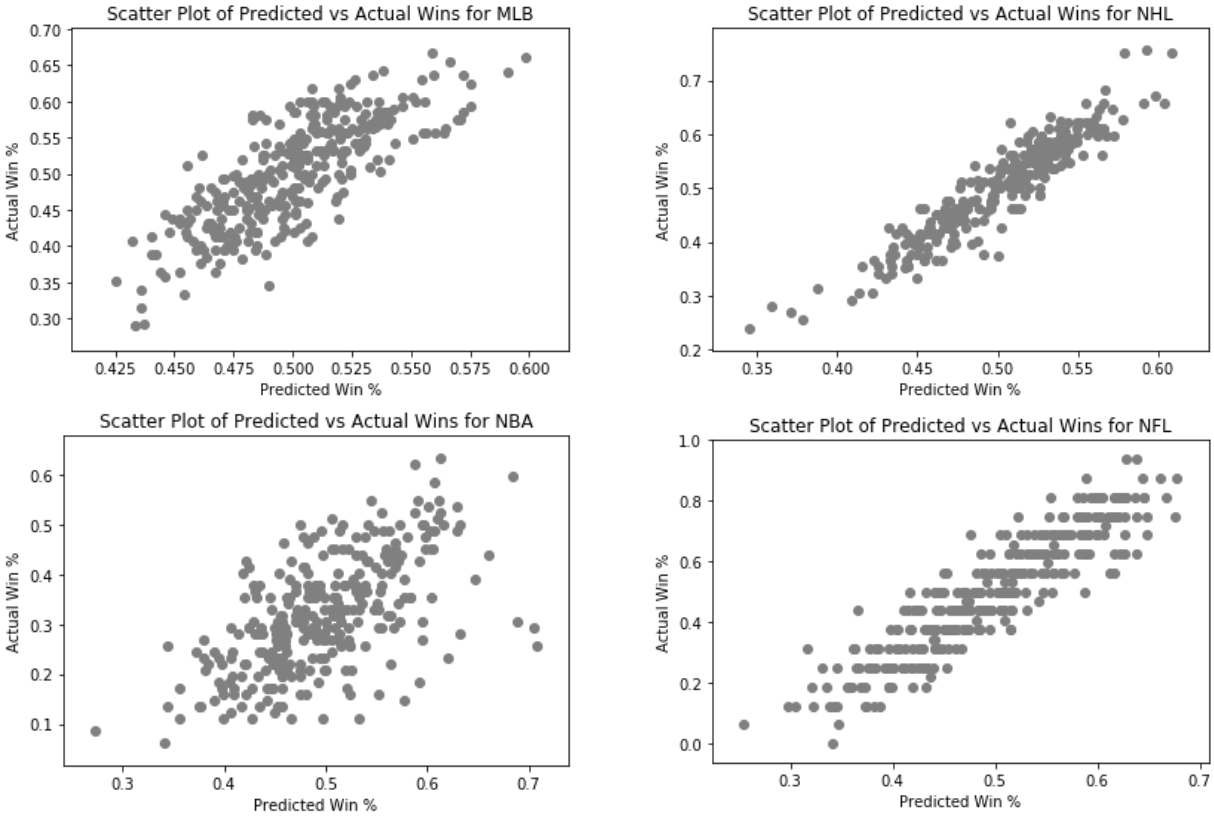
We will also delve into pythagorean win totals, a statistic derived to measure true performance, used by Bill James in baseball and then adapted for basketball by Daryl Morey and now widely used all over the sports industry. The Pythagorean Win Total (or expectation) represents how many games a team should have won without those lucky or unlucky bounces in close games during a season. So, the logic follows that the Pythagorean total should be more predictive of future wins than actual wins as those odd bounces will cancel out. A team will play more like its Pythagorean Win total than real life total in future games and even the next year. So we then look to see how certain teams win (those who win the championship and make it far in the playoffs) and what causes a majority of the teams to lose (what features negatively affect win total).

The Pythagorean win expectation is a simple, yet pretty powerful tool that can be used to predict a team's winning percentage in a given year just using the points a team scored and the points that were scored against the team. For this study, we used each team's yearly point-scoring counts to predict their winning percentages for the next year. How many points or runs were scored vs. how many points or runs were allowed. The equation for the predicted winning percentage can be seen in Figure 4, however the exponent (called K in this case) used in the equation was treated like a hyperparameter in a machine learning model and different values were used to find the exponent that gave the best root mean squared error (RMSE).



From these figures, it can be seen that baseball is the most predictable sport when using the previous year's scoring data, as all of the RMSE values are below 10% for every value of K in each season. Winning percentages in hockey and football follow a similar pattern (shaped similarly to a hockey stick), with an initial drop in RMSE, an optimal value around 1, and then a continuous increase in error as the value of K increases. Basketball seems to be the least predictable sport when using the Pythagorean expectation. Apart from the very high RMSE values, the shapes of the curves indicate that changing the value of K does not impact the predictability of the expectation all that much.

The optimal values of K that gave the lowest RMSE values for each season were collected and the average of the values was taken. The average optimal K value for each sport was then used to predict the winning percentage for every season. The result of the predicted winning percentage and the actual winning percentage for each sport was then plotted.



	NHL	NFL	NBA	MLB
K Value	1.03	1.18	4.10	1.63
Correlation	0.937	0.911	0.601	0.791
R^2	0.879	0.829	0.361	0.625

As can be seen from these plots, the average value of K for each of the sports did offer significant insights into the actual win percentage. It can also be noted that, as the value for K increased, the correlation between the predicted and actual win percentages decreased. Complementing the results from the RMSE plots, basketball was by far the worst sport to use the Pythagorean expectation as a predictor for, with a correlation almost 0.2 below the second-lowest correlation.

Although baseball was the sport with the best RMSE values when predicting just the next season, the average K value did not do as good of a job at predicting the seasons over the 10 year time period. This suggests that outcomes in baseball do not change much from one year to the next year, but they do change over time. The RMSE values for hockey were higher than those for

baseball, but the correlation between the predicted and actual winning percentages were very high. This suggests that hockey is a sport where outcomes might change slightly from one year to the next year, but does not change much overall.

We wanted to compare these pythagorean expectation results to a neural network with all of the teams records per sport dating back to when the leagues started instead of just taking the previous years statistics. We looked for features like runs scored, errors/ turnovers committed, yards per play, goals scored/allowed, rankings, penalties, etc… all attributes that both positively and negatively affect win total per team in each league. Links to all of our aggregate data is shown in the references. We used supervised learning while building these datasets since the data processing phase is guided toward the class variable while building the model. We used an ANN model (artificial neural networks) to solve this regression (predicting final win percentage over the 2020 season) problem (figure 5). Based on clustering, we used unsupervised methods to distinguish between winning and losing teams. Through experimentation, the best architecture in terms of normalized RMSE of the neural network was found to be 4-3-1 (four input neurons/variables, one hidden layer with three neurons, and one outcome). We achieved 64.5% accuracy and we used all this training data from all years in the past to to test on the testing data set for 2020. The best team in terms of winning percentages are summarized in figure 6. When running regression, we were able to see that the winning percentage normally increased if the team was successful (> .500 Win/Loss record) and generally decreased when the team was unsuccessful (> .500 Win/Loss record). So good teams got better, and worse teams got worse.

Predicted Champions for Each League

	MLB	**NHL**	**NBA**	**NFL**
Pythagorean Expectation	Houston Astros	Tampa Bay Lightning	Portland Trail Blazers	Baltimore Ravens
Neural Network	Atlanta Braves	Philadelphia Flyers	Los Angeles Lakers	New Orleans Saints

The two results were dramatically different as the pythagorean expectation drew from the exact previous year… hinting teams that were good in 2019 would be good in 2020. Regression did the same but to less extremes and we were able to compare these winning percentages generated with actual 2020 data since the NBA actually started. The trail Blazers are a losing team in 2020 even though the first model predicted them to win it all. The next model takes the sample statistics calculated before also into account… showing the best team (figure 6) does not necessarily always win the championship.

References

Aggregate Data (2020, May 1). Baseball Reference, Pro Football Reference, NBA.com, NHL.com. Retrieved from https://cutt.ly/tyk7c7N

Forman, S. (2000, April 1). Baseball Reference. Retrieved from https://www.baseball-reference.com/

Rosenbloom, C. (2003, January 3). Pro Football Reference. Retrieved from https://www.pro-football-reference.com/

NHL.com. (n.d.). Retrieved from http://www.nhl.com/stats/

NBA.com. (n.d.). Retrieved from http://www.nba.com/stats/

Appendix

Year	Lg	Team	Wins	Losses	Ties	Regular Season Winning Percentage	Index	Random Sample of Indexes	
2019	NBA	Toronto Raptors	58	24	0	0.707	1	0.616	
2019	MLB	Washington Nationals	93	69	0	0.574	2	0.673	
2019	NHL	St. Louis Blues	45	37	0	0.549	3	0.597	
2018	NBA	Golden State Warriors	58	24	0	0.707	4	0.599	
2018	NFL	Super Bowl LIII: New England Patriots	11	5	0	0.688	5	0.875	
2018	MLB	Boston Red Sox	108	54	0	0.667	6	0.669	
2018	NHL	Washington Capitals	49	33	0	0.598	7	0.714	
2017	NBA	Golden State Warriors	67	15	0	0.817	8	0.805	
2017	NFL	Super Bowl LII: Philadelphia Eagles	13	3	0	0.813	9	0.756	
2017	MLB	Houston Astros	101	61	0	0.623	10	0.817	
2017	NHL	Pittsburgh Penguins	50	32	0	0.610	11	0.659	
2016	NFL	Super Bowl LI: New England Patriots	14	2	0	0.875	12	0.549	
2016	NBA	Cleveland Cavaliers	57	25	0	0.695	13	0.813	
2016	MLB	Chicago Cubs	103	58	0	0.640	14	0.571	
2016	NHL	Pittsburgh Penguins	48	34	0	0.585	15	0.537	
2015	NBA	Golden State Warriors	67	15	0	0.817	16	0.750	
2015	NFL	Super Bowl 50: Denver Broncos	12	4	0	0.750	17	0.650	
2015	MLB	Kansas City Royals	95	67	0	0.586	18	0.598	
2015	NHL	Chicago Blackhawks	48	34	0	0.585	19	0.580	
2014	NBA	San Antonio Spurs	62	20	0	0.756	20	0.659	
2014	NFL	Super Bowl XLIX: New England Patriots	12	4	0	0.750	21	1.000	
2014	NHL	Los Angeles Kings	46	36	0	0.561	22	0.875	
2014	MLB	San Francisco Giants	88	74	0	0.543	23	0.695	
2013	NFL	Super Bowl XLVIII: Seattle Seahawks	13	3	0	0.813	24	0.549	
2013	NBA	Miami Heat	66	16	0	0.805	25	0.682	
2013	NHL	Chicago Blackhawks	36	12	0	0.750	26	0.750	
2013	MLB	Boston Red Sox	97	65	0	0.599	27	0.630	
2012	NBA	Miami Heat	46	20	0	0.697	28	0.556	
2012	NFL	Super Bowl XLVII: Baltimore Ravens	10	6	0	0.625	29	0.857	
2012	MLB	San Francisco Giants	94	68	0	0.580	30	0.549	
2012	NHL	Los Angeles Kings	40	42	0	0.488	31	0.714	
2011	NBA	Dallas Mavericks	57	25	0	0.695	32	0.625	
2011	NFL	Super Bowl XLVI: New York Giants	9	7	0	0.563	33	0.611	
2011	NHL	Boston Bruins	46	36	0	0.561	34	0.695	
2011	MLB	St. Louis Cardinals	90	72	0	0.556	35	0.662	
2010	NBA	Los Angeles Lakers	57	25	0	0.695	36	0.605	
2010	NHL	Chicago Blackhawks	52	30	0	0.634	37	0.574	
2010	NFL	Super Bowl XLV: Green Bay Packers	10	6	0	0.625	38	0.611	
2010	MLB	San Francisco Giants	92	70	0	0.568	39	0.793	
2009	NFL	Super Bowl XLIV: New Orleans Saints	13	3	0	0.813	40	**0.680**	Sample Mean
2009	NBA	Los Angeles Lakers	65	17	0	0.793	41	**0.012**	Sample Variance
2009	MLB	New York Yankees	103	59	0	0.636	42		

Figure 1

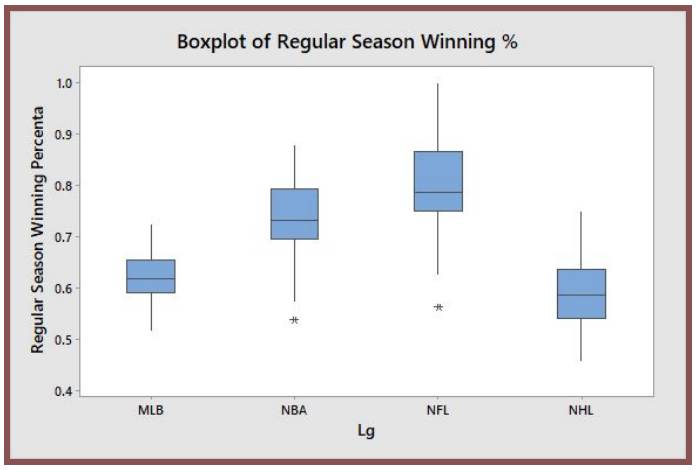


Figure 2

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i \quad = .667$$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L}[N_i^2(1-\frac{n_i}{N_i})\frac{s_i^2}{n_i}] \quad = .00012$$

$$\text{Bound} = 2\sqrt{\hat{V}(\bar{y}_{st})} \quad = 2 * \sqrt{.00012} \quad = .022$$

Figure 3

$$Win\% = \frac{(runs\ scored)^2}{[(runs\ scored)^2 + (runs\ allowed)^2]}$$

Figure 4

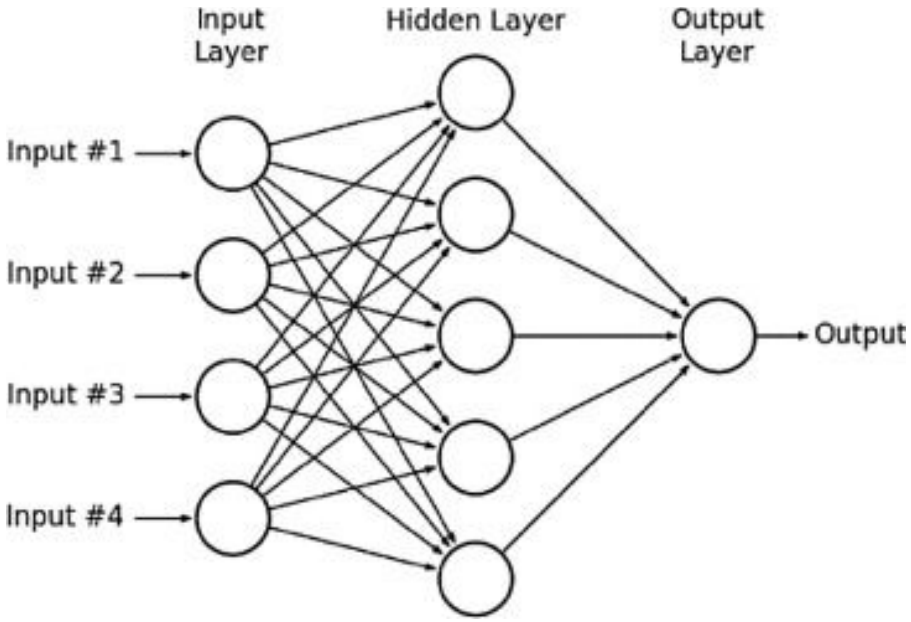


Figure 5

A	B	C	D	E	F
Sport	Year	Tm	W	L	W-L%
Football	2020	San Francisco 49er	14.384	1.616	0.899
Football	2020	Baltimore Ravens*	13.472	2.528	0.842
Football	2020	New Orleans Saint	13.168	2.832	0.823
Football	2020	Green Bay Packers	13.088	2.912	0.818
Basketball	2020	Milwaukee Bucks	66.502	15.498	0.811
Basketball	2020	Los Angeles Laker	63.386	18.614	0.773
Basketball	2020	Toronto Raptors	58.22	23.78	0.71
Baseball	2020	NYY	111.942	50.058	0.691
Football	2020	New England Patri	10.88	5.12	0.68
Hockey	2020	Boston Bruins	55.022	26.978	0.671
Baseball	2020	LAD	108.054	53.946	0.667
Football	2020	Houston Texans*	10.656	5.344	0.666
Hockey	2020	Colorado Avalanch	54.612	27.388	0.666
Basketball	2020	Denver Nuggets	54.202	27.798	0.661
Basketball	2020	Boston Celtics	54.202	27.798	0.661
Hockey	2020	Tampa Bay Lightn	52.644	29.356	0.642
Basketball	2020	Utah Jazz	52.316	29.684	0.638
Basketball	2020	Houston Rockets	52.152	29.848	0.636
Football	2020	Minnesota Vikings	10.128	5.872	0.633
Hockey	2020	St. Louis Blues	51.824	30.176	0.632
Baseball	2020	OAK	102.222	59.778	0.631
Basketball	2020	Oklahoma City Th	51.66	30.34	0.63
Baseball	2020	MIN	100.44	61.56	0.62
Football	2020	Buffalo Bills+	9.776	6.224	0.611
Basketball	2020	Miami Heat	50.102	31.898	0.611
Baseball	2020	HOU	98.82	63.18	0.61
Hockey	2020	Philadelphia Flyers	50.02	31.98	0.61
Football	2020	Kansas City Chiefs	9.6	6.4	0.6
Basketball	2020	LA Clippers	49.2	32.8	0.6
Football	2020	Seattle Seahawks+	9.584	6.416	0.599
Football	2020	Tennessee Titans+	9.568	6.432	0.598
Basketball	2020	Philadelphia 76ers	48.954	33.046	0.597
Hockey	2020	Washington Capita	48.708	33.292	0.594
Baseball	2020	ATL	95.742	66.258	0.591
Football	2020	Philadelphia Eagle	9.44	6.56	0.59
Basketball	2020	Dallas Mavericks	48.134	33.866	0.587
Basketball	2020	Indiana Pacers	47.56	34.44	0.58
Hockey	2020	Pittsburgh Penguin	47.478	34.522	0.579
Hockey	2020	Minnesota Wild	46.494	35.506	0.567
Baseball	2020	TBR	91.53	70.47	0.565
Hockey	2020	Winnipeg Jets	45.92	36.08	0.56
Hockey	2020	Carolina Hurricane	45.756	36.244	0.558
Hockey	2020	Vegas Golden Knig	45.018	36.982	0.549
Hockey	2020	Toronto Maple Lea	44.608	37.392	0.544
Baseball	2020	ARI	87.966	74.034	0.543
Baseball	2020	PHI	84.726	77.274	0.523
Hockey	2020	Vancouver Canuck	42.804	39.196	0.522
Baseball	2020	NYM	84.402	77.598	0.521
Baseball	2020	CLE	84.24	77.76	0.52
Hockey	2020	New York Rangers	42.64	39.36	0.52
Hockey	2020	New York Islander	42.558	39.442	0.519

Figure 6