Exploring and Predicting NFL Games

Elina Choi, Shan Lu, Steven Moen

Abstract

We examine different models to predict NFL game results. Our dataset includes the NFL games from the 2006 through the 2019 regular seasons with 256 games for each season, spanning 17 weeks among 32 teams. We first apply the Bradley-Terry model which is a widely used method in predicting the binary outcome variable of pairwise comparison, winning in this context. Our training set includes the games in the last season. Then, we utilize the penalized logistics regression in which we train the data by either week or entire season. The original Bradley-Terry model and the penalized glm model trained by season yield similar prediction accuracy and the weekly-trained penalized logistic model perform the best. Different sets of covariates have been used in the model but do not improve the accuracy significantly. Also, a model to predict the point spread directly is analyzed, though the results are not quite good enough to turn a profit. Overall, while the results here fall short of outperforming predictions from sports books, they demonstrate that a relatively simple pairwise comparison model can provide insight into the game of football.

1 Introduction

It might be fair to call gridiron football, sometimes called "American" football, an obsession in the United States, at all levels of play - high school, college, and professional. The professional ranks generate eye-popping revenue figures - in 2018, they generated \$16 billion [1]. This revenue figure doesn't fully encapsulate the national phenomenon that is American football. Football encapsulates a gamut of emotions that would rival those from even the greatest Shakespeare play within the 60 minutes on the game clock, and with the ball spotted between the hash marks on the gridiron. As statisticians, we seek to better understand this game that, until now, has been quite enigmatic.

There has been an enormous movement in recent years towards improved sports analytics - how to incorporate modern data analysis tools to better understand outcomes of games and to help teams improve their likelihood of a win [2]. In addition to understanding what happens on the field and how to improve results from the perspective of a coach or player, there is also a growing legal sports betting market in many states. Indeed, there is legal sports betting in 19 states and the District of Columbia as of November 3rd, 2020 [3]. While this industry was shrouded in secrecy in the past, it is now becoming more and more a (legitimate) part of game day.

Predicting and analyzing professional American football games played by the National Football League (NFL) is a tricky business compared to the other "Big 4" sports leagues, namely, the National Basketball Association (NBA), the National Hockey League (NHL), and Major League Baseball (MLB). There are 4 reasons that make this the case.

First off, the NFL only has 16 regular-season games per year, which is significantly less than the NHL and NBA's 82 and the MLB's 162. Clearly, per the sample size alone, predicting NFL games will be much more difficult since any standard error will be larger, other things equal. Moreover, there are more NFL teams (32) compared to 30 teams in the other leagues. The scarce levels of data within a season make direct win-loss comparisons difficult. A team in a weak division (such as the NFC East in 2020, with 8 wins total after Week 8 in the NFL amongst the 4 teams in the division) could hardly be compared fairly to a team in a strong division (such as the AFC North, with 19 wins total in Week 8 for the 4 teams in the division) as the 2020 NFL season approaches its midway point.

Secondly, one might instead point to the margin of victory as a means of properly sorting teams. Again, this is misleading - the same considerations mentioned above still apply. A mediocre team playing a bad team may look better than a great team playing a good team. Also, from 2002 through through the 2014 season, about 64% of NFL games were within two touchdowns [4], and there is a very real problem of "garbage time", where a team that is ahead may modify their offensive strategy to chew clock and increase the chance of a victory instead of scoring more points, because the latter poses a greater risk of a turnover or a fumble [5]. To be sure, it seems quite clear that a team plays to win, not to win by a certain margin in the NFL, as this is what matters for making the playoffs. As such, margin of victory is also deceptive.

Third, one might then look to the microfoundations of the NFL game, but the problem there is that the probabilities needed to understand the game are highly conditional and therefore intractable compared to other sports [6], especially at the level of the individual player. For example, evaluating quarterbacks Deshaun Watson against Patrick Mahomes requires a lot of simplifying assumptions since they play very different styles of football under different head coaches with a different supporting cast of skill position players and offensive linemen.

Fourth, there is also the problem of prior year data. To borrow Vilfredo Pareto's phrase [7], there is a lot of "circulation of the elite" among NFL teams - a great team may find themselves in a cellar quickly, and vice-versa. The NFL draft, with its highest picks and therefore best players awarded to the teams with the worst records, a team that was awful one year can find its fortunes reversed the next [8]. Also, given the high injury risks associated with football, players tend to have shorter careers compared to their Big 4 peers [9] and can have a promising season cut short or hobbled with an injury.

2 Literature Review

A number of studies examine the NFL prediction via a neural networks approach. Purucker (1996) [10] explores several neural network strategies including Hamming, Adaptive Resonance Theory (ART), Kohonen SelfOrganizing Map (SOM), and Back-Propagation (BP). The article also addresses the comparison between supervised and unsupervised training methods. Khan (2003) [11] adopts a back-propagation multi-layer perceptron network among those four strategies. Its advantages include the flexible utilization of parameters and the supervised learning frame which would increase the predictive ability based on training. David, et al. (2011) [12] and Blaikie, et al. (2011) [13] address a neural network model that utilizes the use of statistical differentials to compare teams and to further predict the game results. The model determines which metrics influence the model the most by using PCA and derivative-based analysis. The performance of the prediction is further assessed by analyzing the NFL data over multiple seasons.

Additionally, Uzoma, et al.(2015)[14] introduces a hybridized prediction model utilizing different machine learning algorithms, linear regression, and K-Nearest Neighbour (KNN). The weights of different features were obtained from running the linear regression and used as a basis for the KNN model. It adopts unique features such as the bookmakers' betting spread and players' performance metrics to improve the model accuracy. The prediction accuracy is 80.65% which performs better than other existing methods with accuracy rates around 60-70%. The testing set in Uzoma, et al. is the last two weeks in season 2013. The point spread implied prediction accuracy during this period is 81.25%.

On the other hand, Schlicht (2017)[15] introduces a bias-correcting model to improve the prediction results. While either a subjective or algorithmic decision-making process can be biased, it focuses on the decision biases in an exploitative manner. By utilizing the real-world gambling data to train and test different prediction models, it demonstrates that leveraging oddsmaker biases in an exploitative manner performs the best.

3 Results

3.1 Data description

The dataset for analysis contains NFL games from the 2006 through the 2019 regular season. Each season has 256 games, spanning 17 weeks, among 32 teams. Related variables include game attributes: "game_type", "game_date_time", "location", "stadium_roof", "stadium_surface", "temperature", "wind", "referee"; away team attributes: "away_team", "away_score", "away_coach"; home team attributes: "home_team", "home_score", "home_coach"; and betting variables: "spread_line", "total_line", "under_odds", "over_odds", "away_moneyline", "away_spread_odds", "home_moneyline", "home_spread_odds". The experimental units in this analysis are games.

3.2 Preliminary analysis

Our main focus of the paper is to build a predictive model for the binary game results:

$$Y_i = \begin{cases} 1, \text{ if home team wins game } i; \\ 0, \text{ if home team losses game } i. \end{cases}$$

This problem is naturally high dimensional. If we create one indicator variable to represent each pairwise comparison, then 496 parameters need to be estimated. However, there are only 256 games per season. A team's ability is also expected to change from season to season, or even before-after a major event within a season, e.g. transfer of players, changing coach, or player injuries.

Before conducting the main analysis, we need to assess the temporal dependence of a team's ability. Because of the high-dimensionality of the data, we cannot compute the auto-correlation of the game results. Estimators from the Bradley-Terry model [16] provide a ranking for teams. We use the Spearman correlation of Bradley-Terry parameters between seasons as a surrogate for temporal correlation.

The Bradley-Terry model is a probabilistic model for pairwise comparisons. The probability of event "team i beats team j" is formulated as follows:

$$\{\text{i beats j}\} \sim \text{Bernoulli}(logit^{-1}(\lambda_i - \lambda_j)) \tag{1}$$

where $\lambda_i = \beta_{i0} + \sum_r x_{ir}\beta_r$ and $\lambda_j = \beta_{j0} + \sum_r x_{jr}\beta_r$.

1. The parameter β_{i0} quantifies the ability of team i, i.e. if $\beta_{i0} > \beta_{j0}$, then then team i has higher chances of beating team j.

2. $\{x_{ir}\}_r$ are team specific variables that may influence game results. For instance, with $x_{i1} = I_{\{\text{team i at home}\}}$, β_1 expresses home team advantage.

Denote $\hat{\beta}_i^{team}$ as the Bradley-Terry parameters for the 32 teams in season i. The Spearman correlation is computed between $\hat{\beta}_i^{team}$ and $\hat{\beta}_{i-1}^{team}$, and shown in Figure 1. The average correlation between season (i, i-1), (i, i-2), (i, i-3) are 0.35, 0.24, 0.12 respectively. Because of the relatively fast decay of the Spearman correlation, we will only include games in the last and the current season in model training.

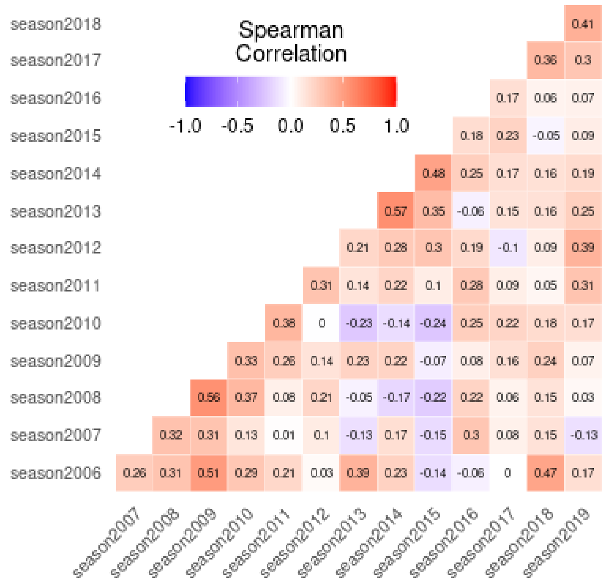


Figure 1: Rank correlation of Bradley-Terry score between seasons

3.3 Model

Following the parameterization of the Bradley-Terry model, i.e. logistic regression, we consider three different training strategies - M1, M2, and M3.

Vanilla Bradley-Terry model with covariates (M1)

- Response variable: $Y_i = \begin{cases} 1, \text{ if home team wins game } i; \\ 0, \text{ if home team losses game } i. \end{cases}$
- Independent variables X_i contains

  1 intercept,

  2 team indicators $\{\text{ARI}_i, \text{ATL}_i, ..., \text{WAS}_i\}$, defined in the following way: $\text{ARI}_i = 1$, if ARI is home team in game i; -1, if ARI is away team in game i; 0 otherwise.

- Training set: games in the last season.
- Model:

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} Y_i \log p_i - (1 - Y_i) \log(1 - p_i) \tag{2}$$

where $p_i = \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}$.

Penalized logistic regression (M2)

- Response, independent variables and training set are defined as in model M1. Model M2 is defined as

$$\min_{\beta} \frac{1}{N}\sum_{i=1}^{N} Y_i \log p_i - (1-Y_i)\log(1-p_i) + \lambda \sum_j \|\beta_j - \tilde{\beta}_j\|_2^2 \tag{3}$$

where $p_i = \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}$, $\{\tilde{\beta}_j\}_j$ are parameters estimated from the previous season. λ is selected through cross validation.

Penalized logistics regression adapted with data in the current season (M3)

- Response, independent variables and model are defined as in model M1. The training set is defined as: (1) games in the last season, for the prediction of games in week $1\ldots,6$; (2) games in the current season in week 1 to k-1, for the prediction of games in week k (k$\geq$7).

3.4 Model comparison through prediction accuracy

We compare the three training strategies (M1, M2, M3) with the spread-implied prediction (Table 1). The spread-implied prediction for game i, denoted as Y_i^{spread}, is defined as

$$\hat{Y}_i^{spread} := \mathbb{1}_{\{\text{The spread for home team in game i is positive}\}}. \tag{4}$$

	Testing accuracy	Training accuracy
spread	67%	
Bradley Terry (M1)	59%	74%
Penalized GLM, training by season (M2)	59%	72%
Penalized GLM, training by week (M3)	63%	

Table 1: Model comparison

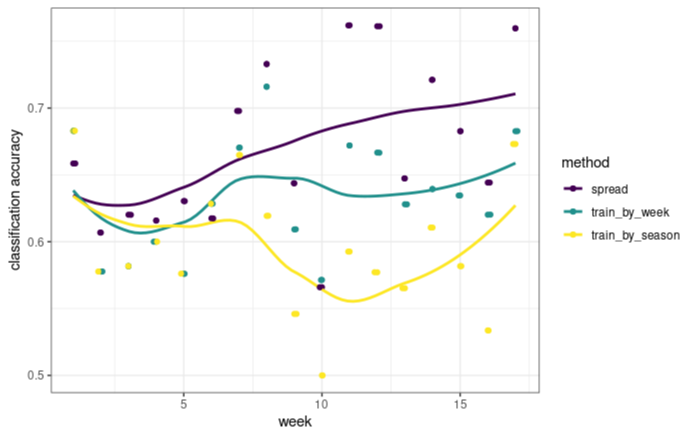


Figure 2: **Comparison of classification accuracy among spread-implied prediction, M2 and M3 by week.** Each dot in the figure is an average of prediction accuracy over all seasons for the week corresponding to the x-axis. The smoothing line is computed by "loess", a nonparametric regression technique.

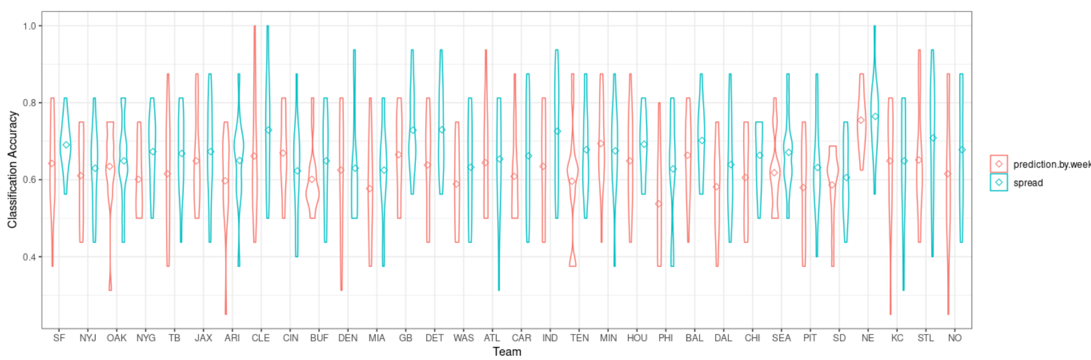


Figure 3: **Violin plot of the prediction for each team.** The violin-shaped density plot shows the distribution of the prediction accuracy for each team over all seasons. The dot in the middle shows the mean of the corresponding distribution. The color indicates different methods (spread-implied prediction – blue, M3 – red).

The Bradley-Terry model (M1) and the seasonly-trained Penalized GLM (M2) performs similarly in terms of training and testing accuracy. The big gap between training and testing accuracy is largely driven by changes of a team's ability between seasons. The weekly-trained Penalized GLM (M3) performs better than M1 and M2, because of the improved performance in games 8 to 17 (Figure2). None of our training strategies beat the spread-implied predictor, which is the outcome that is forecasted by the spreadline (if a team is favored to cover the spread, then they are favored to win). When we look at the prediction performance for each individual team, the spread-implied predictor gives higher average accuracy for 30 out of the 32 teams (Figure 3).

3.5 Other variables

Models M1, M2, M3 only include team names as covariates. We explored the relations of game outcome with a few other variables, e.g. week_days, difference_in_rested_days, change_coach, extreme_weather, spread. Among the variables being tested, change_coach and spread have significant correlation with game results (Figure 4). However, these variables did not improve prediction accuracy when they are included in the logistic regressions (M1-M3).

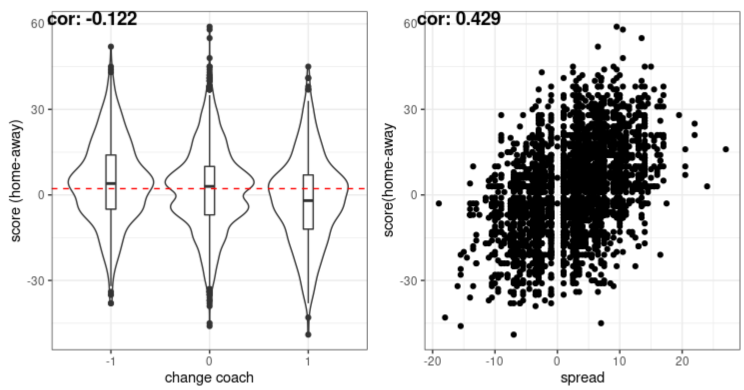


Figure 4: **Correlation between game results and related variables.** The left figure shows the box plot of the game score (home-away) in each change-coach scenario. The right figure shows the scatter between spread and game score.

3.6 Running a Betting Strategy

A natural question that arose when running our model was how it might compare to sports books. Unfortunately, this comparison requires a bit of calculation. The moneyline (which gives the payout of a team winning or losing) has quite a bit of missing data, and so it's necessary to convert a spread to a probability to compare the two, which is somewhat difficult. Luckily, a website analyzing about 40 years of NFL data was available. A subsection of the table is included in this report. Also, note that using a value of 0.999999 is necessary for games with a large spread because there were missing data after 16.5.

Point Spread	Favorite Win Probabilities
0	0.50
0.5	0.50
1	0.513
1.5	0.525
2	0.535
2.5	0.545
⋮	⋮
17+	0.999999

Table 2: How to Convert Point Spread to Odds

3.7 Comparing Model Predictions with Implied Vegas Odds

Included in this report are a few figures that compare the forecasted win probabilities from the model against the win probabilities from the point spread.

Figure 5: The win probabilities implied by the Sports Book spread line diverge significantly from those predicted by our model

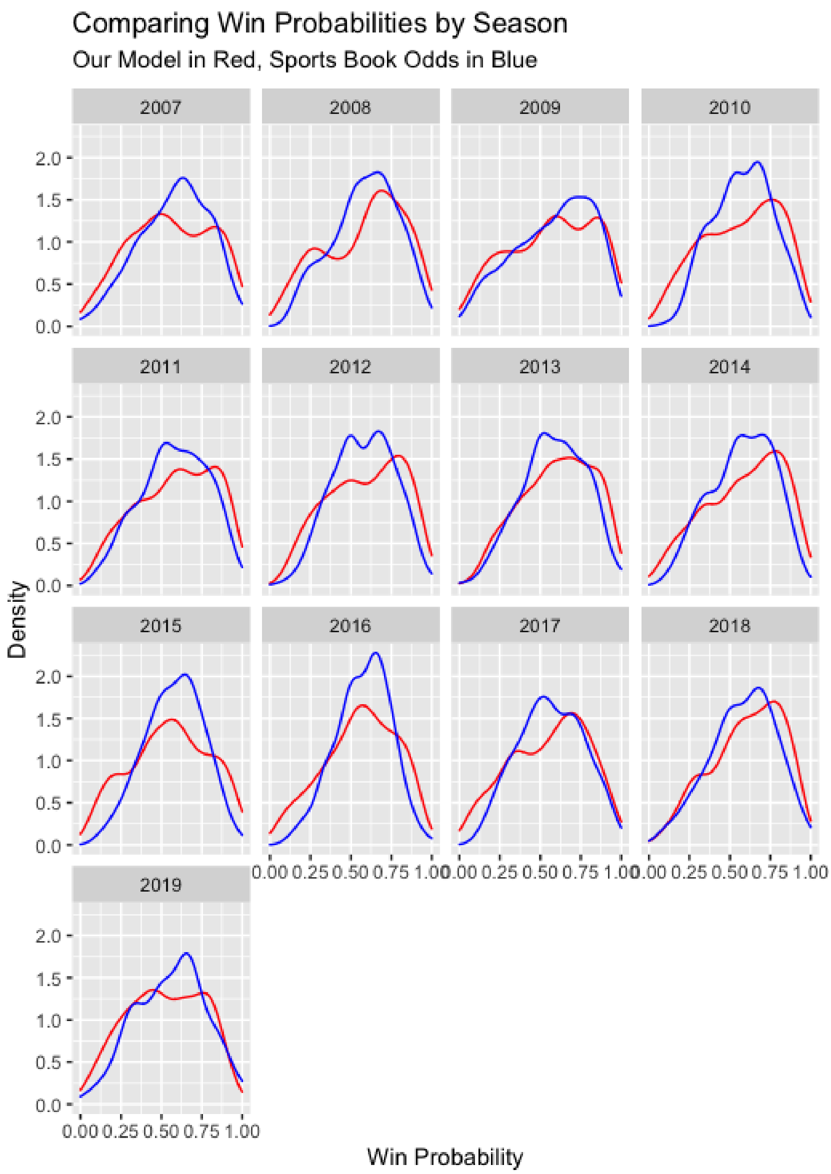


Figure 6: There doesn't appear to be a clear pattern of convergence or divergence across seasons

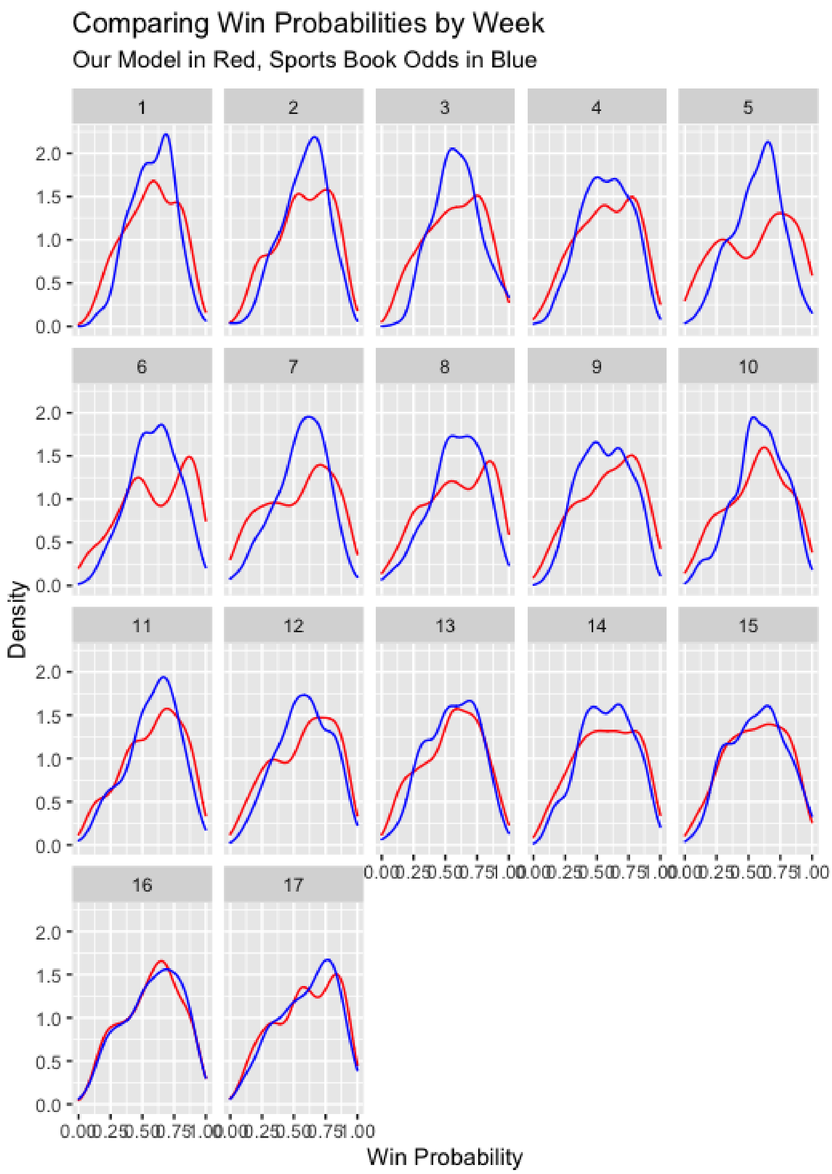


Figure 7: The probabilities tend to converge later in the season

Perhaps the most interesting of these is the one comparing the win probabilities by week. Notice how the two seem to converge towards each other as a season progresses. Remember that in the train-by-week model (M3) described above, there is perhaps the best ability for a model to pick up on the team-level dynamics that are always present in the Vegas sportsbooks.

3.8 The Kelly Criterion

With the conversion of the sports book spread into odds, the next pertinent question is how to determine the optimal bet size. One such method is the Kelly Criterion, which maximizes wealth almost surely compared to other methods. However, as is seen below, a pitfall of this method is that there is a high probability of ruin. Below are details on the method:

- In this equation, f^* is the optimal percentage of your portfolio to bet:

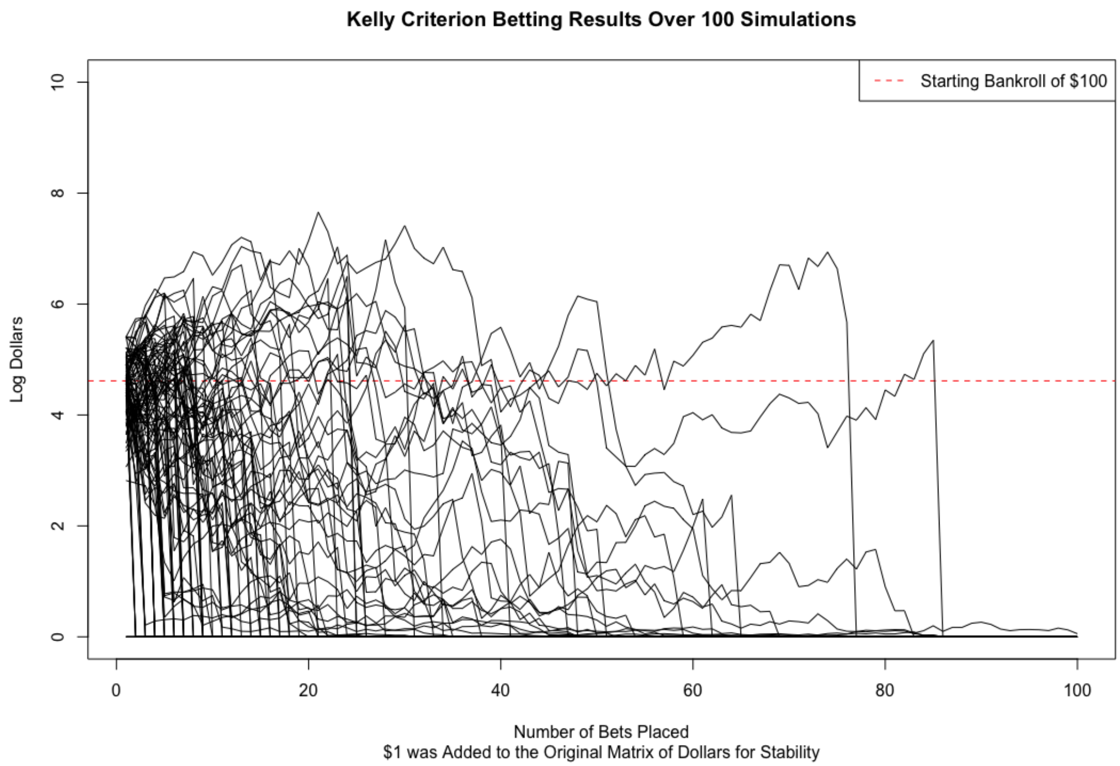
$$f^* = p - \frac{1-p}{b}$$

  where p is the predicted win probability from our model and b is the net gain from a wager. (If you win \$5 from a \$10 bet, then $b = 5/10 = 0.5$)

- We calculate b from the spread line probability calculated earlier $b = \frac{1-p_{spread}}{p_{spread}}$

3.9 Results from the Kelly Criterion Simulation

The probability of ruin is high with the Kelly criterion. Even with a successful strategy, the bets that the Kelly criterion encourages can be quite large.



3.9.1 Notes on the Simulation

- 100 simulations were run
- There is no "spread" taken out of a bet that would exist in Vegas
- It's not possible to lose more than all your money – the dataset is zeroed out at $0
- A bankroll from a bet at time t is used in a bet at time $t + 1$
- 100 random samples are pulled from our data, ordered from earliest to latest, and simulated chronologically. This is to avoid the problem of potentially having bets occuring at the same time with real data

3.10 Modeling the Spread Line Directly

A natural next step would be to try to model the spread line directly. This is a tricky proposition - indeed, teams typically play to win, not to win by a certain amount. However, a benefit of doing this is that a model that is even slightly more accurate than the spread can indeed be quite successful. Below is a discussion of that, which will be used in future considerations.

3.10.1 Breakeven Win Percentages

Sportsbooks need to take a spread of what is bet to make a profit. This spread is called many things, as described here:

"'Vig' (also known as vigorish or 'juice') refers to the fee a bookmaker or sportsbook charges a bettor for placing their wager. The vig allows the bookmaker or sportsbook to make money on every betting line, regardless of the actual outcome of the event. Bettors win and lose, but the right vig guarantees the bookies always win." [17]

In other words, the key for a profitable betting strategy is to be able to earn enough money in order to avoid the cut that the sports book will take. Odds on spreads are often quoted as -110 or -105; this refers to the fact that a wager of \$110 would be required to win \$100 if the odds were quoted with -110. Below, the required win percentage is calculated. Let $\hat{w}$ be the required win percentage for breaking even:

$$100w - 110(1 - w) = 0 \tag{5}$$

Rearranging, we have:

$$210w = 110 \tag{6}$$

$$\hat{w} = \frac{110}{210} \approx 52.38\% \tag{7}$$

Similarly, for a -105 line, the breakeven win percentage ($\bar{w}$) would be given by:

$$\bar{w} = \frac{105}{205} \approx 51.22\% \tag{8}$$

3.10.2 Models Used

The models used here are similar to those used above.

Penalized Spread Line Model with Covariates – Trained by Season

- Response variable: Y_i is the amount that the home team wins by.
- Independent variables, X_i contain:

  1 intercept,

  2 team indicators $\{\text{ARI}_i, \text{ATL}_i, ..., \text{WAS}_i\}$, defined in the following way: $\text{ARI}_i = 1$, if ARI is home team in game i; -1, if ARI is away team in game i; 0 otherwise.

  3 the spread available from the Sports Book (which is known before the game)

- Training set: games in the last season.
- Model:

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j} \|\beta_j - \tilde{\beta}_j\|_2^2 \tag{9}$$

  where $\{\tilde{\beta}_j\}_j$ are parameters estimated from previous season. (For the first season, the penalty term, λ, is equal to 0). For later seasons, λ is selected through cross validation on the training set, and $\|\cdot\|_2$ is the L_2 norm.

Penalized Spread Line Model with Covariates – Trained by Week

- The response, independent variables, and model are defined as in the previous section. The training set is defined as: (1) games in the last season, for the prediction of games in week $1 \ldots, 6$; (2) games in the current season in week 1 to k-1, for the prediction of games in week k (k$\geq$7).

3.10.3 Modeling the Spread Line Directly - Results

Below is a table outlining the prediction results from the spread line:

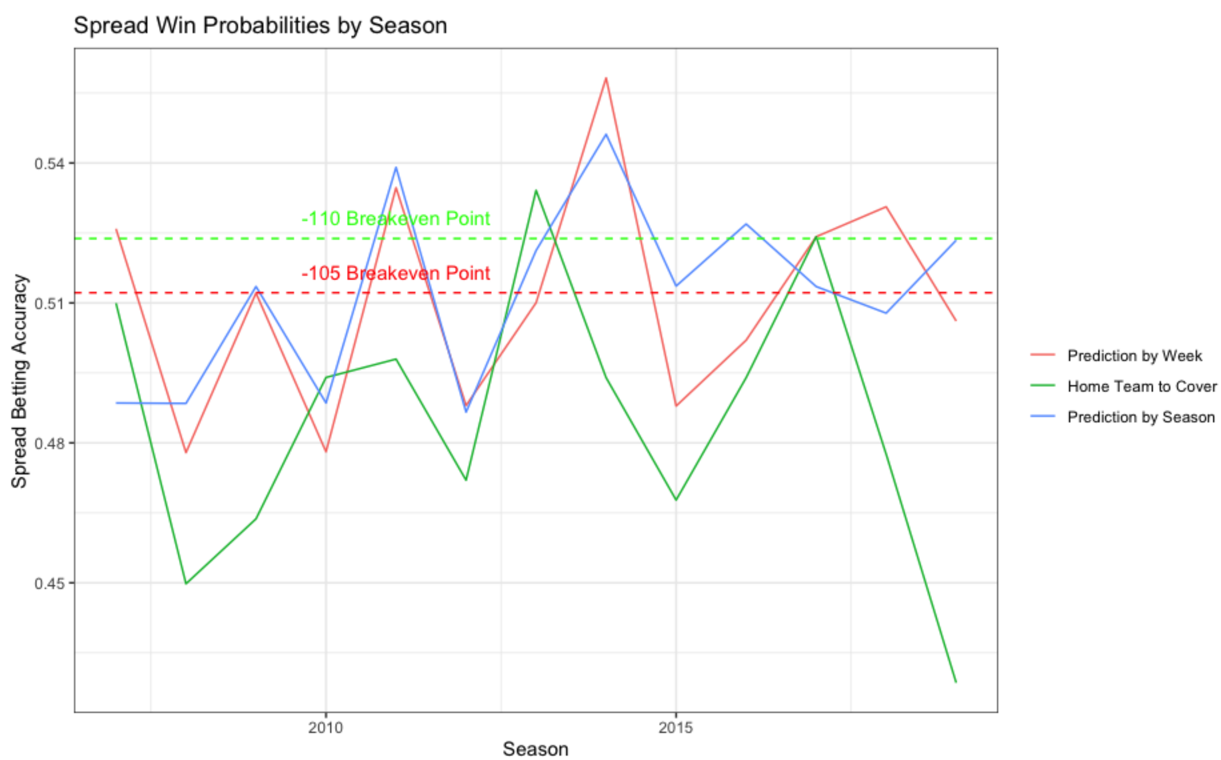
Category	Win Probability
Betting Home Team to Cover	48.52%
Prediction by Season	51.21%
Prediction by Week	51.04%
-105 Breakeven Point	51.22%
-110 Breakeven Point	52.38%

Table 3: Summary of Point Spread Betting Model

There are a few things to note here – while the prediction by season model performs the best, it falls just short of the threshold required to make a profit. Moreover, if one were to simply not be on the home team to cover, then they would win $100\% - 48.52\% \approx 51.48\%$, enough to make money with the -105 odds!

A Note on "Pushes"

An event known as a "push" occurs when the result of the game exactly matches the spread. When a push occurs, the bettor receives their money back. To account for this, pushes are simply dropped from the dataset as if the game were never played.



The two figures above generally confirm the same view that is seen in the table, namely, that while the model seems to outperform the simple bet of taking the home team to cover the spread, it does not consistently outperform the threshold required to turn a profit.

4 Conclusions and Next Steps

There is certainly more work to do here – while there appears to be a rich literature on predicting and understanding games, there appears to be some value in the simplicity of our method. While the focus of our game has been on predicting games and how it compares to odds published in Las Vegas, there is certainly value in interpretation as well. The odds in sports books are generally seen as a benchmark of what is considered good, and what is not, and by testing the models above against that high bar, it lends credence to the quality of the work done here. While an NFL team may like to think about how things occur between the end zones and between the sidelines, there is certainly value in understanding the wisdom of those who seek to be an Oracle into the uncertain future. While an NFL executive, coach, or player may consider it odd (or perhaps insulting) for an armchair statistician to try to forecast a game that they know better than others, we hope that our model brings them insight into where they might want to focus their energies. Indeed, they are able to see - and perhaps change - the reality on the turf to prove our model wrong, whether it be using a fullback to block a blitz-heavy defense, drafting a pass-catching running back, or running different secondary schemes to avoid getting burned by a speedy wideout. Given how the results from this game are etched into the minds of Americans, we hope that our model can become a small part of what makes this game great.

5 Data and code availability

The NFL game data is available at Github folder `https://github.com/leesharpe/nfldata/blob/master/DATASETS.md#games`. The R code for all analysis involved in this paper is available at `https://github.com/altarigilbert/stat992project`.

References

[1] Jay Ojha. The 10 biggest sports leagues in the world by revenue — Pledge SportsPledge Sports, 2020. URL `https://www.pledgesports.org/2020/02/the-10-biggest-sports-leagues-in-the-world-by-revenue/`.

[2] Andrew Witherspoon. NBA three-pointers are leading the sports analytics revolution - Axios, 2019. URL `https://www.axios.com/three-pointers-lead-sports-analytics-revolution-b5613e67-92fe-44a3-897e-ede6780f0edb.html`.

[3] Ryan Rodenberg. The United States of sports betting - Where all 50 states stand on legalization, 2020. URL `https://www.espn.com/chalk/story/{\_}/id/19740480/the-united-states-sports-betting-where-all-50-states-stand-legalization`.

[4] Chad Langager. What is the Most Common Margin Of Victory In The NFL?

- SportingCharts.com, 2014. URL `https://www.sportingcharts.com/articles/nfl/what-is-the-most-common-margin-of-victory-in-the-nfl.aspx`.

[5] Mike Clay. Defining 'Garbage Time' — PFF News & Analysis — PFF, 2012. URL `https://www.pff.com/news/defining-garbage-time`.

[6] Ed Feng. The football analytics resource guide – the top 9 killer articles, 2020. URL `https://thepowerrank.com/top-analytics-articles/`.

[7] Vilfredo Pareto. The circulation of elites. In *Talcott Parsons, Theories of Society; Foundations of Modern Sociological Theory, 2 Vol.*, pages 551–557. The Free Press of Glencoe, Inc., 1961. URL `https://archive.org/stream/theoriesofsociet01pars{\#}page/550/mode/2up`.

[8] Bryan DeArdo. Is the Super Bowl hangover real? How past losers have fared next season, what it means for the 49ers - CBSSports.com, 2020. URL `https://www.cbssports.com/nfl/news/is-the-super-bowl-hangover-real-how-past-losers-have-fared-next-season-what-it-means-for-`

[9] Nick Schwartz. The average career earnings of athletes across America's major sports will shock you — For The Win, 2013. URL `https://ftw.usatoday.com/2013/10/average-career-earnings-nfl-nba-mlb-nhl-mls`.

[10] Michael C Purucker. Neural network quarterbacking. *IEEE Potentials*, 15(3):9–15, 1996.

[11] Joshua Kahn. Neural network prediction of nfl football games. *World Wide Web electronic publication*, pages 9–15, 2003.

[12] John A David, R Drew Pasteur, M Saif Ahmad, and Michael C Janning. Nfl prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports*, 7(2), 2011.

[13] Andrew D Blaikie, Gabriel J Abud, John A David, and R Drew Pasteur. Nfl and ncaa football prediction using artificial neural networks. In *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics, Denison University, Granville, OH*, 2011.

[14] Anyama Oscar Uzoma, EO Nwachukwu, et al. A hybrid prediction system for american nfl results. *International Journal of Computer Applications Technology and Research*, 4(01):42–47, 2015.

[15] Erik J Schlicht. Exploiting oddsmaker bias to improve the prediction of nfl outcomes. *arXiv preprint arXiv:1710.06551*, 2017.

[16] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[17] Beating the juice: Removing the vig from sports betting lines. `https://www.sportsbettingdime.com/guides/strategy/removing-the-vig`. Accessed: 2020-12-14.