Exploring and Predicting NFL Games

Elina Choi, Shan Lu, and Steven Moen

STAT 992

December 8th, 2020

Summary

- American football, while interesting, is extremely challenging to understand and predict
- We tried penalized regression models to improve upon previous results, with some success
- We analyzed how our predictions vary from those from sports books
- We simulated a betting strategy using the Kelly criterion

Introduction and Background

- Football is fundamental to American life and culture
- Sports analytics are in vogue
- The NFL is particularly hard to predict

Model

- 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.
  3. other variables, e.g. difference in rested days, change coach, extreme weather, spread.
- Model:

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} Y_i \log p_i - (1 - Y_i) \log(1 - p_i) + \lambda \sum_{j} \|\beta_j - \tilde{\beta}_j\|,$$

where $p_i = \frac{\exp(X_i^T \beta)}{1+\exp(X_i^T \beta)}$, $\{\tilde{\beta}_j\}_j$ are parameters estimated from previous season.

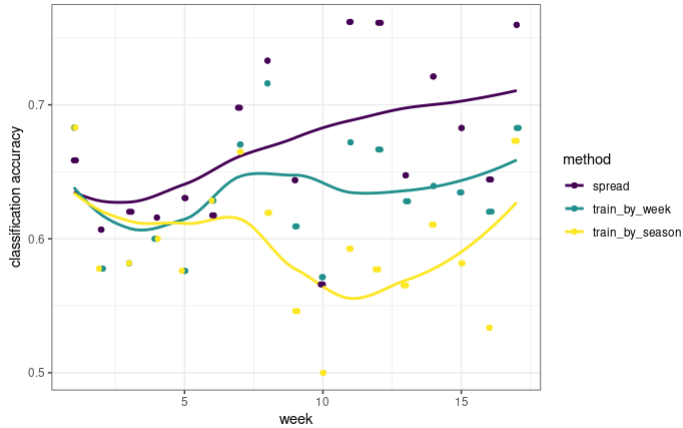
Two training strategies

- Using season i as training set, and predict game results in season i+1;
- For games in the first 6 weeks in season (i+1), predicting using model trained with data in the last season. For games in week j ($j > 6$), use the first (j-1) weeks as training set and season i the parameter as target in the penalty term.

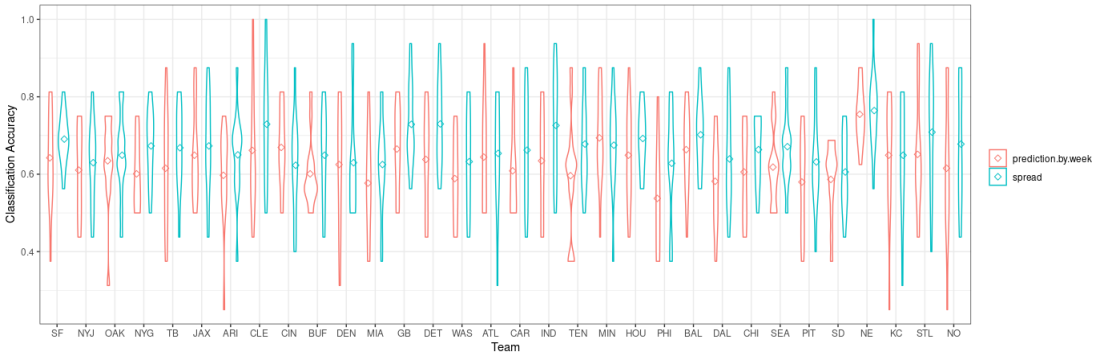
	Testing accuracy	Training accuracy
spread	67%	
Bradley Terry	59%	74%
Penalized GLM (seasonly)	59%	72%
Penalized GLM (weekly)	63%	

Table: Model comparison

Data Analysis



Data Analysis



Data Analysis

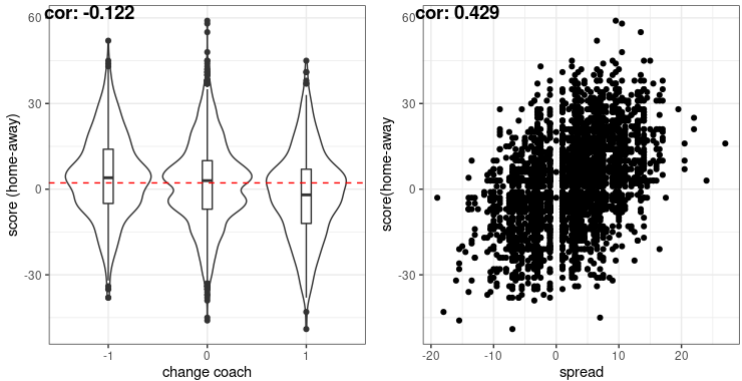


Figure: Related variables

How Our Best Model Compared to Sports Books

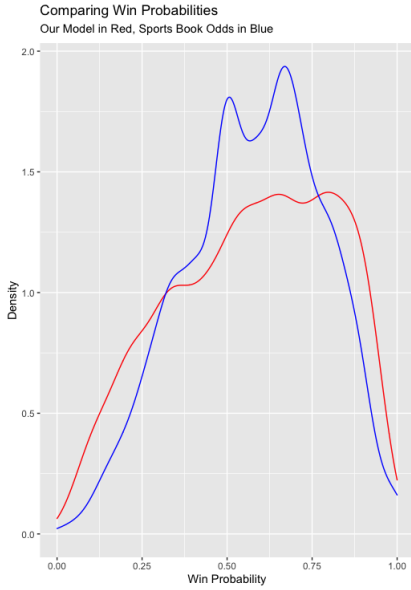
- How do we compare our predicted probabilities to those from the sports book?
- It's necessary to convert a spread to a probability to compare the two, which can be difficult - luckily, a website analyzing about 40 years of NFL data was available

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

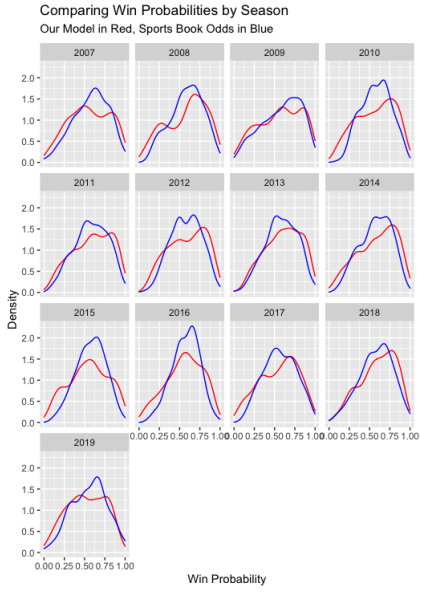
Table: How to Convert Point Spread to Odds

[0]https://www.boydsbets.com/nfl-spread-to-moneyline-conversion/

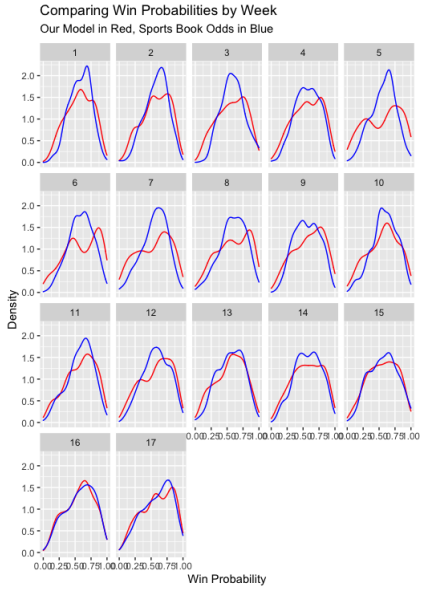
How Our Best Model Compared to Sports Books



How Our Best Model Compared to Sports Books



How Our Best Model Compared to Sports Books



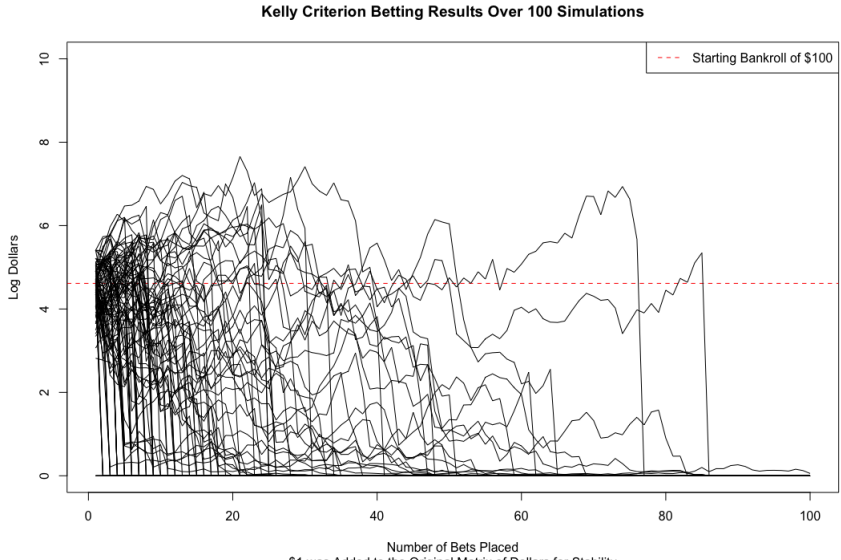
Using our Best Model To Predict Games

- Given our model, can we bet on games?
- One way to determine how to bet is the Kelly criterion, described below, where 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 spreadline probability calculated earlier $b = \frac{1-p_{spread}}{p_{spread}}$
- The Kelly criterion has been demonstrated to maximize expected wealth over the long run, but...

The Probability of Ruin is High



Next Steps

- Explore other options for sparse matrix completion (such as fastadi)
- Find alternative data sources to predict outcomes
- Broaden horizons to other prediction methods (neural nets, random forests, etc.)

Questions and Comments?

Appendix

Notes on the Sports Books

- There is missing moneyline data
- While we could fit a parametric model to this, the data seems to suggest that this fit is tricky
- Using a value of 0.999999 is necessary for games with a large spread because there was missing data after 16.5

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

The Bradley-Terry Model

The Bradley-Terry model (Bradley and Terry 1952) 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, if we set $x_{i1} = I_{\{\text{team i at home}\}}$, then β_1 expresses home team advantage

The model is estimated through maximum likelihood, and is implemented with the R package BradleyTerry2.

Data Used

- We utilized the NFL data from the 2006 through the 2019 regular seasons
  1. Included is information about game date, game time, location, weather, stadium type, coaches, game type (i.e. playoffs, regular season), team score, etc.
  2. For our exploratory analysis, we will model the game result using only team name and home/away information
- The data is structured as a time series. In order to come up with the best prediction model, we initially investigated the temporal structure of the given data

Visualize the NFL game dataset

gameday	home_team	away_team	result	point_spread	...	home_win	home_spread
2006-09-07	PIT	MIA	11	1.5	...	1	1
2006-09-10	CAR	ATL	-14	4.5	...	0	1
2006-09-10	CLE	NO	-5	3.0	...	0	1
2006-09-10	DET	SEA	-3	-6.0	...	0	0
2006-09-10	HOU	PHI	-14	-6.0	...	0	0
2006-09-10	KC	CIN	-13	-1.5	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table: NFL game dataset

Temporal correlation between seasons

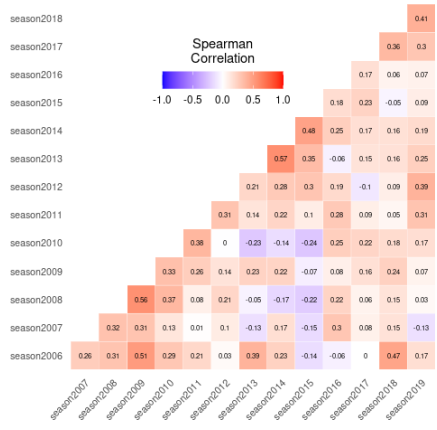


Figure: Rank correlation of Bradley-Terry score between seasons

The average correlation between year(i) with year (i-1, i-2, i-3) are 0.35, 0.24, 0.12.

Prediction Accuracy

Based on the Spearman correlation, we separated the dataset by season, trained the Bradley-Terry model on each season from 2006 through 2018, and then tested the fitted model using the next season's data.

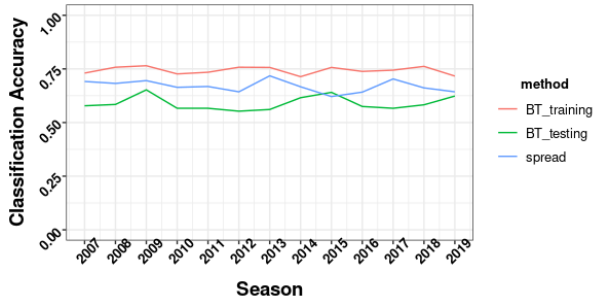


Figure: Classification accuracy for Bradley-Terry model and spread line

References

- Bradley, Ralph Allan, and Milton E Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." Biometrika 39 (3/4): 324–45.
- Clay, Mike. 2012. "Defining 'Garbage Time' | PFF News & Analysis | PFF." https://www.pff.com/news/defining-garbage-time.
- DeArdo, Bryan. 2020. "Is the Super Bowl hangover real? How past losers have fared next season, what it means for the 49ers - CBSSports.com." https://www.cbssports.com/nfl/news/is-the-super-bowl-hangover-real-how-past-losers-have-fared-next-season-what-it-means-for-the-49ers/.
- Feng, Ed. 2020. "The football analytics resource guide – the top 9 killer articles." https://thepowerrank.com/top-analytics-articles/.
- Langager, Chad. 2014. "What is the Most Common Margin Of Victory In The NFL? - SportingCharts.com." https://www.sportingcharts.com/articles/nfl/what-is-the-most-common-margin-of-victory-in-the-nfl.aspx.
- Ojha, Jay. 2020. "The 10 biggest sports leagues in the world by revenue | Pledge SportsPledge Sports." https://www.pledgesports.org/2020/02/the-10-biggest-sports-leagues-in-the-world-by-revenue/.
- Pareto, Vilfredo. 1961. "The circulation of elites." In Talcott Parsons, Theories of Society; Foundations of Modern Sociological Theory, 2 Vol., 551–57. The Free Press of Glencoe, Inc. https://archive.org/stream/theoriesofsociet01pars#page/550/mode/2up.
- Rodenberg, Ryan. 2020. "The United States of sports betting - Where all 50 states stand on legalization." https://www.espn.com/chalk/story/_/id/19740480/the-united-states-sports-betting-where-all-50-states-stand-legalization.
- Schwartz, Nick. 2013. "The average career earnings of athletes across America's major sports will shock you | For The Win." https://ftw.usatoday.com/2013/10/average-career-earnings-nfl-nba-mlb-nhl-mls.
- Witherspoon, Andrew. 2019. "NBA three-pointers are leading the sports analytics revolution - Axios." https://www.axios.com/three-pointers-lead-sports-analytics-revolution-b5613e67-92fe-44a3-897e-ede6780f0edb.html.