stevenmoen / **stat_992_project**

This is where we will store our codes for our STAT 992 project.

0 stars 0 forks

Star Unwatch

Code Issues Pull requests Actions Projects Wiki Security

master

stevenmoen Update README.md ... 4 hours ago 13

View code

README.md

Exploring and Predicting NFL Games Using the Bradley-Terry Model

by Elina Choi, Shan Lu, and Steven Moen for STAT 992

1 Summary

Gridiron football is an integral part of American culture and life, yet the game has proven challenging to understand and predict for a variety of reasons. As statisticians, we seek to better understand this game and see if we can make more informed predictions about the outcomes using techniques that we have learned in STAT 992, taught at UW-Madison by Professor Karl Rohe in the Fall of 2020.

We looked at NFL data from the 2006 through 2019 regular seasons and showed that the Spearman correlations decay quickly between seasons. Thus, we decided to use only one season of training data to fit the Bradley-Terry model, which is a standard model for predicting wins or losses for teams in a tournament-style setting. Our results are mixed in that while the predictions are better than mere chance, they are worse than those predictions implied by the spread line seen in sports books. However, the only covariate we included for this initial pass was whether the teams playing were the home or the away team, and as such, adding additional covariates as well as finding ways to incorporate more recent data may improve our prediction quality. Also, using techniques to incorporate more historical data may also improve our prediction quality.

2 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 (Ojha 2020). 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 (Witherspoon 2019). 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 (Rodenberg 2020). 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 (Langager 2014), 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 (Clay 2012). 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 (Feng 2020), 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 (Pareto 1961), 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 (DeArdo 2020). Also, given the high injury risks associated with football, players tend to have shorter careers compared to their Big 4 peers (Schwartz 2013) and can have a promising season cut short or hobbled with an injury.

One might look at these grim realities and give up - how is it possible to understand America's game? There is hope, and it comes in the form of the Bradley-Terry model.

2.1 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

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

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

The parameter β_{i0} quantifies the ability of team *i*. If $\beta_{i0} > \beta_{j0}$, then team *i* has higher chances of beating team *j*. $\{x_{ir}\}_r$ are team specific variables that may influence game results. For example, if we set $x_{i1} = 1_{\{\text{team } i \text{ is at home}\}}$, then β_1 expresses home team advantage. The model is estimated through maximum likelihood, and is implemented with the R package BradleyTerry2.

3 Results

We have NFL data from the 2006 through the 2019 regular seasons, downloaded from Github folder Data source. The data set contains detailed information, including information about game date, game time, location, weather, stadium type, coaches, game type (i.e. playoffs, regular season), team score, etc. However, for our exploratory analysis, we will model the game result using only team name and home/away information.

3.1 Load data

```
dat = read.table('games.csv', sep=',', header=T,  row.names = NULL, fill=T, quote=''
dat = dat[dat$season>=2006 &dat$season<2020,]

# remove playoff games
dat = dat[dat$game_type=="REG",]
# remove games with ties
dat <- dat[dat$result!=0,]
# Teams change name in season 2016 and 2017. We use their old names for consistent c
dat$home_team[dat$home_team=="LA"]="STL"
dat$away_team[dat$away_team=="LA"]="STL"
dat$home_team[dat$home_team=="LAC"]="SD"
dat$away_team[dat$away_team=="LAC"]="SD"
```

3.2 Check temporal correlation

The data is structured as a time series. We expect that the "ability" of teams varies through time, because of changes in players, coaches, injuries, etc. Therefore, it is better to train a predictive model using more recent data, rather than using all of the historical data. A common technique to inspect the temporal structure is through autocorrelation. For each season, we fit a Bradley-Terry model, and then compute the Spearman correlation of the estimated team score between each pair of seasons. The correlation decays rapidly between years, with a median value of 0.35 between consecutive years, which supports our decision to use a limited amount of historical data for the training model. Thus, we have decided to train the Bradley-Terry model using only data from the previous season.

```
### fit BT by season ###
# BT_score is a matrix storing BT scores for all seasons

BT_score = matrix(NA, nrow = length(unique(dat$home_team)), ncol=length(unique(dat$s
# Sort the team names
rownames(BT_score)=sort(unique(dat$home_team))
# Organize the seasons
colnames(BT_score)=paste0('season',min(dat$season): max(dat$season))


# team ARI is used as control
BT_score['ARI',] = 0

# Loop through for all seasons in the data frame
for (i in min(dat$season): max(dat$season)){
  # Subset the data frame for only one season
  dat_season=dat[dat$season==i,]
  # Count a home team win if result is equal to 0
  home.wins = (dat_season$result>0)*1
  # print(home.wins)
  # Assign variables in a different way
  # Create a list of inputs for the model to run
  football = list(home.team=data.frame(team = dat_season$home_team, at.home = 1),
                  away.team=data.frame(team = dat_season$away_team, at.home = 0))
  # print(football)
  # Try running the model
  mod1 = BTm(outcome = home.wins, player1 = home.team, player2 = away.team,
             formula = ~ team + at.home, id = "team",data = football, family =binomi
  # Extract the model coefficients
  coef=mod1$coefficients[1:(length(mod1$coefficients)-1)]
  # Stores the coefficients in a list
  BT_score[unlist(lapply(names(coef),function(i){substr(i, 5, nchar(i))})),i-min(dat
}


BT_score = BT_score[complete.cases(BT_score),]
cormat = cor(BT_score, method='spearman')
```

```
diag(cormat)=NA

# Reorder the correlation matrix
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}
upper_tri <- get_upper_tri(cormat)
# Melt the correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)
# Create a ggheatmap
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Spearman\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 8, hjust = 1),
        axis.text.y = element_text(size = 8))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = round(value,2)), color = "black", size = 2) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
```

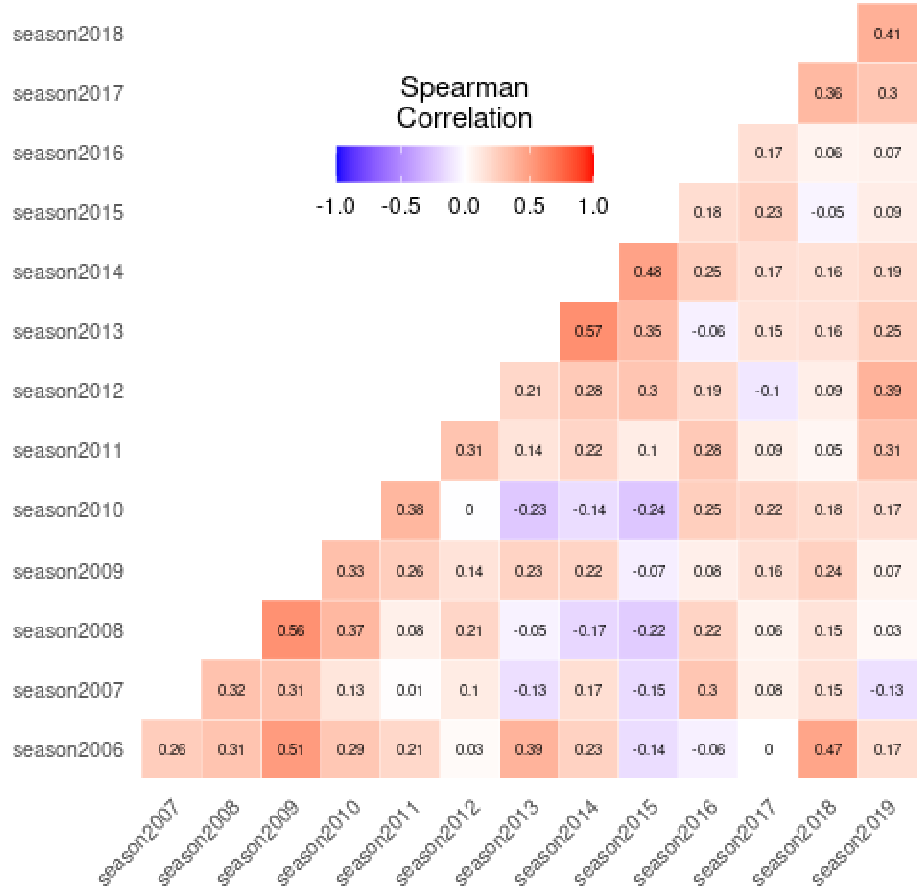

Figure 1: Spearman correlation of the team Bradley-Terry score between each pair of seasons

3.3 Fit BT model and compare the prediction accuracy with spread line

Based on our findings with 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. As shown in Figure 2, the Bradley-Terry model has an average prediction accuracy of about 0.59, or 59%. While this is certainly better than a random guess, it performs worse than what the spread line would predict with an average accuracy of 0.67 or 67%.

```
season = (min(dat$season)+1):max(dat$season)
# test_acc if a matrix storing testing set classification accuracy of Bradley-Terry
test_acc = matrix(0, nrow=length(season), ncol=2)
colnames(test_acc) = c('Bradley_Terry', 'spread_line')
rownames(test_acc) = paste0('season_', season)

for (i in season){
  # Training set
```

```
    dat_train=dat[dat$season==i-1,]
    home.wins = 1*(dat_train$home_score> dat_train$away_score)
    football_train = list(home.team=data.frame(team = dat_train$home_team, at.home = 1
                    away.team=data.frame(team = dat_train$away_team, at.home = 0))
    mod1 = BTm(outcome = home.wins, player1 = home.team, player2 = away.team,
               formula = ~ team + at.home , id = "team", data = football_train, family
    # Testing set
    dat_test=dat[dat$season==i,]
    football_test = list(home.team=data.frame(team = dat_test$home_team, at.home = 1),
                    away.team=data.frame(team = dat_test$away_team, at.home = 0))

    test_acc[i-min(season)+1,]=c(
    mean((predict(mod1, newdata=football_test)>0)*1 == (dat_test$result>0)*1),
    mean((dat_test$spread_line>0)*1 == (dat_test$result>0)*1)
    )
  }
  # reshape the test_acc to plot
  test_acc_long = gather(data.frame(test_acc, season=rownames(test_acc)), method, accu
  test_acc_long$season = as.integer(unlist(lapply(test_acc_long$season, FUN=function(i
  ggplot(test_acc_long, aes(x=season, y=accuracy, color=method)) + theme_bw()+
      geom_line()+theme(legend.title = element_blank())+ylab('Classification Accuracy'
```

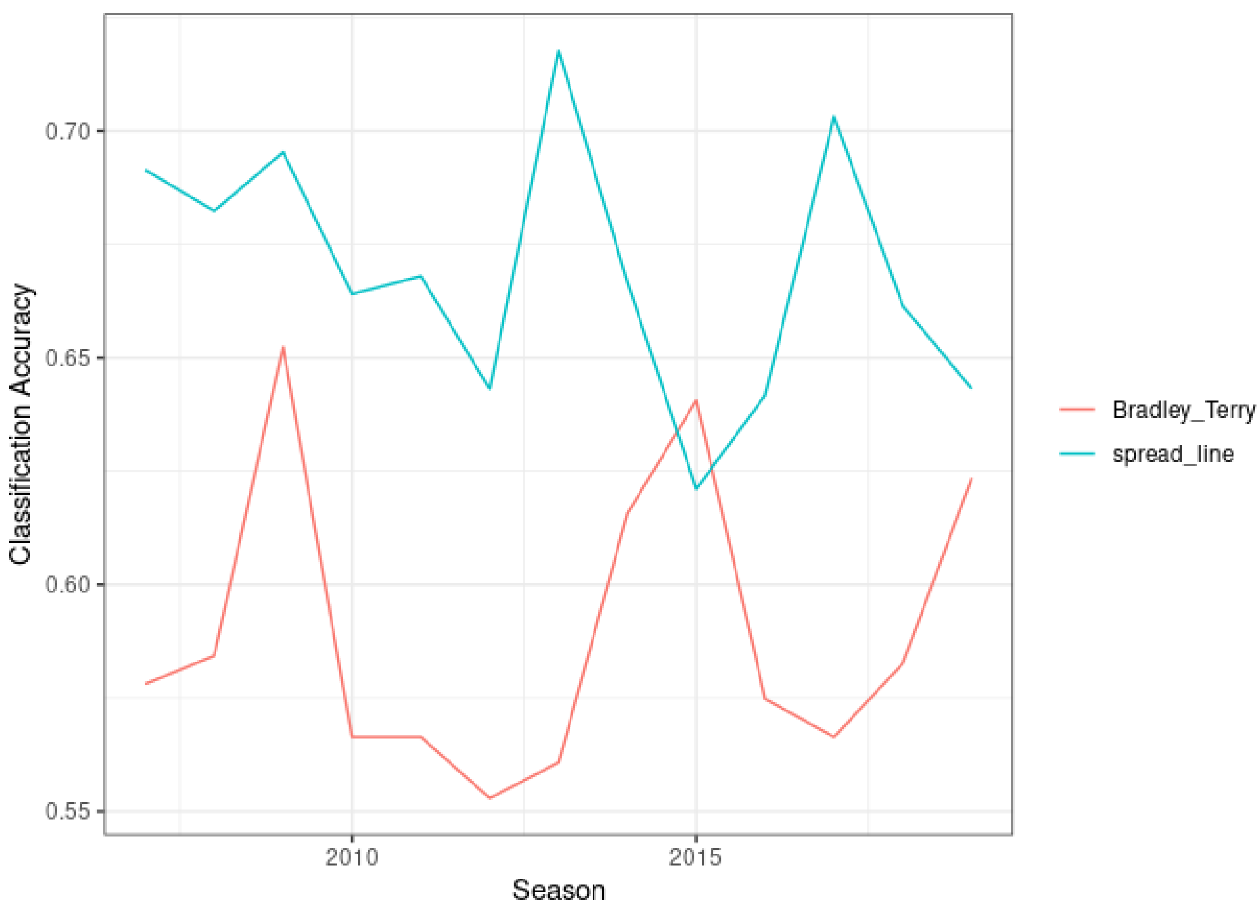

Figure 2: Classification accuracy for Bradley-Terry model and spread line

4 Discussion

After exploring the data with a simple version of the Bradley-Terry model, we have a few directions to improve the model:

- Formulate the Bradley-Terry model in a way that can adjust for ties. While these ties are not particularly common, it is a way to improve the quality of the forecast.
- Train the model in an online fashion, instead of using large batches.
- Take more explanatory variables into account to see if there are areas where the B-T model produces better predictions than do the betting odds.
- Extend the model to account for the temporal dynamics so that we can train with all the historical data.

5 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.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 2


stevenmoen Steven Moen

shanlu01 Shan Lu

Languages

- **TeX** 100.0%