Final project: NFL by half
Advanced Methods III 140.753

1 Introduction

Have you ever wondered if you should keep watching a football game beyond half-time? You are probably having a busy weekend and need to do other things. Specially if the team that you like has a high probability of winning (or losing if they are doing terrible).

Analyzing NFL data and variables has been done before. For example, Carney and Fenn [10] were interested in identifying variables that affect the Nielsen TV rating of the games. Some of the variables they looked at include the game day winning percentage of the two teams involved and the season week. In their analysis, one important variable was whether the local team had gone to the playoffs the previous year which makes sense for explaining the TV rating.

Brian Burke who describes himself as *a former Navy pilot who has taken up the less dangerous hobby of N.F.L. statistical analysis, operates Advanced NFL Stats, a blog about football, math and human behavior* [7] has published many blog posts where he analyzes NFL data. He has devised *narrative statistics* such as a *winning probability model* —it's quite accurate [9]— which he uses to determine the effect specific plays have on the game [8]. Another one that has similar uses is the *expected points* [2] and *expected points added*. By using the *win probability model* he can estimate how exciting the game was [1].

Of special interest, Brian Burke has used logistic regression in a rather exquisite way to predict the probability of winning for each team during the NFL season [3–5, 7].

The goal of this small project is to use the play-by-play data for the first half of NFL games to predict which team will win. To do so a modification of Brian Burke's game probability model [3–5, 7] will be implemented. In addition, part of the goal is to deploy the resulting prediction model on the web so it can be used for the 2013 season.

2 Pre-processing

Brian Burke has compiled NFL play-by-play data for the 2002 to 2012 seasons [6], which I simply downloaded. Processing it was quite another story as play descriptions can be convoluted. Once I sorted out the type of plays, I proceeded to calculate the following variables.

For each season and each team, I got the following information similar to what Burke has described [3].

oPassYdsAtt Net offensive passing yards per attempt. It's the sum of passing yards minus the sack yards, then divided by the number of passing plays (complete passes, interceptions, incomplete passes and sacks).

oInt Offensive interceptions per attempt.

oRun Running success rate. The number of running plays in which the down-distance-togo improved divided by the number of running plays.

oFumble Offensive fumbles per attempt. The number of fumbles divided by the number of passing and running plays.

pen Penalty yards per play.

dPassYdsAtt Defensive net passing yards per attempt. Similar to *oPassYdsAtt*.

dRunAtt Defensive running yards per attempt. Running yards allowed divided by the number of plays the opponent ran.

dInt Defensive interceptions per attempt.

For each season, I then built a data set for model training. It has two rows per game, one where team A is the local team and one where team A is the visiting team as inspired by Burke [3].

- team A's season stats.
- team B's season stats.

local Whether team A was the local team.

win Whether team A won the game.

halfdiff The half-time difference in score. A positive value means team A was winning.

date Date of the game.

resumes Whether team A has the first drive of the second half.

gameA Which game of the season is for team A?

gwrA Game winning percentage for team A in 0 to 1 scale. 0 is used for week 1.

gameB Same as *gameA* but for team B.

gwrB Same as *gwrA* but for team B.

The information can then be easily combined across different seasons.

Finally, I built a data set for evaluating the model's prediction accuracy using the 2012 regular season data. This is similar to the one described above but instead of using team A's and team B's season stats, the stats are calculated per game.

3 Exploratory Data Analysis

3.1 First vs second half

Exploring the differences between the team statistics per season between the first and second half did not reveal major discrepancies. Although some variables have interesting trends as shown in figure 1 where the offensive net passing yards per attempt has become much less variable since 2006 while the inverse is true for the defensive interceptions per play.

After grouping the training data for the 2002 to 2011 seasons, it becomes evident that seasons from 2002 to 2005 are different from 2006 onward. This can be seen in figure 2 when comparing the date versus team A's offensive net passing yards for the first half. This pattern repeats itself with nearly all the variables in the training data set[1]. Thus, when training the models only data from 2006 to 2011 will be used. However, note that there in figure 2 there is a very weak relation between date and win status.

It is also important to note in figure 2 that the score at half-time is the most associated to the win status.

[1]Data not shown but is available at `https://github.com/lcolladotor/lcollado753/tree/master/final/nfl_half/EDA/initial`.

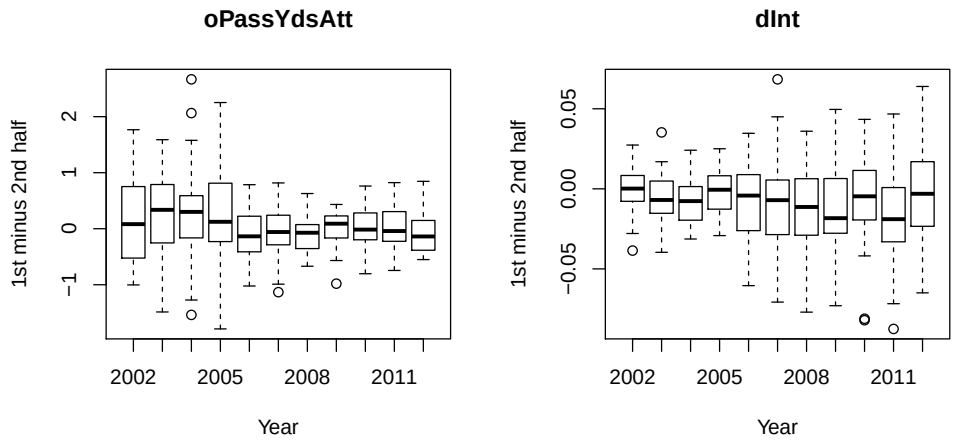

Figure 1: Difference between the first and second half for the offensive net passing yards per attempt (left) and defensive interceptions per play (right). Each year has 32 observations (1 per team).

3.2 Training model

Using logistic regression, a model with all the variables was trained using the data from 2006 to 2011. Single term deletions were explored[2] clearly show that the half-time score is the most important variable as was already explored previously. It is followed by the game day winning percentage for both teams (*gwrA, gwrB*).

Using step-wise AIC selection, the obtained model is the one shown in table 1. In particular, this model is different from the one used by Brian Burke [3, 4] as it considers less team variables but adds the half-time score difference, the game day winning percentages and an indicator for who starts the second half.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.3E-01	1.6E+00	-2.7E-01	7.9E-01
teamAoPassYdsAtt	-1.7E-01	1.2E-01	-1.5E+00	1.4E-01
teamAoRun	6.4E+00	1.8E+00	3.5E+00	5.4E-04
teamAdPassYdsAtt	1.9E-01	1.2E-01	1.6E+00	1.1E-01
teamAdRunAtt	-3.6E-01	9.4E-02	-3.9E+00	1.0E-04
teamBoRun	-5.4E+00	1.8E+00	-3.0E+00	2.6E-03
teamBoFumble	2.5E+01	1.1E+01	2.3E+00	2.2E-02
teamBdRunAtt	3.1E-01	9.3E-02	3.4E+00	7.2E-04
localTRUE	2.2E-01	9.1E-02	2.4E+00	1.7E-02
halfdiff	1.5E-01	6.0E-03	2.4E+01	2.3E-132
resumesTRUE	2.5E-01	9.1E-02	2.8E+00	5.1E-03
gwrA	8.0E-01	1.7E-01	4.7E+00	2.4E-06
gwrB	-9.0E-01	1.7E-01	-5.2E+00	1.9E-07

Table 1: Logistic regression for model trained with 2006 to 2011 data. This model was selected using step-wise AIC selection.

[2]Data not shown but available at `https://github.com/lcolladotor/lcollado753/blob/master/final/nfl_half/EDA/model/`.

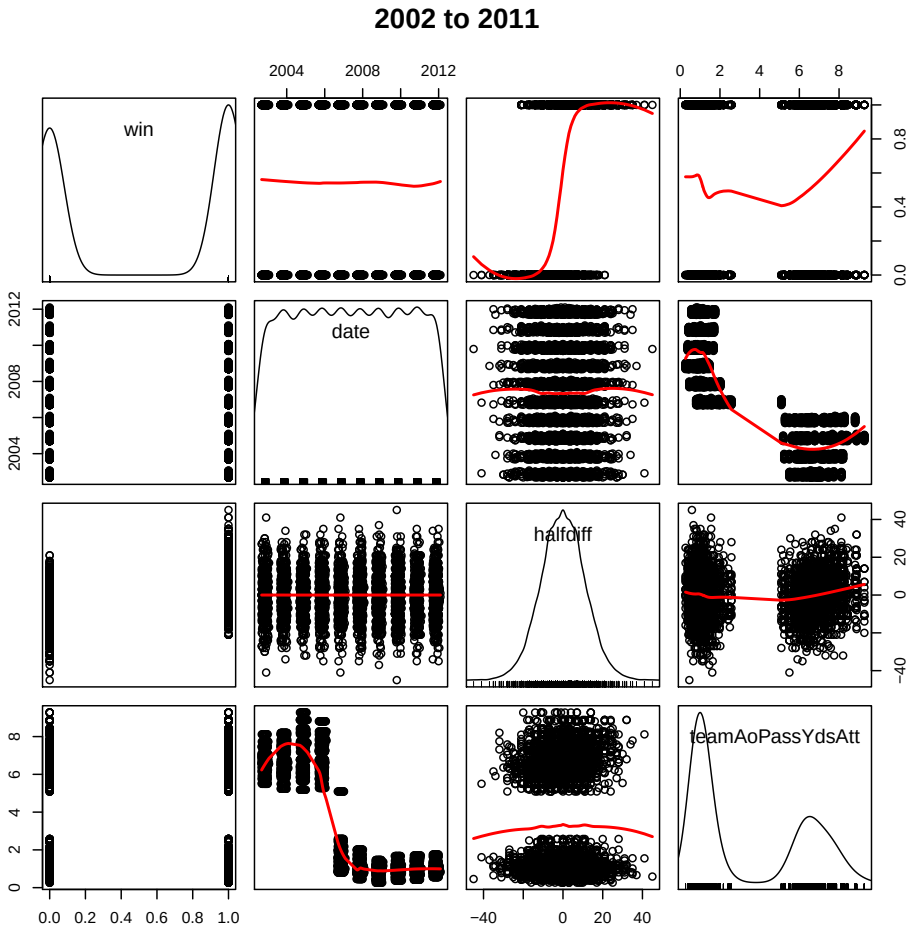

Figure 2: Win status, date, score at half-time and team A's net offensive passing yards shown in a scatterplot matrix. The red lines are smooth lines whose purpose is to illustrate the relation between the variables. Diagonal entries show density plots along with ticks. Data is from the training data for 2002 to 2011 seasons.

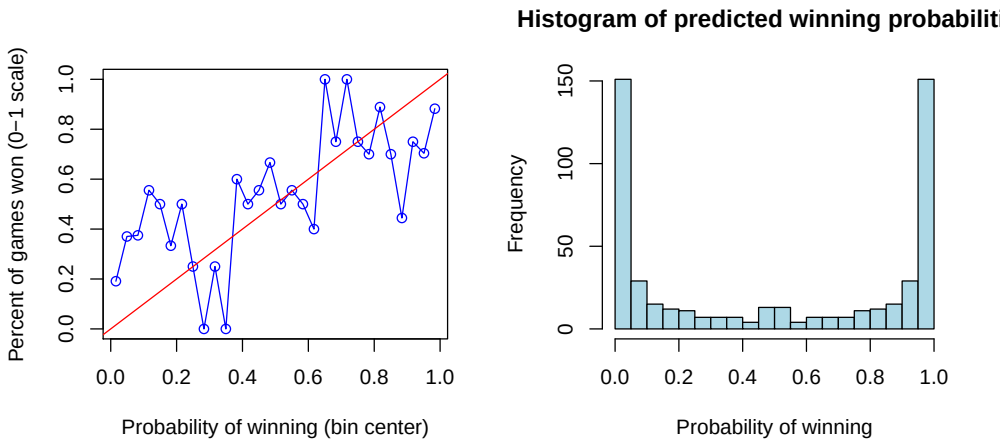

Figure 3: Evaluation of the prediction results for all games in 2012 using model trained on data from years 2006 to 2011. (Left) Predictions are binned, and the actual percent of games won (0 to 1 scale) is calculated and shown in blue. Red line shows the 45 degree line. (Right) Histogram of predicted probabilities of winning using 30 breaks. (Both) Predictions for each team in a game are used.

4 Results

For each game in the 2012 season, the model was used to get a predictor in the logit scale for both teams. Then, the difference of logits was inverted by the inverse logit to get the probability of winning for team A and team B similarly to what Burke does [3]. In more detail, the model is used to get:

$$\eta_A = logit(pr(team_A wins)) = x_{team A}\hat{\beta}, \quad \eta_B = logit(pr(team_B wins)) = x_{team B}\hat{\beta} \tag{1}$$

$$p_A = logit^{-1}(\eta_A - \eta_b), \quad p_B = 1 - p_A \tag{2}$$

Note that $\eta_A - \eta_B$ is the log odds ratio of team A winning over team B. Thus the inverse logit gives the probability of team A winning. By calculating the probability of team A winning in this way, instead of just taking the inverse logit of the model prediction, we guarantee that the probability of team A winning plus the one for team B equals 1. Otherwise this is not guaranteed.

Notably, this model is much simpler than the one used by Burke [4] where he uses iterations of comparing a given team versus the league average to adjust for the strength of each time. The relative strength is simplified in this model by considering the game day winning percentage of both teams ($gwrA$, $gwrB$).

Figure 3 (left) evaluates the performance of the model by using 30 bins. While the correspondence to the real winning percentages wiggles around the diagonal line, it is also important to consider that most of the predictions are closer to 0 and 1 as shown in 3 (right). Thus, there is much less data in the middle of 3 (left) and this can explain the extra variability seen.

5 Deployment on the web

Having identified the variables to use for prediction, a second model was trained with the data from 2006 to 2012 in order to be able to use it when predicting 2013 games. Both the model for predicting 2012 and the model for predicting 2013 have been implemented in a web application

using Shiny [11]. The Shiny code is available at `https://github.com/lcolladotor/lcollado753/tree/master/final/nfl_half/shiny`. However, it is best to see it live which you can do using the following R commands:

```
## This is how you can run the Shiny app
library(shiny)
runUrl("https://github.com/lcolladotor/lcollado753/archive/master.zip",
    subdir = "final/nfl_half/shiny/")
```

The web application shows the prediction, the model information (similar to table 1), diagnostic plots, and lets you download the specified values for future use. Furthermore, it is set with mean values and has sliders with sensible limits. However, it currently requires the user to input the data for the prediction. Future improvements would include live-scrapping the NFL play-by-play data.

6 Conclusions

Notably the game changed at the start of the 2006 season. While this limits the available data for model training, it is certainly possible to build a predictive model using play-by-play data from the first half to predict which team will win at the end of the game. If you are someone who will watch a game depending on how certain it is that a given team will win the match, then you can use the model and the web application to help you make that decision. You make the final decision depending on a given cutoff of your choosing. Plus you might be willing to considering more certain games in certain specific situations.

The accuracy of the model is acceptable, although finer tuning might help and a detailed comparison could be made versus alternative models [3, 4]. Furthermore, the web application is not yet fully user-friendly to use for predicting 2013 games as you need to either track the required data or estimate them from other sources such as the the television half-time summaries.

7 Reproducibility

The code, data, and report is available at GitHub. Specifically here: `https://github.com/lcolladotor/lcollado753/tree/master/final/nfl_half`. The README file explains the order of the scripts.

This report was generated using the following R packages.

- R version 2.15.2 (2012-10-26), `x86_64-apple-darwin9.8.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: car 2.0-16, knitr 1.1, MASS 7.3-23, nnet 7.3-5, xtable 1.7-1
- Loaded via a namespace (and not attached): digest 0.6.3, evaluate 0.4.3, formatR 0.7, stringr 0.6.2, tools 2.15.2

References

[1] B. Burke. *Best Games of 2012 and Best Playoff Games*. URL: `http://www.advancednflstats.com/2013/01/best-games-of-2012-and-best-playoff.html` (visited on 03/17/2013).

[2] B. Burke. *Expected Points (EP) and Expected Points Added (EPA) Explained*. URL: `http://www.advancednflstats.com/2010/01/expected-points-ep-and-expected-points.html` (visited on 03/17/2013).

[3] B. Burke. *How the Model Works–A Detailed Example Part 1.* URL: http://www.advancednflstats.com/2009/01/how-model-works-detailed-example.html (visited on 03/17/2013).

[4] B. Burke. *How the Model Works–A Detailed Example Part 2.* URL: http://www.advancednflstats.com/2009/01/how-model-works-detailed-example-part-2.html (visited on 03/17/2013).

[5] B. Burke. *N.F.L. Week 4: Game Probabilities Are Back.* URL: http://fifthdown.blogs.nytimes.com/2012/09/27/n-f-l-week-4-game-probabilities-are-back/ (visited on 03/17/2013).

[6] B. Burke. *Play-by-Play Data.* URL: http://www.advancednflstats.com/2010/04/play-by-play-data.html (visited on 03/17/2013).

[7] B. Burke. *Week 4 Game Probabilities, From Advanced N.F.L. Stats.* URL: http://fifthdown.blogs.nytimes.com/2009/09/30/advanced-nfl-stats-week-4-game-probabilities/ (visited on 03/17/2013).

[8] B. Burke. *Win Probability Added (WPA) Explained.* URL: http://www.advancednflstats.com/2010/01/win-probability-added-wpa-explained.html (visited on 03/17/2013).

[9] B. Burke. *Win Probability Model Accuracy.* URL: http://www.advancednflstats.com/2009/07/win-probability-model-accuracy.html (visited on 03/17/2013).

[10] S. Carney and A. Fenn. *The Determinants of NFL Viewership: Evidence from Nielsen Ratings.* SSRN Scholarly Paper ID 611721. Rochester, NY: Social Science Research Network, Nov. 2004. URL: http://papers.ssrn.com/abstract=611721 (visited on 03/18/2013).

[11] RStudio and Inc. *shiny: Web Application Framework for R.* R package version 0.4.0. 2013. URL: http://CRAN.R-project.org/package=shiny.