Using Random Forests to Estimate Win Probability Before Each Play of an NFL Football Game

Dennis Lock

Dan Nettleton

New England Symposium on Statistics in Sports

Introduction

- Idea
  - At any specific moment of an NFL game, find the probability a particular team will ultimately win the game.
    - For example, what’s the probability a team receiving the ball on the 20 yard line down 3 with 2 minutes left will go on to win the game?
    - We combine pre-play variables to estimate win probability (WP) at any moment in an NFL game using a random forest methodology.

Introduction

- Idea
  - Demonstrate the use of WP estimates:
    - Fan interest
      - Plot the course of a game using win probability
      - Real time win probability estimation
    - Evaluate plays and play calling decisions
      - Example: Was Baltimore's decision to take an intentional safety late in the 4th quarter of Superbowl 47 a good one?

Introduction

- Motivation
  - Develop an alternative to Brian Burke’s win probability found at www.advancednflstats.com
  - Why?
    1. Estimate WP empirically through objective “binning”.
    2. Include information measuring the quality of both teams competing.
    3. Develop a method that can be easily repeated on a new set of variables, especially in a different sport.

Random Forest Method

- Data
  - Recently (since 2000) the NFL began releasing play-by-play data from all games, regular and post season.
    - We use the seasons 2001 – 2011 as training data and the 2012 season as test data.
    - Raw play-by-play data was downloaded from: www.armchairanalysis.com

Random Forest Method

- Data
  - Observational Unit: A pre-play situation observed with respect to the offensive team.
    - Example: 1st and 10 on the 20 yard line down by 3 with 2 minutes remaining.
    - Score Difference = -3 implies the team with the ball is losing by 3.
    - Win probability is estimated for the offensive team.

Random Forest Method

- Data
  - Variables:
    - Binary Response, $y_i = I(Offense\ Won_i)$
    - Predictor variables: down, yards to go for a first down, field position, seconds remaining, score difference, adjusted score difference, total points scored, time outs remaining, and the Las Vegas point spread

$$\text{adjusted score difference} = \frac{\text{score difference}}{\sqrt{\text{seconds remaining}}}$$

Random Forest Method

- Random Forest
  - A random forest is a combination of either classification or regression trees.
    - A tree is effectively a nearest neighbors method of binning observations on values of the predictor variables in order to maximize within-bin homogeneity of the training responses.
    - We chose to use a random forest of regression trees.
      - A regression tree takes the average of the response values in a resulting bin as a predicted response for future observations in that bin.

Random Forest Method

- Random Forest
  - Each tree of the random forest has two adjustments in order to grow a variety of trees:
    1. Each tree is grown on a bootstrapped version of the original sample.
    2. At each split of the training observations, only a subset of the variables are considered as candidates for deciding the splitting rule.

Random Forest Method

- Why Random Forest?
  1. Allows for complex unknown interactions between predictor variables
     - Example: Score difference and time interact, but we don’t know how.
  2. Predictions are entirely on empirical evidence
     - Minimal dangerous assumptions

Random Forest Method

- Why Random Forest?

3. Nicely handles outliers
    - Blowout victories aren’t overly influential

4. Easily obtain variable importance measurements

5. Good predictability!

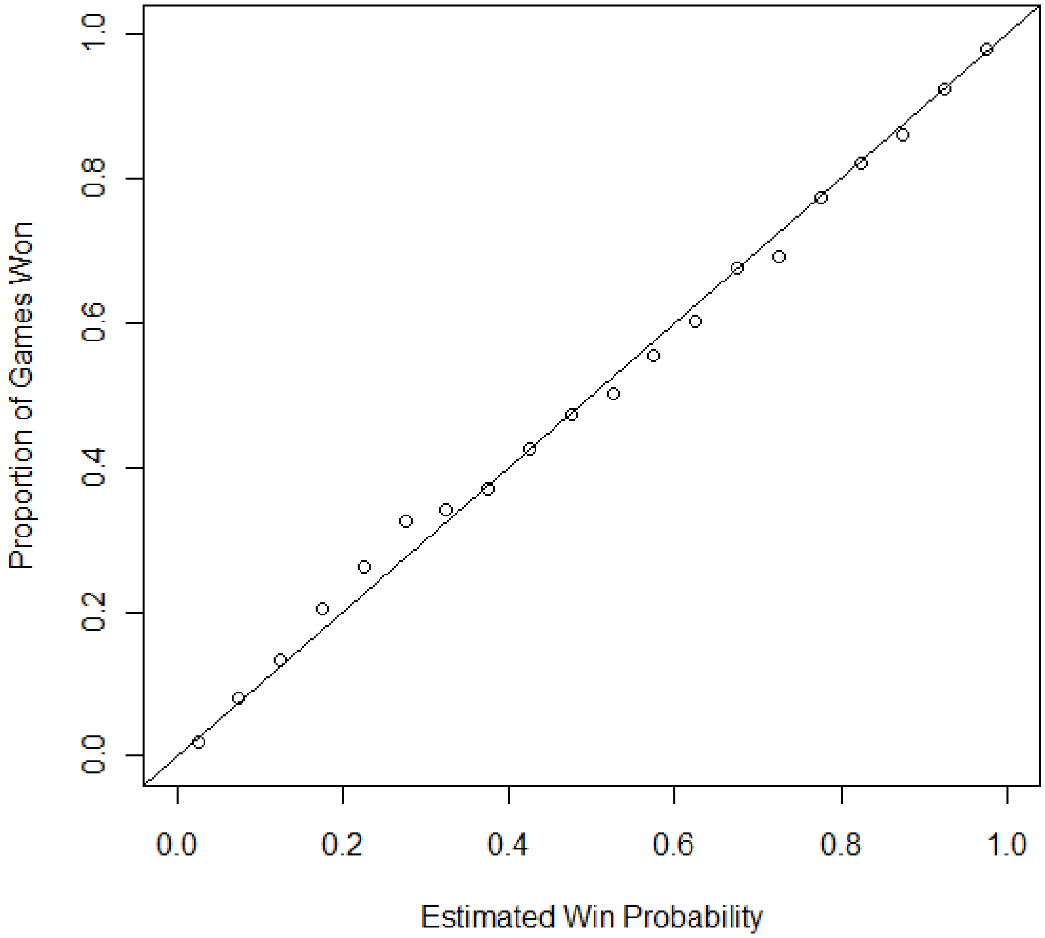
Results

- Performance
  - Test set Mean Absolute Error by quarter:

Quarter:	1	2	3	4
Error:	*0.408*	*0.346*	*0.276*	*0.199*

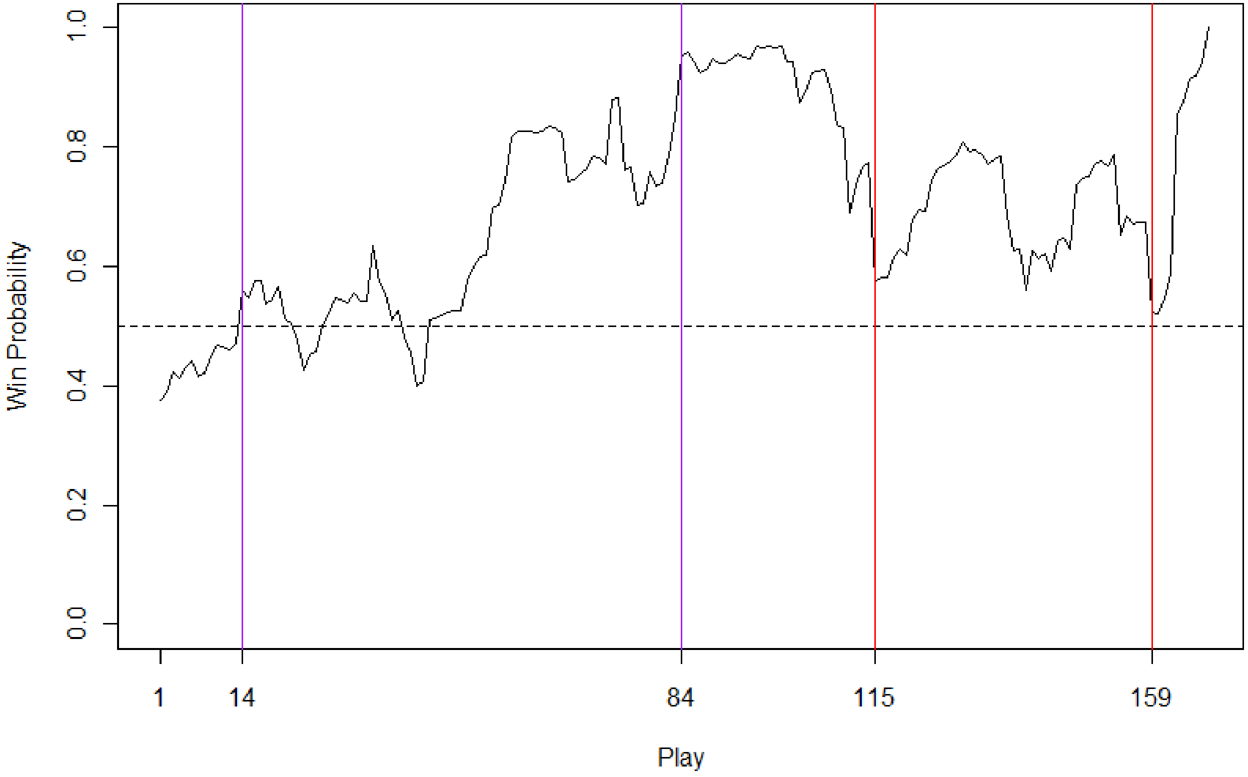
Results

- Performance



Results

- Super Bowl 47 (BAL 34 – 31 SF)



Results

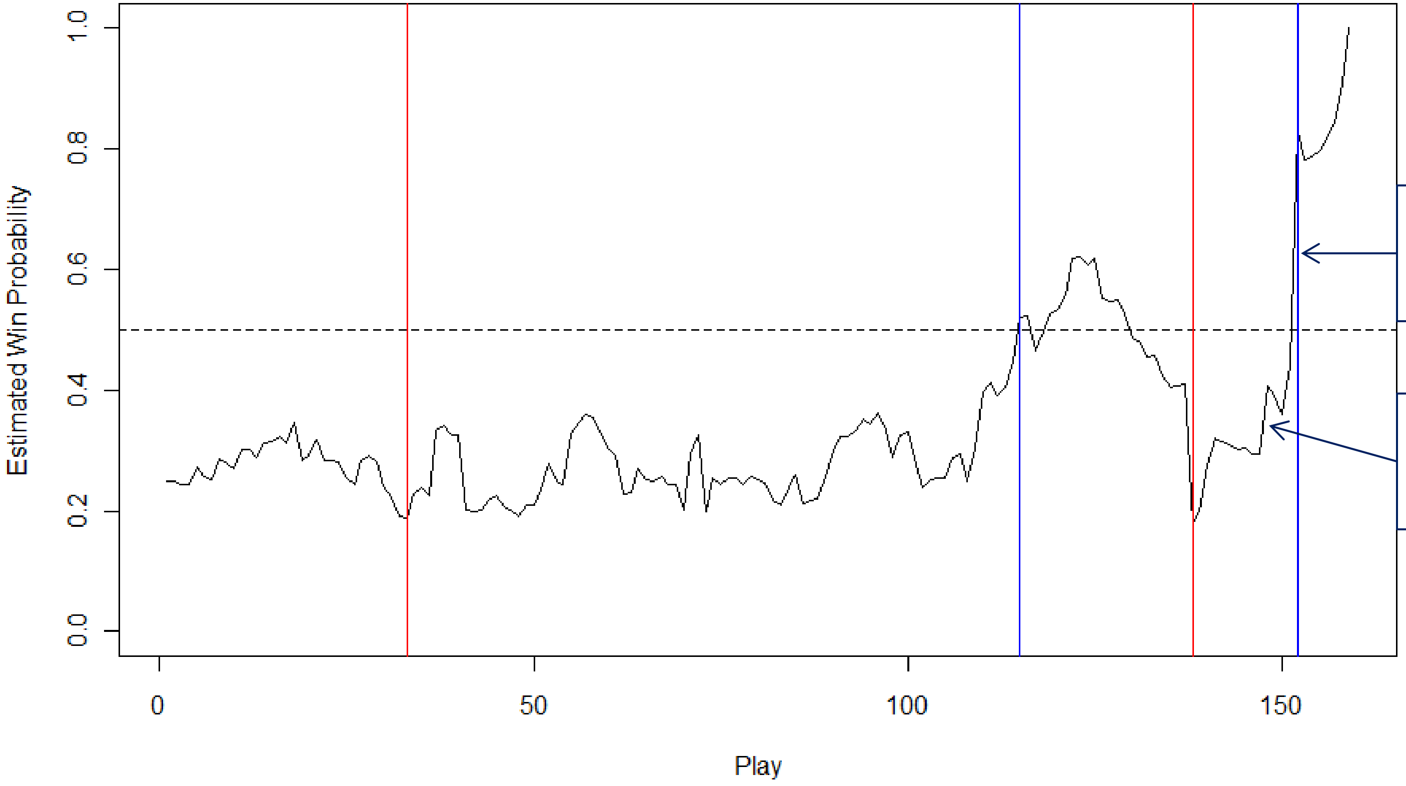
- Play Calling
  - By taking an intentional safety Baltimore increased there WP from 91.8% to 94.2%.
  - Changes in Win Probability (*ΔWP*) such as this can be used to evaluate play calling decisions.
  - For instance by kicking a surprise onside kick (successfully) in Superbowl 44, the Saints increased their win probability by 7%.

Results

- Play Calling
  - We can also use average *ΔWP* to evaluate play calling decisions, examples:
    - The average *ΔWP* for surprise onside kicks is approximately +0.02.
    - *ΔWP* and average *ΔWP* could be used to make real-time decisions on plays such as 4th down decisions.

Results

- Superbowl 42 (NYG 17 – 14 NE)



Burress Touchdown Catch (*ΔWP=0.39*)

Tyree Helmet Catch (*ΔWP=0.12*)

Results

- Influential Plays
  - We can judge the most influential plays from a set of plays (season, game, etc.) using *ΔWP.*
    - The best Super Bowl play of the last 12 years as judged by *ΔWP* was James Harrison’s 100 yard interception return before halftime in 2008 (*ΔWP*=0.510).
    - The best play of the 2012 season was a 39 yard touchdown reception by Cecil Shorts down 5 with 20 seconds remaining (*ΔWP*=0.710)

Future Considerations

- Feature of the data
  - Each game has approximately 150 sequential observations all predicting 1 response value (Won or lost).
    - Independent observations?
      - No
    - Stochastic observations?
      - Maybe not

    - We have attempted methods to account for these possible problems but none improve performance.

Future Considerations

- Other Sports
  - Extending the win probability to other sports
    - Easy in sports that have a clear “pre-play” situation like a possession in basketball or pitch in baseball.
    - May be more of a challenge in flow sports such as hockey or soccer.

Takeaways

- Two Takeaways
  - The Random Forest is a fairly simple and effective way of estimating win probability!
  - Estimated WP can have many uses.
    - *“In any sport win probability is basically the holy grail of analytics.”*

-Brian Burke

Thank You!

Email: Dennis.F.Lock@gmail.com
Website: lockanalytics.com