CS 109A/Stat 121A/AC 209A: Final Project
Protopapas, Rader, Dave, Levine

Win Probability Models in Sports

Problem statement

> *The win probability graphic/discussion on ESPN is literally taking a sword and sticking it through the chest of any fun left in baseball. – Kenny Ducey (@KennyDucey) April 2, 2017*

Win probability models (a within-game metric to predict the probability of each team winning given the circumstances of the game) have become the norm when viewing a broadcast of a professional sporting event. But most models are like black boxes for which little is known as to the factors related to the probability of winning a game (ESPN keeps their methodology secret, for example). In this project you will choose one of three professional sports: baseball (MLB), basketball (NBA), or football (NFL); collect play-by-play data, build a win probability, WP, or expected points/runs, EP, model; and use it to answer a few questions about the sport chosen. For example, which players are the best offensively or defensively as measured by Win Probability Added)? What coaches use the best (or worst) strategy in a high-leverage situation?

Key Challenges: Scraping data, parsing text, engineering features from text.

Important milestone: For milestone 3 (see Project Guidelines), you should have all the data collected, explore the data, and have a baseline win probability model completed (or expected points/runs, if applicable). More importantly, you should provide an important question that you will use your WP or EP model to answer (like ranking players in a specific way, measuring coaches decisions, adjusting models for team strengths, etc?).

Data resources

You will be required to collect *play-by-play* data for your sport of choice (NBA, NFL, or MLB). To fully collect the data, you will need to have a *schedule of games* and *play-by-play* data within each game. Note: some scraped and parsed play-by-play data is available for various sports online, but you will definitely want to engineer some of your own features and add the most recent data no matter what.

1. **MLB** (pitch-by-pitch is available, but not expected for this project)

   Example of play-by-play data (at bottom of page):
   https://www.baseball-reference.com/boxes/LAN/LAN201711010.shtml

   Schedule of games (2017):
   https://www.baseball-reference.com/leagues/MLB/2017-schedule.shtml

2. **NBA**

   Example of play-by-play data:
   https://www.basketball-reference.com/boxscores/pbp/201710200PHI.html

Schedule of games (2017-18 – please scrape up until Nov 26):
https://www.basketball-reference.com/leagues/NBA_2018_games.html

3. **NFL**

   Example of play-by-play data:
   https://www.pro-football-reference.com/boxscores/201709100was.htm#all_pbp

   Schedule of games (2017 – please scrape up until week 11):
   https://www.pro-football-reference.com/years/2017/games.html

   Note that other sites can work for play-by-play data as well. For example ESPN has them as well.

   MLB: http://www.espn.com/mlb/playbyplay?gameId=370612102

   NBA: http://www.espn.com/nba/playbyplay?gameId=400974772

   NFL: http://www.espn.com/nfl/playbyplay?gameId=400951760

   Available play-by-play data (already scraped and parsed):

   MLB ("all" years up to 2016): http://www.retrosheet.org/game.htm (not easy .csv)

   NBA (a very small snippet, must pay for more):

   https://www.bigdataball.com/nba-historical-playbyplay-dataset/

   https://www.nbastuffer.com/access-exportable-nba-stats/

   NFL (2009-2016):

   https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016/data

References

1. Using Random Forests for win probability (in NFL)
   http://homepage.divms.uiowa.edu/~dzimmer/sports-statistics/nettletonandlock.pdf
   http://nessis.org/nessis13/lock.pdf

2. Some interesting discussion on win probability models
   https://statsbylopez.com/2017/03/08/all-win-probability-models-are-wrong-some-are-useful/

3. Discussions about "Win Probability Added"
   https://en.wikipedia.org/wiki/Win_probability_added
   https://www.fangraphs.com/blogs/on-fandom-leverage-and-emotional-barometers/