CS S109A: Final Project
Protopapas, Rader, Tanner

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 4 professional sports: baseball (MLB), basketball (NBA), football (NFL), or soccer (league of your choice, if data are available); 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? What coaches use the best (or worst) strategy in a high-leverage situation?

Project goal: To build a win probability model (or expected points model) from play-by-play or event data in the sport of your choice in order to answer an interesting question about the chosen sport.

Data Recources

1. **MLB**
   Baseball-reference for both play-by-play and game schedule:
   - Play-by-play Example: https://www.baseball-reference.com/boxes/PHI/PHI201909140.shtml
   - Game Schedule: https://www.baseball-reference.com/leagues/MLB/2019-schedule.shtml
   - Retrosheet's Data Repository:: https://www.retrosheet.org/game.htm
2. **NBA**
   Basketball-reference for both play-by-play and game schedule:
   - Play-by-play Example: https://www.basketball-reference.com/boxscores/201810160BOS.html
   - Game Schedule: https://www.basketball-reference.com/leagues/NBA_2019_games.html
   - Kaggle Data Repository: https://www.kaggle.com/schmadam97/nba-playbyplay-data-20182019

3. **NFL**
   Pro-football-reference for both play-by-play and game schedule:
   - Play-by-play Example: https://www.pro-football-reference.com/boxscores/201802040nwe.htm
   - Game Schedule: https://www.pro-football-reference.com/years/2018/games.htm
   - Kaggle Data Repository: https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016/data

4. **Soccer**
   Much more difficult to find data.
   - Kaggle Data Repository: https://www.kaggle.com/secareanualin/football-events/home
   - Other possible sources: https://www.jokecamp.com/blog/guide-to-football-and-soccer-data-a

High-level project goals

1. Choose a sport for application between the NBA, NFL, MLB, or Soccer.
2. Collect and parse 'play-by-play' or 'event' data for your sport of choice.
3. Build a model to predict win probability (WP) or expected points/goals/runs, in-game.
4. Properly evaluate the accuracy of your built WP model.
5. Use your WP model in order to answer a question about strategy or to describe the strengths of various players, coaches, or front offices in your chosen sport.

References

1. Using Random Forests for win probability (in the NFL): http://homepage.divms.uiowa.edu/ dzimmer/sports-statistics/nettletonandlock.pdf
2. Discussion about WP models: https://statsbylopez.com/2017/03/08/all-win-probability-models-are-wrong-some-are-useful/