The Big Dig

Our goal on this assignment is to develop a touch-by-touch point win probability model for Major League Volleyball (MLV). Win probability is an important foundational model for sport analytics with applications for player evaluation and in-game strategy. We will demonstrate how to use the model to improve upon hitting percentage for evaluating attackers.

Instructions

In volleyball, each point begins with a serve and progresses with each team allowed up to three consecutive touches (not including blocks) to return the ball to the other side of the net, until the point ends (typically, with the ball landing in or out of bounds). The team which began the point with a serve is the serving team, and the other team is the receiving team. If the receiving team wins the point, this called a *sideout*.

Use the rvolleydata R package to download play-by-play data and player data for the current MLV season. Each row in the play-by-play data corresponds to a touch. The `action` column indicates the type of touch (S $\mapsto$ serve; R $\mapsto$ reception; E $\mapsto$ set; A $\mapsto$ attack; B $\mapsto$ block; D $\mapsto$ dig; F $\mapsto$ freeball). The `outcome` is a subjective rating of the result of the touch (# $\mapsto$ best; + $\mapsto$ good; ! $\mapsto$ neutral; − $\mapsto$ bad; = or / $\mapsto$ worst).

Model the progression of a point as a Markov chain in which the state updates with each touch. Include in the state: (1) which team (serving or receiving) made the touch; (2) how many allowed touches (out of three) remain for the team; (3) the action (e.g., D, E, A); (4) the outcome (e.g., #, +, −). Define two terminal states corresponding to the serving team or the receiving team winning the point. When modeling transition probabilities, assume a second-order Markov chain, meaning that the probability of the next state depends on the two immediately preceding states, not just one. Estimate the sideout probability (i.e., the point win probability for the receiving team) for each state.

Limit the remaining analysis to attacks only. For each player, sum the change in point win probability (from the perspective of the player's team, not necessarily the receiving team) before and after all of their attacks. Additionally, calculate each player's hitting percentage: (kills – errors) / attacks.

Write a report (max 2 pages) summarizing your findings, assuming the reader is a sports data scientist who does not know your assignment. Tell a STORY about your findings. Your report is a narrative about what the analysis reveals and why it matters. For this assignment, you must follow a specific prescribed format for your report:

- **Introduction.** Explain why your analysis matters (one paragraph).
- **Methods.** Describe your state space and how you estimated transition probabilities (one paragraph).
- **Results.** Display two results:
  - a figure plotting sideout probability over time for the longest point (most touches) in the dataset
  - a table of top ten attackers by total change in point win probability (include attacking percentage)

  Do the results look correct? Explain whether they pass the "sniff" test and why (one paragraph).
- **Discussion.** Write two paragraphs:
  - Escape from Model Land. Highlight one player for whom you think this approach poorly evaluates their attacking contributions, and explain what assumption(s) caused the model to get it wrong.
  - Write a concluding paragraph for your report. Recall that the purpose of the concluding paragraph is to (1) signal the ending; (2) restate the purpose of the report; (3) restate the key results; and (4) state why the findings matter (end with a bang).

Extra Credit

Perform regression to the mean on change in point win probability per attack to estimate the true attacking talent for each player in MLV. Present the top ten attackers by true talent in a table, and interpret your results. Submit your response as a separate PDF file named `extra_credit.pdf`.

Submission Requirements

Anonymize your submission by removing any personally identifiable information (including file paths in your R script that contain things like a username!).

- A PDF report named `the_big_dig.pdf`, summarizing your findings (max 2 pages)
- An R sript named `the_big_dig.R`, containing all of the code you usd to perform the analysis

When uploading these files on Canvas, please upload the PDF file FIRST (to help streamline grading).

Rubric

	Needs Improvement (60%)	Meets Expectations (80%)	Exceeds Expectations (100%)
Modeling	*Your mathematical specification is missing, incomplete, or contains significant errors.*	*You correctly specified the model mathematically, but some notation may be slightly inconsistent.*	*Your mathematical specification is flawless and precise, fully defining the model parameters and structure.*
Coding	*You did not implement the model correctly.*	*You implemented the model correctly, but your code is disorganized, lacks comments, or ignores Tidyverse style.*	*You implemented the model correctly with clean code: well commented, easy to read, and adhering to Tidyverse style.*
Critical Thinking	*Your analysis reveals a misunderstanding, or your results don't pass the "sniff" test, and you do not notice.*	*You discuss the results and identify model limitations in a general sense.*	*You connect specific results to model assumptions, identifying exactly where and why those assumptions limit the findings.*
Communication	*Your writing is disorganized; your figures/tables are messy, unlabelled, or difficult to interpret.*	*Your writing is clear and figures are functional, but the report lacks a polished professional feel or strong paragraph structure.*	*Your figures and tables look polished, and your paragraphs have clear topic sentences with supporting evidence.*