Andrew Marsee
DS2
12/18/18

Midcourse Project Proposal

Executive Summary

For the midcourse project, I would like to explore some basketball data, specifically NBA data. It's always been my favorite sport and these days data is widely used in decision making in the sport. I've read a lot of articles over the years that utilize data in order to show some insight into a team or player. So now I want to try my hand at it. I've heard talk over the years about how rest affects the players. How does the number of days of rest affect teams? How do teams with no days of rest compare to those with 2, for example. What does the margin of victory for no days of rest look like when compared to 3 days of rest? What about win percentage? Another aspect I'd like to look at is the start time of games. Do east coast teams perform worse when traveling west? What about the other way around? I'll be using the ballr package from R, which grabs schedules and stats from the Basketball Reference website. There's also a Kaggle dataset which provides box score data from 2012 to 2018 which could be useful. Wins and losses are zero sum game, as well as point differentials. A challenge could be merging data to include days of rest for the opposing teams. Then I can analyze games in which teams were playing against a team with a different number of days of rest. An assumption will be that any game with more than 4 days rest is a game following the all star break in February. These games won't be included.

Motivation

Basketball has always been my favorite sport and over the last 10 years or so the sport has undergone an analytics revolution. Every NBA team has an analytics department now, and use it with varying levels of success. The Houston Rockets are perhaps the most well known case. Their General Manager, Daryl Morey, was one of the first to implement data analytics in team personnel decisions and in-game strategies. High volume three point shooting is now the norm as a result. Every team has followed suit, with the Golden State Warriors, Philadelphia 76ers, and Toronto Raptors being high profile examples in the past decade.

There was even a challenge that the Sacramento Kings issued a few years ago for amateurs to analyze the draft class and provide recommendations for who they should select. I submitted an analysis that was incredibly basic. I charted player efficiency ratings in college versus those in the NBA. Players with high efficiency in college tended to have high efficiency in the NBA as well. It was a very weak correlation, but I provided recommendations anyway. Now I'd like to try some NBA analysis again with more tools in my tool belt.

Data Question

I'm interested in how the number of days of rest before a game affects outcomes. Do teams tend to lose more often when they play on back to back days? My thought is that games on back to back days is tough on players, so the team with no rest is more likely to lose when playing a team with more rest. I'd also like to look at starting times and see if that has any effect.

Are teams traveling from east to west at a disadvantage? What about the other way around? I'd like to try to predict who will individual games.

A stretch goal, if I have time, is that I'd like to assess which stats correlate to better overall seasons. This can be used to determine the types of players needed for success. Is something like offensive rebound percentage correlated with more wins? I'll focus on days of rest and start times at first, then proceed to the next question.

I did a search to find any analysis previously done on days of rest and found the following pages. How does of rest affect pace and efficiency (link). How rest affects win percentage (link). Explaining how rest affects the league (link). ESPN taking time to identify games which pose a challenge for teams with minimal rest (link).

Schedule (through 1/26)

1. Get the Data (1/1)
2. Clean & Explore the Data (1/6)
3. Build & Deploy your Shiny App (1/15)
4. Document/Pitch your Shiny App with a Presentation (1/20)
5. Individual presentations (1/26)

Data Sources

I'm going to be using data from basketball reference through the ballr package in R. I'm currently thinking I'll use data since 1985, a time which is generally thought of as the modern NBA. I may also look at a subset since 2007, since that is when Daryl Morey was hired with the Rockets. The data since 2007 will represent the current state of the game as more teams have changed their tactics since that time. From this package, I've compiled a dataframe with every team's schedule and results since 1985.

The Kaggle data set with stats from 2012-2018 would be useful as well. The box scores have a lot of information, including individual player stats and the referees in each game.

Known Issues and Challenges

I'd anticipate some challenges with shiny. I'll need to visualize what I want to show and which aspects can be dynamic. I'd like to have a menu where you can select the year, home team, away team, days of rest for each team, and start time. The output would show relevant stats for those circumstances and prediction of which team would win.

There could be some challenges with creating a dataframe with information about the home and away team. The ballr team schedules have the opponent, but not how many days of rest they have. I plan on creating another dataframe with the dates and team abbreviation and merging them.

Turning the information in the data frames into predictions could pose a challenge since I still have limited experience doing that. I'll make sure to get started early and consult DataCamp and instructor for ideas and tips.