Capstone Project Draft 2

Carl Larson

1/31/2018

Stating the problem:

When, exactly, should a coach sub-out their star players at the end of a game?

Even in a post- "Moneyball" world, there is still a drought of statistics in the NBA, and it shows a specific type of weakness in coaching regarding player injury-risk during garbage time.

Dozens of millions of dollars-worth of salaries are represented by the players on any active NBA court, and during game time, coaches can, like a surgeon in the moment, make emotional mistakes, keeping star players in too long. These errors cast an existentially tall shadow of injury-risk over those massive player salaries.

The opportunity of solving this problem is encapsulated by the story of head coach Tom Thibodeau's Chicago Bulls in 2012. As the top seed in the east, starring Derrick Rose and a stacked squad of backups, they took on the the 8th seed 76ers in a first- round game 1 playoff home opener that would rewrite franchise history for those Bulls.

To look at stories in basketball for "Garbage Time" analysis, we need the "Gameflow Chart." For those who don't know yet, this chart is the difference in points for both teams (as the y-axis) over the total time of the game (the x-axis).

For game 1 of their series on April 28, 2012, the Gameflow chart ended up like this:

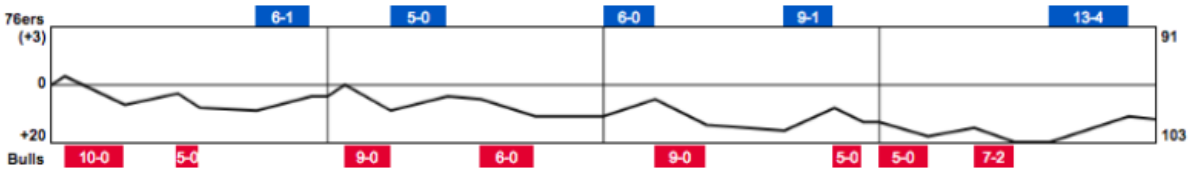

Figure 1:

Basketball is a game of numbers, and the late, great Los Angeles Lakers' announcer Chick Hearn famously coined the phrase "Garbage Time," meaning when a game is in the bag before it's technically over, and you often see the benchwarmers sub in if the coach knows a thing or two.

Beyond just defining "Garbage Time" with mathematical precision, this Data Science project asks, **"Before the game clock expires (or had expired), can we (or could we have) accurately declare(d) an NBA game as decided, with high statistical confidence over 95%?"**

Tom Thibodeau won a championship with the Boston Celtics in 2008 as their defense-savvy assistant coach. That team won on the organization of its defense, and how they effectively double teamed opponents' stars, and then quickly moved to cover the open player after the targeted star passed the ball, assertively forcing the offense into a retreat pattern most teams found unfamiliar, and significantly raised the likelihood of a steal or missed shot.

This was all orchestrated by Thibodeau's excellent strategic coaching, and overall team tenacity showed by the Celtics that year, who starred Paul Pierce, Kevin Garnett and Ray Allen, beating the Lakers in a classic finals.

4 years later, Thibodeau was trying his brilliant defensive strategy as head coach of a very talented, stacked Chicago Bulls team, and had earned the number 1 seed in the conference over the Miami Heat "superteam" which would eventually win the championship that year.

"The score was going the other way," Thibodeau famously said after the game, referring to a run of scoring by the 76ers late in the game.

This particular 76er scoring run, which you can see in the above gameflow chart, would ultimately prove futile; it was garbage time, but as smart as he was, Thibodeau lacked the mathematical precision to define garbage time, and so his substitution-related decisions became emotional.

The road to a championship runs through end-of-the-game strategy, an area where many basketball coaches are weak.

Derrick Rose played through almost half a quarter of garbage time before blowing out his knee with a trivial amount of time left on the clock, less than 90 seconds in a blowout game by the 1-seed (top) over the 8-seed (bottom) in the first round- a series that because of this injury, the 76ers would come back and win, being one of those rare teams in NBA history to be both bad enough to be an 8-seed, yet also good enough to win over the 1-seed in the first round. This has only happened 4 other times in NBA history.

Now Thibodeau and Rose have left Chicago in defeat, both now with other teams, and it basically destroyed those Bulls, which has fully dismantled since then, and is now a lottery franchise which doesn't go to the playoffs.

Despite all the Thibodeau sympathizers and their apologies, it's still true that anything can happen when you ignore the math.

Who Needs The Solution To This Problem?

First and foremost, coaches benefit from this opportunity. Gregg Popovich may be the only coach in the NBA who abides by this "Garbage Time Rule," and he coaches the San Antonio Spurs, who have 5 recent championship banners hanging in their arena. Not to say Poppovich thinks this way, it's more to say his actions reflect an understanding of the principles behind this type of Garbage Time Rule. He is the longest-tenured coach in the NBA and widely regarded as the best active coach, and it's notable that he is an outlier in terms of abiding by this garbage time rule while most other coaches don't.

General managers also need to know this data. General managers, or whoever directly hires and fires the coaches, truly need to know if their coach is making these easy mistakes. This also provides a more mathematically rigorous way of justifying those hiring or firing decisions.

Time is valuable for fans too. Time-frugal fans who aren't afraid of statistics may like to know when a game is still in reach versus when it is truly decided in a more rigorous way. Fans don't want to leave a good game early, which happens a lot- and they also don't want to needlessly stay too long and waste their valuable time if the game is clearly over.

Where Is The Data?

The input data is the Gameflow table for any given NBA game.

This type of data is an ongoing challenge, as in the stats-rich world of basketball today, nowhere easily provides gameflow tables, and it's the kind of data you still need to make from scratch. Popcornmachine.net has this type of data but after corresponding with them, they have indicated it takes perl script to access their gameflow tables.

Nonetheless, observations with three factors were obtained for the 1,350 most recent NBA games (as of December 2017) and will be used the as the sample pool. It was also noted whether the game entered garbage time or not, and whether this "Garbage Time Rule" was violated.

To increase the sample pool beyond coding spreadsheets by hand would require using perl script to custom-parse the html, and to plug it into this project's eventual R program which contains the Garbage Time Rule's algorithm. Future projects that elaborate on this will entail using perl to data mine Gameflow table data from XML files so it can be manipulated in R, and then batch-uploading many games' worth of data (gameflow tables) and judging them all at once. Another later project might involve making a calculator that can read a game's box score (for a game in progress) and return whether the game is in garbage time or not, and if so, a fire-drill to any coaches to get their star players out of the game if they haven't already.

The scope of this project at hand is to make an R program that takes in a gameflow table, and returns whether the game entered garbage time, and whether the garbage time rule was violated. This is essentially two boolean variables one result of which is null, leaving us with a triangular matrix.

Fortunately for now the immediately available data set does consist of 1,350 observations with 3 factors, which are:

- whether the game entered garbage time
- whether the rule was violated
- the date of the game

Including the date enables complete scientific back-verification and reproducibility, as well as interesting ways of sampling and graphing the data.

How This Works

The "Garbage Time Rule" equation for purposes of this project is:

$P = 2M + 6$

Above is the envelope, the maximum lead a team can have without garbage being declared, and below is the inequality series for the boolean variable of whether the game entered garbage time.

$P > 2M + 6$

$s > 15$

$M > 0$

Where the domain M is minutes left (decimalized as a scalar multiplier) greater than zero; the range of P is real integers, equal to the difference of points difference between the two teams; and s is the number of consecutive seconds for which the first inequality is true.

It is stipulated that M must be greater than 0, because at $M = 0$, the game is over and there is no opportunity to make a prediction when the game has expired. Allowing $M = 0$ into the domain essentially ruins the slope of the line and creates the possibility for dividing by zero later which adds to the unnecessary confusion of this very concrete scenario. Thus, creating a more piecewise equation would be within the scope only of future projects.

Another chart to help illuminate this analysis is the descending envelope for the minimum number of points a team would need in order to achieve a Garbage Time guarantee of victory. This line is a decent approximation for our purposes, and actually becomes more precisely accurate in the final minute of the game, slightly showing its true step-stairs type shape, due to the integers-only nature of points in basketball, and the also integer-only nature of seconds remaining in the game, as well as s being a quarter minute, something we can see there in that final 30 seconds where the line below gets a bit granular.

For reasons of mathematical elegance and project scope, this analysis uses decimalized minutes, so that integer s gets largely wiped out as continuous, computationally speaking, in this particular algorithm. However in future projects, it would be plausible to transform the scalars in this algorithm to work for "seconds remaining," and achieve a truly accurate stepwise graph here:

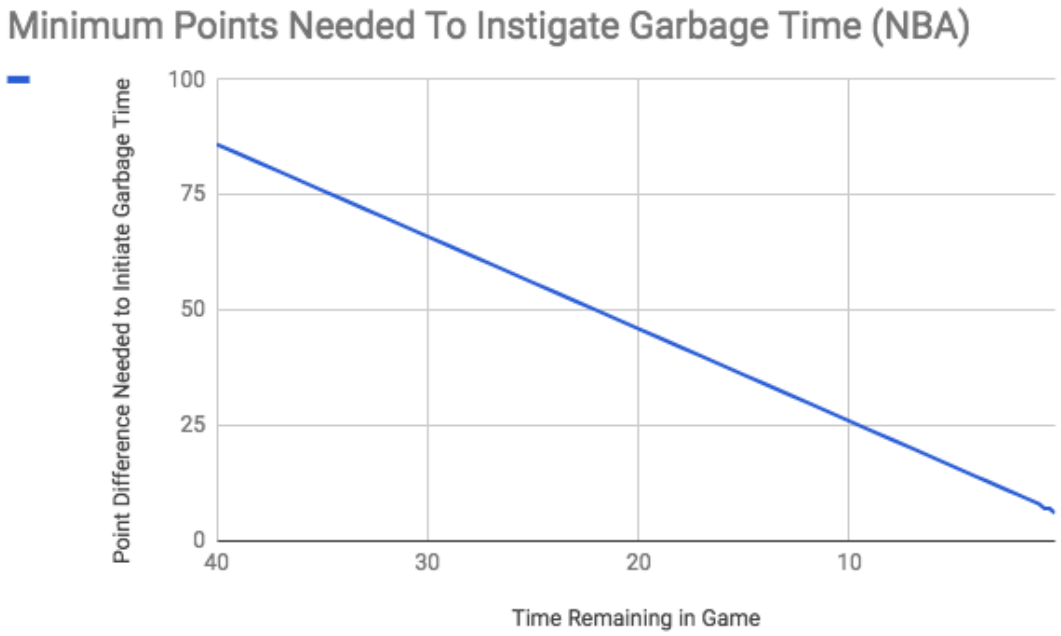

Figure 2:

The axes of this above chart match the axes of a Gameflow chart, so they can be superimposed on the same chart, and wherever it intersects, basically Garbage Time is declared henceforth.

The combined graph looks roughly like Figure 3.

Figure 3 is a fictional test scenario meant to help illustrate the topic (where a team actually comes back and breaks this garbage time forumla, just so we can see what that might look like).

Below are the two hypotheses, one to challenge us on accuracy, and the other to challenge us on applicability:

Experimental Hypothesis A: *Over 95% of basketball games sampled will conform to this Garbage Time Formula.*

Null Hypothesis A: *More than 5% of basketball games sampled will not conform to this Garbage Time Formula.*

Experimental Hypothesis B: *This Garbage Time Rule will apply to over 40% of basketball games sampled.*

Null Hypothesis B: *This Garbage Time Rule will apply to less than 40% of games sampled.*

Deliverables:

- A measure of explanatory power which must be over 40% (i.e. 40% of games sampled experienced "Garbage Time")
- A measure of accuracy which must be over 95% (i.e. this theory was incorrect 4% of the time)
- An analysis of any time the rule was broken
- Scenario analysis the 3 possible outcomes which are again:

1. Rule doesn't apply (close game)
2. Rule applied and was validated
3. Rule applied and was broken

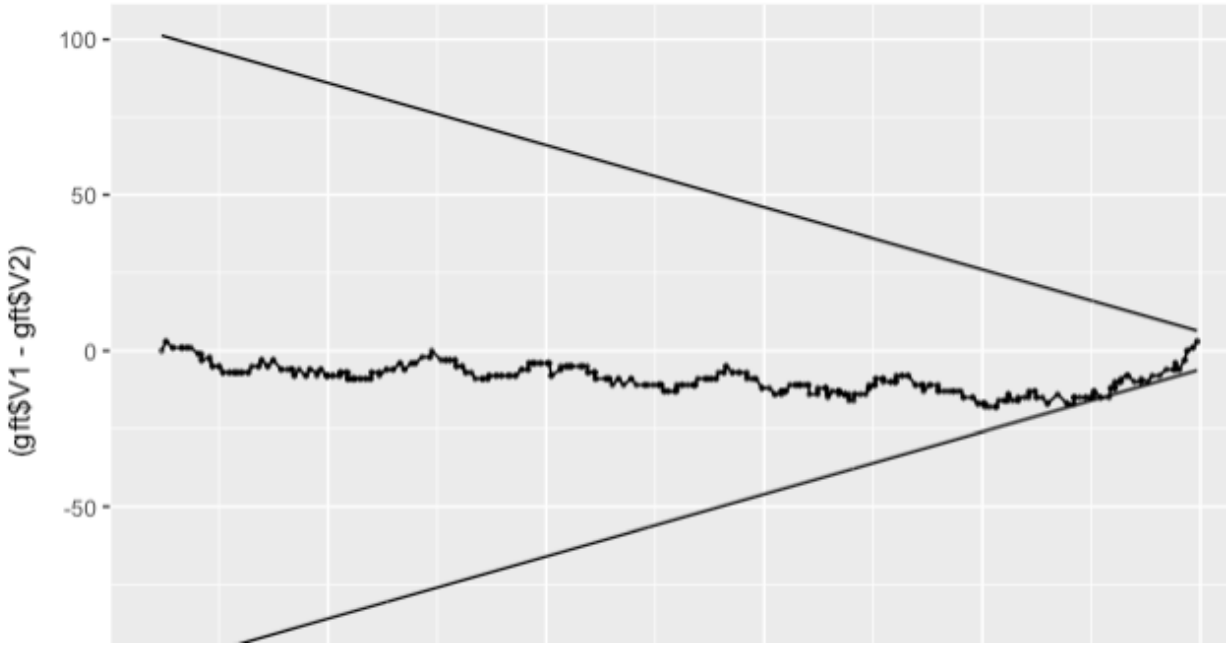

Figure 3: Figure 3