UNIVERSITY OF CALIFORNIA

Los Angeles

Beating the Book: A Machine Learning Approach to NBA Win Probabilities in Search of an Edge Over the Betting Odds

A thesis submitted in partial satisfaction of the requirements for the degree Master of Applied Statistics

by

Guy Dotan

2020

© Copyright by Guy Dotan 2020

ABSTRACT OF THE THESIS

Beating the Book: A Machine Learning Approach to NBA Win Probabilities in Search of an Edge Over the Betting Odds

by

Guy Dotan  
Master of Applied Statistics in

University of California, Los Angeles, 2020

Professor Frederic R. Paik Schoenberg, Chair

In this study, we set out to better understand the voter behavior of different demographic partitions of Los Angeles County (LAC). We developed a methodology to enrich and analyze LAC voter data by joining it with US census data and modelling vote by mail rates and turnout rates using general additive models. In doing so, we were able to forecast where in LAC we expect to see the highest rates of turnout by vote by mail and turnout in person for the 2020 presidential election. These findings are timely: In 2020, LAC will roll out a new and novel voting system called Voting Solutions for All People. The findings of the current study provide guidance on how to best allocate resources to high-turnout populations and target outreach to low-turnout populations in preparation for the first election under the new system. This paper presents a combination of visualizations, predictive models, and summary statistics that can inform LAC of where to focus outreach to expand the electorate and build a more representative electorate while also improving preparation in areas where turnout has been historically high.

Chargeback fraud is a massive problem for e-commerce businesses. Using historical ticket order data, several machine learning models are trained and tested to predict which transactions are high risk for chargeback. The results of this thesis show that many fraudulent transactions can be successfully identified and stopped before they are processed. Using these types of models could significantly reduce chargebacks, saving companies time and money.

The thesis of Guy Dotan is approved.

Vivian Lew

Zili Liu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2020

Table of Contents

1. Introduction
2. The Mathematics of Sports Gambling
3. Data Collection
4. Data Processing and Exploratory Analysis
5. Random Forest
6. Neural Network
7. Betting Results
8. Kelly Criterion
9. Conclusions
10. References

List of Figures

List of Tables

Acknowledgements

Chapter 1

Introduction

Over the past decade there have been two seminal shifts in the world of sports analytics and consequently the entire professional sports landscape. The first: the proliferation and democratization of accessible data. The second, and more recently: the federal legalization of sports gambling within the United States.

While the sports industry might be one of the newest sectors to be disrupted by the emergence of data-driven decisions challenging preconceived notions from “experts”, its impact has been fast and far reaching. Setting aside the unprecedented shock to the economic ecosystem—specifically within sports and entertainment—from the 2020 outbreak of COVID-19, the sports analytics business has been thriving. “The global sports analytics market is expected to reach a revenue of $4.5 billion by 2024, growing at a CAGR [Compound Annual Growth Rate] of 43.5%” [1].

1.1 – Current State of Sports Analytics

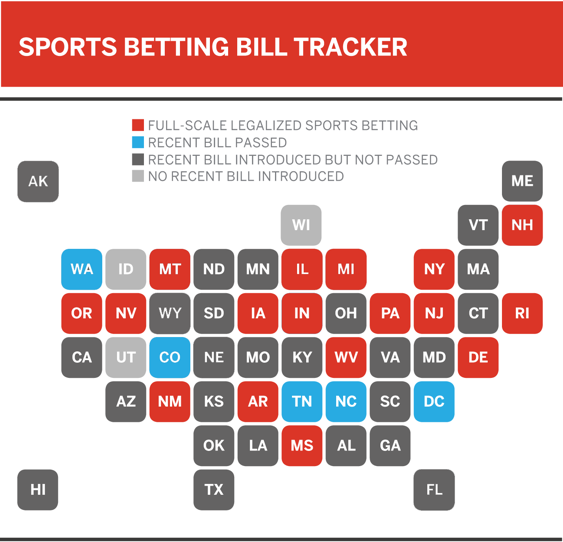
To the general public, the most well-known adoption of analytics into the sports universe was within Major League Baseball, thanks largely to Michael Lewis’ 2003 book *Moneyball* and subsequent movie blockbuster, staring Brad Pitt, in 2011. This story, which chronicles the influence of Bill James, the field of empirical baseball research known as Sabermetrics, and the story of the 2002 Oakland A’s success has been the poster child of how data can create a competitive edge on the playing field. But sports analytics has made its impression in far more avenues than just baseball. The field is responsible for the increased emphasis of the three-point shot in basketball, the use of optical player tracking technology in the NFL, and even the statistical optimization of curling game strategy that led the Swedish women’s national team to a gold medal in the 2018 Winter Olympics [2], just to name a few.

The integration of data analysts and scientists as a crucial element of professional sports organizations appears here to stay, but the acceleration in the field’s adoption can be tied to the increased availability of data. The NFL and NBA hold yearly hackathons to allow anyone the opportunity to dive into their sport’s data and present findings to top league officials with prizes and networking at stake. Conferences such as Sloan Sports Analytics Conference in Boston began as a small gathering of about 100 attendees in 2006, and now in 2020, attracts over 4,000 people. The conference has gained national recognition, notably hosting former President Barack Obama as the keynote speaker in 2018. The industry’s explosion in popularity, though, has been aided by communities such as FiveThirtyEight, Retrosheet, Sports-Reference, and league-offered APIs bestowing data democracy to anyone that desires. Sports analytics has largely become open-source and this hivemind has benefited players, teams, and organizations.

1.2 – The Legalization of Sports Gambling

On May 14, 2018, the Supreme Court case *Murphy v. National Collegiate Athletic Association* reached a landmark decision regarding the federal government’s right to control a state’s ability to sponsor sports betting. In a 6-3 decision, the Professional and Amateur Sports Protection Act of 1992 (PASPA) was overturned, thus opening the doors for every state to make its own laws permitting in-state sports wagering.

In just two years since the ruling there are already 17 states with full-scale legalization and another five that have passed legislation that will take effect in the coming year. [3] And as one would expect, bettors in legal states have flocked to sportsbooks, both digital and brick-and-mortar. (Sportsbooks, or “books”, are places, often times part of a greater casino, where bettors can make wagers on all types of sporting events.) Since the overturning of PASPA, Americans have placed over $20 billion of bets which has generated $1.4 billion of revenue in those legal states. [4] Morgan Stanley projects that in just five years, by 2025, almost three-quarters of US states (36) will have legalized sports betting and the U.S. market could see $7 to $8 billion in revenue. [5]



1.3 – The Intersection of Data and Wagering

Sportsbook operators within casinos have had decades of experience building a complex infrastructure of analytics to help them determine where to set up their gambling lines. Their goal is to set up a bet every game such that there is an even amount of money wagered on both sides of the bet. This allows them to take their cut of the wagers (known in the industry as the “vigorish” or “vig”) and thus drive revenue to their casino, no matter which team wins. For the entirety of their existence, sportsbooks have maintained a significant edge over the majority of bettors. Their advantage was largely based on their access to data and domain expertise building models to determine how to establish the perfect betting line. It is this statistical edge that has keeps income flowing into the sportsbooks within casinos. Money that helps build lavish 50-story casinos and hotels that makes Las Vegas strip world-renowned. That said, surprisingly, sports wagering makes just 3% of gaming revenue in Nevada casinos. [6]

A screenshot of a cell phone

Description automatically generated

But now, with sports wagering becoming more commonplace in the American society and the proliferation of available sports data to everyday consumers, there is an opportunity to close the gap between casinos and bettors. Similar to how stockbrokers use proprietary projection models to systematically “beat the market”, sports wagering has followed suit.

Recall, a sportsbook’s objective on each bet is to account for an even amount of money wagered on both sides. Often times a betting line is skewed by the inherent biases of an average sports bettor. For example, if the Los Angeles Lakers (a TV market size of over five million people) were to play the San Antonio Spurs (TV market size of just 900 thousand), we might expect a sportsbook to make a line that slightly favors the Spurs. Even if the teams were evenly matched, sportsbooks would expect a disproportionate amount of hometown favorite bets supporting the Lakers. Just the smallest marginal edge, demonstrated by this example, could be enough to be exploited by an adept model. A model, when applied to a large enough dataset, could yield a considerate return on investment.

The goal of this study is to determine if applying machine learning methods to vast sports datasets (in this case, within the NBA) can create such a model that would give a bettor the competitive edge over the lines set by a sportsbook.

Chapter 2

The Mathematics of Sports Gambling

In order to understand how to beat the bookmakers, one first needs to understand how to interpret the betting lines they provide. There is a wide array of different types of bets that a person can make at a sportsbook. The three most popular betting styles are: “point spreads”, “over/unders”, and “moneylines”. The following are descriptions of these types of bets using basketball as an example.

2.1 Point spreads, over/unders, and moneylines

Point spreads are mechanisms used to account for the discrepancy between two unevenly matched teams. Usually noted by: Warriors (-5) vs. Clippers or the inverse: Clippers (+5) vs. Warriors. The number in the parenthesis is called the “spread.” Essentially, the sportsbook believes that the Warriors are more likely to win the game, but the book wants to drive an even amount of bettors to wager on the Clippers (even though they are underdogs) as the Warriors. Therefore, their bookmakers suggest placing a spread of five points on the game. So, if I bet on the Warriors they need to win by more than 5 points for me to win the bet. If I bet on the Clippers, they have to lose by 4 points or fewer or win the game, for me to cash in on the wager. If the game ends with the Warriors winning by 5 points, then this is called a “push” and all bettors get their money back. Sometimes a spread will be listed at 0 points (called a “pick-em”), indicating that this is an even matchup and all you have to do is pick the winner to win the bet.

Similar to point spreads, over/unders involve a specific point amount that a bettor needs to wager on the correct side of. However, in this case, the winner of the game is irrelevant. All the bettor must do is guess if the combined score between the teams will be greater than or less than the over/under line. For example, Warriors vs. Clippers (+200). If I bet the “over 200”, I am expecting the combined score between the two teams to reach 201 or more and it does not matter what combination in occurs (Warriors 120-Clippers 81, Clippers 101-Warriors 100, etc.) If the final combined score is exactly 200, again this is called a push and bettors get their money back. For this reason, over/unders (and spreads for that matter), oftentimes use fractional lines (+5.5 or +200.5) to prevent the case of a push. Over/unders can be offered for single quarters, just the first half, just the second half, or even for a single team’s score.

The third type of popular betting type are called moneylines and are the relevant bet type for this study. In a moneyline bet, a person simply needs to determine which team will win the game. But if one team was heavy favorite versus the other team, it would not make sense for a sportsbook to pay out and an equal amount for choosing the favorite as choosing the underdog. As a result, sportsbooks offer a moneyline, which adjusts the amount you win for having your bet hit based on the likelihood that team will win the game. Moneylines are notated in various formats: decimal, fractional, and moneyline. The first two are commonly used in Europe. This paper will use the moneyline odds notation since they are most common to the US (and often called “American” odds). Moneylines are written as follows: Warriors (-235) vs. Clippers, or conversely, Clippers (+185) vs. Warriors.

Essentially, these either positive or negative, three-digit numbers, imply how much money a bettor would profit relative to a $100 bet. +185 means that if a bettor laid $100 on the Clippers, and then they won the game, the bettor would make $185 profit. -235 means that a bettor would have to wager $235 in order to profit $100 from that game. So if a bettor placed $100 on the Warriors at -235 moneyline odds, and the Warriors indeed won, the bettor makes $42.55 profit.

Moneyline Payout – Underdog

ML / 100 = payout / bet-amount

ML \* bet-amount = 100 \* payout

Payout = ML / 100 \* bet-amount

= +185 / 100 \* 100

= $185

Moneyline Payout – Favorite

(ML \* -1) / 100 = bet-amount / payout

Payout \* (-1 \* ML) = bet-amount \* 100

Payout = 100 / (ML \* -1) \* bet-amount

= 100 / (-235 \* -1) \* 100

= (100 / 235) \* 100

= 42.55

In summary, a $100 bet on the Clippers (+185) leads to $185 profit. A $100 bet on the Warriors (-235) leads to about $42 profit. This discrepancy in profits is to discourage enough bettors from taking the favorite Warriors and instead take the potential for upside in profit by betting on the underdog Clippers. Again, the sportsbooks goal is to optimize these moneylines to set it at a line such that an even amount of money is placed on both sides.

2.2 Implied Win Probability

The payout formula for moneylines makes it quite simple to determine how much profit a bettor can make from having their wager hit. Risk averse bettors tend to take favorites (negative moneylines) despite the lower payouts because there is a higher chance that the team they bet on will win. Risky bettors will seek out underdogs with lower probabilities of winning, but they think will actually surprise the public, win the game, and thus provide a larger profit-margin.

In addition to the payout, however, moneylines can actually be converted into a win probability known as “implied probability”. The implied probability formula is defined as the size of the bettor’s wager divided by the return on investment for that wager. Or simply: risk over return.

Implied probability = risk / return

2.3 Derivation of Implied Win Probability

Risk = Bet-Amount

Return = Risk + Profit

Implied Prob for Favorites

Implied Prob for Underdogs

Combined formula

<https://tex.stackexchange.com/questions/47170/how-to-write-conditional-equations-with-one-sided-curly-brackets>

impliedProbablity = \left\{\begin{matrix}

\ \frac{100}{ML \* 100}, & \text{if ML} \geq 0\\

\frac{ML \* -1}{(ML\*-1) +100}, & \text{if ML} < 0\\

\end{matrix}\right.

2.4 The Vig or the Juice

As mentioned previously, the goal of a sportsbook is to have an equal amount of money placed on both sides of a wager so that no matter the results, they will make money once they take their cut of the bets. This cut is known as the vigorish, and more colloquially, the “vig” or the “juice.” So how does one calculate the juice? Let’s use our example from above with the Warriors versus the Clippers.

The Warriors moneyline odds were -235, which after using the derived formula, comes out to an implied probability of 70.15%. The Clippers moneyline odds were +185 and therefore an implied probability of 35.09%. Now the most basic rule of probability states that the sum of all possible probabilities of an event always equals 1. And more specifically, the probability of an event plus the probability of the complement of that event equals 1. If we consider the chances that a team wins a game as the probability while the chances they lose is the complement of that probability, we would expect these two events to sum to 1. See below:

Given:

P(A) + P(AC) = 1

Let…

P(A) = Warriors win

P(AC) = Warriors lose (i.e. Clippers win)

P(A) = 0.70

P(AC) = 0.35

P(A) + P(AC ) = 1.05 ≠ 1

So these two implied probabilities are mutually exclusive, compose the entire space of outcomes, and yet sum to over 100%. This summed probability (in this case 105%) that is greater than 1 is called the “overround” and is how sportsbooks take their cut. By setting the betting lines such that the probabilities result in an overround, the sportsbook effectively ensures that they will gain a profit from this wager. In our example above, a $100 bet on the Warriors pays out $42, while a $100 bet on the Clippers pays out $185. In order to determine how much a sportsbook would expect to pay out from these implied probabilities can be calculated as follows:

Expected Payout E(P) = Profit \* Probability of Winning

Sportsbook’s E(P) on the bet = E(Warriors Win) + E(Clippers win)

E(Warriors win) = $42 \* .7 = ~$30

E(Clippers win) = $185 \* .35 = ~$65

Sportsbook’s E(P) = $30 + $65

= ~$95

As seen above, this overround of the probabilities creates the vig for the casino. If the sportsbook were to take $100 of total wagers on this bet, on average, they would expect to pay out just $95. Thus, ensuring a cut of about $5 for this wager. Now consider the fact that casinos collect millions of dollars on bets, not $100, it is easy to see why sports gambling is such a lucrative endeavor for bookkeepers, especially when the optimal moneylines are established (and therefore the juice is optimized as well).

2.5 “Removing the Juice” - Actual Win Probability

In order to get a clear idea of the sportsbooks’ expectations of how likely each of the two teams is to win a matchup we need to get rid of the guaranteed profit they bake into the lines. The implied probabilities are derived from the betting lines but the actual probabilities are what remains after taking the vig into account. Probability theory claims that the sum of all possible events should always equate to 1. So to get the true probabilities based on the betting odds, the implied probabilities need to be skewed so that they also sum to 1. The method to remove the vig is simple, just divide the probability by the overround.

Actual Probability = Implied Probability / Overround

Actual Probability the Warriors Win = 70% / 105% = 67%

Actual Probability the Clippers Win = 35% / 105% = 33%

We see that the two probabilities now sum up to 100% and therefore represent the true probability that the sportsbook places on each team’s chances of winning. In summary:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Team | Moneyline Odds | Risk | Profit | Return | Implied Win Probability | Actual Win Probability |
| Warriors | -235 | $100 | $42 | $142 | 70% | 67% |
| Clippers | +185 | $100 | $185 | $285 | 35% | 33% |

Chapter 3 – Data Collection

The gambling mathematics discussed in Chapter 2 can be applied to any sport with betting lines. But for this study, we will focus specifically on professional basketball betting data. The NBA is a particularly useful test subject for this research for a few reasons. First off, over the last decade, the NBA has experienced arguably the largest increase in the emphasis put on data analytics of the four major sports (NHL - hockey, MLB - baseball, and NFL - football). More interest in basketball analytics has led to easier accessibility to data that could be useful for research and modeling. Not only is the data availability trending upwards, but the general popularity of the sport of basketball is on the rise as well. A 2017 Gallup poll found that basketball has surpassed baseball as America’s second-most favorite sport to watch. Although, football is still heavily the favorite: at 37%, versus 11% for basketball. [8] Thus, with the sport’s popularity continuing upwards one would expect the betting markets to grow as well. For our purposes, the advantage the NBA has over the NFL, is its sample size. The NBA consists of 30 teams and 82 games, as opposed to the NFL which has 32 teams, but just 16 games. Additionally, basketball games are high scoring and high possession competitions, hence, within a single basketball game, there are a lot of statistics that accumulate. Much more than a baseball, football, or hockey game. Simply put: lots of games and lots of stats within those games makes for a large sample size and a great dataset.

A close up of a map

Description automatically generated

3.1 Betting Lines

The first step to building the dataset required for this study was acquiring betting data for as many NBA games as possible. More specifically, we are looking for moneylines for the full game (as opposed to a singular half) since that is what is necessary to convert betting odds into win probabilities. Fortunately, an archive of betting odds for the NFL, NBA, NCAA football, NCAA basketball, MLB, and NHL were all available on one online resource.[[1]](#footnote-1) The relevant data for this study went back to the 2007-08 NBA season and was complete up to the current 2019-20 season. Each file was formatted into one of 13 downloadable Excel files that contained both regular and postseason odds. Each season’s data was clean and uniform, which seamlessly merged into one large dataset of 32,952 rows (16,476 games since each game had two rows to represent each team’s odds). There were no missing moneylines in the dataset, spanning from the first game of the 2007-08 season until the NBA season was abruptly cut short on March 8, 2020. This was a historic day, not only in basketball, but professional sports history, as Rudy Gobert, a center for the Utah Jazz tested positive for COVID-19, prompting commissioner Adam Silver to cancel the season indefinitely. [9] With a complete set of betting data, the final processing step for the betting data was to convert them into win probabilities. Using the derived probability formulas from Chapter 2, R functions were written to implement these formulas and were simply applied to the entire betting dataset.

3.2 NBA Game Statistics

There were several different approaches for statistics that we could use to build a win probability model to pair alongside our archive for betting odds. The most straightforward approach, and the one used for this research, was to utilize game-by-game, team-level box scores. Game-level detail allowed for the flexibility to aggregate the data in a variety of ways that would be useful for modeling purposes.

Fortunately, the R package *nbastatR* provides a robust interface that easily pulls data from a variety of online basketball data resources such as: NBA.com/Stats, Basketball Insiders, Basketball-Reference, HoopsHype, and RealGM. [10] One of the functions in this package loads game-logs for each team over any desired seasons. All that was necessary was to input the same 2007-2020 season span that we already had in our betting archive. This gave us the same 16 thousand or so games and a variety of raw team statistics to use in our model. A data dictionary of these variables is shown in table XX.

Joining the betting odds data with the game-log box scores was not a trivial task. Because these datasets did not come from the same source there was no linking key between each box score, each betting line, and each team. However, through rigorous data cleaning—which included the manual creation of standardized team IDs and game IDs—we were able to combine the two datasets together. The merging of the two datasets was a perfect match with team statistics available for every single matchup in which betting lines were available save two instances. One, the cancelled 2020 games due to COVID-19 and two, a March 15, 2013 game between the Boston Celtics and Indiana Pacers which was cancelled, and never rescheduled, as a result of the tragic bombing at the Boston Marathon the day before. [11]

3.3 Advanced Metrics

For the majority of basketball history, analytics were driven by these standard box score metrics currently in our dataset. More recently, however, the industry has been shifting away from these raw values toward a more dynamic approach that can take in account the shifts in game tempo season-to-season and even game-to-game. The solution: looking at the normal metrics—points, rebounds, assists, etc.—on a per possession basis, as opposed to per game. “By looking at the game at a per-possession level, it eliminated pace and style of play differences from the equation and put all teams on a level playing field. This way a team that is constantly running and has more possessions each game doesn’t have a statistical advantage compared to a team that plays at a slower speed.” [12]

Possessions used to be estimated through a rudimentary formula that took into account how many field goals, free throws, turnovers, and rebounds a team got over the course of a game. But with the proliferation of play-by-play data, the NBA now has an exhaustive account for true possessions for each team, each game dating back to the 1996-97 season. The NBA actually provides advanced data (of which possessions is included) on per game and per team basis through an open API.[[2]](#footnote-2) Using Python’s *requests* package, we were able to tap into the API and then make the call to the appropriate endpoints to return a data dump of advanced metrics for each game. The data came in a JSON format, so all that was left to do was to use Python’s *pandas* library to format the JSON response into a data frame so it is easier to work with. Since this data was pulled directly from the NBA’s official statistics database, the game IDs and team IDs directly matched those that came from the traditional box score metrics from the *nbastatR* package and was successfully joined together.

Our dataset was now complete with betting data (and corresponding win probabilities), matchup data (outcome, home/away, days rest, etc.), traditional box score metrics, and per-game possessions.

Chapter 4 – Data Processing and Exploratory Analysis

The raw box score statistics provide the backbone for the features that will need to be included in the model. We applied two types of transformations to this dataset. The first, as mentioned previously, was to scale the raw metrics to account for the variability in game-to-game possessions. The second was aggregations of the data, necessary for the accumulation of statistics each team had built up entering that given matchup.

4.1 Seasonal Trends

People who have been fans of the NBA over the past couple decades can attest to how different the game is played in 2020 than before. Whether it is due to the rise in analytics, improvements in training and medical advances, or gameplay strategy implemented by coaches, the fact remains: what it takes to win a basketball game has changed. One of the biggest shifts in game style recently has been the tendency to shoot more three pointers. The secret to unlocking this strategy is no real mystery. NBA teams on average shoot about 35% on three pointers, 45% on two pointers, and 75% on free throws. The math breaks down quite simply as follows:

Expected Points = (Points) \* (Percent chances of making the shot)

E(3PA) = 3 \* 0.35 = 1.05

E(2PA) = 2 \* 0.45 = 0.90

E(FTA) = 1 \* 0.75 = 0.75

There are obviously more subtleties and complexities to these style changes, but the redefining of shot selection has been a major component. With optimizations occurring on court we see it manifesting in season-wide trends on the box score metrics. Figure XX demonstrates these trends.

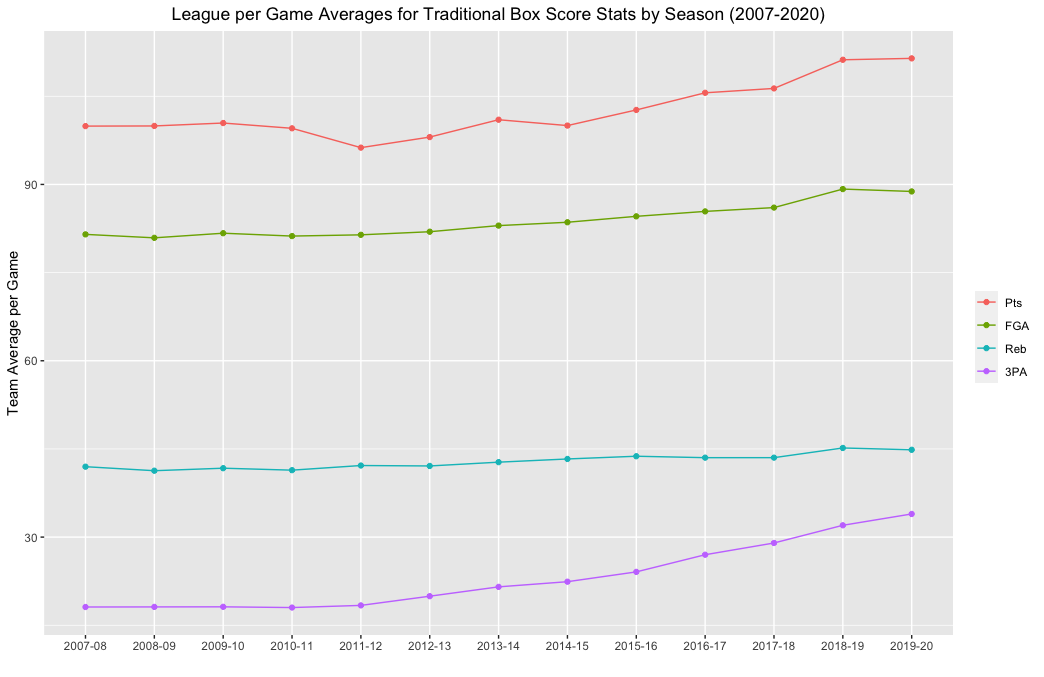


Figure 1 Stats Trends by Season

All four of these metrics are trending upwards, with the most drastic increase being three-point attempts. Just to accentuate this point, the team with the fewest three-point attempts per game in 2019-20 (the Indiana Pacers at 27.5 per game), would have ranked first place by almost a full three-point attempt per game in the 2007-08 season (the Golden State Warriors ranked first with 26.6 per game). The NBA-leading Houston Rockets, who averaged 44.3 threes per game in 2019-20, calculate out to a 145% increase in three-point attempts compared to the league average 18.1 attempts in 2007-08. Going further, 48% of the Rockets field-goal attempts this year came from behind the arc, compared to the league average of 22% back in 2007-08. The league average in 2019-20 was 38%.

Three pointers are up, shooting attempts are up, scoring is up, rebounding is up, all of today’s traditional box score metrics are inflated compared to earlier in the decade. The question that emerges is whether this inflation is due to optimizations in offensive schemes or merely a change in the flow of the game. The eye test suggests that the NBA game is much faster paced today than it was a decade ago. As it turns out, the data backs this belief up.

The more possessions a team gets in a game, the higher tempo that game is being played at. The basketball statistic pace is defined as the average of both team’s possessions per 48 minutes of combined game play.

Pace = 48 \* ((Tm Poss + Opp Poss) / (2 \* (Tm MP / 5))

The season-to-season trend on pace verifies our assumption that the game is played at a higher velocity today. Figure XX shows the positive slope on the regression of pace by season demonstrated as both a scatterplot and series of boxplots.

A close up of a device

Description automatically generated

4.2 Adjusting for Pace

The solution that the basketball analytics community has put in place to this phenomenon of statistical inflation is pace-adjusted metrics. Instead of aggregating stats on a per-game basis, we now examine those same stats in relation to how many possessions that team had in the game. More specifically, the standardized pace-adjusted metric looks at a team’s stats per 100 possessions. Previously, adjusting statistics on a per-48-minute basis was used to account for overtime games, but using per-100 possessions addresses both overtime instances as well as seasonal and daily shifts in game play.

When we build our model, it is important to handle this inflation factor as we don’t want a “good” team back in 2007 to appear like a “bad” team in 2020 standards. The top two charts in Figure XX shows the distribution of scoring and field goal attempts for all games in the 2007-08 season versus the 2019-20 season. As is apparent, the 2019-20 season is shifted to the right verifying this statistical inflation. However, when scaling these two metrics onto a per-100 possessions scale, we see a far great overlap between the two distributions. Clearly, the increase in pace does not account for the entirety of the scoring surges, but it does lessen the gap.

For the purposes of this research we will scale all of these standard box score metrics on a per-100 possession basis. This methodology will help dramatically when trying to train a dataset that spans over a long time frame to standardize across the ever-changing styles of play. In fact, points per 100 possessions (Offensive Rating), points allowed per 100 possessions (Defensive Rating), and the difference between the two (Net Rating) are some of the key metrics used by the analytics community to define the quality of a team’s performance. Figure XX paints a pretty clear picture as to the correlation between the best teams in the league and their performance in those three advanced ratings.

4.3 Aggregation

A complete dataset of game-by-game box scores and all the necessary matchup and team statistics provided the opportunity for a lot of flexibility into terms of how to best aggregate the data. Before aggregating, there was more pre-processing required to handle all aspects for a team and their matchup. First off, the data had to be joined to itself so that we knew that only the team’s own box score in that game, but also their opponent’s. Access to the opponent’s statistics for each matchup makes it much easier to understand the team’s performance on both offense and defense. The number of points and offensive rebounds a team gets will be valuable in a model, but equally as valuable are the number of points they give up and offensive rebounds they allow. Also recall that opponent possessions were part of the formula for pace.

Of course, the statistics that we need to consider when building the model to predict which team won, are not the team’s performance in that game, but leading up to it. We took two different aggregation methods to our dataset. The first, was the team’s year to date performance entering that game. So, for example, if both teams were on their tenth game of the season, we had nine games worth of data aggregated for each squad to make a prediction on that matchup. If it was the last game of the season for both teams, we had 81 games worth of data for each. The one caveat: if the game was the first of the season for that team, we used the entire previous season’s worth of data as our predictors.

The second aggregation method took into account the ebbs and flows of a basketball season. Some teams go on hot streak or cold streaks so using the entire season’s stats to date might not be the best method. As an alternate aggregation option, we looked at the team’s performance in the five-game span entering that game. This rolling aggregation would hopefully be more sensitive than the year-to-date method in reflecting how the team has been performing entering the matchup. The one caveat for this method: if it was the team’s fifth game of the season or earlier, the aggregation would roll back to the previous season. So, if it was the team’s third game of the year, the aggregation would consist of the first two games of that season and then the three last games of the previous. Again, all of these metrics in both methodologies were scaled on a per-100 possession basis, not as the combined raw totals within the aggregation period.

The final processing step required before building the model was necessary to account for the particular response variable for this study. Our goal was to predict the probability that a team would win a game against its opponent, given all the data provided for that matchup. In order to properly build this model, we cannot treat both teams in that matchup as separate entities. If we did, there would be no way to guarantee that the win probability for each team would sum to 100\%. Therefore, the proper way to transform our dataset would be to join the dataset to itself once again. This time, so we have one row per game and build the model so that the response variable (p) is that probability that the home team wins. Then we can simply find the probability that the away team wins by doing 1 – p.

Thus, our dataset went from about 32,000 rows (one row per each team in the matchup) down to about 16,000 rows (one row per game). The sample size may have halved, but the parameter set quadrupled. Before we had just the team’s own statistics for each game. After the processing we now had:

* The home team’s own statistics in each matchup
* The home team’s opponent’s statistics in each matchup
* The away team’s own statistics in each matchup
* The away team’s opponent’s statistics in each matchup

References

[1] “Sports Analytics Market Forecasts up to 2024” https://www.researchandmarkets.com/research/wlpbw8/global\_4\_5?w=4

[2] “Curling may finally be having its Moneyball moment”

https://www.cbc.ca/sports/olympics/winter/curling/curling-moneyball-analytics-1.4623584

[3] “United States of sports betting: An updated map of where every state stands”

https://www.espn.com/chalk/story/\_/id/19740480/the-united-states-sports-betting-where-all-50-states-stand-legalization

[4] “Legal Sports Betting Revenue Tracker”

https://www.legalsportsbetting.com/revenue-tracker/

[5] “Sports betting market expected to reach $8 billion by 2025”

https://www.marketwatch.com/story/firms-say-sports-betting-market-to-reach-8-billion-by-2025-2019-11-04

[6] “Nevada Gaming Control Board Monthly Revenue Report February 2020”

https://gaming.nv.gov/modules/showdocument.aspx?documentid=16630

“Beating the Juice: Removing the Vig from Sports Betting Lines” https://www.sportsbettingdime.com/guides/strategy/removing-the-vig/

[8] https://news.gallup.com/poll/4735/sports.aspx

[9] https://www.espn.com/nba/story/\_/id/28887560/nba-suspends-season-further-notice-player-tests-positive-coronavirus

[10] https://www.rdocumentation.org/packages/nbastatR/versions/0.1.10131

[11] https://www.si.com/nba/2013/04/16/boston-marathon-bombing-terror-attack-celtics-pacers-game-cancelled-nba

[12] https://stats.nba.com/articles/nba-com-stats-switches-to-true-possessions-what-it-means-and-why-its-important/

1. <https://sportsbookreviewsonline.com/scoresoddsarchives/> [↑](#footnote-ref-1)
2. Application Programming Interface [↑](#footnote-ref-2)