Sam Cohen-Devries

CUNY School Of Professional Studies

DATA 698

Capstone Thesis Research Paper

Spring 2020

**Introduction**

In recent years, the use of advanced analytics has seen an exponential rise in professional sports. The most famous example was memorialized in the movie Moneyball (2011), which told the story of how a professional baseball team became world champions through the use of metrics, statistics and data science. It was a novel approach at the time, but today just about every professional baseball team has a dedicated team of data scientists. It has spread beyond professional baseball into other sports, especially basketball, but also soccer, cricket and football. In the NBA, the most recent dynasty was built partially on analytics [*citation*] and many of the top teams in the league rely heavily on data science to evaluate and inform game decisions. Historically, and up until quite recently, statistics have been confined to what could be gleaned from a box score, an aggregation of individual and team statistics, at the game or season level. As noted by Cervone et al., this is “akin to analyzing a chess match based only on the move that resulted in checkmate, leaving unexplored the possibility that the key move occurred several turns before” [1]

In the later part of the twentieth century, basketball statisticians began using these aggregated statistics to create more precise measurements of a player’s value. Two of the most popular, which have become a staple of standard box scores, are plus-minus rating (commonly abbreviated as “+/-“), which is a rough calculation of favorable or unfavorable point differential for the time a given player is on the floor [citation]; and player efficiency rating (“PER”), based on a proprietary formula developed by ESPN columnist [citation]. Plus-minus has since been tweaked for greater “accuracy”, into variations such as Adjusted Plus-Minus, which “reflects the impact of each player on his team’s scoring margin after controlling for the strength of every teammate and every opponent during each minute he’s on the court” [2] and Real Plus-Minus, which similarly “isolates the unique plus-minus impact of each NBA player by adjusting for the effects of each teammate and opposing player” [3].

A big change between standard and “advanced” plus-minus scores is the use of play-by-play, or possession-based, rather than aggregation, in its calculations. Another pioneer of possession-based analytics is Dean Oliver, who wrote the book “Basketball On Paper” (2004), which “introduced possession-based stats and [four factors](https://www.nbastuffer.com/analytics101/nba-analytics-movement/analytics101/four-factors/) with the basketball community” [4][5]. In the years since, the NBA has quickly moved towards spatial analysis, using a tool called SportVU [6]. While these recent developments have opened the doors to fascinating new developments in the game and provided opportunities for even more creativity and analysis in decision making, for coaches, players and managers alike, it also presents a problem in terms of keeping a level playing field.

The use of advanced analytics in college sports is restricted mostly to Division I and some Division II schools. The reason for this is mainly due to a lack of funding. Simply acquiring the data used in the metrics requires an entire staff and/or expensive technology and equipment that is not available to less well-funded schools. This is a threat to the idea of a level playing field, but furthermore, creates a wider gap between the higher and lower divisions. On the NCAA website, coaches, fans and analysts can access metrics for almost all Division I teams and some Division II teams. There are play-by-play datasets freely available for NBA and Division I Basketball on popular data science websites such as Kaggle [7][8]. But there is no such data easily available for Division III.

This can create additional problems for student athletes in lower divisions to continue on to a career in athletics. While it is rare for Division II and III players to go on to the NBA, many have historically gone on to pursue careers as players overseas, or to go in to coaching, training, broadcasting and other media. With the rise of analytics in college basketball, those who are not exposed to it face a larger disadvantage professionally than they have in the past.

**Literature Review**

In his research on different mediums for new research on statistical analysis in sports, professor Tim Swartz states that “best sports statistics contributions 1) contain statistical novelty and (2) address a real sporting problem.” These two goals, however, “can sometimes conflict” as you are in effect attempting to reach two often very distinct audiences; 1) the data science community and 2) the sports community. The focus in preparing for this analysis was in looking for novel and creative approaches to play-by-play analysis in basketball that could be used for my analysis. The target audience in this paper would most likely be #2, as we are not developing new techniques for advanced analytics, but rather incorporating existing methods by creating new data sets. In that regards, our target audience would mostly consist of players and coaches in lower divisions, both current and prospective.

In that sense, my aims differ than what can be found in traditional box score data. Those sorts of statistics (including PER and basic Plus-Minus) are more often used to provide insight to an observer, or fan. These types of statistics and metrics can also be useful for bookmakers and gamblers. Many studies have looked at which individual statistics could be best used as predictors for success, generally measured by the amount of games won. These generally use box-score statistics (aggregated). One such example [9] also found differentiations between different leagues, including professional leagues outside of the United States, suggesting that different variables could be used to predict to success, depending on the makeup of the competition. This is something that I am keeping in mind throughout my research, in allowing for the possibility that the statistical conclusions to be found in Division III could be different than those in studies on higher levels, while still maintaining accuracy.

I am looking for insights that can be used to guide decision making for coaches and players, especially those that use a novel approach. One such example comes from the Journal of Economic Behavior and Organization, in a paper titled “Optimal Stopping in the NBA: Sequential Search and the Shot Clock”. In it, concepts of economic theory are applied to shot clock usage, in determining when a player/team should choose to attempt to score versus hold and wait for a better opportunity.

Another is from the Journal of Applied Behavioral Analysis (“Behavioral Momentum in College Basketball”; [F. Charles Mace](https://www.ncbi.nlm.nih.gov/pubmed/?term=Mace%20FC%5BAuthor%5D&cauthor=true&cauthor_uid=16795791) and [Joseph S. Lalli](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lalli%20JS%5BAuthor%5D&cauthor=true&cauthor_uid=16795791)) in which studies on how humans respond to adversity were applied to determining the impact of momentum on college basketball games. The study was based on the hypothesis that a high amount of “reinforcing” (favorable) events within a short time period leading up to an adverse event increases the probability that another favorable event will follow the adverse event. The idea was that this could inform coaches decision making in terms of if and when to call timeouts to break up an opposing team’s momentum. This paper was particularly interesting in that it was a direct challenge on an early, and widely touted, statistical analysis in basketball, questioning the validity of the “hot-hand” theory [10], by suggesting that momentum does in fact play a statistically-verifiable role in basketball.

Both of these studies are useful for overall coaching decisions and strategies. To examine individual player performance, I looked at “Estimating an NBA player’s impact on his team’s chances of winning”, by Sameer Deshpande, published in the Journal of Quantitative Analysis in Sports (2016). Here the author does a deep statistical analysis and determines a formula to create a new player metric to compete with or complement Player Efficiency Rating and Adjusted Plus-Minus by measuring the change in win probability from the time a player enters the game to the time he is substituted out of the game and then summing these changes over the course of a season.

Finally, I draw upon “A Starting Point For Analyzing Basketball Statistics” by Dean Oliver, et al. from the Journal of Quantitative Analysis in Sports (Berkeley Electronic Press, 2007) as a basis for my research. Among other topics, this paper provides a mathematical formula for approximating possession-level statistics using aggregated statistics.

**Hypothesis**

My hypothesis for this study is very straightforward. I am proposing that there are analytical tools readily available to Division III coaches that can guide their decision making, create insights, and ultimately provide a competitive advantage, whereas traditional box scores and metrics based on such can provide little actionable insight. Additionally, I will demonstrate that participants in Division III basketball can use analytical tools similar to those used in a higher and better funded level of competition, thereby providing better preparation for a professional career in sports.

**Data and Variables**

The data source for this research comes from d3hoops.com, a centralized repository for news, information and statistics on Division III basketball. For every game played in Division III (both Men’s and Women’s, although this paper will focus primarily on Men’s teams), there is a written summary, a box score and a play-by-play. I will be using a number of variable to make predictions, but they will mostly stem from the data points provided in the play-by-play. Thus, the heart of the data will reside in a database, where every record contains:

*Game ID*: this will refer to a single matchup between two teams. If two teams face each other more than once, each matchup will have a unique game ID.

*Home Team ID*: the host

*Away Team ID*: the visitor

*Action ID*: this can be any event in a basketball game, including but not limited to missed shot, made shot, turnover, offensive foul, defensive foul and timeout. Shots can be further broken down into three pointers, two point shots, layups and free throws.

*Time*: the timestamp (in relation to start and end of the Game) during which the event occurred

*Player ID*: the individual performing the action; in the case of a timeout this will refer to the head coach

*Time Of Possession*: this is a calculated field using the difference between subsequent timestamps. This data point is critical in determining and assessing behavior and decision making

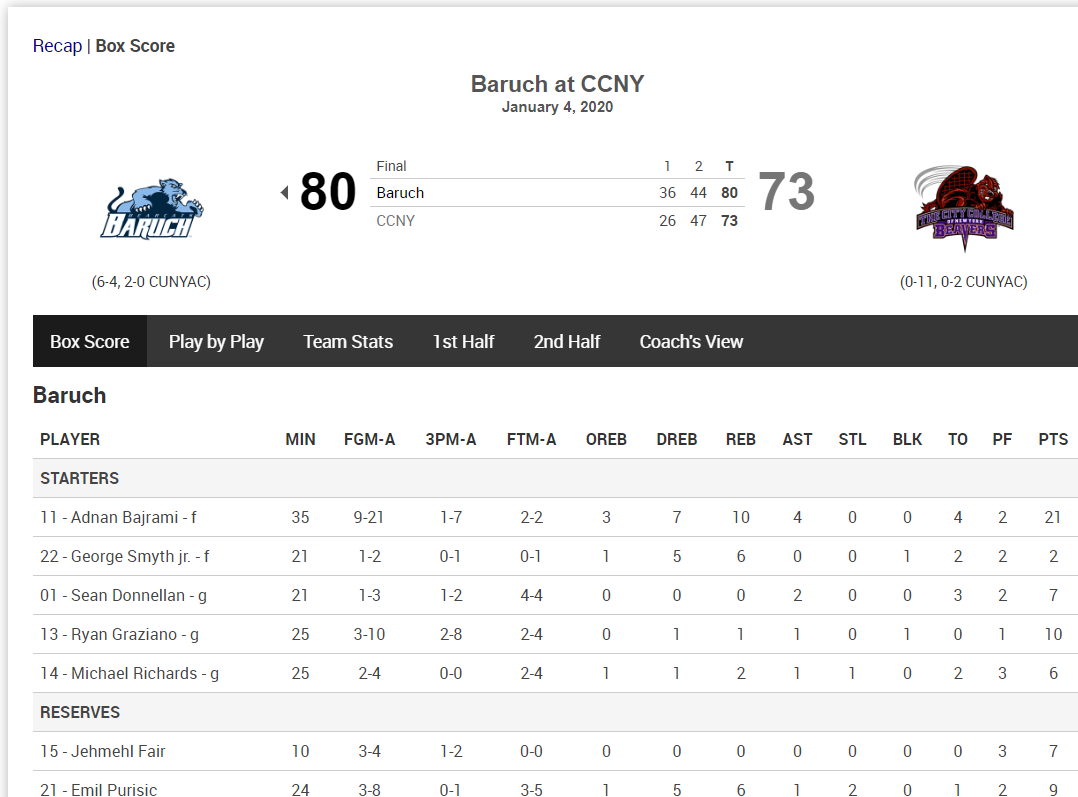
Lineup ID: this is a derived field, and will be gleaned based on substitution data. Every team will have a set of distinct lineups (consisting of five players) which will be both evaluated on performance as well as used to adjust and scale the impact of individual players

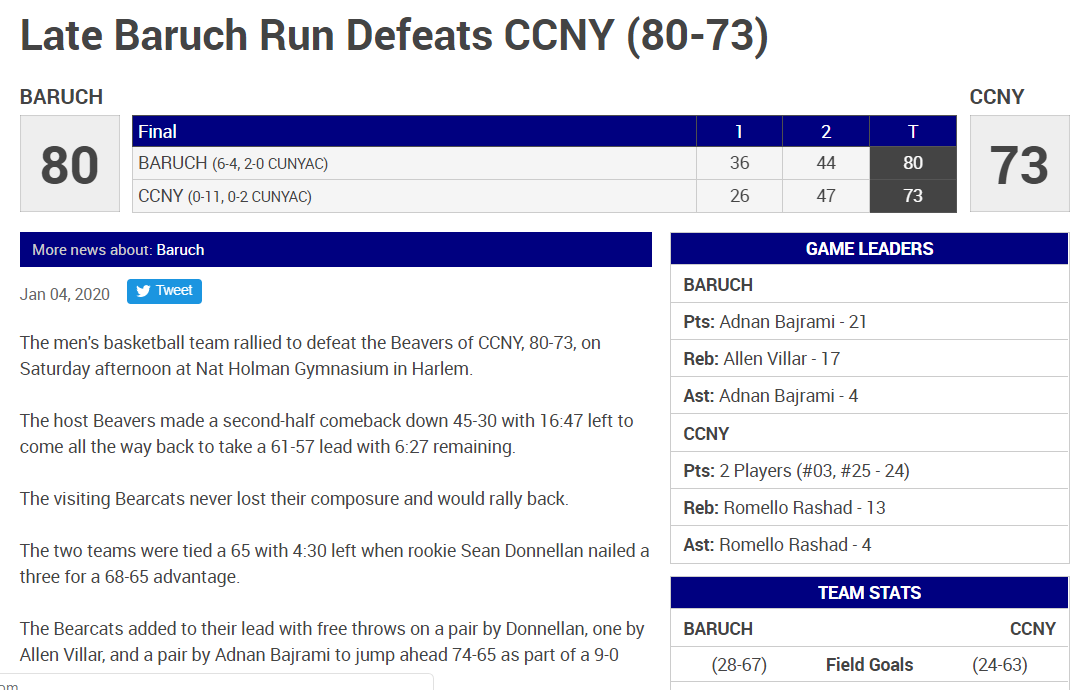
I will store the outcome of each Game ID (winning team only; we are not interested in predicting margin of victory) and the box score, to be used as a null model for comparison.

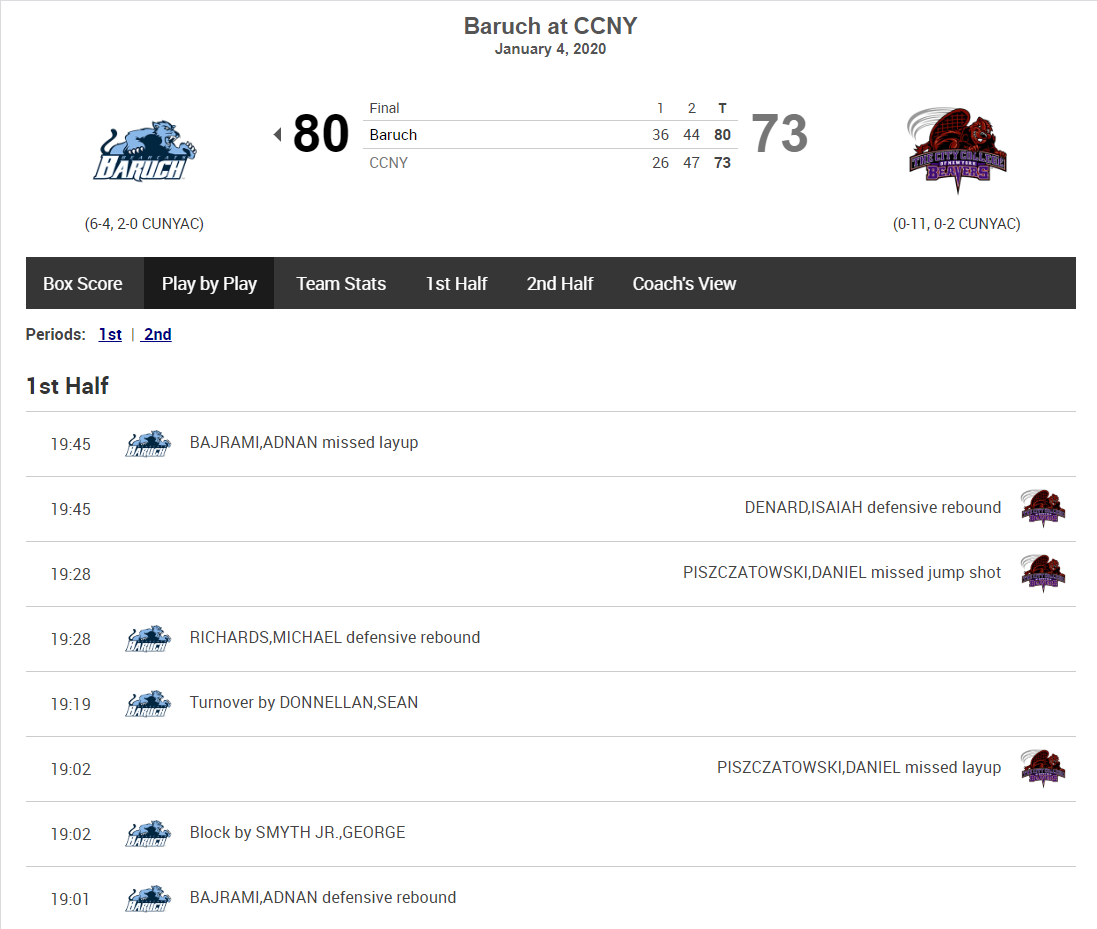
The data will be scraped from the web using the BeautifulSoup (<https://pypi.org/project/beautifulsoup4/>) package in Python and stored in a Relational Database Management System, most likely SQL Lite, possibly using an Object Relational Mapper such as SQLAlchemy (<https://www.sqlalchemy.org/>). I will most likely be using scikit-learn for the prediction models, although multiple methods will be tested and evaluated.

Examples from the source website (d3hoops.com) are given below for a single game featuring Baruch College vs. City College of New York, both participants in the City University Of New York Athletic Conference.

Box score:



Written Summary (Recap):

Play-by-Play:

**Statistical Methods**

Most of the statistical methods employed in my research will be derived from prior research efforts, some of which have already been mentioned here, and which are recognized as having value to basketball decision making and will focus on three areas.

1. Shot clock usage
   1. As mentioned previously, we will be calculating time of possession for each event. Based on the standard 35-second shot clock used in college sports, we can evaluate shot-clock usage and efficiency, similar to the analysis performed by Goldman and Rao. By breaking the clock down into buckets (discrete values will be to sparse to provide meaningful results), we should be able to accurately predict the outcome of a based on the elapsed time in the shot clock. Note that for this method, we are using an alternative definition of “possession”, starting a new possession every time the shot clock is reset, as opposed to defining a possession as when the ball changes hands.
2. Time out usage
   1. Expanding upon research by Mace and Lalli, I will be creating a model to predict the efficiency of a timeout in slowing the momentum of an opposing team. The model will use a timeout event, along with the adverse and reinforcing effects of the previous *n* minutes to predict the number of adverse and reinforcing effects in the subsequent *m* minutes.
3. Lineup usage
   1. Drawing off of research by Joseph Sill [11], I will attempt to create a regression model using a lineup bit matrix, with 1 indicating a player is on the floor, 0 indicating a player is on the bench and -1 indicating that an opposing player is on the floor. This is intended to weight player effectiveness more accurately than simple plus minus by taking into account the effectiveness of the opposing player.

The model will be scored based on its ability to accurately predict the overall outcome (based on point differential) for a game “stint.” A stint is referred to as the time period between lineup changes. For example, let home team’s lineup *l* at time *t=0* (start of the game), when the home team score *H* = 0 and the away team score *A* = 0, giving a margin *m1* = 0. be a vector of length 5, consisting of player IDs [1,2,3,4,5]. This is lineup ID HL1. A substitution is made at time *t=100* and now *l* = [1,2,3,6,8], *H* = 8 and *A* = 4, giving a margin of *m2* = 4. That concludes one snippet, where HL1 receives a score of +4, calculated by *m2 – m1*.

* 1. Each player will be assigned both a basic plus-minus score based off of aggregated box score data, as well as an Adjusted Plus-Minus, based off of possession-based play-by-play data, using the web scraper. Our Adjusted Plus-Minus scores will be evaluated based on how well they are able to predict the outcome of a snippet, compared to the null model.

**Findings**

The outcome of the data scraping operation was only successful at the individual team level, and with some manual intervention.

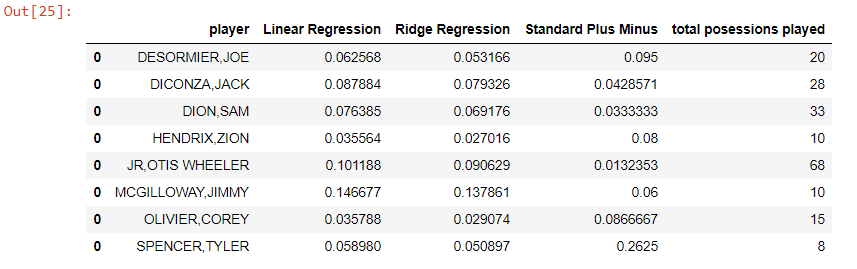
The HTML structure of the play-by-play was fairly reliable, based on the following criteria, after performing some data cleaning operations:

* Player names were consistent within and across games
* Action types had few outliers (see below; distinct action types and counts)
* Substitutions were marked such that our model consistently found 5 players on the floor per team

Where the scraper was ineffective was in automating the scraping process on a per team basis. The reason for this was that in the Game Scores landing page, showing all games on a given date, as well as in the individual play-by-plays, there was inconsistent patterning in how the Home and Away team were identified. In the play-by-play, each “action” HTML node is classified as Home/Away. Given that there was no way found to automatically tie a specific team to either side, the model was run on a team-by-team basis by manually pulling all game identifiers for an individually season into a list, and marking 1/0 in a separate list for if the team was home or away in that game, i.e.:



For the above example, the model output the coefficients for each player, which can be interpreted as the expected value per possession that an individual player would contribute while on the floor.



Running this for 5 separate teams in the CUNY Athletic Conference, our model showed better results for the standard Linear Regression model, as opposed to the Ridge regression model proposed by Sill. One possibility for this is the lack of extensive ratings for opposing players, which would be made possible by an automated solution for scraping all games for all teams. The R^2 coefficient ranged between .42 and .63 for the Ridge regression model and .44 and .66 for the standard Linear Regression model.



I used a KNearestNeighbors classifier to analyze shot clock usage, using the duration of a possession and the current margin, which returned a mean accuracy rate between 60-70% for the current possession. More interestingly, the accuracy rate for the *subsequent* possession was slightly lower, but still in the same range, which can provide a good metric for coaches in determining when to push pace and when to hold back.

**Conclusion**

This process does have potential for coaches to be able to use as a strategic tool. It will be somewhat more time consuming, and may require some technical assistance, however these should be resources that most colleges/universities have available. The ultimate goal of basketball analytics should be to inform decisions, not make them. While this model is not strong enough to automate decision making (i.e. optimal lineups based on opponent), there is enough valuable data extracted on a team-by-team basis to help guide decisions on adjusting playing time on an individual player basis. Similarly, the model does not give an optimized metric for shot clock usage, but it does provide enough information and accuracy to help inform a coach’s strategy.

References:

[1] *POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data*; Sloan Sports Analytics; Dan Cervone, Alexander D’Amour, Luke Bornn, and Kirk Goldsberry (Harvard University) 2014

[2] Adjusted Plus-Minus; NBAStuffer; <https://www.nbastuffer.com/analytics101/adjusted-plus-minus/>

[3] The Next Big Thing: Real Plus Minus; ESPN.COM; <https://www.espn.com/nba/story/_/id/10740818/introducing-real-plus-minus>

[4] Analytics Movement; NBAStuffer; <https://www.nbastuffer.com/analytics101/nba-analytics-movement/>

[5] Four Factors; NBAStuffer; <https://www.nbastuffer.com/analytics101/four-factors/>

[6] Most Innovative Companies: STATS; FastCompany; <https://www.fastcompany.com/company/stats>

[7] <https://www.kaggle.com/schmadam97/nba-playbyplay-data-20182019>

[8] <https://www.kaggle.com/ncaa/ncaa-basketball>

[9] *Investigating the game-related statistics and tactical profile in NCAA division I men’s basketball games*; Daniele Conte, et al; Biology of Sport; 2017

[10] *The Hot Hand in Basketball: On the Misperception of Random Sequences*; Thomas Gilovich et al.; Cognitive Psychology (1985)

[11] *Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing*; Joseph Sill; MIT Sloan Sports Analytics Conference; 2010