



A Starting Point for Navigating the World of Daily Fantasy Basketball

Charles South, Ryan Elmore, Andrew Clarage, Rob Sickorez & Jing Cao

To cite this article: Charles South, Ryan Elmore, Andrew Clarage, Rob Sickorez & Jing Cao (2019) A Starting Point for Navigating the World of Daily Fantasy Basketball, The American Statistician, 73:2, 179-185, DOI: [10.1080/00031305.2017.1401559](https://doi.org/10.1080/00031305.2017.1401559)

To link to this article: <https://doi.org/10.1080/00031305.2017.1401559>



View supplementary material [↗](#)



Published online: 11 Jun 2018.



Submit your article to this journal [↗](#)



Article views: 643



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



A Starting Point for Navigating the World of Daily Fantasy Basketball

Charles South^a, Ryan Elmore^b, Andrew Clarage^a, Rob Sickorez^a, and Jing Cao^a

^aSouthern Methodist University, Dallas, TX; ^bUniversity of Denver, Denver, CO

ABSTRACT

Fantasy sports, particularly the daily variety in which new lineups are selected each day, are a rapidly growing industry. The two largest companies in the daily fantasy business, DraftKings and FanDuel, have been valued as high as \$2 billion. This research focuses on the development of a complete system for daily fantasy basketball, including both the prediction of player performance and the construction of a team. First, a Bayesian random effects model is used to predict an aggregate measure of daily NBA player performance. The predictions are then used to construct teams under the constraints of the game, typically related to a fictional salary cap and player positions. Permutation based and *K*-nearest neighbors approaches are compared in terms of the identification of “successful” teams—those who would be competitive more often than not based on historical data. We demonstrate the efficacy of our system by comparing our predictions to those from a well-known analytics website, and by simulating daily competitions over the course of the 2015–2016 season. Our results show an expected profit of approximately \$9,000 on an initial \$500 investment using the *K*-nearest neighbors approach, a 36% increase relative to using the permutation-based approach alone. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received March 2017
Revised October 2017

KEYWORDS

Bayesian statistics; *K*-Nearest neighbors; Lasso; NBA; Random effects

1. Introduction

Participating in sporting events—as a competitor or a spectator—is one of the most popular pastimes across many societies in today’s world. Over the last 10–15 years, a new method of participation has exploded in popularity: fantasy sports. Fans are able to select players from their favorite teams, compile points based on their players’ performances in real time, and “compete” against other fans who have built teams of their own. It gives fans a feeling of being closer to the players and the games they play.


This research focuses on a particular type of fantasy sports—the daily variety. To call this industry big business may be an understatement; the two largest companies that facilitate daily fantasy sports—DraftKings and FanDuel—have raised hundreds of millions of dollars from investors¹ and have held valuations on the order of \$2 billion dollars. While this explosive growth has not come without some controversy², the top players reportedly earn hundreds of thousands of dollars in winnings³ each year.

As the competitive nature of fantasy sports has evolved along with the game itself, so has the predictive analytics. Numerous entities, both professional and amateur, have attempted to build models and predict how players will perform based on their underlying profiles. This research focuses on modeling player performance in one sport in particular—professional basketball. The combination of a reasonable sample size each season (NBA teams play 82 games, compared to just 16 in the NFL)


along with a reasonable amount of data generated during each game (as opposed to baseball, where most batters only get 3–4 opportunities per game to compile statistics) allow for more flexible analyses.

However, limited research has been published in peer-reviewed journals revolving around the daily performance of NBA players. Casals and Martinez (2013) used mixed models with random effects to study variables that influence both the number of points scored by players and win score (a linear combination of counting statistics that is an indicator of player performance). They used a filtering process to create a balanced study design with repeated measures, resulting in only 27 players (who played 81 games) in their study. Further, they only fit a single model using all the data at once rather than fitting daily models and tracking performance. There have also been attempts at quantifying player ability in a more generic way. For some examples, see Kubatko et al. (2007), Page, Barney, and McGuire (2013), Page and Quintana (2015), Fearnhead and Taylor (2011), Piette, Pham, and Anand (2011), Arkes and Martinez (2011), and Entine and Small (2008).

In this article, we explore a dynamic modeling approach using box score data from the 2013–2014 NBA season, with a linear combination of box score statistics (referred to as fantasy points, or FP) functioning as the response variable of interest. We demonstrate empirically that a Bayesian model with player specific intercepts, slopes, and variances can outperform predictions from a well-known predictive analytics website. Next,

CONTACT Charles South  csouth@smu.edu  Southern Methodist University, P. O. Box 750332, Dallas, TX 75275-0332.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS

¹<https://www.bostonglobe.com/business/2017/03/09/another-million-financing-for-draftkings/oybStfPL1XPfA0jRcDUTJ/story.html>

²<https://www.nytimes.com/2015/10/06/sports/fanduel-draftkings-fantasy-employees-bet-rivals.html?mcubz=0>

³<https://www.bloomberg.com/news/articles/2015-09-10/you-aren-t-good-enough-to-win-money-playing-daily-fantasy-football>

we proceed to the idea of identifying strong lineups: we use a permutation approach to generate an initial pool of lineups, then consider the likelihood that the lineups will exceed a predetermined FP threshold by using k -nearest neighbors to recalibrate this likelihood. The end goal is to identify the best subset of lineups that a user might play on a given night (say, the top 100). The entire system is then implemented in a hypothetical experiment using data from the 2015–2016 season. We show that for an initial investment of \$500, our system would have generated an expected profit of approximately \$9000 over the course of the season.

There are two primary contributions of this work to the literature. First is the presentation of a transparent, complete approach to lineup selection in daily fantasy basketball. That is, we propose an automated data collection and modeling system that generates a pool of potentially successful lineups using freely available box score and betting line data. The overwhelming majority of predictive analytics surrounding daily fantasy sports is proprietary, so the growing daily fantasy community could benefit from a clear look at the performance of modern analytic techniques applied to this problem. We show empirically that our model is competitive relative to those in the industry. A secondary aim of this research is for more general basketball fans or professionals working for NBA teams who want to better understand the driving force behind daily player performance; this audience will primarily be interested in the first half of the paper. R code and data used to produce this research will be made available as supplementary materials.

2. Modeling Fantasy Points

2.1. Daily Fantasy Basketball Background

The objective of fantasy sports is to allow fans to pick a team of real players and compete using the statistics they accrue during games. While there are many websites that host daily fantasy games, the main rules of the game remain the same: under the constraint of a fictional salary cap, pick a set of players that accrue points based on some aggregate measure of observable statistics; the main differences between the websites are in the fictional dollar amounts (and, as a result, player costs) and the weights used to calculate each player's score.

The website whose rules this research used is DraftKings,⁴ arguably one of the two biggest daily fantasy game hosts in the market. Their measure of a player score—call it “fantasy points” (FP)—is defined as

$$FP = Z_1 + 1.25Z_2 + 0.5Z_3 + 1.5Z_4 + 2Z_5 + 2Z_6 - 0.5Z_7,$$

where the variables are as follows:

- Z_1 = total number of points scored by the player,
- Z_2 = total number of rebounds grabbed by the player,
- Z_3 = total number of three-point field goals made by the player,
- Z_4 = total number of assists (a pass to a teammate who immediately scores) made by the player,

- Z_5 = total number of steals (legal takeaways from the opposing team) by the player,
- Z_6 = total number of blocks (shot attempts by the opposing team swatted away) by the player,
- Z_7 = total number of turnovers (giveaways to the other team) by the player.

The FP metric is designed to represent the overall performance by a player on a given day, with the goal of the daily fantasy game being to select a team of NBA players with the largest FP sum for the day. However, the team must be constructed in a way that mirrors a real team—one each of the traditional five players must be selected (point guard, shooting guard, small forward, power forward, center), as well as an additional guard (either point or shooting), forward (either small or power), and utility (any of the five positions). The team of players must be selected from at least two NBA teams, *and* they must come from at least two actual NBA games being playing on the day of interest.

The DFS player choosing the team then decides on the type of daily fantasy game to play as well as the amount of money he/she would like to wager. The two most common types of games are “50–50” and tournament style; in “50–50” games, half of the entrants win money—equivalent to almost double their entry fee—and in tournament style games only the top 20% of participants win money, with the amounts heavily skewed toward the top few places. Typically, fans can enter a large number of teams in tournament style games if they wish, but the max number of entries in the 50–50 games varies (some allow multiple entries, some only allow 1). At the end of the day, the sum of the FP for his/her eight players is compared to all other FP sums of fans that entered the same game and a winner is declared.

2.2. Data

In order to build our models, we scraped player- and team-specific variables for a subset of games played in the 2013–2014 NBA season from www.basketball-reference.com. A full list of these variables is given by South (2016). Data were gathered for a total of 349 players and all 30 teams, with the analysis starting on January 20th (the approximate midpoint of the season) and ending on April 16th (the last day of the season) to ensure that player performances were relatively stable and that teams had established the distribution of minutes among their players.

2.3. The Baseline Model

Because the goal was to predict the total fantasy points scored by player i on day j , each player's previous games had to be summarized succinctly. Perhaps the simplest way to think about a player's performance on a given day is to consider two things: how he has performed in the past and how strong his current opponent is. Thus, define the baseline model for predicting y_{ij} , the fantasy points scored by player i ($i = 1, \dots, n_j$, where n_j is the number of players available for analysis on day j , where $j = 1, \dots, 81$), as:

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij}, \quad (1)$$

⁴ The authors have no relationship with this company, nor any company that facilitates daily fantasy sports. Their system was chosen arbitrarily for use in this research.

Table 1. Variables retained at in at least 50% of cross-validated lasso models.

Variable	Percent of Models
10-game moving average of fantasy points (X_1)	95.1
Game Started Indicator (X_2)	88.9
10-game moving average of turnovers (X_4)	76.5
10-game moving average of defensive rebounds (X_5)	75.3
10-game moving average of free throw attempts (X_6)	58.0
10-game moving average of field goal attempts (X_7)	51.9

where X_{1ij} is a moving average of fantasy points scored by player i updated through day $j - 1$ and X_{2ij} is a moving average of total fantasy points allowed by player i 's opposing team updated through day $j - 1$. For example, if player i 's current opposing team is the Dallas Mavericks, then X_{2ij} would be a moving average of the sum of all individual FP totals the Mavericks had allowed against previous opponents. Note, however, that for any particular day, every team is NOT guaranteed to have played the same number of games. The window size used for the moving averages was 10 games; for more details on how this was determined, see South (2016).

To evaluate model performance, the median absolute prediction error for day j , or MAPE, is defined as

$$\text{MAPE}_j = \text{median}(|y_{ij} - \hat{y}_{ij}|).$$

The idea behind the use of MAPE rather than a traditional residual is a matter of interpretability: irrespective of direction, we wanted to have a measure of center that summarized how close the daily predictions were. We also tracked the standard deviation of the absolute prediction error (SDAPE) to evaluate the consistency of the model.

2.4. Variable Selection Using the Lasso

A fairly large pool of potentially highly correlated variables is available for this analysis and, hence, determining which variables contained the most signal can be challenging. To address this, we used the lasso (Tibshirani 1996) to conduct variable selection. The same process was used to fit models each day, replacing the baseline model with a lasso model fit. However, the fit was sensitive to changes in the data that occurred when daily updating the data frames. To address this sensitivity, we tracked the percentage of lasso models that retained each variable across all days in the study. Upon examination, the variable pool was reduced to contain only those that were retained in at least 25% of the daily models; this reduction removed approximately half of the variables from consideration. The reduced pool was then re-processed through the daily mechanics to make predictions and once again the percentage of models retaining each variable was tracked. Six variables were retained in over 50% of the daily lasso models, summarized in Table 1.

All of these variables are related to volume and aggressiveness. Players with a large number of free throws attempted and field goals attempted have the ball in their hands often and are integral parts of their teams' offense, as do players who turn the ball over often. Players who start the game generally play more minutes than players who come off the bench, allowing more opportunities to contribute and, as a result, increase their fantasy point total. Note that X_2 , the 10-game moving average of fantasy

points allowed by the current opponent, was not retained in a large percentage of the models.

2.5. The Bayesian Model

One of the major problems that arose during the research was the variability in player performance, particularly for players who could be labeled "stars." Figure 1 shows a scatterplot of median fantasy score versus standard deviation of fantasy score, with individual points representing players; the positive monotonic trend indicates that players with larger median scores tended to have larger standard deviations of FP totals.

Casals and Martinez (2013) added random effects to their mixed models to account for player-to-player variability. The Bayesian paradigm provides a simple framework for accounting for this phenomenon by not only allowing specification of individual variances, but also allowing for individual slopes and intercepts. Our proposed model is:

$$y_{ij} \sim \text{Normal}(\mu_{ij}, \tau_i), \quad i = 1, \dots, n_j; \quad j = 1, \dots, 81,$$

$$\mu_{ij} = \beta_{0i} + \beta_{1i}X_{1ij} + \beta_{3i}X_{3ij} + \beta_{4i}X_{4ij} + \beta_{5i}X_{5ij} \\ + \beta_{6i}X_{6ij} + \beta_{7i}X_{7ij} + \sum_{p=8}^9 \beta_{pi}X_{pij},$$

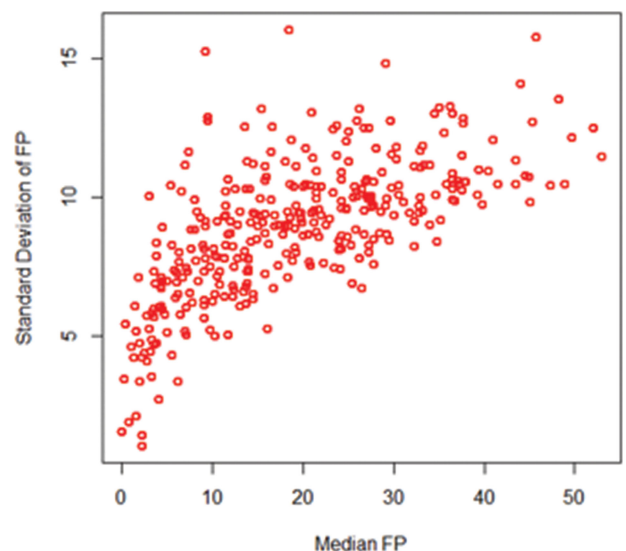
$$\beta_{pi} \sim \text{Normal}(\mu_p, \gamma_p), \quad p = 0, 1, 3, \dots, 9,$$

$$\mu_p \sim \text{Normal}(0, 0.001),$$

$$\gamma_p \sim \text{Gamma}(1, 1),$$

$$\tau_i \sim \text{Gamma}(1, 1),$$

where X_{1ij} , X_{3ij} , \dots , X_{7ij} were the lasso selected variables and X_{8ij} and X_{9ij} represent two indicator variables for opponent strength. Also, note that conventional noninformative priors were assigned to the regression parameters. Recall that the fantasy points allowed variable (i.e., X_2) was not one of the most frequently retained variables in the lasso approach despite the fact that it is well known that opponent strength impacts a player's performance; this suggested the need for a more sophisticated method of quantifying opponent strength. A different approach

**Figure 1.** Scatterplot of standard deviation of fantasy point total vs. median fantasy point total.

to capturing opponent strength was to use subjective clustering: for each player and each day prior to the beginning of the analysis period, we calculated the difference between actual fantasy points scored and the previous 10-day average of fantasy points scored for each player in each game. This difference can be thought of as a measure of opponent strength; for example, suppose player A had a previous 10 game average of 31 fantasy points, but only scored 26 points his next game. That is, his current opponent held him 5 fantasy points below his recent average, suggesting the opponent may have a strong defense. These differences were gathered separately for each team, and the median difference was calculated. From these medians, three teams had a median of 1.5 or higher, five teams had a median of -1.5 or lower, and 22 teams had a median between -1.5 and 1.5 . This criterion was used to split the teams in to three clusters. Indicator variables were assigned to two of the three clusters (X_{8ij} and X_{9ij}), with the third being used as a reference. Because we felt it crucial to add some control of opponent strength to the model, these variables were added after applying the lasso filtering and switching to the Bayesian paradigm, and (though not reported here) did improve the model fit. A combination of R (R Core Team 2016) and WinBUGS (Lunn et al. 2000) were used to fit the proposed model.

2.6. Model Summary

Beginning with the fundamental concept of player strength versus his opponent's strength, a baseline model was developed and expanded upon in the Bayesian paradigm, incorporating the most frequently selected lasso variables and player-specific slopes, variances, and intercepts. The average MAPE and SDAPE from the two models are summarized in Table 2. The Bayesian model provided a reduction in both MAPE and SDAPE, and model diagnostics suggested a reasonable model fit even in the context of the occasional outstanding individual performance. While the reduction is modest, it is often the case that in larger daily fantasy tournaments, one-quarter of a point can make the difference between winning and losing large sums of money; considering that the reduction held over the course of the 81 day sample, we feel the Bayesian model is the more viable option.

3. Lineup Construction

3.1. Strategy

Generating predictions is the first major step when playing daily fantasy basketball; the second step is constructing lineups with a strong probability of success. A "lineup" is simply a collection of players chosen under the constraints defined by the host of the game. Most daily fantasy games follow a similar structure—participants must choose a lineup of NBA players based on their

actual positions⁵ and based on fictional salaries that are assigned by the host. As previously mentioned, this research followed the system developed by DraftKings. Additionally, a fictional salary cap of \$50,000 is imposed, meaning that the sum of the salaries from the eight players mentioned in Section 2.1 cannot exceed this amount.

We gathered the entire season's worth of salary data from the 2015–2016 NBA season to use for analysis.⁶ Each day, we began by generating a pool of lineups that had the largest probabilities of exceeding a particular FP threshold according to the player-level predictions. However, South (2016) found that by incorporating lineup-level metrics, a k -nearest neighbors approach (Altman 1992) could be used to recalibrate the probabilities of exceeding the FP threshold, and re-order the initial pool of lineups to improve the overall quality of the top lineups. The efficacy of this approach—along with the rest of the system—was tested in a hypothetical demonstration of a series of \$5, 50–50 contests using the 2015–2016 data.

3.2. Permutation-Based Lineup Construction

An obvious way to choose a lineup would be to permute all possible combinations of players and simply choose the lineup with the largest predicted point total under the salary constraint. However, this does not take player variability into account—two lineups that have the same predicted point total may have very different lineup variability, which can be assumed to be the sum of that from the individual players composing the lineups. An alternative approach is to incorporate both lineup point total and player variability by maximizing the probability that a lineup will exceed a predetermined threshold. In this research, the threshold was set at 260 points to mimic the reported approximate average cash score of 50/50 games from the 2014–15 season on DraftKings.⁷

An additional benefit of the Bayesian model (2) is that the player-level variances and prediction variances can be saved from the MCMC chain. So, we can easily estimate $P(\sum_{i=1}^8 y_{ij} > 260)$. While it is not computationally feasible to permute every possible lineup, a simplified approach is to use truncated permutations to generate subsets of viable lineups. We employed the following algorithm to generate potential lineups each day:

1. Generate all combinations of point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C).
2. Calculate the combined salary total for each lineup. Delete all lineups with a salary total exceeding \$41,000 or not exceeding \$17,500.
3. Retain the top $x\%$ (with the percent being dependent on the number of games being played and CPU constraints) of the remaining lineups with respect to $P(\sum_{i=1}^5 y_{ij} > 162.5)$.

Table 2. Summary of Section 2 models.

Model	MAPE	SDAPE
Baseline	6.06	6.22
Bayesian Random Slopes/Intercepts Model	5.98	5.96

⁵ There are five positions: point guard, shooting guard, small forward, power forward, and center.

⁶ An example website: <http://rotoguru1.com/cgi-bin/hyday.pl?mon=10&day=31&year=2015&game=dk>

⁷ <https://playbook.draftkings.com/nba/nba-all-star-lesson-05-tournament-vs-h2h-lineup-building>

Table 3. Top 6 Generated Lineups for 11/18/2015.

PG	SG	SF	PF	C	G	F	UTIL	Salary	Prob.	Proj.	Actual
K. LOWRY	R. HOOD	P. GEORGE	T. YOUNG	I. MAHINMI	E. PAYTON	E. FOURNIER	D. SCHRODER	50000	0.958	280.53	318.75
E. PAYTON	R. HOOD	P. GEORGE	PAUL MILLSAP	I. MAHINMI	D. SCHRODER	E. FOURNIER	T. YOUNG	49700	0.948	280.25	340
K. LOWRY	R. HOOD	P. GEORGE	PAUL MILLSAP	L. ALLEN	Z. LAVINE	T. YOUNG	D. SCHRODER	50000	0.929	287.25	304
K. LOWRY	R. HOOD	P. GEORGE	T. YOUNG	I. MAHINMI	E. PAYTON	P. MILLSAP	S. LARKIN	49700	0.929	280.03	318.75
E. PAYTON	R. HOOD	P. GEORGE	T. YOUNG	J. SULLINGER	D. SCHRODER	P. MILLSAP	I. MAHINMI	49200	0.925	281.50	338.5
K. LOWRY	R. HOOD	P. GEORGE	P. MILLSAP	I. MAHINMI	S. LARKIN	T. YOUNG	J. SULLINGER	50000	0.921	282.92	310.5

4. Permute the lineups from Step 3 with all possible guard⁸ options.
5. Calculate the combined salary total for each lineup. Delete all lineups with a salary total exceeding \$44,000 or not exceeding \$25,000.
6. Retain the top $x\%$ of the remaining lineups with respect to $P(\sum_{i=1}^6 y_{ij} > 195)$.
7. Permute the lineups from Step 6 with all possible forward options.
8. Calculate the combined salary total for each lineup. Delete all lineups with a salary total exceeding \$47,000 or not exceeding \$35,000.
9. Retain the top $x\%$ of the remaining lineups with respect to $P(\sum_{i=1}^7 y_{ij} > 227.5)$.
10. Permute the lineups from Step 9 with all possible utility options.
11. Retain all lineups with salary total below \$50,000.

The salary ceiling restrictions were based on the fact that the minimum assigned salary by DraftKings is \$3,000. So, for example, if a lineup from Step 1 exceeded \$41,000 then it would not be possible to fill the last three player slots even with minimum salary players. The salary floors were chosen assuming that any lineup built primarily of players costing in the \$4,000 range would not be viable, as salaries are typically correlated with recent performance. The point cutoff values (i.e., 162.5, 195, and 227.5) were determined by assuming an average of 32.5 FP per player (i.e., 260/8) and multiplying by the number of players in the corresponding step. The percentages retained in Steps 3, 6, and 9 varied depending on the number of players. However, they were most often very small—between 5% and 10%—and sometimes less on days when almost every NBA team had a game. This still resulted in tens of thousands of lineups, but for the purpose of the research only the top 1000 lineups were retained each day. As an example, Table 3 shows output of the top 6 lineups generated for November 18, 2015. The “Prob” column is the estimated probability of exceeding 260 points (note that these values were very high towards the beginning date of the analysis and gradually became more conservative as time passed). All the lineups included Paul George—a superstar with an extremely high ceiling but who occasionally plays poorly—and Rodney Hood—a lower dollar player projected to be a good value. This was a particularly strong day, as all 6 lineups vastly exceeded projections.

3.3. Classification-Based Lineup Construction

Previously, we were concerned with attributes at the *player* level (such as the average number of fantasy points scored over the

previous 10 days), whereas now the main concern shifts to attributes at the *lineup* level. That is, what properties did successful lineups possess that all others did not? While the permutation approach does use information about the expected mean and variance components generated from the Bayesian model, alone it is not flexible enough to determine whether additional signal may be gained by re-considering the available data at the lineup level. To accomplish this, we consider a classification method—that is, identifying “successful” lineups that exceed 260 points by creating a binary outcome and using lineups generated in previous days as training data. We considered lineup averages of six player-specific metrics as predictors for our classification method. These variables were:

1. *Line*—this is a number assigned by bookmakers that is an attempt to estimate how close the two teams are. For example, Dallas played San Antonio on October 28th, 2014. The line was 3.5 in favor of San Antonio. In order for someone placing a wager to win, San Antonio would have to win by more than 3.5 points. In the event of a large line, a star player for the favored team may not play as many minutes as his team is expected to win easily.
2. *Over/Under (OU)*—this is another number assigned by bookmakers. It is the projected combined final score between two teams. In the previous example, the over/under was 203. The final score was actually 101 (San Antonio) to 100 (Dallas), making the total 201.
3. *Average Fantasy Points (last 10)*—this is the same metric used in Model (1).
4. *Average Fantasy Allowed (last 10)*—this is the same metric used in Model (1).
5. *Value*—this is based on a player’s salary relative to his projected point total. Specifically:

$$\text{Value} = \frac{\text{Predicted Fantasy Points}}{\text{Salary}}.$$

This estimate of value is only as good as the model predictions, but it is still a reasonable indicator of how useful a player should be on a given day.

6. *Player-Level Standard Deviation*—this is the value of the estimated $1/\sqrt{\tau_i}$ from Model (2). The larger the standard deviation, the more variable the fantasy point totals. The reason for including this as a metric is to identify lineups that are potentially volatile.

The training data for the KNN algorithm was built by storing all groups of 1000 permuted lineups for all days prior to the day of interest, as well as their outcomes. To re-calibrate the probability of exceeding 260 points for the permuted lineups in the current day of interest, a KNN algorithm was used to find the nearest neighbors in the training data with the distance

⁸ This could just as easily be forward—the order does not matter.

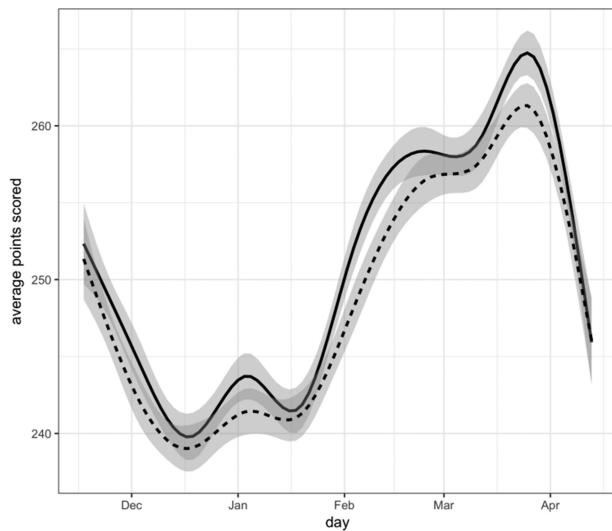


Figure 2. Smoothed averages of the permutation (dashes) and KNN (solid) approaches, 2015–2016 season.

being calculated using the six aforementioned lineup level metrics. The likelihood of exceeding 260 points was estimated to be the percentage of nearest neighbors in the training data who *did* exceed 260 points.

Following the results from South (2016), the KNN and permutation (PM) approaches were implemented using the data from the 2015–2016 season. For each day from November 17, 2015, to the end of the regular season on April 13, 2016, we first generated 1000 lineups using the permutation approach (the start date was chosen to allow a burn-in period; for any players who had yet to accrue 10 games worth of data, moving averages were composed of all available data) and selected 100 random lineups to compose our PM set. In addition, we used the KNN approach to re-order the 1000 lineups, then retained the top 100 as our KNN set. The smoothed averages of each method's lineups are shown in Figure 2; the KNN approach showed modest improvement over the permutation approach when viewed daily.

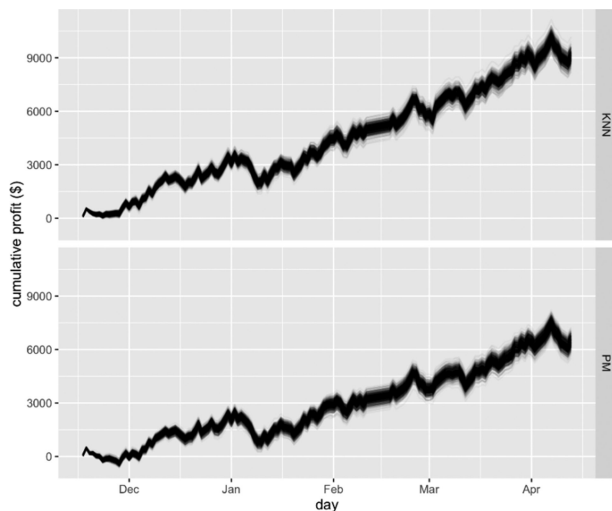


Figure 3. Cumulative profit/loss curves ($n = 500$ replications) for DraftKings 50/50 Experiment, 2015–2016 season.

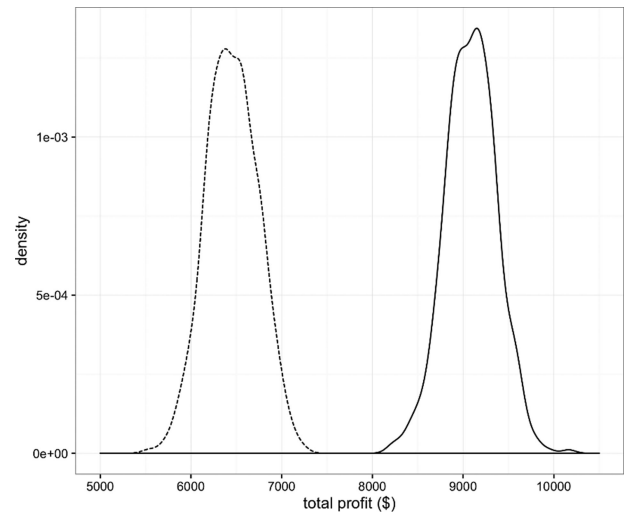


Figure 4. Distribution of total simulated profits ($n = 500$ replications), PM (dashed) and KNN (solid) approaches, 2015–2016 season.

As a hypothetical experiment to test the efficacy of the entire system, we simulated 100 daily, 10-player 50/50 games, each costing \$5. The top five finishers in a 10-player 50/50 game will net \$4 and the remaining players lose their \$5 entry fee. In our experiment, we entered the aforementioned KNN and PM sets and assigned the other nine “players” a point total from the random lineup set. Therefore, we were (hypothetically) at risk of losing \$500 or winning \$400 on a nightly basis under each lineup construction scenario. The experiment was repeated 500 times in order to get a sense of the variability of the process. The cumulative profit/loss curves for each method are presented in Figure 3. As can be seen, each method ends with a substantial profit, though the PM approach operates at a loss for several days uniformly across all simulations. The KNN approach generates a significantly greater profit relative to the PM approach, with average total profits simulated to be \$9,088 ($s = \284.1) and \$6,453 ($s = \284.0), respectively. The distributions of total profits across all 500 simulations for both methods are illustrated in Figure 4.

4. Conclusion

The goal of this research was to create an automated framework for generating player predictions, and use them to construct a subset of potentially successful lineups in daily fantasy basketball games. Specifically, daily player performance is predicted using a Bayesian regression model with player-specific slopes, intercepts, and variances that includes average FP scored over the last 10 games, whether the player started the game, average turnovers over the last 10 games, average defensive rebounds over the last 10 games, average free throws attempted over the last 10 games, and average field goals attempted over the last 10 games, and categorical variables for opponent strength as predictors. A truncated permutation approach and k -nearest neighbor classification using average line, over/under, average FP scored over the last 10 games, average FP allowed by the current opponent over the last 10 games, value, and player-level standard deviations as predictors were found to give the best chance at identifying successful lineups.

To our knowledge, there is not a published method in the literature that addresses this specific question, leaving us no baseline for comparison. However, there are privately owned companies that generate predictions of daily NBA player performance. To show the efficacy of the proposed Bayesian model with respect to player level predictions of FP, we compared our results with the FP predictions provided by www.numberfire.com, a popular analytics website that has worked with ESPN and *Sports Illustrated*. Over a 93-day sample from the 2015–2016 season carried out in real-time, our MAPE was lower than theirs 61% of the time. While this is only one website, we believe that due to its popularity and respect in the analytics community that it does demonstrate the efficacy of the proposed Bayesian model.

To address the efficacy of the entire system, we simulated an experiment that entered our lineups in 100 daily, 10-player 50/50 games (each costing \$5) over the majority of the 2015–2016 season. We found that both the PM and KNN approaches generated several thousand dollars of profit over the course of the season; however, it was quite likely that a hypothetical daily fantasy player would have gone “bust” using the PM method alone and would have needed to re-invest to turn a profit. On the other hand, the KNN approach generated significantly more profit ($t_{998} = 146.67$, $p < 0.0001$) while always staying in the black, netting an average simulated amount of \$9,000 on an initial investment of \$500. Note that this approach suggests the optimal strategy is to enter a number of lineups rather than selecting a single lineup and entering it many times.

There are limitations to this research. First, a number of decisions had to be made that were not data driven, such as determining the volume of data needed to establish the models, a method to quantify opponent strength, cutoff values used at the lineup level, and so forth. Because this analysis problem has many nuances and little-to-no existing peer-reviewed literature to offer guidance, we tried to combine our knowledge of basketball with modern statistical methods; nonetheless, we acknowledge that results might differ under different decisions and hope that other researchers will publish alternative approaches in the future. Further, it is difficult to truly evaluate our system because it is impossible to gauge how informed daily fantasy players are on a day-to-day basis. Lastly, in any automated system it can be difficult to account for short-term changes like injuries or any sporadic “value” plays that come up (such as when a starter

is benched due to poor play, making a great opportunity for his replacement). As daily fantasy players evolve in their ability, future research will need to find ways to incorporate these challenges. However, the proposed system can still be used by analytically-minded daily fantasy basketball players looking to add research tools to their repertoire, or by more general basketball fans or NBA executives who are interested in understanding what contributes to player performance at the micro (daily) level.

References

- Altman, N. S. (1992), “An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression,” *The American Statistician*, 46, 175–185. [182]
- Arkes, J., and Martinez, J. (2011), “Finally, Evidence for a Momentum Effect in the NBA,” *Journal of Quantitative Analysis in Sports*, 7, Article 13. [179]
- Casals, M., and Martinez, J. A. (2013), “Modeling Player Performance in Basketball Through Mixed Models,” *International Journal of Performance Analysis in Sport*, 13, 64–82. [179,181]
- Entine, O., and Small, D. (2008), “The Role of Rest in the NBA Home-Court Advantage,” *Journal of Quantitative Analysis in Sports*, 4, Article 6. [179]
- Fearnhead, P., and Taylor, B. M. (2011), “On Estimating the Ability of NBA Players,” *Journal of Quantitative Analysis in Sports*, 7, Article 11. [179]
- Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. (2007), “A Starting Point for Analyzing Basketball Statistics,” *Journal of Quantitative Analysis in Sports*, 3, Article 1. [179]
- Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000), “WinBUGS—a Bayesian Modeling Framework: Concepts, Structure, and Extensibility,” *Statistics and Computing*, 10, 325–337. [182]
- Page, G. L., Barney, B. J., and McGuire, A. T. (2013), “Effect of Position, Usage Rate, and per Game Minutes Played on NBA Player Production curves,” *Journal of Quantitative Analysis in Sports*, 9, 337–345. [179]
- Page, G. L., and Quintana, F. A. (2015), “Predictions Based on the Clustering of Heterogeneous Functions via Shape and Subject-Specific Covariates,” *Bayesian Analysis*, 10, 379–410. [179]
- Piette, L., Pham, and Anand, S. (2011), “Evaluating Basketball Player Performance via Statistical Network Modeling,” MIT Sloan Sports Analytics Conference. [179]
- R Core Team (2016), R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. available at <https://www.R-project.org/>. [182]
- South, C. (2016), “A Dynamic Modeling and Optimization Approach to Daily Fantasy Basketball,” PhD Dissertation. Southern Methodist University. [180,182,184]
- Tibshirani, R. (1995), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [181]