Beating DraftKings at Daily Fantasy Sports

A statistical approach to estimating the daily fantasy performance of individual players in the National Basketball Association

By: Christopher Barry, Nicholas Canova and Kevin Capiz

(A) Abstract of Project

Our objective is to analyze NBA players and construct our own set of fantasy basketball predictions that minimize the errors between a player's actual fantasy points and our predicted fantasy points that a player scores in daily fantasy sports. We will also compare our estimates against historical salary data determined by the daily fantasy sports website DraftKings; if significant differences are found between our predictions and DraftKings' salaries (the salaries set by DraftKings generally imply what DraftKings' estimated number of fantasy points for a player will be), then over or under-valued players can be identified, and we can create a significant advantage for ourselves versus other individuals playing daily fantasy sports. Our goal is to create these predictions using various regression techniques by estimating a player's expected performance in any given game more accurately than it is estimated by DraftKings in the salaries they set.

(B) From Traditional to Daily Fantasy Sports

Entering into a traditional fantasy sports league has existed well before laptops and modern personal computers. For decades, friends have met in person before the season starts to draft players, called each other up on the phone to trade players, and kept track of their scores with pen and paper. Over the last 15-20 years, the internet has allowed players to compete in traditional fantasy sports leagues online, through websites including Yahoo! Sports and ESPN Fantasy Games. Even with games migrating online, however, sports fans criticized traditional fantasy leagues for both the slow pace of their seasons and the lack of flexibility to change the players on each team. Once an individual drafted his or her team at the start of the season, they could only make changes to their team through trades or free agency. In large part, these shortcomings paved the way for daily fantasy sports games to emerge over the last several years and become highly popular.

Daily fantasy sports (DFS) are offered online as well, primarily through two major websites: DraftKings and FanDuel. As with traditional fantasy sports leagues, in DFS individuals pick players for their team and aim to score the maximum number of fantasy points possible. These points are based on the actual in-game statistics of the players on an individual's team. Unlike traditional fantasy leagues, team selection is not subject to a typical draft. Instead, owners create their teams by "purchasing" players. Each player is assigned a salary by FanDuel or DraftKings, and a DFS participant's only restrictions in amassing his or her team is a "salary cap" that the DFS site imposes and a strict roster size requirement. Individuals are encouraged to create a new team every day or week, and can pick players strategically based on the team and player matchups in the actual sports that week. These subtle differences between traditional fantasy sports leagues and DFS lend competing in DFS to applying statistical techniques for optimizing the most valuable players to select for any given day or week, given their salary.

(C) Daily Fantasy Sports Rules

The following assessment of DFS rules is based on DraftKings Rules and Strategy page[1]. A DraftKings NBA team line-up consists of eight players, with the positions that must be filled as PG, SG, SF, PF, C, (the five main positions), G, F, (where G can be PG or SG, F can be SF or PF), and UTIL, where UTIL can be any of the five main positions. The lineup allows for a total salary cap of $50,000; therefore, the average price per player on a team that uses its entire salary cap is $6,250.

NBA players accumulate points as follows:

- Point = +1 point = PT
- Made 3pt. Shot = +0.5 points = 3PT
- Rebound = +1.25 points = RB
- Assist = +1.5 points = AST
- Steal = Block = +2 points = STL, BLK
- Turnover = -0.5 points = TO
- Double-double = +1.5 points (max 1 per player) = DD
- Triple-double = +3 points (max 1 per player) = TD

(D) Basic Estimation

We begin by asking ourselves "What is the simplest way to estimate fantasy points (F) for any given player in any given game?" From this question, we propose a basic estimate of F, $\widehat{F}$, to be a linear combination using the DraftKings scoring criterion as the weights, multiplied by the player's season averages for each stat as the variables. For the Nth game of the season, our basic estimate $\widehat{F_N}$ for a given player would be:

$$\widehat{F_N} = \frac{1}{(N-1)} \sum_{i=1}^{N-1} PT_i * 1 + \frac{1}{(N-1)} \sum_{i=1}^{N-1} 3PT_i * 0.5 + \ldots + \frac{1}{(N-1)} \sum_{i=1}^{N-1} TD_i * 3 \quad \textbf{(1)}$$

That is, for the Nth game of the season, we estimate that any player will put up stats in that game exactly equal to his season-average stats from the first N-1 games. From this approach, we compared our estimated F with the true F observed in the game, and calculated a mean absolute error of 7.414 and a root mean-squared error of 9.585. We will explain exactly what those metrics mean in our next section, but they will serve as valuable benchmarks throughout our analysis. Our goal moving forward is to make improvements on this basic estimate, and to drive down those error values.

(E) Criterion and Approach to Improving Basic Estimation

As introduced in the previous section, our objective is to minimize the errors that come with predicting fantasy points for NBA players. To measure the size of these errors, we will look at two error measurements: the root mean-squared error (RMSE) and mean absolute error (MAE). Both measurements give a sense of the inaccuracy of the errors in predictions, in the same units as the prediction, with slight

[1] https://www.draftkings.com/help/nba

differences. As the names imply, the RMSE uses a squared-error loss criterion, while the MAE uses an absolute-error loss criterion. Whereas the RMSE penalizes outliers and bad predictions at much higher rate (due to the squared-error criterion), the MAE is rather exactly the average size of each error for each prediction:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{F}_i - F_i)^2} \qquad MAE = \frac{1}{N}\sum_{i=1}^{N}|\widehat{F}_i - F_i|$$

For all but the basic estimate (where coefficients were given, by using the DraftKings scoring criterion), linear models will be fit using the ridge regression technique, using 10-fold cross-validation and the lambda that minimizes prediction error. While looking at both RMSE and MAE, we note that the ridge regression models are fitted to minimize a squared-error loss criterion plus a penalty function. As such, the RMSE may serve as the better error measurement because it more accurately matches the structure of the errors being minimized. Specifically, the ridge regression technique is to:

choose β to minimize $\sum_{i=1}^{N}(F_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=1}^{p}(\beta_j^2)$.

(F) Advanced Estimation: Factors to Take into Consideration

Predicting basic box score statistics for an individual game using only a player's box score statistics averaged over all previous games in the season makes intuitive sense as a basic estimate, but is limited. Additional variables can be included to improve the model and lower prediction errors. After brainstorming, we have come up with five additional factors that can be accounted for when estimating fantasy points for a player:

I. The opposing team's defensive statistics
II. The opposing team's opposing players' (by position) defensive statistics
III. The number of rest days since the team's previous game
IV. Whether the player has been playing well or poorly recently
V. Whether the player's team has home court advantage

Different factors are more difficult to quantify and include in an analysis than others; however, we consider these five factors above all possible for quantification and inclusion in our analysis.

(G) First Improved Model: Same Variables with Best-fit Coefficients

In order to improve upon our original model, we first sought to test whether or not simply using the point values DraftKings assigns to each statistical category as coefficients is the best way to minimize prediction error. Using the same nine variables that DraftKings uses for scoring fantasy points (as mentioned in section C), we fit a player's season averages via a ridge regression (as described above),

solving for improved coefficients (those coefficients $\hat{\beta}$ that minimize the ridge model) rather than setting the DraftKings scoring criterion as the coefficients. That is, we construct our estimate as:

$$\widehat{F_N} = Intercept + \frac{1}{(N-1)}\sum_{i=1}^{N-1} PT_i * \beta_1 + \frac{1}{(N-1)}\sum_{i=1}^{N-1} 3PT_i * \beta_2 + \ldots + \frac{1}{(N-1)}\sum_{i=1}^{N-1} TD_i * \beta_p \quad \textbf{(2)}$$

where p = 9, and β = a vector with coefficients for the 9 DraftKings scoring variables. From this approach, we compared our estimated F with the true F observed in the game, and calculated a mean absolute error of 7.045 and a root mean-squared error of 9.034. This reduction in error values indicated that fitting a ridge regression was a more efficient way to find coefficients for our $\widehat{F_N}$ estimate than our basic estimate using the DraftKings scoring criterion as coefficients.

(H) Second Improvement: Weighting Games, Rest and HCA

In order to further improve our predictions, we investigated the impact of elements III (rest), IV (weighting recent games higher) and V (home court advantage) of section F. For rest, we model an additional parameter RT, the amount of days since a player's most recent game. For home court advantage, we model an indicator parameter HCA set equal to 1 if a player is playing at home, and 0 if playing on the road. Whereas these were relatively intuitive and simple ways to quantify home-court and rest, creating a variable for whether a player was "hot" or "cold" proved to more complex.

Game Num	Weight	Weight
1	1/55	1.8%
2	2/55	3.6%
3	3/55	5.5%
4	4/55	7.3%
5	5/55	9.1%
6	6/55	10.9%
7	7/55	12.7%
8	8/55	14.5%
9	9/55	16.4%
10	10/55	18.2%
		100.0%

We argue that, when predicting fantasy points for a player's 11th game of the season, statistics from the 10th game should be weighted higher than statistics from the 1st game. Our primary rationale behind evaluating recent games more highly was that players on "hot streaks" are more valuable than players in slumps. Therefore, we choose to weight each of the nine DraftKings scoring stats based on game number. More specifically, we divide the game number by the sum of all game numbers that have been played to determine the weighting for that game, and then multiply each of the nine statistics by the appropriate weight. See the table on the right for a small example of how our weightings work.

With these updated variables, we construct our new estimate as:

$$\widehat{F_N} = Intercept + \frac{1}{(N-1)}\sum_{i=1}^{N-1} w_i * PT_i * \beta_1 + \ldots + \frac{1}{(N-1)}\sum_{i=1}^{N-1} w_i * TD_i * \beta_{p-2} + R_N\beta_{p-1} + I_{HCA_N} * \beta_p \quad \textbf{(3)}$$

Where p = 11, w_i is the weight for the i[th] game, I_{HCA_N} is the home court indicator and R_N is the number of days of rest since the player's last game. From this approach, we compared our estimated F with the true F observed in the game and calculated a noticeably improved mean absolute error of 6.614 and a root mean-squared error of 8.559. This led us to believe that there is clear value in taking rest, home-court, and recent play into account.

(I) Third Improvement: Including Opponent Defenses

We felt that we would be remiss not to include the variable that affects athletes most obviously on the court, the defense he is facing, in our model. Thus, in our third improvement, we focus on elements I (opponent's overall defense) and II (opponent's defense vs. a specific position) of section F. To model an opponent's overall defense, we calculate the number of total fantasy points per game that a team has given up to all players on the opposing team, on average, through each game. To model an opponent's defense vs. a specific position, we similarly calculate the number of fantasy points per game that a team gives up to players of each specific position, on average, through each game. We let the variables FAT_i (fantasy allowed total) and FAP_i (fantasy allowed position) be the number of fantasy points a player's opponent allowed in its i[th] game, and the number of fantasy points a player's opponent allowed to other players of the similar position in its i[th] game, respectively. With these updates, we construct our newest estimate as:

$$\widehat{F_N} = Intercept + \frac{1}{(N-1)} \sum_{i=1}^{N-1} w_i * PT_i * \beta_1 + ... + \frac{1}{(N-1)} \sum_{i=1}^{N-1} FAT_i * \beta_{p-1} + \frac{1}{(N-1)} \sum_{i=1}^{N-1} FAP_i * \beta_p \qquad \textbf{(4)}$$

Note FAT_i and FAP_i are not weighted by the recency of game, due to the complexity in doing so. From this approach, we compared our estimated F with the true F observed in the game, and calculated a mean absolute error of 6.603 and a root mean-squared error of 8.541. This error is nearly indistinguishable from the error of our previous model, which took us by surprise.

(J) Comparisons with DraftKings Salaries

Although our third improvement, model (4), showed very little improvement over our second improvement, model (3), we still choose to use model (4)'s predictions when comparing with DraftKings historical salary data.

To determine the value for a specific player, we look at the ratio of expected fantasy points to DraftKings salary; optimizing over this ratio, holding constant the salary cap, should help to optimize total number of fantasy points a roster of eight players is expected to score, given a salary cap. Attached to the right is a table of the top

Player	Average
Jordan Hamilton	6.322
Briante Weber	6.135
Michael Kidd-Gilchrist	5.841
Zaza Pachulia	5.404
T.J. McConnell	5.386
Jordan Farmar	5.377
Michael Beasley	5.353
Jerryd Bayless	5.345
Rajon Rondo	5.275
Andre Drummond	5.267
Manu Ginobili	5.249
Mason Plumlee	5.224
Pau Gasol	5.216
Draymond Green	5.216
Marcus Smart	5.216
Kyle Lowry	5.205
Dwight Howard	5.196
Will Barton	5.176
Brook Lopez	5.166
Eric Bledsoe	5.153
Greg Monroe	5.147
Matthew Dellavedova	5.137
C.J. Miles	5.137
Russell Westbrook	5.123
Paul Millsap	5.120

25 players, ordered by their ratios of expected fantasy points to DraftKings salary, considering all players who played in the 2015 - 2016 NBA regular season. The ratios are averages, and thus consider over the entire season the average number of fantasy points a player was expected to score (given model (4)) per thousand fantasy dollars they costed (again, on average). The list includes both "super-star" players, including Russell Westbrook (24^{th}), Andre Drummond (10^{th}), Draymond Green (14^{th}) and Kyle Lowry (16^{th}), as well as players with significantly smaller roles on their teams, including Jordan Hamilton (1^{st}), Jerryd Bayless (8^{th}) and Matthew Dellavedova (22^{nd}). Generally speaking, if an individual selected these players frequently for their DFS teams throughout the NBA season, they likely performed well.

(K) Additional analyses for building on this project

There were certain factors that were not modeled in our regressions that could have been factored into our analysis, and certain factors that were included in the analysis that could likely be improved upon. In order of their potential improvement to the prediction errors, our opinion of the ordering of these factors is as follows:

1. Coach resting his players - Currently our analysis does not take into consideration a coach intentionally playing his players fewer minutes in a given game. For example, late in the season, coaches (such as Gregg Popovich) may have their best players play significantly fewer minutes than they normally do. This is usually known in advance, since coaches declare their starting line-ups before the game starts and often discuss with the media if a player will be rested. As a result, this could be factored into our analysis in the future if we build predictions in real time.
2. Injuries to players - We should factor into our analysis players coming back from an injury, since a player typically scores fewer fantasy points than average in his first game back after being hurt. Also, injuries to players diminish the usefulness of our rest variable, which currently does not tell the difference between a player returning from injury as opposed to returning from a period of rest. Perhaps we could adapt the model so that players returning from injuries don't register as having rested.
3. Opponents' injuries - If the opposing team's best defensive players are injured, we should expect that a player's fantasy points would be higher than otherwise expected. To the extent that injuries to players can be taken into consideration, this could be extended to a player's opponents as well.
4. Improvement to FAP variable - Our fantasy points allowed by position variable simply aggregates the total fantasy points a team allows to each position. This does not distinguish between fantasy points allowed to starters versus bench players at a specific position, and may be skewed if a team has several players officially listed at one position. For example, if a team has 4 PGs, when really 1-2 of these players typically play in a SG role. Additionally, it could be helpful to make position classifications more granular, distinguishing between shoot-first and pass-first point guards or offensive-minded and defensive-minded centers.
5. Improvement of the recency weightings - It may be the case that a more optimal set of weights can be used when weighing recent games against games that were played earlier in the season.

In addition to improving predictions, an extension towards creating actual line-ups on DraftKings website could be made. It is important to note that simply selecting the eight players each week with the highest

expected fantasy points per salary ratio is not a sufficient way of selecting players for a team, due to the salary cap. That is, if the top eight players' combined salaries exceeds the salary cap, then that line-up would not be allowed. A combinatorial analysis could be performed to construct optimal line-ups, and web automation scripts could be written to automate the construction of a large number of these line-ups on DraftKings' website.

(L) Data Collection and Data Manipulation

Using the package rvest, we wrote a script to scrape box score data from each game of the 2016 NBA season, from http://www.basketball-reference.com/leagues/NBA_2016_games.html. We wrote a separate script to scrape historical DraftKings salary data for each player for each game, from http://rotoguru1.com/cgi-bin/hyday.pl?game=dk.

Significant data manipulation was involved in running this analysis, including:

- The most significant update to the data set was converting individual box score statistics into aggregated season-average statistics for each player. As an example, attached below are two tables displaying (top) Russell Westbrook's box score for his first 6 games and (bottom) Russell Westbrook's season-average box score through each number of games.

DATE	PLAYER	POS	OPP	FP	MP_AVG	FG	FGA	X3P	X3PA	FT	FTA	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
2	Russell Westbrook	PG	SAS	52.50	30.90	12	23	3	6	6	7	1	1	2	10	2	0	5	3	33
4	Russell Westbrook	PG	ORL	74.75	39.73	17	36	1	5	13	16	5	6	11	8	1	1	6	3	48
6	Russell Westbrook	PG	DEN	40.75	18.37	7	13	1	4	0	0	2	7	9	8	3	0	4	0	15
7	Russell Westbrook	PG	HOU	52.25	25.87	10	16	3	6	2	2	2	6	8	11	2	0	7	5	25
9	Russell Westbrook	PG	TOR	51.25	29.25	8	21	3	9	3	4	0	5	5	16	2	0	8	4	22
10	Russell Westbrook	PG	CHI	45.00	32.48	7	18	0	2	6	6	2	6	8	10	1	0	2	2	20

DATE	PLAYER	POS	OPP	FP	MP_AVG	FG	FGA	X3P	X3PA	FT	FTA	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
2	Russell Westbrook	PG	SAS	52.50	30.9	12.0	23.0	3.0	6.0	6.0	7.0	1.0	1.0	2.0	10.0	2.0	0.0	5.0	3.0	33.0
4	Russell Westbrook	PG	ORL	74.75	35.3	14.5	29.5	2.0	5.5	9.5	11.5	3.0	3.5	6.5	9.0	1.5	0.5	5.5	3.0	40.5
6	Russell Westbrook	PG	DEN	40.75	29.7	12.0	24.0	1.7	5.0	6.3	7.7	2.7	4.7	7.3	8.7	2.0	0.3	5.0	2.0	32.0
7	Russell Westbrook	PG	HOU	52.25	28.7	11.5	22.0	2.0	5.2	5.2	6.2	2.5	5.0	7.5	9.2	2.0	0.2	5.5	2.8	30.2
9	Russell Westbrook	PG	TOR	51.25	28.8	10.8	21.8	2.2	6.0	4.8	5.8	2.0	5.0	7.0	10.6	2.0	0.2	6.0	3.0	28.6
10	Russell Westbrook	PG	CHI	45.00	29.4	10.2	21.2	1.8	5.3	5.0	5.8	2.0	5.2	7.2	10.5	1.8	0.2	5.3	2.8	27.2

- Header columns that were scraped needed to be removed, as well as all player-games where a player did not play.
- Columns needed to be constructed or their format manipulated for position, minutes played, date (into game number and rest), opposing team's average fantasy points allowed, and opposing team's average fantasy points allowed by position.

(M) Conclusion

Our efforts to improve upon our basic estimate performed fairly well. The MAE fell from 7.414 to 6.603, and the RMSE fell from 9.585 to 8.541. Significant effort was put into the third improvement, which only

improved MAE and RMSE by roughly 0.01, which was disappointing but highlighted the difficulty in reducing prediction errors past a certain sized error. The challenge of solving DraftKings, and beating the other individuals that play DFS, remains a very tempting one that could be pursued well past the timeline of this project.