Advanced NBA Analytics for Fantasy Basketball Players

Minsub Kim, Siavash Tahamtan, Claudio Fofiu, Kangin Choi, Anh Hoang, Jaeyong Kim

1. Introduction

The prominent rise of data analytics and data science has created a push to integrate this field into any area possible. Data science is now utilized to predict performance of certain stocks or used to classify whether a patient may have a disease, and the NBA is no exception to this trend. Teams are consistently analyzing and searching for the best players that are available to build a championship team. Of course, having the best players is just one area of building a winning team or a winning franchise, but it does heavily influence other aspects like team composition, morale, salary cap and many other aspects. While not everyone can be a front office GM or owner for these NBA organizations, fantasy basketball gives your average fan a chance to build a winning team and compete in leagues with friends or strangers. In fantasy basketball, users will sign up for a league and they have the chance to build a team that will outperform their opponents in the league as well. Often times, a lot of research goes into drafting players to determine what kind of performance they will have in the upcoming season and this is where data science and analytics has been making an impact. Both NBA GMs and fantasy owners both want to predict future performance based on the factors they can see in the athletes. This paper seeks to provide a novel approach to predicting NBA player performance and visualizing the data in a manner where fantasy players can build the best team possible.

2. Problem Definition

With the constantly changing landscape of the NBA, NBA teams and fantasy sports users have noticed the importance of statistics and data in NBA. This is most notably seen stats like this: there is a disappearance of traditional centers and an “increase of 47% [in three-point attempts] over the five years” (Chung 2019). The NBA has changed so much, and analytics has been used to study this change and predict how players will perform with the changing landscape. In 2006, research has shown “fantasy sport fans to be spending approximately $1.5 billions each year”; at the same time, a research in 2010 has also reported that these “players spend approximately 22 hours 40 minutes” a week consuming this form of entertainment (Billings 2013). There is a strong desire for fantasy players to research and predict so they can win their own fantasy leagues; however, most fantasy owners are not statistically literate and often will just draft players based on some subjective opinion formed from the previous performance of the player.

Because NBA analytics are currently used by NBA teams to have the highest advantage against other teams, many advanced analytics created and utilized by NBA teams are hidden away from the public. Moreover, the recent change in players’ and teams’ playstyles influenced by data analytics and rise in load management where players sit out games to recover and decrease the risk of injury or chronic fatigue has shown the importance of not always picking statistically the best player for their fantasy sports teams. In the past, many studies have been performed to tackle the problem of creating a model to predict who would likely be the winning team for upcoming match or to project a player’s potential standing in current season based on newly introduced NBA stats. While it is impossible to get rid of the subjective aspect of a fantasy draft, the desire is to provide a model to predict player performance that is available to fantasy users, so they have more of an analytical background for choosing players.

3. Survey

Since players can be traded or signed by other franchises in the next season, current data analytics for NBA are more focused on the team's current season, such as "collecting sophisticated data points about its players through wearables, sleep monitors, and even saliva samples to assess their fatigue level and predict their performances" (Petra 2020) to predict and prevent injuries for key players who will need to play for more important games.

Being able to predict player performance on the new season will not only benefit fantasy basketball players (Evans 2018), but also front office workers for the NBA teams. With all the different variables in evaluating players, an accurate model of performance gives managers an idea of player’s value (Hwang 2012) and who will be a star and who will be a flop (Moxley 2015). For both Fantasy

players and real GMs, there is another level of strategy that is introduced, where both have insight into how their own players as well as opposing players will perform. One aspect of determining an athlete's performance is the age factor for the players, and age is seen as a common factor of decreasing not only athletic, but cognitive ability in people over time (Vaci 2019). By utilizing basketball statistics like win share, value over replacements, and PER, these writers try to model the decline of athletic ability and value in NBA athletes over a range of age. A similar predicament is evaluating and predicting the performance of rookies in the NBA. An article from 2011 tries to model NBA draft prospects by deriving and modeling formula that incorporates the basketball stats (points, rebounds, etc.) as well as number of years in college and the different divisions that exist (Berri 2010). Finally, for those seasoned veterans or already common place NBA players, models are created to predict their performance in the upcoming season. These models are used to try and identify what kinds variables (team, conference, salary) may affect performance of the athlete in upcoming seasons (Casals 2013). There are so many variables that affect player performance in the NBA and because of the physical nature of the sport, it is not uncommon for players to have fickle performances season to season. In addition, it is difficult to model NBA rookies as they make a jump from a collegiate level to a professional level. Our model will be measured by season-to-season performance of players.

4. Method

4.1. Data

The data was obtained from a third-party API available in Python called nba-api developed by Swar Patel (Patel 2018). The API includes many different endpoints, which takes in specific parameters in order to allow for the access of many datasets that are compiled of relevant NBA statistics. The data inserted into the model included season by season data from 2010 to 2020 for all active players in the NBA today. While we have gone through tens of thousands of records of NBA data from the 2003 to 2020 seasons, the filtered data we've settled on consists of a little over 1,300 records and its disk size of about 500KB. To pull the data for our analysis we've had to first identify and collect the list of active players that have played during the 2010-2020 seasons, and then for each player and season call the APIs to get appropriate data. The collection of the data took over 2 hours to run and we've had to implement some "sleep" time measures between pulls as multiple continuous calls would time-out and interrupt our operation. This was accomplished through trial and error as often our continuous API calls would time-out during or close to the end of the data set resulting in a Python error. The main points of data that we utilized in building the model are included in the table below.

Pre-Features	Features	Output
PLAYER_AGE	GP = Games Played	FGM = Field Goals Made
POSITION	W = Wins	FGA = Field Goals Attempted
HEIGHT	L = Losses	FG3M = Field Goal Threes Made
WEIGHT	W_PCT = Win Percentage	FTM = Free Throws Made
	PLUS_MINUS	FTA = Free Throws Attempted
	MIN = Minutes	REB = Rebounds
	DD2 = Double Doubles	AST = Assists
	TD3 = Triple Doubles	TOV = Turnovers
		STL = Steals
		BLK = Blocks
		PTS = Points

Table 1: Data Points utilized to predict the outputs

Overall, our obtained data's features can be divided into three main categories: pre-features (used to predict the next season features), features (used to predict the outputs) and the outputs. The fields in the output columns are the data points used to evaluate the players fantasy performance in a league. PLUS_MINUS (or +/-) is a measure of the difference in points when the player is in and out of a game. It is a measure of the players impact when they are in or out of a game. Double Doubles are a count of games where a player achieved double digit figures in 2 of the following categories: points, rebounds, assists, steals or blocks. Triple doubles are the same principle but where a player achieves double digits in 3 categories.

The data also required some cleaning to process the data through a model. First thing was to filter out seasons where the player played less than 500 minutes. An NBA regular season spans 82 games, and each game is at least 48 minutes. This means the minimum number of minutes available in a season is 3,936 minutes, excluding overtime and playoffs. If a player plays 500 minutes or less, whether due to injury or due to other reasons, that season is not an effective measure of that players performance. In addition, the height and position field are input as strings so those needed to be converted to numbers. Each position given in the data pull was assigned a 1 to 5 depending on the position. 1 is assigned to the point guard, 2 for shooting guard, 3 for small forward, 4 for power forward and 5 for center. These are recognized universally in terms of position in basketball.

4.2. Models

Because basketball is a positional sport, it made the most sense to model the player performance by splitting it up by position. Traditionally, size and weight dictated what position players would play and what their stat line might look like. For example, Ben Simmons is 6-10, but classified as a guard by the NBA (). Maybe before, he would have been classified as a forward who mostly scores 2 pointers in the paint and grabs rebounds. The same goes for athletic point guards like Russel Westbrook who is shorter but still able to get many rebounds over people taller. Rather than fit the data according to height and weight ranges, we decided that the ability of a player would be based on age and position as classified in the NBA, in addition to height and weight. Our models iteratively fit two models: one to predict the features that are expected in the 2020-2021 season, and another to predict the player performance outputs for the 2020-2021 season. Both were evaluated by position where a new regression model was fitted per each of the 5 positions classified in the NBA.

4.2.a. Linear Regression

One method we have implemented for this project is linear regression. It is a supervised learning method frequently used in machine learning. Linear regression utilizes a series of weight or coefficient (w_n) for each feature x in (x_n) and bias w_0 to compute the predicted value. At the same time, the goal of the model is also to minimize the square difference between the predicted value and actual value (residual sum) during training and actual testing.

$$min_n \left\| y_{predict} - y \right\|^2$$

For our linear regression experiment, the list of features utilized and expected in terms of output for training and predictions are shown in table 1 above. Our linear regression model utilizes player's data for these features across seasons: 2010-2020 to predict the relevant statistics for the 2020-21 season. The output of this model will be fed in as the features for the following iteration of modeling to get the final outputs.

Due to the way we are absorbing the NBA data, we need to predict the inputs of the final model before-hand. We picked 2 features (Age, Position) as the initial starting point in predicting outputs like Games Played, Wins, Losses, Minutes Played, Double Doubles and Triple Doubles. Age and Position are the starting dependent variables they can begin to describe the performance a player may have.

4.2.b. Ridge Regression

Another supervised model implemented is ridge regression. Due to the limited number of datapoints, the introduction of new players and discontinuation of many players' NBA careers, combined with the large number of features used as input, ridge regression is a logical choice as one of its main

utilities is to avoid overfitting. To do this, ridge regression implements L2 regularization on the minimum residual sum squared calculation.

$$min_n \|y_{predict} - y\|^2 + \lambda \sum_{i=1}^{n} w_i^2$$

For this experiment, the features used as input for prediction are shown in table 1. This model is also fitted iteratively depending on the available number of seasons the player has in the data pulled. If the player only has 1 season of data, the model fits data existing for all players in the same position and then generates the outputs. For the players that have more than 1 season of data, each player has his own Ridge model fit based on data existing for that player in previous seasons. Using the output of features expected for the following season, the player performance is predicted.

5. Visualization Dashboard

The visualization dashboard is created in Python using the dash package. The common issue with many fantasy dashboards is that they are either to simple (a single table) or too complex (a dashboard with every stat line possible). We decided to create a dashboard that would interactively intuitive and easy to follow for any user. Because NBA Fantasy Basketball is measured on 9 total categories (shown in table 1 above), we abstracted that data and ordered the best predicted players for that category. When fantasy players go head to head, they are competing to beat each other in each of the 9 categories, so it is in their best interest to draft a team that can provide scoring in all those categories.

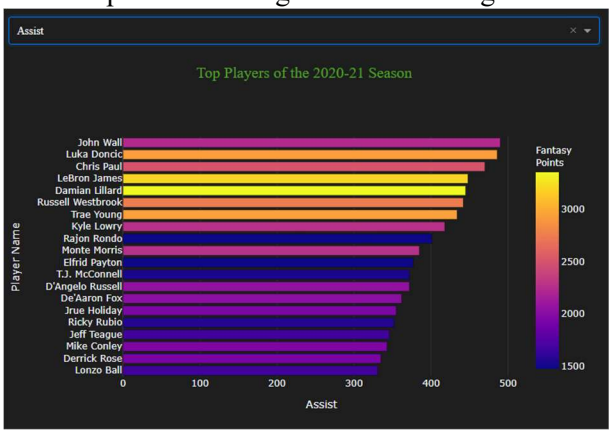

Figure 1: Dashboard with dropdown menu to show each category

Another part of the dashboard is to see what players are projected to be the best in that season. This helps the users keep track of what players are the top for that season and they can filter by team as well to see how the players on the team will perform. Players can be searched by name or through the ranks to help users build an effective draft strategy.

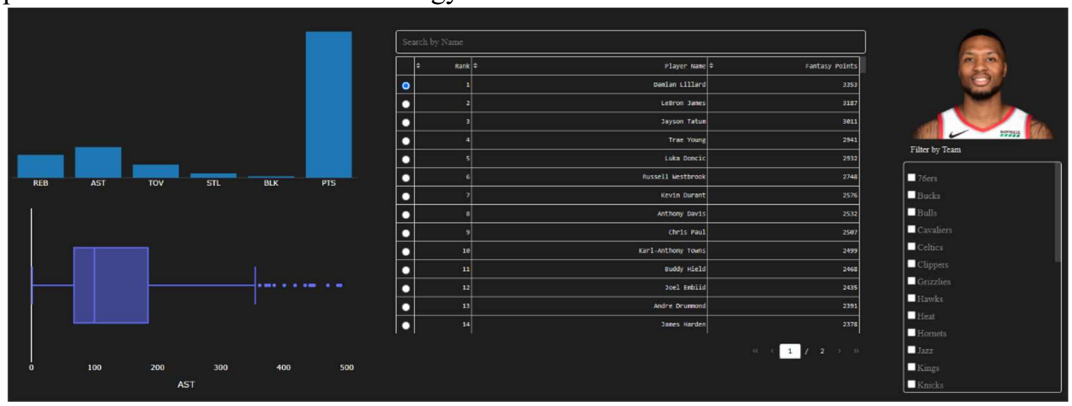

Figure 2: Dashboard with the best fantasy players overall

6. Experiments and Evaluation

The experiments we evaluated were designed to give us the best way of modeling the data to predict the player performance. One issue that we encountered was the data available for NBA players varied per player. We looked at players who are considered active by the NBA today for the upcoming season. Players who are new to the league, like the 2019 Rookie Class, had small amounts of data to work with. The issue here became how to create a specific model to simulate the performance of the player. In addition to this, another topic of debate was deciding whether to create a model based on all players. From our research, it was found that it would be difficult to create a one size fits all model because the NBA has changed over time. As players become more athletic and versatile, they stretch the limits on what players in that position are able to accomplish and so our model also had to be versatile in that manner. Performance would have to be done on a position and player basis to truly understand and predict the future performance of that player.

Our initial efforts began with only one model instead of two models in series. We started with a Linear Regression Model in Python's SciKit Learn package and fed in 12 features (GP, W, L, W_PCT, MIN, PLUS_MINUS, DD2, TD3, HEIGHT, WEIGHT, SEASON_EXP, and POSITION) to achieve the outputs we needed. We quickly discovered that this would not be useful as we did not have access to the features that need to be input in order to predict the 2020-2021 season output. Thus, we had to split the modeling into two models in series with one feeding into another. The initial features were split into two groups: pre-features and features. The pre-features were decided to be the most constant features that we could find in the data. This ended up being the characteristics of height, weight, position and most importantly, age. Those four were used to predict the GP, W, L, W_PCT, MIN, PLUS_MINUS, DD2, TD3 which are also the broader categories of NBA statistics but not the statistics that fantasy basketball looks at. Those features were then the inputs to a final model to determine the outputs such as Points, Field Goals, Steals and categories like such.

A few issues that found in the data were the labels given to the players in terms of position. When data is pulled from the API, the valid positions are Guard, Guard-Forward, Forward-Guard, Forward, Center-Forward, Forward-Center and Center. However, positions are more commonly known as Point Guard (1), Shooting Guard (2), Small Forward (3), Power Forward (4) and Center (5). We assigned values to the position using this table below.

Position	Value
Center	5
Center-Forward	4
Forward	3
Forward-Guard	2
Guard	1
Guard-Forward	2
Forward-Center	4

Table 2: Position vs Value

The positions seem to be assigned in a non-standard format which is why we had to standardize them, but in doing so, our distribution of NBA players by position seemed skewed. We found there were Position 1 had 182 players, Position 2 had 94 players, Position 3 had 142 players, 4 had 67 Players and 5 had 34 players. The error we found in using this was incredibly high and shown in table 3 below. When we used only 3 classifications for Position (1 – Guard, Guard-Forward, 2 - Forward, Forward-Guard, Forward-Center, 3-Center, Center-Forward) we found the error to be large, but still much less than the other method.

POSITION	GP	W	L	W_PCT	PLUS_MINUS	MIN	DD2	TD3
1	13.04%	27.12%	-0.55%	11.25%	81.36%	26.69%	-	-
2	211.94%	539.04%	89.27%	102.75%	-352.26%	302.10%	-	-
3	14.78%	36.56%	-5.55%	18.37%	583.32%	0.40%	51.00%	-
4	23.71%	85.88%	-23.96%	47.01%	-252.06%	-8.37%	-12.84%	-
5	42.19%	-11.68%	251.69%	-38.30%	-91.08%	47.02%	491.36%	-

Table 3: Percent Error in Predicting Pre-Features with 5 positions

POSITION	GP	W	L	W_PCT	PLUS_MINUS	MIN	DD2	TD3
1	13.49%	27.53%	-0.07%	11.07%	70.53%	23.53%	-	-
2	13.96%	36.32%	-6.91%	18.97%	591.93%	-1.41%	44.76%	-
3	21.71%	56.89%	-5.25%	27.22%	-168.92%	-4.84%	11.47%	-

Table 4: Percent Error in Predicting Pre-Features with 3 positions

Having more data with even the non-active NBA players would likely have improved our model for predicting the pre-features using Linear Regression. We only worked with 10 seasons of data but utilizing more data would have improved our accuracy. Unfortunately, we could not pull more data because of connection time out issues via the third-party API. In addition, we utilized the Ridge Model in SKLearn because we could utilize settings like normalizing and fitting the intercept for each model used. Using the Ridge Model gave us non-negative results compared to the regular Linear Regression package available.

7. Conclusion

In most studies pertaining to this problem space, the main focus is on predicting the outcome of a specific game, particularly: who is the winner vs loser; at the same time, more motivated studies push the goal further by trying to predict the difference in scores at the end of the game. Our project focuses on a more novel problem space by predicting future standing of a player in upcoming season by projecting his statistics. As a result, the current database available did not quite provide us with a suitable amount of data for the current endeavor. Had we had more time, we would have pursued cleaning up the positions in NBA data by comparing it to data in other NBA datasets available. Another way to improve the model would be to use Team performance to help develop the pre-features to be inserted into the final model to provide the fantasy outputs. In the current methods, team performance is not being utilized to predict any output. While in fantasy sport, team performance isn't as meaningful as individual performance, it is still a relevant variable and thus, affecting the player's final statistics and his fantasy sport score. Investigating these avenues would be interesting to see what kind of difference it could make in our environment in future developments.

All team members have contributed similar amount of effort

Coding – Claudio, Anh, Jaeyong // Visuals – Siavash, Kangin // Poster, readme.txt - Minsub

References

1. Hore, S., & Bhattacharya, T. (2018). A Machine Learning Based Approach Towards Building a SustainabilityModel for NBA Players. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. doi:10.1109/icicct.2018.8473102
2. Billings, A. C., & Ruihley, B. J. (2013). Why We Watch, Why We Play: The Relationship Between Fantasy Sport and Fanship Motivations. *Mass Communication and Society, 16*(1), 5-25. doi:10.1080/15205436.2011.635260
3. Neu Dong., "Predicting NBA Players Points Scored", University of Wisconsin – Madison.
4. Sethi, Kapil & Gupta, Ankit & Gupta, Gaurav & Jaiswal, Varun. (2019). Comparative Analysis of Machine Learning Algorithms on Different Datasets.
5. Petra. (n.d.). How data analytics is revolutionizing the NBA. Retrieved November 22, 2020, from https://digital.hbs.edu/platform-digit/submission/how-data-analytics-is-revolutionizing-the-nba/
6. Hwang, D.(2012). Forecasting NBA player performance using a Weibull-Gamma statistical timing model. *MIT Sloan Sports Analytics Conference*.
7. Moxley, J. H., & Towne, T. J. (2015). Predicting success in the National Basketball Association: Stability & potential. *Psychology of Sport and Exercise, 16*, 128-136. doi:10.1016/j.psychsport.2014.07.003
8. Evans, B. A., Roush, J., Pitts, J. D., & Hornby, A. (2018). Evidence of Skill and Strategy in Daily Fantasy Basketball. *Journal of Gambling Studies, 34*(3), 757-771. doi:10.1007/s10899-018-9766-y
9. Cheng, G., Zhang, Z., Kyebambe, M. N., & Kimbugwe, N. (2016). Predicting the outcome of NBA playoffs based on the maximum entropy principle. Entropy, 18(12), 450.
10. Landers, J. R., & Duperrouzel, B. (2018). Machine learning approaches to competing in fantasy leagues for the NFL. IEEE Transactions on Games, 11(2), 159-172.
11. Drafting Agent-based Modeling into Basketball Analytics. (2019). *Annual Simulation Symposium (ANSS 2019)*. doi:10.22360/springsim.2019.anss.014
12. Kretzer, M., Maedche, A., & Shah, F. (2015). Designing an Analytics Platform for Professional Sports Teams Designing an Analytics Platform for Professional Sports Teams. *International Conference on Information Systems*
13. Winston, Wayne. (2019). Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football. *Princeton: Princeton University Press*.
14. Vaci, N., Cocić, D., Gula, B., & Bilalić, M. (2019). Large data and Bayesian modeling—aging curves of NBA players. Behavior Research Methods, 51(4), 1544-1564. doi:10.3758/s13428-018-1183-8
15. Berri, D. J., Brook, S. L., & Fenn, A. J. (2010). From college to the pros: Predicting the NBA amateur player draft. Journal of Productivity Analysis, 35(1), 25-35. doi:10.1007/s11123-010-0187-x
16. Casals, M., & Martinez, A. J. (2013). Modelling player performance in basketball through mixed models. International Journal of Performance Analysis in Sport, 13(1), 64-82. doi:10.1080/24748668.2013.11868632
17. Landers, J. R., & Duperrouzel, B. (2018). Machine learning approaches to competing in fantasy leagues for the NFL. IEEE Transactions on Games, 11(2), 159-172.
18. Swar Patel, (2018). API Client package to access the APIs for NBA.com, GitHub repository, https://github.com/swar/nba_api
19. Chung, J. (2019). Explaining the Trends of NBA Strategy through the Lens of Human Risk Tolerance. International Journal of Scientific & Engineering Research, 10(1), https://www.ijser.org/researchpaper/Explaining-the-Trends-of-NBA-Strategy-through-the-Lens-of-Human-Risk-Tolerance.pdf
20. Ben Simmons Stats. (n.d.). Retrieved November 22, 2020, from https://www.basketball-reference.com/players/s/simmobe01.html