Forecasting NBA Careers of NCAA Athletes

Chunpong Lai, Yixin Chen, Ethan Qiu

Data 100 Final Project

Professor Adhikari, Professor Gonzalez

May 13, 2020

Abstract

The goal of our project was to predict how successful NCAA athletes would be in a more professional and competitive league and analyze the performance variation of NCAA athletes between two leagues. We were able to predict with moderate accuracy how well the players will be after being drafted into the NBA in terms of efficiency and start percentile with supervised learning.

Introduction

The National Basketball Association (NBA) is the largest men's professional basketball league in the world. The NBA draws the bulk of its new blood from the league National Collegiate Athletic Association (NCAA) through the NBA draft. Every year, over a thousand college players are eligible for the draft, however, only 60 of them will be entering the NBA. Additionally, the order of being drafted determines their started salary and contract type. According to Spotrac.com, the first pick started with a salary of $6.74 million while the last pick of the first round made $1.34 million for the 2018 NBA draft. Furthermore, being selected in the first round begin their career with a four-year guaranteed contract while second-round picks do not have guaranteed contracts. The draft is paramount for both the league, and the players.

Every team wants to get the best potential player and the NCAA games statistics is one of the most important metrics they rely on to make the draft decision. We have proposed some questions listed below to help us to understand the player's performance variation between two leagues and the relationship between NCAA performance and performance in the NBA.

Questions:

- What is the performance trend of the athletes in the NBA compared to NCAA?
    - Which performance metric(s) increases/decreases between two leagues?
    - Can we forecast the performance of NBA athletes based upon their NCAA statistics?
    - Which position(s) has the most stable performance after joining the NBA?
    - What level of skills development and trajectory should we expect after the NCAA player transferred to the NBA?

Dataset

There are several NBA related datasets supplied with the project but in order for a dataset to further our interests and goals, we select the datasets based on the following criteria: 1.) That the dataset contains game statistics on players' performance. Since we would like to build a model that predicts the future performance of players. 2.) That the dataset contains comprehensive information that can fully describe the performance of players. The metrics to evaluate a player's performance should consider both the positive side like assist statistics and the negative side like turnover statistics. 3.) That the dataset contains game statistics on players for multiple seasons. Since we would like to learn about the skill development and career path of players over years in the NBA, therefore single year data points are not sufficient.

By following criteria above, we select two datasets: 1.) named "College" which contains NCAA games statistics and basic information of players like the height and attended college. 2.) named "Basketball-PlayerBoxScore" which contains the detailed NBA games statistics on players and the game result from 2012 to 2018.

Method & Approach

- **Dropping irrelevant data**: Our goal is to learn about the player's performance variation between two leagues, so we dropped all useless and irrelevant data. For example, we dropped all the players that had 0 value in the NCAA_games column since this type of player has no NCAA game statistic.

- **Decoding position**: After investigating the position naming convention, we have learned that certain players have major position and minor position. For example, C-F indicates playing both center and forward positions but mostly playing a center position. Therefore, we decided to only consider the major position of players and decode the position column into 3 types Guard, Forward and Center.

- **Converting data type:** The data of the players' height is in imperial units and in a string data type, which can potentially be difficult to parse in the later EDA and modeling task, therefore, we converted the height into a numerical feature in the unit of centimeter.

- **Merging dataset:** The two selected datasets, "college" and "Basketball-PlayerBoxScore", contain overall NCAA/NBA statistics and NBA games statistics, respectively. We inner joined these two datasets by the players' name to create a useful table for the later EDA and modeling task. The players that fail to appear in the NBA games statistics are irrelevant to our analysis.

- **Filling missing value:** After joining two datasets and cleaning by the methodology described above, the conspicuously missing value left in the dataset is regarding 3-pointer games statistics. Furthermore, most of the players with such missing values are those who occupy the "Forward" or "Center" positions as Figure 1 shows below. We suspected the real data is zero or very close to zero, considering the rarity of 3-point shots, therefore, we filled all the nan values with 0.

- **Feature engineering** :
  In the dataset, there exist many features to describe the performance of players from different angles like steal, assist and block. However, there's a lack of an overall measurement of how good a player is. So we introduced a new data point "Player Efficiency" to capture the overall performance in one measure unit by applying the formula created by statistician Martin Manley.

  $$efficiency = \frac{(Points + Rebounds + Steals + Assists + Blocked\ Shots - Turnovers - Missed\ Shorts)}{Minutes}$$

  We also noticed that the players' start/bench position were being measured in the NBA game statistics. We then created a feature called start percentile to measure the proportion of games the player spent as a starter compared to the total number of games played.

  $$start\ percentile = \frac{Games\ as\ Starter}{Total\ Games}$$

  Furthermore, we selected schools that produced large amounts of NBA players, players with high efficiency and high start percentile, and labeled them as "NBA preferred school", " good performance school" and "high starter school" to account for high quality college basketball programs. Moreover, we performed one-hot encoding to the position because our EDA (figure 6) shows different positions and has different performance expectations and we would like to capture this information into our model.

- **Detecting & Removing outliers:** We used a boxplot to detect the outliers and verify it with the dataset to see if the outliers are meaningful to our model and EDA . For example, we saw there's an outline in the column of NCAA 3 points field goal percentage. The outliner's value is 1 / 100% which means these players made every single 3 points shoot they attempted. That's amazing! We want to know who they are, therefore, we looked at all the players who had an NCAA 3 points field goal percentage equal to 1. The result seems problematic, the data shows all of them have made zero 3 points field goal and zero 3 points attempts, which is not making any sense, so we suspected this is the error data and removed it from our dataset.

- **Normalize data:** We performed normalization to our dataset since we have certain columns that have very different ranges to the others, for example the sum of players' game play minutes. It can be a very large number like ten thousand while game statistics are usually in the range between 0 to 100, like how many points and assist players made.

EDA & Answering questions:

1. According to "Field Goals Percentage in average" in figure 5, players generally have better scoring performance in NCAA than NBA regardless of position. The trend of decreasing performance also reflects the NBA is a more competitive and professional basketball league.

2.

   2.1. The overall free throw percentage is quite the same while three points goals percentage increased compared to NCAA. The reason could be free throw shooting has relatively less defensive pressure and the NBA's "Small Ball" revolution forces more players to practice three points shooting.

   2.2. We can definitely forecast the performance of NBA athletes based upon their NCAA statistics as the development of the player is following a certain trajectory although the accuracy depends on the predicting target. For instance, the metric of EFF (see formula above) comprehensively weighted offensive and defensive aspects of the player like assist, block and turnover. However, the NCAA statistics lack the defensive data which will decrease the model accuracy.

   2.3. According to Figure 6, the Center position has the most stable performance but also the least growth potential. Juxtaposed against Forwards and Guards, who appear to display dramatic improvement over years. This can be explained by the NBA's "Small Ball" revolution in the recent year as many NBA teams prefer to set up an attacking strategy that relies on Forwards and Guards.

   2.4. For Centers, we can expect very similar performance over years in the NBA. For the Forwards and Guards, we can expect improvement in the first five years and reach their peak around sixth years, then their performance starts to drop as they reach the end of their careers.

3. We notice that a player's start percentile appears to be directly proportional to the efficiency rating of the player, with the residuals of the plotting to be fairly evenly and randomly distributed (Figure 3). From this, it appears that starting players are generally considered more valuable than bench players in that through the entirety of the game, with starting players being expected to have more playtime than bench players.

4. We also plotted the efficiency rating of players against start percentile whilst separating players by their positions (forward, guard, center) (Figure 4). We observe that there seems to be a relationship between efficiency ratings and a player's position, with the evidence of noticeably different regression slope and y-value ranges: it suggests that the expected efficiency of a high performing player is different by position. Curiously, it is the position that we earlier have observed to have the least growth during their NBA careers, the Centers, that have the highest efficiency, with Forwards having lower y-values ranges, and Guards, the least. This could point to a flaw in our metric efficiency in measuring player performance, with it weighing Centers above Forwards and Guards, and thus defensive plays over offensive plays. Efficiency does not weigh different actions differently, and was created before the small-ball revolution, which can explain its flaws in fully describing player performance.

Model Result

We trained two models to predict the athlete's performance in terms of efficiency and start percentile. We first assessed the performance of these models in a random split (80/20) with 446 players in the training set and 112 players in the test. Then, we performed the 5-fold cross validation of the test . The score is calculated in coefficient of determination (R2). Table 1a and 1b summarizes the results.

Model	Training Score (80%)	Test Score (20%)	5-Fold CV Test Score
Ordinary Linear Regression	0.763	0.728	0.617
Ridge Regression	0.763	0.727	0.633

Table 1a: Result on predicting start percentile (score in R2)

Model	Training Score (80%)	Test Score (20%)	5-Fold CV Test Score
Ordinary Linear Regression	0.609	0.548	0.491
Ridge Regression	0.605	0.554	0.519

Table 1a: Result on predicting EFF (score in R2)

Conclusion

1. During our EDA, we noticed that a player's performance seemed to have a noticeable relationship with their NBA efficiency. And when we vectorized the players' positions, and integrated them into our efficiency model, it made a significant impact upon our models' accuracy.

Another interesting statistic that we stumbled upon is the players' start/bench stats per game. In that their percentage of games where the player served as a start player seemed to be correlated with their NBA efficiency. We think that potentially a players' start percentile can also serve as a measurement for their ability on court. This aided us in selecting competitive colleges to account for higher quality college basketball programs, which was then incorporated into our model.

2. The 3pt goal occupies an important cultural status, being commonly associated with demonstrating a player's ability in accurate shots. However, the 3pt goal proved to be a disappointing statistic in our modeling. This might be due to the fact that the 3pt goal was slowly being incorporated into regular strategy with the three-point revolution, with the development most notably catalyzed from 2015 onwards, likely due to Stephen Curry's prowess with the 3pt influencing the league in implementing the 3pt goal into their tactics.

3. There are difficulties for a holistic measurement for player performance in a nuanced game such as basketball. Different positions have similarly differing demands for the player, and consequently, we cannot merely utilize, for example, Free Throw Percent as a holistic metric.

Thus we utilize efficiency, or EFF as the holistic metric for player performance. Though there exists valid criticisms of EFF and alternatives such as PER, it is still an effective system for quantitatively measuring a player's worth. We also look towards the percentage of starter/bench or starting percentile, and play time as a potential measurement of player performance with the argument that effective players would be utilized more as starters, and generally experience more play time.

4. As mentioned above, there was the difficulty in producing a holistic statistic that encompassed a player's value to a team. Though we utilize EFF and a players' starting percentile, they are still not perfect metrics of a player's performance.

Additionally, one of our goals is to analyze the career path and skill development of the rookie. However, the NBA player scores box only includes player's game statistics from 2012 to 2018 while the college dataset covers a wider range from 1960. The perfect scenario will be the player starting their career after 2012, we considered this type of player as a complete data point. But it's not always the case, we will lose data or have incomplete data if the scenarios listed below applies:

- The player started and ended their career before 2012 (data lost).
- The player started their career before 2012 and is still active in the range of 2012 to 2018 (incomplete data).

Our model only utilized data from 2012 and onwards due to the mechanisms of the inner join. Though we risk introducing a sampling bias into our model, considering the changing paradigm of the game overtime, adding overly dated data could produce a model that is unfit for current times.

5. We attempted to predict a player's performance through efficiency and start percentile from their NCAA statistics. This approach can lead to ethical concerns, in that we might be affecting the players' NBA careers through our modeling. Considering the immense difference between salaries of first round draft and second round draft players' salaries, and then the players that are not drafted at all, our models can have very real quality of life impacts upon the players.

We created features to attempt to encapsulate players in colleges that produced large amounts of NBA players/skilled NBA players. This feature could disadvantage those from other colleges with historically prominent basketball programs, leading to lower forecasted performance for such players.

Also, with the changing paradigm of the game, our model can quickly become out of date, leading to incorrect assessments of players' performance upon the field, potentially leading to damage upon the drafts selections, the players involved, and the game itself.

Future Research:

6. It would be interesting to have data regarding player's salaries, and which draft they were picked in (or if they were even drafted at all). In some sense, the player's salary can be a literal metric of a player's worth, and it would also be interesting to study a player's performance through their efficiency or start percentile and its relation to their salaries.

7. To tie a player's salary to their efficiency or performance has a great deal of ethical concerns. There not only exists the issue of the accuracy of the metric utilized to measure efficiency, but also the issue of whether a player's salary should be entirely tied to their efficiency on the playing field. We should take care in overseeing our models' assessments in terms of accuracy and adjust for impartiality. We should consider and utilize our models with reservations and social concerns as these models could affect human wages, and directly impact people's quality of life.

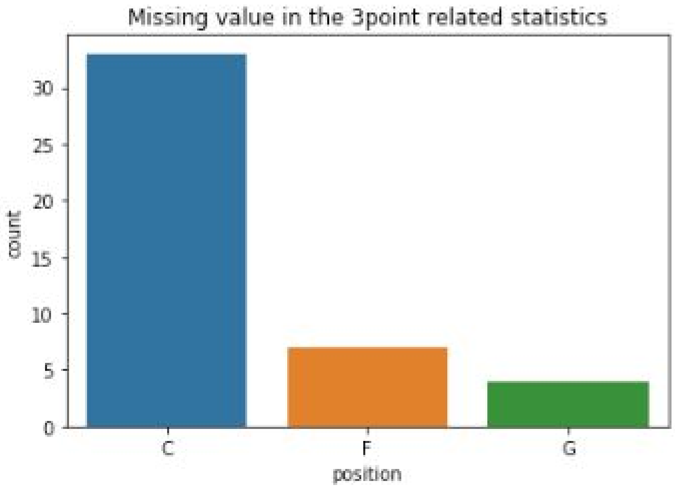
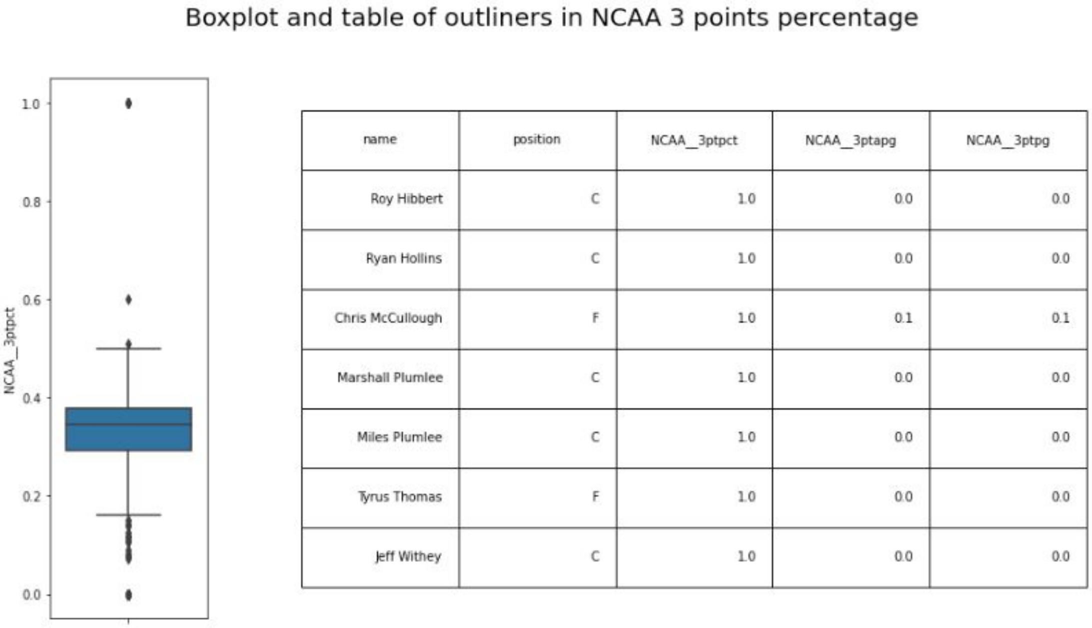


Figure 1. Shows the count of 3points related missing value by position



name	position	NCAA__3ptpct	NCAA__3ptapg	NCAA__3ptpg
Roy Hibbert	C	1.0	0.0	0.0
Ryan Hollins	C	1.0	0.0	0.0
Chris McCullough	F	1.0	0.1	0.1
Marshall Plumlee	C	1.0	0.0	0.0
Miles Plumlee	C	1.0	0.0	0.0
Tyrus Thomas	F	1.0	0.0	0.0
Jeff Withey	C	1.0	0.0	0.0

Figure 2. Shows boxplot and more detail info in table about the outline in NCAA 3 points percentage

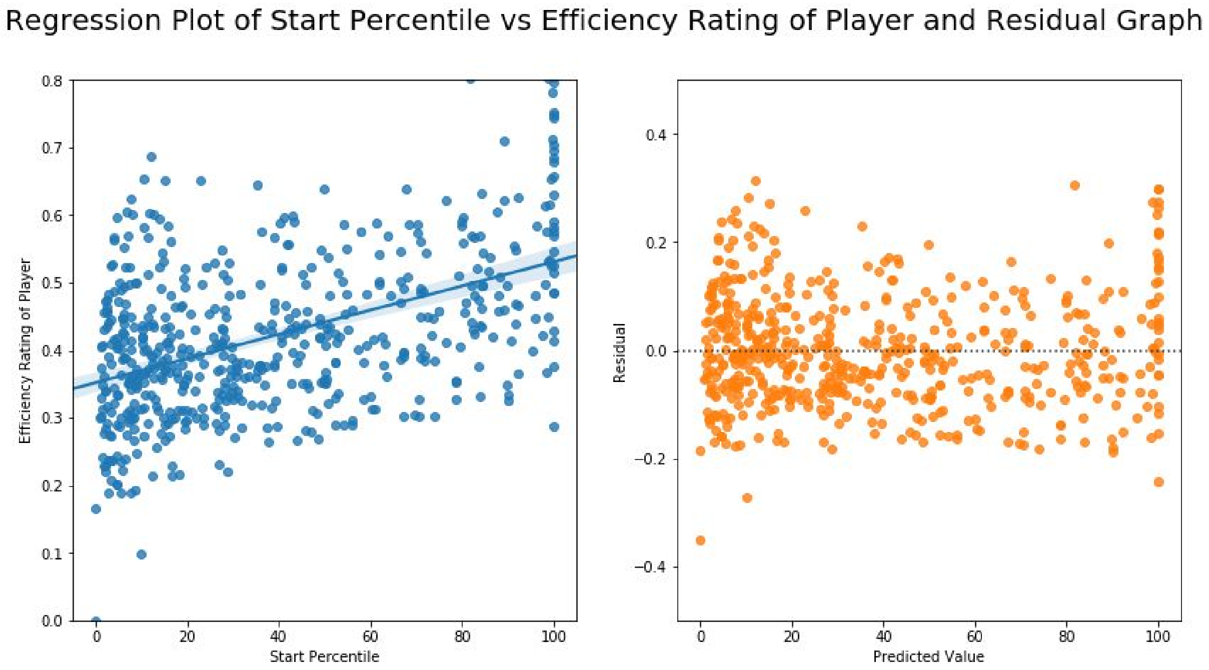


Figure 3. Shows the the starting percentile vs. efficiency

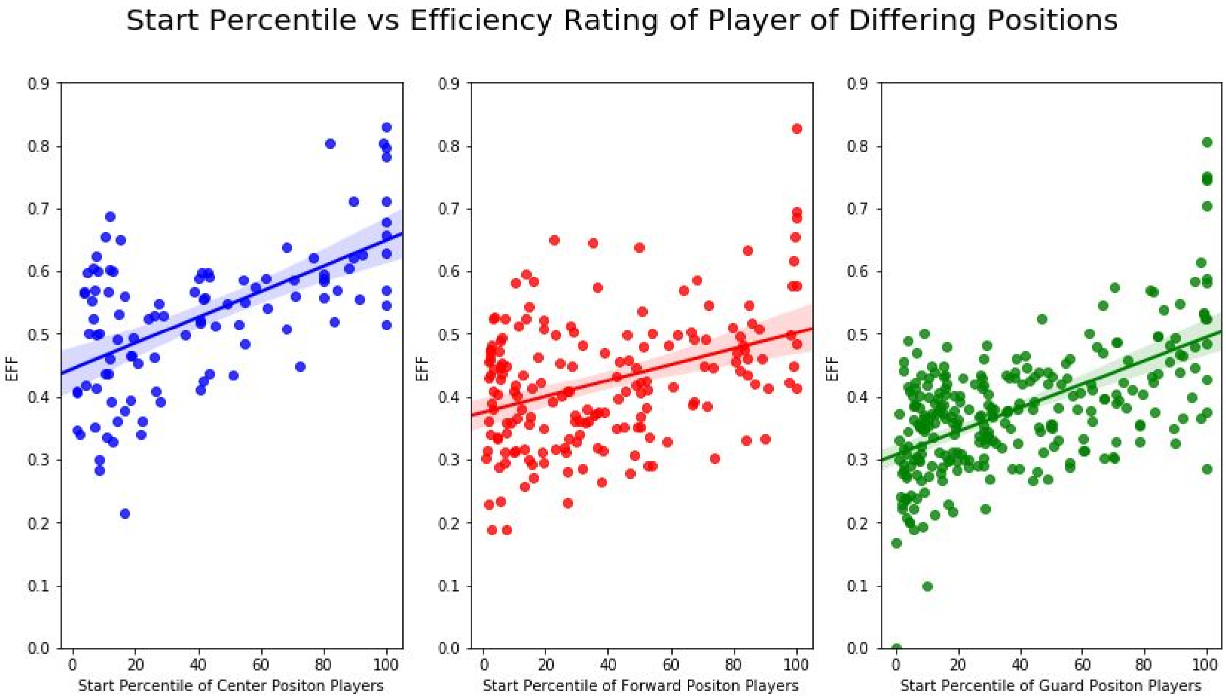


Figure 4. Shows the the starting percentile vs. efficiency, separated by player position

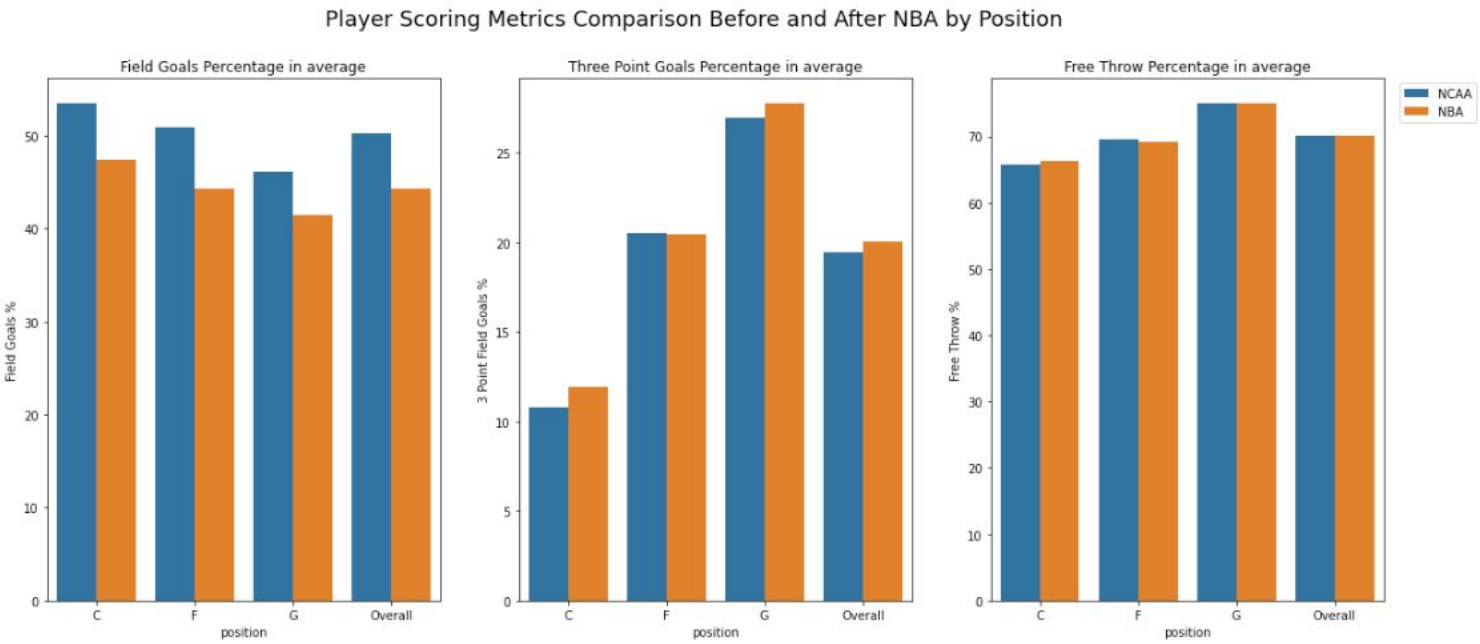


Figure 5. Shows the comparison of player's performance before and after NBA

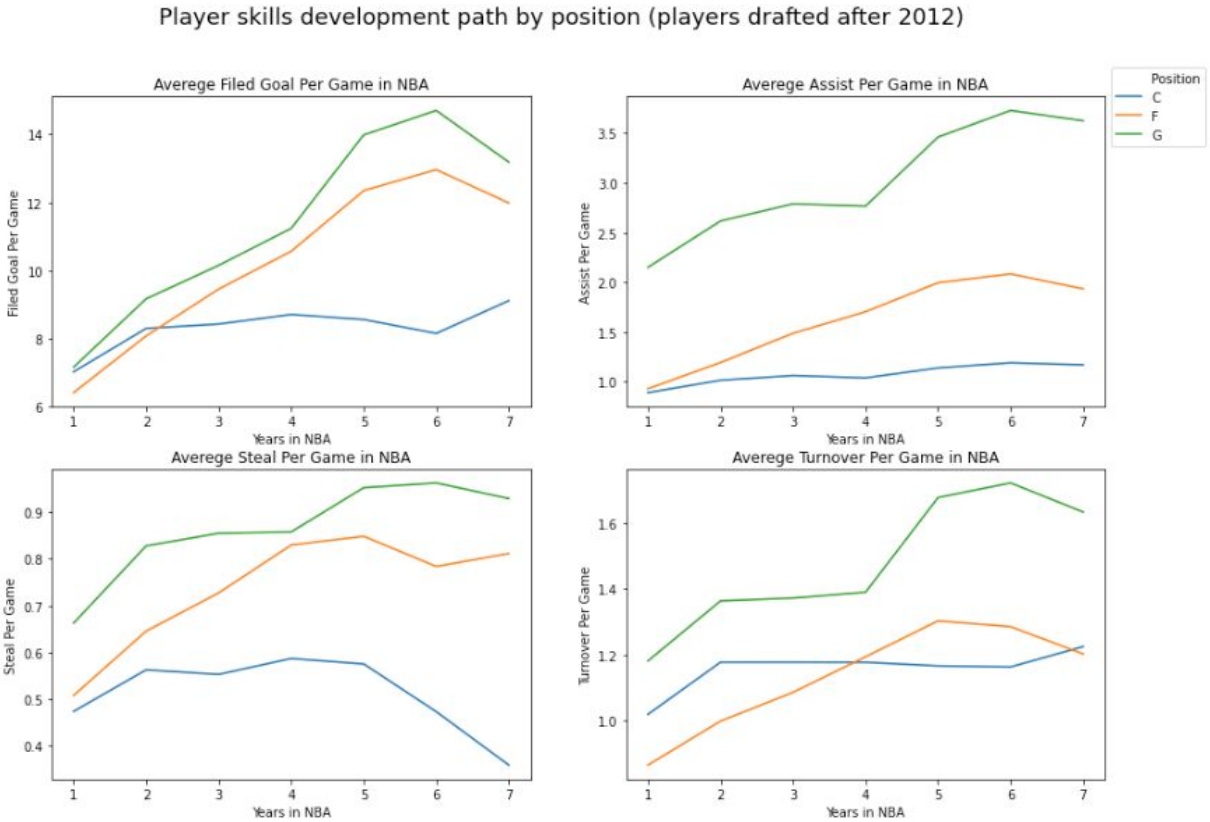


Figure 6: Shows the skill development path of players in the timeline of years they joined NBA

References

https://www.spotrac.com/nba/positional/
https://en.wikipedia.org/wiki/Efficiency_(basketball)
https://shottracker.com/articles/the-3-point-revolution
http://www.bigbluehistory.net/bb/Statistics/ratings.html
https://heavy.com/sports/2019/05/nba-player-average-salary-how-much-highest-paid/
https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba