Predicting NBA Success Through Performance Metrics

Jaiden Ha, Sarah Lodin, Alfred Mastan, Daniel Vargas, Michelle Yoon

Math 42: Introduction to Data-Driven Modeling

Professor Tony Wong

December 18, 2024

PREDICTING NBA SUCCESS THROUGH PERFORMANCE METRICS

Introduction

Background of Modeling Problem

The National Basketball Association (NBA) is a professional basketball league that consists of 30 teams (“About the NBA”). The different teams are split amongst the Western and Eastern conferences, 29 of the teams are from the United States, and one is from Canada, all of which feature “the best basketball players in the world” (“About the NBA”). The NBA’s regular season runs from October to April each year. One of the most important events in the NBA is the NBA Draft, typically held in June following the NBA Finals. The NBA Draft takes place once a year. For 2024, the NBA Draft was held from June 26-27, during the NBA offseason. Newly this year, the Draft was expanded to a two-night format instead of all taking place in one night (“NBA Draft: How…”).

The draft provides an opportunity for teams to take in new players who have not played in the NBA before. Those eligible for the draft are “all players over the age of 19” who have been out of high school for at least one year (“About the NBA”). Additionally, these players must have made an official declaration for the NBA Draft. The pool of applicants also includes international players at least 22 years old who have already committed to playing for an international professional team (“About the NBA”). Another applicant group considered includes US-born players who have completed their four years of college eligibility. In this process, approximately 60 players considered to be prospects for the NBA are drafted by teams during the NBA Draft.

To evaluate the potential of college basketball players to transition into the NBA and predict their future success, we focus on the players’ college statistics from their final year of college. This reflects the player's most recent and relevant performance. Then, the Player Efficiency Rating (PER) formula is applied to obtain a rating for the player. Our approach involves using a normal distribution to set the threshold for different performance states, taking into account the current position of players. As this threshold is determined by the normal distribution of actual players in the NBA, the method acknowledges that most college players may initially fall below the NBA standard and are expected to change over time.

Precise statement of the modeling problem in mathematical language

Statement: Use players' performances in college to simulate their performance progression and career longevity using transition matrices that are based on NBA players' careers from 1980 to the present.

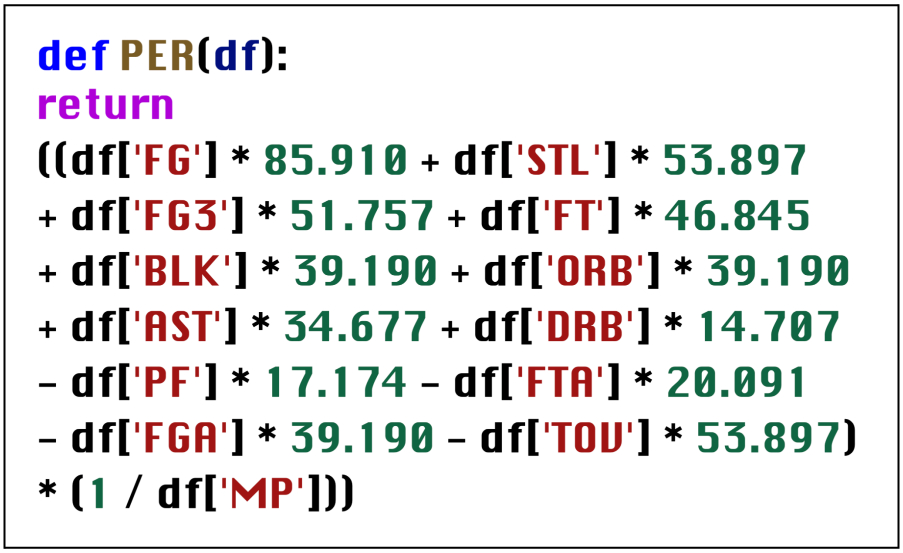
The primary goal of this project is to use college players' performances to simulate their potential NBA career trajectories. We employ transition matrices that are based on NBA players' careers from 1980 to the present, allowing us to model their performance progression and career longevity.

Assumptions and Variables

Assumptions are a crucial part of enabling the creation of a mathematical model. The model is predicated on several fundamental assumptions. One assumption is that a player's performance in their college years affects their performance within the NBA. This assumption allows us to use college statistics to predict players' NBA success. Another assumption being made is that the PER accurately measures the player's efficiency. This is an important assumption, as the majority of the model is based on the efficiency ratings for both college basketball players and NBA players. Furthermore, the model assumes that PER is a consistent, normalized measure over time and that a player's efficiency rating changes from season to season. For the model to be valid across different time periods, a standardized measure of performance is required. Making these assumptions allows the group to move forward with the development of the model.

Building the mathematical model necessitates taking several variables into account. The primary ones consist of the player name and factors which are used to calculate the PER. The formula used to calculate the PER in the model is shown below in Figure No. 1. Measures of the player's performance are used to calculate the PER and consist of number of field goals (FG), steals (STL), 3-Point Field Goals (FG3), free throws (FT), blocks (BLK), offensive rebounds (ORB), assist (AST), defensive rebound percentage (DRB), Personal fouls (PF), free throw attempts (FTA), field goal attempts (FGA), turnovers (TOV), and the minutes played (MP).

Figure No.1



```
def PER(df):
return
((df['FG'] * 85.910 + df['STL'] * 53.897
+ df['FG3'] * 51.757 + df['FT'] * 46.845
+ df['BLK'] * 39.190 + df['ORB'] * 39.190
+ df['AST'] * 34.677 + df['DRB'] * 14.707
- df['PF'] * 17.174 - df['FTA'] * 20.091
- df['FGA'] * 39.190 - df['TOV'] * 53.897)
* (1 / df['MP']))
```

This formula used to calculate the PER for this project follows a form that has been simplified when compared to the original PER equation put out by John Hollinger (Fein, 2009). Note how, in the PER formula, different things that contribute to the overall player performance are given different weights to reflect their varying levels of importance. The defensive rebound percentage has a weight of 14.707, while the field goals are given a weight of 85.910. This means that their defensive rebound percentage is considered less important than the number of field goals a player has made.

Pre-Analysis

Building the model required the team to use two different datasets, both of which are from Kaggle. The first dataset is “College Game Statistics of NBA Players” and contains information about college basketball players and data specifically relevant to calculating the player-efficiency rating. The data spans from 1980 all the way to 2022. The PER for each specific player from this data set is translated into their ‘initial state’ or their ‘starting point’ within the simulation. The second dataset is “NBA Stats (1947-Present)” and is used to help determine how each player will progress through the different basketball seasons. Specifically, this gives the probabilities used for the transition matrices of the Markov chain.

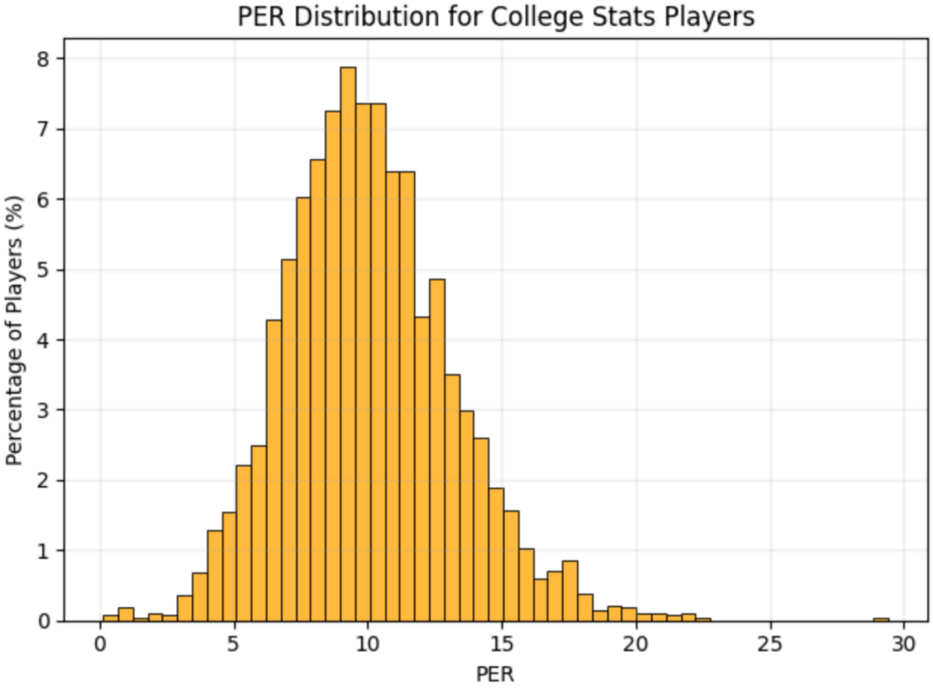
Data Clean-up

One thing the group had to take into consideration is that the player-efficiency rating is inaccurate if there is not enough data about a player. The effect this has is that it artificially skews the PER, especially when the player lacking data is compared to players for which we have a more complete picture. An example of this is a player who is in the game for only 5 minutes but scores several points, which gives them a very high rating but is not accurate when taking into account that they have been there for an insufficient amount of time. There is less overall data on their performance, which gives them an inaccurate PER. The criteria necessary for a player’s information to be kept is that they have played either over 40 games or have over 1,000 minutes played. If a player does not meet these criteria, their information is not used to build the simulation.

Performance Thresholds

Using the "College Game Statistics of NBA Players" dataset to calculate the PER gives the distribution of players. From Figure No. 2 below, it can be seen that the majority of college basketball players have a PER of around 10. The rating that the college basketball players receive here determines their initial state in the simulation.

Figure No. 2

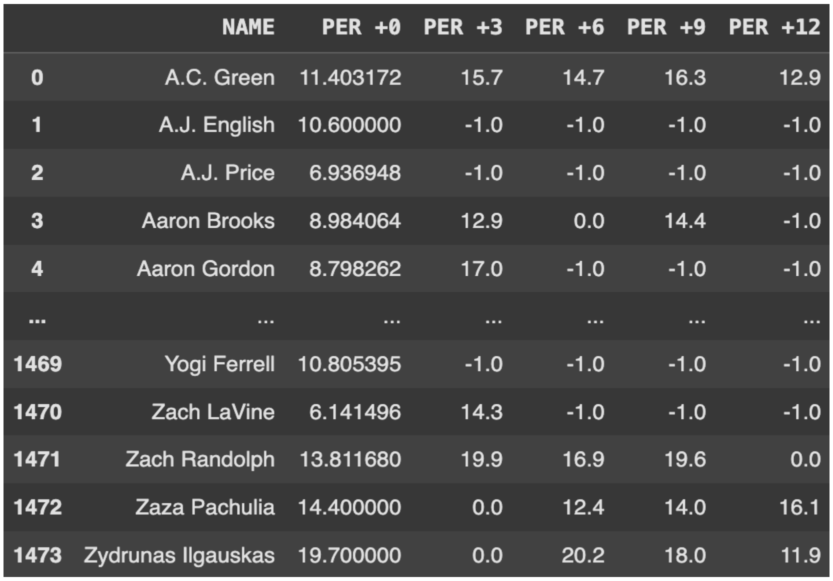


The data derived from the "NBA Stats (1947-Present)" dataset determines the different thresholds for the performance states. After obtaining the PER results for NBA players, it is seen that the average PER for the NBA is around 15. Determining the different thresholds is done by looking at different quartiles, calculating the interquartile range, and using this to then get upper and lower bounds for the thresholds. Overall, it yields that players who have a PER above or equal to 17.5 are placed in a state of "good." Players who are "decent" fall anywhere from a PER of 15 (inclusive) to less than 17.5. "Mediocre" players range from 12.5 (inclusive) to less than 15. The players placed in a state of "bad" have a PER of 0 to less than 12.5. Additionally, an "out of the league" state exists and reflects that the player is not playing for the upcoming season, indicated by the PER value of -1.

Results and Analysis

Computational and Mathematical Analysis

Having the states with their threshold defined, we filtered the data by taking their PER every timestep of *t = 3* years, starting from their latest college years for each player. Note that additional assumptions were made to handle the missing years or data as we took a leap in each timestep. We are considering that the players are "Out of The League" after not playing for 2 consecutive timesteps and also for the rest of their careers. However, if the players return in the next timestep, we consider them "BAD" players with the lowest PER of 0.



	NAME	PER +0	PER +3	PER +6	PER +9	PER +12
0	A.C. Green	11.403172	15.7	14.7	16.3	12.9
1	A.J. English	10.600000	-1.0	-1.0	-1.0	-1.0
2	A.J. Price	6.936948	-1.0	-1.0	-1.0	-1.0
3	Aaron Brooks	8.984064	12.9	0.0	14.4	-1.0
4	Aaron Gordon	8.798262	17.0	-1.0	-1.0	-1.0
...	...	...	...	...	...	...
1469	Yogi Ferrell	10.805395	-1.0	-1.0	-1.0	-1.0
1470	Zach LaVine	6.141496	14.3	-1.0	-1.0	-1.0
1471	Zach Randolph	13.811680	19.9	16.9	19.6	0.0
1472	Zaza Pachulia	14.400000	0.0	12.4	14.0	16.1
1473	Zydrunas Ilgauskas	19.700000	0.0	20.2	18.0	11.9

Based on the example data, PER +0 shows their latest college years, while PER +3 shows their PER 3 years after being drafted and continues until PER +12.

Extracting Probabilities From Data

With the extra assumptions, we can separate the players into different states and keep track of the number of players in each state for every timestep. In order to extract the probabilities, we track each player's progress, starting from their initial state over the time interval, and divide them by the total number of players in their initial state. As an example, the probability in the first time interval (Year 1 to 4) of players in the "BAD" state (Year 1) progressing to "MEDIOCRE" (Year 4) was calculated by taking the total players who were both in the "MEDIOCRE" state and were in the "BAD" state and dividing them by the total players in the "BAD" state. These were then calculated for each state and each time interval, achieving different transition matrices that are dependent on the time.

Building the Markov Chain

Now, with the probabilities extracted from the data on actual NBA players and how their performance changed over the course of their respective careers, we are able to build our Markov chain. Given the decision to analyze performance data in time intervals (i.e., Year 1 to Year 4), our probabilities for each timestep are different from one another. Instead of having a consistent transition matrix for our model regardless of time, we have decided to build a time-dependent Markov chain (otherwise known as a 'non-homogeneous" Markov chain where the probabilities that our transition matrices are built on are based on a relevant time variable. A realistic example of this model would involve the usage of a timestep variable *t*, where the equation that represents the model would take *t* as an input, and the probabilities of the transition probabilities between the current state and the next state are dependent on *t*, whether that be exponentially or logarithmically.

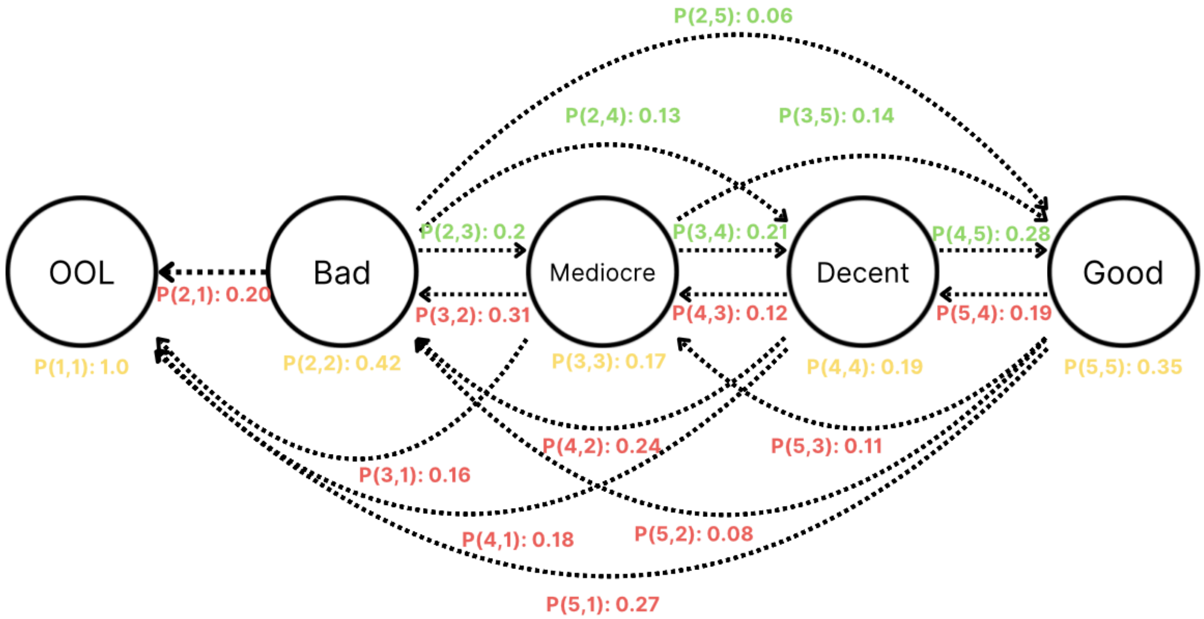
Given that this concept was out of the scope of the covered course material, we decided to build a Markov chain that would use pre-set probabilities that we had extracted in our analysis on real NBA data instead of being determined directly by a variable *t*. The problem with this is obviously that without a representative equation that takes *t* as input, it is impossible to scale our model infinitely, as we have to manually build the transition matrix for each individual time step.

As this will be explored more extensively later, there is actually no need for this at all since the model concludes in an absorbing matrix at a fairly early timestep.

Our initial version of this "time-dependent Markov chain" featured four different transition matrices, with each representing a time interval between Year 1 and Year 13: the first matrix represents Year 1 to Year 4, while the fourth one represents Year 10 to Year 13.

$$\underset{\text{Year 1 to 4}}{\begin{bmatrix} 1.0 & 0.20 & 0.16 & 0.18 & 0.27 \\ 0 & 0.42 & 0.31 & 0.24 & 0.08 \\ 0 & 0.2 & 0.17 & 0.12 & 0.11 \\ 0 & 0.13 & 0.21 & 0.19 & 0.19 \\ 0 & 0.06 & 0.14 & 0.28 & 0.35 \end{bmatrix}} \quad \underset{\text{Year 4 to 7}}{\begin{bmatrix} 1.0 & 0.23 & 0.34 & 0.22 & 0.14 \\ 0 & 0.36 & 0.19 & 0.10 & 0.07 \\ 0 & 0.23 & 0.17 & 0.15 & 0.10 \\ 0 & 0.11 & 0.2 & 0.24 & 0.15 \\ 0 & 0.06 & 0.11 & 0.28 & 0.54 \end{bmatrix}} \quad \underset{\text{Year 7 to 10}}{\begin{bmatrix} 1.0 & 0.39 & 0.39 & 0.28 & 0.23 \\ 0 & 0.29 & 0.25 & 0.18 & 0.05 \\ 0 & 0.18 & 0.18 & 0.23 & 0.10 \\ 0 & 0.09 & 0.15 & 0.21 & 0.15 \\ 0 & 0.05 & 0.03 & 0.09 & 0.47 \end{bmatrix}} \quad \underset{\text{Year 10 to 13}}{\begin{bmatrix} 1.0 & 0.65 & 0.58 & 0.49 & 0.23 \\ 0 & 0.18 & 0.22 & 0.15 & 0.10 \\ 0 & 0.11 & 0.13 & 0.19 & 0.07 \\ 0 & 0.06 & 0.06 & 0.13 & 0.23 \\ 0 & 0.01 & 0.01 & 0.03 & 0.37 \end{bmatrix}}$$

An example visualization of the relationships between each state is shown below in the form of a state diagram specifically representing the first timestep (Year 1 to Year 4).

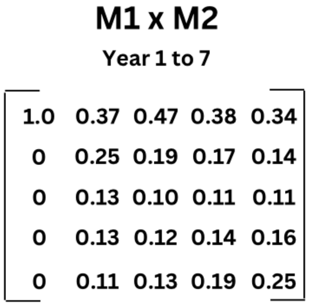


An observation that can be quickly made from this diagram, as well as the transition matrices, is that there is no possibility for a player to leave the state of OOL (or state 1) once they enter that state: this is shown in the state diagram via the absence of arrows leading out of state 1 as well as the first column being [1, 0, 0, 0, 0] in all transition matrices. This is due to the assumption that once a player is "out of the league," they do not return, as players who do return are, as mentioned before, considered the state based on their latest available season statistics during their gap years.

The numbers in the first row of our transition matrices, which represent the probabilities of a player leaving the league from the state respective to the column, increase with every timestep. This is evident in a clear outcome in which the model results in an absorbing state, which we explore further next.

Convergence in Transition Matrix

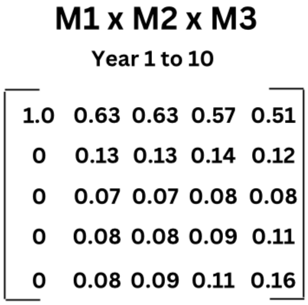
Establishing the presumption that players in the NBA will always ultimately leave the league due to declining performance, old age, permanently damaging injuries, etc., we must verify that the transition matrices in our Markov chain converge into a steady-state matrix with "OOL" being the absorbing state in order to confirm this presumption. As we are dealing with a time-inhomogeneous Markov chain, we used the equation $T(0 -> n) \ = T(0) \ . \ T(1) \dots \ T(n)$, where $T(i)$ is the transition matrix at time interval i. As we already hold the transition matrices over the time interval, we can utilize the equation to verify that the transition matrices will converge into a steady-state matrix. Multiplying the transition matrices determined previously using a 3-year timestep, our matrix products are as shown below:



M1 x M2

Year 1 to 7

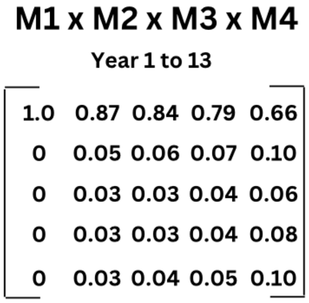
$$\begin{bmatrix} 1.0 & 0.37 & 0.47 & 0.38 & 0.34 \\ 0 & 0.25 & 0.19 & 0.17 & 0.14 \\ 0 & 0.13 & 0.10 & 0.11 & 0.11 \\ 0 & 0.13 & 0.12 & 0.14 & 0.16 \\ 0 & 0.11 & 0.13 & 0.19 & 0.25 \end{bmatrix}$$



M1 x M2 x M3

Year 1 to 10

$$\begin{bmatrix} 1.0 & 0.63 & 0.63 & 0.57 & 0.51 \\ 0 & 0.13 & 0.13 & 0.14 & 0.12 \\ 0 & 0.07 & 0.07 & 0.08 & 0.08 \\ 0 & 0.08 & 0.08 & 0.09 & 0.11 \\ 0 & 0.08 & 0.09 & 0.11 & 0.16 \end{bmatrix}$$



M1 x M2 x M3 x M4

Year 1 to 13

$$\begin{bmatrix} 1.0 & 0.87 & 0.84 & 0.79 & 0.66 \\ 0 & 0.05 & 0.06 & 0.07 & 0.10 \\ 0 & 0.03 & 0.03 & 0.04 & 0.06 \\ 0 & 0.03 & 0.03 & 0.04 & 0.08 \\ 0 & 0.03 & 0.04 & 0.05 & 0.10 \end{bmatrix}$$

The numbers in these matrices represent the probability that a player will ultimately perform at a certain level in a particular year, respective to their initial state. For example, the element in the 3rd row and 2nd column of the product of transition matrices 1 to 3 indicates that there is a 6% probability that a player that enters the league as bad is performing at a mediocre level in Year 10 of their career.

The product matrices also further signal toward the ultimate outcome of a steady-state matrix with probabilities nearing 1 in row 1, which leads us to compute product matrices beyond Year 13. As expected, the computation resulted in an absorbing matrix towards the OOL state.

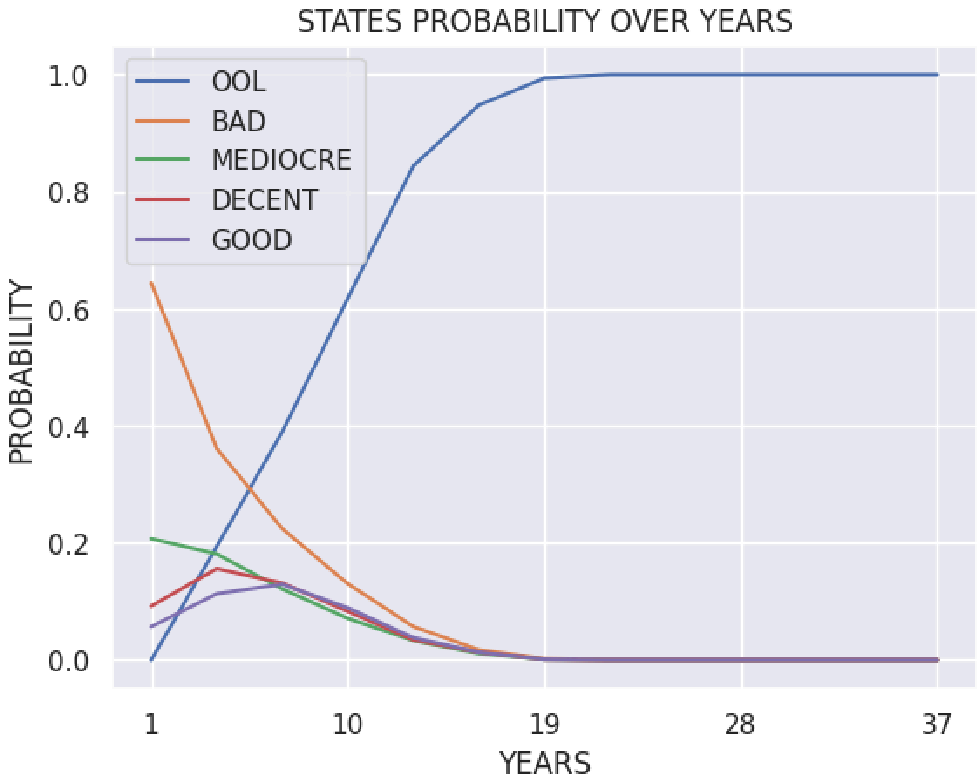
$$\begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Furthermore, the matrix converges at the sixth timestep, indicating that it converges at Year 19 after the players are drafted.

State Probability

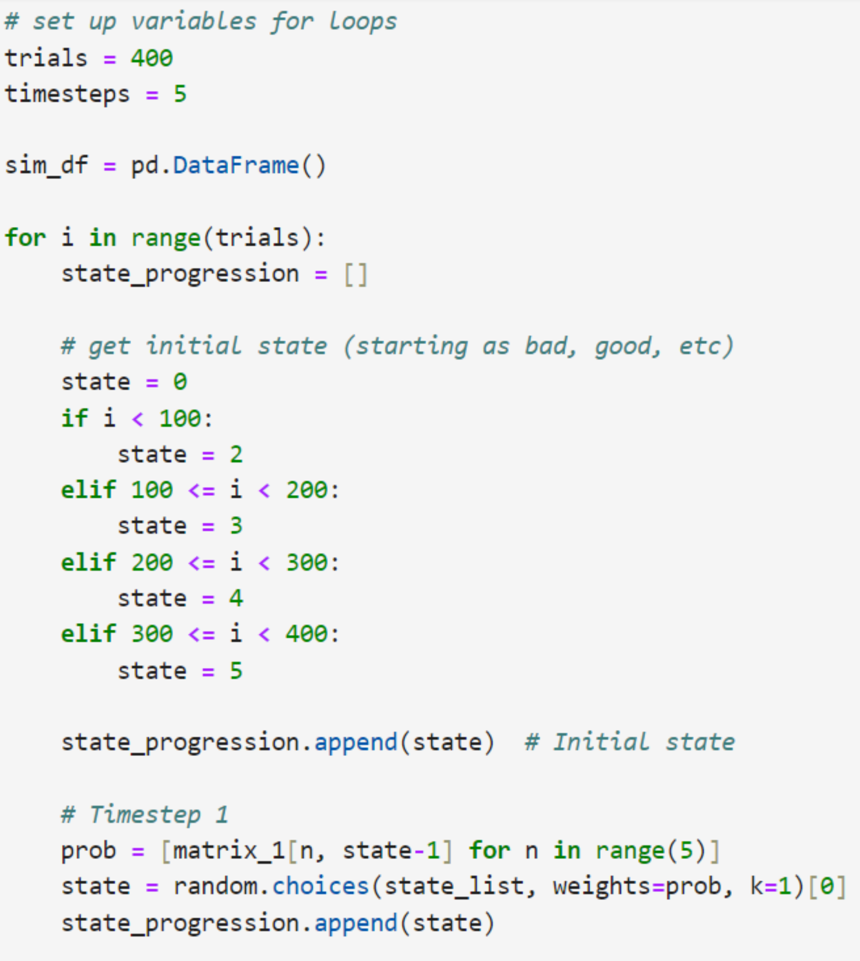
Looking at the bigger picture, we can further use these transition matrices to determine the probability of where players will end up in the upcoming years of their careers. Calculating the initial state probability by taking the number of players in each state and dividing them by the total players in the data, we achieved a probability of [0.000, 0.644, 0.207, 0.092, 0.057] going from OOL to GOOD, respectively. Analyzing it in a more mathematical way, the following equation is used, $x(n)\ =\ T(0\ -> n)\ .\ x(0)$, where $x(n)$ is the state probability at time n, $T(0\ -> n)$ is the product of transition matrices starting from 0 to n, and $x(0)$ is the initial state probability, to achieve the state probability at time n. As an example, figuring out the probability of which state the player will end up 10 years after they are being drafted is easily achieved by simply using the equation, which resulted in [0.618, 0.013, 0.071, 0.083, 0.089] going from OOL to GOOD, respectively. The results aid us in analyzing the overall progress of most players from the beginning of their careers.

The figure below serves as a visualization of the state probability converging at state 1 (OOL) in Year 19.

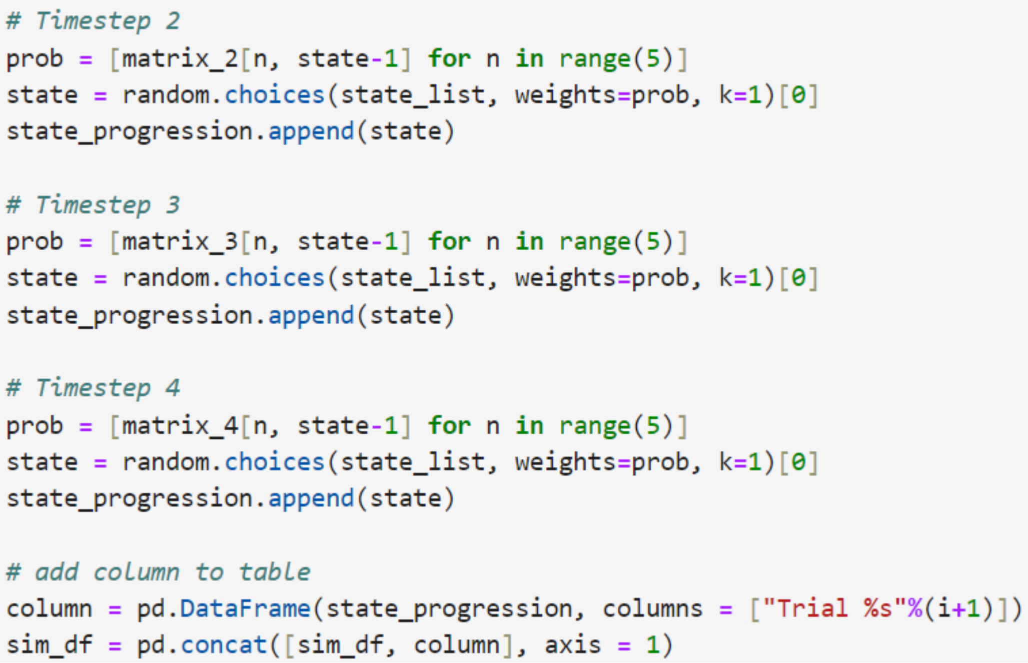


Simulation

Now, to get a better idea of what this model looks like in practice and thus understand the significance of these probabilities, we coded a simulation that ran 400 trials (with every 100 trials being run for a different initial state) through the model. The result of such would ideally give an idea of the distribution. This code is shown below.

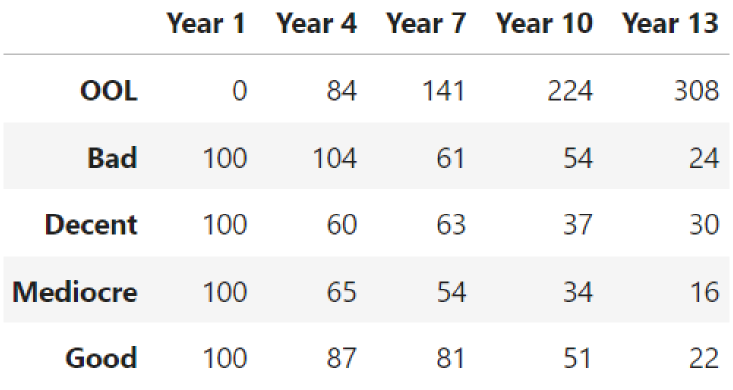


```
# set up variables for loops
trials = 400
timesteps = 5

sim_df = pd.DataFrame()

for i in range(trials):
    state_progression = []

    # get initial state (starting as bad, good, etc)
    state = 0
    if i < 100:
        state = 2
    elif 100 <= i < 200:
        state = 3
    elif 200 <= i < 300:
        state = 4
    elif 300 <= i < 400:
        state = 5

    state_progression.append(state)  # Initial state

    # Timestep 1
    prob = [matrix_1[n, state-1] for n in range(5)]
    state = random.choices(state_list, weights=prob, k=1)[0]
    state_progression.append(state)
```



```
# Timestep 2
prob = [matrix_2[n, state-1] for n in range(5)]
state = random.choices(state_list, weights=prob, k=1)[0]
state_progression.append(state)

# Timestep 3
prob = [matrix_3[n, state-1] for n in range(5)]
state = random.choices(state_list, weights=prob, k=1)[0]
state_progression.append(state)

# Timestep 4
prob = [matrix_4[n, state-1] for n in range(5)]
state = random.choices(state_list, weights=prob, k=1)[0]
state_progression.append(state)

# add column to table
column = pd.DataFrame(state_progression, columns = ["Trial %s"%(i+1)])
sim_df = pd.concat([sim_df, column], axis = 1)
```

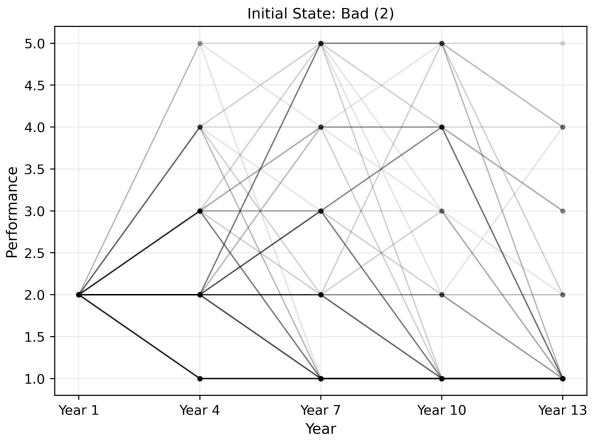
Looking at the simulation results holistically, we are able to track how many players were performing at each state at every timestep, showing data on all 400 players.

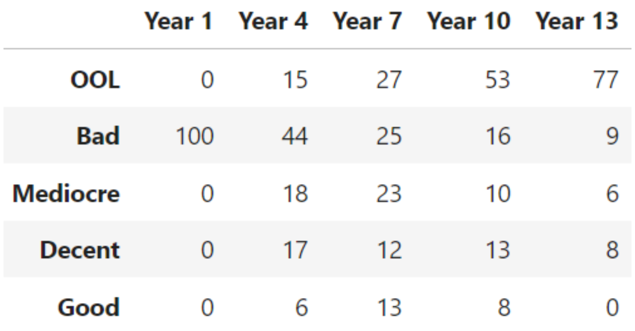


	Year 1	Year 4	Year 7	Year 10	Year 13
OOL	0	84	141	224	308
Bad	100	104	61	54	24
Decent	100	60	63	37	30
Mediocre	100	65	54	34	16
Good	100	87	81	51	22

Beyond this table, we looked specifically at each initial state and what the respective distributions looked like if we were to separate the trials by the initial state.

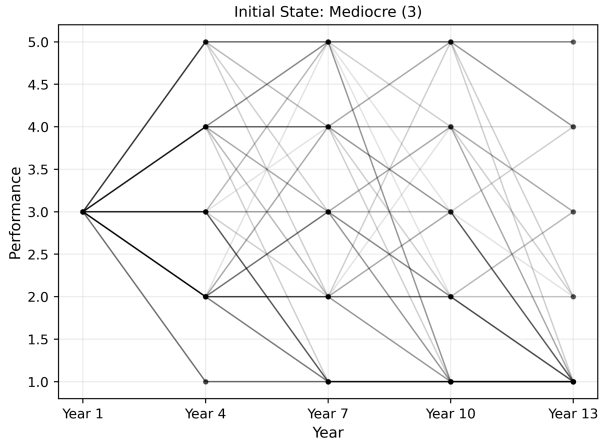
Distribution of Trials with an Initial State of **Bad (2)**

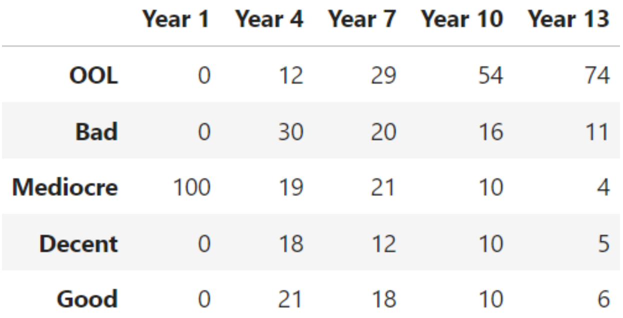




	Year 1	Year 4	Year 7	Year 10	Year 13
OOL	0	15	27	53	77
Bad	100	44	25	16	9
Mediocre	0	18	23	10	6
Decent	0	17	12	13	8
Good	0	6	13	8	0

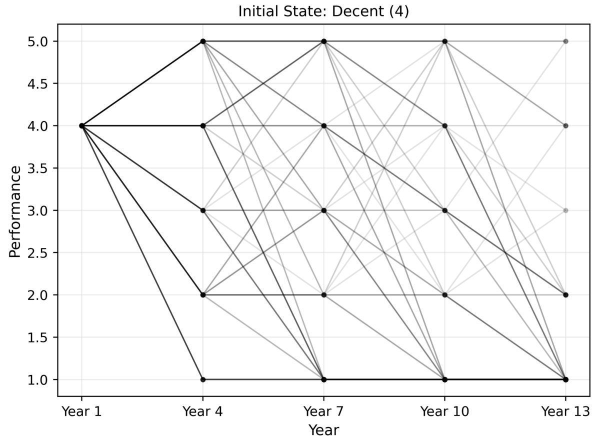
Distribution of Trials with an Initial State of **Mediocre (3)**

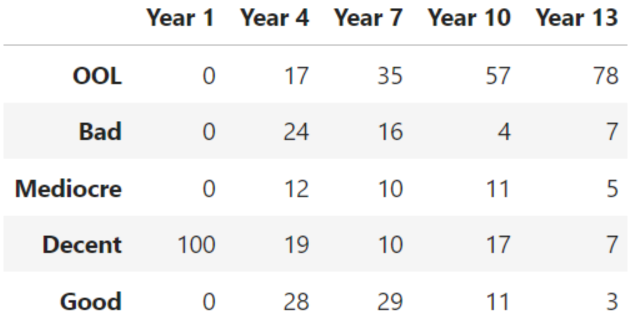




	Year 1	Year 4	Year 7	Year 10	Year 13
OOL	0	12	29	54	74
Bad	0	30	20	16	11
Mediocre	100	19	21	10	4
Decent	0	18	12	10	5
Good	0	21	18	10	6

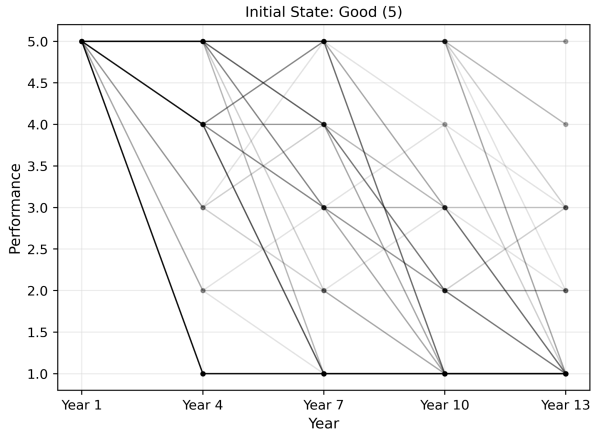
Distribution of Trials with an Initial State of **Decent (4)**

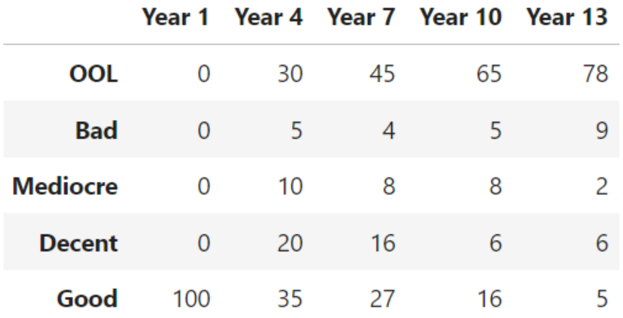




	Year 1	Year 4	Year 7	Year 10	Year 13
OOL	0	17	35	57	78
Bad	0	24	16	4	7
Mediocre	0	12	10	11	5
Decent	100	19	10	17	7
Good	0	28	29	11	3

Distribution of Trials with an Initial State of **Good (5)**





	Year 1	Year 4	Year 7	Year 10	Year 13
OOL	0	30	45	65	78
Bad	0	5	4	5	9
Mediocre	0	10	8	8	2
Decent	0	20	16	6	6
Good	100	35	27	16	5

In each line plot (displayed on the left), the opacity of the lines represents the concentration of trials in which the results are present in that certain state transition. One trend that stays consistent is that the lines representing transitions to state 1 (OOL) are of much thicker opacity in all 4 line plots; with the plot representing trials with an initial state of bad having even thicker opacity for those lines. As the initial state increases, the numbers of players present in higher states indicative of better performance increase for every recorded year: this is further supported by the lines in the plots having thicker opacity in the higher performing states as compared to lower performing states.

Conclusion

Our model aims to predict NBA success based on a player's college performance metrics. We utilized Player Efficiency Ratings and a Markov Chain built from transition matrices to achieve this goal. Over the course of this report, we have collected information on every NBA athlete since 1980, determining their efficiency and performance in games for every season. With this information, we have learned probabilities of changes in efficiency throughout their careers. Noticing that a standard Markov chain was not representative of our research, we placed these probabilities into transition matrices, creating a model similar to a time-dependent Markov chain to simulate possible career trajectories for incoming NBA players based on their efficiency rating in college.

Analyzing the PER of NBA players allowed us to differentiate different thresholds in which college basketball players could be placed in our simulation. The different states include “out of league,” “bad,” “mediocre,” “decent,” and “good.” The transition matrices show the probability that the players move from one state to another. Overall, everyone eventually leaves the league, which is on par with what happens in real life.

Before simulating our results, we wondered: Is a player's initial state reflective of how they will perform throughout their time in the NBA? Our results show a reasonably loose correlation between a player's initial efficiency and progress over time for all states except for players we deemed good. This correlation becomes much stronger for initially good players, shown through the higher probabilities expressed in the individual transition matrices in the 5th-row 5th column as well as the 1st-row 5th column, which describes the probabilities that a good player remains good or retires.

There is also the question of what a typical career progression looks like. Our simulation clearly indicates that "out of league" is an absorbing state. This means that regardless of how a player's career initially starts or what it looks like in the middle, every single player eventually ends up eventually leaving the NBA. This correlates with the real-life NBA career progression, where everyone eventually ends up leaving the NBA, whether this is due to factors such as age, injury, or retirement.

One limitation of our model was a lack of knowledge of the non-homogenous Markov chain, as the final model we created is not very accurate to what a time-dependent Markov chain should resemble. In the future, it would be beneficial to create a continuous equation where *t* can simply be input, and the probabilities are determined using *t.* Another limitation, although not as significant, was the fact that there are anomalies within the actual NBA, with a few players, such as Lebron James, having much higher career longevity than the norm, which in turn slightly skewed our data. A future model should account for outliers, possibly by creating an extra state specifically for "elite" players.

References

Datta, S. (2024). *NBA Stats (1947-Present) [Dataset]*. Kaggle. https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats?select=Player%2BTotals.csv

Fein, Z. (2009, January 19). *Cracking The Code: How to Calculate Hollinger's PER Without All the Mess*. Bleacher Report. https://bleacherreport.com/articles/113144-cracking-the-code-how-to-calculate-hollingers-per-without-all-the-mess

NBA. (2024, June 18). *About The NBA*. NBA News. https://www.nba.com/news/about

NBA. (2024, July 1). *2024 NBA Draft: How to watch, dates, selection order and things to know*. NBA News. https://www.nba.com/news/nba-draft-faq

Tan, C. (2022). *College Game Statistics of NBA Players [Dataset].* Kaggle. https://www.kaggle.com/datasets/calvintancy/college-game-statistics-of-nba-players/data