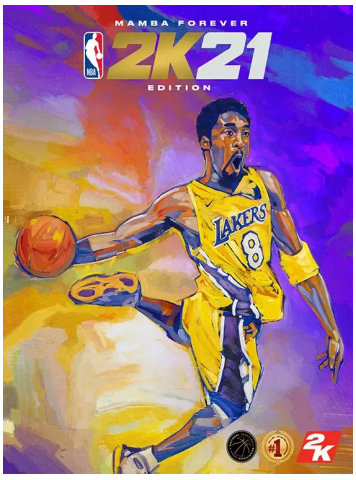
האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Internet & Society program

Course: 47717- Data Mining

Project Title:

How well does the NBA 2K video game predict the players' rating?



Team members:

Aviad Don 311586986 **aviad.don@mail.huji.ac.il**

Lahav Rom 204865398 **lahav.rom@mail.huji.ac.il**

Problem description:

NBA 2K is a series of basketball sports simulation video games that has been developed and released annually since 1999. The goal of each game in the series is to emulate the sport of basketball, more specifically, the National Basketball Association (NBA). In the game, each player is given an overall ranking based on his abilities and performance in real life.

Each year, the NBA 2K game is published in proximity to the upcoming NBA season, for example: NBA 2K20 was published before the 2019-2020 season. Therefore, the 2k company is predicting the players' overall rating based on each player's performances, stats, ability, age and more, in the past year and how well will the player perform in the upcoming season.

The parameters we want to examine in this project are:

1. How well does the game predict players' success?
2. Can we characterize players that the game predicts wrong?
3. Has the prediction become more accurate over the years or has it declined?

General data description:

We found limited data containing the ratings of the players in the NBA 2k video game so we wrote scrapers using python's BeautifulSoup library, to build our own main data set.

We scraped the official NBA website (https://www.nba.com/players) to extract the general NBA player's stats and countries. Then we extracted the NBA 2K ratings from the past 7 years, from a website called 'HoopsHype' (https://hoopshype.com/) and added them to the data. Afterwards, we scraped each player's Wikipedia page to get his age. Some players have the same name as other famous people (Kevin Martin for example), so we had to fill in their age hard-coded.

Finally, we wanted quantitative estimates to compare to the player's ranking without being 'biased'. For example, Points per game can be a 'biased' estimate, because there are players that impact the game in different ways, like defense and more. So, ESPN calculated two estimations for each player called: RPM and WIN-Shares.

- **RPM (Real Plus-Minus):** " A player's estimated on-court impact on team performance, measured in net point differential per 100 offensive and defensive possessions. RPM takes into account teammates, opponents and additional factors".

- **Win-Shares**: "Provides an estimate of the number of wins each player has contributed to his team's win total on the season." (both descriptions are taken from the ESPN website:"https://www.espn.com/nba/statistics/rpm").

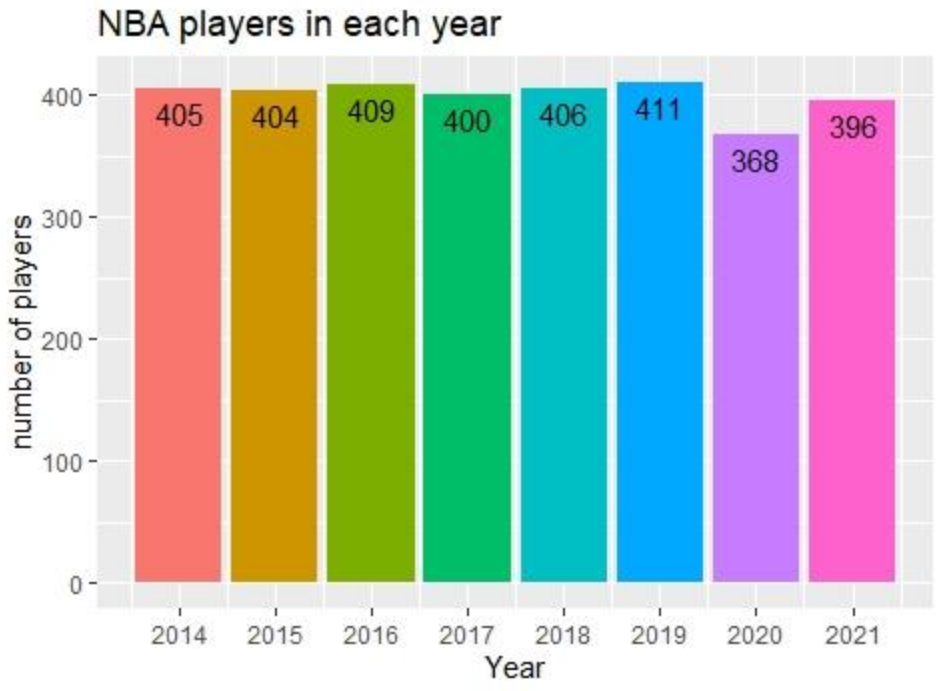
We scraped the players' RPM and Win-Shares from the past 7 seasons and merged them with the other data we had.

Our final data set contains 3200 rows (a row per player per season) and these columns: Name, Year (the relevant season), RPM, WIN-Share, 2k_Rating, Country and Age. Each row has the 2k rating, Win-Share and RPM of the same year, for example: if the year is 2020, the 2k rating of this player is for the 2019-2020 NBA season and it was assigned to him before the season even started, but the Win-Shares and RPM of this row are the real stats of this player for this season.

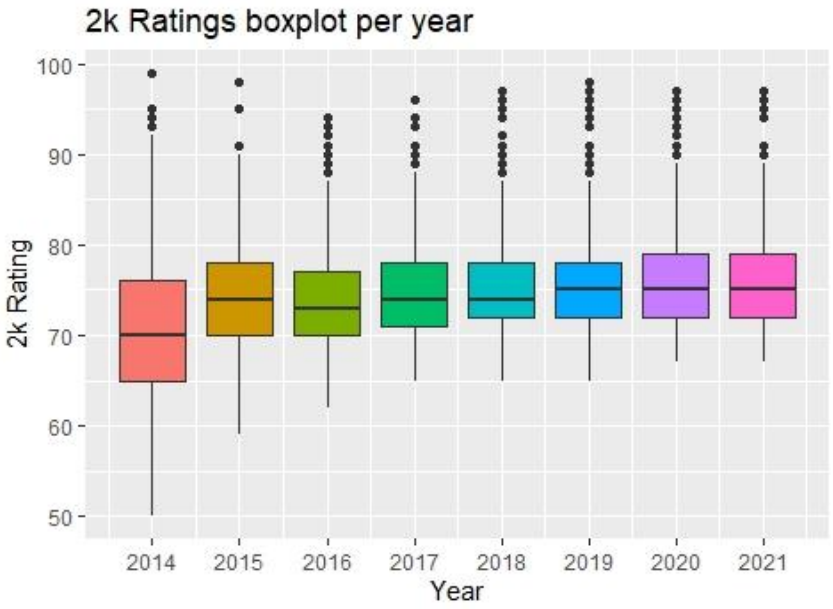
Exploratory analysis:

To check our data and make sure it doesn't contain any anomalies we ran a couple of statistics and graphs.

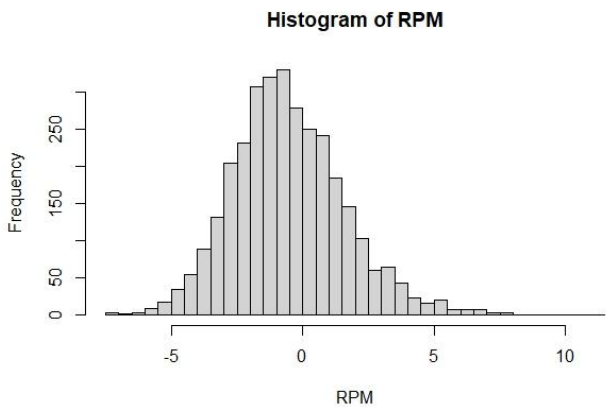
First of all, we checked the number of players in each year, the players that were counted are players that had an RPM and Win-Shares rating, which indicates that they were playing that season. We can see that each year the number is around 400 players, excluding the 2020 season, a season that struggled with COVID-19 and gave fewer opportunities to less experienced players.

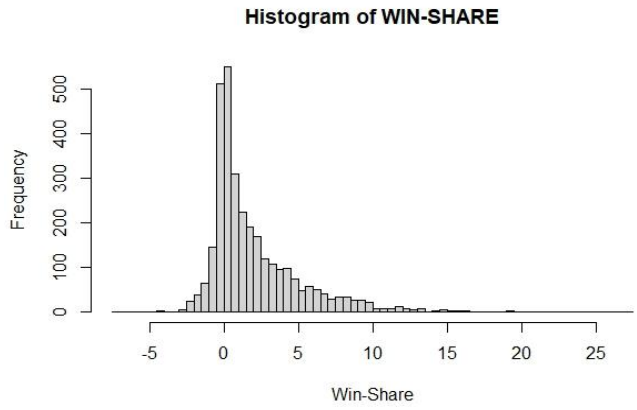


Next, we created a box plot graph showing where most of the players are rating-wise. We can see that in the years 2015 - 2021 the box plots are pretty much the same and go around the same area, where the average overall rating is around 74. 2014 was the last year the lowest rating could be 50, from then on, the lowest rating could be 60 instead of 50. That explains the big difference between the 2014 box plot and the rest of the boxes.



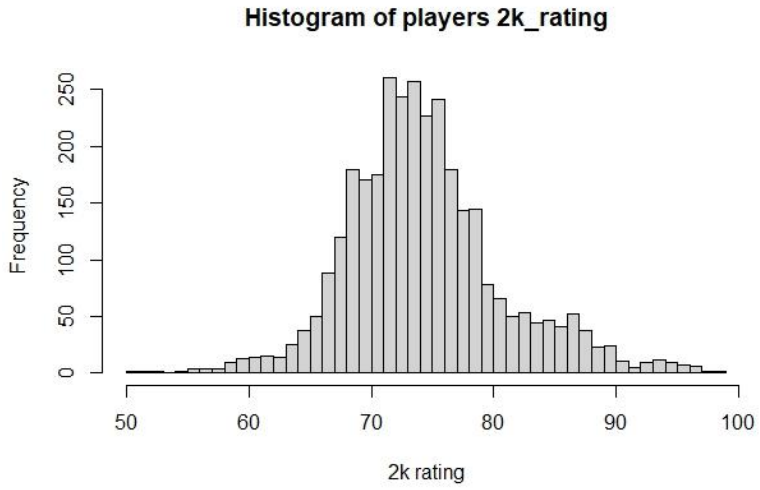
We checked the distribution of the RPM and Win-Shares using a histogram to make sure the data made sense.





As we can see, the RPM distribution is normal and the Win-Shares' distribution is also normal but with a right tail.

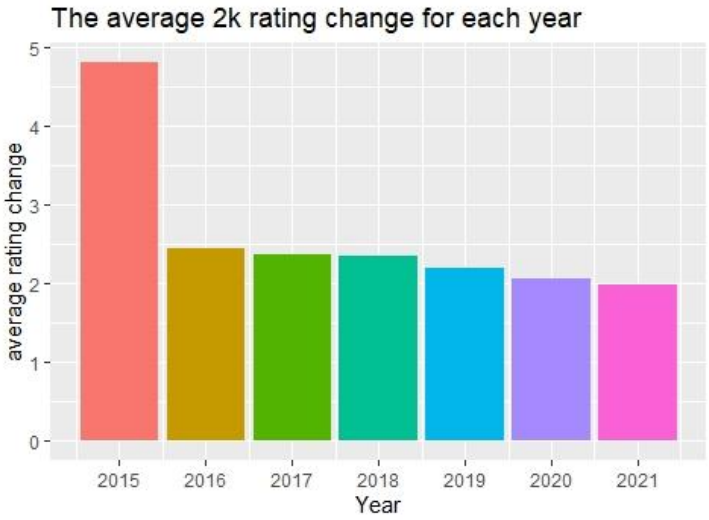
We also made sure that the 2K ranking distribution is normal:



Main analysis:

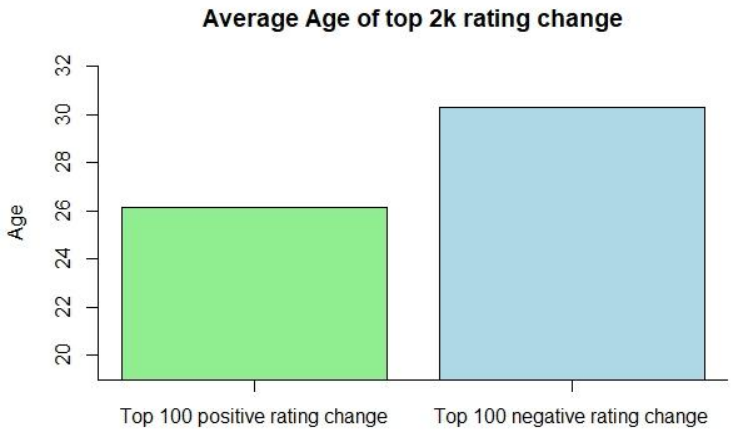
We used a couple of analyses to try and answer our questions:

1. We checked the correlations between the ratings and the win-shares and RPM. The correlation can be between -1 to 1, a higher correlation means linearly related. In other words, the closer the correlation is to 1, means higher rating would lead to higher Win-shares/RPM. The correlation between the 2k ratings and the win-shares is 0.57 and the correlation between the 2k ratings and the RPM is 0.52.
2. Another analysis we performed on the data was checking the change in ratings between years. After calculating the change in ratings between each year for each player, we calculated the average of the absolute change of each year. Being the first data' the year of 2014 doesn't appear because it can't have any rating change.



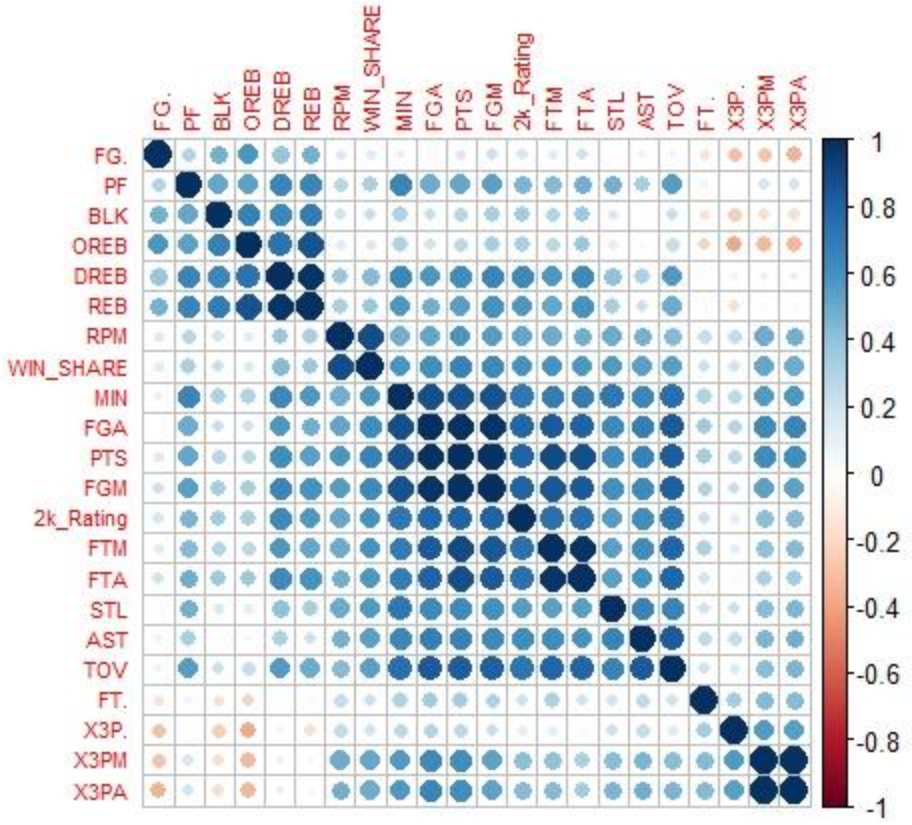
As we can see from the graph, first there is a drop in the average rating from 2015 to 2016 because the lowest rating rose from 50 to 60 from 2015, that's why the average rating change in 2015 is much higher than in the rest of the years. In addition, we can see that in each year, the average rating decreases, meaning that 2k predictions have improved over the years because in each year the average change they need to implement becomes smaller.

3. We wanted to check if we can characterize the players that 2k predicts wrong, meaning with the highest rating change. We thought they might have trouble predicting the correct rating of players outside the USA because there aren't many non-USA players in the NBA. Therefore, we checked how many non-US players there are in our data. The fraction of players not from the USA in all of our data is- 745/3199 = 0.233 = 23.3%. Then, we checked the fraction of players not from the USA in the top 100 players with the highest rating change. The fraction is- 24/100 = 0.24 = 24%. Meaning we can't presume that they fail in non-USA players. Next, we checked age, we found that the average age of the top 100 players with the highest **positive** rating change is- 26.13, and the average age of the top 100 players with the highest **negative** rating change is- 30.32. That makes sense because, 26 is approximately the age the players peak so it is reasonable their rating will get higher, in contrast to age 30 when most of the players start to decline, that's why the average age of the top 100 **negative** rating change players is much higher (4 years in the NBA is a lot of time).



4. Another thing we wanted to see is whether there are certain statistics that affect the players' 2k ratings more than others. So, we created a correlation heat map between all the statistics we could get. As seen, points per game (PTS) and the other stats related to scoring, minutes per game (MIN), RPM and Win-Shares have the highest correlation with the 2k rating. Another stat that is highly correlated to the 2k rating and is interesting because it is a negative stat, is turnovers per game (TOV). Considering it is

an undesirable stat, we thought it should have a negative effect on the 2k rating, however, the players who have the highest turnovers per game are usually the best player's because they hold the ball more than other players.



Evaluation:

Evaluation Criteria:

Our main criteria will be to check mismatches in the 2K ratings and try to understand if we can characterize the players with the big changes.

A mismatch will be defined by a big change in rating between years. We realized that according to our data, a change of 5 points or more is less common. Therefore, we considered a big change as a change of 5 points or more in rating between seasons. A big change will indicate that the game's prediction of the success level of a player wasn't accurate and the game company had to update the player's rating the following year.

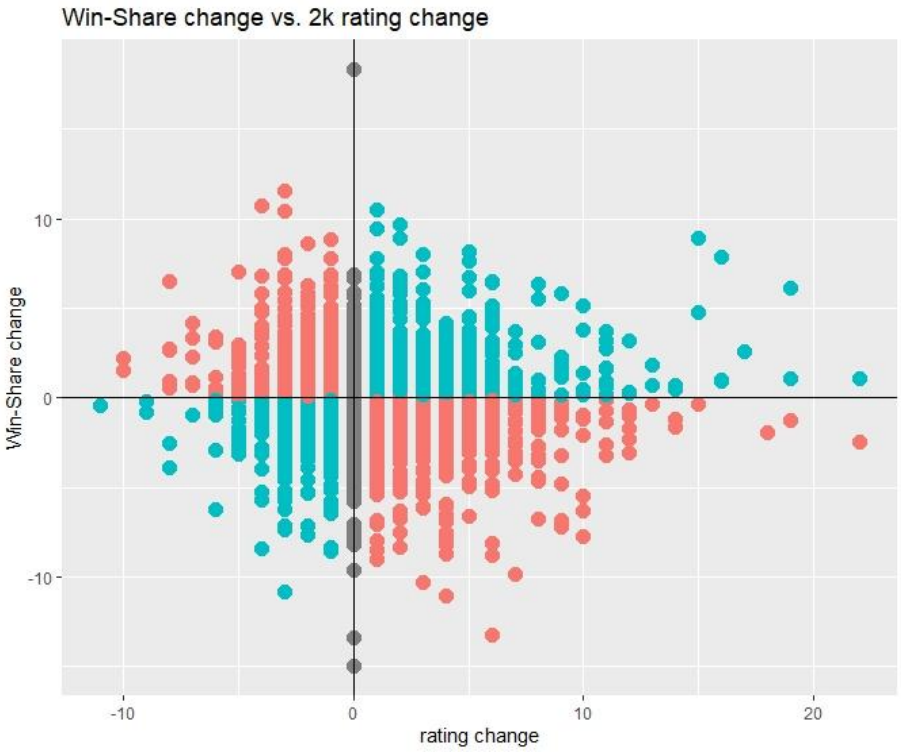
Following these mismatches, we will try to see if we can characterize these players. If the "mistakes" have non-similar statistics to the main data (age percentage, player's origin etc.), it will teach us that 2k tends to make mistakes in players with specific attributes. We would consider our work successful if we characterized the players whose rating prediction was wrong. Another definition of success would be if we can conclude that the 2k company doesn't

make many mistakes in their predictions and that the mistakes made are random and concern players nobody thought would improve (or worsen).

Setup:

As described in the main analysis, after we scraped all of our data, we calculated the rating changes, win-shares changes and RPM changes from year to year. We calculated the correlations to see if the Win-Shares and RPM are making sense. Then, we made a scatter plot to see how many "wrong predictions" 2k made.

Each point represents a player, the X-axis represents the rating change of the player from the past year and the Y-axis represents the Win-Share change of the player from the past year. We assume that the 1st and 3rd quarters (the blue points) are the "good" predictions because the 1st quarter means that the Win-shares of this player increased and his rating as well. Therefore, this player was actually better this season in terms of contributing wins to his team and 2k increased his rating meaning they thought he will be better this year.



The 3rd quarter is exactly the opposite, it concerns players whose win-shares and 2k rating decreased. This means the prediction of 2k was accurate because they predicted these players would decline (hence the 2k rating decreased), and actually, their win-shares decreased this year. However, it seems that the "wrong" predictions (the red points) in the 2nd and 4th quarters have as many points as the "good" predictions, so we can't tell for sure if 2k made more mistakes or not based on the 2k rating change and win-shares change.

Results:

We couldn't characterize the players' rating which 2k had to highly increase or decrease because they predicted wrong. We read about each of the 20 players with the highest rating change to see if there is anything unique we can extract from their stories, but it seems like when 2k predictions are wrong it is because these players really surprised everyone. For example:

- **Hassan Whiteside** - in 2015 his rating was 59 and in 2016 his rating increased to 81. After being selected at 33rd in the 2010 NBA draft, he played only 19 games in 2 years in the NBA due to injuries and lack of opportunities. After no team signed him, for 2 years he played in the NBA development league, in Lebanon and China. In 2015 he signed with the Miami Heat's G-league's team, after Miami's star Chris Bosh and his substitute Josh McRoberts got injured, Miami lacked big men so they gave Whiteside a chance for the rest of the season and he shined.
- **Rudy Gobert** - in 2014 his rating was 52 and in 2015 his rating increased to 71. He played in a mediocre team in the French league and didn't have any interesting stats, combined with the lack of scouting reports and videos of him playing, nobody thought he would succeed in the NBA. However, a young Utah team gave him a chance and he improved and showed his impressive skills.

Each player has his own unique story and how he improved.

We need to remember that basketball isn't an individual sport, meaning each player is dependent on his teammates. Consequently, a player can be great and his 2k rating can increase but if his team doesn't improve as well and succeed, his win-shares can be decreased. In addition, some teams improved and had success, but their best players cannot get higher ratings than other players, even though they had more team success. Therefore, 2k increases the ratings of the role players of these teams, so this team could be good in the video game even without a clear superstar. For example: In 2015, the Atlanta Hawks finished with the 2nd best record in the NBA, but they didn't have one dominant best player, that's why Kyle Korver's rating increased by 15 (!) and Demare Carroll's rating increased by 14 (!).

For these reasons, we believe that 2k ratings' prediction is very accurate (as should be since they make money from it) because the players they predict wrong don't fit a specific typecast and seem random.

Impediments:

The main issue we encountered was how to determine if the 2k rating prediction is right or wrong. There are many stat lines and variables that influence the probability of a single player becoming better or not, and many more that can't be quantified, like leadership, personal

interactions with other team members and more. Hence, it is impossible to predict every player's overall performance with the correct rating.
Another difficulty we faced was acquiring the data. Even though the internet is full of NBA stats and information, we couldn't find data that contains all the parameters we wanted. Therefore, we had to build scrapers and build our own dataset.

Future Work:

The data we scraped is separated from the playoffs stats while playoffs performances highly affect players' ratings, because that's when the competition gets tougher. Therefore, it can be fun to add the playoffs stats to see its impact.
When we created the correlation heat map to see which stat line affects more on the 2k rating, we saw that points are one of the main attributes. It will be interesting to try and build a different rating model based on other attributes and see if we can get a rating model with fewer "mistakes".

Conclusion:

We analysed in different ways NBA players' stats and compared them to their rating in the NBA 2K game, to check how well the game predicts the players' success in the league. Most of the results showed us that the NBA 2k video game rating model is very accurate, their mistakes aren't consistent and seem random and unbiased. We can say that even if they do make mistakes, it doesn't matter because in the past few years, they have updated the players' ratings during the season based on their real-time actual performances. In any case, in the beginning of every season the 2k ratings mainly concern the NBA players themselves, as can be seen in these videos:
https://www.youtube.com/watch?v=HwHYv9wI1II
https://www.youtube.com/watch?v=bRy2vzoy7eY