PERSONAL PROJECT

PREDICT ATHLETES' POINTS AND ANSWERING BUSINESS QUESTIONS

Developer: Marcos Matheus de Paiva Silva

Barra Mansa - RJ

March de 2022

Contents

1 Objective

The growth of technology brings great advances to numerous sectors from the military, education to sports. Currently, athletes have increasingly sophisticated equipment that in some cases is even used in space training. In this way, programming has been proving, little by little, indispensable in anyone's daily life. Therefore in the current project, we fully contemplate the stages of a data science project, so that from the information collected on the website nbastuffer and from the data visualization we can probe some questions that would help athletes, bettors and competitions understand about:

- Which age group, team and position score the most points per game?
- What exactly are the athletes who score the most and give the most assists simultaneously?
- What is the age group, time and position that makes the most assists?
- What exactly are the athletes who have the most turnovers and scores simultaneously?
- What age group, team and position do the most turnovers?

Finally, after creating the model, we provide an online application that allows any athlete to predict what score he would have in a training or real game (Deploy) . Such a procedure can be performed based on information collected or informed by the athlete, such as the number of attempts to fild goal from two points, among others.

2 Proposed Solution

This project was developed following the steps in (GéRON, 2019), and also based on some codes found in (MüLLER; GUID, 2016). To find the answers to the questions raised in the previous section, we first used the Google Colaboratory so that any individual can run the code in python and also better understand all the theoretical argument, insights and methodology used in the project.

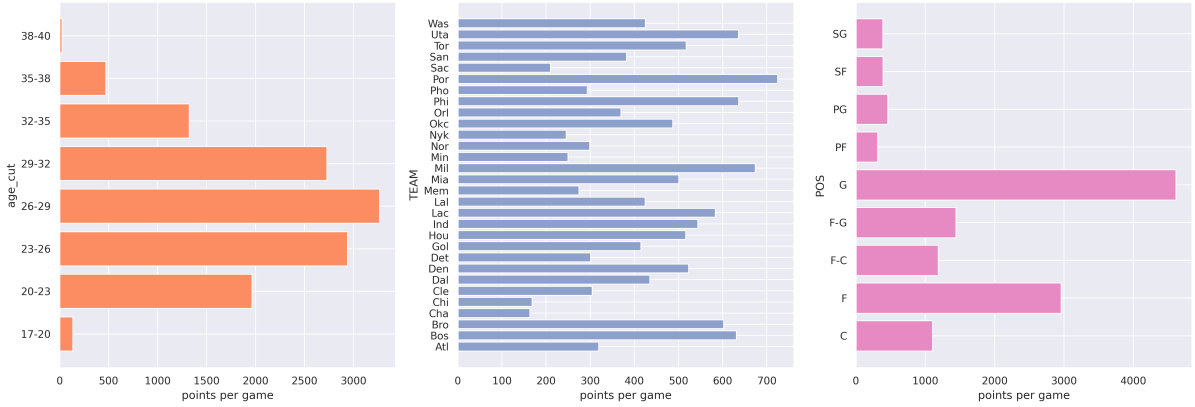
At Google Collaboratory we first web scraped the site nbastuffer using mainly requests and also Beautiful Soup and enter the extracted information to a CSV file. After this procedure, another jupyter notebook was used in which the libraries to operate with pre-processing, data analysis, graphics and files were pandas, scikit-Learn, matplotlib, seaborn numpy, seaborn, imblearn. The library chosen for the machine learning algorithm was extreme-gradient-boost.

Furthermore, after creating the machine learning model in Google Collaboratory, the library pickle was used to serialize the model and write it in a file. With the "trained_model.sav" file in hand, we implemented and locally tested the web application using the python language distribution anaconda and with the help of API streamlit we produce the app in a few lines of code. Finally, we host the web app on the Streamlit platform where you can test our basketball player score prediction app.

3 Results

Our Google Collaboratory project has gone through several steps both to gain initial insights and to understand and explore our NBA athlete dataset. In the data visualization stage, we plot a graph:

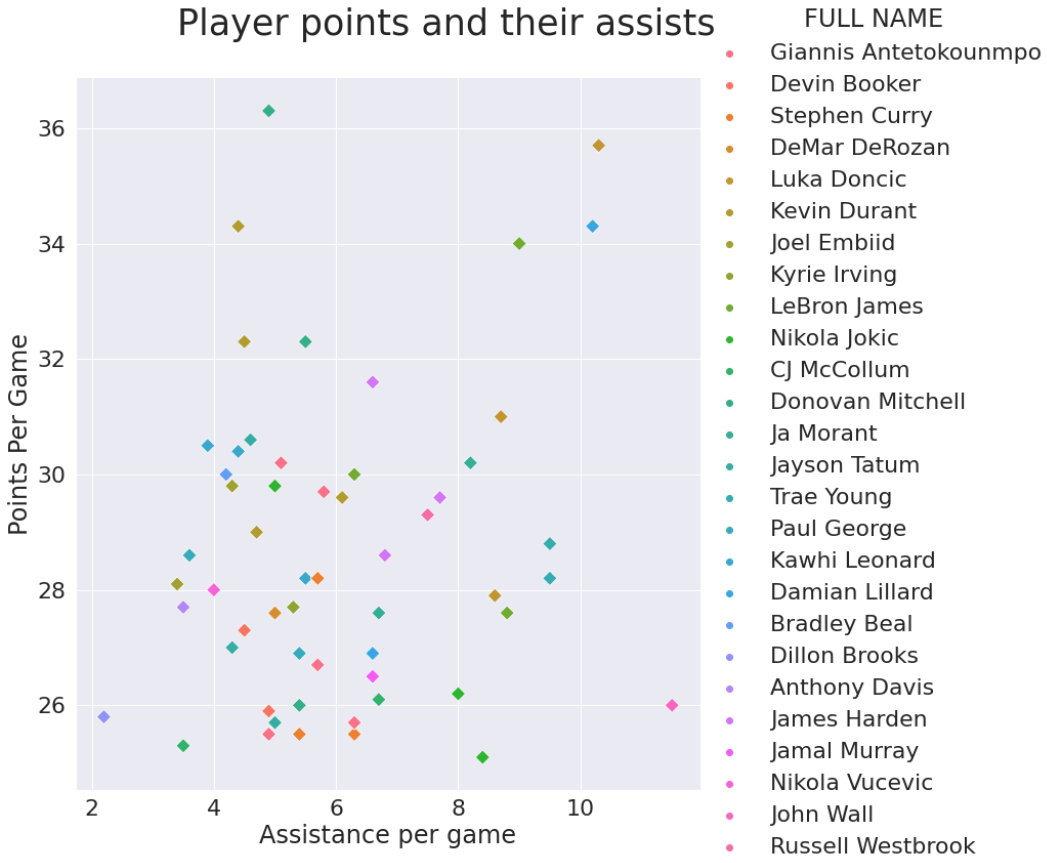
Player Points By Age Group, Team and Position



Source: Author

In this graph we have the classes that score the most points per game, so we understand the age of the players who score the most points varies from 26-29 years, and those who score the least are those with 38-40 years. Similarly, the team that scores the most is the Portland Trail Blazers (POR), and the one that scores the least is the Charlotte Hornets (Cha). Finally, the positions that score the most are Guard (G), and the ones that score the least are Power forward (PF).

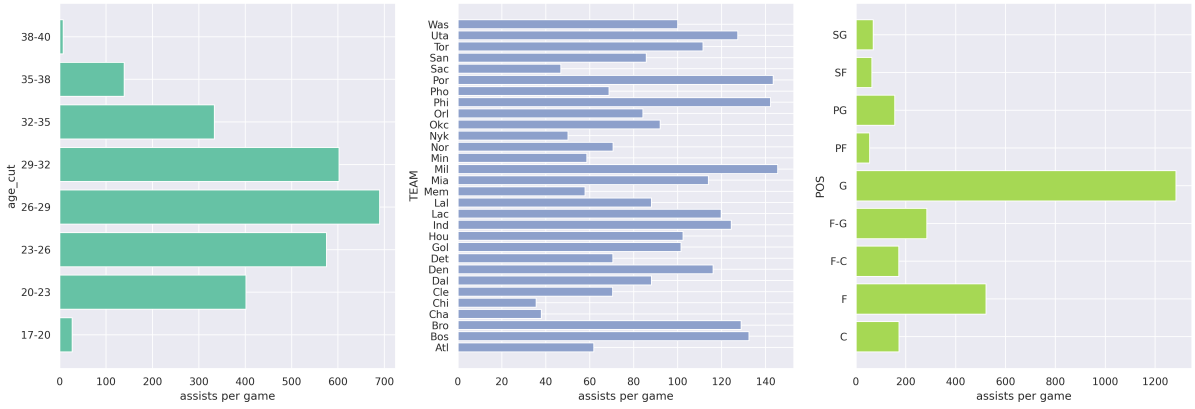
To know exactly which athletes scored the most and gave the most assists simultaneously, we have the graph on the next page. The athlete with the most points is Donovan Mitchell, and the athlete with the most assists and the second most points is Luka Doncic. There are other athletes who have more assistance, but do not add more than 25 points per game, which do not appear in the table, as is the case of Rajon Rondo. Also, some players appear more than once because in our data we have several seasons in which the same player may have participated.



Source: Author

As we've already plotted the assists before, we can see similarly to the previous graphs, which are the age groups, teams and positions of the players who perform the most assists.

Assists Per Game by age group, Team and Position

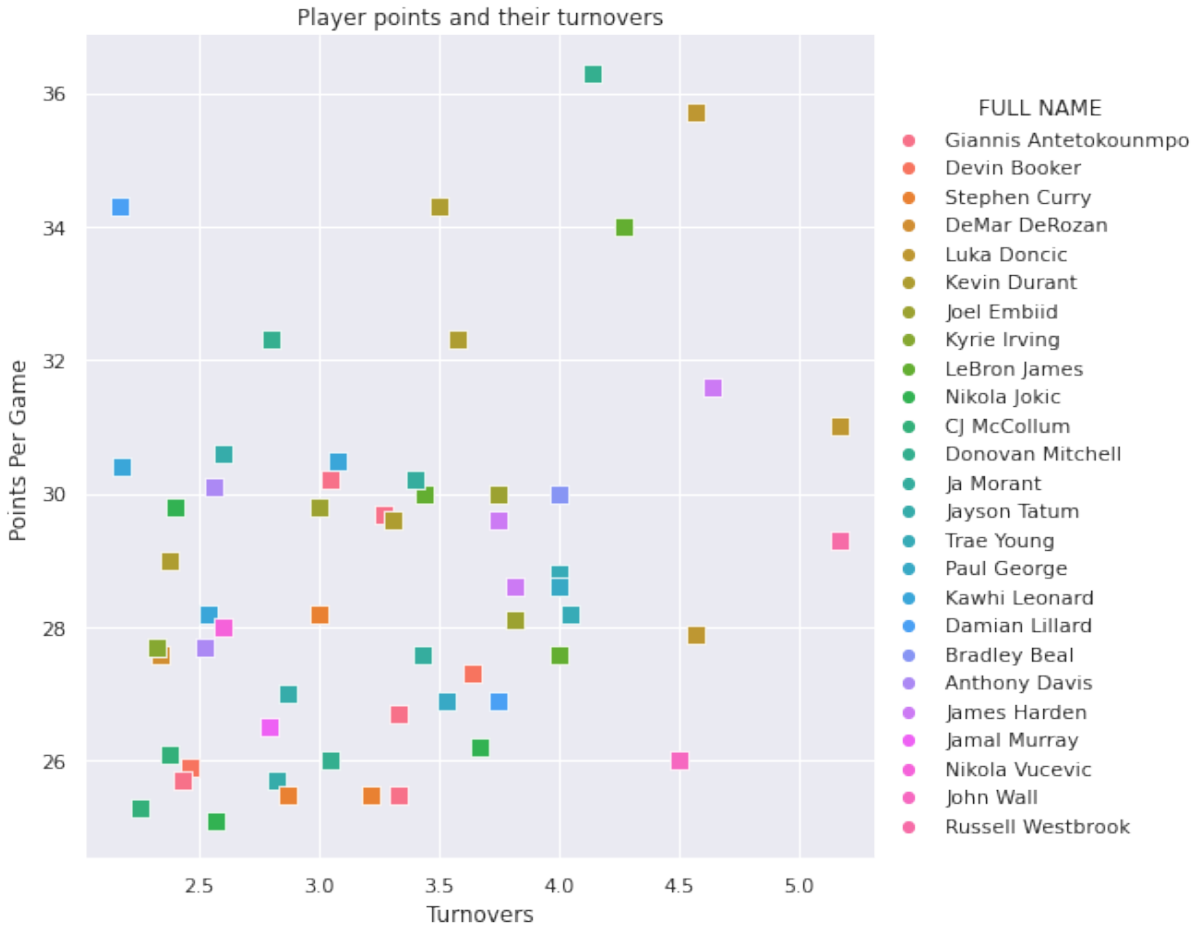


Source: Author

Again, the players with the most assists are 26 to 29 years old, and the least assists are 38 to 40 years old. Similar to the previous chart, the team that has the most assists is the

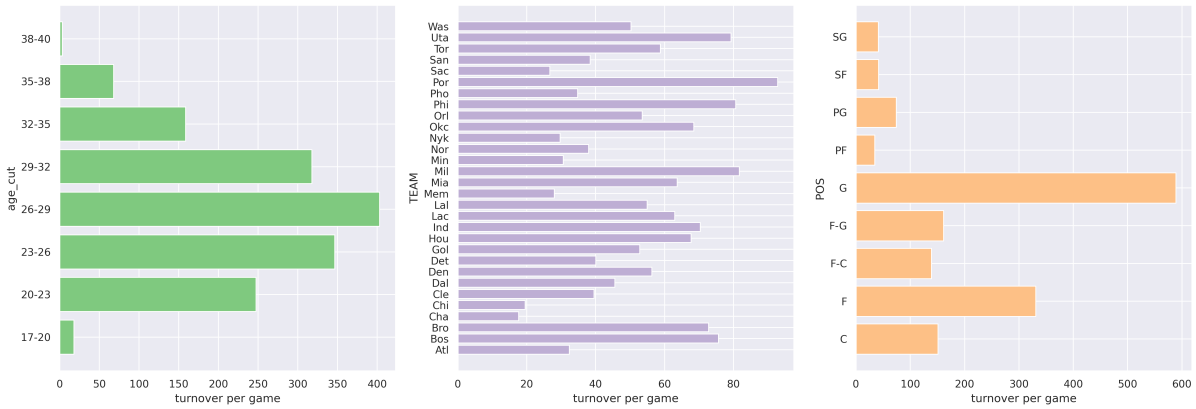
Milwaukee Bucks (MIL), the one that does the least is the Chicago Bulls (Chi). Finally, the position that most assists is guard (G), and the one that least assists is pivot (PF).

One relevant thing is the TOPG attribute, which is the number of turnovers per game. In basketball, turnover occurs when a team loses possession of the ball to the opposing team before a player tries to hit his team's basket, increasing the chance of the team that performed the turnover to score a point, hence its importance. To find out who are the players with the most turnover and points per game, let's retrace the graph that looks like a scatterplot.



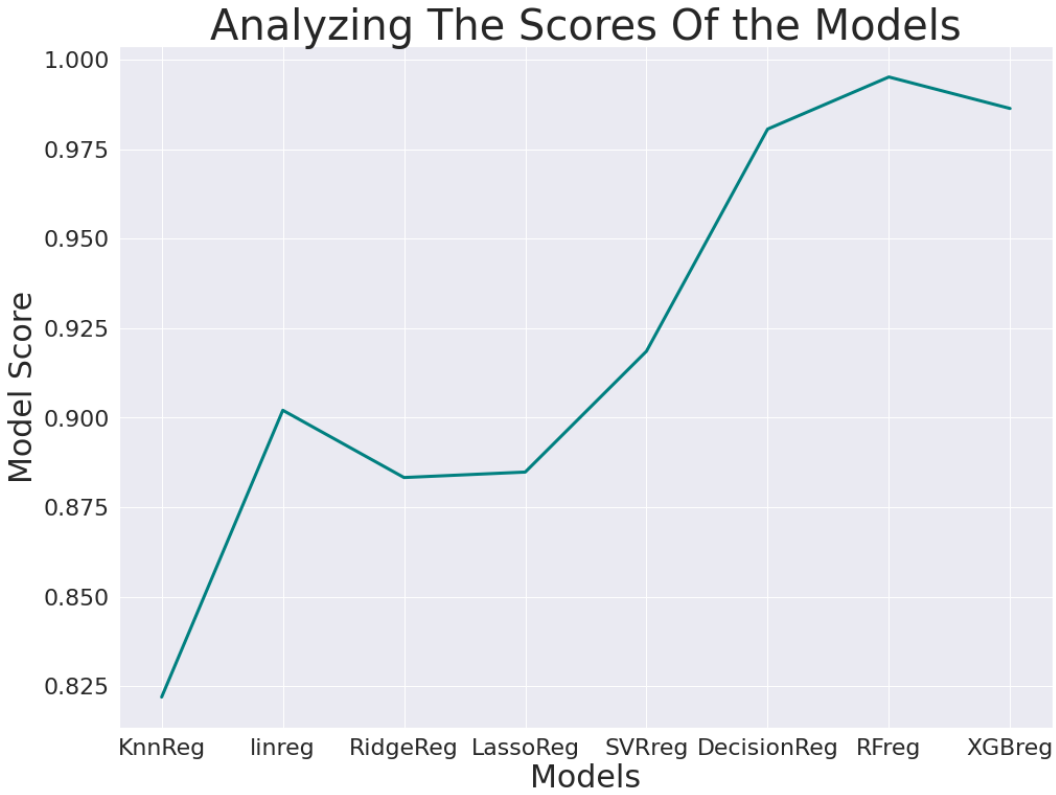
Source: Author

The players who scored the most points we already know, but the one who simultaneously made more turnover and more points on this graph was Luka Doncic. Finally, we would like to know which age groups, teams and positions of players perform the most turnovers.



Source: Author

In the first graph we have the same result obtained previously, in the second graph we find that the team that performed the most turnover was the Portland Trail Blazers (Por), and the team that performed the most turnover was the Charlotte Hornets (Cha). In the last graph, the position with the highest turnover was Guard (G), and the one with the least turnover was the power forward (PF). Finally, we tested several machine learning models: It seems that the best models are Decision Trees, Random Forest Regressor and



Source: Author

Extreme Gradient Boosting for regression, but the best was Extreme Gradient Boosting with 98% score, I chose this one and not the one with 99% efficiency because this model can overfit .

4 Conclusion

In the course of this project, we developed relevant concepts in science and data analysis, in order to suggest the solution to know how many points a basketball player would score without even entering the court, or by data collected during training, therefore, such a project can be used in competitions and betting.

We went through several steps precisely to extract as much information as possible from our dataset. As a result, we found that the individuals who score the most are those between 26-29 years old, the team that scores the most was the Portland Trail Blazers (POR) and the positions that score the most are the Guard (G). Furthermore, we found that the athlete with the most points is Donovan Mitchell, and the athlete with the most assists and the second most points is Luka Doncic. In the later results for assists some results were repeated, however we found that the team with the most assists is the Milwaukee Bucks (MIL). In addition to these results, we plotted a graph of points per turnover and we understand that the player who made the most turnover and points simultaneously was Luka Doncic, so we plotted a graph to know some turniver patterns. Therefore, we found many similar patterns for turnover as previously found, in addition, the team that performed the most turnover was the Portland Trail Blazers (Por), the other results were similar.

In addition, we employ several techniques that can be verified in collaborative google in order to further improve our model and increase its efficiency, so using the extreme gradient boost model we found the score in the training data of 98%, and the test with 97% of score. Finally, we implemented our online model and made it available for anyone, athlete, competition to use. For all these caveats, we understand that one way to improve our model in the future would be to use a much larger dataset and allow the model to be saved and updated to learn more and more as new data is entered into the algorithm.

Bibliography

GéRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2ª. ed. GCanada: O'Reilly, 2019. Citado na página 3.

MüLLER, A. C.; GUID, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 1ª. ed. United States of America: O'Reilly, 2016. Citado na página 3.