Final project

Zhiqian Chen, Yi Zeng, Qihang Liang

5/11/2021

Abstract

The purpose of this research is to study the main factors that affect NBA players' scores. Since the random forest model has the lowest root mean square error, we establish the random forest model to measure the most important factors that affect the scoring level of basketball players, such as age, playing time, height and weight. The results of the prediction model show that NBA players' scores depend largely on the number of times the players play and the playing time, but the accuracy of the model is only about 70%. Factors that are not analyzed in this article but may affect players' scores include home and away games, whether players are injured, and team rankings, etc.

Introduction

Basketball is one of the most popular sports in the world. At the same time, basketball is a popular ball game in people's lives. There are many famous basketball teams in the world, and every basketball team has a large number of fans. Powerful teams and powerful players will represent the country in the Olympics and win trophies for the country. In addition, almost every country has professional basketball leagues, such as the NBA, CBA, and Euro league. These leagues will inspire players to continuously improve and win glory for the country. Moreover, almost every wonderful basketball game has attracted the attention and praise of people from all over the world.

In a basketball game, there are two-pointers scored, three-pointers scored, and each player's total score. The total score is one of the important conditions for evaluating players. Therefore, we are very interested in the scores of the players of each basketball team. We want to find out what factors most affect the scores of the players. Since each player will have a different performance in different basketball games, we will predict the player's total score in each season. Therefore, the question in this report is to build the best predictive model possible for each player's score in each season.

Data Selection Our data comes from Kaggle (https://www.kaggle.com/jacobbaruch/basketball). The data contains 34 variables, and 53,949 observations. Including season, Player, team, GP (# of Game played), Min (# of Minutes Played), FGM (# of Field Goals Made), FGA (# of Field Goals Attempts), 3PM (# of Three Points Made), 3PA (# of Three Points Attempts), FTM (# of Free Throws Made), FTA (# of Free Throws Attempts), PF (# of Personal Fouls), ORB (# of Offensive Rebounds), DRB (# of Defensive Rebounds), REB (# of Rebounds), AST (# of Assists), STL (#of Steals), BLK (# of Blocks), PTS (# of total Points), etc. In addition, we think game played and minuted play are the important variables, which can have the large effect on player's performance.

Methods (Model)

We collect a data from Kaggle.com that about each statistical about the players. As we mentioned before, there are 53,949 observations, In this data set, but we find that there are many blank observations, hence after collect the csv. data set, we first drop the missing value. After we dropping the missing value, there

are only 10,136 observations left. And there is not variable about players' age. Hence we need to create a variable about age. In the data set, we notice that there are two time variables, the first one is season and the another is player's birth year. Also, the season is a range, hence we create a new variable year to indicate which year is in that season, and then we use the variable "year" and "birth year" to generate players' age.

Model Selection

First, we choose our best model to predict the PTS (points). For the model's choice, we consider three option: 1. linear model 2. boosting model and 3. random forest model. We can directly determine which model is best, hence we build all three model and compare their RMSE, then we find that random forest model has the lowest RMSE, so then we use the random forest model to predict the PTS. After prediction, we make three plots. The first plot is the plot of the original data, using a color scale to show points versus GP(Gamed Played) and MIN(Minuted Played). The second plot is the random forest model's prediction of PTS versus GP and MIN and the last plot is the random forest model's residual versus GP and MIN. In our model, there are many variables, like age and height. The reason that we plot the points versus gamed play and minuted play is that in the model we find points is more relevant to game play and minute play.

Below are the RMSE of three model. The first one is RMSE of linear regression model, second one is RMSE of random forest model and the last one is the RMSE of boosting model. We compared these three models with RMSE. Hence, we can find that random forest model has the lowest RMSE. Therefore, we decide to use the random forest model as our best predictive model.

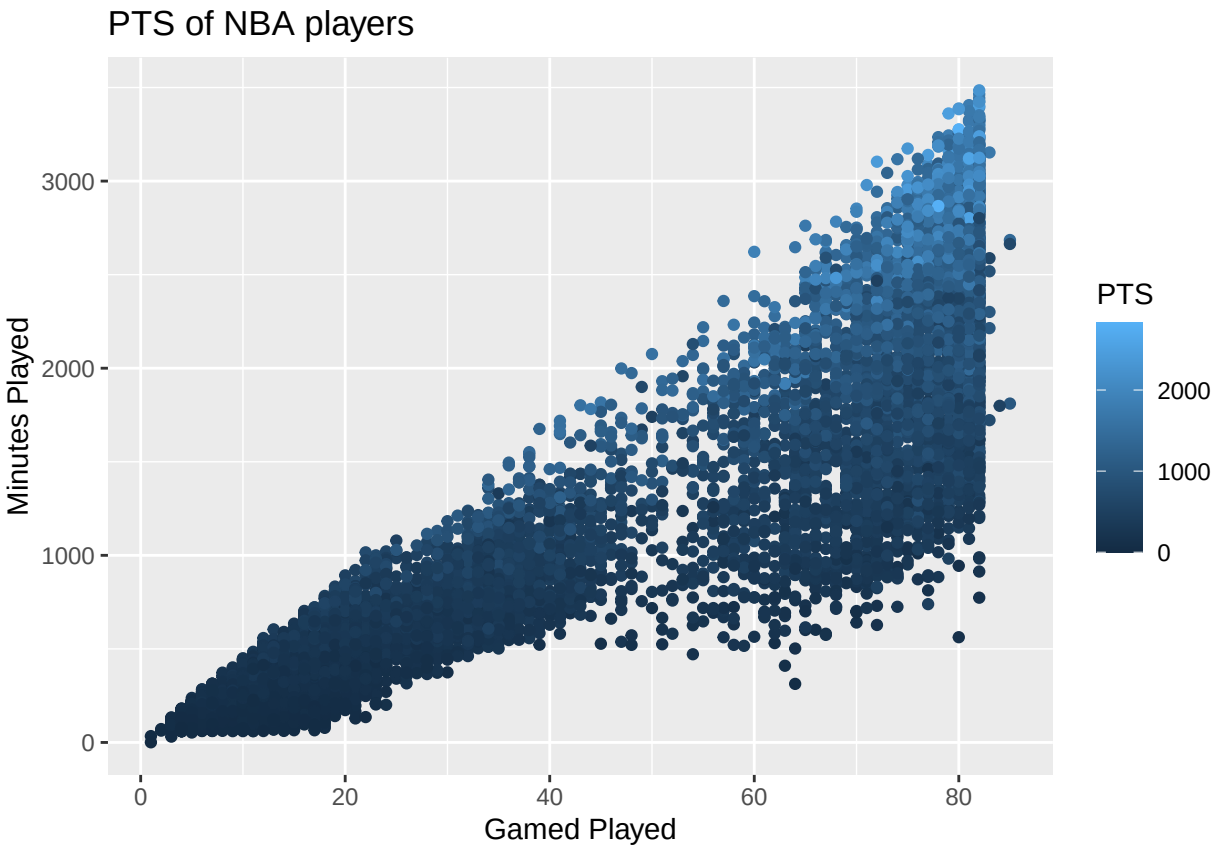
```
##               Model     RMSE
## 1 linear regression 160.4533
## 2     random forest 151.2717
## 3          Boosting 167.8842
```

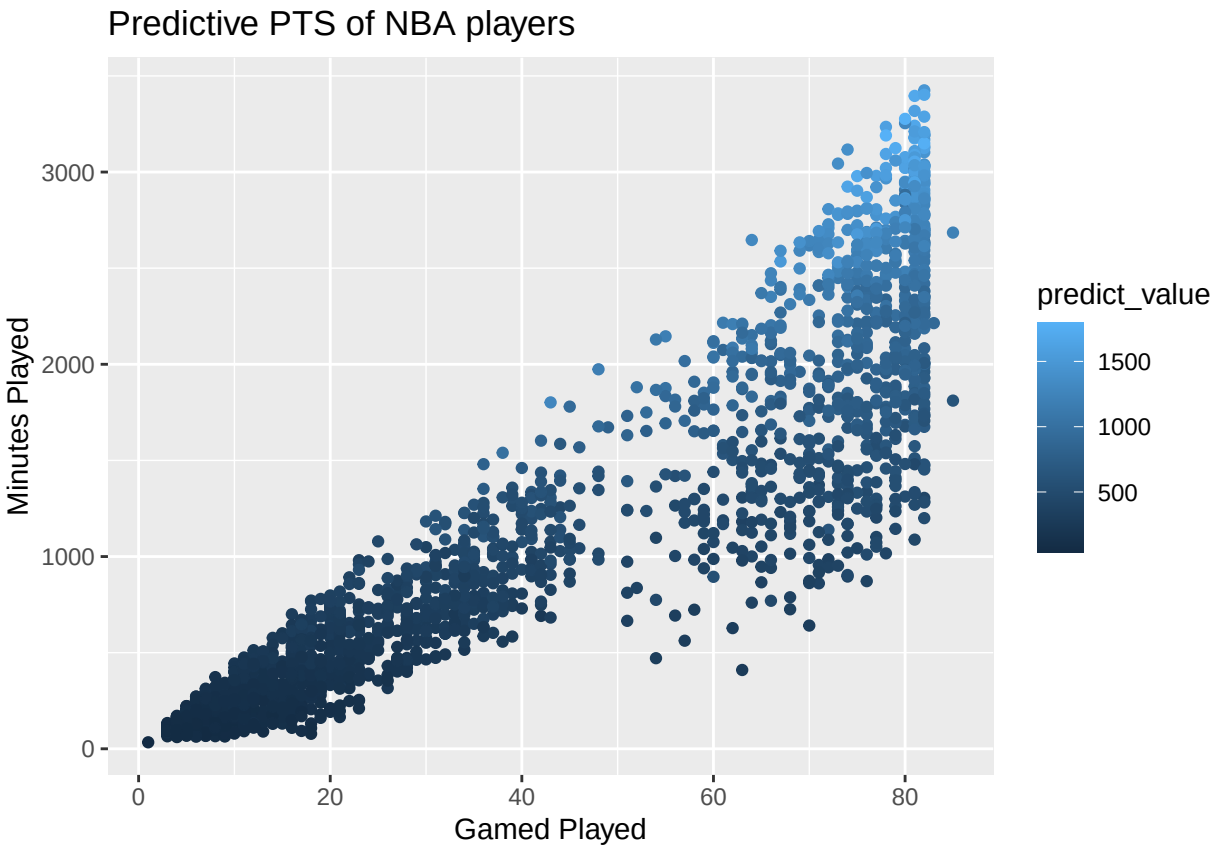
Result

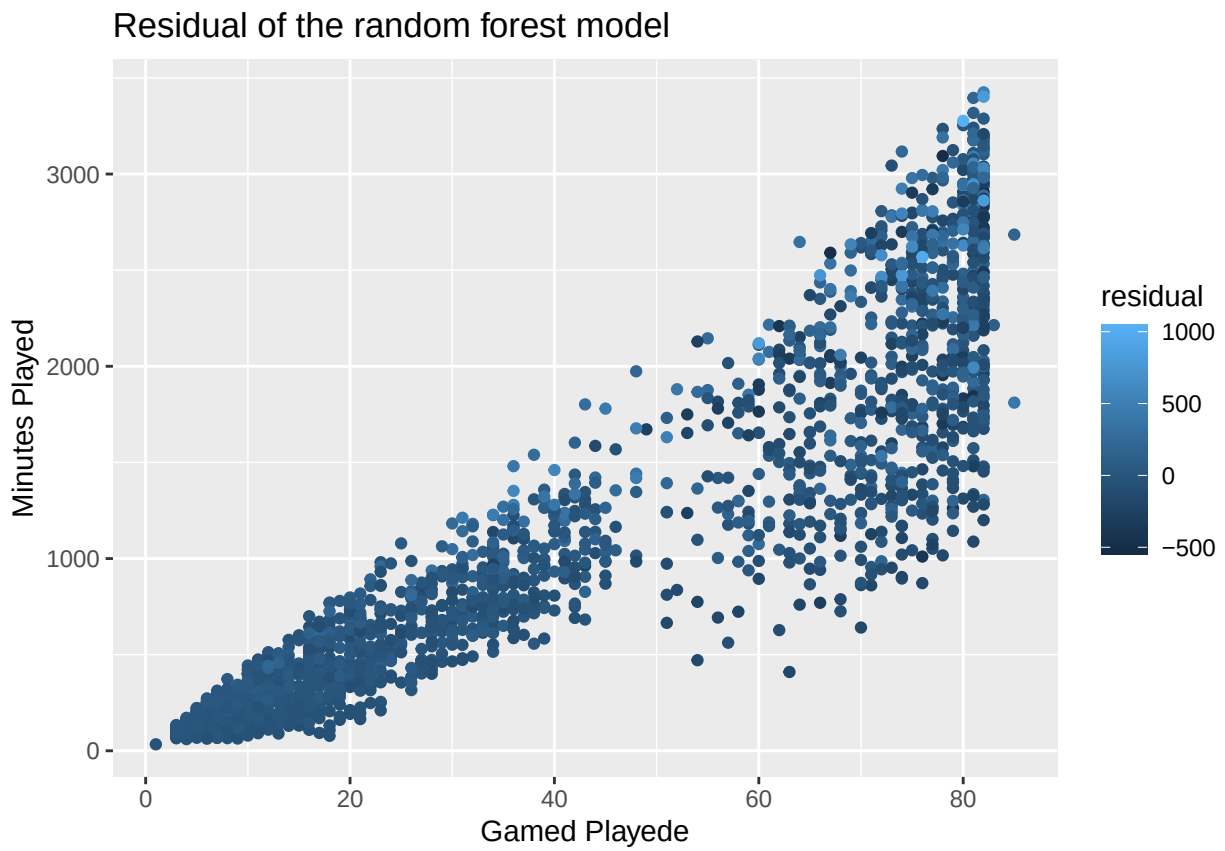
Below is the accuracy of the random forest model, the accuracy is nearly 0.7 which means that our predict value of the points meet 70% of the actually value of the points.

```
## [1] 0.6826539
```

The first plot is the plot of the original data, using a color scale to show points versus GP(Gamed Played) and MIN(Minuted Played). The second plot is the random forest model's prediction of PTS versus GP and MIN and the last plot is the random forest model's residual versus GP and MIN. In the first plot, we can find that in the area of game played below 20 games and minutes played below 1000 minutes, the points are at a very low level. Also, even if some players played many game, but if they have low minutes played, they still get a low points. For the players who played many games and minutes, they get a high level of points. Since we use the train set to predict, hence the second plot is have a better look than the first plot, it is not as density as the first plot. But the result are the similar, players with low game played and minutes played have low level of points, players with high game played and miutes played have high level of points.


PTS of NBA players
Minutes Played
3000
2000
1000
0
0
20
40
60
80
Gamed Played
PTS
2000
1000
0


Predictive PTS of NBA players
Minutes Played
3000
2000
1000
0
0
20
40
60
80
Gamed Played
predict_value
1500
1000
500



Conclusion

Through our model, we found that a player's score largely depends on the number of games attended and the length of time played. Age and physical fitness also affect scoring, but the impact on scoring is not as great as the number of games attended and the length of time played. Therefore, when we predict a player's score, we should first consider whether the player will start, usually the starting player gets more playing time. At the same time, we have to consider the playing time of the players.

In general, we chose the best model we think, but the accuracy of our model is still only about 70%, which shows that there are still other factors that affect the player's score. Our analysis believes that in addition to the variables mentioned in the model, the factors that affect the player's score may also include: 1.Home and away games. 2.Whether the player is injured. 3.The ranking of the team. The first and second factors are very easy to explain. When players are playing at home, they are encouraged by home players to increase their points. The second factor is obvious. The condition of injured players will decline, leading to a decline in scoring. The third factor is very complicated. The relationship between the ranking of the team and the score of the player is not great, but it cannot be ignored. The rule of the NBA is that the lower the ranking team has a higher chance of picking the No. 1 pick in the draft, so some In order to get a high-quality rookie, the team will choose to be bad, so it will affect the players' score.