The impact of draft round on the player’s overall rating

Abstract

This report contains a brief analysis about how the draft round, player’s height and weight, averaged points player scored per game, and other factors affect the overall rating of the player. Mainly focus on the effect of draft round on the rating, after performing propensity score and linear regression for the final model, I found that draft round and other factors have positive influence on player’s rating, while the height has no significant effect on the rating.

Key words

Propensity score, linear regression

Introduction

Basketball is one of the most popular and intense sports in the world. Specifically, NBA represents the highest level of professional basketball league and they would record all the relevant statistics for each player in every season. Then sports media and websites will perform some further analysis for players based on those information. It is worth to mention that a few websites or video games such as 2k sports, publish the ratings in the beginning of the season based on previous year’s data.Finally, they would make a conclusion about the rating for each player, usually scale from 0 to 99. Rating represents an overall power and basketball skill for each NBA player. This report will be focused on identifying the important player’s statistics and analyzing how those factors would affect player’s ratings, especially I am trying to find out the effect of draft round on the ratings of players.

Data wrangling

I obtain the two data sets from Kaggle website, they provide some basic body measurement (e.g.height,weight) and seasonal statistics about each NBA player from 1996-2019. Specifically, I focus on the analysis of 2018-19 season so I make an updated data set called “updated_season” only containing the statistics from this season. In addition, in order to perform the modeling and further analysis, I convert draft round, country and college to dummy variables. Then I add the player’s rating from the second data set to my updated data set. Because some player’s have missing value of statistics, I eliminate those player from the data set.

Now take a glimpse of the updated dataset

Glimpse of updated dataset

player_name	age	player_height	player_weight	college	country	draft_year	draft_round	draft_number	gp	pts	reb	ast	net_rating	oreb_pct
Goran Dragic	33	190.50	86.18248	0	0	2008	0	45	36	13.7	3.1	4.8	1.3	0.019
Glenn Robinson III	25	198.12	100.69742	1	1	2014	0	40	47	4.2	1.5	0.4	-0.8	0.026
Gerald Green	33	200.66	92.98636	0	1	2005	1	18	73	9.2	2.5	0.5	4.6	0.019
Georges Niang	26	203.20	104.32616	1	1	2016	0	50	59	4.0	1.5	0.6	-7.7	0.022
Gordon Hayward	29	203.20	102.05820	1	1	2010	1	9	72	11.5	4.5	3.4	5.1	0.026
George Hill	33	190.50	85.27530	0	1	2008	1	26	60	7.6	2.5	2.3	5.6	0.028

Variables Explanation

Player’s Rating (corresponding variable name: rating) is a discrete numeric variable, scaling from 0-99. It represents an overall power and basketball ability for each player. There are only a few player received an honor of rating 99.

Draft Round Reference

Draft round	Numeric Level
Second round & Undrafted	0
First round	1

The round of player being drafted (variable name: draft_round) is originally a categorical variable, it consists of three categories, player was drafted in first round, second round, or undrafted. However, we need to use it to determine the its effect on player’s rating, so it is reasonable to convert it to dummy variable, use 1 to represent the players drafted in the first round, use 0 to represent those drafted in the second round or undrafted.

Each round has 30 spots, and usually high-skilled players will be drafted in the first round.

College Reference

College	Numeric Level
No college experience	0
College experience	1

The college player attended (variable name: college) is a categorical variable and I change it to dummy variable for further analysis, 0 means the player did not go to college, and 1 means he did attend college.

Player Homeland Reference

Country	Numeric Level
Not from USA	0
USA	1

The country that player comes from (variable name: country) is a categorical variable and I change it to dummy variable for further analysis, 0 means the player did not come from the America, and 1 means he is from America. Height and weight of the player (variable name: player_height, player_weight) are measured in centimeters and in kilograms respectively. They are both continuous variables. The number of game each player play in the season (variable name: gp) is a discrete variable and the maximum number is 82 games. The average points per game that each player score, average number of rebounds player get and average number of assists they give to teammates (variable name:pts, reb, ast) are all continuous variables. Net rating of each player (variable name: net_rating) is a continuous variable, representing the difference in team’s point per 100 possessions when the player is on the court, here we could simply treat it as how much the team would get better or worse when the player is on the court. Positive net rating means it is better for the team, negative means worse. Percentage of team plays used by the player while he is on the floor, the measure of the player’s shooting efficiency that takes into account free throws, 2 and 3 point shots are continuous variables, their variable name are “usg_pct” and “ts_pct” respectively.

Potential drawback

According to the original owner’s note for the data set, he manually filled the missing value by using data from Basketball Reference website, so there might be some typos or incorrect values during the data transformation. This might affect the accuracy of modeling analysis.

Apply propensity score matching

Now I perform the method of propensity score matching to the probability of getting into first round and the treatment is the first draft round. Firstly, it is to build a logistic regression model, with a binary outcome variable draft round. I choose age, the height and weight of players, their college and homeland because those are the relative information available before the NBA draft, other statistics such as average points scored, are recorded after they enter the league, so these cannot be used to predict draft round. The logistic model would be as following:
$log(Y_{draft\ round}) = \beta_0 + \beta_1 \cdot X_{age} + \beta_2 \cdot X_{player's\ height} + \beta_3 \cdot X_{player's\ weight} + \beta_4 \cdot X_{country} + \beta_5 \cdot X_{college}$

1. I check the Variance Inflation Factor(VIF) for each predictor variable, and all the values are less than 5, so it is likely to conclude that there is no multicollinearity among those variables. 2) Also 339 observation is considered as large sample size. 3) The outcome variable draft round is binary because it only takes value zero or one, zero means the player was drafted in the second round or undrafted, one means he was drafted in the first round. 4) We assume the predictor variables are linearly related to the log odds. All the assumptions are satisfied so this is a valid model for logistic regression.

Next I create the matches by using my prediction. Based on the similar propensity score, I want the untreated players (those were not drafted in the first round) to match with the treated players (who were drafted in the first round). Using the match function, I discover that 274 players are matched and 65 players are unmatched.

Next I reduce the dataset to those which are matched

```
##        player_name age player_height player_weight college country draft_year
## 1      Tim Frazier  28        185.42      77.11064       0       1  Undrafted
## 2  Brandon Goodwin  23        187.96      81.64656       0       1  Undrafted
## 3 Patrick Beverley  30        185.42      83.91452       0       1       2009
## 4   Jemerrio Jones  24        195.58      78.92501       0       1  Undrafted
## 5      George Hill  33        190.50      85.27530       0       1       2008
## 6  Anfernee Simons  20        193.04      83.91452       0       1       2018
##   draft_round draft_number gp pts reb ast net_rating oreb_pct dreb_pct usg_pct
## 1           0    Undrafted 59 5.3 2.8 4.2       -2.7    0.035    0.106   0.132
## 2           0    Undrafted 16 1.4 0.2 0.9        8.0    0.016    0.038   0.220
## 3           0           42 78 7.6 5.0 3.8        4.0    0.035    0.135   0.120
## 4           0    Undrafted  6 4.5 8.2 2.2       10.7    0.100    0.198   0.108
## 5           1           26 60 7.6 2.5 2.3        5.6    0.028    0.075   0.147
## 6           1           24 20 3.8 0.7 0.7      -17.6    0.022    0.067   0.240
##   ts_pct ast_pct  season rating   .fitted cnts
## 1  0.544   0.286 2018-19     72 0.2740611    1
## 2  0.413   0.298 2018-19     68 0.2887023    1
## 3  0.561   0.183 2018-19     79 0.2892286    1
## 4  0.398   0.115 2018-19     70 0.2992986    1
## 5  0.554   0.146 2018-19     75 0.3038184    1
## 6  0.535   0.135 2018-19     72 0.3045155    1
```

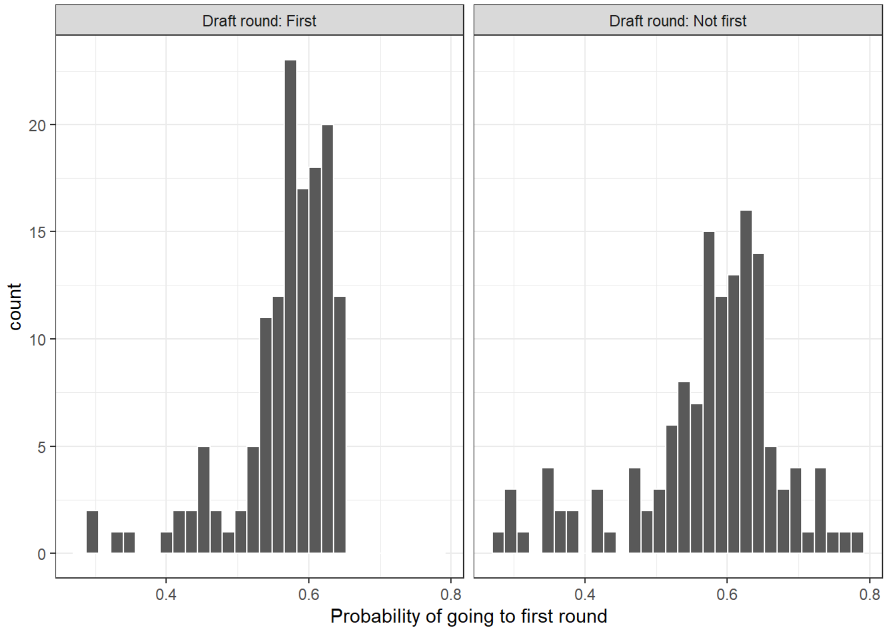
I remove those who are unmatched and keep those 274 players who are matched, so 137 players are being treated. The reason why I use propensity score matching is to eliminate the effect of confounding factors. When the treatment is not randomly assign to the groups, propensity score matching is able to balance the observed factors for two groups and then receive the estimates of treatment effect with smaller bias.

Draw the distribution of two groups.

```
library(tidyverse)

labs <- paste("Draft round:", c("First", "Not first"))
updated_season_matched %>%
  mutate(draft_round = ifelse(draft_round == 1, labs[1], labs[2])) %>%
  ggplot(aes(x = updated_season_matched$.fitted)) +
  geom_histogram(color = "white") +
  facet_wrap(~draft_round) +
  xlab("Probability of going to first round") +
  theme_bw()
```

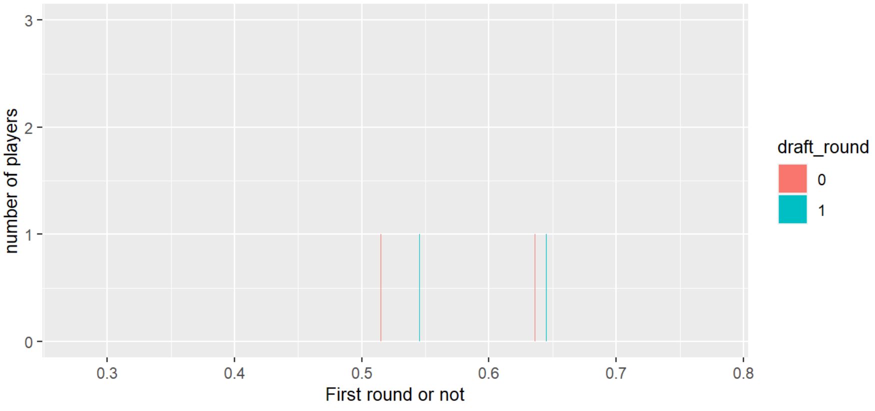
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(updated_season_matched, aes(x=.fitted, fill=draft_round)) + geom_bar() +
  labs(x="First round or not", y="number of players") +
  ggtitle("Figure: The distribution of player being drafted in first round") +
  theme(title = element_text(size=10),aspect.ratio = 0.5)
```

```
## Warning: position_stack requires non-overlapping x intervals
```

Figure: The distribution of player being drafted in first round



The two histograms above show the the number of players with different probability of being drafted in the first round in the control and treatment group. Two plots are not similar due to the inaccuracy of matching.

Now we perform the regression.

Modeling

Now I perform the linear regression to find the effect of draft round and other relative factors on the player’s rating. I use both backward AIC and BIC for the variable selection, as a result, the backward AIC produces a model with smaller value of AIC, which indicates it is a better fitted model.

```
#huxtable::huxreg(PSM_reg)
#kable(rating_backBIC$coefficients)
kable(rating_backAIC$coefficients)
```

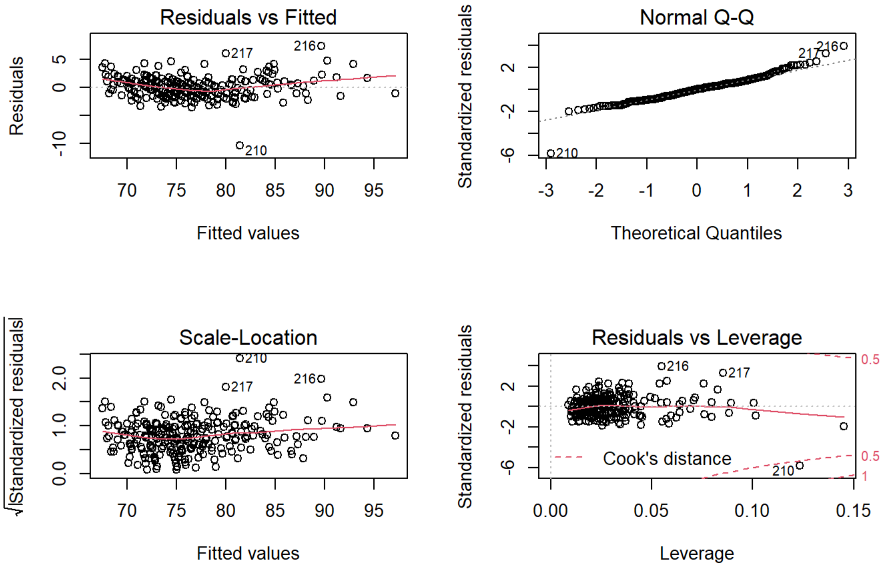
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	61.0385031	1.4755337	41.367068	0.0000000
player_weight	0.0598812	0.0153681	3.896466	0.0001236
draft_round1	0.7222513	0.2530883	2.853752	0.0046605
gp	0.0167878	0.0064185	2.615521	0.0094179
pts	0.5914879	0.0309738	19.096417	0.0000000
reb	0.2972685	0.0746146	3.984051	0.0000875
ast	0.5368999	0.0853034	6.294005	0.0000000
net_rating	0.1236370	0.0194113	6.369339	0.0000000

Result

The final model is
$Y_{rating} = 61.0385031 + 0.0598812 \cdot X_{player's\ weight} + 0.7222513 \cdot X_{draft\ round} + 0.0167878 \cdot X_{gp} + 0.5914879 \cdot X_{pts} + 0.2972685 \cdot X_{reb} + 0.5368999 \cdot X_{ast} + 0.1236370 \cdot X_{net\ rating}$

According to the regression summary table above, the p-value for all the independent variables are smaller than the benchmark 5%, so the coefficient estimates are all statistically significant. All the coefficients are positive, which means they have proportional positive effects on player’s rating. Specifically, if a player was drafted in the first round instead of second round or undrafted, his overall rating will increase by 0.72 units. The average pointed scored and average assisted distributed also have large impact on the rating, the increase in one point or one assist will result in 0.59 or 0.54 units increment respectively in player’s overall power. The net rating of the player and averaged rebound grabbed also have moderate level of impact on player’s overall power, while the number of game played has the least effect, with approximately 0.017 units rise in player’s rating if he plays for one more game.

Discussions and Limitations



Based on the plot of residual vs fitted value, there is no patter found in the graph, which means the model is well-fitted and the assumption of constant variance is satisfied. Similar result could be observed in standardized residuals vs fitted values plot. From the normal QQ plot, it is obvious that normal error MLR assumptions are being satisfied because most values are fitted perfectly on the line.

One of the major limitations is the multicollinearity among independent variables. For example, the correlation between average assist distributed and average point scored is 0.69, which means they are highly correlated with each other. As a result, the coefficient estimation would be less accurate and it reduces the statistical power of the regression model. Another limitation is that the outliers should be removed because they will rise variability of the data, then diminish the statistical power of the model and also weaken the significance of estimated coefficient.

Reference

https://www.kaggle.com/justinas/nba-players-data (https://www.kaggle.com/justinas/nba-players-data) https://www.kaggle.com/isaienkov/nba2k20-player-dataset (https://www.kaggle.com/isaienkov/nba2k20-player-dataset)