Data mining - NBA Analysis

Nikola Tomažin

18. siječnja 2021.

1 Dataset

For my tasks of analysing the NBA data I chose 3 datasets:

- The *Season stats* dataset which is a pretty big dataset and has almost all statistics for a player and it had some information i thought might be usefull for analysis

Rk	Year	Player	Pos	Age	Tm	G	GS	MP	PER	TS%	3PAr	FTr	ORB%	DRB%
24057	2016	Dwyane Wade	SG	34	MIA	74	73	2258	20.3	0.517	0.037	0.343	4.3	10.7

TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	PTS
7.6	27.4	1.8	1.4	13	31.6	2.4	2.6	4.9	0.105	1409

- The *All NBA 1984-2018* dataset which consists of players chosen for the All NBA teams (honor bestowed on the best players in the league following every NBA season) and their statistics for that season

Rk	Player	Season	Age	Tm	Lg	WS	G	GS	MP	FG	FGA	2P	2PA	3P
1	Michael Jordan	1987-88	24	CHI	NBA	21.2	82	82	3311	1069	1998	1062	1945	7

3PA	FT	FTA	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	FG%	2P%
53	723	860	139	310	449	485	259	131	252	270	2868	0.535	0.546

3P%	eFG%	FT%	TS%
0.132	0.537	0.841	0.603

- the *NBA season 17/18 salaries* which consists of all players in the 17/18 season and their salaries

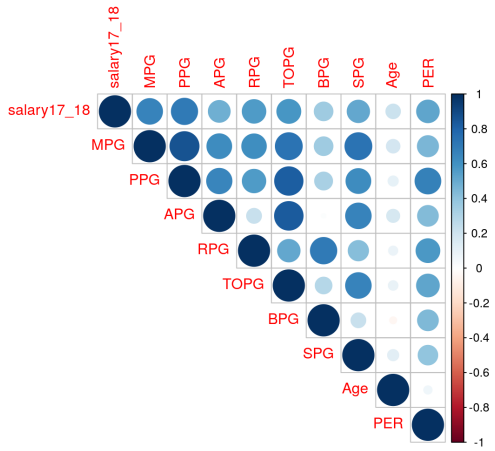
Rk	Player	Tm	season17/18
1	Stephen Curry	GSW	34682550

2 Preprocessing

Since the data from the datasets helds a lot of information and some statistics weren't counted in the first years of the NBA, I had to be careful which data I need. The salary dataset was the only one who didn't have too much information, so I didn't have to clean it, but the other two were extensive. For starters, I extracted only the columns I need and know they meant and afterwards, removed all rows where there was data missing. For example, before 1980s they didn't monitor the 3 point attempts and goals as well as made blocks or turnovers.

Depending on the task i had to combine different datasets (tables), so for example for salary analysis i had to combine the player statistics with their salary.

I also had a problem that some players changed teams in the middle of the season so I had redundant information.



Slika 1: Correlation plot for some important statistics

The interesting part of this is that the number of turnover players make is linked to their salary, and the relationship has a positive correlation. So, I interpreted this relationship like this: "the more turnovers they make" means that they are more involved in ball movements in games, which means that players who make turnovers are, at some extend, important to their team.

There exists a fairly positive correlation between all the performance features with some being highly correlated as seen in 1. Furthermore, we wanted to look into those features that will contribute the most to our analysis. In order to do so, we use the principal component analysis (PCA) technique to reduce the dimensionality. The output can be seen in 2. 82.4% of the variation in the data is explained by PC1 and PC2.

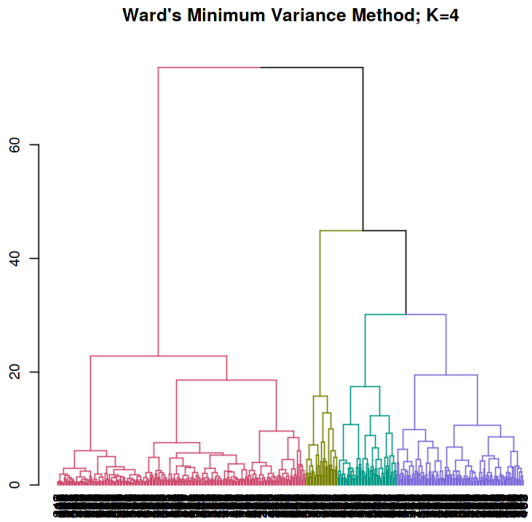
```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     3.0767 1.4379 0.88672 0.80819 0.63364
Proportion of Variance 0.6762 0.1477 0.05616 0.04666 0.02868
Cumulative Proportion  0.6762 0.8239 0.88002 0.92667 0.95535
                           PC6     PC7     PC8     PC9    PC10
Standard deviation     0.52061 0.46651 0.29741 0.16765 0.10779
Proportion of Variance 0.01936 0.01555 0.00632 0.00201 0.00083
Cumulative Proportion  0.97471 0.99026 0.99658 0.99858 0.99941
                          PC11      PC12      PC13      PC14
Standard deviation     0.09065 1.013e-15 4.903e-16 2.995e-16
Proportion of Variance 0.00059 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.00000 1.000e+00 1.000e+00 1.000e+00
```

Slika 2: PCA output

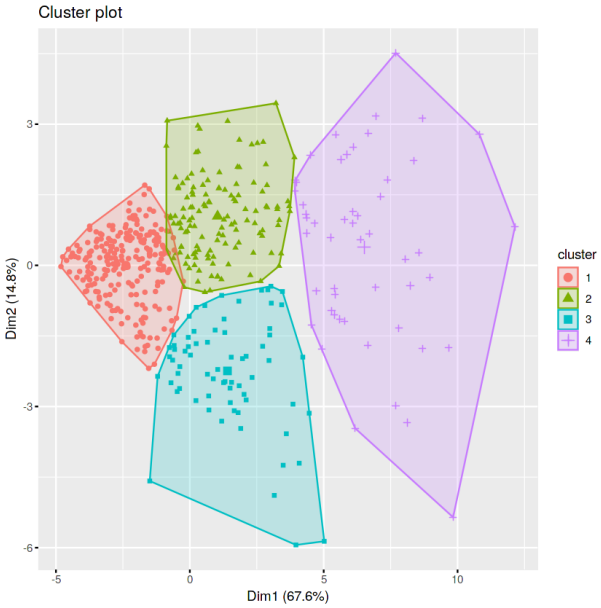
We can see that some variables don't make a difference in the variance, so we can only take the first few principle components.

3 NBA salary clustering

For the clustering I tried to do hierarchical clustering first where I used Ward's minimum-variance method, wherein the clustering criterion is based on the error sum of squares (SSE) and the approach is similar to an analysis of variance problem. Clusters are generated from pairs that yield the smallest SSE. The dendrogram is shown below. I used also the Nearest neighbor method but Ward's method provides a more interpretable result.



Slika 3: Wards minimum variance model for k = 4



Slika 4: K-means cluster plot for k = 4

We segregated the players based on their performance as determined from the k-means clustering into four categories, namely, below average, average, good and excellent (shown as 1, 2, 3 and 4 in the above charts, respectively). We then plotted the player efficiency rating (PER) against their salary within each category as exhibited below. PER is developed by an ESPN columnist and is described as the net rating of a player's accomplishments per minute.

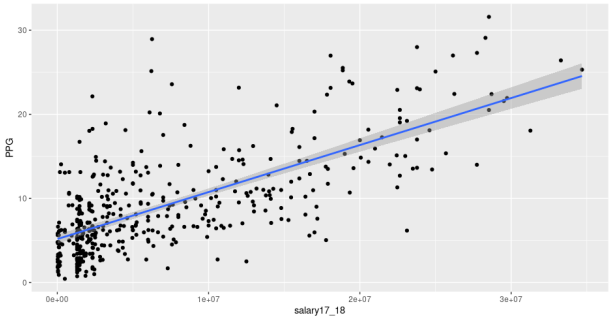
4 NBA salary prediction using regression

As it was seen in 1 the most correlated features with the salary were PPG > MPG > TOPG > RPG > PER > SPG > APG. A simple regression model can be seen in 6 which i used to make the salary predictor. That model only depends on PPG, MPG and TOPG which gave solid results for a specific role, for example shooting guards and point guards, while for example for centers, who mostly rely on rebounds and their defense, it predicted their salary poorly. The model could be improved when there would be a statistic which emphasises defense play more than offense play.

5 Predicting All-NBA teams

A player voted onto the All NBA teams is regarded as one of the 15 top players in the league that season. After data preparation and exploration, I use logistic regression models to estimate the relationships between player statistics (e.g. points, assists, rebounds) and All NBA selection.

Slika 5: PER against their salary



Slika 6: Simple regression model

```
> salary_prediction(model1, 16.7, 31.2, 1.5) # JJ Redick
[1] "PPG:16.7,MPG:31.2,TOPG:1.5 ==> Expected Salary: $13,959,120"
```

Slika 7: Expected salary

5.1 Logistic regression

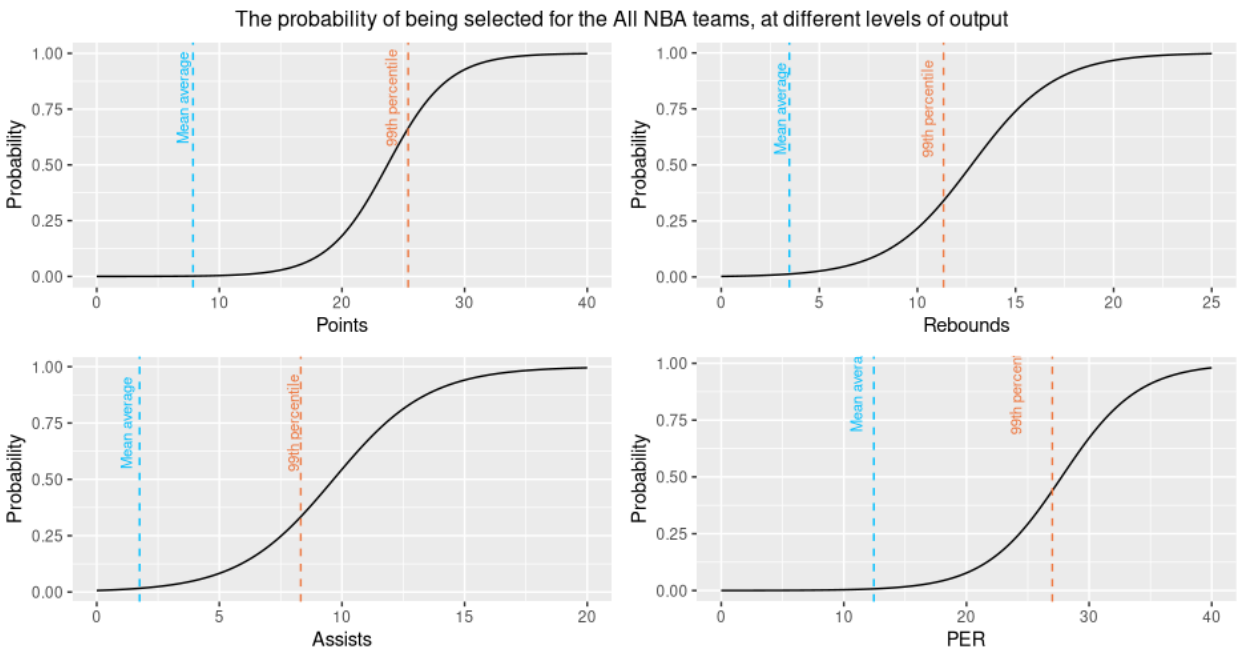
We can see from these graphs that at the mean level of output, players have an almost-zero chance of being selected for the All NBA teams. The vertical lines showing the 99th percentile of output are also telling: we can see that for PER and points, output at this elite level gives players a high (¿70%) chance of selection, this is not the case for assists and rebounds. One hypothesis for this would be that position-level differences are relevant for the extent to which assists and rebounds predict All NBA inclusion, and that a multilevel model with position as a factor would have more explanatory power. As it stands, it is likely that important trends at the position level are obscured by models not sensitive to position.

5.2 Random forest algorithm

Later I trained my random forest algorithm to get better results. I split the data in 2 sets, one to train, one to test it on in a 2:1 ratio. As seen in 9 the algorithm performed close to its batting average (accuracy of 98.2% in testing, versus 98.8% in training). Our algorithm correctly predicted 2674 of a possible 2689 non-selections (99.4%). That's great. However, it was only able to correctly identify 54 of the 90 All NBA selections (60%) which isn't good.

One challenge for the algorithm here is that we haven't build it to be season-sensitive, but selection to the All NBA teams is. The All NBA teams are comprised of 15 players each season, even if the 20th best player the previous season is better than the 12th best player this season. I'm not aware of a way to code this into our algorithm, so we'll have to do it the long way: ask the random forest model to output probabilities of selection - instead of binary responses - and then consider the fifteen players with the highest probabilities in each season as our algorithm's season-specific predictions. That approach gave a result of 76% which is an improvement on our previous 60%.

The problems also was that given the model's tendency to overrate players who are considered defensive liabilities, adding an advanced defensive metric might help. And often in All NBA discussions, team wins



Slika 8: Probability of being selected for the All NBA team

Type	Count
True Positive	55
True Negative	2674
False Positive	14
False Negative	36

Slika 9: Confusion matrix for random forest algorithm

are cited as a factor that journalists considered when they vote. They don't like to reward players on losing teams. Including team winning percentage as a variable could help at the margins.

6 Sources

- https://www.kaggle.com/justinas/nba-players-data
- https://www.kaggle.com/piyush1912/nba-top-players-deep-learning
- https://www.kaggle.com/jonathanbouchet/nba-player-of-the-week-he-s-on-fire
- https://www.r-bloggers.com/2018/03/analyzing-nba-player-data-ii-clustering-players/
- https://rpubs.com/nburke2/636812