Kings analytics project

Chen(Daniel) Xu

2022-09-12

Abstract:

In this report, I will demonstrate how I process and clean the data to make the data collected from different vendor to be usable together. To find the metrics to predict success for players in the most relevant leagues in Europe and highlight the potential players, I use whether if they made into NBA after playing several seasons in Europe as a criteria and use logistic regression model to make prediction. This report consisted of 4 main parts: Abstract, Introduction, Method and Discussion

Introduction:

The analytics staff of the Sacramento Kings needs to provide some recommendations of players currently playing outside of the NBA for the GM to target. The staff wants to know which metrics predict success for players in the most relevant leagues in Europe and specifically which players he should highlight for this season.

Method:

Data processing and cleaning:

The two main difference between the NBA and European data are: different format of player's name and not uniform number of features.

To deal with the first problem, I used the player_lst dataframe as standard and use first letter capitalized for both first name and last name. Doing so, I can also exclude the players who only played for NBA. To deal with the second problem, I found that the NBA data include three more variables than the European data: Plus_minus, calculated_possessions, plays_used. Because the main purpose is to find the players who can be selected from Europe to NBA, I will mainly focus on the European data. Thus, I just get rid of the three variables above.

As I mentioned in the Abstract, I planed to use whether the player made into NBA from Europe as the most criteria to determine if the player success or not. Although my assumption is very subjective, the main purpose of this project is to find European players to watch in the new season and see if they can make it to the NBA. So I think my assumption is reasonable.

I took all the players who made into the NBA from the Euroleague during the 2010-2020 season, and further found the season before they entered the NBA and labeled it as 1. My idea was that because they were successful in the previous season, they were discovered by the scouts to get to the NBA. Any other European season, I put a 0 on it, meaning it wasn't successful enough to make it to the NBA.

I used 2021 season data as the dataframe to be predicted in order to find the players should be highlighted for the coming season.

Modelling

In the modelling part, I mainly used logistic regression model because our outcome would be 0 or 1. 0 represents as unsuccessful and 1 represents as successful.

As another main purpose for this project is to find the metrics that can be used to determine if the player success or not. I tried three different groups of features to fit the model: full features, stepwise selected features and stepwise selected features + mannully added features.

I use AIC as thestandard for my stepwise model.

```
##                      (Intercept)                              age
##                     1.749019e+02                    -3.005857e-01
##                           season                 leagueEuroLeague
##                    -8.443314e-02                    -1.328908e-01
##             leagueItaly - Liga A                leagueSpain - ACB
##                    -6.182052e-01                    -5.463902e-01
##                            games                           starts
##                    -6.157039e-02                    -1.757951e-02
##                          minutes                           points
##                     1.067739e-02                     7.146175e-03
##                  two_points_made             two_points_attempted
##                     1.782243e-02                    -1.714505e-02
##                three_points_made           three_points_attempted
##                    -2.645560e-02                     6.959774e-03
##                 free_throws_made            free_throws_attempted
##                               NA                    -2.516878e-03
##            blocked_shot_attempts               offensive_rebounds
##                    -5.695222e-02                     1.285124e-02
##               defensive_rebounds                          assists
##                     8.155056e-03                     1.609635e-02
##                   screen_assists                        turnovers
##                               NA                    -4.394504e-02
##                           steals                      deflections
##                    -9.249422e-03                               NA
##            loose_balls_recovered                    blocked_shots
##                               NA                    -2.482367e-03
##                   personal_fouls             personal_fouls_drawn
##                     1.812338e-02                     4.214366e-03
##                  offensive_fouls                    charges_drawn
##                               NA                               NA
##                  technical_fouls                   flagrant_fouls
##                               NA                               NA
##                        ejections             points_off_turnovers
##                               NA                               NA
##                  points_in_paint             second_chance_points
##                               NA                               NA
##                fast_break_points                      possessions
##                               NA                    -5.141737e-03
##            estimated_possessions                 team_possessions
##                               NA                     2.820228e-04
##                 usage_percentage         true_shooting_percentage
##                     1.642745e-02                     1.802103e+00
##         three_point_attempt_rate                  free_throw_rate
##                    -9.938896e-01                    -6.901897e-01
## offensive_rebounding_percentage defensive_rebounding_percentage
##                   -8.493174e-02                   -4.532674e-02
##     total_rebounding_percentage                assist_percentage
##                    6.047679e-02                   -1.861264e-02
##                steal_percentage                 block_percentage
##                   -3.963349e-03                    2.087911e-02
##             turnover_percentage        internal_box_plus_minus
##                    2.963488e-02                   -6.023684e-04
```

```
##           (Intercept)                   age                season
##         131.158459738          -0.296639331          -0.062623428
##                 games               minutes       two_points_made
##          -0.048977656           0.012697394           0.042532725
##  two_points_attempted blocked_shot_attempts               assists
##          -0.020685467          -0.054654859           0.009909216
##             turnovers        personal_fouls           possessions
##          -0.037509523           0.013701494          -0.005325987
##   turnover_percentage
##           0.025924341
```

```
## [1] 0.9257576
```

```
## [1] 0.9257576
```

As we can see from the prediction accuracy comparison between two logistic regression model the stepwise regression have selected a reduced number of predictor variables resulting to a final model, which performance was similar to the one of the full model. So, the stepwise selection reduced the complexity of the model without compromising its accuracy.

From the coefficient of Stepwise regression logistic model, I can conclude that the metrics predict success for players in the most relevant leagues in Europe are:

	x
(Intercept)	131.1584597
age	-0.2966393
season	-0.0626234
games	-0.0489777
minutes	0.0126974
two_points_made	0.0425327
two_points_attempted	-0.0206855
blocked_shot_attempts	-0.0546549
assists	0.0099092
turnovers	-0.0375095
personal_fouls	0.0137015
possessions	-0.0053260
turnover_percentage	0.0259243

The next step is to use the stepwise regression model to predict which group of players would be success in 2021 season, thus they need to be highlighted.

Using the stepwise logistic regression to predict, I can conclude that the players staff should highlight are:

x
Hite Bochoridis
Craft Cline
Rodney Jones
Kalinoski Simonovic
Loy Mokoka
Andersson Jenkins
Kelley Waters
Barrett Stackhouse
Simanic Webster
Fortas Sims
Kobe Satterfield
Jabari Brown
Jamel Saybir
Ognjen Ennis

Discussion

As we can see from the metrics we got using stepwise logistic regression model. This model highlighted two points made and two points attempt. In order to be more consistent with the concept and characteristics of modern NBA basketball, I decided to manually add variables related to three-point shooting and inside scoring on the basis of retaining the original variables.

```
## [1] 0.9257576
```

The accuracy remain the same, let's do the prediction:

Using the modified stepwise logistic regression to predict, I can conclude that the players staff should highlight are:

x
Craft Cline
Barrett Stackhouse
Simanic Webster
Jabari Brown
Jamel Saybir

Note: Because of the limited time, I can only apply AIC as the criteria to perform stepwise selection. Just using one standard is not comprehensive enough because the successful rate is only around 8%. The next step, I will try different criteria such as RSS and apply cross validation to the model in order to get the best group of features.

Also, feature engineering such as method for dimension reduction, feature selection are also a try able approach.

Last but not least, thank you so much for letting me put my hands on this interesting project. Through this practice, I also found that I still have a lot of shortcomings. I am looking forward to hearing back from you about the most professional advice.