Predicting NBA Player Salary

Young Jeong
April 20, 2019

Project Overview

I chose to predict player salaries in the NBA out of personal interest. I've always been interested in how salary dynamics in professional sports work and wanted to see if player statistics can help predict the salary for the next season. Finding the right model could mean a marketable product to the NBA front offices, whom are always fighting a battle to acquire the right player at the right price.

With the design in mind, I chose to scrape data from various sports website (all listed in the Data Section). One of the richest basketball data website is Basketball-Reference.com, where I obtained in-season statistics of all the NBA players by season; I used that to accumulate a 3-season statistics for players. I picked 3-season accumulation to avoid having outliers due to injured seasons or benching (some players can be benched due to conflicts with management).

I ran a linear regression model on 19 features that showed strong correlation with the target variable (Salary) and added features using polynomials (degree 2 terms and feature interaction terms). One thing of note is that I transformed the target variable with cube root transformation because the target variable distribution was extremely right-skewed. Then I made some improvements with Ridge Regularization. Cross Validation was used to compare all the models that were generated. Since this was a time-sensitive information, validation set had to be from the same year (i.e. 2017), and the train set from the 3 previous years (i.e. 2016,2015,2014). This somewhat limited the randomness of Cross-validation split, but I did not want to risk leakage of information.

The best model produced was a linear regression model with ridge regularization applied. The alpha value for the ridge regularization was 8.21, the mean absolute error on the training set was $3.04M, and the R^2 score of 0.5619.

The test set was the 2017-18 NBA season and the following season salaries, and using the model produced I made predictions on the salary. With the average of $11M, my prediction error using mean absolute error was about $4.4M.

Tools Used

- Web Scraping: Selenium
- Data Storage: Pandas, pickle
- Data Visualization: seaborn, matplotlib
- Data Analysis: Statsmodel, Scikit-learn
- Presentation: Google Slides

I was glad to have used both statsmodel and scikit-learn; they are both statistical modeling tool that gives different forms of reporting that was useful in doing Exploratory Data Analysis.

Data

The data was obtained by scraping the following websites:

1. Basketball-Reference.com: Player stats by year from 2008-09 seasons to 2017-18 seasons, Rookie lists from 2008-09 seasons to 2017-18 seasons
2. HoopsHype.com: Player salaries from years 2009-10 to 2018-19 seasons
3. Spotrac.com: Free Agent lists from 2011 to 2018

The data were accumulated and then combined to create a single dataset of 2260 total observations. The features used were accumulated are listed in the Appendix.

Conclusions & Improvements

I was glad to have made improvements throughout the data modeling part of the project as the initial R^2 score on the linear regression model was at 0.453 (Cross Validation average). However I think the prediction error indicates that this model is not the best predictor of NBA player salaries. Here are some of the things I would implement as part of the next iteration in order to improve the model:

1. Player Information - Where players sign can be affected by non-playing aspects such as their affinity to a city or family history. I want to find a way to capture some of that and quantify it as a feature or features.
2. Player popularity - I would like to scrape information from twitter or instagram to determine a player's popularity amongst fans. I believe player popularity can be a great predictor of salary, as players are marketable product for teams and with increasing popularity, a player can be rewarded for bringing more merchandise income for teams. Also great players usually have higher exposure and therefore a higher following fanbase.
3. Financial Structure of the NBA - Certain players are bound by what they can sign as part of the NBA Collective Bargaining Agreement. CBA has restrictions on salary terms on both the player and the team's sides. Therefore I want to capture that as well in the model.

I will make these changes in the near future and improve the model as well. By then 2019 information will be

available too (the free agency starts July 1) so I hope to test on the 2019 data.

Appendix: Features Used

Features	Type	Description
G	Int	Number of Games the player played in the past 3 seasons
GS	Int	Number of Games the player started in the past 3 seasons
MP	Float	Minutes played per game in the past seasons
2PA	Float	2 Points attempted in the past 3 seasons
FGA	Float	Total field goals attempted in the past 3 seasons
FT	Float	Total free throws made in the past 3 seasons
3P	Float	Total 3 points field goals made in the past 3 seasons
DRB	Float	Total defensive rebounds grabbed in the past 3 seasons
ORB	Float	Total offensive rebounds grabbed in the past 3 season
AST	Float	Total assists in the past 3 season
BLK	Float	Total blocks in the past 3 season
TOV	Float	Total turnovers in the past 3 season
PTS	Float	Total points in the past 3 season
VORP	Float	Total defensive rebounds grabbed in the past 3 season
USG%	Float	Usage rate: how much a player possesses the ball on offense
STL%	Float	A player's steal rate in a game
DBPM	Float	Defensive Plus/Minus: measures a player's defensive contribution by looking at how much his team outscores (or is outscored by) the opponent
TOV%	Float	A player's turnover rate in a game
DWS	Float	Defensive Win Share: measure's a player's defensive contribution by looking at how many wins he is responsible for