Project Luther

Topic: Predicting NBA Player Performance

Given the rising international popularity of the National Basketball Association (NBA), this attention guarantees that interested businesses will want to invest in a team in one way or another. The main attraction of watching the NBA, or any sport rather, is to see the players perform. Therefore, it is in a team's best interest to pay the correct players the right amount and reduce risk when doing so.

Project Design

The first step for this project was to determine what response variable could translate as player performance. Choosing this was a design choice for the project. The typical metric in the sport of basketball for performance is how many points in a game they score. Therefore, *Points Per Game* would be the basis for the response variable. Data collection for our predictor variables (features) was restricted to the game stat sheet. The range of data collection started from 1979, as this was the date the 3-point line was introduced onto an NBA court and we wanted to ensure that contemporary player performance is relevant to data of earlier years, to 2018. The following are additional qualifications for data collection:

1. Acquire average stats for the first 3 years of a player's career
2. Acquire 4th year *points per game* as our true value; 0 assigned if player was not active

In building the model, linear regression was used to predict a player's performance. To test the validity of the model, k-fold cross validation was implemented between training set and test set. Performance of the model is measured by R^2 with normalized data in addition to the Mean Square Error.

Tools

Data was acquired via web scraping using BeautifulSoup and Selenium. Regression methods and cross validation was done through the use of statsmodels and sklearn.

Data

After performing a 10 fold cross validation for the model over 950 observations, the best average R^2 obtained given the current model is 0.67 +/- 0.06. Consequently, the best average MSE obtained was 1.41 +/- 0.06. What this translates to is that our model can account for 67% of the variation of the response variable *points per game* in the predicted season. Although the error between predicted and true value of points is not drastic, improvements can be made to reduce this error and gain a better fit to the model. In addition, a degree 1 fit is optimal for the given data as results show that difference between train and test MSE grows with increasing degree.

Future

In predicting performance, the model can be refined in a couple ways. First, extra data outside of the stat sheet needs to be collected, hours of sleep and distance traveled between games, for example. This will give a better understanding of how a player feels physically which could strongly affect their performance. Second, further exploration of certain operations on a feature such as applying log function of examining any possible interactions between features could result in a higher R^2. In addition, regularization should be implemented to account for the possibility of overfitting. Overall, what this project has enabled is a pathway to apply the statistical methods mentioned to similar situations and offer stakeholders a quantifiable measure of risk in their investments in trying to attract future businesses.