NBA Salary Predictions

Name: Jonathan Lawrence
Semester: Spring 2020
https://jonathan813.github.io/Portfolio/

Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you're doing with these data.

The data will come from the NBA. The assumption of most teams, and players as well, is that the more money a particular player demands, the better the player. This isn't always the case however, as there are many instances where players are either underpaid or overpaid. The purpose behind this course project is to use predictive analytics to predict what a player's salary should be by looking at player statistics.

References:

Gleeson, S. (2019, June 19). NBA draft: Ranking the five biggest busts as the No. 1 pick. Retrieved from https://www.usatoday.com/story/sports/nba/draft/2019/06/19/nba-draft-ranking-five-biggest-busts-no-1-pick/1488407001/.

Davis, S. (2019, June 19). Where are they now? The biggest NBA Draft busts of all time. Retrieved from https://www.businessinsider.com/nba-draft-busts-2015-6.

Ye, D. (2019, July 4). Top 10 Most Underrated NBA Players of the 2018-19 Season. Retrieved from https://howtheyplay.com/team-sports/Top-10-Most-Underrated-NBA-Players-of-the-2018-19-Season.

Casalan, A. (2018, October 20). NBA player stats 2017-18. Retrieved from https://www.kaggle.com/acasalan/nba-player-stats-201718.

Barrabi, T. (2020, March 6). How much do NBA players earn? Retrieved from https://www.foxbusiness.com/sports/how-much-do-nba-players-earn

Nicoll, C. (2020, February, 26). NBA salary bonus watch: Who's getting that extra cash? Retrieved from https://hoopshype.com/2020/02/26/nba-salary-bonus-watch-gobert-jokic-capela-rumors/

Voss, K. (2020, March, 9). Pistons: How Christian Wood is a dark horse to win NBA Most Improved Player. Retrieved from https://clutchpoints.com/pistons-how-christian-wood-is-a-dark-horse-to-win-nba-most-improved-player/

Swartz, G. (2020, March, 6). Every NBA Team's Biggest Flight Risk This Offseason. Retrieved from https://bleacherreport.com/articles/2878743-every-nba-teams-biggest-flight-risk-this-offseason#slide0

Pitts, W. (2020, February, 7). 10 Highest-Paid NBA Players of 2020. Retrieved from https://www.thebiglead.com/posts/10-highest-paid-nba-players-of-2020-01e0g96vsb80

These are the 2016/17 salaries of all NBA teams. (n.d.). Retrieved from
https://hoopshype.com/salaries/2016-2017/

Which Data?

What is the dataset you'll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

Since this is a continuation project, I will begin with the original Kaggle dataset that I used in DSC 630. Kaggle stated that the data were from the 2017-2018 season, but upon further investigation, the rosters were more in line with the 2016-2017 data. Additionally, the Kaggle data contained 88 missing salary values. These were obtained from hoopshype.com and manually added into the original dataset.

Dataset 1: https://www.kaggle.com/acasalan/nba-player-stats-201718

Codebook 1: The dataset did not have a specific codebook associated with it. I used the following sites to interpret each of the dataset features:

- https://www.basketball-reference.com/about/glossary.html
- https://www.basketball-reference.com/about/bpm2.html
- https://www.breakthroughbasketball.com/stats/definitions.html

The new dataset will be built from data obtained from the basketball-reference.com website for multiple seasons. I will be scraping the data with beautiful soup so that we can combine multiple years of data into one dataset. I will only be using the 14 statistical categories I identified as the most correlated to salary as determined from the original project. These variables are: Age, Player Efficiency Rating, Blocks, Turnover Percentage, Steals, Assists Percentage, Games Started, Games Played, Total Rebounds, Field Goals, Defensive Box Plus Minus, Defensive Rebound Percentage, Usage Percentage, and Free Throws.

Dataset 2: https://www.basketball-reference.com/leagues/NBA_2019.html

Codebook 2: The codebook(s) identified for dataset 1 will be used for dataset 2 as well.

Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you'll be proposing your research questions as well as justifications for why you'd offer these data in this way.

My goal is to use the previously developed model and apply it to more recent player statistics to test the generalization of the developed model. I plan to analyze this data by only selecting the features that are most correlated to salary. My analysis will focus on two questions:

1. Is my model generalized? Can it be used to predict salaries during other NBA seasons?
2. Based on performance, which NBA players from the 2018-2019 season are underpaid or overpaid?

What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

I will use the models from my original project to determine the best fitting model. The original project tested five different models and a similar approach will be used for this effort. The five models I will look at are:

1. Ordinary Least Squares
2. Ridge Regression
3. Lasso Regression
4. ElasticNet Regression
5. Extreme Gradient Boosting Regression.

Comparisons of R^2 and root mean square error (RMSE) values will be used to determine the best fitting model.

Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

The main concern of this project is that the model will not generalize well. While using multiple years of data should help, it is possible that the original model is only able to accurately predict player salary for the data it was built with. If that turns out to be the case, I will have learned a valuable lesson on model generalization and how to recognize when models cannot be used outside of their original domain.

This project could go off schedule if I am not able to get all of the necessary data for the 14 features identified above. I may end up having to do some extensive internet searches if all of the data are not available on basketball-reference.com. Additionally, this may cause me to have to use techniques to fill in missing data, such as using the mean value or setting a particular value to zero. This could end up being a very time consuming process.

Concluding Remarks

Tie it all together. Think of this section as your final report's abstract.

The National Basketball Association (NBA) is a men's professional basketball league comprised of 30 teams. Each of these teams have multi-million-dollar budgets that they use in order to build a talented team that will hopefully win enough games to bring home the NBA title. To do this, the teams build their rosters based on talent and budget. While team payrolls are large, they are not unlimited and strategic decisions must be made to develop the most talented team while maintaining budget constraints. The assumption of most teams, and players as well, is that the more money a particular player demands, the better the player. This is not always the case, however. There are numerous instances where players have been drastically overpaid and dramatically underperformed. Conversely, there have been players that have performed well above their higher-paid counter-parts. Ideally, there would be a mechanism where player salary could be determined by consistent player performance. It is reasonable to assume that NBA front offices do indeed subscribe to this premise, but some player salaries would suggest otherwise. The purpose behind this course project is to use predictive analytics to predict what a player's salary should be by looking at player statistics. This project outlines the process of developing a salary predicting model that can be generalized for future year use.