Predicting an NBA Player's Points per Game

Olivia Docal & Yuksel Baris Dokuzoglu

DATA 602

Purpose

The purpose of this study was to determine which method of analysis gave the most accurate result. The purpose of the models was to see how closely they could predict a player's points per game. Studies like this are important for predicting player's performance and helping people to make selections for their fantasy teams and other types of sports betting. The data set being analyzed is categorized by each player and includes minutes played, field goal percentage, free throw percentage, total rebounds, assists, steals, blocks, turnovers, and points. Taking these statistics, we wanted to find the best model for predicting points per game for an individual athlete.

Background

Fantasy leagues have seen a tremendous amount of growth in recent years. DraftKings, one of the most prominent platforms for online betting and fantasy play, estimated their revenue for 2021 to be between 900 million and 1 billion dollars. They boast over 1.5 million payers, contributing an average of roughly $65 each month in revenue. Stock for the company has also made a significant jump over the past year after DraftKings was able to go public in April of 2020 (Bursztynsky, 2021). This is just the financial reporting of a single online betting platform. There are many similar sites which also bring in a good deal of revenue. With the advancement of technology, sport betting and fantasy leagues have grown in popularity as well as accessibility. Most people have access to platforms, such as DraftKings, at the tip of their fingers through

smartphones and apps. Thus, the importance of analyzing player data to ensure the best results for further economic growth and stimulation.

There are many data sets available in regards to NBA statistics and many different types of studies that have been done with this data. Many studies revolve around creating visualizations and infographics using data from the NBA to help people make sense of how the statistics relate to one another and find correlations. I was unable to find a study similar to ours, which involves finding the best model to predict the amount of points for a player for the use of fantasy drafting. We hypothesized that there will be no significant difference in the predictive capabilities of each model. The alternate hypothesis being that there will be a difference in accuracy depending on the model.

Hypothesis

In order to test the hypothesis, we cleaned the data before trying out different models. To do this we filled any blank or missing values with zero because if, for example, a player had no steals and the slot was left blank, what that really means is that that player had zero steals. Next, we got rid of the seconds for minutes played so that the data was a whole number instead of a decimal.

H0: There is no difference in the accuracy of performance among the selected models.

H1: There is a difference in the accuracy of performance among the selected models.

We had our confidence level at 95%. According to the calculated p-value we did our hypothesis testing. We used Kruskal Wallis Test to check the difference in accuracy coefficients between the different models.

Linear Regression

The first method we tried was a linear regression model. Linear regression can be used to find a "linear relationship between target and one or more predictors" (Swaminathan, 2018). We began by importing necessary libraries such as atplot, , and . For the first test we made the X variable equal to minutes played and assigned points to the Y variable, before splitting, training and testing. We then printed the coefficients, mean squared error, and coefficient of determination. Coefficients came out to be 0.98, mean squared error was equal to 43.36, and the coefficient of determination was .33. We want the coefficient of determination to be close to 1, as 1 signifies perfect prediction. Next, we used a for loop to go through and find these features for each category with points as the Y variable. None of the categories proved to have a strong correlation to points when compared one by one. We then tried using multiple categories as the X variable, those being minutes played, field goal percentage, assists, and total rebounds. The coefficient of determination came back slightly stronger using this method at .56. Lastly, we used all the categories besides points as the X variable and kept Y as the points. This resulted in our highest coefficient of determination with .69. We went on to calculate the R-squared score, mean absolute error, and mean

squared error for linear regression. Mean absolute error equaled 3.34, mean squared error was 19.61, and the R-squared score came out to be 0.70. We then printed the predicted Y value versus the actual Y value, or predicted points compared to actual points. For many, the results were close to accurate. Below is the graph with actual values in orange and predicted values in blue.

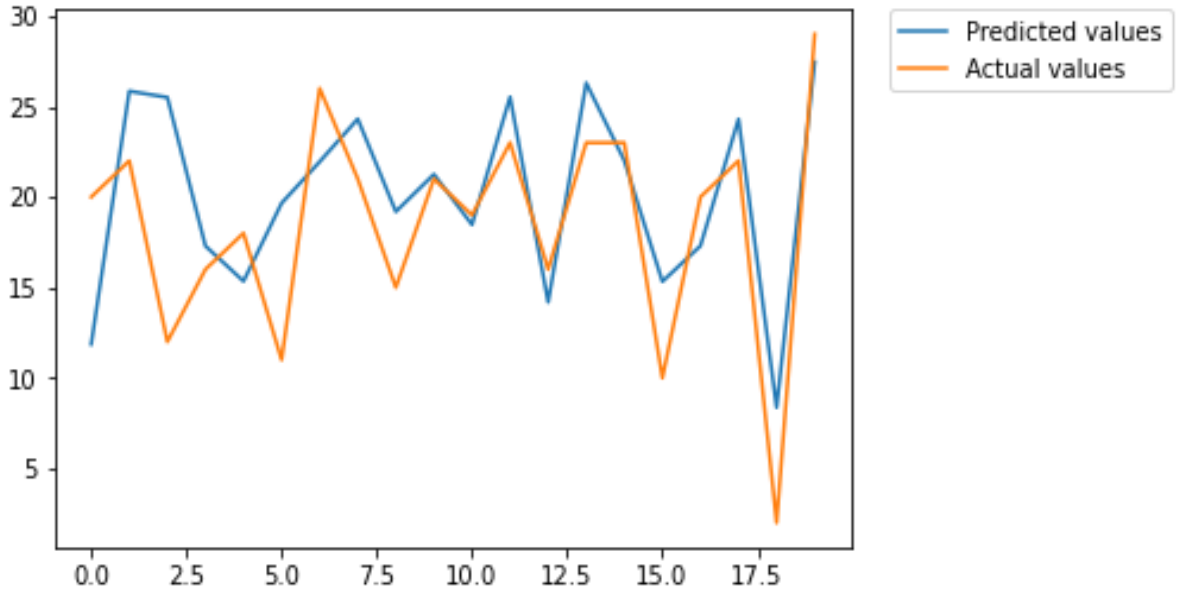

Fig. 1 Shows comparison between the predicted values and actual values.

Lasso and Ridge Regression

The next method we tested was Lasso regression. Lasso regression is a model that aims to achieve shrinkage by removing unnecessary attributes and setting them to zero (Diana, 2021). We calculated mean absolute error, mean squared error, and R-squared score for every model. R-squared score is the coefficient determination. It is an important statistical measure which shows how close the data are to the fitted

regression line. It explains to what extent the variance of one variable explains the variance of the second variable. To implement the Lasso regression, we defined the model and the model evaluation method. We continued by evaluating the model and then made it so that the scores would always be a positive number, and printed out the mean absolute error, which was 4.82. To conclude, we also found the mean squared error for the model equalled 34.5 and the R-squared score was 0.47.

Next we tested Ridge regression on the data set. Ridge regression is similar to Lasso, in the sense that they both want to . However, Ridge regression "uses a penalty to shrink all model parameters" (Diana, 2021). For this model, we followed the same steps we used for the Lasso regression model. The mean absolute error came out to be 3.14, the mean squared error was 14.9, and the R-squared score equaled 0.77.

Random Forest

The last model we chose to implement was the random forest model. Random forests are a type of supervised learning that utilizes ensemble learning for classification and regression. Ensemble learning involves combining the predictions from numerous machine learning algorithms. Thus, allowing the predictions to have better accuracy in comparison to a single model (Chakure, 2019). We defined, evaluated, and fit the model. We found the mean absolute error was 1.65, mean squared error was 4.02, and the R-squared score was equal to 0.94.

Comparing the performance metrics of the models to one another, we can see which ones have superior predictive capabilities. The R-score, mean absolute error, mean squared error, and root mean square error are what we used to determine how well the model functioned. Mean absolute error measures error between paired observations expressing the same phenomenon. Mean squared error measures the average of the squares of the errors. It is the average squared difference between the predicted values and the actual values. Root-mean-square error is used to measure of the differences between the predicted values and the actual values. Equations for R-squared score, MAE, MSE, and RMSE are below:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

$$mae = \frac{\sum_{i=1}^{n} abs\,(y_i - \lambda(x_i))}{n}$$

$$\mathbf{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Below we see how each model performed in predicting the points for a random sample of players. The random forest model had the best performance and accuracy according to these metrics because the mean absolute error is the lowest and R-squared score is the highest compared to other models.

Kyrie Irving	**R-Squared Score**	**MAE**	**MSE**	**RMSE**

Linear Reg.	0.69	3.34	19.61	4.42
Lasso Reg.	0.46	4.82	34.50	5.87
Ridge Reg.	0.77	3.14	14.90	3.86
Random F. Reg	0.93	1.65	4.02	2.00

Fig. 2. Shows the regression metrics for each model used with Kyrie Irving's dataset.

Victor Oladipo	**R-Squared Score**	**MAE**	**MSE**	**RMSE**
Linear Reg.	0.29	5.22	36.72	6.06
Lasso Reg.	0.48	4.36	26.88	5.18
Ridge Reg.	0.63	3.47	19.15	4.37
Random F. Reg	0.92	1.50	3.93	0.92

Fig. 3. Shows the regression metrics for each model used with Victor Oladipo's dataset.

Nikola Jokic	**R-Squared Score**	**MAE**	**MSE**	**RMSE**
Linear Reg.	0.12	4.24	32.49	5.70
Lasso Reg.	0.32	3.86	24.88	4.98
Ridge Reg.	0.44	3.26	20.35	4.51
Random F. Reg	0.87	1.73	4.53	2.13

Fig. 4. Shows the regression metrics for each model used with Nikola Jokic's dataset.

Conclusion

Our model can be easily implemented to different players, since we choose the dataset randomly from a list of players. The features can be changed easily in the model, if somebody else wants to check the effect of other features on certain target. After running the code for different players we got the best results for the random forest model. The mean absolute error was always the minimum compared to other models. Moreover, the R-squared score was higher than the other models. Thus, we can say that we got the best results from the random forest regression model. When we calculated the p-value with Kruskal Wallis Test, it came out as 0.71. It is higher than our significance level (0.05), so we will have to reject the null hypothesis (H0), and accept H1. For that reason, we can say that there is a difference in the accuracy of performance among the selected models.

References

Bursztynsky, Jessica. 2021 February 26. "DraftKings Shares Rise After Reporting a Beat on Revenue, More Growth in Paying Customers". CNBC. https://www.cnbc.com/2021/02/26/draftkings-dkng-q4-2020-earnings.html

Chakure, Afroz. 2019 June 29. "Random Forest Regression". Medium. https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Diana, Tony. "Introduction to Machine Learning DATA 602 Lecture 5 [PowerPoint Presentation]". https://drive.google.com/drive/u/0/folders/1cXzFx9S_bDlwU2BSpDytzERPnJKyieGE

Swaminathan, Saishruthi. 2018 February 26. "Linear Regression- Detailed View". Towards Data Science. https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86