Data 200 Final Report

Anyi Chen, Cheng Ren, Xinyuan Tu
May 13, 2020
Presentation Link: https://youtu.be/uKPS9cufEmI

Abstract

In this report, the basketball (NBA & NCAA) dataset was picked and three research questions were explored, including the 3-point shooting trends in the NBA, how well college data helps to predict NBA performance and who are going to be the next NBA superstars. Several steps were applied such as EDA and feature selections on the data. After fitting with Linear Regression, Random Forest, KNN and SVR, and evaluated by MSE, R^2 scores, and percent accuracy. Finally, 3-point shooting is more popular in recent years according to data visualization, Random Forest Regressor performs the best to predict NBA performance. An unsupervised machine learning K-means to cluster the players into groups and surprisingly find the potential next NBA superstars might be Giannis Antetokounmpo, Benjamin Simmons and Andrew Wiggins.

1 Introduction

For this final project, we picked up the basketball (NBA & NCAA) dataset. Before we start, we are all big fans of basketball and how data could be applied in the sports industry. Moreover, we are all interested in who is going to be the next NBA superstar. It's always exciting to guess who's going to be the next Kobe Bryant or LeBron James. To explore and reveal our curiosity, we separated our project into the following three phases of questions:

- Does three-point shotting become popular in the league in recent years and contribute more wins?
- If so, we may assume scoring ability including 3-point shooting accuracy matters their future. Before the players join the NBA, how well we can apply players' college basketball data to predict the player's performance in the NBA?
- After players joining the NBA, besides score, are other abilities (such as block, teamwork, steal, etc.) that may affect one's performance? Thus, we are going to explore if we consider comprehensively who are the all-stars and who are potential league icons next like LeBron James.

2 Description of Data

2.1 Exploration Data Analysis

2.1.1 Three points shot data on box score

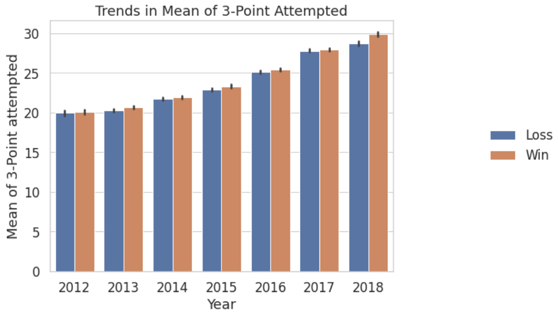


Figure 1: Distribution of total three points attempts made by team over years

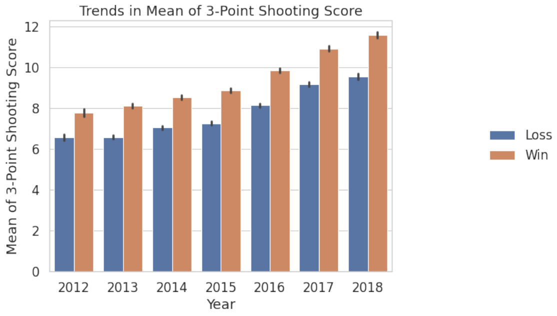
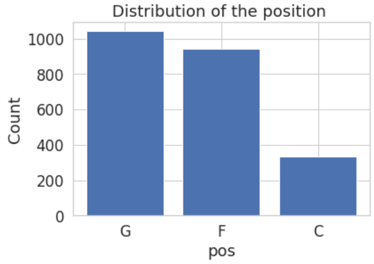
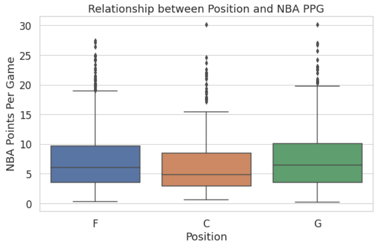


Figure 2: Distribution of total three points shots made by team over years

From the figures above, we can observe there's a trend that three-point shots have become more popular in recent years. Also, between loss and win, the goal for three points contributes more than attempts. We all know three points is the largest point(without foul) one can gain for a shot in the basketball game. To gain more points, it makes sense that players attempt to do more three points shots.

2.1.2 College/NCAA data





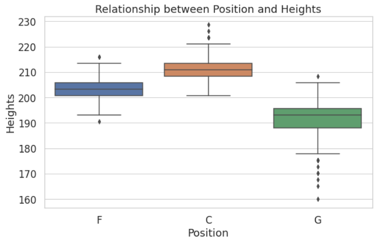


Figure 3: Value counts for different positions

Figure 4: Box plot of score and different positions

Figure 5: Box plot of height and different positions

As we can see, different positions have different numbers of players in the data set. Even though a person can fit into different positions, the need varies based on different situations. For example, forwards and guards may gain more points than a center because they are supposed to take more responsibility on offense have more attempts on shots and get more points. Thus, for our models and other visualizations, we would like to split with positions.

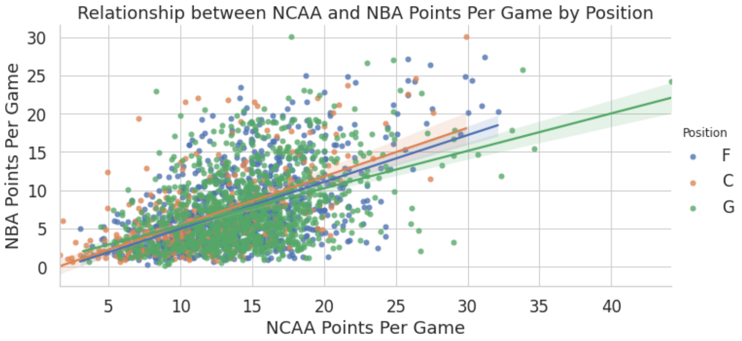


Figure 6: Scatter plot of NCAA points per game and NBA points per game

From this figure, we can see there's a positive growth of NCAA points per game and NBA points per game. So players' overall performance at NCAA has some relationship with the overall performance at the NBA.

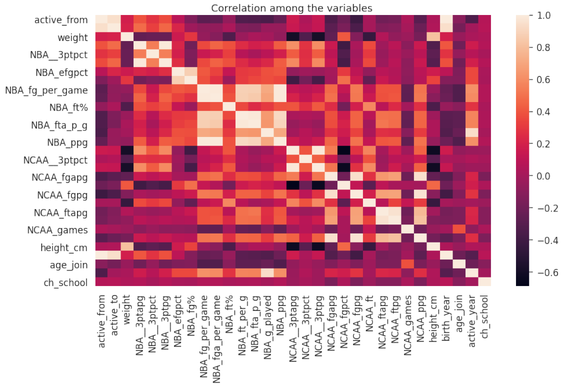


Figure 7: Correlation heat-map of college data set

After comparing the correlation between features, we would like to remove one of two features that correlate higher than 0.9. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features.

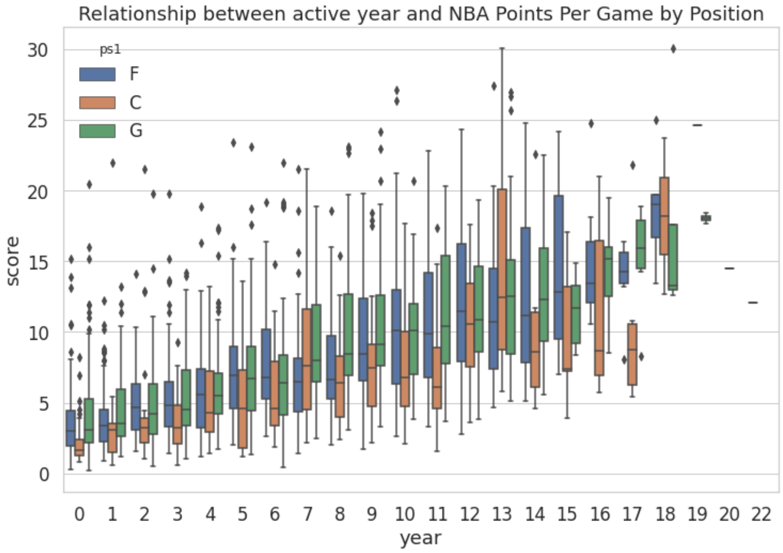
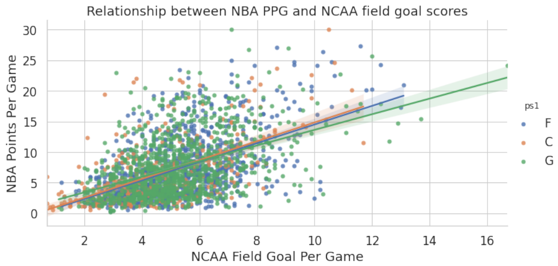


Figure 8: Box plot of active year and NBA points per game

The above figure indicate interesting relationship between the active year and the NBA points per game. The overall trend of the graph shows that if a player serves longer, then he gains more points. However, as we look into specific positions, the center position is not that varies as other positions. Centers' bodies take a heavy toll when they on the court, especially in the paint area.



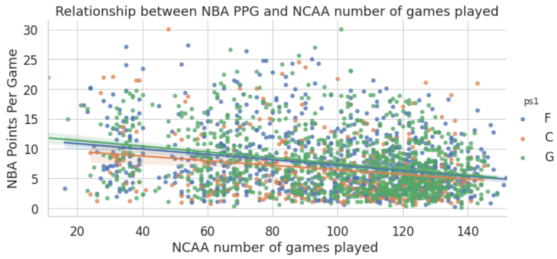


Figure 9: Scatter plot of NBA points per game over field goal per game made by team

Figure 10: Scatter plot of NBA points per game over NCAA games

Last but not the least, from this two graphs, we can also tell relationship of other features from NCAA data and NBA points per game.

2.2 Data Cleaning and Transformations

In this section, we mainly applied data cleaning and transformation with several steps:1)Clean NAs, 2)Inspect some odd columns or values, 3)Create some features. In terms of the "Box Score" and "Play Box Data", these two data sets are almost clean so we only pick the variables we need without much cleaning.

As for the "college" data set, following steps are applied:

- Check the data structure and calculate the missing status of each column.
- Create their main position. Since some players with dual positions, such as F-G, we searched several players via Google and cross-checked their information. Thus, we can assume that the first position is their main position. We create "pos1" to indicate their main position and "pos2" for their minor position.
- Transfer height in numeric. The data type of players' heights is "object". Thus, we turn their heights into numeric.

- Get birth year, the age to join NBA and active years. The year of players' birth, joining and retiring from the NBA was applied to create some new variables.
- University dummy. In this part, we originally thought players who graduated from universities with great NCAA basketball records should be more competitive. Thus, we create a dummy variable called "ch_school" and universities with at least two NCAA championships will be labeled as 1.
- Deal with NAs in the data set. First, players with at least 10 NBA games will be kept. Second, there are some columns such as "NCAA_3ptpg" which has round 59% missing values but these indicators may be important to predict NBA Point per Game. Meanwhile, according to EDA, each position may have different abilities. Thus, we divide the players into three different positions(C, F, G) and fill the missing value by their mean of each position.

3 Description of Methods

After cleaning the data, we were going to create models to show whether NCAA data has a relationship with NBA performance, and how we might predict based on NCAA data. First of all, we tried to select important features based on the features' correlation. As a result, we dropped "ch_school" whose correlation with NBA points per game (NBA_ppg) is very low. Then, we created models for overall NCAA data to predict NBA_ppg with selected features. To be more specific, we tried with Linear Regression, Random Forest Regression, K Nearest Neighbour Regression, and Support Vector Regression (SVR). We compared their mean squared error (MSE), coefficient of determination (R2) score and percent accuracy to select the optimal model. Then, we had a closer look at the "center" subset to see whether the prediction would be more accurate within one particular position.

The percent of accuracy is a customized function written by us. Mainly, this function will return what percent that predicted score is within 15% of the actual score. For example, if one player's predicted NBA PPG is 9 and his actual score is 10, 9 is within 15% of the actual score. Thus, the function will return 1(True). Finally, this function will return the percent of "good guess", in other words, the percent of 1(True).

Furthermore, aiming to find the potential league icons, we used K-means to cluster the cleaned NBA player box score data set. We selected K as 6 based on the minimum squared distance(elbow plot) within clusters. After the clustering, we regard the group which LeBron James belongs to as a group of talented players. Eventually, we selected the young players in that talented group to be our predicted potential league icons.

4 Models and Assumptions

- Our best model is a Random Forest Regressor. We used cross-validation to find the optimal number of estimators to be 8. Also, we set the minimum samples leaf to be 5 and the maximum depth to be 8 to avoid the issue of over-fitting. Though Random Forest Regressor doesn't make assumptions on the underlying distribution of the data, it is still based on the assumption that the sampling for each tree is representative.
- For the Linear Regressor, we assume that the features we chose are not linearly dependent and the residual plot is evenly distributed around the horizontal axis.
- For KNN Regressor, the cross-validation results showed that the best K is 13. Similar to Random Forest Regressor, KNN Regressor is also non-parametric which means it makes no distributional assumptions of the data. It only assumes that similar data points are close to each other in the feature vector space.
- For SVR, we found the optimal epsilon to be 1.9 and optimal C to be 0.4. Again, it is a non-parametric model and only depends on its own kernel function.

5 Summary of Results

In terms of R1, the data visualization in EDA has shown that three-point shots style have become more popular in recent years.

The following table answers R2 and shows the performance of different regression models using different data. We can tell that the Random Forest Regressor with the entire NCAA data set has the best performance in all three metrics.

Model(Test Set)	MSE	R^2	+-15%
Linear Regression(NCAA PPG only)	19.9	0.27	0.19
Linear Regression	12.4	0.54	0.25
Random Forest	**10.3**	**0.62**	**0.29**
KNN	12.8	0.53	0.25
SVR	12.9	0.52	0.25
Linear Regression(Center only)	14.9	0.56	0.24
Random Forest(Center only)	14.07	0.58	0.23
KNN (Center only)	13.8	0.59	0.23
SVR (Center only)	15.5	0.54	0.18

Table 1: Compassion table of MSE, R2 and percent accuracy(+-15%) among different models

According to the table above, we concluded that simple linear regression between NCAA PPG and NBA PPG is not enough. When we include all the selected variables with mixed-positions. Random Forest returns the best MSE(10.3), R^2(0.62) and percent accuracy(0.29). When we test the single position(Center), KNN returns the best MSE and R^2 but liner regression returns the best percent accuracy(0.24). However, percent accuracy is not as good as the mixed-positions model. Somehow, we would recommend predicting score with mixed-position and Random Forest is a sound method among the models we tested.

To present our clustering result in 2 dimension and answers R3, we applied PCA to reduce the dimension to 2. The clustering result is shown is Figure 11.

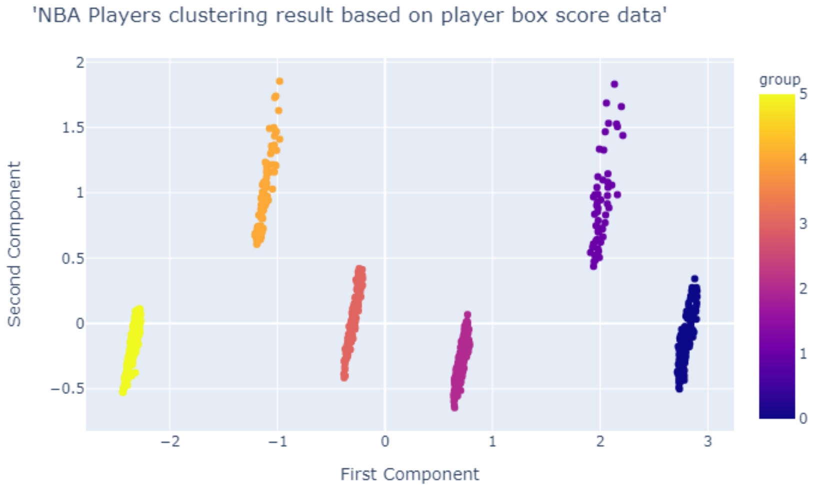


Figure 11: Cluster NBA stars into different groups based on player box score data

According to Figure 11, the cluster analysis provides a very clear 6 clusters. Since we are going to predict who are potential league icons next like LeBron James, we explore who is in LeBron's cluster and what is the characteristics of this cluster. We found many all-stars are in his cluster like Steph Curry, Kevin Durant. There share some similarities like high PPG, a high number in assists, a pretty high number in field-goal attempts and so on. Then, we think the next super start should at least 10 years younger than LeBron James. Meanwhile, when we consider their position, we predict the candidates might be Giannis Antetokounmpo, Benjamin Simmons and Andrew Wiggins based on the data by 2018.

6 Discussion

6.1 Specific Questions

- (i) NCAA_points per game, active year, university, and points per game are the most interesting features.
- (ii) The college feature is the feature we thought would be useful but turned out to be ineffective. We thought champion colleges would produce more superstars and better performances. However, when we do feature selection, this feature has the lowest correlation with NBA points per score, and this feature has been dropped.
- (iii) For the data itself, since we split them into positions, the size of the train and test data set is small. Also, it doesn't include all the basketball data. We also found it was difficult to start the project and make a clear statement of what questions we want to explore and answer since we do not have much very professional domain knowledge in basketball.
- (iv) We thought positions would help and produce a better result, however, it turned out to be less effective.
- (v) We assumed the mean for null data on three-point field goals per game. Maybe some players just didn't have the data and they may have better performance than the mean.
- (vi) Other NCCA data besides scores. For example, does the player receive some awards such as the Region's Most Outstanding Player? Also, this data doesn't include the year of a play attend NBA draft, because some players may participate in the draft when they are freshmen or sophomore. One other data is injury history because this affects a lot to the performance.
- (vii) It seems like we don't have any ethical problems unless the players don't want their data to be open to everyone. If we have injury history data for the players, some of them may not willing to publish the data. Also, how you define the severance of the injury is also a concern. It's hard to solve. Maybe it's a good choice not including it in the data set.

6.2 Evaluation

The evaluation metrics for our models are MSE, R2 score and percent accuracy. A smaller MSE means better performance while a larger R2 score or percent accuracy means better performance. We also use the train- test split to test the data set. There is no very big difference between the results when we use the train or test dataset.

Generally, we selected features based on feature correlations and applied four regression methods on the data set. It turned out that the Random Forest Regressor outperformed other models. However, it still has the risk of over-fitting. When we evaluated our model, the residual plot was applied for linear regression. Meanwhile, since the random forest is easy to over-fit in this dataset, we compared the difference between the training set and test set to evaluate the results and fine-tuning the model to regulate it.

Besides, the limited number of features also constrains the performance of our model.

6.3 Future approach

We assume each position has different obligations. For example, under the current circumstances, shooting guards or small forwards may take more responsibility on offense like scoring. Thus, we assume some scoring patterns may vary based on their position. Therefore, we divided players into each position and try to fit a model for prediction. However, after testing, we realized that the mixed-position model has a better performance than the single-position model. One assumption is that the single-position data set makes the sample much smaller than the mixed-position dataset.

With the current data, some models may be applied to evaluate players' performance. In the NBA, some commentaries always say players' ”effectiveness” and ”efficiency” and point how one player could influence the whole team. If we can learn these models and create new features, we might be able to make a better prediction like Research Question 3.

In the future, if possible, NCAA could record more data like steal, block in their data set. This may help make better predictions not only in scoring but also in other aspects both in offense and defense.

7 Reference

Kannan, Adarsh, et al. ”Predicting National Basketball Association Success: A Machine Learning Approach.” SMU Data Science Review 1.3 (2018): 7.

Kannan, Siddhesvar. “Predicting NBA Rookie Stats with Machine Learning.” Medium, Towards Data Science, 30 June 2019, towardsdatascience.com/predicting-nba-rookie-stats-with-machine-learning-28621e49b8a4.

“NCAA Division I Men’s Basketball Tournament.” Wikipedia, Wikimedia Foundation, 11 May 2020, en.wikipedia.org/wiki/NCAA_Division_I_Men’s_Basketball_Tournament.