Predicting NBA Shots

CSE 160: Final Project
Professor Davison
Group 4: Matthew Calvin, Regina Lee,
Gabriella Nuzzolese, Carl Saba

Executive Summary

Watching the NBA is a favorite pastime for many people for many reasons including the fast paced nature of the game and popularity of the players. With players generating huge fan bases and support, NBA betting is becoming more popular amongst the audience as fans try to guess the outcome of a certain matchup between teams. The task we wanted to tackle was to be able to predict whether or not the ball would go in whenever someone took a shot.

We collected the data off of one of the datasets we found on Kaggle called "NBA Shot Logs" which provided data on all shots taken during the 2014-15 season. Some of the included information in the dataset that we found to be relevant were matchup, touch time, shot distance, closest defender, and the shot result. The data was found to be mostly clean with the exception of some NA values in the shot clock category. The NA was initially thought to be buzzer beaters but due to the uncertainty of it, the values were omitted after some discussion.

Using the data, we applied it to two different data science operations (Random Forest and Naive Bayes) to see whether we could predict the shot result based on some if not all of the other factors. Our results showed that although the Random Forest provided better accuracy (Random Forest had an accuracy of 60.7% vs Naive Bayes had an accuracy of 58.6%), the Naive Bayes model outperformed the Random Forest in other aspects including precision and recall which are known to provide more relevant results. In the end we decided that the Naive Bayes method gave us the best predictions.

Ultimately, our goal for this project was to provide a helpful tool for those involved in the NBA whether they were NBA teams, coaches, players, fans, etc. We thought that based on the statistics, players could adapt their personal strategy with a more data-driven approach to practicing for future games and matchups. Coaches could also come up with better tactics that could benefit the team overall and fans could find our analysis useful for placing bets. Through our model, we hoped to provide a tool for various target groups that would elevate the NBA experience.

Introduction

Data Overview

The NBA Shot Log dataset contains data on shots taken in NBA games in the 2014-2015 season. The data is owned and was collected by a Kaggle.com user named Dan B. The data was taken using SportVU technology. SportVU is the official data tracking technology of the NBA. SportVU has a modernized camera system which records 25 data points a second which records all the attributes of the data. There are 128,069 samples in the data. The data was last updated August 18, 2016.

Data Cleaning

The dataset was fairly clean overall. The only noticeable issue was some NA values on the shot clock. It was unclear whether these NA values were supposed to be zeroes or something else. However, the shot clock was not used in any of the models, so the NA values did not become an issue.

Data Science Problem

The goal of this project was to create models that can predict the classification of shots using other data points within the dataset, such as shot distance, closest defender distance, touch time, and points type.

Customer Benefit

The target customers we are focusing on are NBA teams/coaches/players, and fans. Our analysis will allow for better predictions on how many shots the player or team will make, which teams will win the game, and more. Players can use the model outcomes to develop a more data-driven approach to practicing. Fans can use the analysis for placing bets. Coaches could use the modeling to improve their tactics.

Modeling

Model Selection

The goal of our model was to predict the classification of shots in the 2014-2015 NBA season. We decided to create multiple models and compare performance to determine the best models. These models included Random Forests and Naive Bayes. It seemed unreasonable to

focus only on one model as these models have unique advantages for predicting classification of a specific attribute, in our case, whether each shot went in or not.

Validation

To test the results produced by our Random Forest and Naive Bayes models, we used 10-fold cross validation. In each of the folds, 90% of the data available was randomly selected to be used for training the model while the final 10% was used to test the prediction. For each of our model results, we used the average value of all of our performance measures across all of the 10 folds. We found that this technique for cross validation produced much more consistent results with smaller variance than using a single training set and testing set across both models by reducing both variance and overfitting.

Attribute Selection

For every one of our models, we classified the shot result based on the shot distance, closest defender distance, points type, and touch time.

1. **Shot Distance:** If the shot is very close to the hoop (a layup), an NBA player is expected to make this shot almost every time. Alternatively, if the shot is very far away from the basket, say 30 or more feet, this is considered a deep three, and this is a kind of shot you wouldn't expect to go in very often.
2. **Closest Defender Distance:** Generally a player is much more likely to make a shot when the closest defender is far away meaning they are wide open, having an uncontested shot.
3. **Points Type:** We considered the points type, meaning whether the shot was a 2 pointer or 3 pointer, because it is well known in the NBA that the average shooting percentage for 2 pointers is much better than the average shooting percentage for 3 pointers
4. **Touch Time:** The amount of time the player touched the ball before they shot because generally players make a much higher percentage of shots where they catch the ball and shoot very quickly, rather than taking time to dribble which would put off their shot.

Performance Measures

To measure the performance of each of our models, we took the average accuracy, precision, recall, and f-measure of each of the 10 confusion matrices produced by cross validation for each of our results.

- **Accuracy:** Most intuitive measure of performance, simply the percentage of predictions which were correct
  - $(TP + TN)/(TP + FP + FN + TN)$
- **Precision:** the percentage of predictions labeled as positive where actually true positives

    - $TP/(TP + FP)$
- **Recall:** the percentage of actual positives which we correctly labeled as positive
    - $TP/(TP + FN)$
- **F-measure:** weighted average of precision and recall
    - $2 * (Recall * Precision) / (Recall + Precision)$

Random Forests

For our first model, we created a random forest model which generates 100 decision trees to vote on classification output for each instance in the test set. The classification with the most votes gets predicted. In this model, we included all 128069 observations from the original data set. A drawback from this model is that our model predicted classification shots of many players based on training data of possibly different players.

Naive Bayes

We created multiple Naive Bayes models, using all 128069 observations from the dataset. Each fold generated a model which predicted the classification of the shot results assuming all the attributes we selected were independent, then used this model to predict each value in our test set. Similarly to Random Forests, possible drawbacks is that our model predicted classification shots of many players based on training data of different players. Therefore, we also decided to gather 3 subsets of the original dataset: data consisting of only shots from Lebron James, then Kyrie Irving, then Steph Curry. We created similar Naive Bayes models to predict shot classification based on the same attributes for each of these 3 players. Our goal was to see if some players had "preferences", meaning that it could be possible that our models would be more accurate in predicting our target variable due to trends analyzed by each player.

Analysis

As stated, we performed 10-fold cross validation using both a random forest and Naive Bayes. The cross-validation is used to efficiently validate the performance of the designed model. For each, 10 confusion matrices were generated which display the summary of predictions for each classification problem. From each of these matrices, we can view the predicted versus actual shot results. In the top left, we see the number of shots that were made, and the bottom right is shots missed. The top right shows shots predicted to be made but actually missed, and vice versa for the bottom left. With this information, we can calculate the accuracy, precision, recall, and f-measure for both methods as seen in Figure 1.

	Random Forest with 100 Trees	Naive Bayes

n	120869	120869
Accuracy	0.6067589	0.5856296
Precision	0.3070654	0.4208446
Recall	0.4120346	0.5451079
F-Measure	0.3518886	0.4749834

Figure 1

The accuracy of the Random Forest model in this model is slightly higher than Naive Bayes. This may be due to the fact that Random Forests can handle a large number of different attributes due to embedded feature selection in the generation of the model. Furthermore, the accuracy for Naive Bayes, while this algorithm can also handle large datasets, may be slightly less given the assumption that all attributes are independent, which may not necessarily be true in real world data.

While the accuracy of the Random Forest is slightly higher, accuracy does not distinguish between the numbers of correctly classified examples of different classes, meaning alone it can lead to erroneous conclusions. When we account for precision and recall, we see that the Naive Bayes values are significantly higher than Random Forest. This is important to consider because precision refers to the percentage of results which are relevant and recall refers to the percentage of relevant results that are correctly classified by our chosen algorithm. As a result, taking all performance metrics into account, we conclude that Naive Bayes 10-Fold Cross Validation produces a better model when predicting NBA shots made and missed with this given data set.

Similarly, we performed Naive Bayes Cross Validation for 3 specific players: Lebron James, Kyrie Irving, and Steph Curry. We created similar Naive Bayes models to predict shot classification and calculated the associated performance measures as seen in Figure 2.

	Lebron James	Kyrie Irving	Steph Curry
n	978	942	968
Accuracy	0.6206291	0.5381747	0.607442
Precision	0.6279489	0.5089246	0.6093684
Recall	0.5560297	0.4801776	0.5463222
F-Measure	0.5898050	0.4941334	0.5761256

Figure 2

When comparing individual player performance measures to the entire dataset, we can see that some of the percentages for specific players are higher than the overall. As an example, Lebron's performance measures are all respectively higher than the overall percentages for the entire NBA dataset, giving us evidence to believe that it could be possible our model is more accurate in predicting the target variable when analyzed by each player. The same can be said for Steph Curry's statistics as seen in Figure 2.

Accounting for Error

One factor that could potentially affect the performance of our model is the attributes we selected to classify the shot result. We chose shot distance, closest defender distance, points type, and touch time when predicting the shot result, not taking into account the final margin, shot number, period of the game, game clock, shot clock, and number of field goals made. Including or omitting any of these other attributes could have affected the performance measures of our models either positively or negatively, however we chose the attributes we found most relevant to our target variable.

Another factor that could have negatively impacted our results for the Naive Bayes model is the fact that Naive Bayes assumes all attributes are independent of each other, which rarely happens in real life. As a result, this may limit the applicability of the Naive Bayes algorithm in real-world use cases.

Furthermore, our dataset has data from specific players in specific games, which may not always be indicative of their best performance, therefore lowering their predicted probability of making a shot when in reality that player may have had an off game or two. As a result of this discrepancy, we could see a lowered percentage for the four performance measures for both specific players such as Steph Curry and for the entire dataset overall.

Another important aspect that our group would have liked to consider is the player's exact location on the court so that the model could predict with even more accuracy whether or not it is likely that a shot will go in. While this information was not exactly provided in the given dataset, this would be an interesting point to consider in order to give fans, coaches, and players deeper insight into specific points out in the field that shots are made from more often than not as well as locations in the field that are not conducive to making shots.

Conclusion

In the end, although we found the Random Forests model to be more accurate, the precision and recall of the Naive Bayes was much higher and with only a 2% difference in

accuracy between the two, we determined that the Naive Bayes model was what gave us the best results in terms of shot predictions. Overall the Naive Bayes gave us more relevant results that the Random Forest couldn't give us. As mentioned before, the Naive Bayes model also proved to be more accurate when predicting shots for a specific player which was useful to know since one of our target goals was to create a model that could help players identify and improve on their weak points and create a more efficient way to practice. To conclude, while it is not recommended to be used on its own, we believe that in conjunction with other tools the Naive Bayes model we created to predict NBA shots could be of service as we originally intended it to be, providing useful predictions to various people including the players as well as coaches and fans.

Appendix

Matthew Calvin - Wrote the introduction which contained data overview, data problem, data cleaning, and customer benefit.

Regina Lee - Wrote the executive summary and the conclusion

Gabriella Nuzzolese - Conducted analysis on each of the two models in order to determine which model produced better results overall. Determined the errors associated with each model and how each error could have impacted the results obtained.

Carl Saba - Found and cleaned dataset, created sub datasets for individual players, and created and discussed all models used in the report.