COLUMBIA UNIVERSITY
School of Professional Studies

NBA Shot Probability Analysis

Project Deliverable #2

December 10th, 2017

Eric Pridgen, Fan Yang, Kun Yang, Wangyi Zhang, Zixin Jia

Table of Contents

1. BACKGROUND INTRODUCTION

1.1 Research Questions

NBA teams are constantly searching for better players and court strategy. There are many aspects of an NBA player that a general manager must consider when performing their analysis: offensive and defensive ability, makeup, personality, etc. Many of these skills, especially those pertaining to offense and defense, can be quantified using statistical measures. One of the most common measures in basketball pertains to shot results. Our project aims to improve the court performance from both offensive side and defensive side with following research questions:

- **How do offenders affect shot results?**
- **How do defenders affect shot results?**

1.2 Data

We obtained original shot log dataset from Kaggle contains 128069 shot observations that occurred during NBA 2014-2015 season. These observations include 21 variables. Some of them are counted as missing, while other useful variables related to player stats we cannot find. We merged shot logs with the NBA Players Stats dataset and did necessary cleaning process. The proportion of our sample that is missing is about 4%, but left of our sample has plenty of data for our analysis. The remaining concern is that a systematic pattern to the messiness might be present. We test its validity by randomly sampling, and we did not find any evidence that the two populations, missing and non-missing, are different, so we are confident that no biases are introduced in our calculations by dropping the missing observations from the dataset.

2. METHOD EXPLANATION

Note: The specific figures, R code, and outputs are in the Appendixes.

2.1 Method Choice

2.1.1 Descriptive Analysis

Based on the key variables in NBA games and the historical data collected in the datasets, we developed a descriptive analysis on shot results, which were described as "SHOT_RESULT" in our final database. We explored how factors in a basketball game (like shot distances, dribble times before a shot and the nearest defenders' experience in NBA) effect the shot results and to which extend. In this way, we could give NBA teams and their general managers detail descriptive indexes which will help them select most suitable players, develop court strategy, and improve shot hit rate in NBA games.

2.1.2 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a binary variable (in which there are only two possible outcomes). Since our dependent variable "SHOT_RESULT" is binary, and was determined by multiple independent variables, logistic regression is a great fit for our project. Moreover, we utilized the stepwise method to help us select variables in the regression model.

2.2 Pros and Cons

	Pros	Cons
Descriptive Analysis	• Reveal the relationship between variables • Easy to conduct • Provide insights for further modeling	• Can only reveal the surface layer relationship of variables. • Cannot make accurate prediction of the future.
Logistic Regression	• As a classic modeling method, logistic regression method is clear and easy to be understood • Suitable for our dataset as our dependent variable is categorical	• May not perform as accurate as other more complicated models like Decision Tree and Random Forest • Based on certain assumptions

3. DATA ANALYSIS

3.1 Analysis Processing Summary

Before starting the data analysis, **we separated defensive and offensive independent variables apart into two sub-datasets and built logistic regression model for each side**. That is because one team cannot be the offensive side and defensive side at the same time. Furthermore, aims of the offensive side and defensive side are different. By building separate model for each side, we could provide more straightforward insights for both sides which has more practical purport. We conduct same process for the both data sets.
Our data analysis process contains three steps which reflect our work flow to dig deeper into our research question.

- The first one is to explore the variables we would use in the modeling with descriptive analysis. Before the modeling, we should have an overall understanding of our variables. The relationship between independent variables and dependent variable is worth exploring, and the correlation and interaction between the independent variables may impact the validity of the model.
- The second one is our modeling process. As illustrated above, we input our variables into logistic regression and see the results. For the sake of practical insights, we categorized the variables into offensive side and defensive side.
- The third step is to evaluate the validity of our model. As we split the data set into train set and test set, we could estimate the prediction effect of the model. In addition, ROC curve is also a commonly used method to assess models, especially for classification methods.

3.2 Variables Exploration

Before running the logistic regression model, we explored the correlation between variables by using the correlation plot. The darker the blue, the higher the positive correlation. The darker the red, the higher the negative correlation.

Variables in offensive side:

Figure 1: Correlation between variables

In our dataset, we have a high correlation between *TOUCH_TIME* and *DRIBBLES*.

Variables in defensive side:

Figure 5: Correlation between variables

From the correlation plot, we can see that BMI and Weight are positively correlated**,**

This logically makes sense as longer you hold the ball, more dribbles you would make. This kind of correlation may influence the modeling effect. Rather than remove a column or merge the information, we will make a note to assess the validation results later.

Figure 2: Distribution of numeric variables over dependent variables

The boxplots show the distribution of numeric independent variables among the different value of dependent variable. We could see that the *SHOT_DIST, SHOT_CLOCK,* and *GAME_CLOCK* have relatively significant difference. It may indicate that these variables could be significant in the modeling.

Figure 3: Distribution of category variables over dependent variables

Pick *PERIOD* as an example, the mosaic plot shows us the distribution of shots in different periods. With the game going on, (the period going from 1 to 4, or even more), the shot hit rate seems to decrease. It may could be explained by the physical strength decreasing of players.

Figure 4: The interaction between variables

Take the interaction between SHOT_CLOCK and PERIOD as an example, the Shot Clock has a positive relationship with shot result, which means the faster they shot, the higher possibility they could put the ball in. However, as game goes on, especially in the overtime (period 5,6,7), players tend to shot the ball more carefully.

Weight and Height are positively correlated; Position and Height are negatively correlated and Weight and Position are negatively correlated. To sum up, these variables are correlated to each other which could weaken their significance in the model.

Figure 6: Distribution of numeric variables over dependent variables

We could see that the *Height, SHOT_CLOCK* and *GAME_CLOCK* have relatively significant difference. It may indicate that these variables could be significant in the modeling.

Figure 7: Distribution of category variables over dependent variables

Pick *Pos(Position)* as an example, the mosaic plot shows us the distribution of shots facing defenders in different positions. With the position goes from 1 to 5 (1 stands for Point Guard, 2 stands for Shooting Guard, 3 stands for Small Forward, 4 stands for Power Forward and 5 stands for Center), the shot hit rate seems to decrease. It may could be explained by the height of the players, as Point Guard (Pos1) is shorter and the Center (Pos5) is generally the tallest player in the team.

Figure 8: The interaction between variables

Take the interaction between Position and Shot Clock into example. the Shot Clock has a positive influence on shot result which means the faster they shot, the higher possibility they could put the ball in. And as discussed above, the position has a negative impact on the shot result. But there is no strong interactions between the two independent variables.

3.3 Logistic Regression

We selected a binary predictor, SHOT_RESULT, as the dependent variable, as different independent variables for offensive side and defensive side. The process of modeling is the same except different variables.

Offensive Side:

First, we separated the data into train and

Defensive Side:

With seven independent variables, we ran a

test data and used 90% data as the train data. From the variables exploration part, we realize that some variables may not be appropriate for the model. Rather than selecting the input variables by our intuition, we used stepwise method to help us select variables in the regression.

step analysis using a null model to determine which independent variables to use for the final model. Our final model output consisted of two significant variables, Position and closest defender distance (Position is a factor with five levels, indicating the five NBA positions).

3.4 Modeling Assessment

We used ROC curve and confusion matrix to evaluate our models. As for offensive side, we graphed the ROC curve and computed the Area Under the Curve (AUC). In addition, we know the accuracy of predicting test set from the confusion matrix. The overall assessment indicates that our model is not the best one but is above the average.

	Area Under the Curve	Accuracy
Offense	0.6192.	62.47%
Defense	0.5816	53%

3.5 Modeling Conclusion

Offensive Side Conclusion:
Variable LOCATION is not significant in the model. Other variables, PERIOD, GAME_CLOCK, SHOT_CLOCK, DRIBBLES, TOUCH_TIME, and SHOT_DIST are significant in the model. There into, period, game clock, dribbles, touch time and shot distance have a negative relationship with shot result while shot clock has a positive relationship. The specific model information is in appendixes.

Defensive Side Conclusion:
Many variables are not appeared in the final model after the stepwise method. It may result from the high correlation between the variables. For others variables, CLOSE_DEF_DIST (closest defender distance), Pos4 (Power Forward), and Pos5 (Center) are all important. Moreover, the CLOSE_DEF_DIST has a positive relationship meanwhile Pos4 and Pos5 influence the result negatively. The specific model information is in appendixes.

4. RECOMMENDATIONS

Offensive Side

- Don't be afraid of "Away Games"

"Away Games" means the games not at home. For example, if New York Knicks fight Chicago Bulls in Chicago, that is an away game for New York Knicks and a home game for Chicago Bulls. As shown in the model, the location is not significant. Many teams are afraid of fighting in away games and they usually blame the failure on fighting away. However, the true reason may be the court strategies or cooperation rather than location.

- Having more "Catch and Shoot"

"Catch and Shoot" means the player shoot the ball at once after he receive it. The backup reason shows in the model. The dribbles, touch time and shot clock have a significant negative relationship with shot result. This indicate that, the more time you use to dribble the ball, or

hesitate to shoot, the more failure you could make. Therefore, using court strategy to create space for "Catch and Shoot" is a high efficiency way to earn more score.

- "Early Fighters" win the game

By saying "Early Fighters", I refer to the team put more emphasis on the early periods of the game and earn more scores in first two periods. Many NBA players think the last period is the most important one. There are some commonly used indices measure the performance of players in last period. And many coaches have made specific court strategies for the last period. However, the variables "Period" and "Game Clock" have a significant negative relationship with shot result. A possible reason is the physical strength loss during the game. Therefore, "Early fighters" may have greater advantages in the game.

Defensive Side

- Keep closer

The defensive court strategy should put emphasis on defensive players' relative distance to NBA offensive players. In our model, the variable "CLOST_DEF_DIST" (closest defense distance) is significant and has a positive relationship with shot result. It suggests that the odds of a shot being made decreases as the distance from the offensive and defensive players are closer. We suggest that coaches develop plays specifically to keep defensive players as close as possible to the player with the ball.

- The taller, the better

Assign more defensive missions for power forward and center positions in tactics. As tallest positions in a basketball team, players on these two positions could make most effective defense which reduces opponents' goal numbers to the maximum degree. Our model showed that position 4 (Power Forward), and position 5 (Center) are two significant variables that effect opponents' shot results. So, NBA coaches should pay attention to train power forward and center players' defensive skills in daily training.

5. LIMITATIONS AND FUTURE IMPROVEMENTS

5.1 Limitations

- Data aspect

Our data originated from games in 2014-2015 season. So the conclusion may not be updated for now. And as discussed in Deliverable 1, our data has some problems and we lost some observations after processing the problems.

- Modeling aspect

We choose the logistic regression to fit our data. However, the performance of our model is not very satisfied. The prediction accuracy rate is lower than 60%. So our next step is to further adjust our model. In addition, there may exist some better deep learning algorithm for a better performance.

5.2 Further Solutions

We can determine which NBA defender has the best odds of a player not completing a shot successfully. Next, we could build a prototype based on the specific player's: 1) Height, 2) Weight, 3) Experience, and 4) BMI. We recommend that NBA scouts recruit defensive specialist who model such analyses. We can think further about such idea and develop a R code to do it.

6. APPENDIXES

- Appendix1: Variables Exploration Figures 1-8
- Appendix2: Rscript and Output