NBA Shot Probability Model

Younes Bensouda Mourri, Ankit Goyal, Eli Shayer
CS 229 | Stanford University

Introduction

We analyzed a collection of 632 NBA games to create a shot probability model for each player in the NBA. Each of our data sets consisted of a player's xy coordinates, as well as the ball's xyz coordinates, sampled every 40 milliseconds.

Using this data, we extracted features such as the shooter's distance from the basket, the angle of the shot, whether the shooter was on the home or away team, the shooter's distance to the nearest defender, how much time the shooter has had the ball for, and whether the shooter has dribbled or not.

We created a shot probability model using logistic regression and boosting, achieving 64% accuracy through the boosting method.

Objective

Utilize SportVU basketball player tracking data to create a shot probability model on the basis of player movement and interactions.

Data Set and Processing

The data came in the form of xyz position data of the ball and xy position of the 5 players on the court 25 times per second throughout each of the 632 games in our sample in the first half of the 2015-2016 season.

First we extended the position of the ball into the velocity and acceleration of the ball at all times throughout the game in each dimension. We then identified shots as those times the ball moved through the air with only the force of gravity and ended near the basketball rim. Among these shots, we classified shots that passed through the area directly below the center of the hoop within the next short period of time as made shots. In finding this response variable, we also identified the moment at which the shot reached the rim and the time the shot was released.

The features we extracted are the distance of the shooter to the hoop, the angle of the shot, whether the shooter played for the home or away team, the distance to the nearest defender, the length of time the shooter had the ball, and whether the shooter had dribbled since receiving the ball.

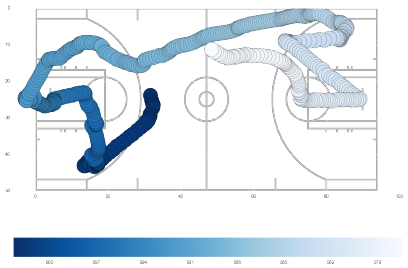
Methods

After creating the features, we used a logistic regression to predict whether a shot was made or not. We got all the shots attempted from the 632 games and used 70% of those shots to train our models. We then used them to predict on the remaining 30% of the data. After using logistic regression, we used Support Vector Machines and compared the two models.. Finally we tried boosting which provided us with the best results.

Analysis

Support Vector Machine performed worst. We got an 15% accuracy using the model. Following SVM, logistic regression performed slightly better with 35% accuracy. Finally, boosting performed best with 64% accuracy.

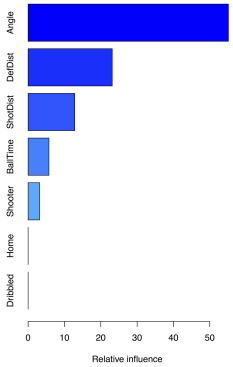
Our shot probability model did not perform as well as we desired. One reason for this low accuracy is that the labels of whether shots were successful were generated by our own algorithms, and have some discrepancies from reality. Additionally, we did not include all features relevant to shot probability, such as the velocity of the shooter.



A visualization of the movement of the ball in a single event in the SportVU data, where the ball moves from the dark blue to the light blue section during the event.

Results

Boosting relative influence graph



	Shot Made	Shot Missed
Predicted Made	0.160	0.358
Predicted Missed	0.130	0.353

Variable	Coefficient	p-value
Distance	−0.015	0.035
Home	0.079	0.191
Defender Distance	−0.008	0.641
Angle 0 to 15	1.047	0.004
Angle 15 to 30	1.008	0.005
Angle 30 to 45	0.882	0.015
Angle 45 to 60	0.809	0.028
Angle 60 to 75	0.585	0.113

Logistic regression accuracy and coefficients

References

1. Tavish Srivastava, Analytics Vidhya. .09.2011. Business Analytics R http: //www.analyticsvidhya.com/blog/2015/09/complete-guide-boosting-methods/
2. Stephen P. Borgatti. Centrality and network flow. Social Networks, 27(1): 55–71, jan 2005. ISSN 03788733. doi: 10.1016/j.socnet.2004.11.008. URL http://www.sciencedirect.com/science/article/pii/S0378873304000693.
3. Andrew Borrie, Gudberg K Jonsson, and Magnus S Magnusson. Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. Journal of Sports Sciences, 20(10):845–52, 2002. ISSN 0264-0414. doi: 10.1080/026404102320675675. URL http://www.ncbi.nlm.nih.gov/pubmed/12363299.