Using Spatio-Temporal Data To Create A Shot Probability Model

Eli Shayer, Ankit Goyal, Younes Bensouda Mourri

June 2, 2016

1 Introduction

Basketball is an invasion sport, which means that players move freely during every second of the game. In order for us to evaluate a player, it is highly beneficial to consider the movements of all the players on the basketball court. A more complete understanding of a player's performance can be achieved by taking into account the spatio-temporal considerations of movement and player interactions.

With basketball a lucrative and competitive sport in the US, teams have incentives to accurately project player quality. The NBA competition has become so fierce that minor details about players could have a dramatic impact on game results. For example, taking shots under pressure, from different angles, and at different times of the game are very hard to accurately assess qualitatively. Using statistical methods to assess these shots would help us identify the quality of players and would provide us with the ideal situations of taking shots. To do so, one would need to analyze every player at different points in time, and luckily we could do so by using the spatio-temporal data acquired by special techniques recently developed.

We sought to create a probability shot model in a basketball game. We started by getting a single NBA data set which breaks down every second into 25 moments in which every moment consists of the player's and the ball's locations. Based on this data, we created criteria such as the distance between the player with the ball and the closest defender, the ball's velocity, acceleration, maximal height, along with many others features to determine whether a shot was taken or not. Once we knew whether a shot was taken or not, we trained our model on 70% of the shots attempted in 632 basketball games and created a probability shot model that we tested on the remaining 30% of attempted shots. To focus on our analysis, we only considered jump shots, and excluded other shots such as alley oops and lay-ups.

2 Related Work

A few papers have been published on this field mainly dealing with analyzing specific players or the team as a whole. A common technique called Network Analyses turns teammates into nodes and passes into paths (or archs) thus creating a flow chart. Using these flow charts one could analyze the most frequent

paths that the ball went through. Based on this model one could mathematically justify why the triangle offense works and why the winning team tends to have more entropy. Topological depth, entropy, price of anarchy, and power law distributions are assigned to each player to assess outcome classification. Other techniques used are known as Intensity Matrices and Maps which transform the playing area into polar space and induce subdivisions in the space. This common technique uses Matrix factorization on the intensity matrices to produce a compact low-rank representation. It thus models shooting behaviors with the insight that similar types of players shoot from similar locations and then maps each type to an area within the court. Other papers were written on tactical group movement and how they affect the play. Special techniques were used to identify formations such as clustering coefficients and different forms of centrality.

3 Data and Processing

The data were obtained from public GitHub account that had scraped the publicly available SportVU player tracking data in basketball. This data contains the xy coordinate of the 10 players, 5 for each team, that are on the court, as well as the xyz coordinates of the ball 25 times per second. We transformed this raw data into a csv file with each row containing the location of the ball, and the location and identities of the 10 players on the court.

We then worked on processing this data into our response variable, the shots that were taken and whether they were successful. In order to identify when a shot was taken, we relied on the physics of the ball's flight through the air. We computed the position, velocity, and acceleration of the ball. We identified shots as moments in which the ball traveled through the air with no x acceleration or y accerleration, and z acceleration only due to gravity, with the ball's flight ending at the rim.

To determine whether a shot was successful or not, we check whether, in the fifth of a second immediately following the ball being in the area of the rim, the ball passes through the area immediately below the basketball hoop. If so, we mark the shot as a make.

Through manual checking of this procedure against the actual play-by-play of several games, we know that the process successfully captures a large majority of shots, and successfully classifies a large majority of the shots. There are nonetheless some errors, however, with about 10% error in each of the steps. This is because we attempt to identify shots on the basis of a definition approach. It would perhaps be more effective to manually label a certain number of games, and extrapolate from these labels with a machine learning approach.

Over the 632 games for which we had data, this approach identified 42,034 shots, of which 29.6% were classified as successful.

Next, we extracted a feature set from the data. These features 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.

4 Technical Approach

Model Selection: We randomly selected 70 percent of the shots to train logistic regression and SVM. We then tested the models on the remaining 30 percent. When we implemented support vector machine, we got 20 percent accuracy. On the other hand, when we used logistic regression, we got 50 percent accuracy.

Finally we used boosting and got 64 percent accuracy. To implement boosting, we evaluated, using resampling, the effect of model tuning parameters on performance. We then chose the optimal model across these parameters and then estimated the model performance. We used repeated training/test splits (4), with a 60 percent partition on the data. This algorithm predicted whether the shot would be successful with 64 percent accuracy.

5 Current Results and Analysis

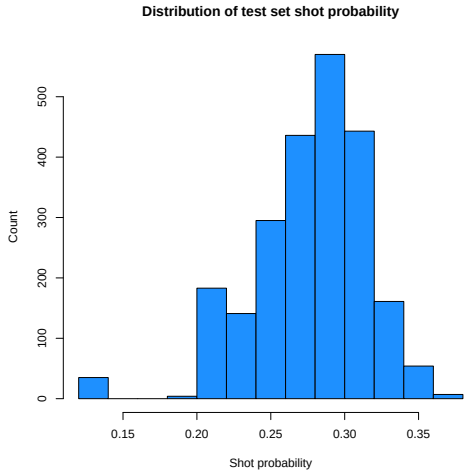
After training our data on logistic regression we found that logistic regression performed at a mediocre rate. Here are the classification results:

	Shot Made	Shot Missed
Predicted Made	0.189	0.394
Predicted Missed	0.106	0.310

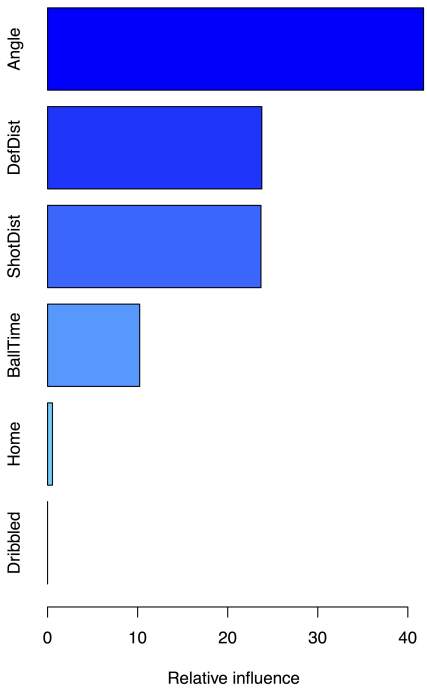
The following table gives the coefficients and p-values associated with each of the features included in the model:

Variable	Coefficient	p-value
Distance	−0.012	0.000
Home	−0.008	0.786
Defender Distance	−0.020	0.013
Time With Ball	−0.000	0.153
Dribbled	−0.000	0.994
Angle 0 to 15	−0.675	0.633
Angle 15 to 30	−0.692	0.625
Angle 30 to 45	−0.822	0.562
Angle 45 to 60	−0.978	0.490
Angle 60 to 75	−1.161	0.412

The following figure gives the distribution of projected shot probabilities on our test set.



The following figure indicates the relative influence of our various features on the boosting algorithm.



6 Conclusion

From the results we conclude that Adaptive boosting works best with such a data set. Since knowing the location of the player, his nearest defender, angle to

the basket, along with the many other features only provides us with a slightly better guess than random, it makes total sense why boosting might out perform logistic regression and SVM. From boosting's relative influence, or "weights" assigned per feature, we could see that the major components to predicting a shot are the angle to the basket, closest distance to the defender, and the ball's distance to the basket. Determining whether the player dribbled before or not was not of much importance.

7 Where To Go From Here

This project could be improved and extended in several ways. First, it would be highly beneficial to apply machine learning to the identification of shots and the classification of their success. This would require labeling a certain number of shots, and then using those labels to identify other shots.

We could also improve our model by adding additional features. These other features include defender angle, as well as defender distance and angle for the second nearest defender. A new feature that we think would be valuable is the velocity (speed and angle) of the shooter at the moment at which they take their shot. These are features that we learned are used by Second Spectrum, a sports analytics company focused on analysis of SportVU data.

Another area of expansion is into other elements of the game of basketball, such as rebounding and passing. Predicting the outcome of increasingly many elements of the game of basketball would build into a more comprehensive model of the game, that could be used to thoroughly evaluate basketball performance.

Finally, this type of analysis can be applied to more sports. The NFL has tracking chips in players' shoulder pads; many European soccer leagues have optical player tracking; the NHL has experimented with player and puck tracking technology. The methods used in analyzing basket ball player tracking data could also be applied to other sports to more thoroughly understand player performance.

8 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.