Team 187 (ATL/BOS) **CSE6262 Project Proposal**

Adam Warner, Youjian Chi, Calvin Leather, Robert Curran, Sanchari Roy

H1: What are you trying to do? Non-jargon: Understand how player statistics and position on the court impact the success of making a shot. Technical: Build a model to predict the probability of Shot Quality, or the likelihood of a shot being made by an average NBA player, and a web page visualization to enable a user to interact with the features of the model. Main approach: Build a model using scikit-learn, lme4, or another library to predict the probability of Shot Quality, the likelihood of a shot being made by an average NBA player, serve the model using PlumbR or Flask and engineer a web page visualization using d3.js to enable a user to interact with the features of the model.

H2: How is it done today; what are the limits of current practice? Sports data analysis has proven a powerful approach to learn from past games and optimize strategies for future games. Many applications of sports analytics have been researched in recent years, especially for NBA games. A recent study was done using play-by-play NBA data to optimize shot selection strategy, including how often to make 3-point vs. 2-point shots (Fichman & O'Brien, 2019; Chang, 2018). Another used gameplay data to choose between taking or passing on a shot, given time remaining on the shot clock (Goldman & Rao, 2017). These are relevant studies because they highlight shot type and shot clock as potentially important features in predicting shot success. An additional study centered around basketball shot factors that chose to consider include a study that used shot tracking data from the 2013-2014 NBA season to propose and illustrate "Effective Shot Quality", or the rate at which an average NBA player would make a shot given distance from the basket and distance of defender (Chang et al., 2014). This study provides evidence that distances are relevant for predicting shot likelihood, allowing us to design a spatial visualization around those features.

The ability of an individual to make a successful shot is a key component to winning a game. "A Shot Recommender System for NBA Coaches" used a recommender system to predict shot success with shot log data and underscores some of the challenges of using supervised learning, the main approach in our research. We also considered another study that analyzes the optimality of the shot distribution and which players on a team are contributing the most shots (Sandholtz et al., 2019). As our project approach seeks to analyze shot quality, shot location is likely to be a factor we need to take into consideration. Players must also determine which shot opportunities are likely to lead to success, and which should be passed as considered in "The Problem of Shot Selection in Basketball", which analyzed which shot opportunities were most likely to result in points scored and should be taken. Lastly, the fourth study to consider analyzed how team formations had an effect on getting shot takers open and its effect on shot quality (Lucey et al., 2014). Each of these papers brings to light factors we will consider when determining the features our model needs to consider. Our project seeks to improve upon these studies by creating a real-time, interactive visualization based on player identity and location on the court.

Spatial variables also play a key role when considering analytics basketball models. One such study used player data to build an economic model of how a player's actions will change the Expected Value of Possession. The model assigns a probability and value to each decision (pass, shoot, drive) and the resulting EPV after the decision using Markov chains (Cervone et al., 2014). This article provides perspective on creating a framework to determine how changes in spatial attributes affect outcomes in an NBA game. Another study identified the "best" shooters by identifying the range of court area a player effectively shoots from (Goldsberry, 2012). This study addresses "who is the best shooter?", rather than "how does the average NBA shooter perform?", making it less useful to the project. Two additional papers we reviewed include one that measures shot success using multinomial regression using spatial and shot type variables which provides a specific example of how to create a model to fit our data (Erčulj & Štrumbelj, 2015) and another that attempts to answer the "hot-hand" fallacy, if a player is more likely to make a shot given a previous miss or a previous make (Chang, 2018). These two studies can be integrated into our research because they examine how shot type and previous shot attempts affect the probability of a successful attempt. A fourth paper provides an interesting example of how environment and time affects the score outcome throughout the game. Fitting a power law model attempts to find the scoring rate for a particular game (Ribeiro, 2016). All three studies fail to recognize that there is more than a binomial outcome to a shot attempt.

Lastly, we assessed papers that surveyed other disciplines for spatial analysis approaches. Chi and Zhu (2007) assessed modeling techniques including conventional OLS, OLS with adjacency-based autoregression, and spatial autoregressive moving average (SARMA), and found SARMA had the best AIC/BIC with the least spatial error autocorrelation. This study also provides factors to measure spatial model fit. A second study we reviewed applied several advanced methods, including wavelet regression, and Bayesian approaches, finding GLM convergence issues, but that the Bayesian autoregressive approach yielded the best data fit (C. M. Beale et al., 2010). This paper is relevant as it suggests Bayesian approaches. Lastly, we reviewed a methods paper for a Bayesian spatial modeling system (Lee, 2013). They identify mechanisms to speed up spatial MCMC sampling, and benchmark results against OLS methods. This paper provides a specific method to use to analyze our data. Each of these papers have the shortcoming that they analyze data with pre-existing spatial discretization (e.g. county boundaries), whereas basketball only has the overly broad 2- and 3-point zones, meaning we will have to identify a novel strategy to discretize the space.

H3: What's new in your approach? Why will it be successful? Our innovation will allow users to visually interact with a model by manipulating nodes around a basketball court to better understand how spatial features affect the quality of a shot. This will allow users to build intuition regarding the marginal changes in the spatial features', which can be difficult to do with the current state of literature and research. Our model will innovate by using logistic regression and regression trees. We expect to branch based on 2- vs 3-point attempts, low post vs jump shots and time remaining in the shot-clock or game clock.

H4: Who cares? NBA teams including coaches, players, analysts and owners, as well as sports journalists and NBA fans.

H5: If you're successful, what difference and impact will it make, and how do you measure them? If successful, our research will provide a helpful tool that NBA teams and fans can leverage to simulate shots and formulate game strategy. Its success could be measured by web traffic of the webpage where we publish our research, citation count if published, interviews with NBA analysts who use our tools, and reactions from online communities.

H6: What are the risks and payoffs? The complexity of the proposed visualization and analytics present a risk in execution. However, the payoff will be an engaging, interactive webpage.

H7: How much will it cost? All datasets and tools to be used are free.

H8: How long will it take? This project is estimated to take between 170-300 hours. Adam will focus on data importing, joining and cleaning, (30-60 hours). Youjian and Sanchari will focus on analytics (30-60 hours each). Calvin and Robert will focus on visualization using D3 (40-60 hours each). We have all participated in the literature survey and composition of Proposal document; Youjian produced the presentation slides, and Youjian, Robert, Calvin and Adam worked together to produce the presentation video.

H9: Mid/final exams? The mid-term deliverable of our project would include an initial predictive model built on clean dataset with all necessary features, as well as a visualization prototype based on dummy prediction data. The final deliverable would be the final application with integrated prediction and visualization.

Distribution of Team Effort: All team members have contributed a similar amount of effort.

References

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. Ecology Letters, 13(2), 246–264. doi: 10.1111/j.1461-0248.2009.01422.x

Cervone D, D'Amour A, Bornn L, Goldsberry K. (2014). POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data. *MIT Sloan Sports Analytics Conference 2014.*

Chang, S.-C. (2018). Capability and opportunity in hot shooting performance: Evidence from top-scoring NBA leaders. Plos One, 13(2). doi: 10.1371/journal.pone.0179154

Chang, Y.-H., Maheswaran, R., Kwok, S., Levy, T., Wexler, A., Squire, K., & Su, J. (2014). Quantifying Shot Quality in the NBA. *MIT Sloan Sports Analytics Conference 2014.*

Chi, G., & Zhu, J. (2007). Spatial Regression Models for Demographic Analysis. Population Research and Policy Review, 27(1), 17–42. doi: 10.1007/s11113-007-9051-8

Erčulj, F., & Štrumbelj, E. (2015). Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. Plos One, 10(6). doi: 10.1371/journal.pone.0128885

Fichman, M., & O'Brien, J. R. (2019). Optimal shot selection strategies for the NBA. *Journal of Quantitative Analysis in Sports*, *15*(3), 203–211. doi: 10.1515/jqas-2017-0113

Goldman, M., & Rao, J. M. (2017). Optimal stopping in the NBA: Sequential search and the shot clock. *Journal of Economic Behavior & Organization, 136,* 107–124. doi: 10.1016/j.jebo.2017.02.012

Goldsberry, K. (2012). CourtVision : New Visual and Spatial Analytics for the NBA. *MIT Sloan Sports Analytics Conference 2012.*

Lee, D. (2013). CARBayes: AnRPackage for Bayesian Spatial Modeling with Conditional Autoregressive Priors. Journal of Statistical Software, 55(13). doi: 10.18637/jss.v055.i13

Lucey, Patrick, et al. (2014). "'How to Get an Open Shot': Analyzing Team Movement in Basketball Using Tracking Data." MIT Sloan Sport Analytics Conference, http://www.iainm.com/publications/Lucey2014-How/paper.pdf.

Ribeiro, H. V., Mukherjee, S., & Zeng, X. H. T. (2016). The Advantage of Playing Home in NBA: Microscopic, Team-Specific and Evolving Features. Plos One, 11(3). doi: 10.1371/journal.pone.0152440

Sandholtz, Nathan, et al.(2019). "Chuckers: Measuring Lineup Shot Distribution Optimality Using Spatial Allocative Efficiency Models." *MIT Sloan Sports Analytics Conference*, http://www.lukebornn.com/papers/sandholtz_sloan_2019.pdf.

Skinner, Brian. (2012). "The Problem of Shot Selection in Basketball." *PLoS ONE*, vol. 7, no. 1, 2012, doi:10.1371/journal.pone.0030776.

Wright, R., Silva, J., & Kaynar-Kabul, I. (2016). Shot recommender system for nba coaches. *KDD Workshop on Large Scale Sports Analytics*. doi: 10.475/123 4