Team 187 **Progress Report**

Adam Warner, Calvin Leather, Robert Curran

Plan of Activities:

Original Plan of Activities: 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. By the progress report, we hope to have an end to end working prototype including a web browser based visualization in d3, Javascript/HTML/CSS, experiments with predictive models, cleaned data and a solution for interacting with a trained model from the web browser.

Revised Plan of Activities: Our group has been reduced to three members over the course of the semester. At this time, we feel that we have achieved our midterm goal of a working end to end prototype. Our plan of activities going forward: Calvin will focus on the web browser based visualization, Adam will focus on finding and cleaning additional features for the same data as well as contributing to experimentation with the predictive model, and Robert will focus on experimenting with and refining the predictive model.

Introduction - Motivation:

The motivation behind our project is to provide a way for users to interact with a model that predicts the likelihood of an average NBA player making a shot. There is plenty of literature on prediction and modeling of NBA games, and a number of static visualizations of the likelihood of a shot success or the effects of a decision to pass on the outcome of a possession, but these solutions all lack one thing: user interaction.

Problem Definition:

We want to allow a user to interact visually with a predictive model trained on features of more than 270,000 shots from an NBA season in a highly accessible way. Allowing a user to interact with the model in real time by moving relevant players around on an NBA court will facilitate intuition building; compared to static visualizations, the user can move players around the court, update values of non-spatial model features, and receive a probabilistic output as a measure of shot quality. This functionality is of interest to many groups including basketball coaches trying to design and implement new plays, managers making player roster decisions, basketball trainers trying to determine which player skills are most important to improve, and statistically infatuated basketball fans that want an entertaining and rigorous basketball experience. Accessibility is another issue with current approaches - we want to provide a solution that does not require a background in programming or statistics, the installation of additional software or exploration of scientific papers to build a probabilistic intuition of the game of basketball.

Survey:

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 S-C, 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.

Adam Warner, Calvin Leather, Robert Curran

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 most to the lost 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 take into account when determining the features our model needs to consider. Our project seeks to improve upon these studies by creating an real-time, interactive visualization based on player identity and location in the field.

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 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 some additional papers that surveyed other disciplines for spatial analysis approaches. Chi and Zhu (2007) assessed modeling techniques including conventional ordinary least squares (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 generalized linear model (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 markov chain monte carlo (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 discretizations (e.g. county boundaries), whereas basketball only has the overly-broad 1, 2 and 3 point zones, meaning we will have to identify a novel strategy to discretize the space.

Proposed Method:

Adam Warner, Calvin Leather, Robert Curran

Intuition: We believed our visualization would be better than the state of the art because we knew that static, non-interactive visualizations could impede learning. Using an interactive visualization, the end user can easily run experiments of different basketball situations, fostering an intuition of the impact of various features on a shot's outcome. We believed our model would provide value over modern shot prediction models for two reasons. First, our model would not only use cutting edge player location data, but would also be joined with features of players such as height and wingspan. Secondly, we knew we had a number of modeling approaches in our tool bet that we could experiment with to provide a cutting edge prediction.

Description of Approaches: The original data set contains 21 features of approximately 120,000 shots taken during the 2014-2015 NBA season. This data was formatted as a csv for download. There are two popular open source packages that pull in NBA data, ballR and statsnba for R and Python, respectively. ballR allowed for the pulling of the shot location for the 2014-2015 by player id, which is a unique key that is standardized across many NBA data sets. To join the seasons worth of x, y shot location data generated by the ballR package we had to create our own unique shot key id. To create a unique shot id both datasets were grouped by the player id and game id and then sorted by the game clock in descending order. We then created a running tally of the shot number allowing us to join the datasets on the combination of game id, player id and shot number. Player attributes such as player height and weight were gathered using the statsnba package in Python, but not all player attribute data was successfully pulled from the API and needed to be manually added. To increase the size of our dataset, we added shots from the 2015-2016 season to make our final data set 270,000 rows and 49 columns.

To build a model to predict the likelihood of a shot going in we experimented with a few models including logistic regression, decision trees, a random forest, and boosting algorithms. Logistic regression was initially proposed due to its probabilistic output. **Under construction

To create an accessible visualization, we decided to build the visualization using JavaScript, HTML and CSS, along with the D3 and JQuery libraries. The foundation of the user interface is the lines of a basketball court expressed as GeoJSON sourced from a publicly-available GeoJSON repository. This GeoJSON set contains the courts sidelines, baseline, the key area, half court line, the location of the basket over the court and various hash marks. On top of the map of the court are two moveable circles that represent an offensive player and a defensive player. There are lines between the two players and the offender player and the basket that are redrawn as the circles are moved. We theorized that the distances between players and the basket would be relevant in our model, so we decided that they would need to be visualized. **Under construction

To allow a user to interact with our model with just a web browser, we experimented with a few solutions, but decided to use the Pyodide package. This package makes a Python interpreter available in the browser by compiling key scientific and statistical packages such as numpy, scipy, statsmodels and sklearn into WebAssembly. We used the Pickle package to serialize trained models and load them into the browser when the page is opened. As a user moves either player on the page, or when changing one of the non-spatial feature values, the model's parameters are passed to our model's predict function, returning the measure of shot quality predicted by the model in real time.

List of Innovations: 1. Novel interactive spatial visualization to convey the likelihood of successfully making a shot. 2. Use of Pyodide and WebAssembly to allow interaction with trained models using only a web browser. 3. Multiple models trained on a combination of datasets including shot by shot information and player's physical attributes. 4. An ultra-accessible interactive visualization enabled by Pyodide and GitHub Pages.

Experiments/Evaluation:

Testbed: We used packages from both R and Python for our analytics and modeling experimentation including scikit-learn, statsmodels, xgboost and glm. For data manipulation and cleaning, we used

Pandas, NumPy dplyr, and data.table. The subset of features includes shot distance from basket, closest defender distance, shot clock or game clock remaining, shot count of shot by player and game, amount of time a player held the ball before shooting, and if the shot was a two or three point attempt. We leveraged glm, sklearn and statsmodels to build models and measured model accuracy using random 80/20 splitting of train and test data, and used McFadden's R2 to measure logistic model quality. Visualizations were tested with an http server, web browser and JS/HTML/CSS stack. **Under construction

Experimental Questions: How can we facilitate a user's interaction with a trained model in R or Python? Which models should we experiment with and which models will provide the most accurate models? Which set of features and interactions of features will result in a simple but accurate predictive model? How can we effectively visualize the findings of our spatial based model development?

Details of Experiments: To find a solution to allowing a user to interact with a trained model from a web browser, we experimented with two packages - PlumbR and Pyodide. PlumbR allows you to expose methods in R as http endpoints. For example, we could expose a method that takes the model parameter and calls predict on a trained model, returning the model's prediction in an http response. Using the PlumbR solution would require the end user to run to install R on their machine, but would allow us to access the functionality of an R development environment. Pyodide is a package that makes a Python interpreter available in the web browser through the use of WebAssembly and exposes important scientific and statistical packages available, such as numpy, scipy, sklearn and statsmodels. We experimented by loading relevant statistical packages into our web browser embedded Python environment at page start up, loading a Pickled model into the environment as a global variable, and then calling predict on that model with parameters arranged in JavaScript as the user interacted with the web page. This proved to be accessible, practical and efficient; it took less than a minute for the page to load the Python environment, packages and model and provided low-latency model interaction.

In developing an unbiased and consistent predictive model, we experimented with a few types of models: logistic regression, decision tree, random forest and principal component analysis. After experimenting using a logistic regression model with various sets of features and interaction terms, forward selection, and linear and non-linear models, it was clear that finding a highly explanatory model would be challenging. Although some of our logistic models had accuracies in the range of 50-60%, we were unable to arrive at a model that produced McFadden's Pseudo R2 of greater than 0.1. This implied that our feature selection did not provide significant value over the null intercept only model.

Decision trees, random forests, PCA + logistic regression. **Under construction

To evaluate end user experience, we used a convenience sample of basketball fans (n=5) from the project team's friends and colleagues and provided a link to the visualization system on github. We provided a simple questionnaire - Were you able to load and interact with the visualizations? Were there any technical issues? Did you learn anything new from the visualization? How do you think fantasy sports fans or team managers could use this information to alter decisions? What questions would you be interested in answering about shot success that you were unable to answer? **Under construction

Conclusion:

**Under construction

Distribution of Team Effort:

All team members have contributed a similar amount of effort.

Bibliography:

[E1] 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
[E2] 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
[E3] 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
[C1] G. Chi and J. Zhu, "Spatial Regression Models for Demographic Analysis," Population Research and Policy Review, vol. 27, no. 1, pp. 17–42, 2007.
[C2] C. M. Beale, J. J. Lennon, J. M. Yearsley, M. J. Brewer, and D. A. Elston, "Regression analysis of spatial data," Ecology Letters, vol. 13, no. 2, pp. 246–264, 2010.
[C3] Lee, D. (2013) CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. Journal of Statistical Software, 55(13), pp. 1-24.
[R1] 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.*
[R2] Goldsberry, K. (2012). CourtVision : New Visual and Spatial Analytics for the NBA. *MIT Sloan Sports Analytics Conference 2012.*
[R3] 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.*
[A1] Erčulj, F., & Štrumbelj, E. (2015). Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. *PLoS One*.
[A2] Chang S-C (2018). Capability and opportunity in hot shooting performance: Evidence from top-scoring NBA leaders. *PLoS ONE 13(2): e0179154*. https://doi.org/10.1371/journal.pone.0179154
[A3] Ribeiro HV, Mukherjee S, Zeng XHT (2016). The Advantage of Playing Home in NBA: Microscopic, Team-Specific and Evolving Features. *PLoS ONE 11(3): e0152440*. https://doi.org/10.1371/journal.pone.0152440
[S1] 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.
[S2] Skinner, Brian. (2012). "The Problem of Shot Selection in Basketball." *PLoS ONE*, vol. 7, no. 1, 2012, doi:10.1371/journal.pone.0030776.
[S3] 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.

**Under Construction, need to add citations for: numpy, sklearn, pandas, statsmodels, glm, ballR, statsNba, dypler, data.table, xgboost, R, Python, Team Carto GeoJSON, Pyodide, PlumbR