Executive Report - Shot Prediction with Machine Learning Models

Grant Culbertson and Andrew Mayer

Background:

We were interested in exploring a relatively under-researched topic. The academic papers we found were doing similar things to us, but the majority of the papers only used one simple dataset (the kaggle one we also used) with shot information like distance, shot type, etc.. We wanted to see if we could improve upon these research papers by combining multiple datasets to create a shot log with more variables, and most importantly a shot log that included detailed defender information. We felt that including greater defender and shot information was an intuitive and straightforward way to get more accuracy out of shot prediction models. The datasets we combined were a kaggle shot log, an nba_api shot log, and player info from nba_api all for the 2014-15 season.

Beyond just predicting whether a shot goes in, we also wanted to give our model some real world use, so we developed a function in R to create a "scouting report" for guarding a given player. With models that could take both defender information and shot information it seemed an obvious next step to put the model to work predicting what defenders would hold a player to a lowest predicted field goal percentage.

In our modeling we chose to look specifically at LeBron James, Steph Curry, and Marc Gasol because in the 2014-15 season they were all players that ranked decently high in MVP voting as well as being players that cover a range of positions (PG, PF, C).

Methods:

- **Support Vector Machines (SVMs)**
  - The first machine learning modeling method we explored was Support Vector Machines. This supervised machine learning method is useful in binary classification, for example a made or missed shot. After fitting and tuning these models for our players of interests we were able to improve upon the research and get accuracies between 62%-67% when tested with testing data.
- **Gradient Boosting (XGboost):**
  - We chose to use xgboost decision tree models because xgboost is generally one of the best models when it comes to gradient boosting, and it was the type of model that had the greatest accuracy in the academic papers we read. The models we created for all 3 players had an RMSE of ~.36. Our LeBron model was 61% accurate, with a 95% confidence interval of 56% - 64%. Both other models had similar performance (R was crashing when trying to find their accuracy).

Conclusion:

We were able to replicate and in some cases improve on the research papers we used as our baseline. While the extra factors included such as the type of shot and defender information were not the most important factor in any of our models, they did seem to lead to models with higher predictive accuracy (seen more with the SVM models). While the academic papers generally found XGboost models to be most accurate, that was not the case for us. This is most likely due to limits in computational power to fully tune the XGboost models. Our "scouting reports" provided somewhat useful information, but required some intuition about basketball to be interpreted most effectively. In the future, the use of individual matchup information rather than just shot logs would probably provide "scouting reports" with results that make greater sense, or are closer to what we hoped the function would do–predict the best defender for a whole game, not just in the moment someone is shooting.