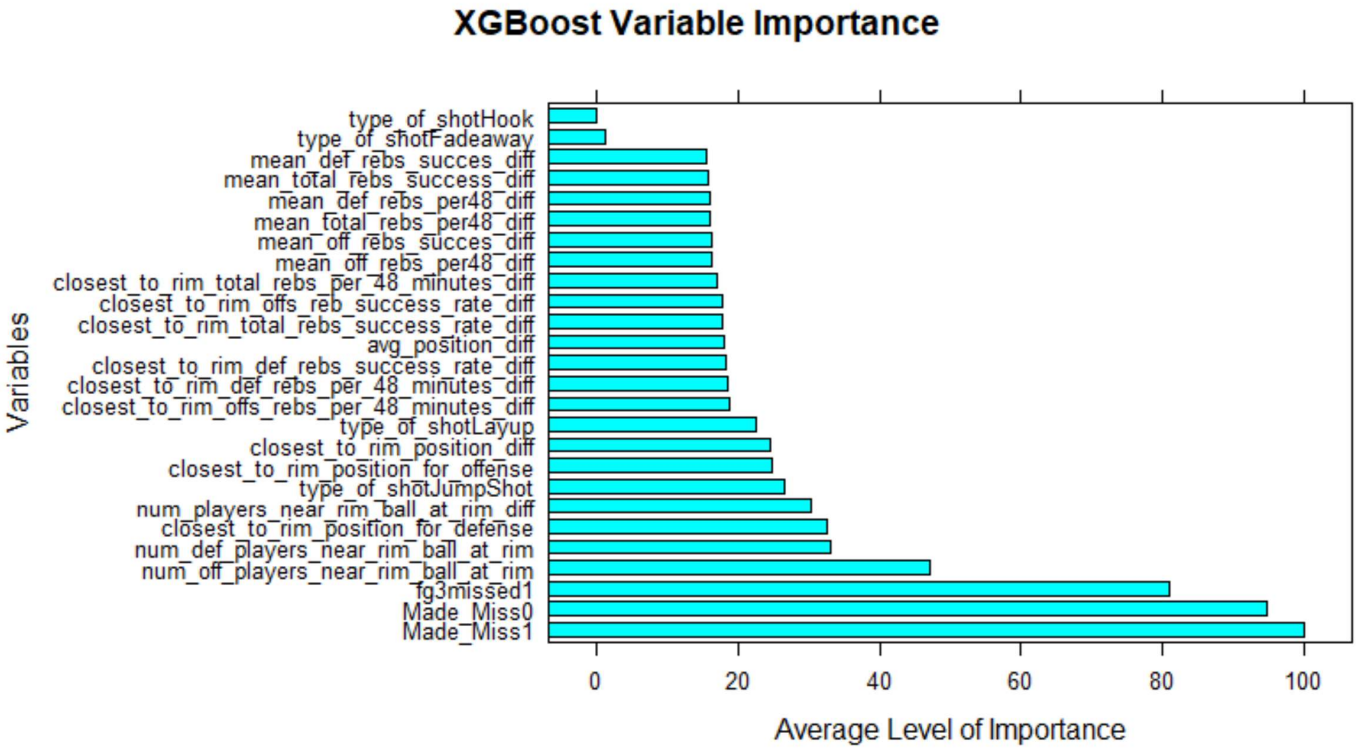
Basketball Analytics Research Project

1) After reading in the data, I performed some data inspection checks to find issues with missing values and class imbalances. For the training set, the percentage of rebounds collected after each shot by the offensive team only occurs about 14% of the time. Since the training data is large, I felt it appropriate to omit rows with missing player coordinates or missing shot labels to eventually have a data set for modeling with no missing values.

For data engineering, I used the positions and rebounding stats data to create individual player level stats such as rebounds per 48 minutes (off, def, total), rebound success rate (off, def, total), and position. I will use these stats as an aggregate (weighted by minutes) to describe the rebounding characteristics of the offensive team versus the defensive team at each shot. Next, I decided to focus on only the player locations at the time the ball hits the rim and not at the time of the shot in order to get a more accurate description of player location at the time of rebounding after they get in position to rebound. For each team (offense and defense), I calculated the number of players that are within a six feet radius from the offensive basket as well as the rebounding statistics and position for the player who is closest in Euclidean distance to the rim. Lastly, I differenced the data by team (offensive – defense) and will use those variables as inputs to the models. A quick note to mention is whether a player makes a shot is the most important predictor since rebounding is not possible on made shots. The goal is to find some subset of feature engineered variables that can help explain some additional variation in offensive rebounds that shot making cannot explain.

To address the class imbalances, I used a random sampling technique called SMOTE to rebalance the data set by random sampling less observations from the majority class and synthetically creating more samples representative of the minority class. I investigated four different models (Penalized Logistic Regression, XGBoost, Random Forests, and Stochastic Gradient Boost). Due to runtime issues, I subsetted the overall data to 100,000 rows and created a 70/30 train, test split from those rows. For each model, I used 5-fold cross validation on the training split for parameter tuning and estimating log-loss error on unobserved data and made predictions on the test split to evaluate the models. I chose the model for which log-loss was minimized on the unobserved 30% testing set and made predictions from that model on the testing set given (removed rows for which player coordinates were missing).

2)



2) Above, is the variable importance chart from the chosen model, xgboost. This horizontal bar chart ranks in order from lowest to highest the influential rank of each variable on the model. As stated before, the biggest predictor in whether an offensive player snatches the rebound for any shot is whether the offensive team made the shot. This makes sense as hitting a shot makes it impossible for any rebound to occur. Three-point shots have a lower make percentage than two-point attempts, so we also expect three-point misses to yield more offensive rebounds than missed two-point shots. Some of the other important predictors are number of players for each team within a six feet radius of the basket, aggregated player statistics for the closest offensive and defensive player to the basket, and layups. In other words, offensive rebounds are influenced by how many offensive players crash the boards and which ones closest to the basket are skilled rebounders (usually positions 4 and 5). Also, offensive players have a better chance of rebounding their own missed layups since they are right under the basket and have better intuition on the trajectory of their shot compared to defenders. These variables all make intuitive sense on the success rate of offensive rebounds and as we move along the variable list from bottom to top, we can make intuitive sense based off basketball knowledge on why they influence rebounding. Lastly, we observe team-level rebounding statistics (offense vs defense teams) as having an importance rank above zero.

3) Dear Head Coach,

I am one of the intern analysts on the team and have been assigned to investigate a basketball related question of interest for you. The project of interest is to investigate what factors influence the ability to grab offensive rebounds and what can help us predict an increase in offensive rebounding chances. To answer this question, I looked at lots of data of individual shots from past games and used information such as player location on the floor, player rebounding stats, and shot information to help predict or explain successfully obtaining offensive rebounds. I used those factors mentioned as inputs to a model that predicts with over 80% accuracy the chances of securing an offensive rebound at any possession. There are some key findings that can help us understand offensive rebounding.

1) Shot Making:
    - The biggest predictor of being able to grab more offensive rebounds is having more chances to grab offensive rebounds. If we have players on the floor that typically miss more shots than others, then we can crash the boards since we expect more opportunities to grab rebounds.
2) 3-pt Shooting:
    - Since 3-pt shooting has the lowest shooting percentage among shot types, we expect more misses when we take more 3-pt shots which also yields to more chances of securing offensive rebounds. Teams that rank high in 3-pt shooting attempts should expect more opportunities to grab offensive rebounds.
3) Player Positioning:
    - It's no surprise that the more players we have crashing the boards, the better chance we have of securing the rebound. That is one also one of the largest factors in being able to secure offensive rebounds. More specifically, having more of our players near the rim after the shot is up than the opponents will yield to more offensive rebounds. Also, having the closest player to the rim be a big man (PF or C) will yield to more offensive rebounds.

Depending on our rotation and opponents' strengths, there may be moments within a game that would be beneficial for us to focus on offensive rebounding to score off second chance opportunities.

Ruslan Davtian