Sriram Ravula (sr39533)

Varun Sridhar (vks377)

Ramakrishna Sai Annaluru (rsa662)

EE461P Project Proposal

We propose to use a dataset of NBA field goal (shot) attempts to model the probability that any shot will go in. Shot quality is a very important metric in basketball, as coaches and players always try to find the best scenario possible to have an accurate shot at the basket. Factors like the closest defender’s distance to the shooter, time remaining on the shot clock, current quarter of the game, and shot distance all weigh significantly on whether or not a field goal attempt is successful. While some situational factors have an intuitive effect on shot outcome (e.g. if a defender is closer to a shooter, then a shot attempt has a lower chance of success), some other factors like time remaining in the quarter have a less clear impact on a shot. Determining which factors positively or negatively affect shot success would be very useful in improving the quality of shots attempted by NBA players.

Our group intends to use a dataset of 2014-2015 NBA season shot attempts to train, validate, and test our model. We have found a dataset from Kaggle which lists 128,070 shot attempts from the 2014-2015 NBA season [1]. The set describes 21 features for each shot including player name, shot distance, distance to closest defender, time left in the quarter, and number of dribbles before shooting. Our intended target/prediction value is shot success which is a binary categorical variable: 0 is a miss and 1 a make. We can also add other features that may be relevant like player field goal percentage (i.e. proportion of shot attempts that are successful) from each distance range for the prior season (e.g. a player shoots 55% when 0-3 feet from the goal), player height in inches, and player free throw percentage for the prior season[2]. We will have to pre-process the features extensively to mitigate collinearities between features and reduce the amount of predictors our model will consider.

One of the problems we will have to address is the fact that our modeled output is based on a binary target, shot success, whereas the regression models generally model continuous outputs. The data are more suited to classification (i.e. will a shot go in or not) than to regression. However, it is more useful to know the probability that a shot will go in rather than simply a binary prediction of whether it will go in. The problem with using regression is that output values can be lower than 0 or greater than 1, so we must find some method to normalize outputs to the range [0,1]. One method to constrain the outputs is to use a logistic regression model, which maps input predictors to an output probability by using a linear combination of the features as an input to a logistic function [3]. Another method we may consider is the probit regression, which is similar to logistic regression in that it maps inputs to a probability, but uses a gaussian distribution instead of a logistic one [3]. Our group will have to try several different methods to compare test errors and qualitative differences between the models, and select the best one.

To summarize, we want to predict the probability that an NBA player will make a shot attempt given predictors like shot distance and player field goal percentage. We plan to fit a regression model to our data, but will have to research further to find a suitable way to predict shot probability. We believe that this project will be useful in identifying the factors that lead to high-quality shot attempts for NBA players.

**REFERENCES**

[1] https://www.kaggle.com/dansbecker/nba-shot-logs

[2] https://www.basketball-reference.com/

[3] http://courses.umass.edu/pubp608/lectures/l22-2.pdf