Colin Menz, Shaoyi Zhang, Emily Zheng
Team: Lakers
PSTAT 231, Spring 2016

Project Proposal

Question: We want to predict whether Kobe Bryant successfully makes the shot.

Dataset: The dataset has every shot attempted in Kobe Bryant's NBA career. It contains the location and circumstances of each goal attempted, as well as whether or not the shot was successful (shot_made_flag). In 5,000 random records the shot_made_flag is removed, which creates the test set. The remaining records make up the training set. This dataset contains 30,697 records of 25 variables.

Variables:

- Action_type: factor, the action performed during the shot
- Combined_shot_type: factor, the type of shot
- Game_event_id: numeric, corresponds to event within each game
- Game_id: numeric, corresponds to which game the shot was attempted in
- Lat: numeric, latitude
- Loc_x: numeric, position on the court horizontally
- Loc_y: numeric, position on the court vertically
- Lon: numeric, longitude
- Minutes_remaining: numeric, How many minutes are there left in the period
- Period: factor, description of which part of the game(goes up to 7, meaning 3 overtime)
- Playoffs: factor, description of whether the game is a playoff game
- Season: factor, a more general description of the date
- Seconds_remaining: numeric, How many seconds on the shot clock
- Shot_distance: numeric, the distance between shot location and baseline
- Shot_made_flag (this is what you are predicting): factor, 1 for shot made, 0 for shot missed
- Shot_type: factor, 3-point shot or 2-point shot
- Shot_zone_area: factor, describes general area of the court the shot is made from
- Shot_zone_basic: factor, description of shot zones on the court
- Shot_zone_range: factor, levels of distance from the basket
- Team_id: factor, Team ID that Kobe Bryant is playing for
- Team_name: factor, Team that Kobe Bryant is playing for
- Game_date: factor, day of the game
- Matchup: factor, matchup of teams playing
- Opponent: factor, Opposing team abbreviation
- Shot_id: numeric, Number of the shot taken

Techniques: To identify significant variables we will use clustering, principal component analysis, attribute-focusing, and random forests. Logistic regression can be used to give the probability of a shot's success.

Comments/Concerns: We will add or remove variables because the existing variables might be highly correlated or will benefit from using an indicator variable.

We are concerned about predicting shots with very little prior data. There are variables not addressed in this data set such as personal life and injuries.