Shannia Fu
Professor Albert Y. Kim
May 23, 2017
Final Project Report

The competition I ended up choosing was Kobe Bryant Shot Selection, which is a competition on predicting whether Kobe Bryant makes a shot based on 25 variables. While some of the variables are categorical, the ones I ultimately used in my final version were all numerical. I mutated the shot types and the times to be more numerical predictors.

I briefly considered using regular regression, but decided ultimately on logistic regression. It seemed the most appropriate to predict the probability of a shot, and then assigning an outcome based on a magical determining value (p*). It was also pretty reminiscent of what was given in Problem Set 7, which was a logistic regression predicting survival of passengers on Titanic.

The whole analysis was based on finding the optimal p* value. In the cross validation portion of the code, I split up the data into 1:4 parts test:train data (5-fold CV), and created a model based on the training data. There was also a nested for loop of p* values to use to determine whether to predict the outcome as 1 (shot made) or 0 (shot not made), based on whether the predicted probability was higher than p*. The range of possible p* values actually depended on how the model was trained; it didn't seem necessary to test any p* values that were out of range. The actual cross validation came in when the p* value was compared to the correct outcome in the pseudo test set.

Based on the nested CV, the approximate optimal p* value was ~0.554, which gave the lowest error (false negatives + false positives) compared to every other possible p* value. I then set the model back to be trained from the full train set, and assigned predictions to the test set using only the optimal p* value. My score, based on the Log Loss method on Kaggle, was 13.51 (which put me on the lower half of the leaderboard).