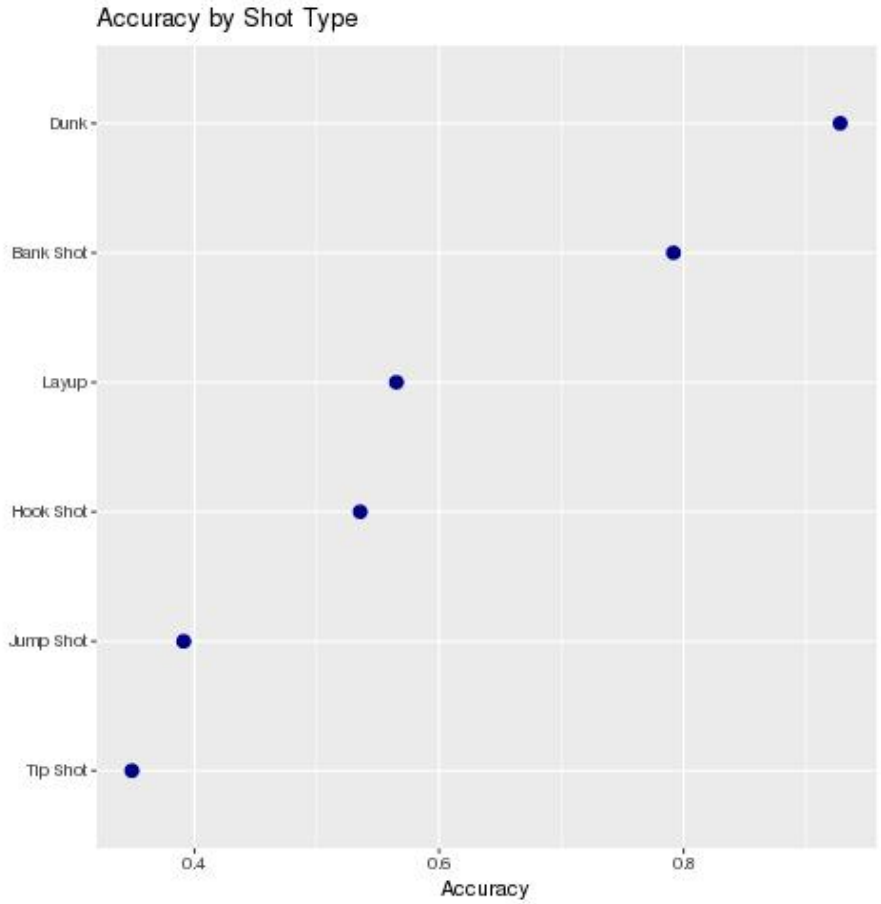
Harrison Marick, Meredith Manley, Christien Wright
STAT 495 Final Project

Executive Summary of Kobe Bryant's Shot Selection

For our project, we decided to participate in the Kobe Bryant Shot Selection Kaggle competition. We were given a dataset consisting of each of the 30,697 shots Kobe Bryant took in his 20 year NBA career. Our task was to accurately predict the outcome, make or miss, of 5,000 randomly selected shots over the course of Bryant's career. In this project, we examined the ability of three different strategies to predict shot success on the test set: K-NN, Logistic Regression, and CART.



Exploratory Analysis

The predictor we found most useful was the "combined_shot_type" variable. Notice in the graph to the right that there is clear shot type effect, as dunks and bank shots are markedly more accurate than jump shots and tip shots.

K-NN

In our analysis of potential predictors for the K-NN model, we quickly removed categorical variables from consideration, since creating a distance metric for categorical variables is difficult when the categorical variables are not ordinal. Ultimately, we settled on the x and y locations of the shots, distance, period, and the season as our predictor variables. In order to optimize our selection for K, we conducted 10-fold cross-validation. Since binary log-loss was the score Kaggle gave us, this is also what we used to measure cross-validation performance. We found that K=250 provided us with the lowest binary log-loss. After predicting using K=250 on the test set, we were left with a log-loss of 0.74653, which put us in the middle of the leaderboard on Kaggle, as the best score was approximately 0.56. The major pitfall of this method was that we did not use the "combined_shot_type" categorical predictor, which proved to be useful in our other models.

Regularized Logistic Regression

Our initial LASSO model formula was composed of the nine variables we felt were representative of the data as to avoid any redundancies that might result if all variables

were to be included. The main advantage of this model was the fact that variable selection is embedded in the process of model development. The variables that contribute most significantly to the model are those whose $\hat{\beta}_i$ remains positive the longest as our constraints on model complexity increase (λ increases). Through the use of cross-validation, we computed the value of λ, which outputs the minimum binomial deviance. We then proceeded to use the λ, which yields a binomial deviance within one error from the minimal binomial deviance to preserve model simplicity. We found that the most important variables were "combined_shot_type Dunk", "combined_shot_type Jump Shot", and "shot_distance" in order of importance. The Kaggle-reported binary classifier log-loss value of this model was 0.65159. Ultimately, our logistic model actually performed better on the Kaggle test set than it did on the given training set. As was mentioned above the top Kaggle score was 0.56 and thus we are confident in the accuracy and strength of this logistic regression model with L1 regularization.

CART

Kaggle asks for $\hat{p}$ values instead of binary values, so we built a regression tree. Given our exploratory analysis, it was clear "combined_shot_type", "shot_distance", "season", and "period" were most useful in our regression tree. CART is greedy, in that is always branches according to the variable that provides the most information gained at each moment. Interestingly, "combined_shot_type" was always the variable of choice for this model. Given that our tree only yielded three terminal nodes, it did not make much sense to enforce strict control parameters. If our tree were more complex, we would've limited the depth or the number of terminal nodes in the tree to reduce the likelihood of overfitting. Ultimately, our final CART model yielded a Kaggle score of 0.65462, which finished as a close second to LASSO. Although it only split on one predictor, we can be confident in this score, but with the development of defensive statistics that could be added to the data set, we could potentially enhance this type of a model and account for more variability in the dataset. Having said that, since this model scored so close to LASSO, we were comfortable with its overall accuracy and strength.

Takeaways

Ultimately, we submitted the LASSO model because it yielded both the lowest score and it also incorporated regularization, which ensured we optimized the bias variance tradeoff.

We feel that the biggest takeaway from this project is that transforming and manipulating variables can go a long way. Perhaps certain categorical variables are too specific and have too many levels, so a broader category could prove to be useful. At times, we felt as if we were limited to the variables and dataset we were given. We often did not look to step outside the box, and look at this dataset in a different fashion. As a result, we feel like we were limited to only a few very useful variables, but with a bit more manipulation, we could've improved our models by using more variables.

Overall, the CART and Logistic models performed best because they utilized useful categorical variables. Additionally, both models were very simple, relying heavily, and in CART's case solely, on the "combined_shot_type" predictor. With this in mind, the next primary takeaway is that simplicity of the utmost important when predicting out of sample. For future research, we hope to both improve these models and find ways to generalize our efforts to a broader dataset and perhaps include certain defensive metrics.