Introducing dataset:

Our dataset was analyzed in order to predict how many points an NBA team might score in a game based on 5 variables: Field Goal Percentage, Free Throw Percentage, 3 Point Field Goal Percentage, Rebounds, and Assists. We gathered a large dataset which contained these scoring metrics for thousands of NBA games that occurred from 2005 all the way through this current season that ended prematurely in 2020.

Model Selection train/validation/test splits:

Our dataset contained 23,195 observations. Prior to testing a variety of approaches to model our data, we split our dataset into training, validation, and test subsets. The training dataset contained 50% of the data while the validation and test datasets both contained 25%. We fit all of our model approaches on the training set and obtained their test RMSE by predicting y values in the validation set. We then compared these out of sample test errors and selected the model with the lowest RMSE. We proceeded to fit the optimal model to a combined training and validation dataset and then verified its accuracy by testing it on the remaining test dataset.

Univariate analysis:

As a first step for analyzing the data we took each variable separately and tried to fit a linear model on the training data set with PTS\_home being the response variable. So, essentially we had 5 different models with each predictor being tested against the response variable. For each of these models we calculated the RMSE values separately and observed that the RMSE lied in the range of 9.623471 to 12.79566. The model with the lowest RMSE corresponded to the model with Field Goal Percentage as a predictor and the model with the highest RMSE corresponded to Rebounds as a predictor. Since these RMSE values were a little high we proceeded to multivariable analysis of stepwise regression.

Stepwise regression:

In the stepwise regression model we took all the variables as potential predictors and tried to iterate the model in forward stepwise direction, backward stepwise directions and both th directions at the same time. This ensured that we ended up with a model that consists of only the important variables and as it turned out all the 5 variables were equally important. Next we ran the models to calculate the RMSE for each of the models and found that each of the models had an equal RMSE of 8.025279.

Optimal model lasso: lambda, regression, final equation:

As the Lasso model provided the lowest validation set RMSE of 8.025041, we selected this approach as our optimal model and fit our combined validation and training dataset with Lasso regression. We selected an optimal lambda value of 0.01 as it produced the lowest test error of 7.831 when compared with the lambda value 1 standard error away from the minimum. Our final multilinear model is expressed in the following equation: Ŷ = -10.702 + 120.552(FG\_PCT) + 22.557(FT\_PCT) + 13.551(FG3\_PCT) + 0.592(AST) + 0.5(REB). This model accounted for 63.3% of the variation in points per game. When analyzing the regression distribution of residuals plot, it can be observed that our model’s residuals are close to normal but exhibit a leftward skew. However, this negative skew can be explained by the nature of our dataset. Our dataset contains game statistics ranging from 2005 to 2020. Around 2015, the majority of NBA teams adjusted their game strategy, opting for a greater number of three point shots, thus dramatically increasing the total number of points scored. As our model accounts for games within this total range of years, it does have a tendency to underestimate the number of points scored per game.

Boosting:

As a precautionary measure we also ran the gradient boosting algorithm to ensure that we had covered the base with all the models, we ran a gbm model on the test data set found with 100 optimal trees and tried to ascribe an RMSE value for the model. The RMSE came out as 16.14957 which was pretty high relatively. We also took a summary of the model and tried to gauge the relative importance amongst all the variables and found that FG\_PCT home and AST\_home had relatively higher importance compared to all the other variables.

Interpreting results:

In order to test our multilinear regression formula, we obtained team stats of 4 teams from the most recent NBA season: Los Angeles Lakers, Orlando Magic, Miami Heat, and Milwaukee Bucks. Our model was fairly accurate in determining the points per game of these teams. For the Magic and the Heat, our model overestimated by 5 points, while for the Lakers our model underestimated by 7 points and for the Bucks our model underestimated by 11 points. These values lie fairly within our RMSE of the formula. We also ran more tests on games that occurred on July 25. There was significant error for predicting the Magic and the Bucks scores; we underestimated the Magic game by 25 points and underestimated the Bucks by 28 points. This is due to the fact that both teams had low field goal percentage and 3 point percentage, but they most likely had a large volume of shots taken which led to them scoring a large amount of points.

Dataset shortcomings:

The data set that we chose does not account for the volume of shots taken. A 50% field goal percentage could represent a team shooting 3/6 or 20/40. However, if we knew the volume of all shots taken and the percentages, it might be too easy to predict how many points were scored. Additionally, our model does not account for the style of offense played by each team. The Houston Rockets have been one of the best offensive teams in the league, yet are one of the worst in assists. Our model has assists as one of the most important offensive statistics, meaning that it is likely our model would predict the Rockets are not a good offensive team. By understanding what our model is actually doing, we can predict its shortcomings.