Predicting Season 3PT Percentage

Nehal Chigurupati

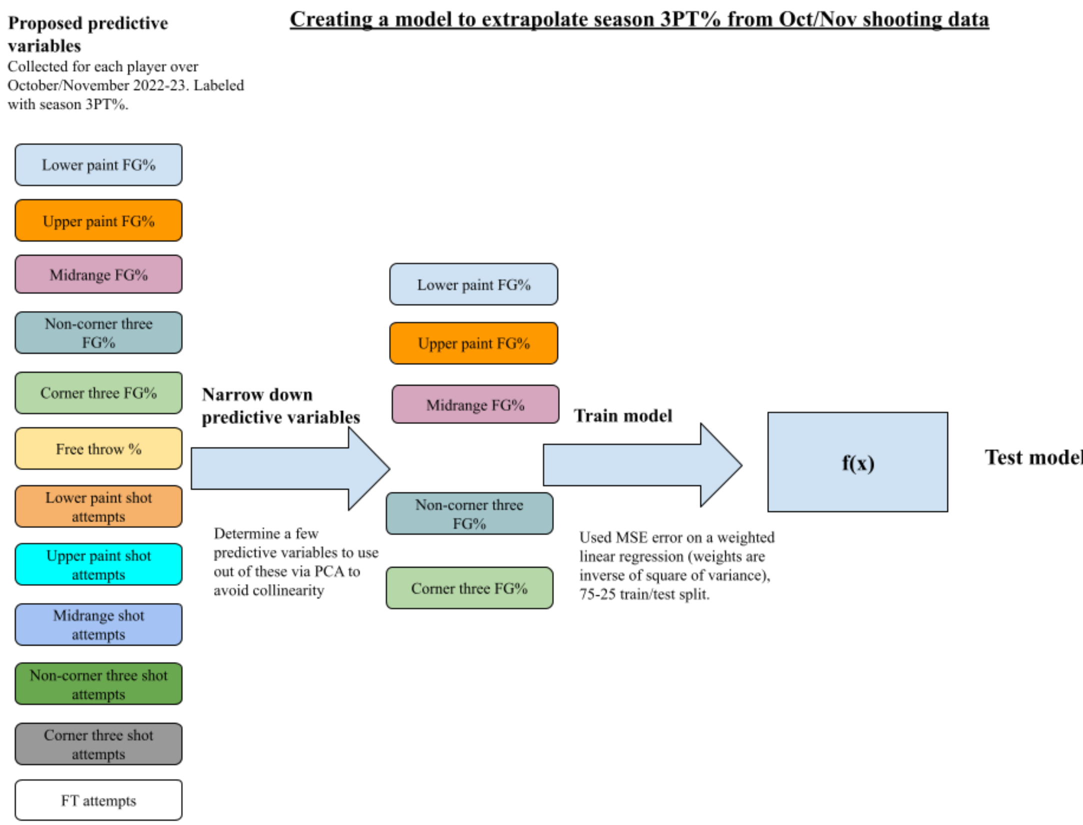
1 Objective

This analysis was done recently as my submission for the final round of the NBA's application process for the Future Analytics Stars program (I hear back next week). The assignment was to take efficiency data on player shot attempts from certain locations on the court during October/November of last season, and fit a model to extrapolate season 3 point percentage. This model might be interesting from the perspective of a front office, searching for younger players to trade for to add perimeter shooting quickly.

2 Methodology

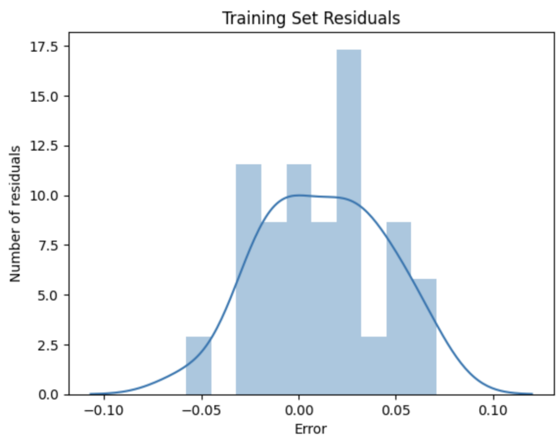
Twelve predictor variables were given, collected over the span of October/November 2022-23 for each player. First, I used PCA to narrow down the predictor variables to five statistics. I then tested four different models: linear, weighted linear, random forest, and gradient boosting. For the random forest model, the max depth was optimized via cross-validated grid-search. For the gradient boosting model, both the max depth and the learning rate were determined via the same technique.

Cross-validation demonstrated that the non-linear models were too complex for the data. Given that the MSEs of the two linear models were similar, I decided to plot residues vs. fitted values for the unweighted model, and noticed that there seemed to be some sort of dependent relationship developing. As a result, I felt homoskedasticity was too strong an assumption, and adopted the weighted linear model, resulting in the following pipeline:



3 Results

In training, the weighted linear model achieved an R^2 value of .992. In testing, the model had an **MSE of .001**. The model seems to work decently well at taking location-based FG percentage profiles from October/November and extrapolating season 3PT percentage.


Training Set Residuals
Number of residuals
Error
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
−0.10
−0.05
0.00
0.05
0.10