Using NBA Advanced Metrics to Predict VORP

Liam Rosengren & Will Gibbs

As fans of the NBA, it feels like we can't listen to a broadcast without the mention of various advanced statistics. Thus, we decided to focus our project on NBA advanced metrics and the statistics that go into calculating them. Our curiosities were piqued by two articles, one published by the University of Maryland and the other by the University of Arizona, which discussed player evaluation methods using advanced metrics such as WS, PER, and, the focus of our project, VORP.

Using advanced data for 5000 observations from 2014-2024 collected from the Basketball Reference website, we began with an exploratory data analysis (EDA) through various parametric and nonparametric regression methods. The goal was to predict the effect of the following 11 variables on a player's VORP: PER (player efficiency rating), TS% (true shooting percentage, 3PAr (3-pt attempt rate), FTr (free throw rate), TRB% (total rebound percentage), AST% (assist percentage), STL% (steal percentage), BLK% (block percentage), TOV% (turnover percentage), USG% (usage percentage), WS/48 (win shares per 48 minutes).

Our initial full model yielded an R^2 of 0.8846, where every variable in the model was significant at the 99% confidence level. However, after controlling for multicollinearity and removing TS%, PER, and shifting WS/48 to WS, our R^2 settled at 0.8801. Curious about the effect of these statistics on a positional basis, we created positional subsets that yielded varying R^2 values from 0.859 to 0.933. Disappointingly, there were no notable differences in the significance of each variable by position. A violation of normality led us to a nonparametric rank-based multiple linear regression (MLR), which diminished the R^2 to a value of 0.761. Rank-based regressions are susceptible to overfitting, so we expanded our analysis into random forests in an effort to mitigate any which we were potentially suffering from.

We used 100 decision trees for our model as ten times the number of predictors is a standard starting point, and adjusting it further would not have improved accuracy to a degree that outweighed the increase in computation time. The hyperparameters that produced the lowest RMSE included M_{try} equal to 2, sample fraction of 0.8, sampling without replacement and minimum node size of 5. The low M_{try} value demonstrates that each predictor leads the trees to be more different and less correlated, and most predictors can positively impact the model. Sampling without replacement encourages the model to include different types of players in the sample. The sample fraction was less than all of the data which increases variance amongst the trees, but 0.8 is relatively high which also shows how much the data samples can vary. A minimum node size of 5 is very low and shows that the decision trees are very deep, as each split will continue splitting the sample until there are 5 points or less per subgroup. Our random forest model has an R^2 of 0.824. The feature importance was also interesting to look at, as methods found the most valuable predictors to be PER, then WS/48, then USG% or AST% depending on the method of calculation.

Future work could include using our models of VORP to look at different years not included in the dataset to see how well our model applies to other years, as well as seeing how the game has developed throughout the years. We attempted to subset by position but it didn't develop significantly as we would have hoped, so it would be interesting to expand upon that idea. It also would be interesting to learn about other machine learning techniques and try applying them. Lastly, it would probably be more relevant to NBA teams to evaluate players earlier in their careers and try to predict how good they will become.