**Predicting the NBA MVP**

By Casey Li, Anjali Shrivastava and Aprillia Judokusumo

## I. Motivation

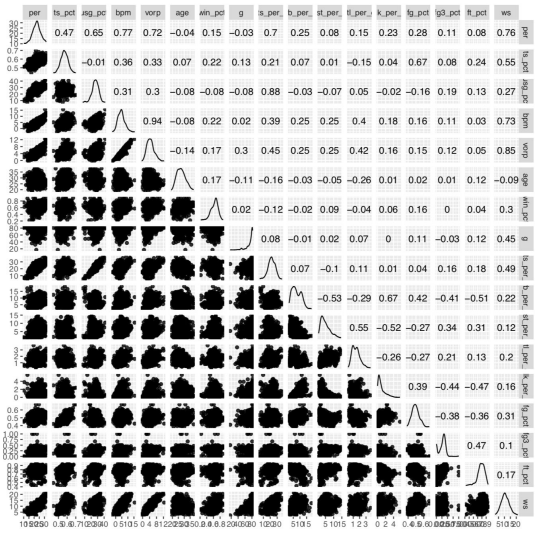
Every season, one out of 450 NBA basketball players is named Most Valuable Player (MVP), the most prestigious award given for an individual's regular season performance. The winner is chosen through the weighted votes of over one hundred media members. Aside from Stephen Curry's 2016 MVP season (where he achieved 100% unanimous MVP votes), there often arise arguments between fans and media about which factor is most important for a player to be named MVP. Some would argue that the MVP is the player who has the best individual performance and statistics. Others believe that the player that brings the most impact to the team and make their teammates better should be crowned the MVP. Both arguments are valid in their own ways, but **we aim to use various machine learning methods to help understand which player attributes make MVPs and consequently, predict who will be this season's MVP.**

## II. Data collection

We scraped voting and basketball player performance data from a third party website — basketballreference.com — from the 1980-81 season until now. Prior to 1980, voting for the NBA MVP was done by players; since the 1980-81 season, members of the media voted for the MVP. Therefore, it made sense to train our model only on data from seasons that picked the MVP using the same criteria as that of today. We also only chose to consider the top ten candidates for MVP from each season for two reasons: firstly, if we were to consider all players, those that do not play often usually have pretty extreme advanced statistics because of their limited playing time, which leads to extremely noisy data. Secondly, these top ten MVP contenders are usually easily identifiable early in the season, even by humans, so our decision to limit our model to ten players a season is easily approximable and reproducible.
In addition to voting data, we scraped each of the player profiles for advanced performance statistics like Player Efficiency Rating (per) and True Shooting Percentage (ts_pct). The scraper identified which 10 players got the most votes for MVP in each season, and scraped each of those individual's performance statistics.

## III. Variable Selection

Our final, cleaned data set was 43 variables — some basic statistics (variables recorded directly from the game), and some advanced statistics (variables derived from the basic statistics). Since we are only considering the top 10 candidates for the MVP title, we assumed that they have adequate statistics in terms of games, games started, and minutes played to qualify them for the MVP shortlist. Next, we saw some correlation among predictor variables, such as among Field Goal Made, Field Goal Attempt, and Field

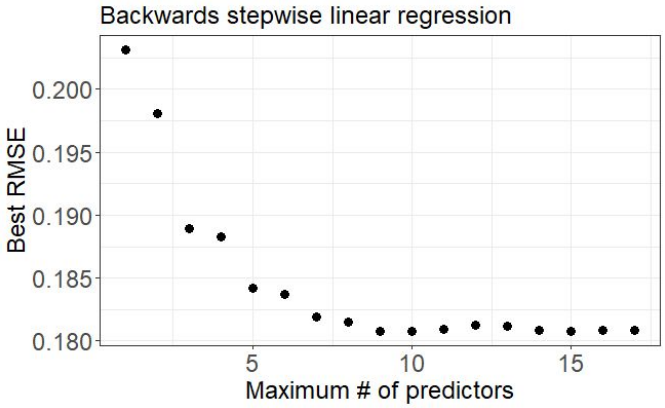


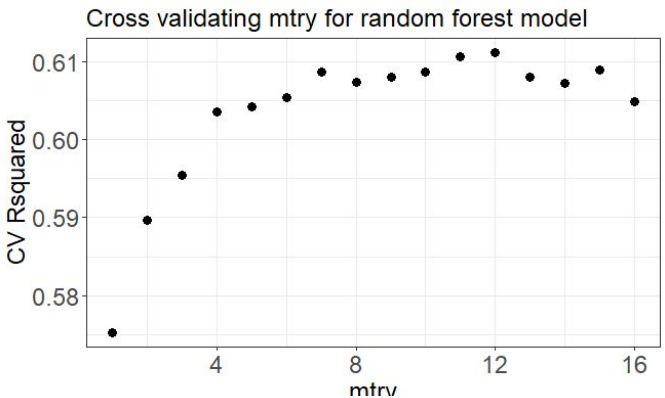
Goal %. Since Field Goal Made and Field Goal Attempt are reflected in Field Goal %, we only keep Field Goal % as a predictor to reduce multicollinearity. We also applied this to 2 Points, 3 Points, Free Throws, Box +/-, and Win Share. Besides using basic statistics that are shown in the box score, we decided to also use advanced statistics, that are formulaically derived from basic statistics. Note that, even though advanced statistics are derived from the basic ones, they are not linearly correlated with each other. Advanced statistics are essential to our model prediction because they do a better job of quantitatively capturing impacts which are not necessarily shown in the box score. However, we noticed some of the advanced statistics such as Assist % and Block % are statistics that only compares an individual's performance in respect of their team and thus won't accurately capture their performance with other players as team percentages highly fluctuate from season to season. Finally, we graphed the correlation matrix (above) and noticed that there are a few variables that are slightly correlated with each other, but we decided to keep the variables because we believe that those variables will still be useful in predicting the MVP winner.

## IV. Models

We chose to model the MVP selection process through a combination of parametric and non-parametric methods. In particular, we chose linear regression for its accessibility and interpretability. We chose to run both naive and the so-called "ideal" (where the number of and exact predictor variables are chosen through a backwards stepwise process) linear regressions to assess not just how accurate a parametric model can be, but to understand which variables did not have much predictive power. We chose boosting and random forests as the non-parametric methods because of existing literature of data analytics in sports that identified these methods as the best-performing ensemble learning models.

First, we built a baseline linear model (RMSE: 0.3091) by taking the average value of every player in the test set's award share. Our naive linear model took every predictor variable and improved slightly over our baseline model on both the train (RMSE: 0.2030) and training data ($OSR^2$: 0.5685). Next, our backwards stepwise linear regression identified 9 out of 16 features as the ideal number of predictor variables because that model had the smallest RMSE compared to models with more or less predictors. The best model of all 9-predictor linear regressions identified the strongest 9 predictor variables as fg_pct (field goal percentage), trb_per_g (total rebounds per game), ast_per_g (assists per game), stl_per_g (steals per game), ts_pct (true shooting percentage), ws (win share), bpm (box plus/minus), win_pct (win percentage), and pts_per_g (points per game). Even so,

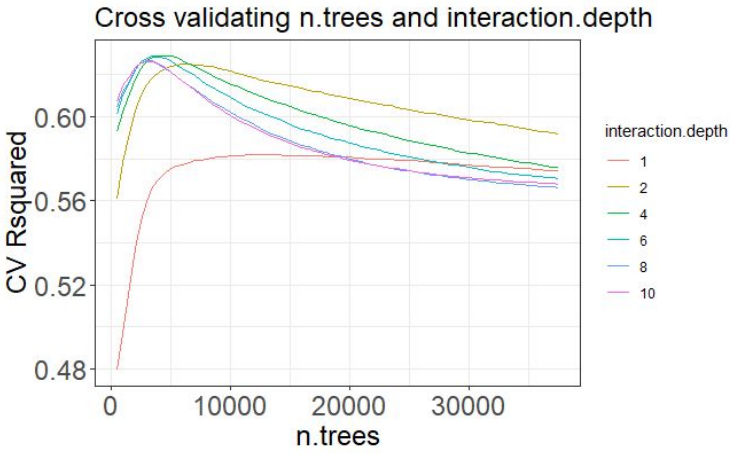


our best linear regression only performed slightly better than our naive model (RMSE: 0.2029; $OSR^2$: 0.5691).

Next, we implemented non-parametric statistical learning methods through random forest and boosting models. Upon running a basic random forest model with mtry = 5, we discovered that the most important features in the model are ws, vorp, win_pct, bpm, and pts_per_g. This is interesting because the first 4 are team contribution features, and the only individual feature that has great influence is the number of points they score in the game. Using cross-validation on 5 folds, we decided that the optimal amount of features to be chosen (mtry) is 12. The graph on the right shows the relationship between mtry values and cross-validated R-squared values. On our final model, the feature ws has very significant importance, followed by win_pct, per, and vorp. Tuning the model keeps the same top 4 influential features and just rearranges the order of their relative influence; this means that these 4 features are crucial in gaining insight to audience voting behaviors.

Lastly, in the initial boosting model, we started with parameters n.trees = 1000 and interaction.depth = 2. Running this initial boosting model, we discovered that ws, vorp, per, win_pct are the most influential features for this model (in order from highest relative influence). Then, we used cross-validation to determine the optimal values for parameters n.trees and interaction.depth. We discovered the best performance on boosting model is achieved when n.trees = 4000 and interaction.depth = 4. The graph above depicts the relationship between chosen values for boosting parameters (n.trees and interaction.depth) and cross-validated R-squared value. On the final boosting model, the features with highest influence are ws, win_pct, vorp, per. Tuning the boosting model keeps the same top 4 influential features and just rearranges the order of their relative influence; furthermore, these 4 features are consistent with the most influential features for the random forest model. Therefore, we can deduce that the 4 features above really factor in to audience voting behaviors.




The following table outlines the performance metrics for each of our models:

| Model Type | $OSR^2$ | MAE | RMSE |
|---|---|---|---|
| Baseline | 0 | 0.2150102 | 0.3090948 |
| Linear Regression | 0.5684896 | 0.1772665 | 0.2030427 |
| Backwards Stepwise Linear | 0.5691312 | 0.1787127 | 0.2028917 |

| | | | |
|---|---|---|---|
| Regression | | | |
| Random Forest | 0.7677164 | 0.1248506 | 0.1489708 |
| Boosting | 0.8231889 | 0.09692254 | 0.09692254 |

## V. Assessing performance

Using bootstrap on 10000 bootstrap replicates, we get the following confidence intervals for $OSR^2$, MAE, and RMSE:

| Model Type | $OSR^2$ | MAE | RMSE |
|---|---|---|---|
| Baseline | ( 0, 0 ) | ( 0.0776, 0.3208 ) | ( 0.1625, 0.4977 ) |
| Linear Regression | ( 0.4050, 1.9182 ) | ( 0.1165, 0.2289 ) | ( 0.1363, 0.2635 ) |
| Backwards Stepwise Linear Regression | ( 0.4146, 1.9315 ) | ( 0.1201, 0.2296 ) | ( 0.1397, 0.2602 ) |
| Random Forest | ( 0.6469, 1.7677 ) | ( 0.0775, 0.1690 ) | ( 0.1068, 0.1989 ) |
| Boosting | ( 0.7047, 1.8199 ) | ( 0.0434, 0.1399 ) | ( 0.0746, 0.1925 ) |

From these confidence intervals, we can conclude that both random forest and boosting models perform better than linear regression - especially in reducing MAE. However, confidence intervals for RMSE still have similar upper bounds.

Next, we return to our original motivation to try to understand how well each type of model did in successfully predicting MVP. In other words, did the player our models predicted the largest award share for actually win MVP? For the 2018-2019 season, both our linear models and the cross-validated boosting model predicted the #2-MVP contender James Harden as the MVP, when the award went to Giannis Antetokounmpo. Our cross-validated random forest model narrowly predicted the largest award share (0.6607) as going to Giannis over Harden (0.6517), but not the magnitude of the margin by which Giannis would beat Harden in the media voting (0.932 vs 0.768). None of our models successfully predicted the ranks of the top three MVP-contenders by award share (in order: Harden, Giannis, and Paul George), let alone the exact rank of all ten players in the 2018-2019 by award share.

## VI. Impact

One of our biggest motivation in using machine learning techniques to predict the future NBA MVP is **to find objectivity in a subjective, group decision-making process**. Throughout the years, objections about the fairness of the current MVP voting system have been brought up. Critics argue that media members may consider factors such as season narrative, nationality, and voter's fatigue (a tendency to prefer those who have never won the award over those who have won in the past), which overshadows the true objective of the MVP title in recognizing best player in the NBA. In building our models, we have identified the attributes players should focus on—individual or team—if they are aiming for an MVP title. Additionally, by being able to

quantitatively predict a numerical value that describes an individual's season long performance, we identify the role (or lack thereof) that variables only human votes consider—e.g., narrative, nationality, and voter's fatigue, as mentioned above—by comparing our model to the true results of MVP selections.

Based on our best performing models, the independent variables ws, vorp, win_pct, bpm, and pts_per_g for random forest and ws, vorp, per, win_pct for boosting are found to be the most impactful in predicting an individual's MVP chances. The variables ws (Win Share), vorp (Value over Replacement Player), and per (Player Efficiency Rating) are advanced statistics designed to capture how much value an individual could contribute in team wins and their individual impact compared to an average player. Lastly, win_pct (Team Winning Percentage) represents how the individual's team performed in that current NBA regular season. From these results, we can conclude that voters pick MVP candidates who come from a highly seeded team, contribute heavily to their team's wins, and do so with above average statistics and high efficiency in their gameplay. For instance, in our Harden-Giannis example, human sports analysts identify Giannis' edge over Harden as primarily due to narrative (Giannis was the centerpiece of a successful but lackluster-without-him team, whereas Harden, while still the strongest player on his team, did not stand out as much in comparison) and voter's fatigue, since Harden was the 2017-2018 MVP. Even so, it remains unclear that human voters conclusively relied on these two factors, since even our more complex non-parametric machine learning methods were able to successfully predict Harden over Giannis as MVP. For future iterations, we'd like to more closely investigate the role of subjectivity in the voting process, either by creating measurements and statistics ourselves or by more deeply studying other machine learning methods. Even though one of our models did successfully predict MVP, we'd also like to explore modeling the ranking of players as a more general problem.

Beyond predicting the MVP winner, we are also interested in seeing how much predicted MVP winners correlate to their team's success in future seasons, to even predicting the effect of the team in raising the chances of any individual winning the title. By capturing their individual impact towards a team success, we could also forecast various NBA teams' success given different scenarios (u.e. free agency, team injuries). In the end, we hope to use data analytics to gain more insights into why the NBA is one of the most famous sports leagues in the world.