Predicting NBA's Most Valuable Player

STAT 6021 Linear Models Final Project

Team 5

Shannon Overly Mitchell <som3dq@virginia.edu>
Andrew Kromkowski <ajk5nh@virginia.edu>
Kaustav Saha <ks3hw@virginia.edu>
Yifeng Song <ys8mz@virginia.edu>

Executive Summary

Analytics have been playing an increasing role in professional sports over the years. Whereas there are still disbelievers that value talent scouts over numbers, teams are increasingly hiring statisticians to assist in decision making. The NBA Most Valuable Player (MVP) Award honors the top player in the league each season. The voters consist of NBA players and media members. The goal of this project was to identify if a linear model can predict the MVP award winners which would suggest there is an underlying, possibly unconscious, model that the award voters tend to follow. A linear regression model was was created to predict the percent share of votes a player receives in a season based on their statistics for that season. Historical player statistics and MVP voting results from www.basketball-reference.com were used to perform regression analysis to uncover patterns and statistical correlations in this model. Basketball-reference has an in-house calculated "win shares" statistic which estimates the proportion of wins each player contributes to their team. The MVP prediction capability of this win shares statistic was 53% and was used as a baseline for this project to improve upon. Multiple models were created, and the top performing models were compared for prediction accuracy. The top model was able to improve upon the prediction capability of the win shares statistic to 57% by adding minutes played, free throws and personal fouls as predictors. Voting predictions were also calculated for two separate time periods as the voting panel changed in 1981. The resulting MVP player qualities can be used by team owners and managers in order to predict rising MVPs. With future analysis, this work can also be used to determine whether the cost of those potential MVPs are worth the associated increased revenue.

Introduction to APBRMetrics

APBRMetrics is the basketball-related offspring of Sabermetrics, which refers to the analysis of baseball through a systematic and quantitative lens. One of the most important domains in APBRMetrics is player evaluation, which consists of using statistical analyses to assess the performance of players. These analyses are used ubiquitously in the National Basketball

Association by managers who control player acquisition, coaches to determine starting lineups and playing times, and scouts who evaluate draft prospects.

Several advanced analytical methods are widely used in the basketball sphere. Many of these methods combine common statistics such as Points (PTS), Assists (AST), Rebounds (REB), Steals (STL) and many more to quantify a player's efficiency. Others attempt to evaluate the Wins above Replace (WAR) of a player, or the estimated wins he added when compared to a player of replacements level.

There are many highly notable critics of the overemphasis of advanced statistics in basketball, Hall of Famers such as Larry Brown and Charles Barkley. These critics argue that statistics are a useful but largely imperfect technique for analyzing a complex game. The only true technique for player evaluation, they claim, is using humans with deep subject matter knowledge to individually evaluate talent.

These critiques are at least partially supported by evidence. The table below shows the end of year PER (a widely used advanced statistic) rankings for the last 13 years of each of the last Most Valuable Player (MVP) award winners in the NBA.

Year	**MVP Winner**	**PER Ranking**
2015	Stephen Curry	3
2014	Kevin Durant	1
2013	LeBron James	1
2012	LeBron James	1
2011	Derrick Rose	9
2010	LeBron James	1
2009	LeBron James	1
2008	Kobe Bryant	7
2007	Dirk Nowitzki	2
2006	Steve Nash	14
2005	Steve Nash	18
2004	Kevin Garnett	1
2003	Tim Duncan	4 (Tied)
	Average	**4.9**

While the model performs well usually (and in several cases is able to pick the MVP), there are obvious years where there existed a tremendous disconnect between the statistical score and the opinion of a large body of basketball "experts." On average, the MVP predicted by these experts placed around 5th in the PER Rankings.

Project Objectives

Every year since 1956, a large pool of NBA players, sportswriters, and broadcasters have voted on the top 5 most valuable players. This report aims to pool together this 60 years of basketball subject matter knowledge to derive a quantitative metric that is more effective in modeling the complexities of player evaluation. The underlying assumption of this project is that there exists an underlying quantitative model that drives this voting process, and thus this report aims to derive this model through statistical analysis. An effective model will consistently predict the MVP given a testing data set that contains all necessary statistics and players for a given year.

Data and Data Pre-Processing

The data for the MVP award winners for all seasons during 1956-2015 were scraped from www.basketball-reference.com using the 'XML' package in R and then merged into one single data frame. The tables containing the NBA player statistics data during the same time period were downloaded from www.basketball-reference.com and saved as .csv files. The player statistics were then read into R and stacked together as a single large data frame. In this data frame, there were separator rows that did not contain any player statistics, so they were searched and removed from the data frame. The Season value for each row was converted into the year value when the season ended, for simplicity and consistency with the annual MVP votes data. Then the data during seasons prior to 1956 or subsequent to 2015 were deleted from the data frame. Finally the duplicated rows are checked and removed, resulting in the final table for all player statistics.

Next the player statistics data frame was combined with the MVP votes data frame on the players' names and the teams and seasons they played in, using the left outer join method. Additional checking was performed in order to verify that there is no misspelling within the names, teams and seasons present in both tables. The issue of more than one player sharing the same name was also checked and it was confirmed that among the players who were chosen to be the MVP contenders in the voting there were no identical names found in the non MVP contenders group. Then the entire data frame for the modeling was generated, which includes a total of 36 columns and 18,026 rows.

The response variable that is used in the model is the "Share" of the player, which is the points won by the player in the annual MVP voting divided by the maximum possible points that one could earn. For the rest of the players who were not qualified for the MVP voting, their "Share" values will all be set to zero. All other columns will be considered as the predictors except the

Season, Team, Player Name and League, which are all categorical variables and would probably have no influence on the model. Prior to the Season 1981-1982, the statistics for number of games started, 3-pointer, rebounds, assists, steals, turnovers are missing for some players. So we used the “Amelia” package in R to impute all of the missing values. The players who have the imputed values that obviously do not make sense in the real life scenario were not included in the modeling.

Analysis -- Model Building and Evaluation

We have subsetted the data by players who have played 20 minutes per game for 82 games in a season. The dataset has been further split into two halves containing pre-1980 era data and post-1980 era data. This has been done to take into account the change in the style of voting. Pre-1980 era had players voting whereas post-1980 era had the media voting. We have checked for Multi-Collinearity through VIFs(Variance Inflation Factors) and removed some variables such as FGA,FG,TRB,PtsWon,PtsMax,Share etc. Construction of a multiple regression model has been done and we have used analytic methods involving test for significance of regression. We have performed variable selection procedures on the full model (including forward / backward selection, lasso regression). The baseline model for our analysis contained Win-Shares. This is a widely used metric which assigns a score on each player based on the contribution of the player in the winnings share of a team.We have compared our model against this baseline model. We have essentially looked into 4 models. Model A had 16 variables including quadratic and interaction terms. Model B contained Minutes Played^2, Win Shares^2,Blocks,Personal Fouls.Model C had Minutes Played ,Defensive Rebounds,Personal Fouls^2,Free Throws,2 Pt Field Goals,3 Pt Field Goals ,Assists,Blocks.We determined the most important variables and construct a partial model which has a better performance than the initial full model. We checked the normal probability distribution of the residuals and perform R-Student Residual Analysis. We find and remove Influential & Leverage Points using DFFITS, DFBETAS, Cook’s D and COVRATIO. We have used Leave-One-Out Cross-Validation. We prepared the Training Set on all seasons except the season that we want to predict. We iterated over all the seasons and compared the top 3 players according to predicted vote shares to actual MVP winners for each season.

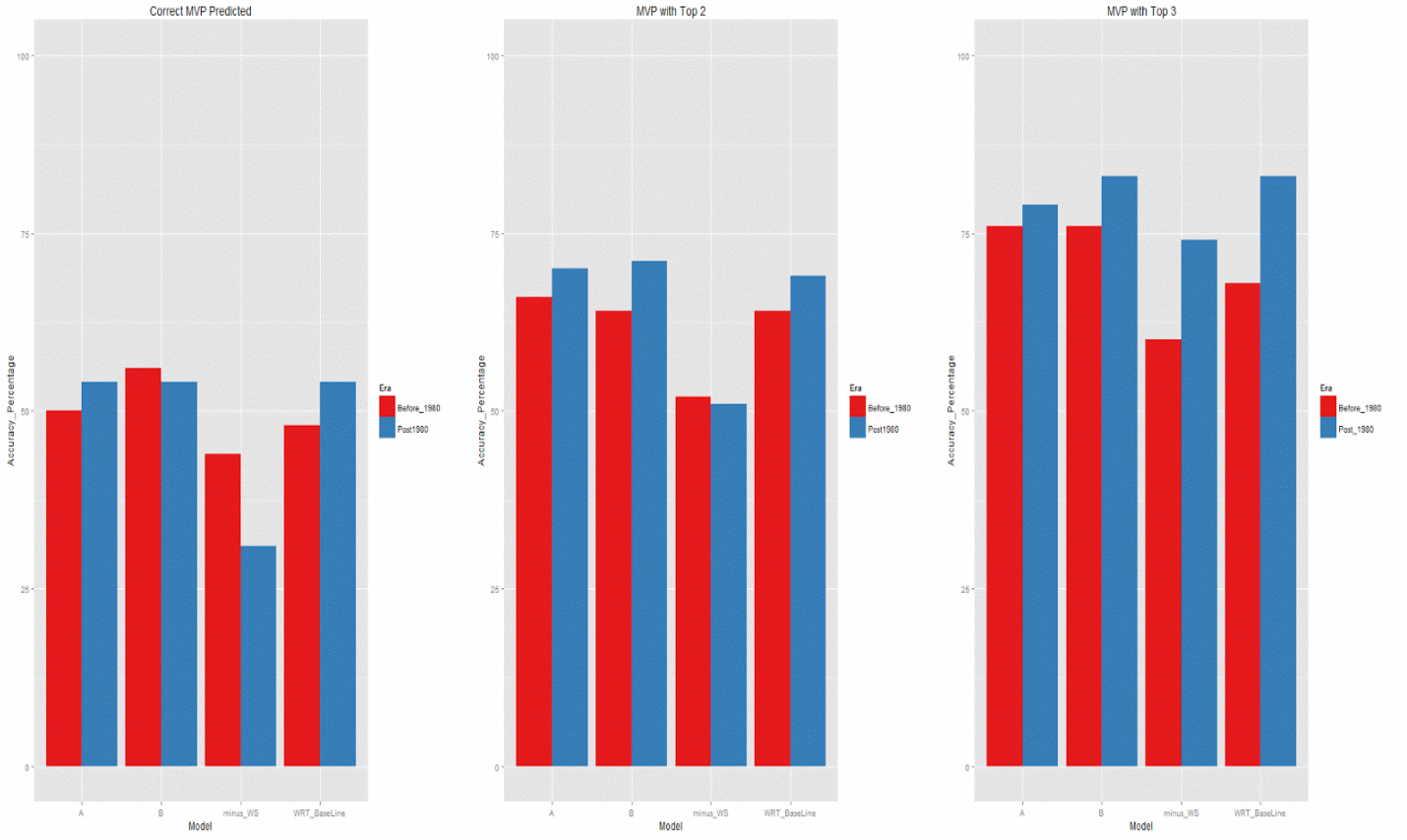
Results

The performance of each of the top models was evaluated by predicting which three players received the most MVP votes each year, and comparing those predictions to the actual MVP award winner.

Model	Predictors	Correct MVP Predicted	MVP within top 2	MVP within top 3
A	BLK * STL * REB Minutes Played^2 2 Point FG ^2 3 Point Attempts ^2 Free Throws ^2 Free Throw Attempts ^2 Offensive Rebounds ^2 Defensive Rebounds ^2 Assists ^2 Steals Field Goals 3 Pt FG % Free Throw % True Shooting Percentage Win Shares ^2	56%	70%	80%
B	Minutes Played^2 Win Shares^2 Blocks Personal Fouls	57%	70%	80%
C	Minutes Played Defensive Rebounds Personal Fouls^2 Free Throws 2 Pt Field Goals 3 Pt Field Goals Assists Blocks	50%	67%	83%
Baseline	Win Shares	53%	68	77%

The “Win Shares” variable is a calculated statistic quantifying a player’s contribution to a team’s wins. It was able to predict the MVP fairly well alone, however improved results were achieved by adding additional factors of minutes played, blocks and personal fouls. There was not a single model that performed the best in the three categories of correct MVP, top 2 and top 3, therefore multiple models have merit depending on what prediction is desired.

The MVP voting panel changed after the 1980 season. Through 1980, NBA players voted for MVP. Starting in 1981, sportscasters and broadcasters were given votes. As different types of voters could use different criteria, the models were rerun using the data in the two time periods separately.



The top performing model was again model B, with the actual MVP being predicted in the top 1, 2 and 3 vote-receivers more frequently than the other models. Even though Model C performed similarly to models A, B and the baseline in the full 1956-2015 season range, it performed more poorly than the other models with the separated time periods. If one time period had performed consistently better than the other it would imply that the model was more heavily influenced by a particular type of voting style, however that does not seem to be the case.

Besides the comparison of voting eras, two additional queries were run using the predicted percentage vote shares. The top 5 player seasons as defined by highest predicted vote share are as follows:

1964	Wilt Chamberlain
1972	Kareem Abdul-Jabbar
1962	Wilt Chamberlain
1988	Michael Jordan

1964 Oscar Robertson

The single lowest player season was also identified as Tyson Wheeler in 1999. This makes sense as his career was a single season only, and he only played in one game for 3 minutes.

The ***Biggest Snub*** happened in 2011 when the MVP was awarded to ***Derrick Rose*** who actually had a ***Score of 1.3*** while actually the best player was ***LeBron James*** with a score of ***1.9.***

Conclusions/Recommendations

In conclusion, the Win Shares statistic is a fairly accurate predictor of MVP winners, however it can be improved by adding minutes played, blocks and personal fouls. The change in voting panel in 1981 does not appear to change the predictive power of the models selected. However, it is possible that new models built specifically for the two different time periods might be able to achieve improved results. One other main change that could be evaluated is the addition of the 3-point field goal, which did not exist until 1979. As the timing of that change corresponds closely to the change in voting panel, the two differences might be confounded but would still be worthwhile to explore.

This model can be used by team owners and managers in order to predict rising MVPs. With future analysis, this work can also be used to determine whether the cost of those potential MVPs are worth the associated increased revenue.

---------- END OF REPORT----------