DSC 423 - Winter 2021

Sports/Video Game Group: El Dorado

Devin Carroll - Week 7 Individual Milestone

The first category pundits discussion in assessing NBA players is their ability to score often and reliably. But how can front offices tell if a player's scoring streak is an aberration or a valid trend? What areas should coaches instruct their players to improve their offensive game to maximize their potential and most efficiently increase their scoring output? By developing a model to predict a player's total points in a season, we hope to provide some insight into these questions. It is easy, but not productive to ask a player to improve their shooting accuracy, we aim to identify specific areas where players focus on developing their skills to most efficiently increase their scoring. However, because of the different positional requirements for the five main positions in basketball, Point Guard, Shooting Guard, Small Forward, Power Forward, and Center, we will develop models specific to each position group. This will allow us to provide more targeted insight to players of each position. Players who play at multiple positions will be included in datasets for all position groups they are listed at at basketball-reference.com, where our data was scraped from. As a result, players will often be included in the model building for multiple positions.

This report will focus on Small Forwards, a position unique in its versatility and range. Small Forwards are considered part of the frontcourt part of the team, along with the "Big Men" in Power Forwards and Centers, but also considered perimeters or wing players like shooting guards. Practically this means there is a wide range in the style of play for small forwards. Some players are shooters who specialize in off-ball movement and taking quick shots, like Khris Middleton. Others prefer to have the ball in their hands and create by driving or cutting directly towards the basket, like a young LeBron James or Jimmy Butler. Other players like to play in the receive in the low-post and create mid-range or layups from there, like DeMar DeRozan. This report will attempt to find what areas Small Forwards operate in and how they most frequently or efficiently score.

As a group we narrowed our focus onto 14 variables that we considered the most predictive of scoring across all positions based on their correlation with our dependent variable: Games Played, Games Started, Minutes Played, Assists, Steals, Turnovers, Free Throws Made, Free Throws Attempted, Offensive Rebounds, Defensive Rebounds, and Total Rebounds, 2 Point Shots Attempted, 3 Point Shots Attempted[1], and Total Field Goal Attempts. The advanced metrics were left out due to their limited usability as compared to box score metrics. Additionally, variables which like Assist Percentage or Rebounding percentage were found to not be as strongly correlated as the total counting version of those variables.

Due to my intuitive concern over multicollinearity among these variables, this potential issue was explored first to provide a guide as to which variables would interfere with one another. Not surprisingly shooting statistics were highly correlated with one another in a 1-to-1 manner. The one exception to this fact was 3 Point attempts, which didn't have a correlation coefficient with any other independent variables higher than 0.76. In comparison, most other shooting metrics had a correlation above 0.88 with all other shooting metrics, showing clear signs of collinearity. Rebounding statistics were similarly highly correlated with one another and Total Field Goal attempts, but were not a concern with other explanatory variables. The playtime statistics were also correlated with one another.

Based on these findings, I considered the best path forward was to experiment with one variable at a time from each group and see which variables were most useful in the model. Among the shooting metrics I eventually found Free Throw **Attempts** and 3 Point Attempts to be the most useful to include

[1] Note: In the data set 2 Point Attempts and 3 Point Attempts are listed as X2PA and X3PA respectively, the X is only there to follow R's variable name rules. It does not present any information.

in the model. These two features are not collinear with one another and have a great deal of predictive power. These two variables alone have created a model with an Adjusted R^2 of 0.9433. 3 Point Attempts was chosen because of its predictive strength and its lack of collinearity with other variables. It also provides a specific area of the court players should take and practice their shots from, making it a highly useful variable in terms of insight and explainability. Other shooting metrics were similarly predictive as Free Throw **Attempts**, but I selected Free Throw **Attempts** because it provides a more clear area of focus for users of the model. Unlike 2 Point Attempts or Total Field Goal Attempts, Free Throw **Attempts** provides a specific type of shot that players should focus on. First, this variable tells us, players who take and make their Free Throws are more productive. Secondly, it tells us that creating Free Throw Attempts is in itself productive, which tells us a specific type of non-3 Point shot is highly productive, namely close to the basket shots. Intuitively this makes sense, players who drive to the basket, draw fouls, and score despite being fouled put more pressure on a defense. The potential for a near basket to not only get the player 2 points but also additional points in free throws, outpaces the lower potential of a tougher outside shot. Plus by drawing fouls, defensive players are forced to either come out of the game or play defense more conservatively, making scoring later in the game potentially easier. These two variables also had significant t-tests with extremely small p-values. In all these factors make me confident these two variables should be included in the model, possibly even constituting a model by themselves.

Next I experimented with including Assists, Turnovers, Steals, and Rebounding statistics. Assists and Turnovers are highly correlated, and Assists were not particularly useful in the model. Total Rebounds was found to be the most useful and least multicollinear Rebounding metrics, but unfortunately still collinear with Turnovers. One solution was to apply a natural log, along with adding 1 to avoid errors, to either or both Total Rebounds or Turnovers. Another solution was to replace the variables with Total Rebound Percentage and Turnover Percentage respectively, these variables present probability of a player getting a rebound or committing a turnover rather than volume of those metrics. Both solutions have the unfortunate side effect of increasing the perceived importance of these two variables in the model due to the large difference in scale of values compared to the shooting metrics already in the model. Ultimately, I decided against including a rebound as it contributed little to the model's predictive strength despite having a significant t-tests. In the final model, I opted to use Turnover Percentage[2] over using a natural log transformation because I believe this metric to be easily understood and interpreted by potential users of the model. Additionally, Turnover Percentage has a negative beta coefficient and correlation coefficient, unlike total Turnovers. This difference in direction, may have led people to the illogical conclusion that Turnovers are a productive or acceptable outcome to a play. Steals were not found to be a useful variable in the model, likely because there is a large number of defensive specialists in the NBA who don't score particularly often but do accrue an impressive number of steals.

The next step was to determine which, if any, of the play time metrics should be included in the model: Games Played, Games Started, and Minutes Played. This choice is relatively straight forward, Minutes Played has the most predictive power of the three options, a linear model with just minutes played as an R^2 of 0.8544, compared to 0.5 and 0.69 for Games Played and Games Started respectively. While Minutes Played does not contribute massively to the Adjusted R^2 of our existing model, bringing it up to 0.9654, it does contribute significantly to reducing the Mean Square Error. Our previous model has a Mean Square Error of roughly 12541.5 but including Minutes Played reduces that by about 40% to

[2] Turnover Percentage is listed as 'TOV.' in the data due to variable naming rules. It presents the number of Turnovers a player is expected to commit over 100 possessions.

7680.5. This reduction in the residuals is likely due to the fact that playing time helps scale the players shooting volume statistics, it has the smallest standard error showing it has the smallest impact of the variables so far. However, I believe the reduction in residuals is reason enough to include Minutes Played in the model.

Stepwise model building methods were also experimented with as a check for the model individually constructed. Backward Elimination was attempted starting with all of the counting variables, Assists, rebounds, steals, blocks, etc along with the percentage version of those variables and various draft metrics, draft round, draft pick, height, and weight. Advanced metric-style variables were still excluded from this process due to the difficulty in utilizing them in prediction without a deeper historical analysis of the player. Using a player's PER or True Shooting Percentage for example for that season is more reactive to their points rather than vice versa. The Backward Elimination model was found to be much more complex and messy than the model built from scratch. It contained 17 variables, 10 of which had VIF higher than 10, indicative of a great deal of multicollinearity. Additionally many of the variables failed the t-tests and had beta coefficients that could not be trusted. Although the residuals were a good deal smaller than our model, the Adjusted R^2 was negligibly improved. The Backward Elimination model did contain the variables in the model already developed, but would require a similar amount of feature selection that was done to develop that model. Forward Selection had similar results to Backward Elimination, 15 variables were in the final model for that process with 8 having VIF scores of 10 or greater. For both methods, multicollinearity was confirmed in a number of cases where the sign of the beta coefficient in the model was opposite from the variables correlation with the dependent variable. Again, a number of the variables in the final model showed non-significant t-tests. Interestingly, the Mean Square Error is higher in Forward Selection as compared to Backward Elimination but the residual standard error is smaller. It seems the overall residuals are higher in this method but more tightly clustered near their mean. I think the reason for the complexity and messiness of these methods is due to the overlap among a majority of the explanatory variables. For example there is Usage Rate, Field Goal Attempts, 2 Point Attempts, 3 Point Attempts, Free Throw Attempts, and Free Throw Rate, all of which capture an aspect of the player's shooting volume. But the similarity in these variables makes it difficult for AIC and thus the algorithm to differentiate which variable to keep/add and which to discard. It also does not have an understanding of which variables are useful for front offices or coaches. In the end the Mean Square Error, explainability, and relative strength led to ignoring the stepwise models and continuing with the model already developed.

The final model for predicting a player's total points in a season among Small Forwards included explanatory variables for Turnover Percentage, beta coefficient = -2.38, Free Throw **Attempts**, beta coefficient = 1.99, 3 Point Attempts, beta coefficient = 0.68, and Minutes Played, beta coefficient = 0.17, with a intercept of 6.92. The final model has an Adjusted R^2 of 0.9654 and a Mean Square Error of 7680. As a formula: Small Forward Points = 1.99*(FTA) + 0.68*(3PA) + 0.17*(MP) + 6.92 - 2.38*(TOV%). Note in the dataset, 3PA is expressed as 3XPA and TOV% is expressed as TOV. due to variable naming rules. The overall model has an F-test of less than $2.2*10^{-16}$, showing something in the model is truly predictive of points scored over the season. The t-tests for each variable is similarly small and therefore, there is confidence in the beta coefficients that have been calculated and presented.

In conclusion, for Small Forwards in the NBA the best way to increase scoring output for a season is to focus on 3 point attempts and shots that lead to free throw attempts, such driving to the

basket and developing finishing skills after contact. In addition to limiting the rate or frequency with which a player turns the ball over will have a useful impact on their point totals. These findings are intuitively logical and in keeping with contemporary NBA thinking. 3 Point shots and attempts near the basket that result in free throws, have a higher scoring potential than mid-range 2 point attempts which rarely lead to free throws. Additionally turnovers have a negative effect on scoring because it lessens the number of potential attempts a player has over the course of the season. The number of minutes played helps correct for the range of possible outcomes in terms of points scored, adjusting for players who either play very few or the majority of minutes in a season. There is some overlap between Minutes Played and the shooting variables, the more a player is on the court, the more shots are possible for the player.

There are a few possible avenues to refine the model from here. The model could be refined to use variables scaled to per minute metrics, which could potentially give us greater insight into the balance of drives and 3 points shots a player should attempt. Additionally, we could use a player's previous year's metrics to predict a players point output for the next season. A model such as that one would provide front offices with a better tool for assessing which players to attempt to acquire, retain, or deal away. A model such as this one would also allow us to incorporate advanced metrics with intellectual honesty.