Performance in the NBA: Data 100 Final Project

Shankara Srikantan, Mohak Buch, and Adiyan Kaul

Abstract:
Successful modern NBA teams have used data analytics to improve their team performance. Though analysts would love to predict player performance and value purely based on numbers, intangible attributes such as mental strength, teamwork, and work ethic also undoubtedly influence the success of a career. We were curious about whether we could make meaningful predictions about the success of a player, measured by All Star selections and career average salary, based on information known prior to their league debut. Our results show that it is quite difficult to make these predictions based on NCAA statistics and height/weight/position information. After working on these models, we then moved on to explore what factors are most significant for performance within NBA games. We conduct analysis of key features for winning NBA games, and isolate field goal attempts (team and opponent) as being the primary stats that NBA players and teams should seek to boost in order to win games.

1. Introduction

We are interested in exploring the question of whether we can predict NBA stars based on early career metrics such as NCAA numbers. Players at the high school and college level are often acclaimed as future NBA stars. Sometimes, these players (who are early draft picks) do indeed meet or exceed expectations. Frequently, they underperform or flop as well. Intangible factors, such as natural athleticism (which often refers to a subjective quality of a player's ease and grace) and mentality are used to make these early judgements. Could we rely on simple data to make accurate predictions of success?

One natural way of assessing player success is to look at All Star selections. A handful of players are chosen as All-Stars in the middle of each season for their performances and impact. The All-Star selection process is not perfect. There have been numerous well known "snubs", where the public feels a deserving player was left out of the list. Nevertheless, an All-Star selection is generally regarded as a measure of a high-calibre player.

Another way of assessing success is to look at player salary, which can be viewed as the amount that a team values a certain player. The NBA is a highly competitive marketplace, and teams effectively must bid for their players.

We then move onto the somewhat different topic of what in-game metrics seem to be most important for winning. We are interested in how this inquiry can lead to the question: what skills should players and teams focus on developing. For example, are high field goal percentages or high field goal attempts more predictive of wins?

2. Data

We are provided with datasets by the Data 100 course staff about basic player information, and detailed box score information on the team and player level.

To conduct analysis of All Star selections and Salaries, we merge in external datasets. NBA salaries have consistently increased over time, making comparisons between the past and present meaningless unless we can normalize for this inflation somehow. We use a strategy of dividing salaries by the season's team salary cap (another external dataset). The normalization does not result in a perfectly straight line, because teams can exercise various exceptions to exceed the salary cap, but is considerably better than the non-normalized graph. Furthermore, when teams exceed the salary cap, it is usually because they value their players highly enough to find sources of revenue to pay them. These are some assumptions we made.

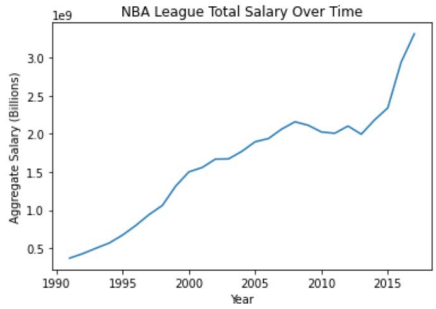

Figure 1: NBA Total Salary Over Time

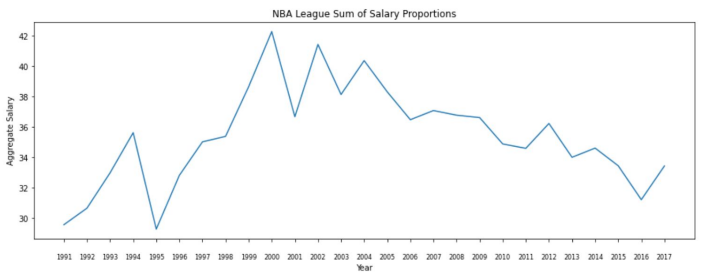

Figure 2: NBA League Salary Normalized By Salary Cap

In the data cleaning stage, we also take several other steps to ensure that the data we later analyze is relevant and complete. We notice that there is more incompleteness in data from the 80s. Combined with our knowledge that the 80s NBA was a very different game from today's, we eliminated pre-1990s data. Also, we did not have pre-NBA data for International players, so they too were eliminated. **This was a sort of ethical dilemma for us -- since International players come from various leagues around the world it is not trivial to fill in their statistics or understand the quality of these leagues.** However, they are a large and significant portion of NBA players, and knowledge about them is highly valuable to teams, and we have forsaken that. Finally, players who skipped college too were eliminated, but this is such a small subset of players that it is most unlikely to affect our research.

In terms of handling missing values, we were forced to remove some players from the dataset for whom college statistics were not available (only 40 players out of over a thousand in the dataset). The rest of the NaN values were in place of 0s, and we replaced these values accordingly.

2. Exploratory Data Analysis

We start out by exploring our basic assumption that there are correlations between player NCAA and NBA statistics. Here, we see that while there appears to be a weak correlation between points-per-game and field-goal-attempts in both leagues, there is no discernible relationship between the number of games played in both leagues. This lines up with our intuition that great players may spend any number of years in college before entering the draft.

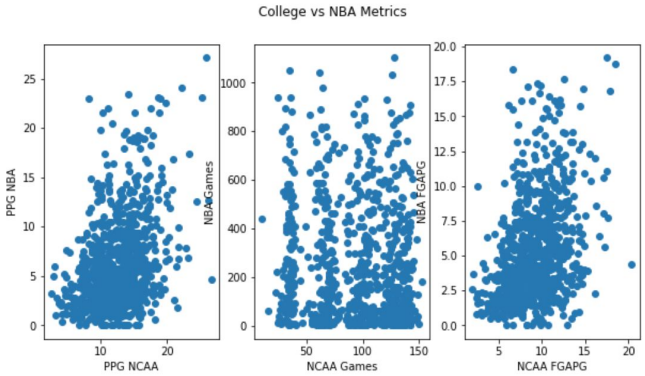

Figure 3: Player Statistics NCAA vs NBA

We also observe, interestingly, that All-Star players and non-All-Star players hardly differ in their physiques. The average height and weight comparisons are 78.8 vs 78.6 and 222.0 vs 218.0 respectively. Our visualization confirms this.

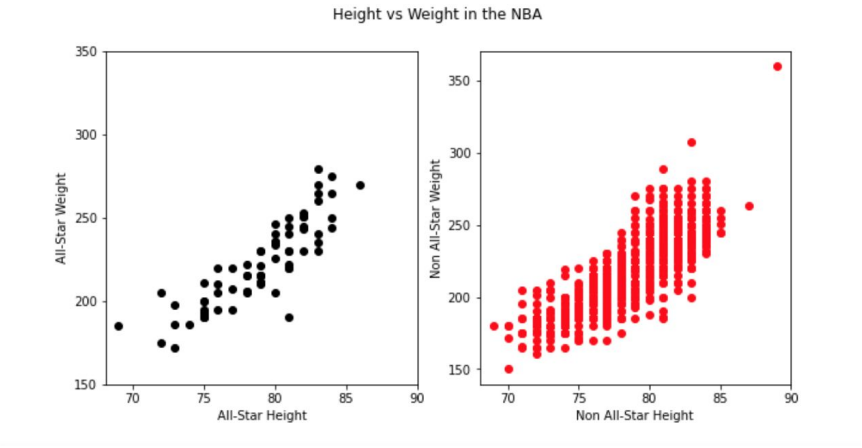

Figure 3: Height vs Weight for NBA Players, separated by All-Star selections

As we move onto modeling, we wanted to get an idea of the relationships between variables and features in our college dataset. This gorgeous heatmap shows the correlations between all variables. There are interesting observations here that confirm our intuitions and give us new insights. As expected, height and weight negatively correlate with three-point attempts as well as "guard" positions. Field goal attempts and points per game positively correlated, presumably because high-usage players get enough opportunities to score. Very interestingly, three-point attempts and field goal percentage are negatively correlated. This is not necessarily surprising, but the relationship is very strong. Position information was one-hot encoded.

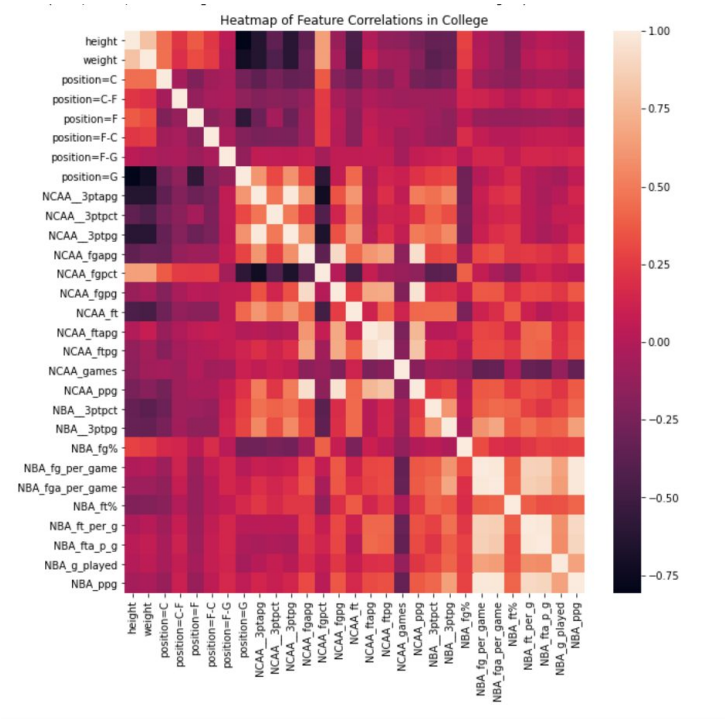

Figure 4: Heatmap of College Features

3. Methods, Models and Analysis

Salaries: In our first model, we try to predict salaries based on information known before players enter the league. We trained a Linear Regression model, gradually adding features and complexity. Multiple Linear Regression seemed like the right approach for this task because we need to predict a quantity based on inputs. Important assumptions of Linear Regression for us to keep in mind are that there are linear relationships between independent and dependent variables, that independent variables are normally distributed, that there is not multicollinearity, that observations are independent, and that residuals are homoscedastic. **There is room to believe that these assumptions may not be perfectly correct, especially regarding the residuals, which we would test in greater detail in further study. We specifically address multicollinearity in in greater detail below.**

The below residual plots will be referred to as exhibits (a) through (f) from top left to bottom right.

Picture (a) shows the residuals for our simplest model, which uses only height, weight, and the player's position. (b) uses all available NCAA statistics, including but not limited to points per game, field goal %, field goal attempts per game. (c) uses this data as well as the height/weight/position data. (d) is a Ridge Regression. We use Ridge Regression because of the multicollinearity between features (as we saw in the heatmap), and Ridge Regression would minimize the effect of that. (e) is more for our curiosity than for training the model. For our assumptions about the model to hold any weight, we should expect that we can predict salaries accurately with NBA data. Indeed, our mean squared error is much lower using NBA data. Finally, since Ridge Regression was the best model we trained, we test it out, and the results are in picture (f). For all of these instances, we used cross validation to ensure a robust value for RMSE, and have more confidence before testing our model.

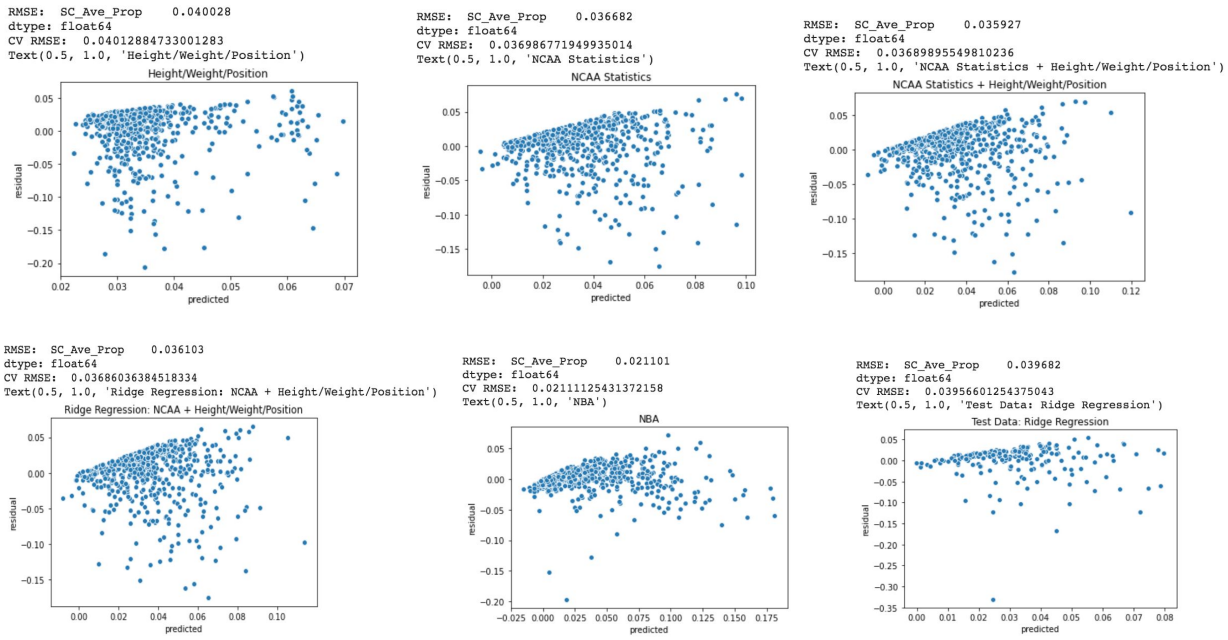

Figure 5 (a-f): RMSE, CV RMSE and Residuals for Linear Models to Predict Salary

All_Star Model: The goal of this model was to predict whether a player would eventually become an All Star based on a few college statistics, height, weight, and their enrolled university. After attempting to fit our training data to Logistic Regression, Random Forest, and Decision Tree models, we found that the DecisionTree model was the best model for predicting true positives. In this scenario, those were players correctly predicted to be NBA All Stars. As we expected, this model also greatly overfitted, yielding a perfect training accuracy and only an 89% test accuracy. While our Random Forest model yielded a greater test accuracy and a near perfect training accuracy, we found that this model could rarely ever yield a true positive value, and would predict an unreasonably large number of Non All Star Values. We are hesitant to trust this model much because we know that Decision Trees are prone to overfitting.

Decision Tree Model:

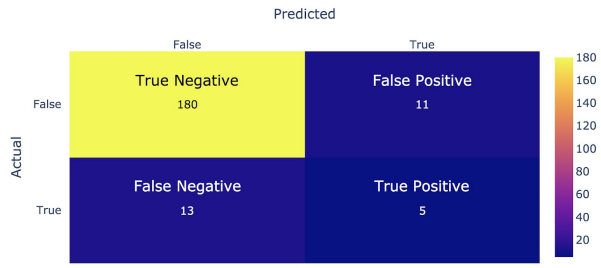

Figure 6: Test Predictions for Decision Tree

Random Forest Model:

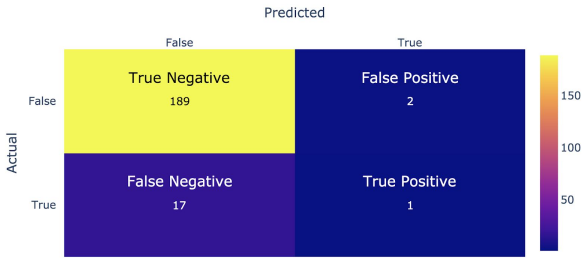

Figure 7: Test Predictions for Random Forest

Some of the notable players that fell under the false positives for our Decision Tree model were De'Aaron Fox and Jamal Murray, both of which are only 2-3 years into their careers and are expected to become All Stars in their careers.

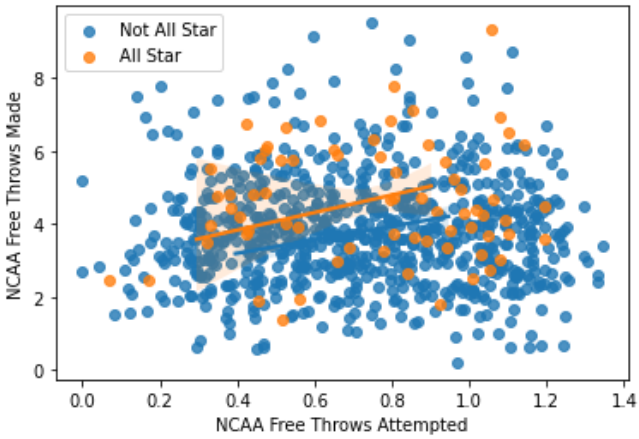

Figure 8: Free Throw Attempts and Makes

When we dove into the different features, we consulted our heat map created above and created several visualizations in order to break down the table into its most important and non-collinear features. Much to our surprise, we found data on free throws offered no correlations between

All Stars and regular players as shown in the figure. **Initially, we believed that the player who attempts the most free throws is likely the star player on any given team and would therefore have a greater chance of becoming an All Star.** However, our data showed otherwise. We also found that several features such as 3s per game, 3 percentage, and 3s attempted per game were far too similar and led to greater overfitting. Therefore, we removed the 3's per game, and field goals per game features along with free throw-related features.

For our college attended feature, we decided to use a technique called frequency encoding in order to assign values to the college column in our dataset. Much to our surprise, the model did 3% worse simply by removing this feature. Upon doing more research, we found that there is a loose correlation between the college attended and future All Stars given that a University of Kentucky and Duke account for the most All Stars in the duration of this data set (2003 - 2018).

One of the greatest drawbacks we faced was the lack of important features in the given dataset. We felt that data on assists, rebounds, blocks, and steals are extremely important to creating a more perfect model. Through our preexisting knowledge on the NBA, we knew that we would also have benefitted from having this same dataset of each NCAA player's final year in college, which is generally the most indicative of a player's potential. Seeing that many of the false positives in our dataset were late 2nd round draft picks and were never likely to be All Stars, we also came to the conclusion that knowing the draft position of each player would have helped. The primary assumption we made was that all the given statistics held equal weight. Our model does not account for key factors such as performance in March Madness, performance in conferences of different quality, and the growing emphasis on the 3 point shot.

NBA Game Win Predictions Model:

This part of the project wasn't as fleshed out as the rest but it was something we wanted to test with the data we had. We attempted to create a model that would predict who would win games simply by looking at the raw stats like rebounds,assists, field goal percentage,etc and predict who would win the game without knowing the points made by each team. This we believe would give remarkable insight into facets of the game are the most important when considering how to maximize the amount of wins a team has.

After training a random forest classifier on the box score data we were able to obtain a 100% accuracy on the training data and a 92% on the test data. Although the model overfitted a little bit, the accuracy on the training set was really high and we were impressed with that. We were able to analyze the feature importances which revealed that the two most important categories were field goal percentage and opponent field goal percentage. Each of these features had a weight of 0.11 which means they were both important indicators for success in the game. We were surprised that the algorithm determined that the opponent 3 point percentage was given a higher weight than the teams 2 point and 3 point percentage. This confirms a trend that we have seen recently in the NBA that teams have been trying to really limit their opponents ability to score three pointers and it was really cool that the model was able to pick up on this and gave it a high importance.

Concluding Thoughts and Ethical Concerns

Much of our analysis has already been conducted on a model to model basis. On a holistic basis, our models for salary and All-Star predictions left much to be desired. A lot of the room for improvement, we attribute to needing more sophisticated features, as is discussed previously in greater depth. Thus, with the limited data available about college players in these datasets, it does appear difficult to make a meaningful prediction of player success in the big league. Nevertheless, we have been able to reach some interesting conclusions about which features might be more important than others. We also came to interesting conclusions about which features were most important to winning NBA games.

There weren't any ethical concerns in the strict sense of us being worried about the impact of the analysis in a biased or potentially harmful manner. There were some decisions about handling our data that we had to question, such as, we noted before, the exclusion of international players from our analysis of salaries and All-Stars. A future iteration of this type of report may want to include that data as well to get a bigger picture of the NBA. Lastly, since we were not gathering some of the data ourselves or from the course staff and were simply taking it from an online source, we made sure that our sources were cited.

Data Sources:

1. **Data 100 Course Staff**
2. **https://www.basketball-reference.com/allstar/NBA-allstar-career-stats.html**
3. **https://data.world/datadavis/nba-salaries**
4. **https://basketball.realgm.com/nba/info/salary_cap**

Thanks to these helpful websites/individuals for sharing data!