Mitchell Wilson

Final Project Report

CS4347.001

Using Machine Learning Algorithms to Predict NBA Player Performance

Introduction:

Machine learning algorithms and predictive analysis are widely used throughout the sporting world. This report will focus on its applications to the National Basketball Association (NBA). One of the most well-known applications is the MVP Tracker from basketball-reference, which is used to track player metrics in real time and predict the probabilities that given players will win the Most Valuable Player (MVP) award at the end of the season. Machine learning is also increasingly being used in practice gyms to film the players while they practice. This type of software uses facial recognition to note where each player is, while tracking their shots in order to generate useful data for them to reflect on.

While some of the applications have been the inspiration for this project, this report details a predictive analysis on one specific player. The objective was to determine how well Russel Westbrook, an eight-time all-star and former two-time MVP basketball player, will perform after being traded from the Oklahoma City Thunder to the Houston Rockets. The Oklahoma City Thunder drafted him prior to the 2008-2009 season, and he has spent his entire eleven-year career with them until his recent departure. This season is currently underway, which is a good thing; I will have some mid-season stats to compare my predictions to.

I am interested in attempting this analysis, not only because I am admittedly a big Houston Rockets fan, but because this Rockets team is quite different than his Thunder. For this reason, a lot of controversy arose from this trade, with many people claiming that he will not mesh well with the Rockets at all. I suspected that this analysis may be difficult, because the Rockets are a unique offensive team and my objective is to predict some offensive statistics of his. The Rockets have an offense and team that is unlike any other franchise in the NBA, and my goal was to take some player metrics from this team to make predictions regarding this high-profile trade.

My approach to this problem, seeing as that the catalyst of the situation is Westbrook playing with a new, differently styled team, is to use data collected on his previous teammates to train a model that will predict how Westbrook performs. Of course, his performance is not entirely dependent on his teammates; for this reason, I have included the season number in all my models. My theory is that his performance is affected a significant amount by his teammates. For instance, if his teammates tend to turn the ball over a lot, that may leave less offensive opportunity for Westbrook to capitalize on. If his teammates tend to score a lot, that may leave Westbrook with the opportunity to get more assists or perhaps even score less. Those examples are not concrete predictions of mine, I am simply using them to convey my overall reasoning. I will be predicting the "big three" statistics of his: points per game (PPG), assists per game (APG), and rebounds per game (RPG). I determined that regression analysis would be fit for this type of problem, so I've made models using multiple linear regression, partial least squares regression, and random forest regression. Below is a summary of the results from each different type of model created. Bear in mind that Westbrook's stats are from less than a quarter of the current season and therefore are subject to change.

Westbrook Real			MLR			PLS			Random Forest		
PPG	APG	RPG	PPG	APG	RPG	PPG	APG	RPG	PPG	APG	RPG
22.5	6.9	7.6	23.9	9.1	7.4	23.6	8.8	7.4	23.1	8	6.7

Problem Description:

In order to have a one-dimensional set of dependent variables that will predict a one-dimensional output (PPG, APG, or RPG), I have averaged his teammates statistics together, each time for each season that Westbrook has been in the NBA. A normal averaging of these statistics, however, would result in some problems. For instance, a teammate that got to play one minute of basketball in an entire season, would matter just as much to the model as a player that played in every game and perhaps even logged more minutes played than Westbrook himself. To combat this, I chose to take a weighted average; one that is weighted based on the minutes played by each of his teammates. This way, a player with a lot of minutes played will have a greater impact on the model than a player with less minutes played. I also decided to take the top ten players by minutes played from each season (excluding Westbrook himself). While

choosing ten players inevitably results in a smaller set of training data, it also better represents how he will be playing with the Houston Rockets. The main reason being that another quirk of the Rockets is that their head coach likes to play a much smaller set of his roster than any other coach in the NBA.

The statistics of Westbrook's teammates that I chose to begin the process with are as follows: field goal percentage (FG%), three-point percentage (3P%), two-point percentage (2P%), total rebounds per game (TRB/RPG), assists per game (AST/APG), steals per game (STL), blocks per game (BLK), turnovers per game (TOV), points per game (PTS/PPG), minutes played (MP), and the season number (numbers 8-18 to represent seasons 2008-2009 through 2018-2019). Not all of these were determined to be relevant or consistent, which will be discussed later-on in this paper. Using these statistics as a starting point, I constructed nine different models that predict Westbrook's performance. Three models each for the regression methods mentioned above, each one predicting either the PPG, APG, or RPG of Westbrook.

I collected the data that I needed from basketball-reference, by copying data from each season into Excel spreadsheets. I also copied data from the Rockets most recent (2018-2019) season into a spreadsheet, for the purpose of making the final predictions. I saved each of these sheets as .CSV files, to make them easier to work with in Python. In the home directory of this project, you will find a file called 'dataPrep.py'. This file contains the code that prepared all the raw data for the models to be trained on. The code first reads in all of the raw data, calculates the weighted averages mentioned above, and saves the resulting data frame as three separate .CSV files, each one with either the PPG, APG, or RPG of Westbrook appended as the last column to serve as the dependent variable. The resulting files are 'PPGdata.csv', 'APGdata.csv', and 'RPGdata.csv'. The data preparation file also creates the file called 'RocketsData.csv', which is just the weighted averages from the 2018-2019 Rockets data, with no dependent variables. You can find all these outputted files in the 'Data' subdirectory.

For the multiple linear regression models, I first started by splitting the training data into training and test sets. The models were initially trained on all possible independent variables. Including all the dependent variables resulted in wildly inaccurate predictions for all three models. It was then obvious that if there was a correlation between this data and the performance of Westbrook, that the correlation was not present in all variables. So, I then calculated the p-values for each of the variables, began removing the variables with the highest p-values from the data, and retraining the models. I removed variables until I minimized the root mean square error (RMSE) value that was calculated for each model's test set, since that value indicates how well-fitting the models are. As I suspected, different variables mattered for each different model (PPG, APG, and RPG). There were still some commonly relevant variables between the models, a notable one being the season number. This was expected, as any good NBA player tends to get progressively better, to an extent,

following their rookie season. Below are p-values for each model, when trained on all the training data.

	PPG	APG	RPG
FG%	0.98	0.4	0.13
3P%	0.85	0.77	0.8
2P%	0.1	0.32	0.44
TRB	0.98	0.39	0.64
AST	0.53	0.98	0.63
STL	0.62	0.37	0.15
BLK	0.57	0.03	0.02
TOV	0.17	0.17	0.3
PTS/G	0.16	0.79	0.77
Season	0.06	0.009	4.15

Rather than focus on a specific cut-off point for the p-value of a variable to be included in a model, I chose to take the backwards elimination route. I made this decision because with a relatively small sample size, I did not want to put too much trust into the exact p-value.

For the partial least square regression models, I again split my data into training and test sets. I threw out variables that disrupted the model in my multiple linear regression models, and trained partial least square regressors on the remaining data, again trying to minimize RMSE values. I chose a partial least square regression for a few reasons. One reason being that there is a large number of dependent variables relative to the number of observations, and a PLS model will ideally be able to better traverse the multidimensional direction in the space of independent variables that ultimately explains the variance in the dependent variables. I also suspected that there may be some multicollinearity in the data, given that basketball is a team sport. For example, an assist for one player means that another player scored points.

For the random forest regression models, there was no need to split my data into training and test sets. I chose to use random forest regression, because it is an overall powerful regression technique. I hoped for greater accuracy in my predictions by averaging multiple deep decision trees, since these trees are trained on different areas of the same training set. I first created the forest/tree structure using 10 trees, but the results were subpar. I steadily increased the number of trees used, finding that around five-hundred trees were the sweet spot for optimizing the accuracy of my results. Any more than five-hundred trees and the return on accuracy was negligible and diminishing.

Results:

Westbrook Real		
PPG	APG	RPG
22.5	6.9	7.6

MLR		
PPG	APG	RPG
23.9	9.1	7.4

(Multiple Linear Regression) The variables that seemed to most affect Westbrook's PPG are his teammates total rebounds per game, points per game, and the season number. The variables that seemed to most affect his APG are his teammates total rebounds per game, blocks per game, turnovers per game, and the season number. The variables that most affected Westbrook's RPG are his teammates field goal percentage, total rebounds per game, steals per game, and the season number. The only variables that my models found relevant that I might not have initially intuited are blocks and rebounds per game affecting Westbrook's APG, and steals per game affecting Westbrook's RPG. Possible reasons could be that more blocks lead to more fast breaks (an area where Westbrook excels) that allow him more assist opportunities, and that more offensive rebounds lead to more assist opportunities well. In addition, his teammates stealing the ball might lead to less chances for him to get rebounds. However, that is just speculation and I have no data-driven analysis to confidently make those claims. In conclusion, my multiple linear regression models over predicted Westbrook's PPG by 1.4 points, over predicted his APG by 2.2 assists, and underpredicted his RPG by only 0.2. Aside from the RPG prediction, these were generally my least accurate models.

Westbrook Real		
PPG	APG	RPG
22.5	6.9	7.6

PLS		
PPG	APG	RPG
23.6	8.8	7.4

(Partial Least Square Regression) These models over predicted PPG by 1.1 points, over predicted APG by 1.9 assists, and underpredicted RPG by 0.2. That is a 0.3-point improvement in PPG, a 0.3-assist improvement in APG, and no improvement in RPG from the multiple linear regression model.

Westbrook Real		
PPG	APG	RPG
22.5	6.9	7.6

Random Forest		
PPG	APG	RPG
23.1	8	6.7

(Random Forest Regression) These models over predicted PPG by 0.6, overpredicted APG by 1.1, and under predicted RPG by (-0.9). That is a 0.8-point improvement in PPG, a 1.1-assist improvement in APG, and a 0.7-rebound reduction in RPG from the multiple linear regression model. It is also a 0.5-point improvement in PPG, a 0.8-assist improvement in APG, and a 0.7-rebound reduction in RPG from the partial least square models.

Conclusions and Future Work:

Overall, I am satisfied with results that I was able to come up with, as they closely approximate the actual performance statistics of Russel Westbrook so far this season. As the season is not even a quarter of the way through, it will be interesting to see how much closer (or further away) his actual statistics get to the predictions of my models.

It is interesting that while the random forest approach was the most accurate in predicting PPG and APG, there was actually a decline in accuracy in predicting RPG. I suspect this could be due to relations in the variables more correlated with RPG that are not present in those that are more correlated with PPG and APG. Perhaps some of these relationships are linear, while others are nonlinear in nature. Overall, my least accurate predictions were typically the APG predictions. Last season Westbrook had a career high 10.7 APG, so maybe that influenced my model more than I would've liked. Including some different data to represent the fact that the offense would no longer be run through Westbrook when he is a Rocket might have helped to improve APG accuracy.

Some future work could include trying entirely different regression techniques than the ones that I've used or just improving on the models that I have made. Including more statistics such as the number of shots taken, fouls, or even more advanced statistics like true or effective field goal percentage might yield more accurate results. It

might also be advantageous to include some team statistics, such as their total record. While a big part of this project was factoring in the fact that players take on a roll based on how good their teammates are, perhaps including their teams total win-loss record would make for even better results. In addition, I suspect my methods biggest drawback to be the size of the data. Other machine learning problems, such as image classification, can have huge datasets to train on, whereas I am limited by the number of seasons that Westbrook has played in the NBA. For that reason, it may also be helpful to use these models on a number of different NBA players. This would allow for more data to be used in order to determine the relationships between one players performance and that of their teammates. While I'm sure it will affect each player in a slightly different way, a metanalysis of a large number of these models on different players may provide some more clarity regarding these relationships. In other words, maybe there are some trends that will become more obvious, and changes can be made to my 'Westbrook models' to better accommodate for them. Another approach could be to start small, and construct models for each individual variable to better determine the nature of its relationship with the target player's performance metrics.

From this project I learned that while using machine learning to do predictive analysis can help you discover intricacies of data that you would otherwise be completely ignorant to, it also is not perfect. It is impossible to consider every possible contribution to a certain independent variable when it comes to real world problems. On the contrary, I'm sure that as the databases in the world grow not only in size but in quantity, that our ability to leverage that data to make useful predictions and decisions will only continue to improve.

References:

Quantifying Shot Quality in the NBA (Chang et al. 2014)

Prediction of NBA Games Based on Machine Learning Methods (Y.H. Hu. 2013)

Recognizing and Analyzing Ball Screen Defense in the NBA (McIntyre et al.)

Using Automated Machine Learning to Predict NBA Player Performance (Miller. 2018)

Artificial Intelligence in NBA Basketball (Woo. 2018)

https://www.basketball-reference.com/