Final Report - Team 103

Introduction/Background

Studies using previous performance data have shown that college experience may not improve performance in the NBA and that players with limited or no college experience perform better than those who attend college. Another argument is that a player who enters the NBA directly from high school lacks maturity, measured by one's ability to choose goals which highlights the need for important psychological needs, such as the need for autonomy, relatedness, and competence [1]. Fewer than 2 percent of NCAA student-athletes go on to become professional athletes even though on average NCAA student-athletes graduate at a higher rate than the general student body. Hence an analysis based on game statistics parameters like Win shares, VORP, defensive and offensive would provide better insights into a players development and how much college experience relates to the player's performance in the NBA.

Problem Statement

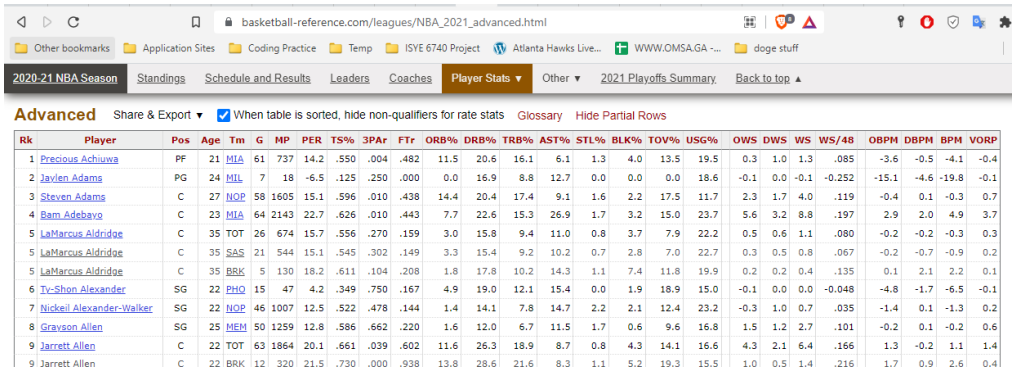
Perform a comprehensive analysis to predict the success of players drafted from NCAA, based on their relationship between college experience and their performance in NBA, using offensive and defensive statistics like point scores, assists, steals, blocks, field goals, Winshare and VORP.

Data Source

We sourced the NBA(1) and College Basketball(2) data for our project from 2 different websites.

1. The NBA data was downloaded from Basketball Reference, and contains advanced statistics on each NBA player from 2011-2021 (example link for 2021). From this dataset we extracted the variables that can summarize overall success in the NBA. The variables we selected are: 'G (games played in a season)', 'MP' (minutes played in a season), 'PER' (Player Efficiency Rating), 'OWS' (Offensive Win Shares), 'DWS' (Defensive Win Shares), 'WS' (Win Shares), 'WS.48' (Win Shares per 48 minutes), 'OBPM' (Offensive Box Plus Minus), 'DBPM' (Defensive Box Plus Minus), 'BPM' (Box Plus Minus), 'VORP' (Value Over Replacement Player), ‘Player’ (Player Name), ‘Year’ (year played) and ‘Pos’ (play position).

   Example of nba advanced data from basketball-reference:



basketball-reference.com/leagues/NBA_2021_advanced.html

Other bookmarks | Application Sites | Coding Practice | Temp | ISYE 6740 Project | Atlanta Hawks Live... | WWW.OMSA.GA -... | doge stuff

2020-21 NBA Season | Standings | Schedule and Results | Leaders | Coaches | Player Stats | Other | 2021 Playoffs Summary | Back to top

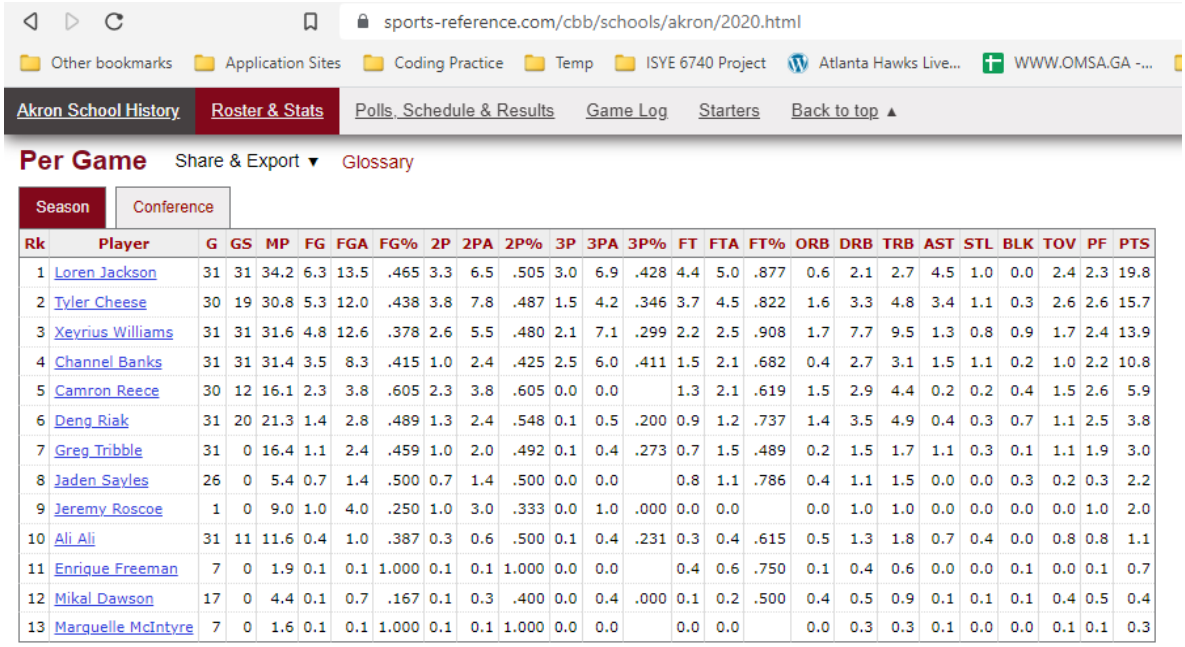
Advanced Share & Export | When table is sorted, hide non-qualifiers for rate stats | Glossary | Hide Partial Rows

Rk	Player	Pos	Age	Tm	G	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP
1	Precious Achiuwa	PF	21	MIA	61	737	14.2	.550	.004	.482	11.5	20.6	16.1	6.1	1.3	4.0	13.5	19.5	0.3	1.0	1.3	.085	-3.6	-0.5	-4.1	-0.4
2	Jaylen Adams	PG	24	MIL	7	18	-6.5	.125	.250	.000	0.0	16.9	8.8	12.7	0.0	0.0	0.0	18.6	-0.1	0.0	-0.1	-0.252	-15.1	-4.6	-19.8	-0.1
3	Steven Adams	C	27	NOP	58	1605	15.1	.596	.010	.438	14.4	20.4	17.4	9.1	1.6	2.2	17.5	11.7	2.3	1.7	4.0	.119	-0.4	0.1	-0.3	0.7
4	Bam Adebayo	C	23	MIA	64	2143	22.7	.626	.010	.443	7.7	22.6	15.3	26.9	1.7	3.2	15.0	23.7	5.6	3.2	8.8	.197	2.9	2.0	4.9	3.7
5	LaMarcus Aldridge	C	35	TOT	26	674	15.7	.556	.270	.159	3.0	15.8	9.4	11.0	0.8	3.7	7.9	22.2	0.5	0.6	1.1	.080	-0.2	-0.2	-0.3	0.3
5	LaMarcus Aldridge	C	35	SAS	21	544	15.1	.545	.302	.149	3.3	15.4	9.2	10.2	0.7	2.8	7.0	22.7	0.3	0.5	0.8	.067	-0.2	-0.7	-0.9	0.2
5	LaMarcus Aldridge	C	35	BRK	5	130	18.2	.611	.104	.208	1.8	17.8	10.2	14.3	1.1	7.4	11.8	19.9	0.2	0.2	0.4	.135	0.1	2.1	2.2	0.1
6	Ty-Shon Alexander	SG	22	PHO	15	47	4.2	.349	.750	.167	4.9	19.0	12.1	15.4	0.0	1.9	18.9	15.0	-0.1	0.0	0.0	-0.048	-4.8	-1.7	-6.5	-0.1
7	Nickeil Alexander-Walker	SG	22	NOP	46	1007	12.5	.522	.478	.144	1.4	14.1	7.8	14.7	2.2	2.1	12.4	23.2	-0.3	1.0	0.7	.035	-1.4	0.1	-1.3	0.2
8	Grayson Allen	SG	25	MEM	50	1259	12.8	.586	.662	.220	1.6	12.0	6.7	11.5	1.7	0.6	9.6	16.8	1.5	1.2	2.7	.101	-0.2	0.1	-0.2	0.6
9	Jarrett Allen	C	22	TOT	63	1864	20.1	.661	.039	.602	11.6	26.3	18.9	8.7	0.8	4.3	14.1	16.6	4.3	2.1	6.4	.166	1.3	-0.2	1.1	1.4
9	Jarrett Allen	C	22	BRK	12	320	21.5	.730	.000	.938	13.8	28.6	21.6	8.3	1.1	5.2	19.3	15.5	1.0	0.5	1.4	.216	1.7	0.9	2.6	0.4

2. The College Basketball data was scraped from sports-reference, and contains per-game statistics for each college player from 2011-2021 (example link for Akron University for 2020), as well as a column labeled 'Drafted', indicating whether or not the player was drafted to the NBA; 0 for No, 1 for Yes.

   We used a Python script to scrape data from this website because it provides player statistics on a separate page for every team. The script iterates over each college team available for every year between 2011 and 2021 and creates a consolidated .csv file containing per-game statistics.

   Example of college player per-game data from sports-reference:



Rk	Player	G	GS	MP	FG	FGA	FG%	2P	2PA	2P%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Loren Jackson	31	31	34.2	6.3	13.5	.465	3.3	6.5	.505	3.0	6.9	.428	4.4	5.0	.877	0.6	2.1	2.7	4.5	1.0	0.0	2.4	2.3	19.8
2	Tyler Cheese	30	19	30.8	5.3	12.0	.438	3.8	7.8	.487	1.5	4.2	.346	3.7	4.5	.822	1.6	3.3	4.8	3.4	1.1	0.3	2.6	2.6	15.7
3	Xeyrius Williams	31	31	31.6	4.8	12.6	.378	2.6	5.5	.480	2.1	7.1	.299	2.2	2.5	.908	1.7	7.7	9.5	1.3	0.8	0.9	1.7	2.4	13.9
4	Channel Banks	31	31	31.4	3.5	8.3	.415	1.0	2.4	.425	2.5	6.0	.411	1.5	2.1	.682	0.4	2.7	3.1	1.5	1.1	0.2	1.0	2.2	10.8
5	Camron Reece	30	12	16.1	2.3	3.8	.605	2.3	3.8	.605	0.0	0.0		1.3	2.1	.619	1.5	2.9	4.4	0.2	0.2	0.4	1.5	2.6	5.9
6	Deng Riak	31	20	21.3	1.4	2.8	.489	1.3	2.4	.548	0.1	0.5	.200	0.9	1.2	.737	1.4	3.5	4.9	0.4	0.3	0.7	1.1	2.5	3.8
7	Greg Tribble	31	0	16.4	1.1	2.4	.459	1.0	2.0	.492	0.1	0.4	.273	0.7	1.5	.489	0.2	1.5	1.7	1.1	0.3	0.1	1.1	1.9	3.0
8	Jaden Sayles	26	0	5.4	0.7	1.4	.500	0.7	1.4	.500	0.0	0.0		0.8	1.1	.786	0.4	1.1	1.5	0.0	0.0	0.3	0.2	0.3	2.2
9	Jeremy Roscoe	1	0	9.0	1.0	4.0	.250	1.0	3.0	.333	0.0	1.0	.000	0.0	0.0		0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	2.0
10	Ali Ali	31	11	11.6	0.4	1.0	.387	0.3	0.6	.500	0.1	0.4	.231	0.3	0.4	.615	0.5	1.3	1.8	0.7	0.4	0.0	0.8	0.8	1.1
11	Enrique Freeman	7	0	1.9	0.1	0.1	1.000	0.1	0.1	1.000	0.0	0.0		0.4	0.6	.750	0.1	0.4	0.6	0.0	0.0	0.1	0.0	0.1	0.7
12	Mikal Dawson	17	0	4.4	0.1	0.7	.167	0.1	0.3	.400	0.0	0.4	.000	0.1	0.2	.500	0.4	0.5	0.9	0.1	0.1	0.1	0.4	0.5	0.4
13	Marquelle McIntyre	7	0	1.6	0.1	0.1	1.000	0.1	0.1	1.000	0.0	0.0		0.0	0.0		0.0	0.3	0.3	0.1	0.0	0.0	0.1	0.1	0.3

Data Preprocessing

We took the average of the per-game stats over a player's college career and the average of the advanced NBA statistics each player had over the time period we are considering. We then combined the two datasets in R using an inner join that gave us our resultant dataset with the NBA players that came from college basketball over the past 10 years (2011-2021).

A screenshot of our final NCAA dataset:

	Player	Position	School	G	GS	MP	FG	FGA	FG_percent	Twos	TwosA	Twos_perc	Threes	ThreesA	Three	FT	FTA	FT_pe	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	Drafted
3	A.C. Reid	Guard	Liberty	52	6	1162	103	301	0.342	25	57	0.439	78	244	0.32	32	45	0.71	6	107	113	79	33	5	62	99	316	0
4	A.J. Astroth	Guard	Overall	37	9	484	45	97	0.464	40	75	0.533	5	22	0.23	29	46	0.63	39	92	131	17	15	5	27	50	124	0
2	A.J. Atkinson	Guard	Davidson	16	1	36	3	5	0.6	2	3	0.667	1	2	0.5	2	2	1	3	8	11	0	0	0	2	6	9	0
1	A.J. Bowers	Forward	Chattanooga	6	0	30	0	1	0	0	0		0	1	0	2	2	1	2	4	6	0	1	1	1	5	2	0
4	A.J. Caldwell	Guard	Chattanooga	69	15	1731	120	294	0.408	43	92	0.467	77	202	0.38	23	38	0.61	37	203	240	129	42	15	56	107	340	0
3	A.J. Cheeseman	Forward	Overall	33	12	559	85	185	0.459	83	173	0.48	2	12	0.17	70	109	0.64	43	76	119	15	14	16	44	68	242	0
4	A.J. Davis	Forward	Overall	61	52	1814	176	461	0.382	126	303	0.416	50	158	0.32	190	271	0.7	87	283	370	113	51	18	130	146	592	0

Data Modeling

Creating the Logistic Regression Model

We first create a logistic regression model using all variables except Position and School to predict a response variable or 0 or 1 for whether or not the player will be drafted by the NBA.

```
model1 <- glm(Drafted ~ G + GS + MP + FG + FGA + FG_percent +
              Twos + TwosA + Twos_percent + Threes + ThreesA +
              Threes_percent + FT + FTA + FT_percent + ORB + DRB +
              TRB + AST + STL + BLK + TOV + PF + PTS,
            family=binomial(link='logit'), data=trainSet)
```

From the model's summary output (refer output in HTML file under folder "Code") we can identify the statistically significant variables. The statistically significant predictor variables are: G (Games), GS (Games Started), MP (Minutes Played), FG (Field Goals), FG_percent (Field Goal Percentage), FT_percent (Free Throw Percentage), DRB (Defensive Rebounds), AST (Assists), BLK (Blocks), and TOV (Turnovers).

Discussion on the Coefficients of our model

These statistically significant predictor variables do outline the attributes that a well-rounded NCAA basketball player should excel in to showcase their potential abilities as an NBA player. A player can be drafted to the NBA if they play and start in many games, and play for the majority of the game (with high Minutes Played). These statistics showcase the consistency, reliability, and health a player has throughout their collegiate career. Field Goals, Field Goal Percentage, Free Throws, and Free Throw percentage are good indicators for a player's ability to score points which, arguably, is one of the most important skills a player needs. Defensive rebounds, Assists, Blocks, and Turnovers are good indicators of the non-point based metrics that illustrate a strong defensive and team player.

One thing to note, which we will discuss later, is the line, "Coefficients: (4 not defined because of singularities)". The "Threes", "ThreesA", "TRB", and "PTS" have coefficients of NA. To understand this we will look into the correlations our predictor variables have with each other.

"Threes" (Three-point shots made) is almost perfectly correlated with "ThreesA" (0.99011326) which makes sense, the number of three points made has a strong linear relationship with how many three point shots are attempted.

"TRB" (Total Rebounds) is the sum of "ORB" (Offensive Rebounds) and "DRB" (Defensive Rebounds) which is why "TRB" is almost perfectly correlated with those predictor variables.

Train and Test the Model(s)

We will now use model1 to predict the Drafted response variable for the 30% test set we created earlier.

```
probabs1 <- predict(model1, testSet, type='response')
```

We then convert these predicted values to 0 or 1 using a threshold of 0.5. If the predicted value is less than 0.5 we say the player won't be drafted to the NBA, and if it is greater than 0.5 we say the player will be drafted to the NBA.

```
preds1 <- ifelse(probabs1 > 0.5, 1, 0)
```

We will construct a Confusion Matrix using the actual Draft values from the test set and the predicted values to assess the model's accuracy.

```
confusionMatrix(factor(preds1), factor(testSet$Drafted))
```

```
          Reference
Prediction    0    1
         0 2903  111
         1    7    4

               Accuracy : 0.961
                 95% CI : (0.9535, 0.9676)
    No Information Rate : 0.962
    P-Value [Acc > NIR] : 0.6352

                  Kappa : 0.0572

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99759
            Specificity : 0.03478
         Pos Pred Value : 0.96317
         Neg Pred Value : 0.36364
             Prevalence : 0.96198
         Detection Rate : 0.95967
   Detection Prevalence : 0.99636
      Balanced Accuracy : 0.51619

       'Positive' Class : 0
```

The confusion matrix shows us that the model predicted the draft status for 96.1% of players correctly. However, there are 111 players that the model incorrectly did not draft, when they should have been drafted.
Our first model has an accuracy of 96.1% However, it did not draft 111 players that should have been drafted. Let's try lowering the threshold of our predictions to see if that make the model correctly draft more players.

We will lower the threshold to 0.4 and construct the Confusion Matrix.

```
preds2 <- ifelse(probabs1 > 0.4, 1, 0)
```

```
          Reference
Prediction    0    1
         0 2891  111
         1   19    4

               Accuracy : 0.957
                 95% CI : (0.9492, 0.964)
    No Information Rate : 0.962
    P-Value [Acc > NIR] : 0.9276
```

Lowering the threshold did not improve accuracy, it lowered it 95.7%.

Let's create a second model using only the statistically significant predictor variables and see if it performs better than model1 (refer output in HTML file under folder "Code").

```
model2 <- glm(Drafted ~ G + GS + MP + FG + FG_percent + Twos + FT_percent +
              ORB + DRB + AST + STL + BLK + TOV,
              family=binomial(link='logit'), data=trainSet)
```

With model2 we will predict draft status using the testSet and using a threshold of 0.5 and construct the Confusion Matrix.

```
probabs2 <- predict(model2, testSet, type='response')
preds3 <- ifelse(probabs2 > 0.5, 1, 0)
```

```
          Reference
Prediction    0    1
         0 3396  119
         1   11    3

               Accuracy : 0.9632
                 95% CI : (0.9564, 0.9691)
    No Information Rate : 0.9654
    P-Value [Acc > NIR] : 0.7849

                  Kappa : 0.0373

 Mcnemar's Test P-Value : <2e-16
```

```
            Sensitivity : 0.99677
            Specificity : 0.02459
         Pos Pred Value : 0.96615
         Neg Pred Value : 0.21429
             Prevalence : 0.96543
         Detection Rate : 0.96231
   Detection Prevalence : 0.99603
      Balanced Accuracy : 0.51068

       'Positive' Class : 0
```

According to the Confusion Matrix output our model's accuracy did improve to 96.32%. However, it still does not correctly draft 119 players.

Predicting the 2022 NBA Draft

Let's use model1 to predict which players from the 2022 NCAA season will be drafted to the NBA.

We used the same python script from earlier to get the relevant player information for the 2022 season only (ignoring draft status since that is what we want our model to predict)

After running our predictions with model1, we constructed a dataframe using player name's and the predicted value, 0 or 1. We sorted to list the 1's first.

Name	Drafted
Chet Holmgren	1
Trayce Jackson-Davis	1
Keegan Murray	1
JT Shumate	1
Norchad Omier	1
Jayveous McKinnis	1
Johni Broome	1
Max Abmas	1
Jalen Pickett	1

Model1 predicted 9 players to be drafted by the NBA; Chet Holmgren, Trayce Jackson Davis, Keegan Murray, JT Shumate, Norchad Omier, Jayveous McKinnis, Johni Broome, Max Abmas, and Jalen Pickett.

Notable Draftees Prior to 2011

Let's see how model1 predicts 10 notable NCAA draft picks prior to 2011: (1) Stephen Curry, (2) Blake Griffin, (3) Derrick Rose, (4) James Harden, (5) Draymond Green, (6) Kevin Durant, (7) Chris Paul, (8) Brook Lopez, (9) Carmelo Anthony, and (10) Russell Westbrook.

Model1 did not predict Derrick Rose, James Harden, Draymond Green, Chris Paul, Brook Lopez, or Russell Westbrook from being drafted to the NBA.

How to make a more effective model

Our model only uses box score stats; there could be other factors that play a role in determining whether a player gets drafted or not, for example: the position that player plays, or maybe even their height.

The NBA teams draft order is decided by using the reverse order of their regular season record. The teams that did poorly will get to pick first, giving these teams a chance to get a high draft pick to improve their team's performance for the upcoming season.

However, what should also be considered is the team's current roster. Did a veteran player just retire? Are there players planning on retiring? Are there players with ending contracts? Are there any players with significant injuries that could affect them for future seasons? These scenarios could also influence what draft picks to choose that go beyond the scope of just box score stats.

Predicting NBA Success Through Advanced Stats

Another way we considered predicting NBA success from a college player's performance is by looking at if we can predict advanced NBA statistics that summarize a player's performance with college per game statistics. The NBA variables we chose to use are Win Shares (WS) and Value Over Replacement Player (VORP). Win Shares is a player statistic which attempts to divvy up credit for team success to the individuals on the team. This metric is designed to estimate the player's contribution in terms of wins.Value Over Replacement Player (VORP) converts the Box Plus/Minus (BPM) rate into an estimate of each player's overall contribution to the team, measured vs. what a theoretical "replacement player" would provide. Where a "replacement player" is defined as a player on minimum salary or not a normal member of a team's rotation. The two models we trained to predict success are a Decision Tree Regressor and a Linear Regression.

Decision Tree Regressor Models

To create our decision tree regressor we used the "rpart" library in R. We created a training and test dataset to build and test the model with. The training dataset contains 75% of the combined college & nba dataset, with the test dataset containing the other 25%. Initially we trained models that predicted WS and VORP using all of the college basketball statistics available (GS, MP, FGs, FGA, FG%, 2pt FGs, 2pt FGA, 2pt FG%, 3pt FGs, 3pt FGA, 3pt FG%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, Tov, PF, PTS). We then fine tuned each model by using Mallow's CP variable selection method. This improved each models' accuracy and gave us insight into what variables are important to NBA success.

Variable Selection:

For predicting Win Shares we found the following college basketball per-game statistics to be the most significant: FG (field goals), 2pt FG, 3pt FG, FTA (Free Throw Attempts), FT% (Free Throw Percentage), TRB (Total Rebounds), AST (Assists), STL (Steals), BLK (Blocks), and PTS (Points Scored). These stats give a good overview of impact in a basketball game both offensively and defensively. Interestingly, the variable selection method placed importance on counting stats for a player's shooting ability, favoring the total number of field goals a player makes vs statistics like FG% or 3pt FG%. It also placed importance on a player's free throw performance, where it did consider the percentage of shots being made. A player shoots a free throw as a penalty shot after being fouled, so a player gets to take standing still 15 feet from the basket with no defense being played. Perhaps the Mallow's CP algorithm found importance with efficiency in free throws because it better projects overall shot efficiency in the NBA then field

goals taken during the game. On the other hand, counting the number of made 2 point and 3 point field goals can show a college player's importance to their team's offense.

For predicting Value Over Replacement Player we found the following college basketball per-game statistics to be the most significant: MP (minutes played), FG, 2pt FG, 3pt FG, 3pt FGA (3pth Field Goal Attempts), 3pt FG%, FT (Free Throws made), FTA, DRB (Defensive Rebounds), TRB, AST, STL, BLK, TOV (Turnovers). Compared to the Win Shares model, the VORP variable selection placed more importance on 3 point field goals and also added in minutes played and turnovers. It makes sense to place importance on 3 point field goals. NBA offenses today are significantly more motion based, in an attempt to set up 3 point field attempts, compared to previous eras of the NBA. Turnovers can show how well a player can read offenses and defenses, which could align with the more motion based offenses as well.

Model Results

Win Share Model:

Model	MAE	SSE
All College Variables	1.15	481
Selected College Variables	1.11	465

Value Over Replacement Player Model:

Model	MAE	SSE
All College Variables	1.12	390
Selected College Variables	0.34	62

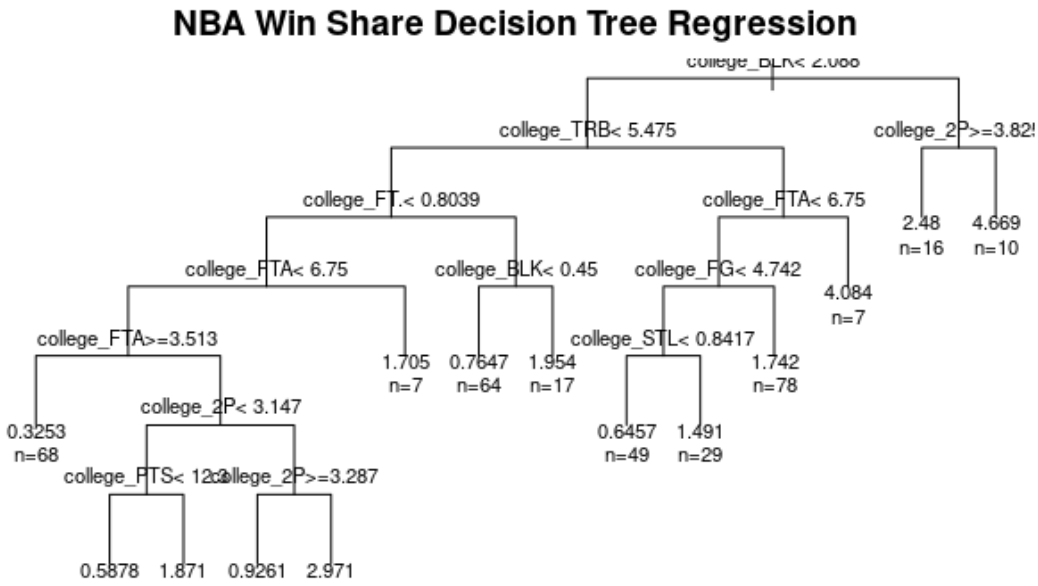
The Mallow's CP variable selection method successfully improved the accuracy of both of our models. There was great improvement for predicting VORP, going from an average absolute error of 1.12 points in Box Score Plus/Minus to 0.34. This means the decision tree model we trained was able to predict how many points a player adds or subtracts compared to a replacement-level player to within a fraction of a point. The WS model saw less of an improvement, going from an average absolute error of 1.15 games to 1.11. While the model didn't improve accuracy as much, it is still predicting Win Shares with only around 1 game of inaccuracy.

Summary based on Decision Tree Regressor Models

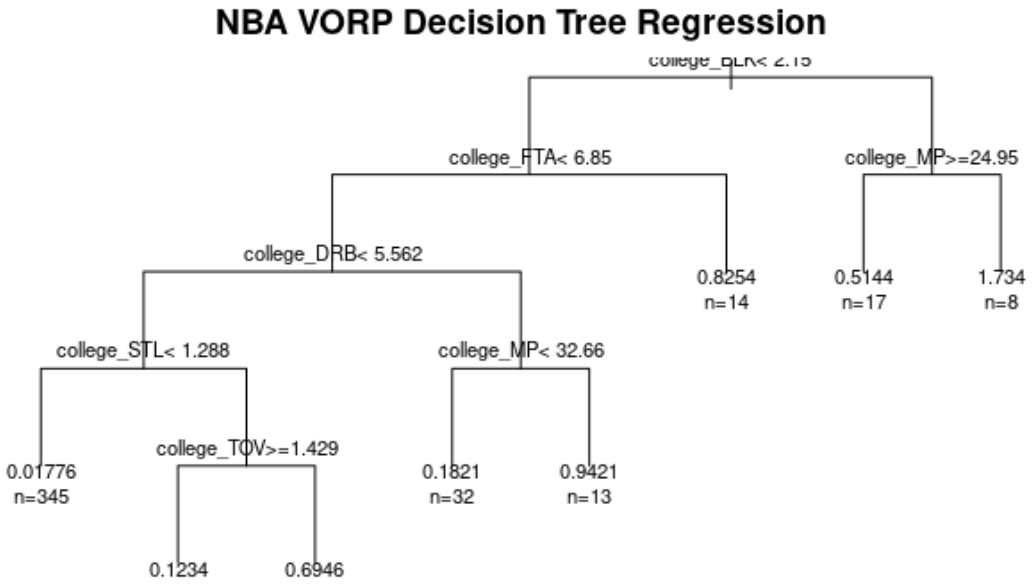
For both the Win Share and Value Over Replacement Player models favored college players with high counting stats, and only considered efficiency for shots like free throws. This makes sense because the best college players will have the ball in their hands often and accumulate a high volume of field goals made, assists, rebounds, etc. College players are still developing so even very talented players can still be inefficient, i.e. miss a significant amount of shots and turn the ball over. The VORP prediction model had better accuracy for a decision tree regressor, and

placed higher importance on 3 point performance which makes sense for the present day NBA, where 3s are more important than they are in college basketball or other eras of the NBA. The decision tree appears to split off defensive players by using variables like blocks and steals as split variables. Other important split variables are Free Throws Attempted and Minutes Played, showing how it's important to be given a large role in your team's rotation.

WS Decision Tree Regressor



VORP Decision Tree Regressor



Linear Regression Models

The combined college and NBA dataset was used to create the training(75%) and test(25%) sets for linear regression models to predict the success of current NBA players using their college experience attributes, after excluding factor variables like ”college_Player” and “college_Team”. A bidirectional stepwise regression method (forward and backward selection) was run to identify college attributes which would be significant in predicting WinShare and VORP response variables. We chose the models with the lowest Akaike Information Criterion (AIC) value from the outputs generated by the stepwise regression method.

WinShare (WS) Regression Model

Fitting the WinShare regression model produced positive coefficients for college predictors - 2-point and 3-point attempts(college_2PA and college_3PA) total rebound (college_TRB), assists(college_AST), steals(college_STL), points (college_PTS) and blocks (college_BLK), while among the remaining predictors negative coefficients like turnovers (college_TOV) and field goals attempted (colege_FGA) affect the overall WinShare measure of the player. However, assists, points, 2-point and 3-point attempts, and total rebounds are the statistically significant predictors which influence the WinShare measure of players in college and in their NBA careers.

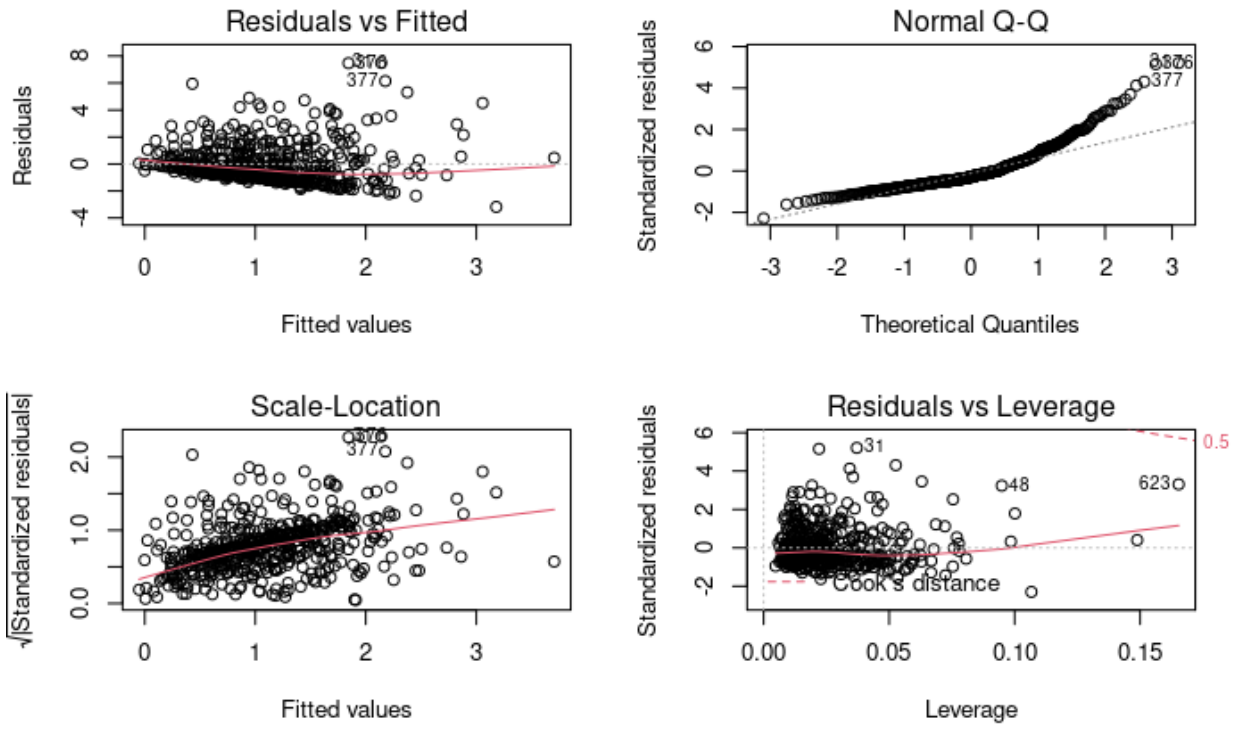


Fig. 1 : WinShare plot(s)

The residual plot shows a nonlinear relationship between the actual and predicted WinShare response variable and the Q-Q plot shows a right-skewed distribution which means that WinShare measure for most player's lies in a fixed range while there are several outlier players with a higher WinShare measure (Eg: Damian Lillard - 10.25). Outliers with greater than 2 standard deviations from the mean, as can be seen in the standardized residual vs. fitted values plot, points to the presence of an extremely unusual range of WinShare measure for some of the players.

Note: the residual std error indicates that on average a player's WinShare measure is 1.466 points away from the predicted values which clearly shows the model isn't an accurate one.

Call:
lm(formula = nba_WS ~ college_G + college_FGA + college_2PA + college_3P + college_3PA + college_DRB + college_TRB + college_AST + college_STL + college_BLK + college_TOV + college_PTS, data = as.data.frame(train_data))

Value-Over-Replacement (VORP) Regression Model

Fitting the VORP regression model produced statistically significant and positive coefficients like assists (college_AST), defensive rebounds (college_DRB), steals (college_STL), 2 & 3-PA (college_2PA & 3PA), which increases a player box-score estimate. But few significant negative coefficients like minutes played (college_MP), field goals attempted(college_FGA), 3-point %(college_3P.) and turnovers (college_TOV) indicate an impact to their overall contribution to the team which could affect their NBA careers.

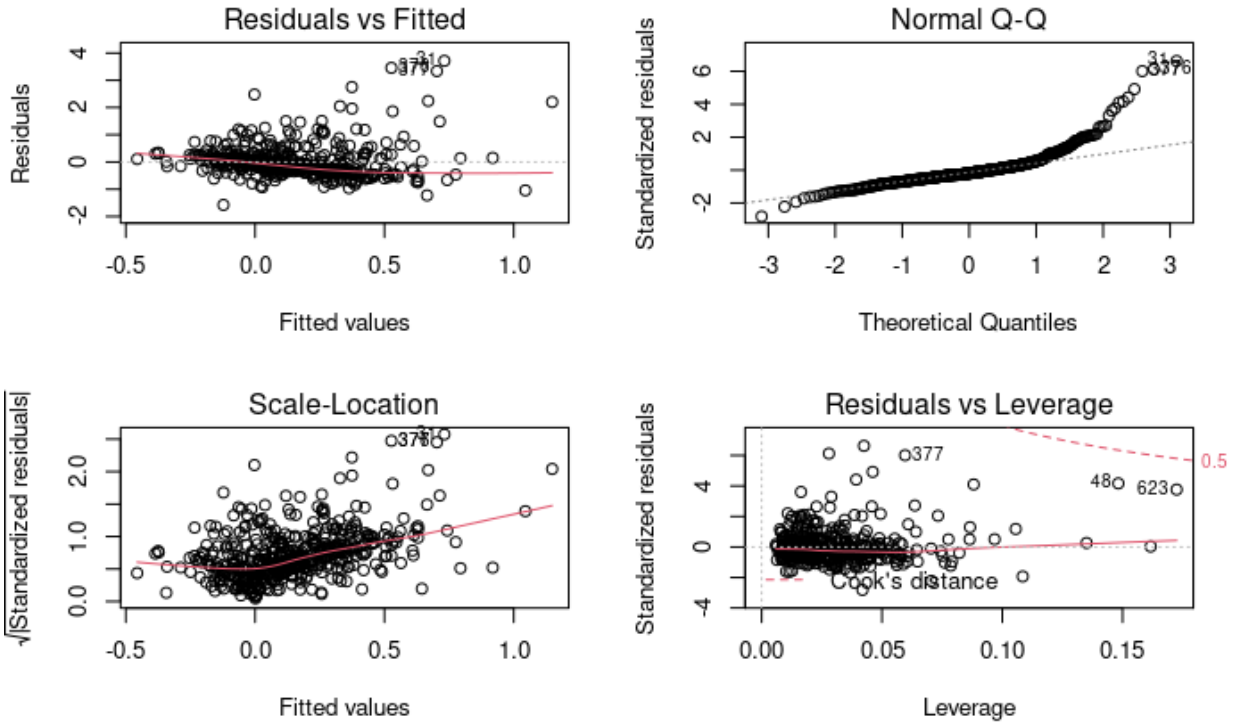


Fig. 2: VORP plot(s)

The residual plot shows a nonlinear relationship between actual and predicted VORP values and the Q-Q plot shows a right-skewed distribution which means that VORP score for most player's lies in a fixed range while there are several outliers who score more runs in a season (Eg: Damian Lillard - 4.76). Though a large number of datapoints appear concentrated near Cook's distance there are only a few influential observations that could significantly change regression the model, as can be seen in the residual vs. leverage plot. These influential points are players with very low VORP score from the mean (for e.g: Norris Cole [-0.5625], Jimmy Butler [3.125], Lonzo Ball [1.2])

Call:
lm(formula = nba_VORP ~ college_MP + college_FG + college_FGA + college_2PA + college_3PA + college_FT + college_FTA + college_DRB + college_AST + college_STL + college_BLK + college_TOV + college_3P., data = as.data.frame(train_data))

R^2, Adjusted R^2 and p-value of the 2 linear models

Models	Using all college predictors for each model			Using different selected college predictors for each model		
	R^2	Adj. R^2	p-value	R^2	Adj. R^2	p-value
WinShare	18.91%	14.93%	4.55e-12	13.52%	11.45%	7.108e-11
VORP	18.53%	14.53%	1.15e-11	13.42%	11.17%	2.413e-10

Both linear regression models have a low R-squared value and a low p-value (p-value <= 0.05). The low p-value for the regression models indicate that the model fits the data quite well even though their respective R-squared value is low.

Conclusion based on linear regression models

Output from the regression models show that the performance of players in college who are better in all offensive(assist, points, rebounds, field goal %, free throws) and defensive(steal, blocks, rebounds) areas may have a influence on their NBA careers but it also shows that players in college tend to commit more turnovers and fouls which affects their VORP and could have potential impact on their careers in NBA.

Interpretation

To know whether the results make sense, asking the tangential question of "how do basketball teams win games?" makes sense. Dean Oliver, the father of NBA basketball analytics, answered this question with what he called the "Four Factors of Basketball Success". These factors are efficient shooting, turnovers, rebounds, and free throw attempts.

When it comes to players being drafted, the logistical model found the most impactful independent variable to be high number of games played. The other variables in order of significance were blocks, assists, turnovers, minutes played, field goal percentage, field goals made, free throw percentage, defensive rebounds, and games started. As far as what is distinctly missing is anything regarding free throw attempts per field goal attempts.

Looking at the decision trees predicting future NBA VORP and WS, the variables that show up in both models are blocks, steals, and free throw attempts. Those last two were not in the drafted logistical model. Number of assists is missing from the decision trees despite being the third most significant factor in being drafted. Other factors that appear once in the decision tree models are two pointers made, free throw percentage, points, free throws made, defensive rebounds, total rebounds, and turnovers.

In the linear regression models, field goal attempts (2 point, 3 point, and total) appears in both VORP and WS models. Also, differentiating the linear regression models from the decision trees is that assists appears in both linear regression models while appearing in neither of the decision tree models. Again, steals and low turnovers makes an appearance in both the linear regression models like they did in the decision tree models. Other variables that appear in these models are blocks, points, 3 point percentage, defensive rebounds, and total rebounds.

So looking at these three different types of modeling there are some areas where the drafted players logistic model and the player performance, decision trees and linear regression, models differ. The main variables that differ are amount of games played, free throw attempts per field goal attempts, steals, and points. This makes sense as NBA teams have been known to look for

offensive stars over efficient scorers and defenders. The models were trained for accuracy, so looking for quality role players more than looking for highest potential.

As far as future work there are a few things we can add such as physical metrics, player positions, and on and off the court splits.

Reference

1. Zestcott, Colin & Dickens, Jessie & Bracamonte, Noah & Stone, Jeff & Harrison, C.. (2020). One and Done: Examining the Relationship Between Years of College Basketball Experience and Career Statistics in the National Basketball Association. Journal of Sport and Social Issues. 44. 019372352091981. 10.1177/0193723520919815.
2. Otten, Mark & Miller, Travis. (2015). A BALANCED TEAM WINS CHAMPIONSHIPS: 66 YEARS OF DATA FROM THE NATIONAL BASKETBALL ASSOCIATION AND THE NATIONAL FOOTBALL LEAGUE 1 ,2. Perceptual and Motor Skills. 121. 10.2466/30.26.PMS.121c25x4.