Mateen Hussaini
On-Uma Lomsomboot
Due 5/11

Analysis and Prediction of NBA MVP Award Winners

Introduction

During each season, 100 NBA reporters and analysts receive ballots from the National Basketball Association(NBA) where they can rank the 1st, 2nd, 3rd, 4th, and 5th Most Valuable Player(MVP) in the NBA. Each vote a player receives awards a point: 1st = 10 points, 2nd = 7 points, 3rd = 5 points, 4th = 3 points, and 5th = 1 point. The ballots have to be submitted before the end of the regular season and the winner is announced by the NBA during the postseason. Voters are allowed to publicly state who they voted for before the deadline, such as Zach Lowe. There are basketball fans who have created an Award Tracker spreadsheet that tracks all votes that have been publicly stated. At the time this was written, 5/10/22, and before the vote was officially announced, the spreadsheet identified 56 1st, 42 2nd, 38 3rd, 29 4th, and 21 5th place votes for MVP. So any fan who read articles written by voters and saw the Award Tracker Spreadsheet would have already determined that Nikola Jokic would become the NBA 2022 MVP weeks before it was officially announced. The goal of this project is to use machine learning to explore new ways to evaluate the MVP Award winners.

Results and Discussion

The training data we used was posted by Kaggle user Vivo Vinco and it was scraped from Basketball-Reference. The data contained player stats, mvp votes, and team stats from the 1991 season to the 2021 season. We used Vivo Vinco's code to scrape the 2022 player stats to use as test data. The total data contained 14092 observations: 31 MVPs, 440 MVP candidates, and 13621 Non-Candidates. There were 44 variables, both quantitative and categorical. We then created the categorical response variable called MVP which we labeled 0,1,2 for Non-Candidates, MVP, and MVP Candidates. In retrospect, we should have made the order Non-Candidates, MVP Candidates, and MVP.

The last step before starting analysis was to standardize the data. Since rules and philosophies evolve season to season, we decided to standardize our data by season. We can see below in figure 1 that while KD, Curry, and MJ both averaged 30.1 points per game(PPG) in different seasons, Curry's PPG is higher once standardized to the season.

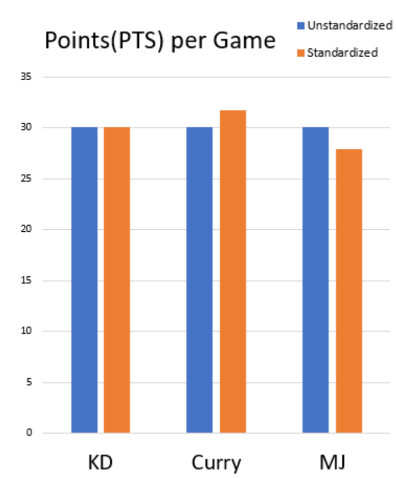

Figure 1

The 2nd step was to narrow down variables down from 44 to a smaller amount (Figure A in the appendix contains all variables in our data). We removed variables we felt were unnecessary such as PTS Won and PTS Max, which calculates the points a player earns from their votes and the maximum number of points a player can earn by vote respectively. We also removed double counting stats e.g we removed offensive rebounds(ORB) and defensive rebounds(DRB) because we already had total rebounds(TBR). We ended up with 15 quantitative variables and we believe that a better variable selection method should be used in the future. The 15 variables we selected are in Figure 2 below. TS% is a variable we added and used instead of 3P%, 2P%, FT% because we wanted to measure overall shot making efficiency in a player as opposed to individual efficiency of each type of shot. W/L% and SRS are team stats and the rest are player stats. In Figure 3 below, you can see the distributions of each stat. Most of them are not normally distributed.

Age	Age of Player
GS	Games Started
3PA	3 Pointers Attempted
2PA	2 Pointers Attempted
FTA	Free Throws Attempted
TRB	Total Rebounds
AST	Assist
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls
PTS	Points Scored
W/L%	Percentage of games teams has won
SRS	Simple Rating System from Basketball Reference. Used to measure value of wins e.g A 20 point win over the best team increases SRS e.g A 20 point lost to the worst team increases SRS
TS%	Combination of % of FGA and FTA PTS/(2*FGA+(0.44*FTA)

Figure 2. 15 variables used in decision tree ranking.

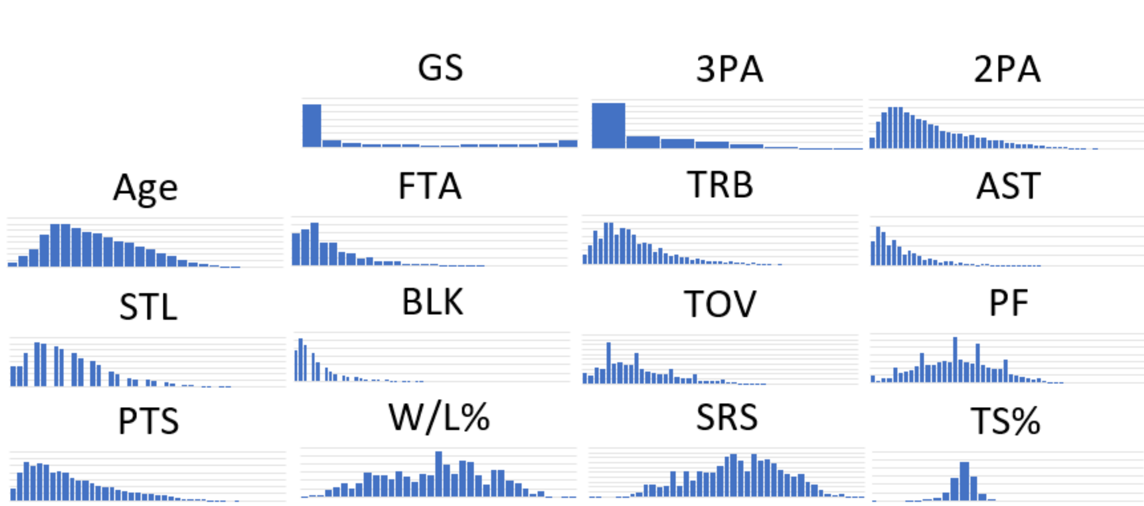

Figure 3

-

-

Next we used decision trees from the scikit-learn package for python to rank the top ten most important variables. We used both the Extra Trees and Random Forest model to rank the most important variables out of the 15. We ran the Extra Trees model 500 times and Random Forest model 100 times and took the average value of importance for each variable. Due to time and technology restraints, we ran Random Forest 100 times instead of 500. We ran the models with the past 31 years, past 15 years, and past 10 years of data to make a total of 6 models. In Figure 4 below we have a table that shows the value of each variable. In both the models, PTS(points) and FTA(free throw attempts), are the 1st and 2nd most important variables. Assists were ranked more important than rebounds, and the two team stats, W/L% and SRS were in the top 10 for variable importance.

Determing Variable Importance using Randomized Decision Trees (Extra-Trees)						Determining Variable Importance with Random Forest Classifier					
Rank	Stat	31 ET	15 ET	10 ET	Avg	Rank	Stat	31 RF	15 RF	10 RF	Avg
1	PTS	0.152	0.150	0.151	0.151	1	PTS	0.182	0.176	0.191	0.183
2	FTA	0.114	0.122	0.116	0.117	2	FTA	0.110	0.118	0.114	0.114
3	2PA	0.090	0.083	0.078	0.084	3	W/L%	0.099	0.094	0.081	0.091
4	W/L%	0.085	0.082	0.077	0.081	4	2PA	0.082	0.079	0.076	0.079
5	AST	0.072	0.077	0.081	0.077	5	SRS	0.077	0.078	0.076	0.077
6	SRS	0.071	0.073	0.070	0.071	6	AST	0.066	0.070	0.075	0.070
7	TOV	0.064	0.068	0.065	0.066	7	TOV	0.066	0.067	0.062	0.065
8	GS	0.054	0.054	0.050	0.052	8	GS	0.049	0.053	0.044	0.049
9	STL	0.049	0.047	0.052	0.050	9	TS%	0.045	0.046	0.047	0.046
10	TRB	0.055	0.045	0.046	0.049	10	TRB	0.053	0.043	0.041	0.045
11	TS%	0.046	0.048	0.050	0.048	11	STL	0.044	0.044	0.047	0.045
12	BLK	0.042	0.037	0.038	0.039	12	3PA	0.030	0.037	0.046	0.038
13	Age	0.035	0.037	0.040	0.037	13	Age	0.033	0.033	0.036	0.034
14	PF	0.035	0.037	0.037	0.036	14	BLK	0.034	0.031	0.031	0.032
15	3PA	0.035	0.041	0.050	0.042	15	PF	0.032	0.032	0.032	0.032

Figure 4.

We also analyzed the difference between each model and the mean of its respective model type. In Figure 4, you can see that the importance of 2PA in the Extra Trees model dropped from .090 to .083 to . 078 as we included less distant data. We made these changes to see in Figure 5 below. The variables that have decreased in importance for more recent data are 2PA(2 pointers attempted), TRB(total rebounds), and W/L%(Win % of team). TS%(True shooting %) has consistently increased when we look at more recent data. One explanation for the trends is that better shot selection means that players make more shots, therefore TS% increases, there are less rebounds to secure, and players need to attempt less shots to score the same number of points.

Model Variable Importance Compared to Extra-Trees Mean					Model Variable Importance Compared to Random Forest Mean				
Rank	Stat	31 ET	15 ET	10 ET	Rank	Stat	31 RF	15 RF	10 RF
1	PTS	0.001	-0.001	0.000	1	PTS	-0.001	-0.007	0.008
2	FTA	-0.004	0.004	-0.001	2	FTA	-0.004	0.004	0.000
3	2PA	0.007	-0.001	-0.006	3	2PA	0.007	0.003	-0.010
4	W/L%	0.004	0.001	-0.005	4	W/L%	0.003	0.000	-0.003
5	AST	-0.005	0.000	0.004	5	AST	0.000	0.001	-0.001
6	SRS	0.000	0.002	-0.002	6	SRS	-0.004	0.000	0.005
7	TOV	-0.002	0.002	-0.001	7	TOV	0.001	0.002	-0.003
8	GS	0.001	0.001	-0.003	8	GS	0.000	0.004	-0.004
9	STL	0.000	-0.002	0.002	9	STL	-0.001	0.000	0.001
10	TRB	0.007	-0.004	-0.003	10	TRB	0.007	-0.003	-0.005
11	TS%	-0.002	0.000	0.002	11	TS%	-0.001	-0.001	0.002
12	BLK	0.003	-0.002	-0.001	12	BLK	-0.008	-0.001	0.008
13	Age	-0.002	0.000	0.002	13	Age	-0.001	-0.001	0.002
14	PF	-0.002	0.001	0.001	14	PF	0.002	-0.001	-0.001
15	3PA	-0.007	-0.001	0.008	15	3PA	0.000	0.000	0.000

Figure 5.

For Principal Component Analysis(PCA), we originally used the top 10 most important variables, as determined by the Extra-Trees model. Then we removed W/L% and SRS because it did not make sense to include those variables when describing playstyles of players. We believe this model can be improved if we increase the number of variables to the 15 variables from the decision trees minus W/L% and SRS, but due to time-constraints we will not make any changes. Therefore we performed PCA for the past 31, 15, and 10 years with variables: PTS, FTA, 2PA AST, TOV, GS, STL, TRB. All principal components and the summary of the components is available in the Appendix as Figure B and C. The PCA analysis showed that the first two principal components summarized more than 70% of the variance in the data as you can see below in Figure 6.

Importance of components	Past 31 Years		Past 15 Years		Past 10 Years	
	PC1	PC2	PC1	PC2	PC1	PC2
Standard deviation	2.4746	1.1292	2.4676	1.1092	2.4654	1.0968
Proportion of Variance	0.6137	0.1278	0.6101	0.1233	0.6089	0.1205
Cumulative Proportion	0.6137	**0.7415**	0.6101	**0.7333**	0.6089	**0.7294**

Figure 6

In Figure 7, you can see the values of the first two components. The rows are colored, where green values are higher and red values are lower. We can see that the values for PC1 do not change much based on the length of the data. The variables ranked in order for 31 yr PC1 are PTS, TOV, 2PA, FTA, GS, STL, AST, TRB, and 3PA. We see PTS, TOV, 2PA, and FTA being some of the most responsible for PC1 scores with TS% being the least responsible. For PC2, we can see that the TS% is important for the past 10 years of data, but not as important for the past 31 and 15 years. We can also see the TRB and TS% are really positive, while 3PA and AST are really negative.

	Visually seeing Magnitudes of Principal Components									
Data	PTS	FTA	2PA	3PA	TS%	AST	TOV	TRB	STL	GS
31 yr PC1	0.388	0.355	0.367	0.211	0.158	0.303	0.370	0.272	0.322	0.334
31 yr PC2	0.013	0.188	0.193	-0.595	0.206	-0.407	-0.061	0.541	-0.248	0.086
15 yr PC1	0.387	0.354	0.368	0.204	0.167	0.297	0.371	0.272	0.326	0.338
15 yr PC2	0.032	0.220	0.192	-0.602	0.106	-0.428	-0.031	0.528	-0.267	0.078
10 yr PC1	0.389	0.356	0.365	0.228	0.140	0.315	0.368	0.275	0.316	0.327
10 yr PC2	-0.005	0.136	0.184	-0.544	0.391	-0.371	-0.119	0.541	-0.207	0.101

Figure 7

It can be seen much more easily using a radar graph. The PC1 values are nearly stacked on top of each other and PC2 have nearly all vertices stacked over each other except for TS% and FTA which have slightly changed. If we were to split the data from 1991-2000, 2001-2010, and 2011-2021 we may see some further patterns, but that is beyond the scope of predicting the 2022 MVP. We wanted to evaluate whether or not more data aids in creating a more accurate algorithm.

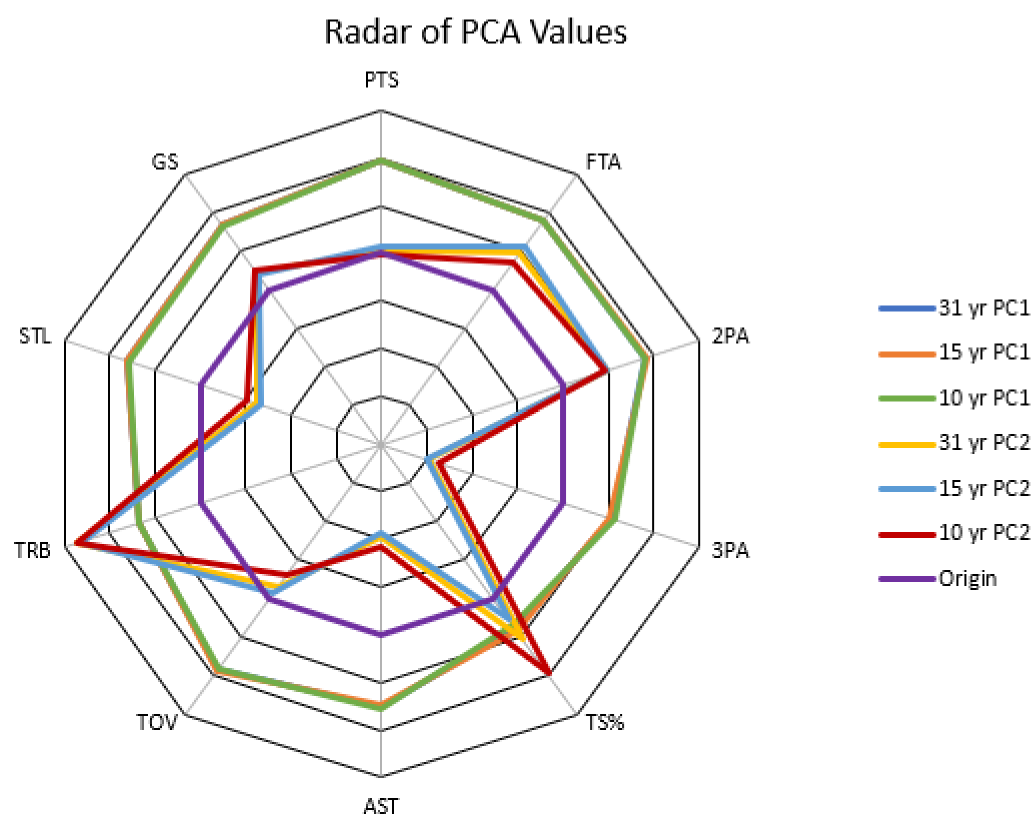

Figure 8

In the biplots created below by the first and second component, we can see the MVP candidates (blue points) are clustered near the high PC1 values, but they are not affected by PC2 values. Rocha de Silva and Rodrigues from [All-NBA Teams' Selection Based on Unsupervised Learning] suggested that on page 162, that PC1 is highly related to player performance and PC2 is related to playstyle where more positive

PC2 values describes guards, values close to zero describe forwards, and more negative values describe big men.

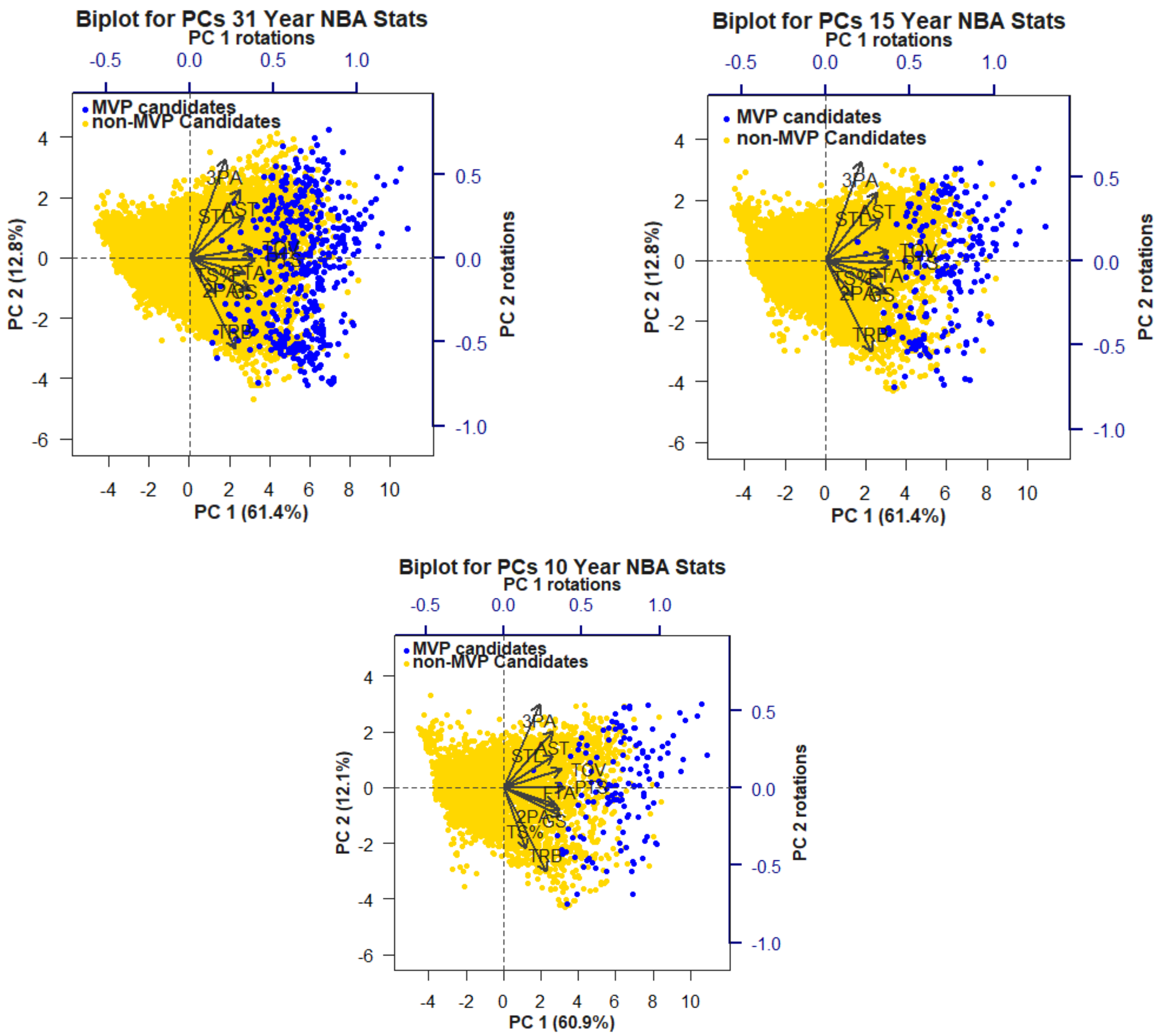

Figure 9

For logistic regression, we aimed to use the algorithm to predict 2022 MVP and 2022 MVP candidate. We utilized the scikit-learn package by two classifiers: extra-tree classifier and random forest classifier to train the model. We created separate notebooks to train different models : past 10 years, past 15 years, and past 31 years. For feature selection, we first manually removed features according to our own basketball's features' analysis. With the remaining features, we reduced dimension using feature importance which is an inbuilt class that comes with Tree Based Classifiers passing the whole independent variables. To make the feature importance even more robust, we ran the fitting for 100 times, and took the average of all runs to indicate the final most important features regarding its testY (true label) or its accuracy to predict the model. We trimmed the number of features down to 10 features by sorting the highest features to the last

10th. We splitted data into training and testing in an 80:20 portion randomly passing X from trimmed features and y as MVP column : 0 = non MVP candidate, 1= MVP, 2= MVP candidate to trainX, testX, trainY, testY. Because our dataset has 3 target classes, we applied Multinomial Logistic Regression or known as Softmax Regression to predict these multiple classes mentioned earlier. For the purpose of analysis, prior to fitting the model, we saved categorical features before removing which are player name columns that will output alongside the 2022's test on the testing process in the dictionary having the index of each sample as the dictionary's key and the according value as the dictionary's value. For the evaluation, y_pred is obtained by the trained model predict() function with testX. The output will be either 0,1, or 2. To get accuracy, we passed testY and y_pred to accuracy_score() function from scikit learn metrics. For visualization, we plotted the graph by the correct classification on our trained models. We splitted the data into class 0, 1, 2 and used matplotlib to see how far apart each class is compared to others as figure 10 below.

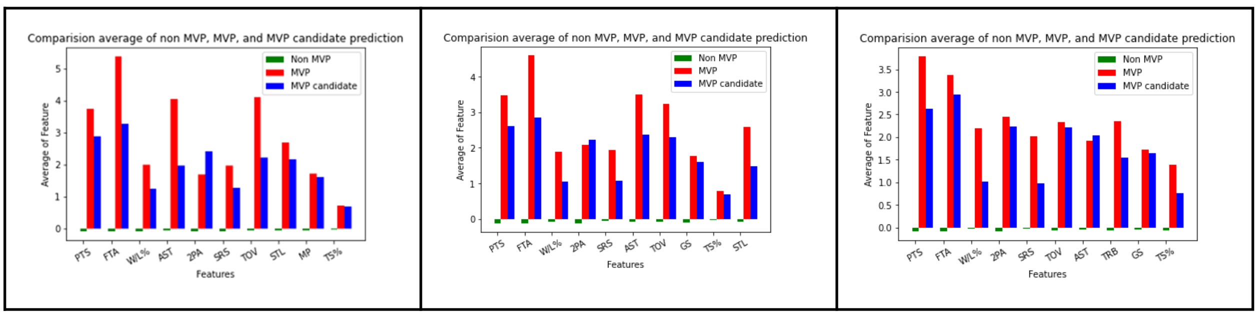

Figure 10

For the testing process of 2022 MVP, we took the whole columns and trimmed the features as when we trained our trained models. It is required to have the same features. As mentioned above, y_pred will output in discrete numbers 0,1,2; however, we would like to get the range of who would be MVP, and MVP candidates so we then called the predict_proba()function. We created the columns MVP% and Candidat% to store the probability of class 1 (MVP) and class 2 (Candidate).

To determine the MVP, we sorted data by column MVP% from highest to lowest reading only the top 20 players in the row; We have that information in the Appendix. The algorithms were best at classifying Non-Candidates, and the worst at classifying MVPs. This may have been caused by imbalance in data. The 31yr algorithms were the most accurate at classifying each category, while 10yr data was worst at classifying each category.

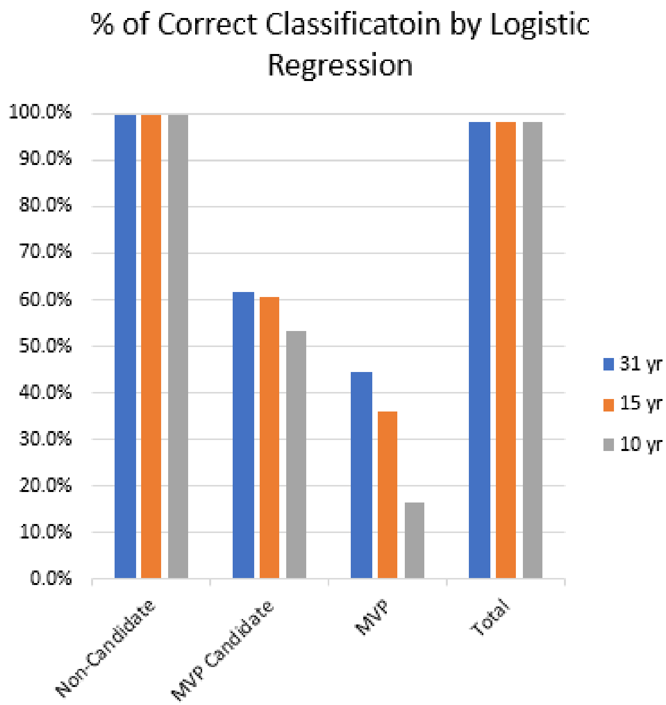

Figure 11

We want to compare our results with Basketball Reference's algorithm and the initial results leaked by voters. The biggest difference between our algorithms and Basketball Reference is that our algorithms value Ja Morant very highly. One reason for the discrepancy is that Morant's team had a 80% win rate when he didn't play in the games, and they had a 63% when he played. Morant's probability may lower in algorithms that use Player win% instead of team win%.

Logisitic Regression

Random Forest Classifier			Extra Trees Classifier			Basketball Reference Algorithm		Votes Counted as of 5/06					
10 yr Prob%	15 yr Prob%	31 yr Prob%	10 yr Prob%	15 yr Prob%	31 yr Prob%	BR Prob%	Player (Team)	1st Place Votes	2nd place votes	3rd place votes	4th place votes	5th place votes	TOTAL POINTS
7.3%	5.4%	30.1%	7.1%	5.3%	23.0%	43.5%	Nikola Jokić	37	10	1	0	0	445
9.8%	15.8%	15.3%	10.4%	16.2%	18.5%	12.4%	Joel Embiid (76ers)	11	15	18	0	0	305
13.1%	19.9%	22.0%	13.8%	20.2%	23.4%	24.3%	Giannis Antetokounmpo	6	15	17	0	0	250
7.5%	21.2%	21.0%	7.5%	21.5%	31.4%	2.2%	Devin Booker (Suns)	0	1	1	17	5	68
5.8%	7.1%	11.7%	6.0%	6.9%	18.9%	4.5%	Luka Dončić	0	1	0	10	15	52
0.8%	0.7%	1.2%	1.0%	0.7%	2.3%	2.5%	James Harden (76ers)	54	42	37	27	20	1120
2.1%	2.0%	4.9%	2.1%	2.0%	7.9%	5.4%	Chris Paul (Suns)	Tabe above curated by Max Croes, @CroesFire, /u/TexasAlaskaMontana					
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.1%	Rudy Gobert	← This player is considered one of the best defensive players in the NBA					
2.1%	3.5%	2.0%	2.0%	3.3%	2.1%	1.6%	Trae Young						
0.0%	1.7%	1.9%	1.2%	0.6%	4.1%	1.5%	Jayson Tatum	← Won 4th and 5th place votes			4	7	19
15.5%	20.3%	15.0%	15.9%	20.5%	16.1%	N/A	Ja Morant	← Team won 80% of games without Ja as opposed to 63% when he played					
2.5%	5.3%	2.7%	2.5%	5.0%	1.9%	N/A	DeMar DeRozan	← Was considered top 5 MVP for 1st half of season					

Figure 12. Our algorithms compared to leaked results

If you look at Figure 13 below, You can see that our algorithm's give the win to Devin Booker and consides Booker, Antetokounmpo, and Morant as the top 3 MVP candidates. These are results we did not expect. Booker did not receive much conversation in the top 3 during the regular season, maybe due to Chris Paul and DeAndre Ayton taking some credit. If you were to look at Figure 14, you can see that the top 3 candidates were decidedly, Jokic, Embiid, and Antetekounmpo since they received only 10 out of the 300 votes outside the top 3.

	1st	2nd	3rd	4th	5th	Points
Devin Booker (Suns)	30	0	5	6	0	41
Giannis Antetokounmpo	0	28	10	0	0	38
Ja Morant	20	14	0	0	1	35
Nikola Jokić	10	0	5	3	2	20
Joel Embiid (76ers)	0	0	10	6	1	17
Luka Dončić	0	0	0	3	2	5

Figure 13

The final results were announced May 11th. Our results did not accurately reflect the MVP race. One positives is that our algorithm was able to predict Ja Morant to have earned votes, when Basketball Reference didn't. While our algorithm wasn't able to predict the ultimate winner and wasn't able to find a correct permutation, it did well in finding the top combinations of players. It was able to correctly predict 8 out of the top 10.

Logisitic Regression													
Random Forest Classifier			Extra Trees Classifier			Basketball Reference Algorithm		Official Vote Count released May 11th					
10 yr Prob%	15 yr Prob%	31 yr Prob%	10 yr Prob%	15 yr Prob%	31 yr Prob%	BR Prob%	Player (Team)	TOTAL POINTS	1st Place Votes	2nd place votes	3rd place votes	4th place votes	5th place votes
7.3%	5.4%	30.1%	7.1%	5.3%	23.0%	43.5%	Nikola Jokić	875	65	27	6	2	0
9.8%	15.8%	15.3%	10.4%	16.2%	18.5%	12.4%	Joel Embiid (76ers)	706	26	39	34	1	0
13.1%	19.9%	22.0%	13.8%	20.2%	23.4%	24.3%	Giannis Antetokounmpo	595	9	32	52	7	0
7.5%	21.2%	21.0%	7.5%	21.5%	31.4%	2.2%	Devin Booker (Suns)	216	0	1	8	49	22
5.8%	7.1%	11.7%	6.0%	6.9%	18.9%	4.5%	Luka Dončić	146	0	1	0	32	43
0.0%	1.7%	1.9%	1.2%	0.6%	4.1%	1.5%	Jayson Tatum	43	0	0	0	8	19
15.5%	20.3%	15.0%	15.9%	20.5%	16.1%	0.0%	Ja Morant	10	0	0	0	1	7
0.2%	0.2%	0.4%	0.2%	0.2%	0.9%	0.0%	Steph Curry	4	0	0	0	0	4
2.1%	2.0%	4.9%	2.1%	2.0%	7.9%	5.4%	Chris Paul (Suns)	2	0	0	0	0	2
2.5%	5.3%	2.7%	2.5%	5.0%	1.9%	0.0%	DeMar DeRozan	1	0	0	0	0	1
0.1%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	Lebron James	1	0	0	0	0	1
0.0%	2.2%	1.8%	0.0%	2.2%	1.3%	0.0%	Kevin Durant	1	0	0	0	0	1
0.8%	0.7%	1.2%	1.0%	0.7%	2.3%	2.5%	James Harden (76ers)	0	0	0	0	0	0
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.1%	Rudy Gobert	0	0	0	0	0	0
2.1%	3.5%	2.0%	2.0%	3.3%	2.1%	1.6%	Trae Young	0	0	0	0	0	0

Figure 14

While this algorithm didn't predict the MVP accurately, we think this algorithm is a great first step in exploring how machine learning can be applied in sports.

For rule based learning, because the goal of rule based learning is to find the minimum qualification of MVP and MVP candidate instead of overall target class to 0,1,2, class 1 and 2 are combined to class 1. Therefore, there are 2 target classes : 0 for non MVP candidate, 1 for MVP and MVP candidate. Rule Based Learning was trained by passing the training and testing set in the portion of 80 to 20 percent. The decision tree created the rules passed by mode type "r" as rule. Binary classification is applied to the model as being specified as "classify" in order to determine the least qualifications to be considered MVP and MVP candidate.

10 year-trained model's rule:

	rule	type	coef	support	importance
0	PTS > 2.314630150794983 & W/L% > 0.8920324742794037	rule	1.714740	0.016667	0.219519
1	PTS <= 2.314630150794983	rule	-4.229046	0.966667	0.759137

The minimum qualification of being an MVP and MVP candidate from the past 10 year-trained model is to have a standardized score of PTS (Points Scored) greater than roughly 2.31 and standardized score of W/L% (Percentage of games teams has won) greater than approximately 0.89 according to the first rule created above. Below is the review of overall class 0 and 1 for this model.

Class 0 :

```
class_0 = df.iloc[np.where(df.Is_MVP_or_MVP_candidate == "0")]
class_0
```

	Age	G	MP	3P	3PA	2PA	FTA	TRB	AST	STL	BLK	TOV	PF	PTS	W/L%	SRS	TS%	Is_MVP_or_MVP_candidate
6829	-0.554215	-1.279530	-1.379770	-0.751925	-0.867548	-0.936267	-0.555623	-0.809237	-0.923072	-1.522622	0.496858	-0.765335	-0.922376	-1.027864	-0.706557	-0.587617	-0.527129	0
6830	-1.244852	-2.044464	-1.832327	-1.016637	-0.970548	-1.168290	-1.076141	-1.179838	-1.036315	-1.522622	-0.940482	-1.021237	-1.871536	-1.391023	-0.706557	-0.587617	-5.835876	0
6831	0.366635	0.330859	-0.816833	-0.619568	-0.764549	-0.936267	-0.729129	-0.768060	-0.526721	0.425372	-0.461368	-0.509434	-0.515593	-0.978342	-0.706557	-0.587617	-0.433768	0
6832	1.517696	0.451638	-0.353238	0.174569	0.265446	-0.646239	0.080565	-0.644526	0.152739	0.181873	-0.700925	-0.253532	-0.108810	-0.285039	-0.706557	-0.587617	0.269501	0
6833	2.668758	0.008781	0.717446	0.968705	0.883443	1.151939	0.138400	1.208476	-0.186991	-0.061626	0.736415	-0.253532	0.569161	0.953004	-0.706557	-0.587617	-0.018536	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13629	1.568638	1.222186	0.534912	1.544233	1.531471	-0.484485	-0.407388	-0.808845	0.221136	-0.045857	-1.015774	-0.112113	-0.573363	0.286494	-0.268188	-0.289204	0.234289	0
13630	-0.378637	-1.095907	-1.474066	-0.997428	-1.105208	-0.836285	-0.591070	-1.268654	-0.851278	-0.561747	-0.771009	-0.709166	-0.968617	-1.024449	-0.268188	-0.289204	-0.105987	0
13631	2.055457	0.980718	0.191129	0.770684	0.696523	0.365698	-0.101252	0.486979	-0.315071	0.212088	0.452815	-0.112113	0.217146	0.379031	-0.268188	-0.289204	-0.124223	0
13632	-1.108865	-0.274916	-1.345147	-0.997428	-1.193097	-0.689702	-0.774752	-1.268654	-0.475933	-1.077636	-1.015774	-0.828577	-1.627375	-0.993603	-0.268188	-0.289204	-0.063138	0
13633	-0.135227	-0.951027	-0.453462	-0.444893	-0.489983	-0.631069	-0.468615	0.027171	-0.744037	-0.819691	-1.015774	-0.947987	-1.363872	-0.608032	-0.268188	-0.289204	0.299021	0

4885 rows × 18 columns

Class 1 :

```
class_1 = df.iloc[np.where(df.Is_MVP_or_MVP_candidate == "1")]
class_1
```

	Age	G	MP	3P	3PA	2PA	FTA	TRB	AST	STL	BLK	TOV	PF	PTS	W/L%	SRS	TS%	Is_MVP_or_MVP_candidate
6919	0.136422	0.894494	1.534256	3.218760	3.252431	1.790002	3.839859	-0.356281	2.304360	0.668871	-0.461368	2.177533	0.704756	3.379568	1.101833	0.567176	1.083640	1
7007	-0.784427	0.854235	1.788130	-0.354856	-0.198051	4.023223	3.897694	3.390901	0.152739	1.642869	4.329765	1.665729	0.704756	3.231003	-0.610208	-0.384683	0.605381	1
7048	0.366635	0.330859	1.490104	1.498130	1.449940	1.964019	2.509647	1.949677	1.681522	1.155870	2.892425	1.409828	0.297973	2.752293	2.369188	2.765632	1.396498	1
7054	0.366635	1.015273	1.490104	4.409966	4.024928	1.035928	1.584282	0.384919	2.700711	2.860365	-0.461368	2.433434	0.840350	2.785308	2.369188	2.765632	1.061002	1
7080	0.366635	1.095793	1.622560	2.292267	2.582935	3.501171	4.938730	2.937944	4.852332	2.373366	0.017745	5.504253	0.840350	3.825263	0.560798	0.299013	0.290828	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13400	-0.378637	0.739250	1.351394	-0.997428	-1.149153	1.626315	1.919248	1.490198	2.634066	2.533592	0.452815	2.276098	1.666412	0.826294	1.321222	1.129245	0.370501	1
13408	0.108182	0.401195	1.211733	0.107642	0.081298	3.004199	5.470428	2.911425	0.435618	0.985923	2.410934	2.395509	1.007655	3.016339	1.321222	1.129245	0.909177	1
13444	0.108182	0.884131	1.415854	0.107642	0.344966	2.945566	4.735701	3.078628	2.097860	1.501812	1.921404	2.753741	1.534661	2.954648	1.021871	1.189209	0.857341	1
13482	-1.108865	1.125599	1.555515	2.096768	2.410364	2.300599	3.266247	1.824605	3.545618	0.985923	0.208050	3.828436	0.875903	2.892957	0.622737	0.504797	0.398510	1
13522	0.595001	1.367067	1.179503	-1.107935	-1.193097	1.098615	2.164157	4.123648	-0.368692	-0.045857	5.592878	0.723761	0.875903	0.826294	1.613446	1.892230	1.299101	1

133 rows × 18 columns

15 year-trained model's rule:

	rule	type	coef	support	importance
0	FTA <= 2.8089007139205933	rule	-4.132595	0.969479	0.710868
1	FTA > 2.8089007139205933	rule	0.097159	0.030521	0.016713

The minimum qualification of being an MVP and MVP candidate from the past 15 year-trained model is to have a standardized score of FTA (Free Throws Attempted) greater than roughly 2.8 according to the second rule created above. Below is the review of overall class 0 and 1 for this model.

Class 0 :

```
class_0 = df.iloc[np.where(df.Is_MVP_or_MVP_candidate == "0")]
class_0
```

	Age	G	MP	3P	3PA	2PA	FTA	TRB	AST	STL	BLK	TOV	PF	PTS	W/L%	SRS	TS%	Is_MVP_or_MVP_candidate
6829	-0.554215	-1.279530	-1.379770	-0.751925	-0.867548	-0.936267	-0.555623	-0.809237	-0.923072	-1.522622	0.496858	-0.765335	-0.922376	-1.027864	-0.706557	-0.587617	-0.527129	0
6830	-1.244852	-2.044464	-1.832327	-1.016637	-0.970548	-1.168290	-1.076141	-1.179838	-1.036315	-1.522622	-0.940482	-1.021237	-1.871536	-1.391023	-0.706557	-0.587617	-5.835876	0
6831	0.366635	0.330859	-0.816833	-0.619568	-0.764549	-0.936267	-0.729129	-0.768060	-0.526721	0.425372	-0.461368	-0.509434	-0.515593	-0.978342	-0.706557	-0.587617	-0.433768	0
6832	1.517696	0.451638	-0.353238	0.174569	0.265446	-0.646239	0.080565	-0.644526	0.152739	0.181873	-0.700925	-0.253532	-0.108810	-0.285039	-0.706557	-0.587617	0.269501	0
6833	2.668758	0.008781	0.717446	0.968705	0.883443	1.151939	0.138400	1.208476	-0.186991	-0.061626	0.736415	-0.253532	0.569161	0.953004	-0.706557	-0.587617	-0.018536	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14087	-0.091724	1.113670	1.025399	0.617089	0.763783	0.587060	0.557555	0.723273	0.570653	0.877153	-0.182383	0.660981	0.628128	0.730006	-0.537997	-0.676847	0.302459	0
14088	-0.555920	-0.563355	-1.363893	-0.774400	-0.660391	-1.038473	-0.818088	-0.967234	-0.698884	-0.495522	-0.585917	-0.663611	-1.272335	-1.066681	-0.537997	-0.676847	-0.908281	0
14089	-0.555920	-0.604258	-1.099466	-0.619790	-0.541710	-0.886079	-0.563339	-1.128235	-0.809279	-0.724301	-0.585917	-0.904446	-1.628672	-0.907682	-0.537997	-0.676847	-0.480678	0
14090	-1.484312	-0.358839	-0.806706	-0.155960	-0.185666	-0.759085	-0.919988	-0.685483	-0.698884	-1.181859	-0.384150	-0.904446	-0.559662	-0.685084	-0.537997	-0.676847	0.202020	0
14091	-0.091724	0.541028	0.619314	0.617089	0.467080	0.282272	-0.053842	1.045274	0.073878	0.190816	0.422918	-0.181941	1.103244	0.332509	-0.537997	-0.676847	0.370827	0

7059 rows × 18 columns

Class 1:

```
] class_1 = df.iloc[np.where(df.Is_MVP_or_MVP_candidate == "1")]
  class_1
```

	Age	G	MP	3P	3PA	2PA	FTA	TRB	AST	STL	BLK	TOV	PF	PTS	W/L%	SRS	TS%	Is_MVP_or_MVP_candidate
6919	0.136422	0.894494	1.534256	3.218760	3.252431	1.790002	3.839859	-0.356281	2.304360	0.668871	-0.461368	2.177533	0.704756	3.379568	1.101833	0.567176	1.083640	1
7007	-0.784427	0.854235	1.788130	-0.354856	-0.198051	4.023223	3.897694	3.390901	0.152739	1.642869	4.329765	1.665729	0.704756	3.231003	-0.610208	-0.384683	0.605381	1
7048	0.366635	0.330859	1.490104	1.498130	1.449940	1.964019	2.509647	1.949677	1.681522	1.155870	2.892425	1.409828	0.297973	2.752293	2.369188	2.765632	1.396498	1
7054	0.366635	1.015273	1.490104	4.409966	4.024928	1.035928	1.584282	0.384919	2.700711	2.860365	-0.461368	2.433434	0.840350	2.785308	2.369188	2.765632	1.061002	1
7080	0.366635	1.095793	1.622560	2.292267	2.582935	3.501171	4.938730	2.937944	4.852332	2.373366	0.017745	5.504253	0.840350	3.825263	0.560798	0.299013	0.290828	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13949	0.372472	0.950058	1.469260	0.617089	0.467080	2.466582	2.493645	2.172279	0.901836	0.190816	0.826451	1.022234	0.271791	2.606192	2.476057	2.000638	0.989980	1
13972	0.836668	1.031864	1.270939	-0.774400	-0.779072	2.212592	2.493645	2.856531	0.901836	0.419595	4.054722	1.865156	0.628128	1.874797	1.625939	2.290023	0.730476	1
13973	-0.555920	0.909155	1.119838	-0.465180	-0.541710	2.136395	1.168951	-0.121981	2.060979	1.105932	-0.585917	1.503904	-0.203325	1.652199	1.625939	2.290023	0.669302	1
14015	0.140374	0.663737	1.431485	2.008578	2.365979	2.593576	2.391745	0.723273	2.612951	1.563491	0.221151	2.105991	-0.084546	2.606192	1.061770	1.394824	0.049740	1
14022	0.836668	0.868252	1.771463	-0.465180	-0.423028	2.949162	2.238896	3.742035	1.288217	1.334712	2.642354	1.744739	0.509349	2.256395	-0.823946	-0.822892	0.368273	1

204 rows × 18 columns

31 year-trained model's rule:

	rule	type	coef	support	importance
0	PTS > 1.667323112487793	rule	0.0	0.081411	0.0
1	PTS <= 1.667323112487793	rule	0.0	0.918589	0.0

The minimum qualification of being an MVP and MVP candidate from the past 31 year-trained model is to have a standardized score of PTS (Points Scored) greater than roughly 1.67 according to the first rule created above. Below is the review of overall class 0 and 1 of this model.

Class 0 :

```
class_0 = df.iloc[np.where(df.Is_MVP_or_MVP_candidate == "0")]
class_0
```

	Age	G	MP	3P	3PA	2PA	FTA	TRB	AST	STL	BLK	TOV	PF	PTS	W/L%	SRS	TS%	Is_MVP_or_MVP_candidate
0	-0.886218	-0.716531	-1.501674	-0.726261	-0.796777	-1.057710	-0.606015	-0.758695	-0.969130	-1.214828	-0.420182	-1.004570	-1.330463	-1.064343	0.472185	0.509455	0.021818	0
1	0.174886	-1.903423	-1.558538	-0.726261	-0.796777	-1.258329	-0.771329	-0.796254	-1.020713	-1.214828	-0.786541	-1.483782	-1.450498	-1.277519	0.472185	0.509455	-1.687118	0
2	0.174886	0.232983	0.100016	-0.726261	-0.796777	-0.079690	-0.220282	0.405639	-0.659630	-0.011881	0.312536	-0.165949	0.109958	-0.277230	0.472185	0.509455	0.293066	0
3	2.562368	-2.061675	-0.828774	-0.726261	-0.796777	-1.158019	-0.881538	-0.458222	-1.020713	-1.415319	-0.237003	-1.603585	0.710134	-1.228325	0.472185	0.509455	-1.371853	0
4	-0.355666	-0.558279	-0.080055	-0.726261	-0.735300	-0.355542	-0.440701	0.405639	-0.659630	-0.813845	-0.237003	-0.764964	0.590099	-0.392018	0.472185	0.509455	1.134430	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14087	-0.091724	1.113670	1.025399	0.617089	0.763783	0.587060	0.557555	0.723273	0.570653	0.877153	-0.182383	0.660981	0.628128	0.730006	-0.537997	-0.676847	0.302459	0
14088	-0.555920	-0.563355	-1.363893	-0.774400	-0.660391	-1.038473	-0.818088	-0.967234	-0.698884	-0.495522	-0.585917	-0.663611	-1.272335	-1.066681	-0.537997	-0.676847	-0.908281	0
14089	-0.555920	-0.604258	-1.099466	-0.619790	-0.541710	-0.886079	-0.563339	-1.128235	-0.809279	-0.724301	-0.585917	-0.904446	-1.628672	-0.907682	-0.537997	-0.676847	-0.480678	0
14090	-1.484312	-0.358839	-0.806706	-0.155960	-0.185666	-0.759085	-0.919988	-0.685483	-0.698884	-1.181859	-0.384150	-0.904446	-0.559662	-0.685084	-0.537997	-0.676847	0.202020	0
14091	-0.091724	0.541028	0.619314	0.617089	0.467080	0.282272	-0.053842	1.045274	0.073878	0.190816	0.422918	-0.181941	1.103244	0.332509	-0.537997	-0.676847	0.370827	0

13621 rows × 18 columns

Class 1 :

```
class_1 = df.iloc[np.where(df.Is_MVP_or_MVP_candidate == "1")]
class_1
```

	Age	G	MP	3P	3PA	2PA	FTA	TRB	AST	STL	BLK	TOV	PF	PTS	W/L%	SRS	TS%	Is_MVP_or_MVP_candidate
5	1.235989	0.826429	1.436338	2.329278	2.523001	1.324645	1.928798	1.044144	1.455287	2.193521	0.312536	1.271686	0.710134	2.182496	0.472185	0.509455	0.629882	1
6	1.235989	0.668176	1.787004	-0.726261	-0.673822	3.907619	2.865577	2.734306	0.784703	2.193521	5.441561	2.349913	1.670415	3.166386	0.472185	0.509455	0.449002	1
18	0.970713	0.509924	1.351041	1.042735	1.170499	2.001735	2.865577	2.846983	1.094203	1.792539	0.495715	1.032080	1.070239	2.379274	1.394055	0.837225	0.568083	1
83	1.235989	1.063807	1.351041	1.203553	0.924589	0.321548	0.771601	-0.157748	5.324037	3.396468	-0.237003	2.349913	0.590099	1.018225	1.469310	1.667292	1.410709	1
84	0.970713	1.063807	1.644842	-0.565443	-0.489390	3.205451	3.416624	2.659187	0.784703	1.792539	1.045254	1.870701	1.430344	2.986006	1.469310	1.667292	0.747133	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13949	0.372472	0.950058	1.469260	0.617089	0.467080	2.466582	2.493645	2.172279	0.901836	0.190816	0.826451	1.022234	0.271791	2.606192	2.476057	2.000638	0.989980	1
13972	0.836668	1.031864	1.270939	-0.774400	-0.779072	2.212592	2.493645	2.856531	0.901836	0.419595	4.054722	1.865156	0.628128	1.874797	1.625939	2.290023	0.730476	1
13973	-0.555920	0.909155	1.119838	-0.465180	-0.541710	2.136395	1.168951	-0.121981	2.060979	1.105932	-0.585917	1.503904	-0.203325	1.652199	1.625939	2.290023	0.669302	1
14015	0.140374	0.663737	1.431485	2.008578	2.365979	2.593576	2.391745	0.723273	2.612951	1.563491	0.221151	2.105991	-0.084546	2.606192	1.061770	1.394824	0.049740	1
14022	0.836668	0.868252	1.771463	-0.465180	-0.423028	2.949162	2.238896	3.742035	1.288217	1.334712	2.642354	1.744739	0.509349	2.256395	-0.823946	-0.822892	0.368273	1

471 rows × 18 columns

Conclusion

We found the top most important variables to MVPs were points, free throw attempts, 2 point attempts, team wins, assists, simple rating system, and turnovers. The importance of variables did not change significantly between Extra Trees or Random Forest classifying methods.. Some of the variables have consistently increased and decreased with time. The importance of TS% has increased, while rebounds, 2 point attempts, and team wins have decreased. When performing a PCA analysis on MVP, we can see that PC1 is related to player quality, and PC2 is related to playstyle. Our logistic regression model was able to accurately classify Non-Candidates with 99.6% accuracy, MVP Candidates with ~60% accuracy, and MVPs with ~40% accuracy. Our Logistic model decreased in accuracy when it was trained with the past 10 years of data. Our final prediction was able to correctly predict 8 of the top 10 votes, but did not correctly predict the rank of the top 10 votes. Rule based learning (RuleFit) captured mostly the overall of each target class match.

Future Work

We want to find alternative ways to trim down 44 variables to 15 variables. We should have performed the PCA with the top 10 variables, or should have found an alternative way to decide on which variables to use. We believe that a player's winning percentage may be more important than a team's winning percentage, so if we continue research in the future, we may need to find a way to acquire that data. We are interested in models that would allow us to create an algorithm that can predict the MVP one-year ahead of time. We were also interested in using the prediction models built into the Scikit Learn decision tree packages. Due to the severely imbalanced dataset, the models possibly learned heavily on the large portion from trained data such as class 0 (non MVP candidate). The accuracy could have been misleading. The better approach to normally train the data, we could further our project by oversampling the data or undersampling. The evaluation metrics could be changed to AUC, a better measurement to classification task.

Splitting data into training/testing and then splitting training into training and validation dataset so that the model can be verified whether it is overfitting prior to the testing process. If the model is overfitting, the regularization can be applied.

Appendix

player_stats.csv	
Variable	**Explanation**
Rk	Rank
Player	Player name
Pos	Basketball Position of player
Age	Age of player
Tm	Team
G	Games Played
GS	Games Started
MP	Minutes Played
FG	Field Goals Made
FGA	Field Goats Attempted
FG%	Field Goal Percentage
3P	3 Pointers Made
3PA	3 Pointers Attempted
3P%	3 Point Percentage
2P	2 Pointers Made
2PA	2 Pointers Attempted
2P%	2 Point Percentage
eFG%	Efficient Field Goal Percentage
FT	Free Throws Made
FTA	Free Throws Attempted
FT%	Free Throw Percentage
ORB	Offensive Rebounds
DRB	Defensive Rebounds
TRB	Total Rebounds
AST	Assist
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls

mvps.csv	
Variable	**Explanation**
Rank	Rank in MVP voting
Player	Player name
Age	Age of player
Tm	Team
First	Number of First Place Votes
Pts Won	Points won
Pts Max	Max Points that one player can earn
Share	First/Pts Max
G	Games Played
MP	Minutes Played
PTS	Points Scored
TRB	Total Rebounds
AST	Assist
STL	Steals
BLK	Blocks
FG%	Field Goal Percentage
3P%	3 Point Percentage
FT%	Free Throw Percentage
WS	Win Shares
WS/48	Win Shares per 48 minutes
Year	NBA Season data was collected from

PTS	Points Scored		
Year	NBA Season data was collected from		
Pts Won	Points won		
Pts Max	Max Points that one player can earn		
Share	First/Pts Max		
Team			
W	Games Team has won		
L	Games Team has lost		
W/L%	Percentage of games teams has won		
GB	Games behind best team in the conference		
PS/G	Points per game		
PA/G	Opponent’s points per game		
SRS	"Simple Rating System; a team rating that takes into account average point differential and strength of schedule. The rating is denominated in points above/below average, where zero is average."		
MVP	0 for Non-Candidate, 1 for MVP, 2 for MVP Candidate (but did not win)		

Figure A. All variables in the dataset where colored cells are variables we expected to be important before we began research. Green and red cells denote variables we expected to be positively correlated and negatively correlated to probability of winning MVP respectively.

PCA on past 31 years of data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PTS	0.3882014	-0.01341528	0.03216517	0.23793415	0.241616883	-0.023598967	0.16676271	-0.240531428	0.09024732	0.799204022
FTA	0.3551944	-0.18827999	-0.06047116	0.04884832	0.454263985	-0.117436069	0.17461444	0.689892296	0.26576638	-0.187414872
2PA	0.3670439	-0.19336444	-0.17117536	0.05068295	0.233984452	-0.000459556	0.28468700	-0.631164319	0.01471534	-0.511506213
3PA	0.2112668	0.59488682	0.25079048	0.63963971	0.006570546	-0.067611941	-0.25614255	0.008019483	0.01371130	-0.242827747
TS%	0.1575905	-0.20559445	0.93069486	-0.23575388	0.016171978	0.041972125	0.02232921	-0.048810508	-0.04426061	-0.065269530
AST	0.3034796	0.40697839	-0.08426939	-0.52918323	0.004794675	0.339433026	-0.27791393	-0.083643542	0.50560922	0.004652775
TOV	0.3701483	0.06102056	-0.13671783	-0.21890301	0.205630076	0.142261294	-0.32592814	0.108399551	-0.78206855	0.022882579
TRB	0.2723268	-0.54096094	-0.08624450	0.16759547	-0.316342991	-0.190467759	-0.64735417	-0.044513684	0.20067809	0.001249359
STL	0.3224547	0.24771360	-0.03986176	-0.28768364	-0.420336757	-0.710346771	0.24705520	0.044450898	-0.07676541	0.011379861
GS	0.3342131	-0.08640470	-0.04581145	0.18684226	-0.600957146	0.550449867	0.36190704	0.206576913	-0.07364506	-0.024662550

PCA on past 15 years of data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PTS	0.3889570	0.004105882	0.05492097	0.16010679	0.2947981	-0.05712334	0.13728386	-0.2541651487	0.077742254	0.800240867
FTA	0.3562328	-0.147465564	-0.09013761	-0.07634813	0.4424454	-0.18572384	0.25444879	0.6735425555	0.237392431	-0.179966812
2PA	0.3664523	-0.187572491	-0.19912349	-0.02199809	0.2282755	-0.02396098	0.25662320	-0.6521164754	0.005417999	-0.496558748
3PA	0.2183230	0.569615626	0.35770052	0.55494317	0.1876094	-0.07615319	-0.27497144	-0.0001442914	0.018797085	-0.273865466
TS%	0.1489024	-0.323762734	0.88122292	-0.28855332	-0.0429685	0.05481574	0.03099521	-0.0367914220	-0.039347070	-0.066898101
AST	0.3100267	0.384224387	-0.07701123	-0.51156175	-0.1335865	0.35739268	-0.24904034	-0.0524758394	0.527502494	0.004509107
TOV	0.3690549	0.095108021	-0.14109191	-0.28366802	0.1217959	0.15041344	-0.29877700	0.1298576747	-0.781252284	0.020837324
TRB	0.2728791	-0.547163286	-0.13369089	0.24875402	-0.2015197	-0.14666413	-0.66767424	0.0203738233	0.195069019	-0.004589387
STL	0.3191504	0.220237957	-0.01047849	-0.12692920	-0.5845472	-0.66591436	0.20725186	0.0140800311	-0.070345290	0.013417092
GS	0.3302248	-0.095932300	-0.02974111	0.39582240	-0.4617332	0.58099857	0.36599519	0.1868178439	-0.063684415	-0.022861651

PCA on past 10 years of data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PTS	0.3888909	0.004924439	0.067499629	0.103346318	0.31148587	0.08567849	-0.14223612	-0.251470758	0.072532041	-0.800255665
FTA	0.3563807	-0.136337525	-0.101811963	-0.151010587	0.39477246	0.21824202	-0.27921606	0.683564690	0.204993053	0.174687326
2PA	0.3651999	-0.183667983	-0.210974945	-0.070546068	0.21606059	0.02831130	-0.28041786	-0.646459317	-0.003241403	0.489254410
3PA	0.2280617	0.544472188	0.393439186	0.484532954	0.28478497	0.11947948	0.28656832	-0.012401758	0.027704165	0.289051506
TS%	0.1396167	-0.391149226	0.855077829	-0.275786356	-0.09057393	-0.06326557	-0.02642262	-0.027731146	-0.042309143	0.069497123
AST	0.3153321	0.371251212	-0.051782583	-0.480554682	-0.19756528	-0.36822895	0.22607220	-0.043889558	0.549044258	-0.003854224
TOV	0.3675810	0.119078896	-0.134064568	-0.311222425	0.06348181	-0.15926736	0.30096838	0.119580039	-0.775195541	-0.025811466
TRB	0.2750106	-0.541365578	-0.182575152	0.253029581	-0.13209562	0.16011487	0.67216643	0.016392050	0.196798706	0.006532103
STL	0.3157940	0.207010560	0.005602915	-0.008408684	-0.66485794	0.60430992	-0.21046372	0.006073771	-0.075055650	-0.011265456
GS	0.3270687	-0.101335653	-0.024400562	0.508744947	-0.32953854	-0.61168971	-0.32025030	0.184764810	-0.058378550	0.021772501

Figure B. Values for all 10 Principal Components

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.4746 1.1292 0.93864 0.70256 0.67046 0.54016 0.46358 0.36846 0.32514 0.08716
Proportion of Variance 0.6137 0.1278 0.08829 0.04946 0.04505 0.02924 0.02154 0.01361 0.01059 0.00076
Cumulative Proportion  0.6137 0.7415 0.82975 0.87922 0.92426 0.95350 0.97504 0.98864 0.99924 1.00000
```

Summary of 31yr PCA

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8    PC9    PC10
Standard deviation     2.4676 1.1092 0.94168 0.71944 0.68914 0.55676 0.47028 0.37693 0.3190 0.08559
Proportion of Variance 0.6101 0.1233 0.08885 0.05186 0.04758 0.03106 0.02216 0.01424 0.0102 0.00073
Cumulative Proportion  0.6101 0.7333 0.82217 0.87403 0.92162 0.95267 0.97483 0.98907 0.9993 1.00000
```

Summary of 15yr PCA

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.4654 1.0968 0.9426 0.72818 0.69990 0.55708 0.47505 0.38542 0.31636 0.08773
Proportion of Variance 0.6089 0.1205 0.0890 0.05312 0.04907 0.03109 0.02261 0.01488 0.01003 0.00077
Cumulative Proportion  0.6089 0.7294 0.8184 0.87155 0.92062 0.95171 0.97432 0.98920 0.99923 1.00000
```

Summary of 10yr PCA

Figure C. Summary of each P

```
result.sort_values(by='MVP%', ascending=False).head(20)
```

	PTS	FTA	W/L%	SRS	2PA	AST	TOV	STL	TS%	3PA	MVP%	Candidate%	Player
541	27.4	7.3	0.683	5.37	16.2	6.7	3.4	1.2	0.575340	4.5	0.154800	0.953885	Ja Morant
477	29.9	11.4	0.622	3.22	15.0	5.8	3.3	1.1	0.633045	3.6	0.130636	0.983956	Giannis Antetokounmpo
31	30.6	11.8	0.622	2.57	15.9	4.2	3.1	1.1	0.617135	3.7	0.098257	0.971598	Joel Embiid
48	26.8	5.3	0.780	6.94	13.9	4.8	2.4	1.1	0.576791	7.0	0.074797	0.879950	Devin Booker
15	27.1	6.3	0.585	2.16	13.8	7.9	3.8	1.5	0.661880	3.9	0.072744	0.941794	Nikola Jokic
463	28.4	7.5	0.634	3.12	12.8	8.7	4.5	1.2	0.570281	8.8	0.058164	0.867176	Luka Doncic
272	29.9	7.4	0.537	0.82	14.8	6.4	3.5	0.9	0.634658	5.5	0.025885	0.862740	Kevin Durant
196	27.9	7.8	0.561	-0.38	18.3	4.9	2.4	0.9	0.590301	1.9	0.025263	0.828253	DeMar DeRozan
46	14.7	3.1	0.780	6.94	8.3	10.8	2.4	1.9	0.580385	3.1	0.021369	0.772114	Chris Paul
511	28.4	7.3	0.524	1.55	12.3	9.7	4.0	0.9	0.603947	8.0	0.020573	0.829692	Trae Young
439	21.4	8.0	0.646	4.23	12.5	5.5	2.1	1.6	0.593785	2.0	0.011081	0.766823	Jimmy Butler
30	22.0	8.2	0.622	2.57	8.4	10.3	4.4	1.3	0.581764	6.9	0.008206	0.681025	James Harden
72	26.9	6.2	0.622	7.02	12.0	4.4	2.9	1.0	0.576560	8.6	0.005914	0.509869	Jayson Tatum
407	22.8	5.6	0.585	2.38	14.5	5.3	2.7	1.3	0.562574	3.2	0.002689	0.391962	Pascal Siakam
273	27.4	4.4	0.537	0.82	13.0	5.8	2.5	1.4	0.592151	8.2	0.002492	0.388278	Kyrie Irving
516	25.9	4.7	0.598	5.67	10.8	5.3	3.0	1.5	0.573821	9.8	0.002456	0.329947	Donovan Mitchell
431	19.1	6.1	0.646	4.23	12.9	3.4	2.6	1.4	0.608901	0.1	0.002120	0.371857	Bam Adebayo
294	25.5	4.7	0.646	5.52	7.4	6.3	3.2	1.3	0.602324	11.7	0.002058	0.351313	Stephen Curry
326	30.3	6.0	0.402	-3.08	13.8	6.2	3.5	1.3	0.619885	8.0	0.001445	0.354041	LeBron James
47	17.2	2.4	0.780	6.94	11.7	1.4	1.6	0.7	0.658701	0.3	0.001199	0.236291	Deandre Ayton

Figure D. Results sorted by descending MVP probability for 10yr Random Forest

```python
result.sort_values(by='Candidate%', ascending=False).head(20)
```

	PTS	FTA	W/L%	SRS	2PA	AST	TOV	STL	TS%	3PA	MVP%	Candidate%	Player
477	29.9	11.4	0.622	3.22	15.0	5.8	3.3	1.1	0.633045	3.6	0.130636	0.983956	Giannis Antetokounmpo
31	30.6	11.8	0.622	2.57	15.9	4.2	3.1	1.1	0.617135	3.7	0.098257	0.971598	Joel Embiid
541	27.4	7.3	0.683	5.37	16.2	6.7	3.4	1.2	0.575340	4.5	0.154800	0.953885	Ja Morant
15	27.1	6.3	0.585	2.16	13.8	7.9	3.8	1.5	0.661880	3.9	0.072744	0.941794	Nikola Jokic
48	26.8	5.3	0.780	6.94	13.9	4.8	2.4	1.1	0.576791	7.0	0.074797	0.879950	Devin Booker
463	28.4	7.5	0.634	3.12	12.8	8.7	4.5	1.2	0.570281	8.8	0.058164	0.867176	Luka Doncic
272	29.9	7.4	0.537	0.82	14.8	6.4	3.5	0.9	0.634658	5.5	0.025885	0.862740	Kevin Durant
511	28.4	7.3	0.524	1.55	12.3	9.7	4.0	0.9	0.603947	8.0	0.020573	0.829692	Trae Young
196	27.9	7.8	0.561	-0.38	18.3	4.9	2.4	0.9	0.590301	1.9	0.025263	0.828253	DeMar DeRozan
46	14.7	3.1	0.780	6.94	8.3	10.8	2.4	1.9	0.580385	3.1	0.021369	0.772114	Chris Paul
439	21.4	8.0	0.646	4.23	12.5	5.5	2.1	1.6	0.593785	2.0	0.011081	0.766823	Jimmy Butler
30	22.0	8.2	0.622	2.57	8.4	10.3	4.4	1.3	0.581764	6.9	0.008206	0.681025	James Harden
72	26.9	6.2	0.622	7.02	12.0	4.4	2.9	1.0	0.576560	8.6	0.005914	0.509869	Jayson Tatum
407	22.8	5.6	0.585	2.38	14.5	5.3	2.7	1.3	0.562574	3.2	0.002689	0.391962	Pascal Siakam
273	27.4	4.4	0.537	0.82	13.0	5.8	2.5	1.4	0.592151	8.2	0.002492	0.388278	Kyrie Irving
431	19.1	6.1	0.646	4.23	12.9	3.4	2.6	1.4	0.608901	0.1	0.002120	0.371857	Bam Adebayo
326	30.3	6.0	0.402	-3.08	13.8	6.2	3.5	1.3	0.619885	8.0	0.001445	0.354041	LeBron James
294	25.5	4.7	0.646	5.52	7.4	6.3	3.2	1.3	0.602324	11.7	0.002058	0.351313	Stephen Curry
516	25.9	4.7	0.598	5.67	10.8	5.3	3.0	1.5	0.573821	9.8	0.002456	0.329947	Donovan Mitchell
345	24.6	6.3	0.561	2.53	11.5	3.6	3.1	1.0	0.641561	4.9	0.000965	0.308918	Karl-Anthony Towns

Figure E. Figure D. Results sorted by descending MVP-Candidate probability for 10yr Random Forest

```
] result.sort_values(by='MVP%', ascending=False).head(20)
```

	PTS	FTA	W/L%	2PA	SRS	AST	TOV	GS	TS%	STL	MVP%	Candidate%	Player
48	26.8	5.3	0.780	13.9	6.94	4.8	2.4	68	0.576791	1.1	0.212411	0.956715	Devin Booker
541	27.4	7.3	0.683	16.2	5.37	6.7	3.4	57	0.575340	1.2	0.203349	0.972692	Ja Morant
477	29.9	11.4	0.622	15.0	3.22	5.8	3.3	67	0.633045	1.1	0.199123	0.993778	Giannis Antetokounmpo
31	30.6	11.8	0.622	15.9	2.57	4.2	3.1	68	0.617135	1.1	0.157751	0.991717	Joel Embiid
463	28.4	7.5	0.634	12.8	3.12	8.7	4.5	65	0.570281	1.2	0.070662	0.940237	Luka Doncic
15	27.1	6.3	0.585	13.8	2.16	7.9	3.8	74	0.661880	1.5	0.053876	0.954619	Nikola Jokic
196	27.9	7.8	0.561	18.3	-0.38	4.9	2.4	76	0.590301	0.9	0.053068	0.954648	DeMar DeRozan
511	28.4	7.3	0.524	12.3	1.55	9.7	4.0	76	0.603947	0.9	0.034669	0.926479	Trae Young
272	29.9	7.4	0.537	14.8	0.82	6.4	3.5	55	0.634658	0.9	0.021849	0.859892	Kevin Durant
46	14.7	3.1	0.780	8.3	6.94	10.8	2.4	65	0.580385	1.9	0.020405	0.774862	Chris Paul
72	26.9	6.2	0.622	12.0	7.02	4.4	2.9	76	0.576560	1.0	0.017023	0.754667	Jayson Tatum
439	21.4	8.0	0.646	12.5	4.23	5.5	2.1	57	0.593785	1.6	0.009387	0.812390	Jimmy Butler
30	22.0	8.2	0.622	8.4	2.57	10.3	4.4	65	0.581764	1.3	0.007183	0.816808	James Harden
407	22.8	5.6	0.585	14.5	2.38	5.3	2.7	68	0.562574	1.3	0.003305	0.583561	Pascal Siakam
516	25.9	4.7	0.598	10.8	5.67	5.3	3.0	67	0.573821	1.5	0.002890	0.428347	Donovan Mitchell
294	25.5	4.7	0.646	7.4	5.52	6.3	3.2	64	0.602324	1.3	0.002347	0.402658	Stephen Curry
71	23.6	4.8	0.622	11.4	7.02	3.5	2.7	66	0.575273	1.1	0.001273	0.313837	Jaylen Brown
345	24.6	6.3	0.561	11.5	2.53	3.6	3.1	74	0.641561	1.0	0.001047	0.488447	Karl-Anthony Towns
47	17.2	2.4	0.780	11.7	6.94	1.4	1.6	58	0.658701	0.7	0.000968	0.262119	Deandre Ayton
431	19.1	6.1	0.646	12.9	4.23	3.4	2.6	56	0.608901	1.4	0.000812	0.401910	Bam Adebayo

Figure F. Results sorted by descending MVP probability for 15yr Random Forest

```
result.sort_values(by='Candidate%', ascending=False).head(20)
```

	PTS	FTA	W/L%	2PA	SRS	AST	TOV	GS	TS%	STL	MVP%	Candidate%	Player
477	29.9	11.4	0.622	15.0	3.22	5.8	3.3	67	0.633045	1.1	0.199123	0.993778	Giannis Antetokounmpo
31	30.6	11.8	0.622	15.9	2.57	4.2	3.1	68	0.617135	1.1	0.157751	0.991717	Joel Embiid
541	27.4	7.3	0.683	16.2	5.37	6.7	3.4	57	0.575340	1.2	0.203349	0.972692	Ja Morant
48	26.8	5.3	0.780	13.9	6.94	4.8	2.4	68	0.576791	1.1	0.212411	0.956715	Devin Booker
196	27.9	7.8	0.561	18.3	-0.38	4.9	2.4	76	0.590301	0.9	0.053068	0.954648	DeMar DeRozan
15	27.1	6.3	0.585	13.8	2.16	7.9	3.8	74	0.661880	1.5	0.053876	0.954619	Nikola Jokic
463	28.4	7.5	0.634	12.8	3.12	8.7	4.5	65	0.570281	1.2	0.070662	0.940237	Luka Doncic
511	28.4	7.3	0.524	12.3	1.55	9.7	4.0	76	0.603947	0.9	0.034669	0.926479	Trae Young
272	29.9	7.4	0.537	14.8	0.82	6.4	3.5	55	0.634658	0.9	0.021849	0.859892	Kevin Durant
30	22.0	8.2	0.622	8.4	2.57	10.3	4.4	65	0.581764	1.3	0.007183	0.816808	James Harden
439	21.4	8.0	0.646	12.5	4.23	5.5	2.1	57	0.593785	1.6	0.009387	0.812390	Jimmy Butler
46	14.7	3.1	0.780	8.3	6.94	10.8	2.4	65	0.580385	1.9	0.020405	0.774862	Chris Paul
72	26.9	6.2	0.622	12.0	7.02	4.4	2.9	76	0.576560	1.0	0.017023	0.754667	Jayson Tatum
407	22.8	5.6	0.585	14.5	2.38	5.3	2.7	68	0.562574	1.3	0.003305	0.583561	Pascal Siakam
345	24.6	6.3	0.561	11.5	2.53	3.6	3.1	74	0.641561	1.0	0.001047	0.488447	Karl-Anthony Towns
516	25.9	4.7	0.598	10.8	5.67	5.3	3.0	67	0.573821	1.5	0.002890	0.428347	Donovan Mitchell
294	25.5	4.7	0.646	7.4	5.52	6.3	3.2	64	0.602324	1.3	0.002347	0.402658	Stephen Curry
431	19.1	6.1	0.646	12.9	4.23	3.4	2.6	56	0.608901	1.4	0.000812	0.401910	Bam Adebayo
326	30.3	6.0	0.402	13.8	-3.08	6.2	3.5	56	0.619885	1.3	0.000442	0.316799	LeBron James
71	23.6	4.8	0.622	11.4	7.02	3.5	2.7	66	0.575273	1.1	0.001273	0.313837	Jaylen Brown

Figure G. Results sorted by descending MVP-Candidate probability for 15yr Random Forest

```
result.sort_values(by='MVP%', ascending=False).head(20)
```

	PTS	FTA	W/L%	2PA	SRS	TOV	AST	TRB	GS	TS%	MVP%	Candidate%	Player
15	27.1	6.3	0.585	13.8	2.16	3.8	7.9	13.8	74	0.661880	0.301020	0.998366	Nikola Jokic
477	29.9	11.4	0.622	15.0	3.22	3.3	5.8	11.6	67	0.633045	0.219752	0.997792	Giannis Antetokounmpo
48	26.8	5.3	0.780	13.9	6.94	2.4	4.8	5.0	68	0.576791	0.209579	0.957733	Devin Booker
31	30.6	11.8	0.622	15.9	2.57	3.1	4.2	11.7	68	0.617135	0.153177	0.995686	Joel Embiid
541	27.4	7.3	0.683	16.2	5.37	3.4	6.7	5.7	57	0.575340	0.149719	0.961416	Ja Morant
463	28.4	7.5	0.634	12.8	3.12	4.5	8.7	9.1	65	0.570281	0.116547	0.993208	Luka Doncic
46	14.7	3.1	0.780	8.3	6.94	2.4	10.8	4.4	65	0.580385	0.048573	0.877833	Chris Paul
196	27.9	7.8	0.561	18.3	-0.38	2.4	4.9	5.2	76	0.590301	0.026957	0.810964	DeMar DeRozan
511	28.4	7.3	0.524	12.3	1.55	4.0	9.7	3.7	76	0.603947	0.019936	0.926824	Trae Young
72	26.9	6.2	0.622	12.0	7.02	2.9	4.4	8.0	76	0.576560	0.019128	0.861385	Jayson Tatum
272	29.9	7.4	0.537	14.8	0.82	3.5	6.4	7.4	55	0.634658	0.018472	0.944880	Kevin Durant
47	17.2	2.4	0.780	11.7	6.94	1.6	1.4	10.2	58	0.658701	0.012961	0.558406	Deandre Ayton
30	22.0	8.2	0.622	8.4	2.57	4.4	10.3	7.7	65	0.581764	0.011844	0.961471	James Harden
407	22.8	5.6	0.585	14.5	2.38	2.7	5.3	8.5	68	0.562574	0.005725	0.689483	Pascal Siakam
439	21.4	8.0	0.646	12.5	4.23	2.1	5.5	5.9	57	0.593785	0.004139	0.664738	Jimmy Butler
294	25.5	4.7	0.646	7.4	5.52	3.2	6.3	5.2	64	0.602324	0.003717	0.781564	Stephen Curry
345	24.6	6.3	0.561	11.5	2.53	3.1	3.6	9.8	74	0.641561	0.003630	0.734249	Karl-Anthony Towns
431	19.1	6.1	0.646	12.9	4.23	2.6	3.4	10.1	56	0.608901	0.002443	0.522646	Bam Adebayo
516	25.9	4.7	0.598	10.8	5.67	3.0	5.3	4.2	67	0.573821	0.001706	0.468178	Donovan Mitchell
71	23.6	4.8	0.622	11.4	7.02	2.7	3.5	6.1	66	0.575273	0.001039	0.325609	Jaylen Brown

Figure H. Results sorted by descending MVP probability for 15yr Random Forest

```
result.sort_values(by='Candidate%', ascending=False).head(20)
```

	PTS	FTA	W/L%	2PA	SRS	TOV	AST	TRB	GS	TS%	MVP%	Candidate%	Player
15	27.1	6.3	0.585	13.8	2.16	3.8	7.9	13.8	74	0.661880	0.301020	0.998366	Nikola Jokic
477	29.9	11.4	0.622	15.0	3.22	3.3	5.8	11.6	67	0.633045	0.219752	0.997792	Giannis Antetokounmpo
31	30.6	11.8	0.622	15.9	2.57	3.1	4.2	11.7	68	0.617135	0.153177	0.995686	Joel Embiid
463	28.4	7.5	0.634	12.8	3.12	4.5	8.7	9.1	65	0.570281	0.116547	0.993208	Luka Doncic
30	22.0	8.2	0.622	8.4	2.57	4.4	10.3	7.7	65	0.581764	0.011844	0.961471	James Harden
541	27.4	7.3	0.683	16.2	5.37	3.4	6.7	5.7	57	0.575340	0.149719	0.961416	Ja Morant
48	26.8	5.3	0.780	13.9	6.94	2.4	4.8	5.0	68	0.576791	0.209579	0.957733	Devin Booker
272	29.9	7.4	0.537	14.8	0.82	3.5	6.4	7.4	55	0.634658	0.018472	0.944880	Kevin Durant
511	28.4	7.3	0.524	12.3	1.55	4.0	9.7	3.7	76	0.603947	0.019936	0.926824	Trae Young
46	14.7	3.1	0.780	8.3	6.94	2.4	10.8	4.4	65	0.580385	0.048573	0.877833	Chris Paul
72	26.9	6.2	0.622	12.0	7.02	2.9	4.4	8.0	76	0.576560	0.019128	0.861385	Jayson Tatum
196	27.9	7.8	0.561	18.3	-0.38	2.4	4.9	5.2	76	0.590301	0.026957	0.810964	DeMar DeRozan
294	25.5	4.7	0.646	7.4	5.52	3.2	6.3	5.2	64	0.602324	0.003717	0.781564	Stephen Curry
326	30.3	6.0	0.402	13.8	-3.08	3.5	6.2	8.2	56	0.619885	0.000842	0.770447	LeBron James
345	24.6	6.3	0.561	11.5	2.53	3.1	3.6	9.8	74	0.641561	0.003630	0.734249	Karl-Anthony Towns
407	22.8	5.6	0.585	14.5	2.38	2.7	5.3	8.5	68	0.562574	0.005725	0.689483	Pascal Siakam
439	21.4	8.0	0.646	12.5	4.23	2.1	5.5	5.9	57	0.593785	0.004139	0.664738	Jimmy Butler
47	17.2	2.4	0.780	11.7	6.94	1.6	1.4	10.2	58	0.658701	0.012961	0.558406	Deandre Ayton
431	19.1	6.1	0.646	12.9	4.23	2.6	3.4	10.1	56	0.608901	0.002443	0.522646	Bam Adebayo
530	15.6	6.7	0.598	7.6	5.67	1.8	1.1	14.7	66	0.732532	0.000591	0.482854	Rudy Gobert

Figure G. Results sorted by descending MVP-Candidate probability for 15yr Random Forest

Reference

AGRON Stats. “Biplot using base graphic functions in R.” Agron Info Tech. 26 June 2018, http://agroninfotech.blogspot.com/2020/06/biplot-for-pcs-using-base-graphic.html.Written Accessed 1 May 2022.

Fuchs, Michael. “Multinomial Logistic Regression - Michael Fuchs Python.” *Michael-Fuchs-Python.netlify.app*, 15 Nov. 2019, michael-fuchs-python.netlify.app/2019/11/15/multinomial-logistic-regression/.

Galarnyk, Michael. “Logistic Regression Using Python (Scikit-Learn).” *Medium*, 29 Apr. 2020, towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a.

Raheel Shaikh. “Feature Selection Techniques in Machine Learning with Python.” *Medium*, Towards Data Science, 28 Oct. 2018, towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e.

Vinco, Vivo. “1991-2021 NBA Stats.” *Kaggle.* 15 Apr. 2022, www.kaggle.com/datasets/vivovinco/19912021-nba-stats. Accessed 1 May 2022.

Yalçın, Orhan G. “Interpretable Machine Learning in 10 Minutes with RuleFit and Scikit Learn.” *Medium*, 24 June 2021, towardsdatascience.com/interpretable-machine-learning-in-10-minutes-with-rulefit-and-scikit-learn-da9ebb925795. Accessed 3 May 2022.

“RuleFit: A Modeling Method for Automatically Extracting Interactions.” *HACARUS INC.*, 5 Jan. 2022, https://hacarus.com/ai-lab/20211208-rulefit/ Accessed 1 May 2022.

Silva, João V.R.d., and Paulo C. Rodrigues. 2022. "All-NBA Teams’ Selection Based on Unsupervised Learning" *Stats* 5, no. 1: 154-171. https://doi.org/10.3390/stats5010011