Decoding the MVP: An Analysis of Player Statistics, MVP Votes, and Team Standings in NBA Regular Season MVP Selection

Aniket Khetan, Aashay Khot, Sahil Dhamelia, Yash Bhatia
Student, Student, Student, Student

Abstract

This research paper presents a comprehensive analysis of the factors influencing the selection of the NBA regular season MVP. By utilizing player statistics, MVP votes, and team standings data from 1991 to 2023, a regression model is developed to predict the MVP winner. The study compares the performance of three different models, namely Ridge Regression, Random Forest, and Bayesian Regression, and evaluates their effectiveness in predicting the winning share of the top 5 ranked players in the past 8 years. The findings reveal that the Ridge Regression model outperforms the other models, emphasizing the significance of defensive contributions in MVP selection. However, the model faces challenges in evaluating players who impact the game beyond traditional statistics or assists. This research contributes to the understanding of the MVP selection process and highlights the importance of considering both offensive and defensive contributions in evaluating player performance.

1.Introduction

The prediction of the Most Valuable Player (MVP) in the National Basketball Association (NBA) has long been a topic of interest for basketball enthusiasts and analysts alike. The NBA MVP award is one that recognizes the player who made the most impact on their team's success during the regular season. Predicting the MVP winner, as well as studying the statistics of MVP winners in previous seasons provides valuable insights into factors that contribute to a player's, as well as a team's success.

In this paper, we aim to contribute to the existing body of research by developing a regression model for predicting the NBA regular season MVP. To achieve this, we used a comprehensive dataset from Basketball's most trusted source of statistics, basketball-reference.com using web scraping algorithms. The dataset includes player statistics, MVP votes and team standings for the seasons spanning from 1991-2023. By exploring this data, we aim to identify the factors that have historically influenced MVP selection

2.Methodology

2.1 Data collection

This section will now begin to detail the development of a regression model for predicting the regular season MVP for a given NBA season. The first step in developing a model involved deciding the necessary data required for being able to predict if a player will win MVP that season or not. Given that there are several factors that affect this decision, the ones that would be most important to judge a player's performance would be the regular season statistics for each player in an NBA season. Next it was important to also consider the MVP votes each player that got nominated and the winner of the award received. This would be key in defining a target variable for our model to understand how the statistics and performance of a player translate to the votes each player gets and how the winner is decided. Lastly the standings and performance of each team would also be recorded as the trend that has been set is that the title is generally awarded to a player that performs exceptionally well and is also a member of one of the top performing teams that season. The years that were considered for the data were those between 1991 and 2023 for no other reason but the fact that any more data or less would be either too much or too little respectively. All the data consisting of the player statistics, MVP votes and team standings were gathered from 'basketball-reference.com'. To save time, simple web scraping algorithms were applied and implemented with the help of libraries such as 'Beautiful Soup' and 'Selenium'.

Once all data was successfully gathered it was cleaned and processed and finally merged into one

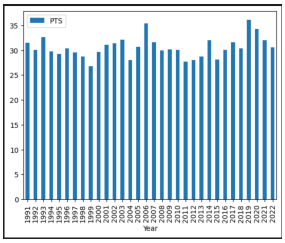
table. Individual columns were chosen from certain tables based on their relevancy, such as only the information about voting points and the win share for the votes were merged from the MVP votes table. Several missing values existed in the Player statistics data through the years in cases when players failed to record a successful shot or an attempt in a season.
For the sake of accuracy in model development and training, these missing values were converted to 0 as dropping the respective players could mislead the predictions in the model.
The final data after preprocessing can be observed below:

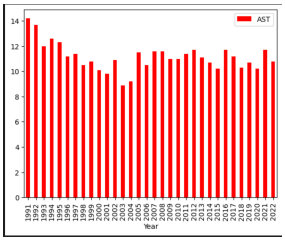
	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	...	Team	W	L	W/L%	GB	PS/G	PA/G	SRS	Npos	Ntm
15189	Hollis Thompson	SG	23.0	PHI	71.0	23.0	25.0	3.2	7.6	0.413	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	13	28
15190	Isaiah Canaan	PG	23.0	PHI	47.0	21.0	20.0	3.1	8.0	0.386	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	6	28
15191	Ish Smith	PG	28.0	PHI	55.0	14.0	15.1	2.6	6.7	0.390	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	6	28
15192	JaKarr Sampson	SF	21.0	PHI	74.0	32.0	15.3	2.0	4.7	0.422	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	9	28
15193	JaVale McGee	C	27.0	PHI	23.0	0.0	11.1	1.8	3.4	0.532	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	1	28
15194	Jason Richardson	SG	34.0	PHI	19.0	15.0	21.9	3.3	9.4	0.348	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	13	28
15195	Jerami Grant	SF	20.0	PHI	65.0	11.0	21.2	1.9	5.4	0.352	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	9	28
15196	Larry Drew II	PG	24.0	PHI	12.0	1.0	18.3	1.7	4.8	0.345	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	6	28
15197	Luc Mbah a Moute	PF	28.0	PHI	67.0	61.0	28.6	3.7	9.5	0.395	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	3	28
15198	Malcolm Lee	SG	24.0	PHI	1.0	0.0	2.0	0.0	1.0	0.000	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	13	28
15199	Malcolm Thomas	PF	26.0	PHI	17.0	0.0	11.4	1.1	2.4	0.450	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	3	28
15200	Nerlens Noel	C	20.0	PHI	75.0	71.0	30.8	4.0	8.7	0.462	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	1	28
15201	Robert Covington	SF	24.0	PHI	70.0	49.0	27.9	4.3	10.8	0.396	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	9	28
15202	Thomas Robinson	PF	23.0	PHI	54.0	4.0	14.8	2.3	4.8	0.485	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	3	28
15203	Tony Wroten	PG	21.0	PHI	30.0	15.0	29.8	5.8	14.5	0.403	...	Philadelphia 76ers	18.0	64.0	0.220	31.0	92.0	101.0	-9.04	6	28
15204	Elliot Perry	PG	30.0	NJN	60.0	5.0	13.4	2.1	4.9	0.435	...	New Jersey Nets	31.0	51.0	0.378	21.0	98.0	99.0	-1.18	6	21
15205	Evan Eschmeyer	C	24.0	NJN	31.0	5.0	12.0	1.2	2.3	0.528	...	New Jersey Nets	31.0	51.0	0.378	21.0	98.0	99.0	-1.18	1	21

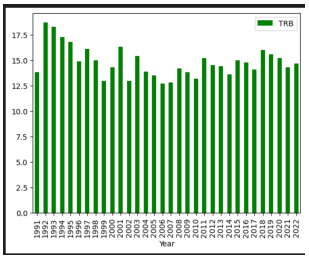
Table 2.1.1

2.2 Data Exploration

Due to the large size of the training data that had been collected, It is necessary that we understood it before using it to train a predictive model. Since the data gathered is across 30 years and like any sport, team tactics, valuable qualities in a player and many other factors could have changed, too much variance can lead to unpredictability and inaccuracy. To better understand what exactly we had collected we will visualize the data to answer important questions that are correlated to the MVP winner of a given NBA season.
The top point scorer, assister or rebounder of a NBA season is more or less likely to be nominated for the title if not win it, and to make sure that the data is reliable in this essence we will visualize and study how much the benchmark to be the top point scorer in a season has changed.
The figure below is a bar chart showing the points per game scored by the leading scored of each season:







Only players who played 70+ games were considered for this since a player with a lesser total of games will naturally have an abnormally higher ‘per game’ metric. Here we observe that the lead scorers of all three criteria’s each year don’t differ much which is good for when we train our model.

2.3 Output Variable

While the features or predictors will vary model to model the target variable for all 3 models will remain the same which is predicting the ‘Share’ variable which represents the winning share of the MVP votes for each season.

3.Implementation

3.1 Ridge Regression Model

The first model we consider uses Ridge Regression to predict our winning share. Ridge regression is a

model tuning method that is used to analyse any data that suffers from multicollinearity. Since we will have several features which will be related to each other, choosing ridge regression allows us to tackle this problem while avoiding overfitting our model by setting an alpha parameter.
The formula for Ridge regression is as follows:

Min(||Y – X(theta)||^2 + λ||theta||^2)
Equation 3.1.1

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced. Y will be our target variable "Share" and X will be our set of predictors.

The predictors for this model are the following features:

'Age', 'G', 'GS', 'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%', 'eFG%', 'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV', 'PF', 'PTS', 'Year', 'W', 'L', 'W/L%', 'GB', 'PS/G', 'PA/G', 'SRS'

3.2 Random Forest Model

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

It works on the following foundation:

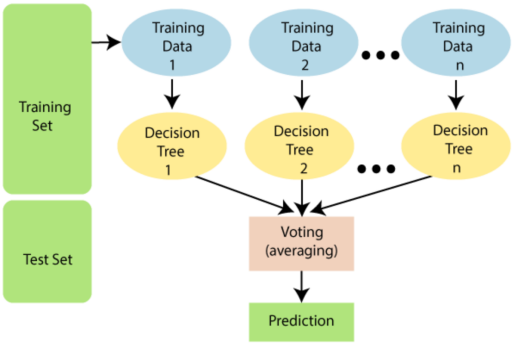


Figure 3.2.1

Since the foundation involves classification we were able to include the players positions and the teams they belong to after label encoding them to the list of predictors

The predictors for this model are the following features:

'Age', 'G', 'GS', 'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%', 'eFG%', 'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV', 'PF', 'PTS', 'Year', 'W', 'L', 'W/L%', 'GB', 'PS/G', 'PA/G', 'SRS', 'Ntm', 'Npos'

3.3 Bayesian Regression Model

Bayesian regression is a type of linear regression that uses Bayesian statistics to estimate the unknown parameters of a model. It uses Bayes' theorem to estimate the likelihood of a set of parameters given observed data. The goal of Bayesian regression is to find the best estimate of the parameters of a linear model that describes the relationship between the independent and the dependent variables.

The formula it uses to calculate the probability of each hypothesis is

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Equation 3.3.1

The predictors for this model are the following features:

'Age', 'G', 'GS', 'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%', 'eFG%', 'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV', 'PF', 'PTS', 'Year', 'W', 'L', 'W/L%', 'GB', 'PS/G', 'PA/G', 'SRS'

4. Results

To calculate the performance of each model, we will compare the predicted ranks after the MVP votes to the actual ranks each players were placed in. Along with this the mean squared error will also be found for each Model. To best evaluate each model, a back testing algorithm will be employed to calculate the average performance of each model when tested for the last 8 years.

4.1 Backtesting

Backtesting will allow us to run our models on a set of years and return an average performance across the inputted years.

```
aps = []
all_predictions = []
years = list(range(1991,2024))
def backtesting(stats, model, years, predictors) :
    for year in years:

        train = stats[stats['Year'] < year]
        test = stats[stats['Year'] == year]
        model.fit(train[predictors],train['Share'])
        predictions = model.predict(test[predictors])
        predictions = pd.DataFrame(predictions, columns=['predictions'], index=test.index)
        predictions_combined = pd.concat([test[["Player", "Share",'Year']], predictions],
        predictions_combined = add_ranks(predictions_combined)
        all_predictions.append(predictions_combined)
        aps.append(find_ap(predictions_combined))
    return sum(aps)/len(aps), aps ,  pd.concat(all_predictions),all_predictions
```

Figure 4.1.1

Here we input the dataset, the model, the split consisting of the training data and the predictors

Then the model will be used on all remaining years and the mean AP, and rankings of each tested year will be returned .

4.2 AP metric

Since the target variable will be a predicted winning share value, it will be easy to compare each players actual ranking to their predicted ranking. Besides just checking if the MVP was predicted as the MVP. Each model can be best evaluated by checking how many of the top 5 ranking players did it predict to be in the top 5 of a seasons MVP list. We will calculate this using the following algorithm.

```
def find_ap(combination):
    actual = combination.sort_values("Share", ascendi
    predicted = combination.sort_values("predictions"
    ps = []
    found = 0
    seen = 1
    for index,row in predicted.iterrows():
        if row["Player"] in actual["Player"].values:
            found += 1
            ps.append(found / seen)
        seen += 1

    return sum(ps) / len(ps)
```

Figure 4.2.1

Here for every player that was placed in the actual top 5 is found to be in the predicted top 5 list, the found variable is incremented by 1. If a player in the top 5 is predicted outside this boundary, the seen variable is incremented by the number of places the model got wrong.

Finally, these values are totaled and averaged for each player, and we are returned the AP metric for a model.

	Model Name	Mean AP
0	Bayesian Regression	0.803467
1	Ridge Regression	0.804044
2	RandomForestRegressor	0.785648

Table 4.2.1

This shows us that the Ridge Regression model performed the best when it comes to predicting the winning share of the top 5 ranked players in the past 8 years.

4.3 Mean squared Errors

Mean Squared Error (MSE) measures the amount of error in a statistical model. Evaluate the mean squared difference between observed and predicted values. If the model has no errors, the MSE is zero. Its value increases as the model error increases.

The following is the MSE that was observed when the model was tested on the last 8 years

	Model Name	Mean Squared Error
0	Bayesian Regression	0.002633
1	Ridge Regression	0.002665
2	RandomForestRegressor	0.002264

Table 4.3.1

The MSE is almost indistinguishable, Ridge Regression again slightly outperformed the other 2 models.

4.4 Interpretations and Findings

After observing the results of all three models, another back test was performed on the Ridge Regression Model, this time only training it on years from 1991-2001 and testing it on the past 23 years and the observations are discussed below.

The Model favours those players with higher contributions to Defence and defensive stats over players with majority contributions in offence. This is confirmed by the model predicting players like Giannis Antetokounmpo to win MVP every year between 2019-2023 and Shaquille O'Neal to win in the years 2001-2005 (Shaquille O'Neal didn't win in any of those years) and place second in 2006 where in reality he ranked 380th that year. Majority of the Correctly predicted MVP winners are players also known for their defensive prowess with notable exception being Stephen Curry

The model struggles in evaluating the performances of those players that impact the game through more than just statistics or through assists notably those who were most successful in the 'Point Guard' Role . This can be seen by looking at the highest differences between actual and predicted ranks when looking only at top 5 players of each year after 2001

	Player	Share	Year
1441	Jason Kidd	0.712	2002
5674	Steve Nash	0.839	2005
9277	Peja Stojaković	0.228	2004
13800	Joakim Noah	0.258	2014
4010	Chauncey Billups	0.344	2006

Figure 4.4.1

All of these players were more subtle in their performances when it came to purely just racking up statistics.

When looking at the weight that the model gives to each feature, EFG% ranked highest, this helps its case as it allows for the model to recognise in the increase of the importance of the 3 point show in the last 10 years due to the rise of players such as Stephen Curry, James Harden, Klay Thompson etc. The model successfully predicts Curry to win the MVP in 2016 when he wins the award unanimously however, fails to give him credit for his other MVP victories backing the conclusion in the model favouring Defensive players over point guards.

13	1.042605e-01	eFG%
29	5.884854e-02	W/L%
18	3.358452e-02	DRB
17	2.055335e-02	ORB
10	1.729904e-02	2P
21	1.204751e-02	STL
22	1.058104e-02	BLK
15	1.031958e-02	FTA
12	8.545880e-03	2P%
20	6.923249e-03	AST

Figure 4.4.2 Weightage given to each feature (Top 10)

Conclusion

In conclusion, our findings revealed that the model favoured players with higher contributions to defense and defensive statistics over those with majority contributions in offense. This was evident in the model's prediction of players like Giannis Antetokounmpo to win MVP consistently between 2019 and 2023, as well as Shaquille O'Neal being predicted to win in the years 2001 to 2005, despite not winning in any of those years. Notably, the model struggled to evaluate the performances of players who impacted the game through more than just statistics or assists, particularly those in the "Point Guard" role.

Since the model is based on statistics, it is unable to account the intangible qualities a player possesses. For example, in the 2011 season, where Derrick Rose won the regular season MVP, Lebron James outperformed him in almost every single major stat category. But since He had won the award the previous two years and it is based on voting, James did not win the award due to what the media claims is ‘voter fatigue’

Overall, this research contributes to the existing body of knowledge by providing insights into the factors that contribute to a player's MVP selection. It highlights the importance of considering both offensive and defensive contributions, as well as the impact of team performance. Future research could further refine the model by incorporating additional variables and exploring the evolving dynamics of the NBA