NBA MVP: Predictor for top MVP Candidates

Introduction

1. Problem Statement

The NBA MVP award is a hotly debated award that members of the media vote on every year. Media members will vote for their top 5 candidates and the award goes to the player who receives the highest share of the total votes. I want to create an algorithm that takes all historical statistical data and predicts who the top 5 vote getters would be for any given season.

2. Background

The most common predictors used to determine the MVP winner are per game statistics and team performance. For example, a player may have excellent individual statistics but if their team does not win enough games, they are unlikely to win the MVP. The opposite can be true as well. In recent years, the media members who vote for MVP have also taken advanced statistics and shooting percentages into account and we will be testing those variables as well.

3. Goal

The goal of this project is to understand which variables determine the MVP award and also create an algorithm that can predict the MVP share for any given season.

4. Datasets

The data that I used was sourced from Kaggle which was originally pulled from Basketball Reference. There were multiple separated datasets that were included however I narrowed it down to the few that we wanted to explore. I first used the Player Per Game data to pull in all the stats for every player since 1947. Then the next piece was to isolate the players who had received any MVP votes and bring those columns into the data which was done so through the Player Award Shares dataset. I also brought in the teams win/loss statistics from the Team Summaries dataset. Lastly, the Advanced dataset was used to pull in the remaining advanced statistics that I wanted to test.

The final list of predictors that we used were age, games played, minutes per game, points per game, field goals per game, attempts per game, field goal percentage, three pointers per game, three point attempts per game, three point percentage, effective field goal percentage, free throws per game, rebounding per game, assists per game, steals per game, blocks per game, fouls per game, team wins and losses, usage percentage, offensive win shares, defensive win shares, total win shares, box plus minus, and value over replacement.

Data Cleaning and Wrangling

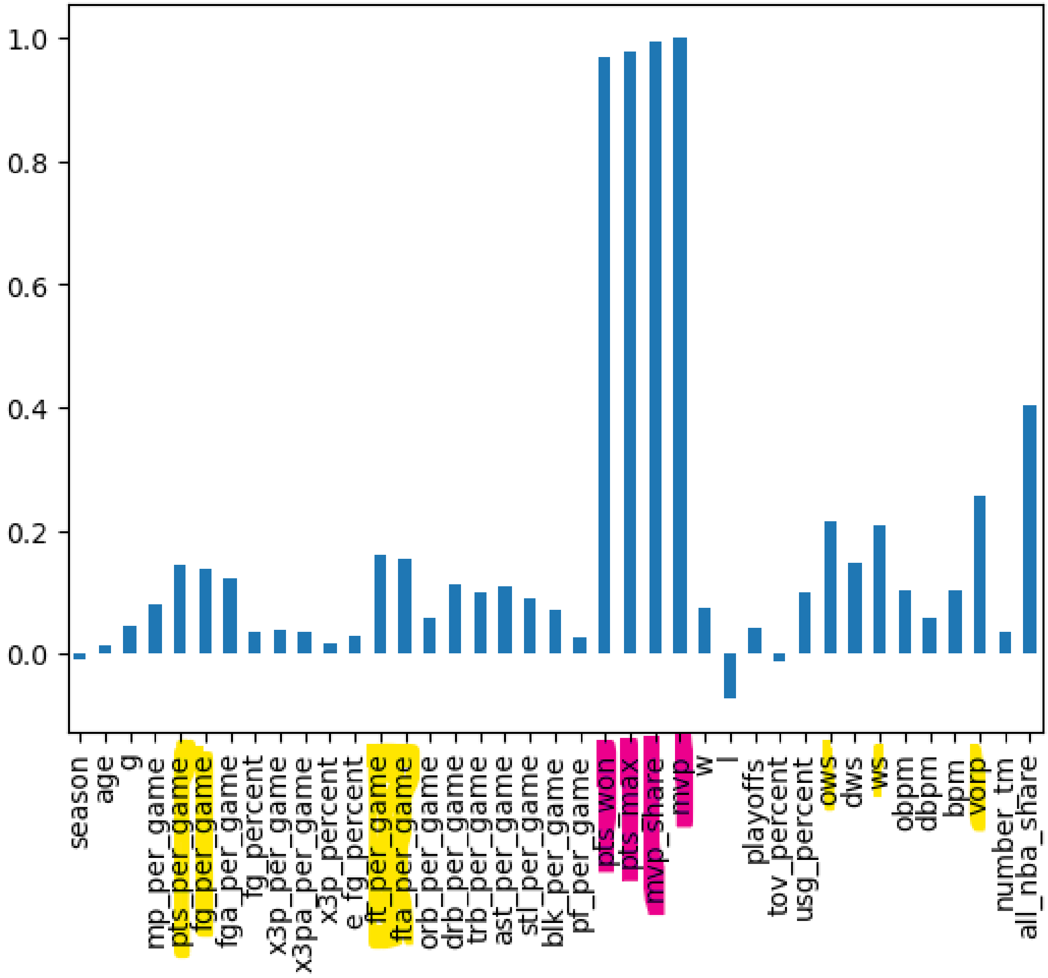
1. Cleaning

I had to create one dataset to preprocess the data to make it compatible for machine learning algorithms. In this situation, I used the per game statistics as the base. I chose this data to build off of because it had data for every player in every season going back to 1947. From there, I removed any nulls, removed duplicate columns for players who had been traded in the same season, and dropped any columns that were not relevant for the model. I then used the awards dataset to filter for any players who had received MVP votes and merged this with the per game statistics. I merged the team win/loss columns to the working dataset so that each row would indicate how that specific player's team performed that season. I used the same process to combine the advanced statistics with the working dataset.

Initially, I wanted to use data going back to the inception of the NBA back to 1947. However, there were several statistics that were not tracked before 1980 such as specific rebounding stats, three pointers, and other defensive metrics. I decided to filter the dataset so that everything before 1980 would be excluded, that way I could minimize errors while modelling.

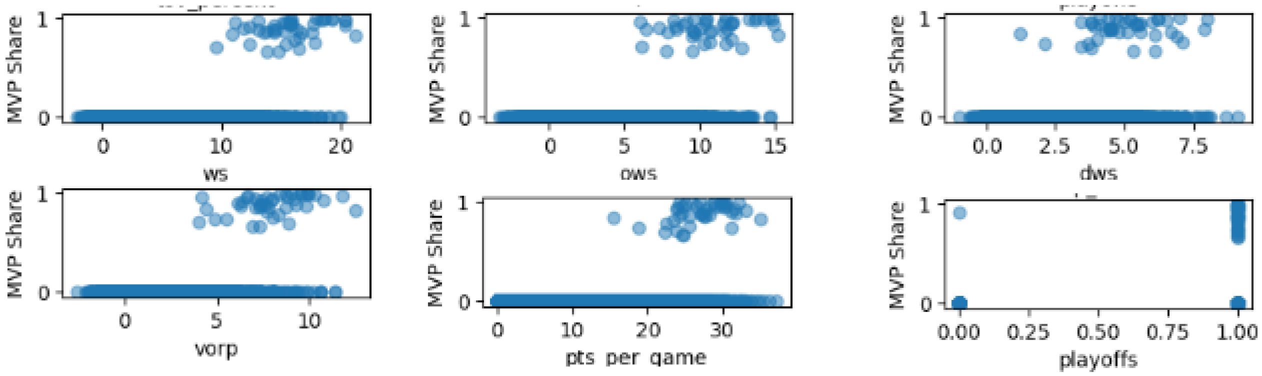
2. Data Analysis and Modelling Results

I first wanted to see the correlation between the different features and MVP winners.



The chart above shows the correlation and we can see that points per game, fg per game, ft per game, offensive win shares, total win shares, and value over replacement correlated most highly with a player winning MVP. The four columns highlighted in red can be excluded as they are the variable that we are testing against.

I also wanted to see the correlation with the selected features and the MVP Share. MVP share is what we will be predicting during modeling.



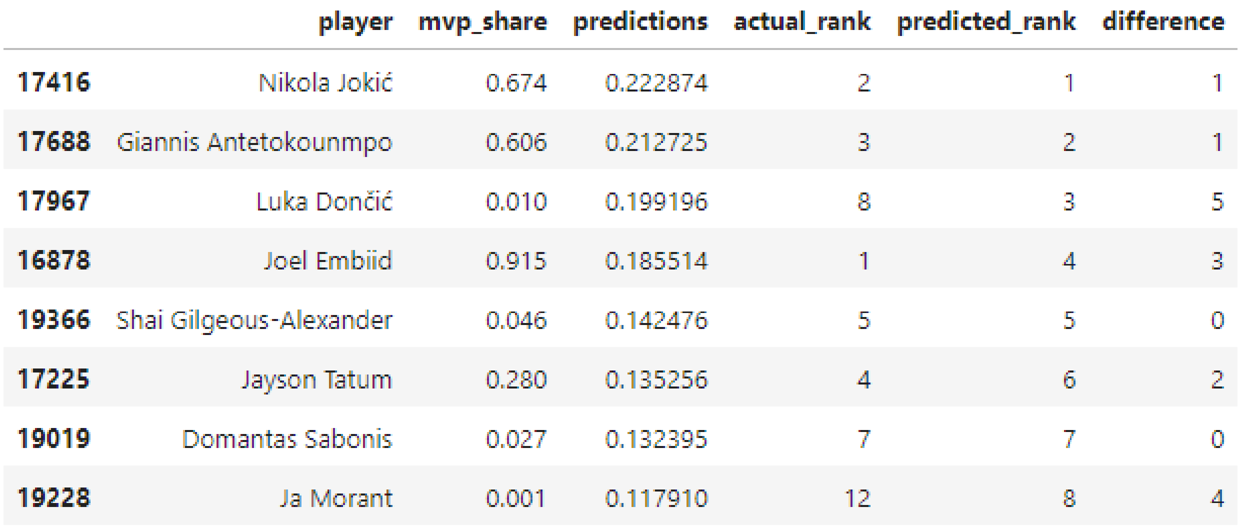
The scatter plots above show the correlation between win shares, value over replacement and points per game with the MVP share. Most players who have a high MVP share also have a generally high score in the above-mentioned stats. Additionally, we can also see that almost every player who has a high MVP share was on a team that made the playoffs.

3. Modelling

To create a machine learning model, I used the combined dataset that was used for exploratory analysis in the previous step. I first identified the predictors that I wanted to use for the model.

To split the data into training and testing, I wanted to use statistics from previous seasons to predict the MVP results for the most recent season. The testing data would be the combined dataset for the 2023 season and the training dataset would be all subsequent seasons. Because all the data was already numerical, I did not have to create any dummy variables.

I decided to use a ridge regression model as it would avoid overfitting the data and also because I had a large number of parameters. The ridge model predicted the MVP share for each player and I compared that to the actual MVP share those players received. Using this, I created an actual rank and predicted rank column to compare the results.



	player	mvp_share	predictions	actual_rank	predicted_rank	difference
17416	Nikola Jokić	0.674	0.222874	2	1	1
17688	Giannis Antetokounmpo	0.606	0.212725	3	2	1
17967	Luka Dončić	0.010	0.199196	8	3	5
16878	Joel Embiid	0.915	0.185514	1	4	3
19366	Shai Gilgeous-Alexander	0.046	0.142476	5	5	0
17225	Jayson Tatum	0.280	0.135256	4	6	2
19019	Domantas Sabonis	0.027	0.132395	7	7	0
19228	Ja Morant	0.001	0.117910	12	8	4

I created the following error metrics to test the accuracy of the data. I limited this to the top five vote getters in every season as media members only vote for the top five.

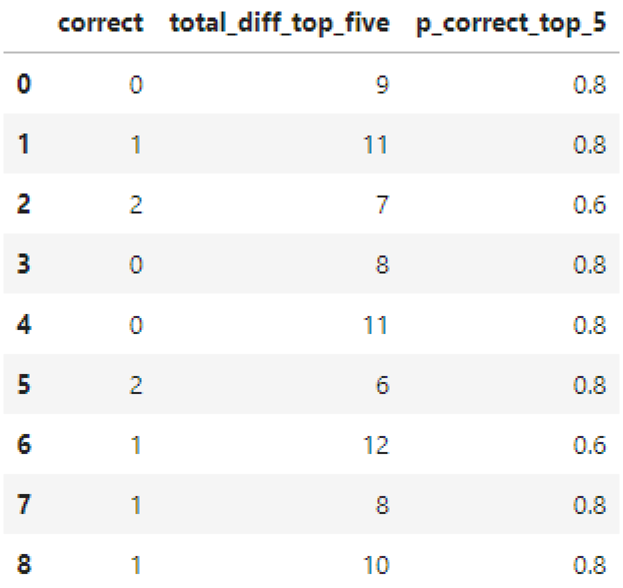
- Check if the model got the exact rank correctly
- The total sum of the difference between the predicted rank and actual rank
- How many of actual top five did the model have in its top five

I ran the model for the past 9 seasons and these were the results on average.

Exact Rank: **0.88/5 (17.6%)**

Total Difference: **9.1**

Top 5 Correct: **75.5%**



	correct	total_diff_top_five	p_correct_top_5
0	0	9	0.8
1	1	11	0.8
2	2	7	0.6
3	0	8	0.8
4	0	11	0.8
5	2	6	0.8
6	1	12	0.6
7	1	8	0.8
8	1	10	0.8

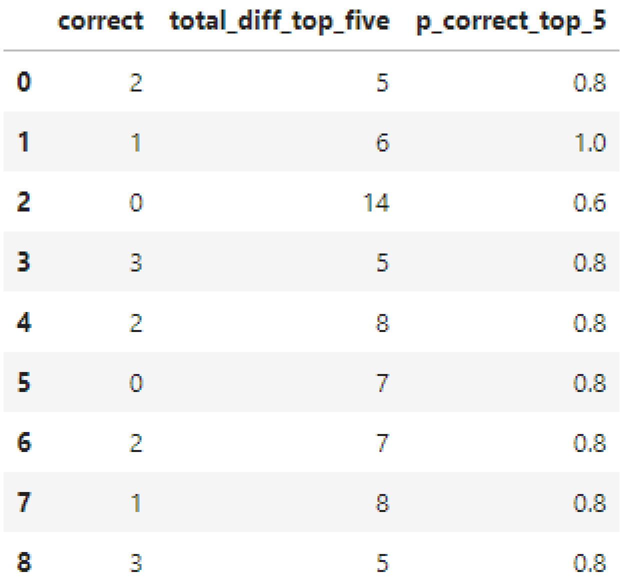
I ran a Random Forest Regressor to compare against the Ridge model to see if it would be more accurate.

There are the results for the Random Forest model.

Exact Rank: **1.56/5 (31.2%)**

Total Difference: **7.2**

Top 5 Correct: **80%**



	correct	total_diff_top_five	p_correct_top_5
0	2	5	0.8
1	1	6	1.0
2	0	14	0.6
3	3	5	0.8
4	2	8	0.8
5	0	7	0.8
6	2	7	0.8
7	1	8	0.8
8	3	5	0.8

The random forest model is significantly better at predicting the exact MVP rank of the player but neither is reliable. The random forest model also predicts the top 5 players 5% more accurately than the Ridge model.

Findings and Next Steps

1. Findings

The model was generally accurate to determine who should be in the top 5 for each season. The MVP award is largely influenced by stats however several voters take the narrative and storylines of the NBA season into account as well and that cannot be quantified in a machine learning model. I think media members and fans can use this model to understand who is having the best season in terms of statistical impact. It can be a tool that can be used to help influence who to vote for or who to maybe place a bet on for fans.

2. Next Steps

I think this model can be used to test the results of other NBA awards such as defensive player of the year, rookie of the year, and All NBA teams. The parameters can be adjusted as well to maybe fit the data better. A deeper dive can also be done to look into which players impact winning the most.