Kamaal Bartlett
NBA MVP Prediction Project
April, 9th 2023

Problem Statement

The National Basketball Association (NBA) Most Valuable Player (MVP) award is the most prestigious individual award in the National Basketball Association. This award is given to the player that makes the greatest individual contributions to their respective team. There are various statistical categories that are leveraged when league officials vote for the MVP of the league at the end of a season. There are 82 games in the regular season and the data used to select the MVP 82 games in the regular season and the data used to select the MVP can only be pulled from those 82 games. Over the years, the way the game of basketball is played has changed. There was an era when players dominated the area underneath the rim (the paint), but over the last few years, the league has shifted to a more perimeter-focused game. Players are taking more three-pointers per game now than in the history of the entire league. In fact, some years ago the three-point shot didn't exist in the NBA.

This is part of what makes training a model to consistently and accurately predict the winner of the MVP award difficult. How can we properly leverage the data available to us from both the past and present, to develop a model that is able to accurately predict the MVP award winner of the 1982 season, while also being able to predict the winner of the 2021 NBA season? The application of this model could be retrained and expanded year after year to predict future MVPs.

Method

The dataset contains data on every NBA season from 1982 – 2022, over the years the NBA has added new statistical categories and some fundamental changes to the game. The addition of the three-point line is one of the biggest changes in the history of the league. Due to this major change, earlier seasons in this dataset do not have values in this category, which may affect the models' ability to accurately predict the MVP of each season. There are a few advanced metrics that utilize the three-pointer in their measurements, making this statistic an important factor. However, due to the nature of the NBA voting process and its heavy dependence upon those advanced statistics, filling in missing values with the standard deviation or mean method would skew the data and cause the model to return even more inaccurate results. Therefore, for this project, the approach taken was to simply fill in all missing values with 0s and remove data only when necessary.

Data Wrangling

Data Wrangling Notebook

The dataset pulled in from Kaggle needed a few changes made to it. First, the column names were updated to be more readable and also more consistent, then all NaN values were filled with 0s, as mentioned under Method this choice was to preserve the integrity of the data in making predictions. Next, the data was used to find the actual MVP winners for each season and create a new dataset that would be used to contain the MVP statistics for later reference. The intention of doing this would be to create a benchmark, as changes were made to the broad dataset, those changes could be compared to or applied to the data frame

containing the MVP data. By taking this approach, it was easier to visualize results as the known award winners for each season were already separated into different datasets. Through data wrangling, once the heatmap was used (Fig 1), the results returned from it could be visualized in both the general data frame and the MVP data frame. There was a strong correlation between award share and value over replacement player (VORP) for example. Having a separate data frame-ready, it helped to look at that data and see what was the award share of past winners when having a higher VORP.

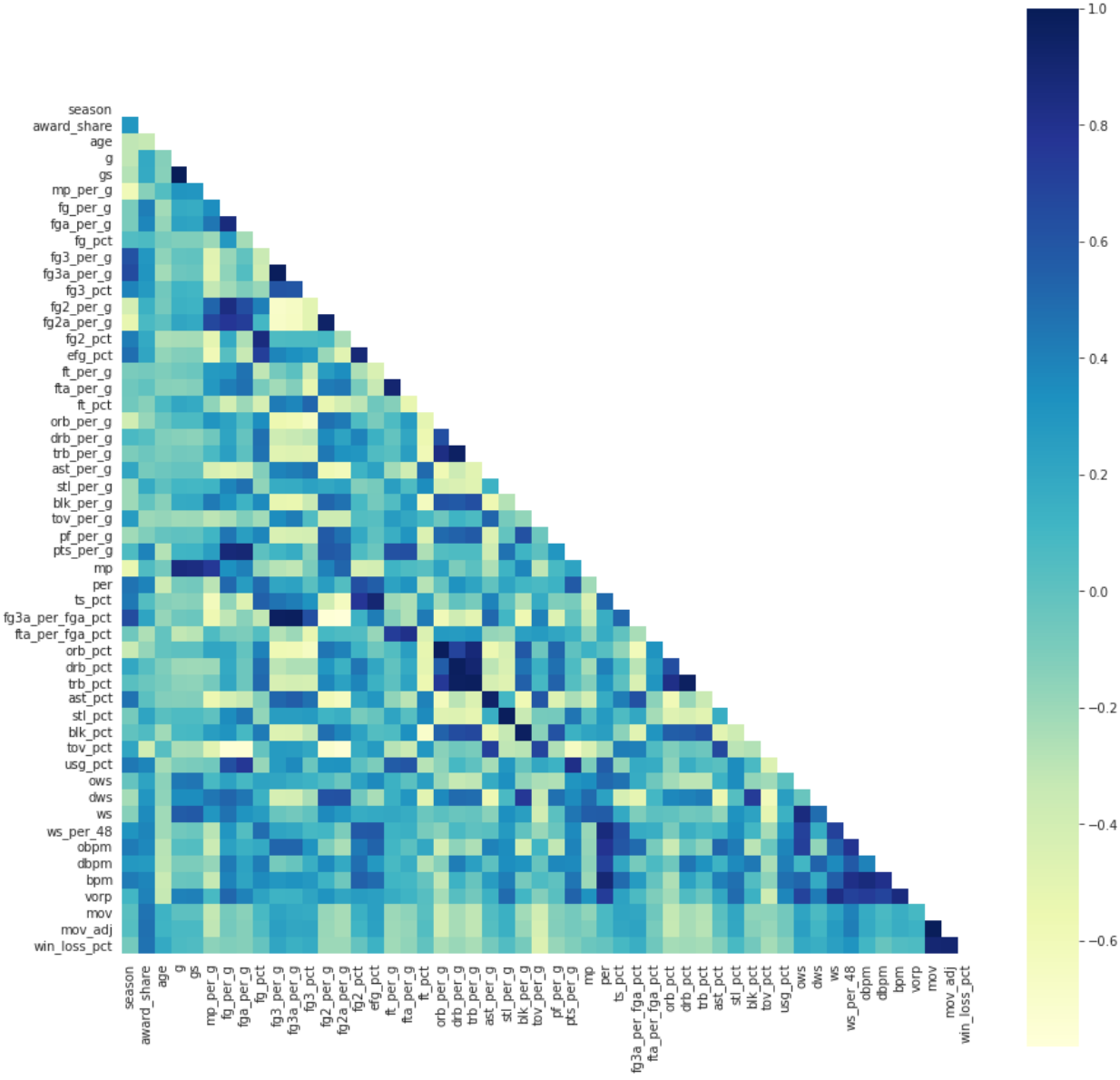


Figure 1: Correlation Heatmap

Next, the major statistical categories based on the heatmap were visualized for past MVP winners, the question here was ‘What stats contributed most to the award share a player received and is there a drop off’? Figure 2 shows just how high some of the winners ranked in various categories, but the tables immediately following answer that question.

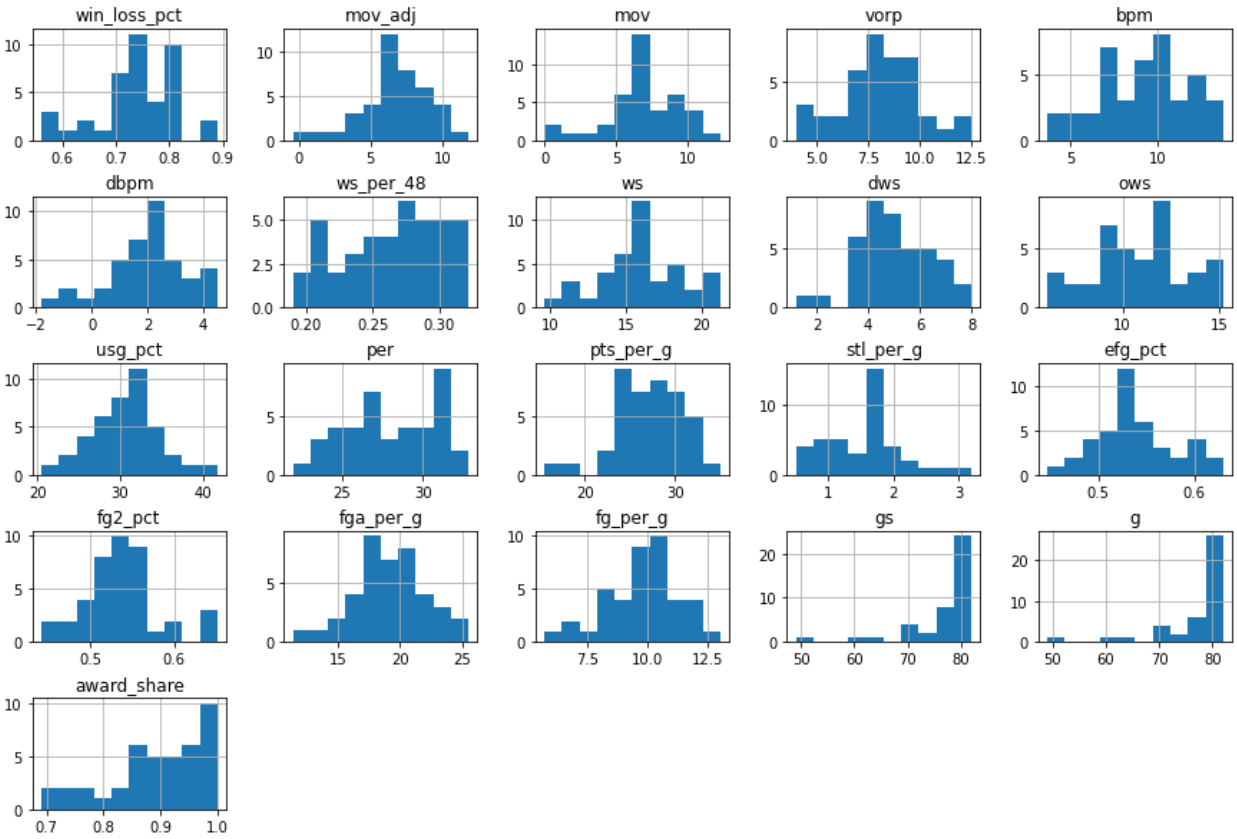
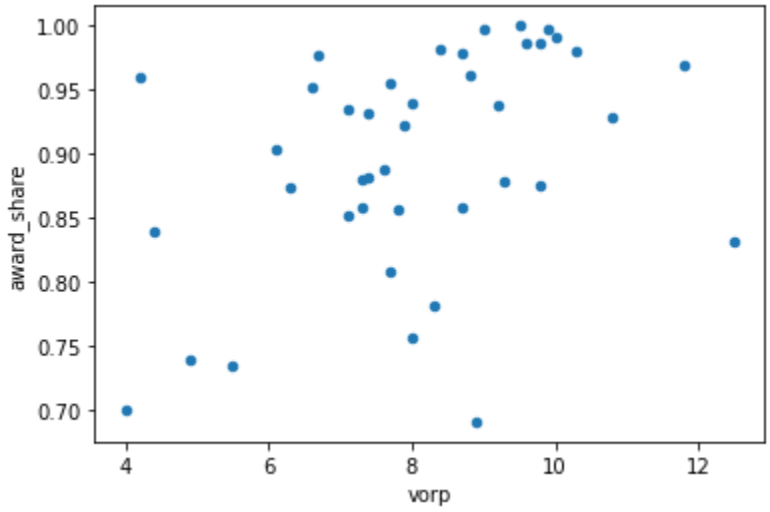


Figure 2: MVP Winners Target Categories

Figures 3 & 4 visualize award share's relation to VORP and Player Efficiency Ratio (PER) present what was already expected. The higher the stat, the higher a player's award share was for that particular season, however, after a certain number the weight of that category dropped. This confirms that no one statistic can determine the winner of the award.

Figure 3: Award_share to VORP relationship



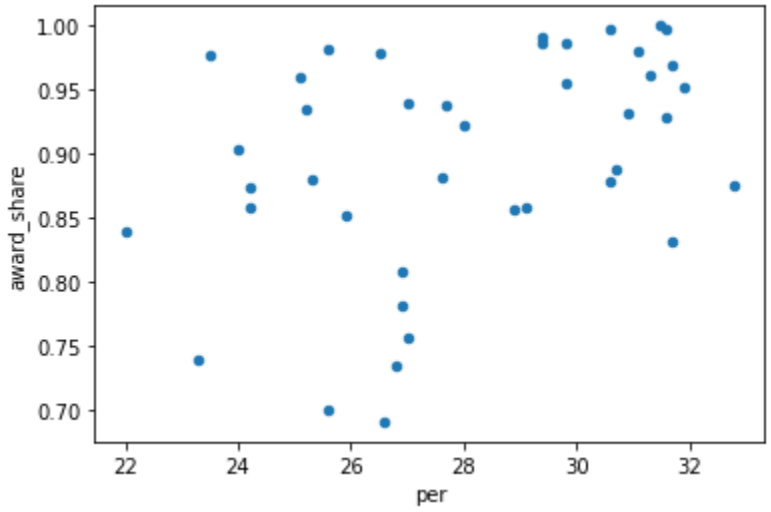


Figure 4: Award_share to PER relationship

Exploratory Data Analysis

Exploratory Data Analysis Notebook

This phase of the project examined the relationships between 'award_share' and other statistical categories. In this phase, the data was also adjusted to a 35-minute-per-game basis. This allowed the comparison of stats for players that have played different minutes to be compared on a per-minute basis. This adjustment was applied to the general dataset and also the MVP data frame, scaling both datasets to a 35-minute-per-game basis. Next, the data in the general dataset was normalized, and the columns containing the stat categories most correlated to the 'award_share' metric were split off from the data. This was done to perform a mutual information score test between the categorical values and the target variable. In doing this, a few stats that were somewhat correlated to the key metric were dropped. Based on the results from Figure 5, only the categories with the highest mutual information scores were saved to the testing data frame that would be used for modeling.

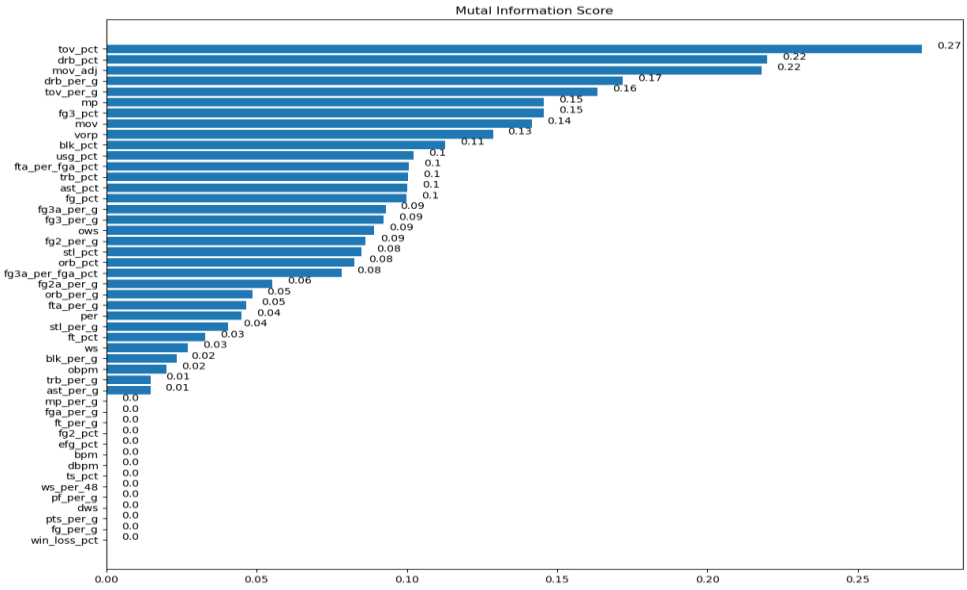


Figure 5: Mutual Information Scores

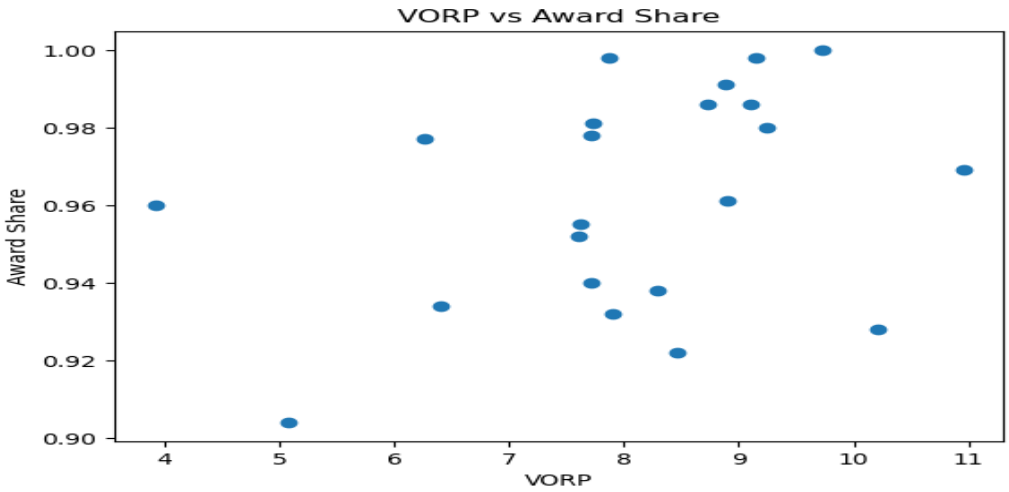


Figure 6: VORP vs Award_share

Preprocessing

Preprocessing Notebook

This phase of the project was all about preparing the data for modeling. Here the data was reviewed and all changes made so far were double-checked to make sure there wasn't any noise that could interfere with the results. Finally, the results obtained from the mutual information score test were used along with the master data frame to create a new data frame that would be ready for testing. In this phase, the dataset that would be used to answer the main problem statement, 'How can we predict the NBA MVP given all of these variables?' was prepared. Data from the 2022-2023 (2023) NBA season was scrapped and prepared for modeling.

Modeling

Modeling Notebook

In the final phase of the project, with the data already preprocessed and ready for testing, 4 different methods were tested. Linear Regression, Random Forest Regression, XGB Regressor, LGBM Regressor. Each model was tested on every season from 1982 – 2022 using the stats deemed most appropriate through the heatmap from the data wrangling phase and mutual information scores from the EDA phase. Based on the results shown in Figure 7, the models are only predicting the correct MVP around 60% of the time. While these results are acceptable, they can be improved upon.

	Model	average MAE	average R squared	accuracy
0	Linear Regression	0.020504	0.161206	0.575000
1	Random Forest Regressor	0.004980	0.626316	0.600000
2	XGBoost Regressor	0.007613	0.629775	0.550000
3	LGBM Regressor	0.005126	0.630324	0.550000

Figure 7: Model Results Summary

Future Improvements

In the future, I plan to use hyperparameter tuning to quickly find the best parameters for each model in order to obtain higher accuracy scores for each model. The goal of this project is to get as close to 100% as possible, that is predicting the MVP each season for as many seasons as possible. I would also like to fine-tune the models so new data can be imported after every season and the MVP for the upcoming season can be predicted based on the latest data.

Conclusion

In conclusion, the model has predicted that the winner of the NBA MVP award for the 2022-2023 season will be Nikola Jokic (4/9/23). However, there are a large variety of statistical factors that can affect a computer's ability to predict the NBA MVP. The award is handed out based on human judgment in tandem with player statistics and watching players perform on a nightly basis. In the project, the goal was to remove human judgment and make sound predictions based solely on statistics, eliminating as much bias as possible. As it stands, a computer is only able to match the predictions of human judges 60% of the time when considering similar factors. The outcome of this project raises yet another question, how many of the past MVP winners won the award solely because of inherent bias instead of statistical measures?

1	Model	1st Place	1st Place Share	2nd Place	2nd Place Share	3rd Place	3rd Place Share	4th Place	4th Place Share	5th Place	5th Place Share
0	Random Forest	Nikola Joki?	0.586809	Shai Gilgeous-Alexander	0.364843	Joel Embiid	0.355137	Luka Don?i?	0.33465	Giannis Antetokounmpo	0.258198
0	Linear Regression	Nikola Joki?	0.114897	Luka Don?i?	0.102383	Giannis Antetokounmpo	0.08859	Joel Embiid	0.084646	Shai Gilgeous-Alexander	0.077095
0	XGBoost	Nikola Joki?	0.692494	Luka Don?i?	0.444728	Shai Gilgeous-Alexander	0.390531	Giannis Antetokounmpo	0.309286	Joel Embiid	0.30599
0	LightGBM	Luka Don?i?	0.625549	Nikola Joki?	0.448166	Joel Embiid	0.307691	Shai Gilgeous-Alexander	0.270732	Giannis Antetokounmpo	0.197185

Figure 8 2023 MVP Prediction