Utilizing Neural Networks and LASSO Regression to Predict MVP

Nicholas Liagridonis, Lawson Vitosh
DATA 22100

Section I: Abstract

This project delves into predicting NBA Most Valuable Player (MVP) award winners and understanding the evolution of MVP voting trends over time. Utilizing a custom dataset compiled from Basketball-Reference.com, we collected yearly MVP voting data starting from 1982. We compared our approach to a similar Kaggle project by Robert Sunderhaft, which employed Random Forest and XGBoost models. Our work builds on this by implementing neural networks and Lasso regression, aiming for improved prediction accuracy and interpretability. Exploratory data analysis revealed a significant class imbalance, addressed by setting cutoffs and employing Synthetic Minority OverSampling Technique (SMOTE). We then trained a neural network model, achieving an 80% accuracy in predicting the correct MVP and a 99% accuracy in the top two predictions. Additionally, decade-specific models were developed to assess MVP preferences over time, albeit with lower accuracy due to limited data. The Lasso model was introduced to provide insights into variable importance across different eras. The model achieved an R^2 value of 0.76, demonstrating fairly decent predictive power. Visualizations showcased a positive correlation between actual and predicted MVP voting shares.
By combining machine learning techniques with domain knowledge, this project contributes to the understanding of MVP selection dynamics and offers valuable insights for basketball enthusiasts and analysts alike.

Section II: Dataset

This dataset was constructed by the team for this project. A web scraper using the python package 'requests' collected data from 3 databases on Basketball-Reference.com [1]. We acquired data on player's statistical totals per game, player's advanced statistics, and their MVP votes, which were all on different pages. We collected these full tables for all active NBA players from the 1982-83 season through the 2022-23 season [2]. Following this process was a cycle of data cleaning and processing, adjusting the tables to synthesize into 1 data frame. We then filtered out some players from the dataset that played minimal minutes/games that would skew our results significantly.

Since modern NBA MVP voting had not started until 1982, only data after 1982 was used. NBA MVP candidates receive first through fifth place votes from 100 different judges. First place votes are worth the most, and the player with the largest voting share is selected to be the MVP.

While accurately predicting the NBA MVP using the entire breadth of data interested us, we also attempted to understand how the MVP voting has changed over time. This data will allow us to explore what the ideal basketball player has looked like in different time periods.

Section III: Literature Review

We compared our process and results to a similar project found on Kaggle by Robert Sunderhaft [3]. The author of that project predicted the NBA MVP using Random Forest and XGBoost models. They also dealt with the class imbalances in similar ways we did (by setting cutoffs and using SMOTE). Their best model, the XGBoost model, was able to predict the correct NBA MVP for each year about 78% of the time, and that the top 2 predictions contained the real MVP 98% of the time. We attempted to build off their work by using different cutoffs for class imbalance, using different features for prediction, using neural networks, and predicting by decade.

Section IV: Exploratory Data Analysis

When we first started to explore the data, we realized that there was a large class imbalance: most NBA players do not ever receive an MVP vote. We needed to find a way to balance the classes, or else a model might just predict all players to receive 0 votes, and this would have over 96% accuracy since only 4% of players ever receive a vote. However, we suspected that MVP candidates and average NBA players would on average have much different stats.

Figure 1: How Many Minutes NBA Players Play Per Game

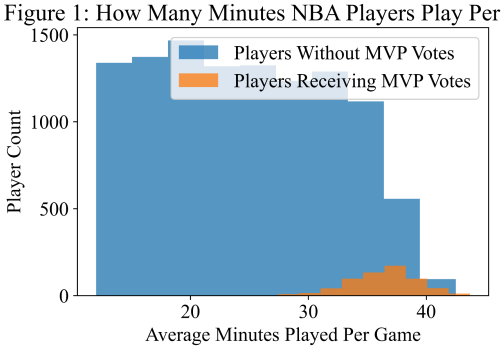


Figure 1 is a histogram, showing how many minutes each player averaged per season, with players who do not receive an MVP vote in blue and players who do receive an MVP vote in orange. It is obvious that the groups differ, so in order to balance the

groups we set cutoffs to get rid of observations in the majority group. For example, we set a cutoff of minutes played at 25 and greater, since MVP candidates do not play less than that. After setting cutoffs for PER at 15 and VORP at 2, the MVP candidates made up 30% of the observations. Next, we used Synthetic Minority OverSampling Technique (SMOTE) to create synthetic observations for MVP candidates. This method performs better than random undersampling of the majority group and random resampling of the minority group. Now the classes have perfect balance, although predictions will not be made on the synthetic players (1339 of each).

Section V: Model 1 - Neural Network

In order to predict the NBA MVP, a neural network was trained. Because a neural network is capable of its own feature engineering, the input to the model was all 20 advanced statistics. The MVP was estimated each year by leaving one out cross validation. The data was trained on the other 40 seasons, while the MVP voting share was estimated on the last season. The predicted MVP was the player with the most predicted MVP voting shares each season.

The best performing model was a linear network with one hidden layer with 30 nodes, which was capable of predicting who won the NBA MVP correctly 80% of the time, and predicted the correct MVP in the top two scores about 99% of the time, slight improvements to the XGBoost model found by Sunderhaft.

Nodes	10	20	30	40
MVP	59%	59%	80%	65%
Top 2	93%	95%	~99%	90%

Model	Complex NN	XGBoost
MVP	34%	78%
Top 2	51%	98%

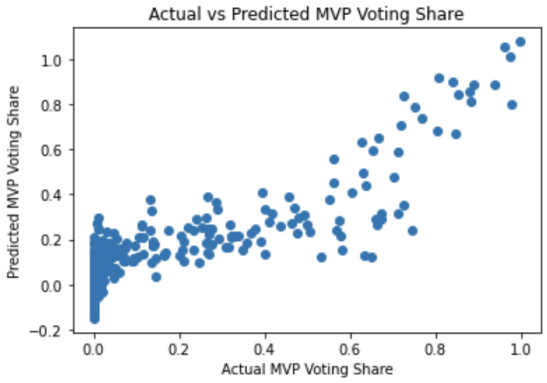
The neural networks seemed to be very capable of predicting the real MVP to be in the top two players with the highest predicted MVP voting share score. There seems to be a trend that around 30 nodes is enough to accurately predict; anymore, and the model starts to overfit. This is the same with the complex neural network, which also overfit with 195 fully connected nodes in 7 layers., and performs worse.

Next, a model was fit to train on each decade (1980's - 2010's) to see how MVP preferences have changed over time. By training on one decade, then testing on the others, how voting preferences over time have changed. These models were less accurate. It is likely that the one quarter of observations in each decade is not enough data to accurately train the model to predict the MVP. While every decade had roughly similar top2 prediction rates, the 2010s had the best MVP prediction rates.

	80s	90s	00s	10s
MVP	32%	17%	24%	34%
Top 2	71%	76%	76%	78%

Section VI: Model 2 - Lasso

We explored this idea more with a LASSO model, wanting coefficients to help interpret what features are more or less important to an MVP candidate now than in the 90s. After creating the model and cross-validating for an alpha value, we inspected which variables remained. The LASSO model ended up dropping a large set of variables, and some of what remained was quite interesting. Certain team column dummy variables remained, suggesting that the team the player was on played a role in whether or not they won MVP historically. Both per-game stats and advanced stats were dropped, particularly if there was an advanced stat that was derived from a per game stat, typically only one would be kept. The model peaked at an R^2 value of 0.76. Below is a plot of the actual MVP Voting shares vs the predicted ones, which show a convincing positive correlation.



Leave One Out Model:

Picked the correct MVP: 56%
Top two: 90%
Top features for all decades: VORP, WS/48, OWS, USG%, TOV%
Shrunk to 0: BPM, WS, DWS, TRB%, DRB%

BY DECADES:

Top features for 80s: VORP, WS/48, PER, TOV%, OWS
Shrunk to 0: DRB%, WS, BPM

Top features for 90s: PER, TOV%, DBPM, DWS, VORP
Shrunk to 0: TRB%, BMP

Top features for 2000s: VORP, WS/48, TOV%, OWS, USG%
Shrunk to 0: PER, DRB%, DWS, WS, BPM

Top features for 2010s: VORP, WS/48, PER, USG%, 3PAr
Shrunk to 0: TRB%, OWS, BPM

For all models, VORP (value over replacement), WS/48 (the amount contributed to a win for scaled for a 48 minute game), PER (player efficiency rating), and USG% (how much a team uses that player) were among the top features contributing to MVP vote share, all of which one would expect to show how good a player is and how much they make their team win. Many rebounding stats (such as TRB% and DRB%) might not matter as much, since the Lasso model shrunk those coefficients to zero. While most features are constant across all models, the model trained on the 2010's shows that 3PAr (percentage

of field goal attempts made from behind the 3 point line) is more important. Modern basketball favors 3 pointers more, and MVP voting reflects that, giving more voting share to players who shoot more 3 pointers than 2 pointers.

Section VII: Comparison of the Two Models

The best neural network model has much higher predictive accuracy for predicting the actual MVP than the best Lasso (80% compared to 56%). However, Lasso could predict the real MVP in the top two predicted vote shares with 90% accuracy which is lower than the neural network's (99%), but is not a bad prediction. If predictive power is the only factor, the neural networks should be used. The neural networks seem to be able to handle this kind of data well, especially considering there are synthetic observations in the training data.

The Lasso model has two distinct advantages. First, it is much quicker to train. The neural network took 3.5 minutes to train while the Lasso took 2.2 seconds to train. If the goal was to make quick predictions and we only cared about predicting the MVP in the top two of voting share, a Lasso could do almost as good as the neural network, but is trained much quicker.

The second advantage of the Lasso model is the coefficients. This lets us know what features are important. Since the actual MVP voting committee considers stats, this method is good at representing how voters might think.

Section VIII: Conclusion

In conclusion, we were able to create models to accurately predict NBA MVPs with up to 80% accuracy, even after having to create synthetic observations to balance the groups. We were also able to track how different stats have become more or less important for selecting MVPs over different decades.

Section IX: Bibliography

[1]: ***Basketball Statistics & History of every Team & NBA and WNBA players*****. Basketball. (2024, March 8). https://www.basketball-reference.com/**

[2]: ***2022-23 NBA awards voting*****. Basketball. (n.d.). https://www.basketball-reference.com/awards/awards_2023.html**

[3]: Sunderhaft, R. (2022, July 13). ***Predicting the NBA MVP*****. Kaggle. https://www.kaggle.com/code/robertsunderhaft/predicting-the-nba-mvp#Test-Set**