Predicting NBA Defensive Player of the Year Awards: A Machine Learning Approach

Hannah Kim, William Mayes

2023-12-12

Motivation

This paper explores the application of various machine learning models to predict the NBA Defensive Player of the Year (DPOY) award winners from 1983 to 2023. Utilizing a comprehensive dataset from Kaggle, which includes individual player statistics, team success metrics, and award information, we examined models like Ridge Regression, Lasso, Random Forest, Linear Regression, Classification Tree, Logistic Regression, and radial SVM. Our analysis reveals that more complex models, particularly the radial SVM, significantly outperform simpler models, with some key predictors being playoff appearances and defensive win shares (DWS).

The NBA Defensive Player of the Year award, a prestigious accolade in professional basketball, recognizes the league's best defensive player each season. Predicting the winner of this award not only has implications in the realm of sports analytics, but also offers insights for sports betting enthusiasts. Our project aims to apply machine learning techniques learned in our class to predict DPOY winners, bridging the gap between academic knowledge and practical application in sports analytics and gambling. This project is motivated by a fascination with the intersection of sports, data analytics, and betting, combined with an academic interest in validating the effectiveness of these machine learning models in real-world scenarios. The dynamic, unpredictable nature of sports achievements, or awards, presents us with a unique challenge for predictive modeling.

Methodology

We sourced our data from a Kaggle dataset comprising player and team performance metrics from 1983 to 2023. Some key features included team playoff participation (Playoffs), defensive win shares (DWS), and defensive box plus minus (DBPM), minutes played(MP), etc. Our methodology involved training various models (Ridge Regression, Lasso, Random Forest, Linear Regression, Logistic Regression, and radial SVM) and splitting the data into training and testing sets. Furthermore, we adjusted our training data to only include candidates for the award, so our model could get better acquainted with the parameters that actually mattered for predicting the winner.

The rationale for including Ridge Regression, Lasso, and Linear Regression was to explore how models typically not tailored for classification tasks would perform in predicting the DPOY winner. Interestingly, while there was some overlap in the influential predictors identified by both sets of models, the regression techniques also highlighted some unexpected choices as the most influential predictors, such as "ws_48", which no classification models had as an impactful parameter. Only proving that they weren't as well suited for the task.

The core of our analysis centered on the classification models, with SVM emerging as the most effective in predicting the DPOY winner. This finding was supported by the performance of Random Forest and Classification Tree models, which also proved to be more useful than the regression models.

To assess the accuracy and performance of these models, we employed confusion matrices and compared the predicted outcomes against actual results. Our methodologies involved an initial phase of training the models with training data, followed by optimization processes to fine-tune their predictive capabilities. Subsequently, the models were tested with a separate set of test data to validate their predictive accuracy. The SVM model consistently demonstrated the highest accuracy in these tests, affirming its suitability for predicting the DPOY winner in the NBA.

Results

The radial SVM model demonstrated the highest accuracy, with ~ 98% on training data and 100% on the test set. Comparatively, other models showed varying degrees of accuracy, from ~ 84% - 93%, with complex, classification, models generally outperforming regression models (Linear Regression, etc.). The most significant predictors of DPOY winners were found to be playoff appearances and DWS, and DBPM. We found a lot of the defensive predictors we thought would be influential weren't. This we found, was do to the fact the statistics DBPM and DWS already take into account these other defensive stats. These stats included defensive rebounds, steals, blocks, etc. This led to what looks like few things playing into who the winner is, when in reality there are many factors rolled into these few.

The superior performance of the SVM model is attributed to its efficacy in classifying binary outcomes, which aligns well with the "win or lose" nature of the DPOY award. Our project demonstrates the potential of advanced machine learning models, particularly SVM, in accurately predicting sports awards. We believe these methods can be applied to any sports awards wherein you have the adequate data, winner information, individual stats, team stats, etc. That being said, these findings have implications for sports betting and the broader field of sports analytics. Future research could explore the incorporation of more diverse data sources, exploring different predictors and different classifications models.

References

https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats/data