Predicting NBA MVP Using Machine Learning

Wahyunan Andika

August 2024

1. **Introduction**

When we talk about the NBA, we always hear about famous basketball players like Michael Jordan, Allen Iverson, Kobe Bryant, Shaquille O'Neal, LeBron James, and Stephen Curry. So what do these players have in common? Is it because they are amazing scorers or have the ability to make clutch shots at critical times? Well, those are all true, and one thing they also have in common is that they are winners of the NBA MVP award. So what’s the NBA MVP? The NBA MVP is one of the most prestigious honors in the NBA. This award is given annually to the player who is considered the most valuable to their team during the regular season, based on their performance, impact on games, and overall contribution.

Predicting the MVP winner is a popular activity, especially for betting purposes, and involves analyzing various factors like player stats, team dynamics, and historical performances. This task can be quite complex, given the vast amount of data and the need for accurate predictions.

In my work, I’ve tackled this challenge by extending the dataset from 1991 to 2024. I use data from the 90s era because one of the players holds the record for the highest career three-point shooting percentage, and three-point shooting is closely associated with the modern basketball era.

To enhance the prediction process, I’m using advanced Machine Learning techniques, including time series cross-validation (TSCV) and additional metrics like MAP@K. These methods not only aim to predict the MVP winner but also identify the top 5 candidates.

Furthermore, I’m experimenting with Principal Component Analysis (PCA) to see how dimensionality reduction impacts model performance. This approach may help refine the accuracy of predictions and capture evolving player trends more effectively.

1. **Related Works**

Predicting NBA MVP using Machine Learning by Gabriel Pastorello

Gabriel Pastorello aimed to predict the NBA MVP for the 2022 season using machine learning models. His approach involved analyzing player performance data from the NBA from 2000 to 2021.

The model that he used was Support Vector Machines, Elastic Net, Extreme Gradient Boosting, Random Forest, Adaboost, Gradient Boosting, Light Gradient Boosting Machine, with evaluation metrics RMSE and R².

His result were:

* SVM: RMSE = 0.087, R² = 0.867
* Elastic Net: RMSE = 0.153, R² = 0.585
* Random Forest: RMSE = 0.096, R² = 0.837
* AdaBoost: RMSE = 0.119, R² = 0.752
* Gradient Boosting: RMSE = 0.108, R² = 0.794
* LGBM**:** RMSE = 0.107, R² = 0.797

His strength for his works was:

* Diverse Model Comparison (he knew that the dataset was high dimensional, so he used complex model for diversity and comparison).
* Using both RMSE and R² allows for a thorough evaluation of model performance. RMSE measures prediction accuracy, while R² provides insight into the variance explained by the model.
* Wide Dataset (He used a dataset where physicality is less emphasized compared to the 90s season, and the 2000s are characterized by the beginning era of more flexible players.)
* High Performance

His Weakness for his works was:

* Model Complexity (Some Model like XGB and LGBM, are complex and require high computation resource and times.)
* Evaluation Metrics (RMSE and R² are useful but may not fully capture the nuances of MVP prediction. Including additional metrics like MAP@K could provide more insights into model performance.)
* High Dimensionality Dataset

Opinion:

In my work, I’ve expanded the dataset to cover the years 1991 to 2024. I chose Steve Kerr as a benchmark because his record for the highest career three-point shooting percentage marks a significant milestone in the evolution of three-point shooting, representing the beginning of the modern NBA era. I’m also using time series cross-validation (TSCV) and additional metrics like MAP@K to not only predict the MVP winner but also to identify the top 5 candidates. Additionally, I’m experimenting with PCA to explore how dimensionality reduction affects model performance.

1. **Dataset & Features**

The dataset is sourced from web scraping Basketball Reference (https://basketball-reference.com/). It consists of four major components:

* Player Statistics per Season (1991-2024): Detailed stats for individual players each season.
* Player Advanced Statistics per Season (1991-2024): Advanced metrics that provide deeper insights into player performance.
* Team Performance per Season (1991-2024): Data on how teams performed each season.
* MVP Candidates per Season (1991-2024): Information about players considered for MVP each season.

Preprocessing begins after exporting the data from web scraping to CSV files. Missing values are addressed, particularly in shooting percentages where some players didn’t attempt certain shots (like free throws, three-pointers, or field goals). Player names are cleaned by removing any asterisks that indicate MVP winners. For the position column, where players may have multiple positions listed, it is standardized to reflect their primary role or the position they played most during that season. In the GB column, which shows the discrepancy between the highest win and the selected team’s win, '-' is replaced with 0 to simplify the data.

For the train-test split, data from 1991-2023 is used for training and 2024 for testing, with time series cross-validation (TSCV) applied using 5 splits due to the temporal nature of the data. Data normalization is performed as part of the preprocessing pipeline once all datasets are merged.

The dataset includes a comprehensive set of features related to player and team performance, with a total of 61 features:

* Player Information: Includes player name ('Player'), position ('Pos'), age ('Age'), and team ('Tm').
* Game Statistics: Covers games played ('G'), games started ('GS'), minutes per game ('MP'), field goals made ('FG'), field goals attempted ('FGA'), and shooting percentages ('FG%', '3P%', '2P%', 'eFG%', 'FT%').
* Shooting Stats: Details three-point shooting ('3P', '3PA', '3P%'), two-point shooting ('2P', '2PA', '2P%'), and free throws ('FT', 'FTA').
* Rebounding and Defense: Includes offensive rebounds ('ORB'), defensive rebounds ('DRB'), total rebounds ('TRB'), steals ('STL'), blocks ('BLK'), turnovers ('TOV'), and personal fouls ('PF').
* Scoring and Efficiency: Metrics such as points scored ('PTS'), player efficiency rating ('PER'), true shooting percentage ('TS%'), and various advanced stats including usage rate ('USG%') and win shares ('WS', 'WS/48').
* Advanced Metrics: Features like offensive box plus-minus ('OBPM'), defensive box plus-minus ('DBPM'), box plus-minus ('BPM'), and value over replacement player ('VORP').
* Team Performance: Team-specific data such as team wins ('W'), losses ('L'), win/loss percentage ('W/L%'), games behind ('GB'), points scored per game ('PS/G'), points allowed per game ('PA/G'), and simple rating system ('SRS').

These features provide a thorough view of player performance, team dynamics, and advanced metrics over the seasons from 1991 to 2024.

1. **Methods**

* X