Modeling Approach for 2024 MVP Prediction

Introduction

My primary goal was to forecast player performance using the PIE metric and, ultimately, predict the season’s Most Valuable Player (MVP) for 2024. Initially, I set out to build a regression model to estimate PIE values but later pivoted to a classification approach after the regression model identified a different candidate than expected. This document summarizes my process, the results I obtained, and outlines next steps for refining the models.

Data Preparation and Feature Engineering

I consolidated multiple datasets that included player statistics, team on/off-court metrics, records, and historical MVP voting data. Key steps in this phase included:

* Data Aggregation: Merging player-level data by team and season and standardizing column names to create a unified source.
* Metric Computation: Engineering advanced performance metrics like PIE (as a share of total season performance), points per game, rebounds per game, assists per game, and various shooting percentages.
* Handling Missing Data: Applying mean imputation to features such as weight and height, and binning these values to create meaningful categorical groups.
* Data Transformation: Converting percentage strings into numerical values to ensure consistency in model input.

Modeling Approach

Regression Model with XGBoost

My first approach was to predict the PIE metric using an XGBoost regressor. I trained the model on pre-2024 seasons and then tested it on the 2024 data.  
Key Evaluation Metrics: RMSE, R², MAE, and MAPE all indicated that the model captured player performance nuances fairly well.

Feature importance from this model highlighted:

* mvp\_rank as the most influential,
* Followed by min\_per\_game and pts\_per\_game.

Below are the XGBoost model-based importances:

feature importance

14 mvp\_rank 0.543672

3 min\_per\_game 0.245622

0 pts\_per\_game 0.067443

11 pts\_share 0.059989

13 pts\_won 0.019075

2 reb\_per\_game 0.010680

4 fg\_pct 0.009854

6 win\_pct 0.007084

8 eFG% 0.005776

7 TS% 0.005729

Classification for MVP Prediction

After noting that the regression model did not identify the MVP candidate I expected, I turned my attention to a classification strategy. Using a cost-sensitive RandomForestClassifier—with SMOTENC to oversample the minority MVP class—and Bayesian hyperparameter tuning (via BayesSearchCV), I built a classifier that improved precision and recall on the MVP predictions.

The classification model ultimately identified A'ja Wilson as the top candidate for the 2024 MVP season, with an MVP score of approximately 0.9250. Key performance metrics for this candidate include:

* PIE: 0.03
* Points per Game: 26.87
* Rebounds per Game: 11.87
* Assists per Game: 2.32
* Minutes per Game: 34.41
* Field Goal Percentage: 0.52
* Win Percentage: 0.68

For reference, here are the top 5 MVP candidates for 2024:

player\_id player\_name team\_id team\_code mvp\_score PIE pts\_per\_game reb\_per\_game ast\_per\_game min\_per\_game fg\_pct ft\_pct win\_pct wins losses

0 1628932 A'ja Wilson 1611661319 LVA 0.925010 0.033747 26.868421 11.868421 2.315789 34.411281 0.518170 0.843636 0.675 27 13

Permutation Importances

To further understand which features were driving my predictions, I computed permutation importances. Here are the results for all features and for numerical features only.

Regression Results

My final regression predictions for the 2024 season are as follows:

player\_id predicted\_PIE target\_PIE abs\_error player\_name

0 1629483 0.024782 0.021816 0.002966 Napheesa Collier

1628932 0.024050 0.033747 0.009697 A'ja Wilson

1627668 0.023990 0.023968 0.000022 Breanna Stewart

While these predictions gave me insight into the PIE metric, the classification model proved more effective for pinpointing the MVP.

Classification Results

The MVP classifier, after tuning and resampling, identified A'ja Wilson as the top candidate:

🏆 Predicted MVP for 2024 Season 🏆

Player: A'ja Wilson (ID: 1628932)  
Team: LVA (ID: 1611661319)  
MVP Score: 0.9250

Key Performance Metrics:

* PIE: 0.03
* pts\_per\_game: 26.87
* reb\_per\_game: 11.87
* ast\_per\_game: 2.32
* min\_per\_game: 34.41
* fg\_pct: 0.52
* win\_pct: 0.68

And here is a snapshot of the top 5 MVP candidates:

player\_id player\_name team\_id team\_code mvp\_score PIE pts\_per\_game reb\_per\_game ast\_per\_game min\_per\_game fg\_pct ft\_pct win\_pct wins losses

0 1628932 A'ja Wilson 1611661319 LVA 0.925010 0.033747 26.868421 11.868421 2.315789 34.411281 0.518170 0.843636 0.675 27 13

Key Observations

* Top Features: Across both permutation and model-based importances, min\_per\_game and pts\_per\_game stood out as key predictors, with mvp\_rank being particularly influential in the XGBoost model when using regression for PIE and PIE when using our classification for MVP.
* Model Pivot: Initially using regression to predict PIE, I observed a divergence between predicted and expected MVP candidates. This led me to employ a classification model better suited to handling the imbalance in MVP instances.
* Evaluation Techniques: Beyond traditional accuracy metrics, I used SHAP and Shapash for a deeper understanding of feature contributions.

Future To-Dos

Based on my findings and lessons learned from this season, my future plans include:

* Enhanced Feature Engineering: I will spend more time refining and experimenting with additional features to capture nuances in player performance.
* Incorporate Time-Series Data: I plan to develop a wins time-series LSTM/RNN model to integrate wins/losses dynamics into the PIE regression, as team performance is a crucial factor in MVP selections.
* Deepen Model Interpretability: I will focus further on SHAP/Shapash analyses to better understand and build on the most impactful metrics.
* Ongoing Model Tuning: I will continue to refine the Bayesian optimization and oversampling techniques to improve precision in MVP classification.

Conclusion

This integrated approach—combining rigorous data preparation, targeted feature engineering, and iterative model development—has allowed me to derive actionable insights from raw sports data. While the regression model provided a solid baseline for predicting PIE, my classification framework ultimately identified A'ja Wilson as the top MVP candidate for 2024. I am excited to build on these results with additional feature enhancements and time-series integration in the future.