Mystics Data Analyst Role – Technical Exercise Part 2: Modeling Evaluation

Vitor Freitas

The goal of this exercise was to build a model to predict the likelihood of a player winning the WNBA MVP award in a given season using data from the 2016 through 2024 regular seasons. The provided datasets included MVP voting results, team records, player biographical information, and player/team play-by-play and box score stats aggregated at the season-team level. I began my approach by analyzing and understanding the structure of the provided datasets. The data included player stats and team performance metrics, so my first step was to merge these datasets into a single dataframe using common identifiers like nba\_person\_id, nba\_team\_id, and nba\_season. I did not find it necessary to use any additional dataset, as the data provided is already pretty large and complete.

One key challenge was handling players who played for multiple teams in a single season. While those players are unlikely to contend for MVP, I decided to aggregate their stats across all teams for that season to maintain data consistency to ensure that their total contributions were fully captured, even if they switched teams mid-season. I also standardized player and team stats within each season to account for differences in performance across years. I then created a binary column called ‘MVP’ to be used as the target variable, which was set to 1 if a player won the MVP in that season (mvp\_rank == 1) and 0 otherwise. This allowed me to frame the problem as a classification task.

I selected a Random Forest classifier for this task, because of its capacity to handle non-linear relationships and its ability to calculate feature importance. To improve the model, I utilized GridSearchCV to adjust two hyperparameters: the quantity of trees (n\_estimators) and the maximum depth of the trees (max\_depth). The model ultimately chose max\_depth = 5 and n\_estimators = 100. I also experimented with a Genetic Algorithm for feature selection and dimensionality reduction, as I had used this approach in my master’s thesis, where it significantly improved model accuracy. However, in this case, the results were very similar to those without the algorithm. Given the added time complexity of the Genetic Algorithm and the fact that the Random Forest’s predictions were already fast and accurate, I decided to remove it. I also tried using XGBoost as my prediction algorithm, however the runtime was considerably larger, and the results were about the same.

The model was trained on data from the 2016–2023 seasons and evaluated using the ROC AUC score and a classification report. The model achieved 100% accuracy in predicting all MVP winners from 2016 to 2023, which gave me confidence in its ability to identify the winners. After that, I used the 2024 season as an out-of-sample test set to test the model’s performance on unseen data. The model predicted A'ja Wilson as the 2024 MVP, which aligned with the actual results. However, I noticed that her predicted probability was relatively low considering her dominant individual stats. Upon further analysis, I realized this was likely due to her team’s performance. One of the visualizations I created showed that the Las Vegas Aces (A'ja’s team) finished 4th in the standings in 2024, while all other MVP winners’ teams finished 1st or 2nd. Despite this, A'ja’s MVP probability was almost twice as high as the second-place player, Brenna Stewart, which I believe show the model’s success in balancing individual and team performance.

Finally, I created a couple visualizations to support the analysis, using matplotlib, including a bar plot showing the predicted probabilities for the top 5 players, a horizontal bar plot showing the top 10 most important features for MVP prediction and a bar plot showing the winning percentages and league rank of MVP winners’ teams over time. I also saved the final merged dataset and the full 2024 MVP prediction to .csv files.

Overall, I believe this model successfully demonstrated the ability to predict WNBA MVP winners using player and team performance data. The model highlighted the importance of both individual performance and team success in determining MVP winners. I’m satisfied with the results and believe this model has practical applications in predicting future MVPs and can even be adapted for other tasks such as scouting, fantasy basketball, or team strategy.