Predicting NBA Player Retention: Executive Summary

Erdős Institute Data Science Boot Camp, Fall 2024

Team: Alex Pandya, Peter Johnson, Andrew Newman, Ryan Moruzzi, Collin Litterell
Project Mentor: Nadir Hajouji

Objective: Predict if a given NBA player will play in the next NBA season based on their current season on-court performance, age, experience, salary, and transactions.

Stakeholders: A model for predicting which NBA players will play in the NBA in the next season is valuable to many decision makers including, but not limited to,

- NBA front offices planning roster changes and deciding which players to invest in,
- Sport bettors making long term bets,
- Advertisers planning sponsorships.

Data Collection and Cleaning: Counting statistics data was collected from the official NBA site using Python APIs and custom web scraping tools. Advanced statistics were collected from Kaggle. Transaction data and salary data were scraped from Basketball-Reference.com and hoopshype.com, respectively, using BeautifulSoup. The data was cleaned and missing data (e.g. missing salaries) were imputed. The data was split into a training set consisting of seasons 1990-2016 and a test set with seasons 2017-2022.

Exploratory Data Analysis: We initially aimed to predict player transactions (whether a given player would be traded/waived), but this proved challenging due to weak correlations between statistics and transaction data (magnitude ~0.05). We found appreciable correlations with whether a player stayed in the NBA in the following season (*NBA player retention*), however.

Modeling: Predicting NBA player retention in the following season is a classification problem with time series structure and imbalanced classes. We evaluated 10 models using walk-forward cross validation with the balanced accuracy score as our central metric. Most models were trained on an augmented training set produced using the Synthetic Minority Oversampling Technique (SMOTE) to balance classes.

Final Results: The best-performing model was XGBoost with hyperparameters chosen to maximize balanced accuracy on the (augmented) training data. We evaluated this model's performance on the test set using walk-forward testing (with an expanding window) and achieved a balanced accuracy of about 81%. The model achieved a precision of about 95%, reflecting high confidence in predicting players who stayed, and a recall of about 78%, ensuring most players who stayed were correctly identified. Its specificity of about 84% further highlights its ability to correctly identify players who left. While the negative predictive value (NPV) of about 50% appears lower, this is expected given the model's prioritization of minimizing false positives (high precision) and achieving balanced accuracy (strong recall and specificity) on such imbalanced data. Incorporating additional factors such as G-League data, contract terms, and injuries could likely improve the NPV. In any case, looking at a random sample of our model's false negative predictions, many of these players have at some point gone down to the G-league, suggesting that although these players were misclassified, our model was still able to successfully identify them as “fringe” players who were at risk of leaving the NBA in the near future.

Future Directions: To build on our existing model, some additional data that we could incorporate include:

- Contract terms,
- Injury information,
- G-League data,
- Salary cap and Collective Bargaining Agreement (CBA) data.

Additionally, we could expand our model of NBA players to incorporate other professional basketball leagues, for example the EuroLeague, and track movement across leagues. We could also expand the problem to focus on assigning probabilities that a player will be waived, traded, or not retained in the league at all, rather than pure classification.