Xubing Lai & Jiaxin Luo
Professor Koehler
Final project
2025/12/17

Predicting NBA All-Star Selection Using Season-Level Player Statistics

Introduction

This project studies whether NBA All-Star selection can be predicted using season-level player performance data. Using box score statistics, the task is formulated as a binary classification problem: predicting whether a player is selected as an All-Star in a given season.

Several predictive models are trained and compared, ranging from simple linear methods to tree-based models and a neural network. Model performance is evaluated using ROC-AUC due to the strong class imbalance in All-Star selection. Beyond model comparison, the project also examines cross-season generalization and demonstrates a practical application by producing ranked All-Star candidate predictions.

Data Description

The data used in this project come from a publicly available dataset published on Kaggle, which contains detailed game-level box score statistics for every NBA player. The dataset provides per-game performance data across the entire season, making it a reliable and comprehensive source for modeling player performance.

The data consist of NBA game-level box score statistics from 2023 to 2025, with the 2023–24 season used for cross-season evaluation and application. Each row in the raw dataset represents a single player's performance in one game.

To prepare the data for modeling, game-level statistics are aggregated to the player–season level by computing per-game averages. The final dataset includes common performance metrics such as points, assists, rebounds, shooting percentages, turnovers, and game score. The target variable is a binary indicator of All-Star selection, where players selected as All-Stars are labeled as 1 and all other players are labeled as 0.

Models and Methods

Five models are implemented in this project: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, and a simple neural network (MLP). These models represent different levels of complexity and modeling assumptions, allowing for a broad comparison of approaches.

Logistic Regression provides a simple and interpretable baseline and performs strongly, showing that All-Star selection is largely driven by a few key performance metrics. K-Nearest Neighbors (KNN) predicts All-Star status based on similarity to other players but performs relatively worse due to high dimensionality and class imbalance. Random Forest captures nonlinear relationships and feature interactions and achieves the best overall performance, making it the primary model used in this project. Gradient Boosting incrementally corrects prediction errors using sequential trees and performs competitively, though slightly below Random Forest. Neural Network (MLP) offers a flexible modeling approach but does not outperform tree-based models, consistent with common results on structured tabular data.

All models are evaluated using ROC-AUC to account for class imbalance. In addition to reporting AUC values, ROC curves are used to compare classification behavior across models, and feature importance visualizations are used to compare how Logistic Regression and Random Forest prioritize different performance metrics. These analyses help explain performance differences and guide model selection.

Results and Interpretation

All models achieve strong predictive performance, indicating that All-Star selection is closely related to season-level performance statistics. Among the tested models, Random Forest performs best with a test ROC-AUC of 0.996. Logistic Regression and Gradient Boosting achieve similar performance (both approximately 0.989), followed closely by the neural network (0.987), while KNN performs relatively worse (0.974).

Random Forest performs best because it is well suited to capturing nonlinear relationships and interactions among performance metrics. NBA All-Star selection is not determined by a single statistic, but rather by combinations of factors such as scoring, playing time, and overall efficiency. By aggregating many decision trees, Random Forest can model these complex patterns while remaining relatively robust to noise, giving it an advantage over simpler approaches.

Although Logistic Regression is a linear model, its performance remains very strong. This suggests that All-Star selection is largely driven by a small number of dominant performance indicators, such as points per game and overall game contribution, which can already be captured by linear relationships. Gradient Boosting achieves similar performance, indicating that sequential tree-based corrections add limited additional benefit beyond what is already explained by these strong primary signals.

Despite being a more complex model, the neural network does not outperform the simpler or tree-based models in this setting. This outcome is consistent with common findings for structured tabular data, especially when sample size is limited and class imbalance is severe. With only a few hundred player-season observations and roughly twenty All-Star selections per season, the neural network has limited opportunity to fully exploit its higher capacity and may be more prone to unstable learning or overfitting. In contrast, tree-based models tend to be more data-efficient and better suited for heterogeneous tabular features.

K-Nearest Neighbors performs worst among the tested models, likely due to its sensitivity to high dimensionality and extreme class imbalance. When positive cases are rare, nearest-neighbor comparisons often fail to provide reliable discrimination.

All models achieve high AUC partly because All-Star selection is strongly tied to performance statistics, but also because the dataset is relatively limited in size. With a small number of All-Star observations, models tend to learn stable but coarse patterns, which can inflate ranking-based metrics such as ROC-AUC. ROC-AUC measures the probability that a randomly chosen All-Star is ranked higher than a randomly chosen non–All-Star. In a dataset with over 500 players and only around 20 All-Star selections, this ranking task is inherently favorable: as long as true All-Stars are generally placed above the majority of non–All-Stars, the AUC can be very high even if the exact selection is imperfect. Therefore, the large AUC values reflect strong ranking performance under severe class imbalance rather than flawless prediction, and they reinforce the interpretation of the models as ranking and decision-support tools rather than exact selection mechanisms.

Cross-Season Evaluation

To evaluate generalization across seasons and the efficiency of the model, the Random Forest model is trained on the 2024–25 season and tested on the 2023–24 season. This temporal split provides a more realistic evaluation than a random within-season split.

Under cross-season evaluation, the model achieves a ROC-AUC of approximately 0.988. Although this represents a slight decrease relative to the within-season result, performance remains strong, suggesting that the model captures stable patterns underlying All-Star selection rather than season-specific noise.

Prediction on 2023–24 Season to prove model efficiency

Using the trained Random Forest model, players from the 2023–24 season are ranked by predicted All-Star probability. This ranking demonstrates how the model could be used in practice to identify top All-Star candidates rather than making a strict binary decision.

The Random Forest model shows excellent performance, successfully identifying the vast majority of true All-Stars and generating highly reliable probability rankings. An examination of prediction errors shows that false positives often include players with strong per-game statistics whose actual selection may have been influenced by injuries, limited games played, or team context. False negatives tend to involve players whose selection reflects subjective factors such as fan voting or team success. Cases where high-probability players are not selected do not reflect weaknesses in the model, but rather due to factors outside box score data. Overall, the model effectively captures the statistical foundations of All-Star selection and is best interpreted as a data-driven ranking tool that complements, rather than replaces human judgment in a selection process influenced by contextual and subjective factors.

Conclusion and Next Steps

This project shows that NBA All-Star selection can be predicted with high accuracy using season-level player statistics. Among the models tested, Random Forest provides the strongest performance and generalizes well across seasons. However, the analysis also highlights that All-Star selection depends on factors beyond on-court performance alone.

Future work could incorporate multiple seasons of training data, include advanced metrics or team-level variables, and evaluate predictions in a forward-looking setting. These extensions could further improve the robustness and practical usefulness of the model.