Predicting NBA standings on day-one of regular season

Martin Bogaert, Valentin Brekke, Lucas Leforestier & Clara Schneuwly

Problem statement

Every season, the popular sports media ESPN publishes its predictions for the number of wins of each team in the NBA. The goal of this project is to better ESPN's predictions by leveraging historical data of performance, while identifying the key features that make an NBA team successful.

Data scraping and pre-processing

The first step of the project was to build a dataset that encodes historical performance of teams and players by scraping Basketball Reference. The predictors span three main categories: (a) team historical performance in the last 3 years (number of wins, net points, team rating), (b) historical performance of players in the roster on opening day (points per game and advanced metrics weighted by minutes played and injury risk), and (c) calendar difficulty (average rest, number of back-to-back games, distance traveled). In the dataset, each row represents a team in a given year, and has columns that quantify the team's past performances, the strength of its roster, and the toughness of its schedule. Finally, the target variable is the team's actual win ratio at the end of the season. The final dataset consists of 33 independent features and over 1.1k observations.

Methodology

To predict the number of wins, the project evaluated the performance of a range of regression models. Initially, simpler and more interpretable models such as Linear Regression, Holistic Regression, and CART are employed. Subsequently, the project explored more sophisticated tree-based models, including Random Forest and XGBoost. Finally, it incorporated an MLP neural network. For each relevant model, a grid-search approach is employed to fine-tune parameters. Moreover, time-series cross validation is used to account for the temporal nature of the data and prevent data leakage.

A model that predicts a team's number of wins in the previous season is used as the baseline. Then, the objective is to compete with a model that uses forecasts provided by ESPN. The metrics used to assess the regression include R^2, MAE, and RMSE. The Spearman's ρ and Kendall's τ coefficients are then used to assess the quality of the predicted standings inferred from the win predictions. This comprehensive evaluation allows the project to gauge the efficacy of each model across different dimensions, ranging the prediction itself but also the ensuing standings.

Results

All observations before 2013 are used within training or validation sets while all seasons between 2013-2014 and 2022-2023 are used for model testing. Figure 1 presents a summary of average performance over the testing seasons for all models, where the predictions are gauged through the MAE, and the inferred standings are assessed using Spearman's ρ, where perfect standings achieve a score of 1.0. It is observed that the Random Forest and Neural Network models both outperform the baseline, whereas CART is less performant. Furthermore, the Linear Regression and XGBoost models are the most performant. Relative to the baseline, they reduce the MAE by 17% and 14% and improve the Spearman's ρ by 16% and 11%, respectively. Lastly, no model is able to match ESPN's expertise. However, it is encouraging to observe that the two best models are very close to ESPN's predictions and even outperform them at predicting conference standings for multiple seasons.

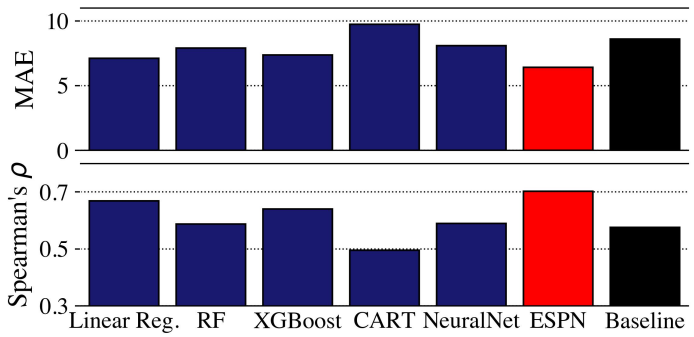

Figure 1. Model performance on predicted wins (MAE) and standings (Spearman's ρ)