Factors influencing offensive success in NBA games

Abstract

My client, a current NBA team, is currently in a rebuilding phase – the team had a poor season record and lost their marquee players to free agency. The general manager and coaching staff are seeking guidance on the types of players to pursue to increase their chances at obtaining a spot in the playoffs.

Design

Looking at box score information provided by https://www.basketball-reference.com/, I examined data from five NBA seasons: the 2014-2015 season through the 2018-2019 season. This provides a large dataset for analysis while avoiding any complications resulting from teams changing names or being impacted by COVID during the seasons spanning 2019 through 2022.

I looked only at regulation games (i.e., no games that went into overtime).

Data

The dataset contains 23,137 rows of data (with 22 columns), where each row represents a team's box score from a game. The columns include numerical data that indicate a team's performance during the game on metrics such as shooting (three points, field goals, free throws), assists, and turnovers, among others. For this analysis, I focused on two sets of features to build predictive models:

1. Shooting related metrics
2. Secondary metrics

Algorithms

Feature Engineering

1. Calculating two-point field goal features (removing three-point features from combined field goal features)
2. Removing collinear features (e.g., dropping field goal features after creating two-point field goal features, dropping features that were totals in favor of percentage features)
3. Separating features into two groups

Models

Linear regression, ridge regression, lasso regression, and ElasticNet regression were used during analysis, with the final model being a linear regression model.

Model Evaluation and Selection

The dataset was split into 60/20/20 train/validation/test sets and then cross-validated across 5 folds for training. The validation set was used to determine predictive accuracy. Once a final model was selected, the model was re-trained on the combined train and validation sets of data and then assessed using the test dataset.

A second model was evaluated and selected looking at “lesser” features that are relevant to a team’s performance. Removing the features related to scoring points, namely the three-point, two-point, and free throw features, this model incorporates secondary metrics of a team’s performance: offensive and defensive rebounds, assists, steals, turnovers, and fouls.

Final linear regression model cross-validation scores: (7 features)

- *On combined train/validation dataset:*
    - Mean R2: 0.95889, std: 0.00129
    - Mean RMSE: 2.53018 , std: 0.04081
- *On testing dataset:*
    - Mean R2: 0.95535, std: 0.00255
    - Mean RMSE: 2.62429, std: 0.07841

Secondary linear regression model cross-validation scores: (5 features)

- *On combined train/validation dataset:*
    - Mean R2: 0.40687, std: 0.01343
    - Mean RMSE: 9.6113, std: 0.10251
- *On testing dataset:*
    - Mean R2: 0.41199, std: 0.02331
    - Mean RMSE: 9.52588, std: 0.19547

Tools

- BeautifulSoup was used to scrape the data
- Numpy and Pandas were used for manipulating the data
- Scikit-learn and Statsmodels were used for modeling the data
- Matplotlib and Seaborn were used for plotting the data

Communication

Please see slides for the final presentation.