NBA Capstone

December 6, 2021

1 Can you buy wins?

An examination of NBA salary cap distribution.

1.1 Overview

Nba roster construction is important. How a team allocates funds in order to produce the basketball team changes each season and often during the season. The NBA currently uses a soft-cap system. This means that each NBA franchise can choose how much money they wish to spend on players with the exception of a salary floor or minimum. Individual players do have a cap or maximum salary. The purpose of this examination of salary cap useage is to determine whether spending more money, or allocating money in a specific manner directly influences a team's win total.

This could be used as a determination for whether a team over or under performs. An evaulation of roster construction or a franchises front office performance.

Another use could be for sports gambling. Every season odds makers set win totals for teams based on all the data available or deemed important, and can also include public perception. This examination is a simple, by only taking into account salaries, form of what book makers do, in an attempt to predict win totals.

1.2 1. Data

Basketball Reference is an encylcopedia of stats with records for each team going back all 75 years of the NBA and ABA. This was the source of Player salaries.

Salary Cap history for the league. was also acquried through basketball reference.

Historical Vegas win totals These were used to determine an accuracy of my model. 2016-17 2017-18 2018-19 2019-20 2020-21

1.3 2. Method

Take player salaries for the past 5 NBA seasons and order them in descending order by: total salary, and cumulative salary against both total team spending for that season and the league's salary cap for that season. Each team can roster up to 15 players at one time, and could roster near 30 different players during a season. At any time during a game, each team will have 5 players playing. And an NBA game lasts 48 minutes before potentially needing an extended period or overtime to determine a winner. Each team has a different approach to minutes allocated for players, however a constant is, your best players play the most minutes. That is why this examination will only focus on the top 10 most expensive salaries.

1.4 3. Data Cleaning

Eastern conference Western conference Together with dataframe munipulations

For future predictions 2021-22 Season

Problem 1: The data collected was in text form. Extraneous characters needed to be removed. Then formated into dictionaries.

Problem 2: Converting salary information into a dataframe of the desire format. Reshaping and orienting the data so there is a column for each feature (player salary) and a row for each entry (team by season).

Problem 3: Change data types of salary information.

Problem 4: Creating features. From the salary information develop columns for each player's salary: as a cumulative percent of a team's total salary spending, as a cumulative percent of the cooresponding season's salary cap.

Problem 5: Converting team win totals into win percentages. Over the past 5 seasons the number of regular season games has been different. E.g. Winning 50 games in an 82 game season is not the same as winning 50 games in a 72 win season.

1.5 4. EDA

EDA

Seaborn pairplot showing each feature plotted against win percentage

1.6 5. Modeling

Pre-processing

Modeling

For pre-processing, StandardScaler, PowerTransformer, and MinMaxScaler were tested. StandardScaler provided the best results.

After StandardScaler was selected, Linear regression, Ridge regression, Lasso, and Random Forest models were created. Random Forest had the best results. I chose to use mean absolute error as the criteria to judge models. This is because MAE doesn't penalize large error. Large error is not uncommon for NBA win totals. Should the best player on a team miss an entire season due to injury that team's win total will not be predicted very well.

These are the difference summaries of the models. Absolute value of y_pred - y_test. Random Forest had the best average while also having the single worst prediction. This is the situation I was looking for.

Random forest specifics Trees with less depth had worse r2 but did predict better. Ultimately a tree with depth 2 did the best predictions on average.

Random Forest Regression Estimators weighting Sample tree Different tree in 'fancy' mode

1.7 6. Predictions continued.

Y-test against Y-pred plot My model against Las Vegas.

When Las Vegas sets betting odds they take into account non-basketball stats. This included public perception. e.g. If they know People love betting on the Lakers win total over (win more games) they can adjust their model predictions for that. So when their model might say 41 wins, they could list 43 wins. Theoretically you would have an advantage if you bet the under. They do not disclose adjustments.

Accuracy of Las Vegas win total predictions: (average difference between actual wins and predictions) 2016-17: 4.167 2017-18: 5.467 2018-19: 6.833 2019-20: pandemic shorted season, their predictions are based off 82 games. 2020-21: 5.217

The Random Forest predictions against real results

6.99952 (0.08536*82)

The 2019-20 Golden State Warriors top 2 salary players missed the season. This team was part of the test set and were the models worst prediction, and it should be. If you removed this team as those injuries were known about before the season began the model's accuracy would be:

6.08 (0.07419469*82)

This will be revisited in future adjustments.

1.8 7. Future Predictions

Now that a model has been created. What does it believe will happen with the current season?

1.9 8. Adjustments

This model only considers player salaries in an attempt to predict win totals. That was the original idea and I believe it accomplished that decently well.

Injuries and dead cap salary are not being accounted for. Dead cap is when a player's contract is bought out to save some money. This is typically players who are near the end of an expensive contract on a team that is not living up to expectations. This leaves the salary on their books while that player can go play for another team.

An example of this would be Blake Griffin on the Pistons. His ~35m were on the Detroit's books while they played for Brooklyn. My model over estimates Detroit in cases like this.

A potential way to correct for injuries would be to adjust their salaries for games played. So if a player making 10m plays 41/82 games the model would only take into consideration 5m.

Talented rookie contract players are not accounted for. Luka Doncic on the Dallas Mavericks earns around 10m because that is how rookie contracts are structured, however he is playing like a 35m player.

1.10 9. Credits

Thank you for reading through my first data project. When I developed this project proposal I was not at all aware the power of data science. I had no idea what machine learning was and manually acquired the data instead of scrapping.

Special thanks to my mentor Ricardo Alanis. They continued to give words of encouragement and advice throughout the development.

[ ]:

```
```