Carnegie Mellon University

NBA Project - Progress Report 3

Team: Andrew Liu, Willis Lu, Reed Peterson
Advisor: Brian MacDonald
Client: Kostas Pelechrinis

Introduction - Main Questions

Definition: Plus-Minus (+/-) is a statistic measuring point differential in the NBA.

1. Is there a way to more accurately measure an NBA player's performance?
    - One current method is box score +/-.
        - *Problem with this method is it is skewed by who the player is on the court with*

2. Can we use additional data such as contract value, team rating, and player history, to better calculate +/- for a player?
    - How do we balance these terms?

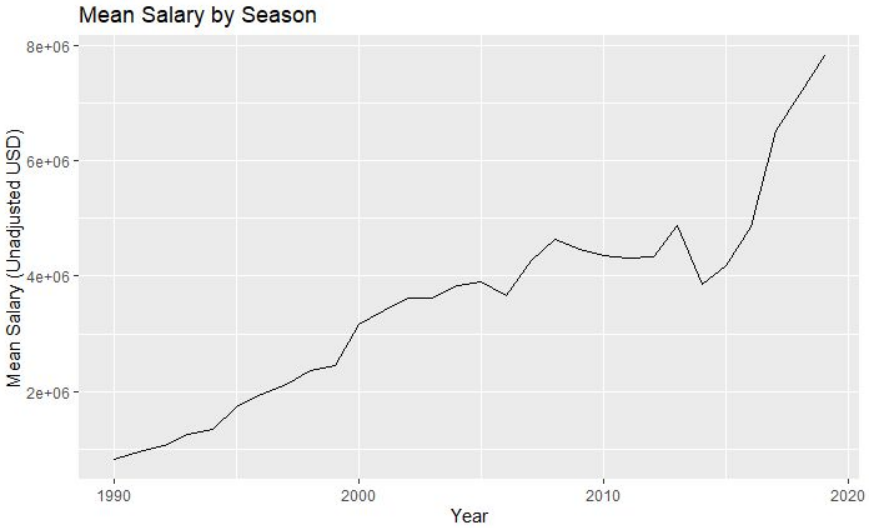
Introduction - Additional Questions

1. Taking contract value into account for the prior, how do those on rookie contracts fare in our model? How do we correct for this?
    - Players who outperform their rookie contract tend to be extremely underpaid (i.e. Luka Doncic)

2. How can we use previous seasons to predict player performance in future seasons?
    - Since first year players do not have prior data, how do we account for them?

Data - Contract Data

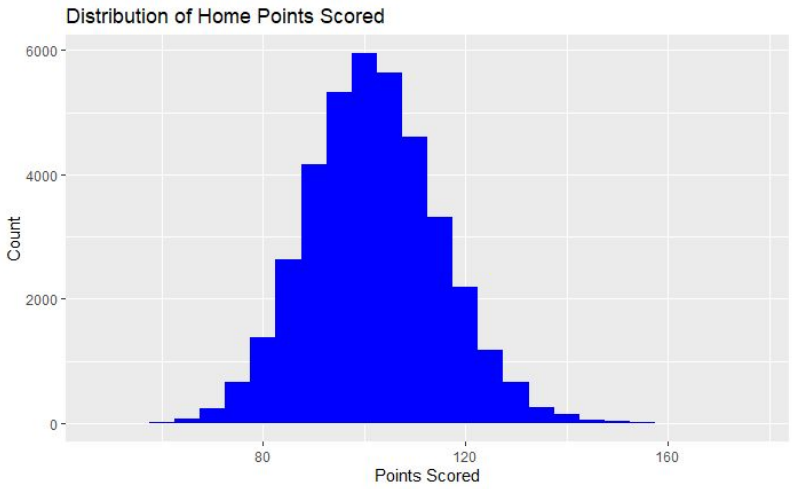
- 2018 and 2019 season data scraped using Python and BeautifulSoup Package. Original source: spotrak.com.
- 1990 - 2017 data found on Kaggle, and joined with 2018/2019 seasons
- In total data accounts for 1990-2019 seasons:
  - 12,724 total contracts (2406 unique players, 32 unique teams)
  - Variables: Name, Contract Value, Year, Team, and Type (Rookie/Non-rookie)

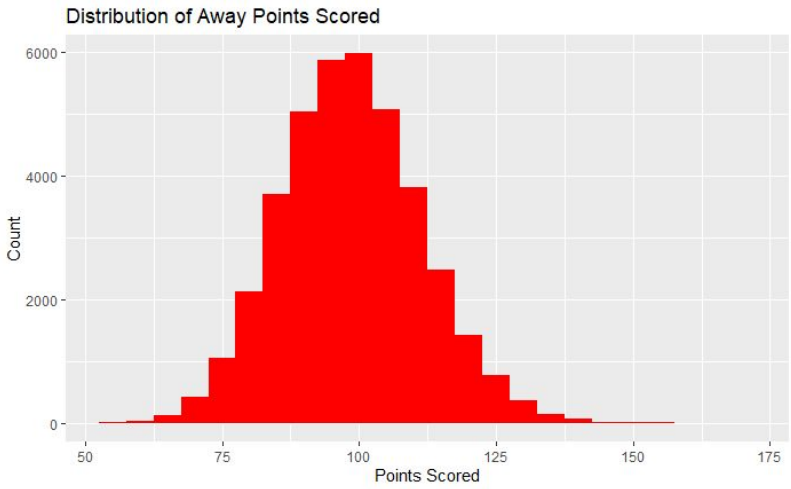


Carnegie Mellon University

Data - Games Data

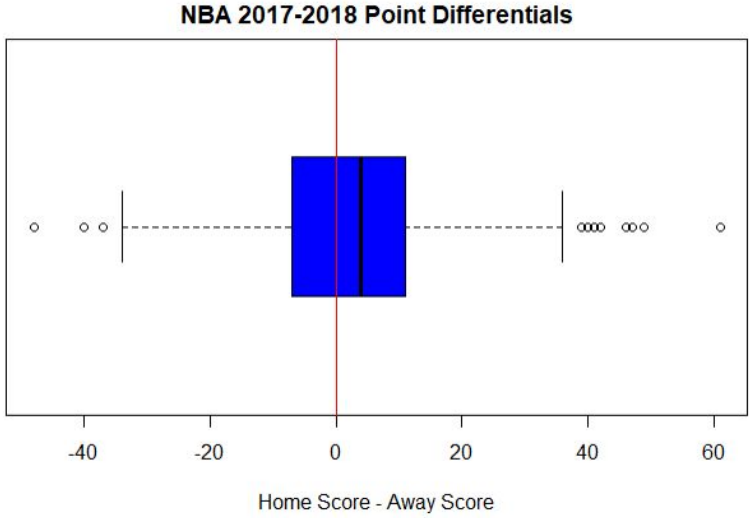
- Our games data comes from 538's study on NBA Elo rankings. (https://github.com/fivethirtyeight/data/tree/master/nba-forecasts)
- This dataset contains game by game elo ratings all the way back to the 1946 NBA Season.
  - The only variables we used are the game scores from 1990 to 2019.
- Used to create team ratings for our prior

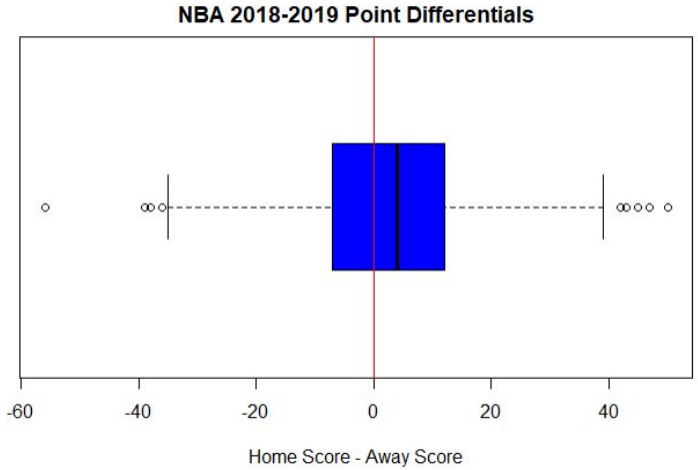




Data - Games Data

- Average home court advantage is worth 2.367 points in 2017
- Average home court advantage is worth 2.793 points in 2018
- Need to control for home court advantage in our dataset

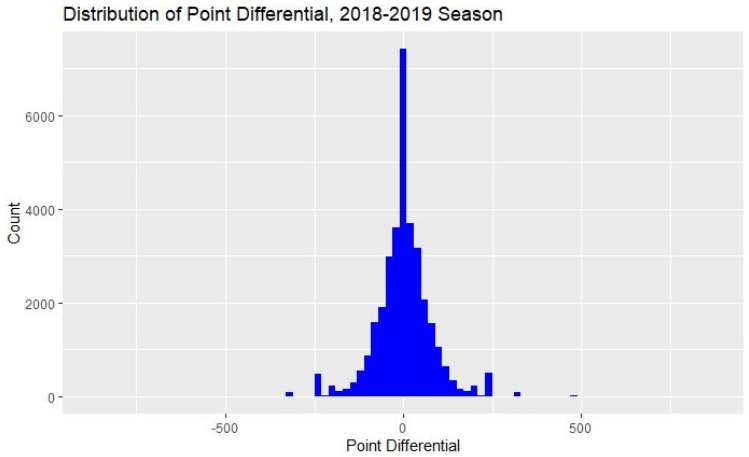




Carnegie Mellon University

Data - Shifts Data

- A “shift” is a period of time in an NBA game where the same 10 players are on the court with no substitutions
- We reformatted play-by-play data from eightthirtyfour as shift data to track the +/- of each shift (https://eightthirtyfour.com/data)
- Shifts are normalized by recording +/- per 100 possessions, where the number of possessions in each shift is calculated from this common formula: https://www.nbastuffer.com/analytics101/possession/
- Variables: Point Differential per 100 Possession, Home Team, Away Team, One-hot encoding of players on the court (1 for home, -1 for away)



Carnegie Mellon University

Methods - Overview

The following steps work together to help us answer our research questions. The balance between team rating and contract is found with Ridge Regression, and the rest of the questions are answered with Bayesian Regression.

- Simple Linear Regression
  - Used to acquire initial team ratings

- Ridge Regression
  - Used to obtain our final priors: balance between team rating and contracts

- Bayesian Regression
  - Using prior estimated with Ridge Regression, produce an estimate of +/- posterior
  - Also used to analyze rookie contracts separately

Methods - Linear Regression

We use simple linear regression to create team rating for priors. We regress point differential on two variables (team and location).

Applications:

- to be used in our Bayesian regression priors
- can tell us how good teams are in the regular season
- will allow us to adjust player ratings in accordance to their team ratings.
- produces standard errors for each player's prior

Methods - Ridge Regression

We use ridge regression to estimate the prior means for each player. To achieve this, we utilize a nested Ridge Regression.

1. Take the player coefficients produced by an initial Ridge Regression performed on a sparse matrix of point differentials for an nba season.
2. Use Ridge regression to try to predict the player coefficients from step 1 using prior season data. Idea is to use previous season's contract and team rating prior to predict next year performance.
3. Predict using the ridge model to get a new coefficient for each player. These coefficients will now serve as prior means for another Bayesian model.
4. Since standard errors aren't very applicable to regularized models, we run the same model with penalty parameter 0 to achieve basic linear regression. This allows us to produce standard errors.

In this manner, we have achieved a new mean and standard deviation prior for each player that we can now use in another Bayesian regression.

Carnegie Mellon University

Methods - Bayesian Regression

- Regressing point differential on the players present on the court (one hot encoding of all players in the league), with a prior distribution for each player.
  - Priors obtained through Ridge Regression

- Output is an estimated distribution of a player's +/-, adjusted for team ratings and contract value.

- Ran a model for all players as well as a model for just players on rookie contracts

Carnegie Mellon University

Results - New priors from Nested Ridge Regression

This is how we incorporated previous seasons data (team ratings and contract data) and balanced the two terms.

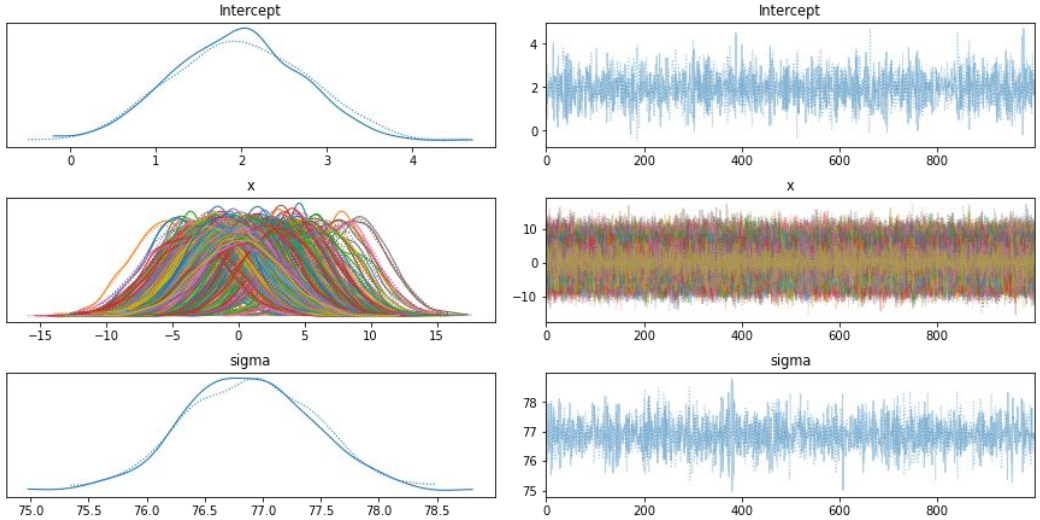
rating	mu	name	coefs	finalpriors	finalse
-12.68894695	11.56085	Stephen Curry	10.2252196	4.863816469	0.761081404
-6.237947831	11.09523633	LeBron James	7.284820557	4.621640899	0.677228176
-6.899803629	10.423077	Paul Millsap	6.290998011	4.272983299	0.627803243
-8.267971906	9.9093	Gordon Hayward	6.974004242	4.006562474	0.595820884
-5.484227899	9.837633333	Blake Griffin	4.529567932	3.969117137	0.58736324
-11.46367947	9.567901333	Kyle Lowry	2.103909634	3.829746062	0.61139066
-8.526489449	9.510202667	Russell Westbrook	4.330412351	3.799532845	0.569743397
0.49881092	9.510202667	Mike Conley	8.643611381	3.798676183	0.649069604
-13.21905832	9.433133	James Harden	8.733328601	3.759994033	0.638908518
-11.46367947	9.246658333	DeMar DeRozan	-0.419637724	3.663083204	0.591246181

Carnegie Mellon University

Results - Bayesian Regression (All Players)

Using priors from our nested Ridge Regression, we were able to rank the players from the 2018-2019 season. This is our improved box-plus minus metric.

Our top 5 players were: Steph Curry, Paul Millsap, James Harden, Damian Lillard, and LeBron James.



Carnegie Mellon University

Results - Bayesian Regression (Rookie Contracts Only)

- Similar to the model with all players, but only run on players on rookie contracts.
- Our top 5 players on rookie contracts were: Markelle Fultz, Donte Divincenzo, Donovan Mitchell, Karl Anthony-Towns, and Luka Doncic.
- High draft picks and team strength significantly impact these ratings due to small standard errors - an issue we will look to address in the future
- Note - results from this model are slightly out of context since removing veterans from the dataset allows rookies to take greater credit/blame for their team's success

Carnegie Mellon University

Results - Other Expected Results

- We currently have promising results for 1 season (2018-2019) that are based on the prior season.
  - We expected to have more results pertaining to other seasons utilizing similar steps
- One other issue is in how we address the rookie situation:
  - We are currently discussing ways to better address the fact that team strength appears to be more important that player performance.

Carnegie Mellon University

Discussion - Impression on Results

- Current results appear promising: star players are near the top, valuable role players fill out the above average portion.
  - Our current model seems to correct for players who are consistently playing with really good teammates
  - One inconsistency is that those on rookie contracts are consistently undervalued by the model
- Rookie problem - rookies on strong teams are overvalued
  - Witnessed by Donte Divincenzo (Milwaukee Bucks) and Markelle Fultz (Philadelphia 76ers) being top 2 rookies.
  - In the process of addressing this issue.
- We still need to address the question of how previous seasons can be used in the prior more rigorously.

Discussion - Next Steps

- We have mostly addressed our main questions at this point. We are constantly brainstorming possible improvements, but what we have is promising.

- To address how we use previous seasons to predict player performance in future seasons, we would like to build models in a similar fashion for seasons that we have data for.

- Our question about how to account for rookies is still in progress - we have discussed accounting for draft position/ minutes played in some way.

Carnegie Mellon University

Discussion - If Time Allows...

- Does a player changing teams change their ratings?
- Can we include coaching in the player rating?
- Does resting players increase the player's performance?
- Predict team offensive/defensive ratings at end of season. Compare this to weighted average player rating based on minutes played to overall offensive/defensive ratings.

Carnegie Mellon University

Thank You!

Q & A

Carnegie Mellon University

Extra Results - Team Ratings

We utilize a simple linear regression that uses game data and tries to predict point differential given the variables team and home court advantage. This creates a coefficient for each team that we use as player ratings.

- Dimensions: 30 observations and 2 variables (team, rating).

	team	rating
17	MIL	16.3807222
14	LAL	14.6040696
13	LAC	14.5170573
2	BOS	14.2242103
28	TOR	13.8739727
7	DAL	12.0453683
16	MIA	11.5500537
11	HOU	11.0840935
29	UTA	10.6306580
8	DEN	10.2444927
23	PHI	9.6917090
21	OKC	9.5921958

Carnegie Mellon University