Carnegie Mellon University

NBA Project - Development of Bayesian Contract Plus-Minus (BCPM)

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 +/- (BPM)
        - *Definition - BPM uses a player’s box score information, position, and the team’s overall performance to estimate the player’s contribution*
        - *Problem - does not consider other players on the court*
    - Another method - Real Plus Minus (RPM)
        - *Definition - A statistical measure of a player’s performance calculated from net point differential per 100 offensive and defensive possessions.*
        - *Works well (accounts for teammates, opponents, box stats priors) but no contract data, measures of uncertainty not made public.*

2. Can we use additional data such as contract value, team rating, player history, etc. to better calculate +/- for a player while simultaneously obtaining a measure of uncertainty?

Carnegie Mellon University

Introduction - Additional Questions

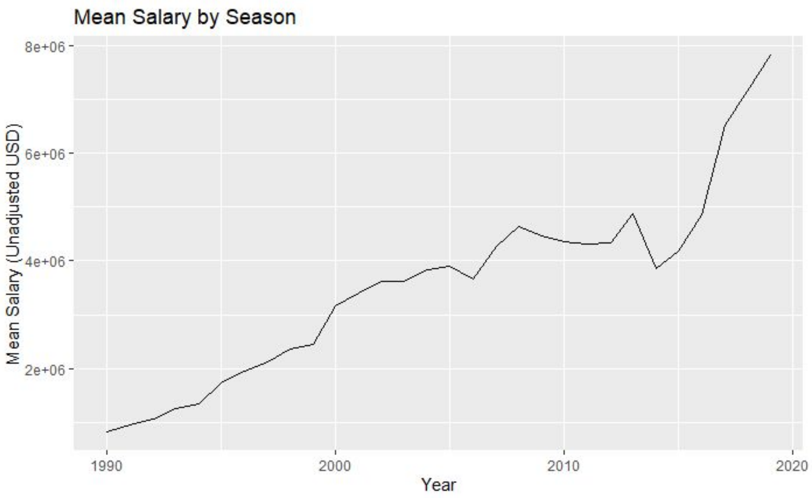
1. Taking contract value into account for the prior, how do those on rookie contracts fare in our model? How do we correct for this?
   - Players who outperform their rookie contract tend to be extremely underpaid (i.e. Luka Doncic)

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?

Carnegie Mellon University

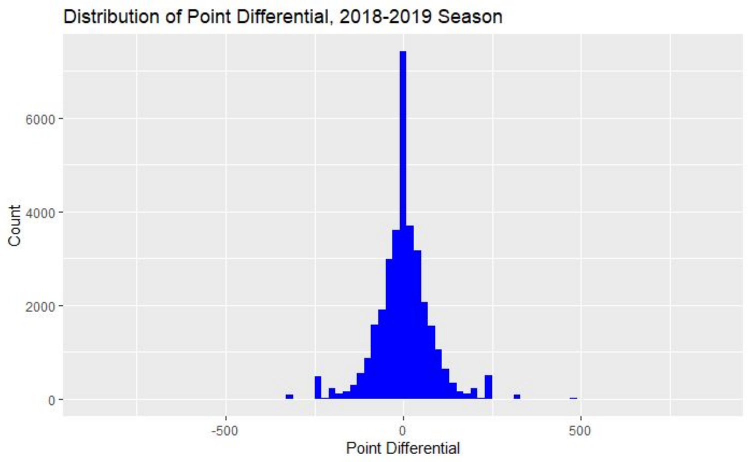
Data - Contract Data

- 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)



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 eighthirtyfour 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:

- Ridge Regression, Model Selection, Random Forest Regression
  - Used to obtain our final priors based on contract values for players
  - Separate models for rookies and non-rookies due to large discrepancies in contract values
- Bayesian Regression
  - Produces an estimate of +/- posterior distribution for each player

Methods - Ridge Regression, Random Forest Regression for Priors

Ridge Regression → Random Forest Regression → Final Priors

Ridge Regression

- Run ridge regression on past data to obtain coefficient estimates for each player

Random Forest Regression

- Train random forest regressor to map contract value to coefficient estimates
- **Separate models for rookies and non-rookies**

Final Priors

- Use the random forest model to calculate final prior means and standard deviations for a new season

Carnegie Mellon University

Methods - Prior Model Selection

- Before training the random forest regressor, we tested and validated a number of different models to identify an optimal model for prior distributions
  - Prior models were built/validated on 5 seasons of past data (i.e. priors for the 2015/16 season were developed using data from 2010/11 up to 2014/15 seasons)
    - *For model selection - each candidate model was trained on the first 4 years of data and validated on the 5th year of data*
    - *For computing final priors - once the optimal model was selected, all 5 years of data were used to compute final priors*
    - *Prior means are output by the random forest model*
    - *Prior standard deviations for rookies/non-rookies are set to the RMSE of the respective model*

Methods - Model Selection Cont'd

- Four main candidate models (**note - all models include contract value as a predictor**):
  - Random Forest Regressor (including team rating as a predictor)
  - Random Forest Regressor (excluding team rating as a predictor)
  - Gradient Boosting Regressor (including team rating as a predictor)
  - Gradient Boosting Regressor (excluding team rating as a predictor)
- Based on MSE and manual inspection of results, random forest regressor excluding team rating was the optimal model for deriving priors

Methods - Bayesian Regression

- Bayesian Regression model of the following form:

$$y = \mu + X\beta + \varepsilon$$

Which becomes for each shift:

$$y_i = \mu + \beta_{H1} + \ \dots\ + \beta_{H5} - \beta_{A1} - \ \dots\ - \beta_{A5} + \varepsilon$$

Where y_i is the point differential for the i^{th} shift

μ is a constant corresponding to home court advantage

β is a vector of coefficients for all players

β_{Hj} is the coefficient for the j^{th} player on the home team

β_{Aj} is the coefficient for the j^{th} player on the away team

X is our sparse design matrix (shifts data)

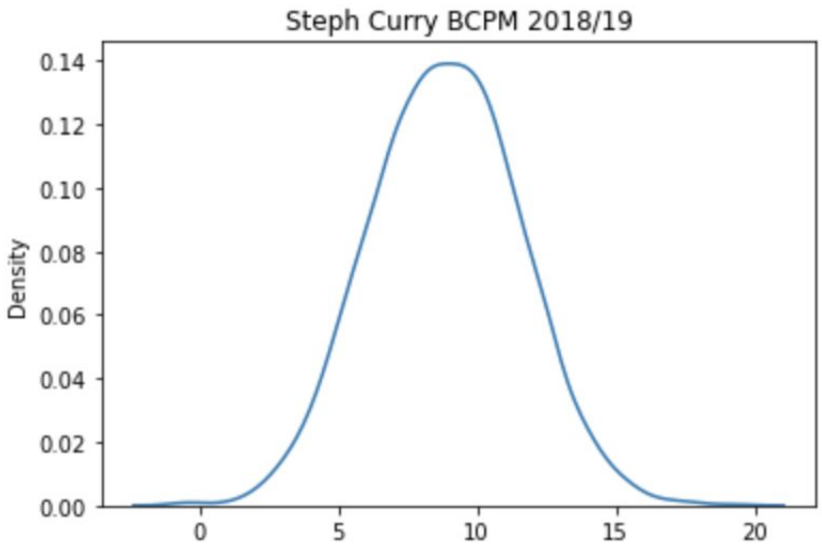
ε is random error

- Learns distributions for each $\hat{\beta}_j$ which is the BCPM for the j^{th} player (each distribution assumed to be normal)

Carnegie Mellon University

Results - Bayesian Regression

- We ran the Bayesian regression model on 4 seasons of data individually (2015/16 up to 2018/19)
- We can inspect the resulting distributions of player BCPM estimates
  - We call this metric Bayesian Contract Plus Minus (BCPM)
- Top 10 players in 2018/19:
  - Jrue Holiday
  - Steph Curry
  - James Harden
  - Paul George
  - Damian Lillard
  - Giannis Antetokounmpo
  - Al Horford
  - Gordon Hayward
  - LeBron James
  - Mike Conley



Carnegie Mellon University

Results - Interactive App (Player Distributions)

NBA Player Distributions According to BCPM

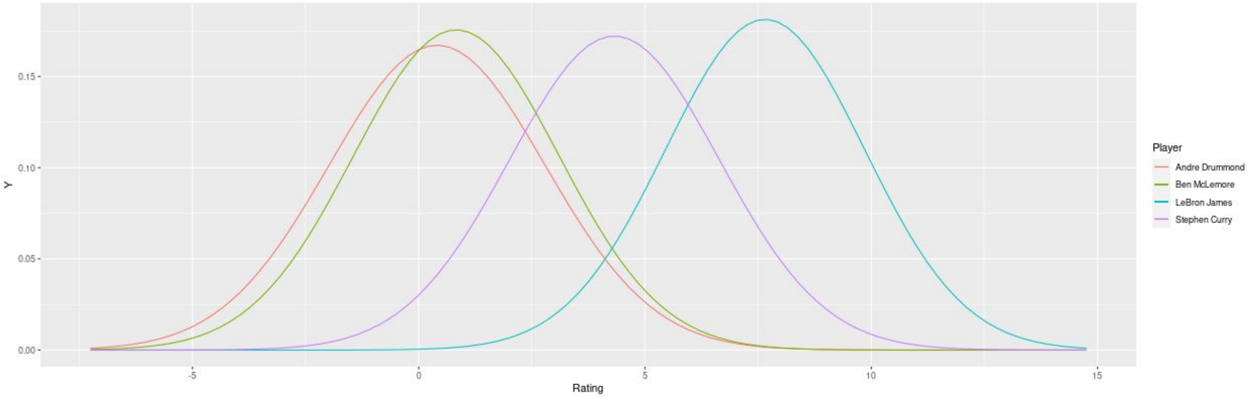
User Selects Season and Players. Player distribution, assumed to be normal, is displayed. Mean and Standard Deviation from our BCPM model.

Select Season

2015-2016

Select Player

LeBron James Ben McLemore Stephen Curry Andre Drummond



	Name	Team	Rating	SD
19	Andre Drummond	Detroit Pistons	0.406	2.387
39	Ben McLemore	Sacramento Kings	0.830	2.272
284	LeBron James	Cleveland Cavaliers	7.654	2.200
409	Stephen Curry	Golden State Warriors	4.325	2.317

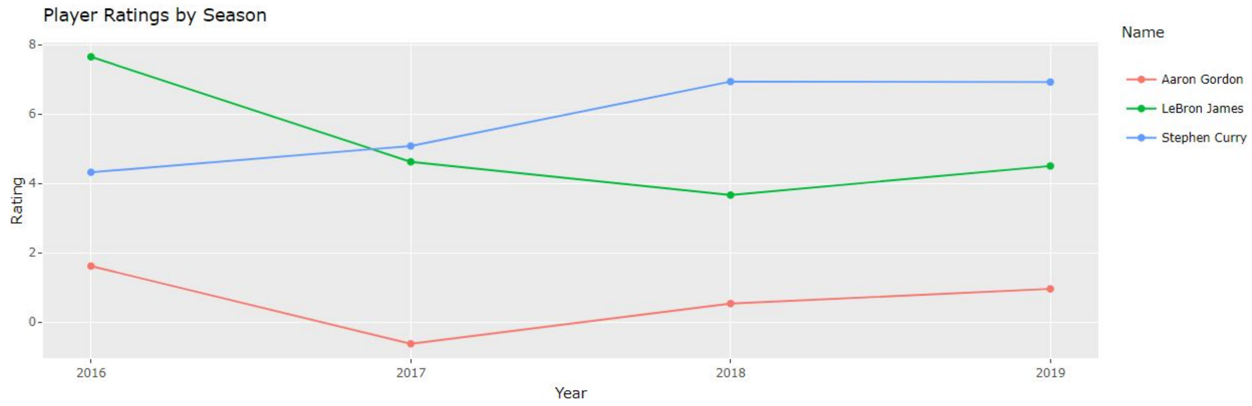
Results - Interactive App (Time Series)

Mean BCPM Rating by Season

User selects Players. Player Ratings from our BCPM model are displayed as a time series across 4 different seasons.

Select Player(s)

LeBron James Stephen Curry Aaron Gordon



Results - Interactive App (Rating vs Prior)

BCPM Player Rating by Contract Prior

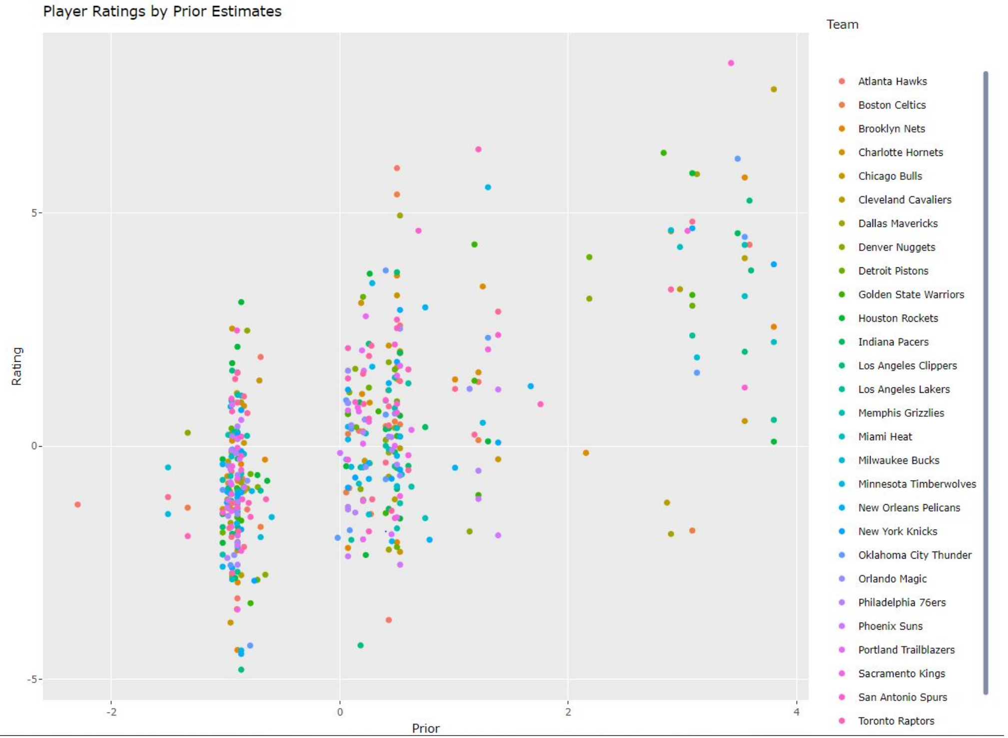
User selects Season and Team. Displays a scatterplot of our final BCPM Rating against Contract prior used to train the model.

Select Season

2015-2016

Select Team

All Teams



Results - Interactive App (Probability Matrix)

NBA Player Comparisons

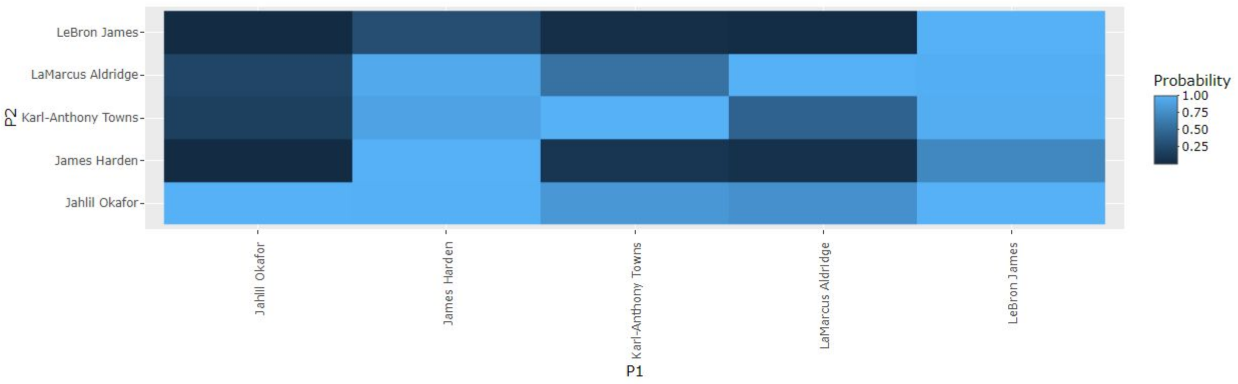
User selects Season and Players. Displays the probability that Player 1 (P1) is better than Player 2 (P2). Probability obtained by comparing 2000 samples from relevant distributions given by BCPM model.

Select Season

2015-2016

Select Player

LeBron James · James Harden · Jahlil Okafor · Karl-Anthony Towns · LaMarcus Aldridge



Live Demo

https://coly1119.shinyapps.io/NBA_Project/

Carnegie Mellon University

Discussion

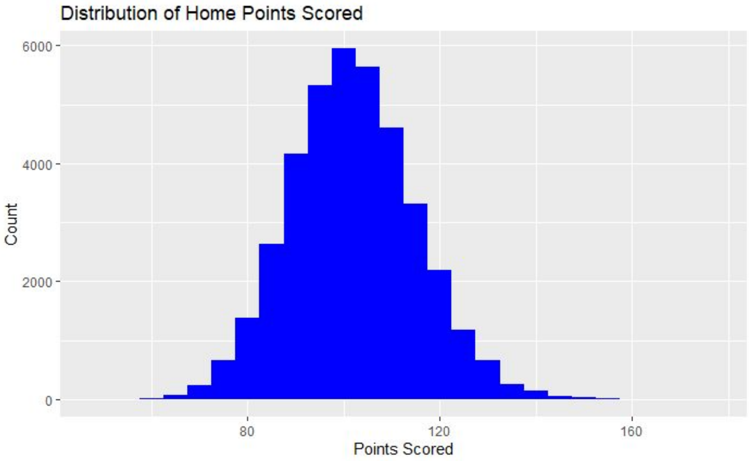
- Current results appear promising: star players are near the top, valuable role players fill out the above average portion
  - Our current model seems to correct for players who are consistently playing with really good teammates
- Rookies still appear to be undervalued despite separate prior models
  - Inspecting the prior values shows that rookie priors are noticeably lower than veteran priors despite separate models
    - *Something to address in future work*

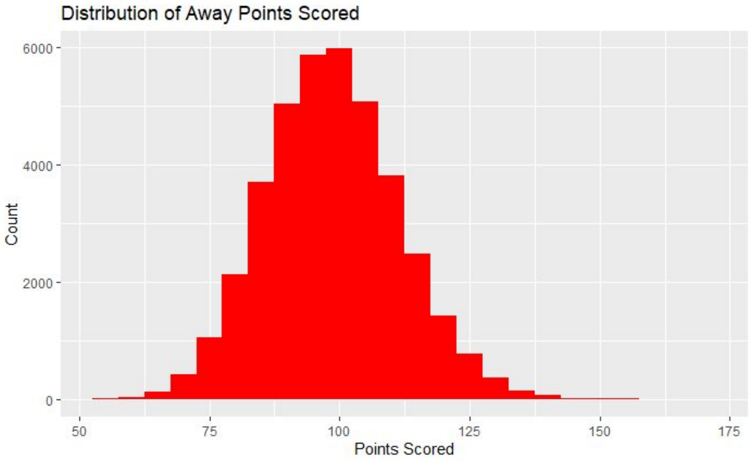
Thank You!

Q & A

Extra - 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.

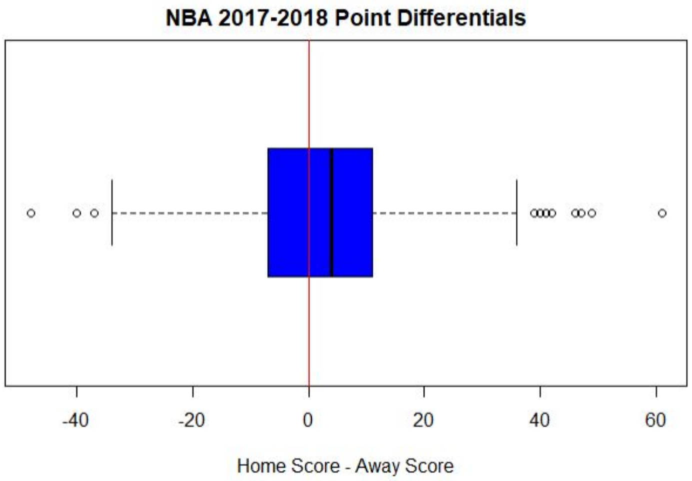


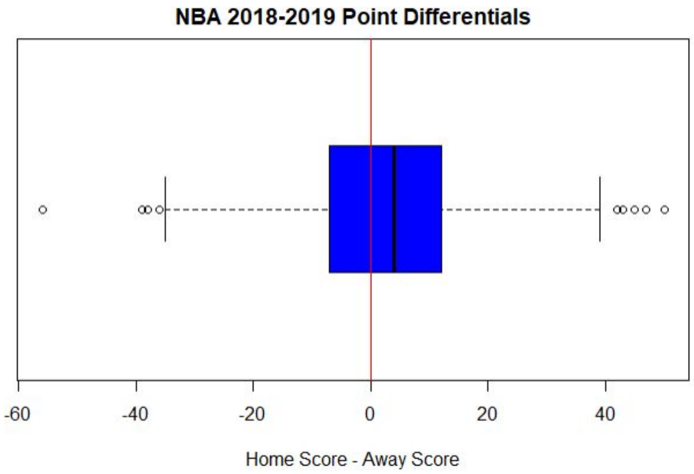


Carnegie Mellon University

Extra - 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

Extra - Linear Regression

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

Applications:

- Potential 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.

Carnegie Mellon University