High Anticipation

Exploring Trends Between Public Perception and Player Value

Alana Willis, Fiona Dunn, and Sahana Rayan
October 25, 2020
~ What is the best metric for defining player performance?

~ How do we quantify public perception?

~ Do any players or subset of players stand out?

~ Can we predict public perception using player performance and vice versa?
Area of Focus

- Narrowed NBA players to 2018 first round draft class
  - All data had to be collected separately
  - Includes a randomized mix of players
    - International
    - Highly anticipated
    - Underrated
Constructing the Data Set

01 Reddit
Scraped Reddit user comments using RedditExtractoR

02 Web Hits
Scraped views from YouTube, Wikipedia, and Google (Images and News)

03 Game Data
Scraped Box score data from NBA stats page during the 2018 - 2019 season

04 Combined
Rows aligned by weeks a player participated in at least one game over rookie season.
General EDA
Interesting Player Trends

Anfernee Simons’ Season Game Score Time Series

Anfernee Simons’ Sentiment Time Series
Clustering

- Hierarchical Clustering was done using positive score, negative score, Wiki views, Google web hits, Google news hits, and Youtube hits
- Cluster 1: Not very popular, low sentiment
- Cluster 2: A bit more popular, high sentiment
More EDA with the clusters
XGBoost: Predicting Clusters

- These clusters were used as a response variable for a binary classification problem.
- XGBoost model was trained using AUC as an evaluation metric and the test predictions has an AUC of 0.887
Partial Dependence Plots

Partial Dependence plot with Average Minutes

Partial Dependence plot with Average Points
Final Thoughts

- Data wrangling and cleaning is HARD!
- There is a relationship between a player’s performance, public sentiment, and overall popularity.
- One metric needs to be used to define public perception: Social Score and Clusters
- Average minutes played and average points scored are the most important in predicting popularity and sentiment clusters
Continuing the Research

- Introduce more variables to the models.
  - Reddit: comment score and controversiality
- Expand to more draft classes and to more social media/chat rooms
  - Other subreddits, blogs, Facebook, etc
- Explore specific comments in the Reddit data
- Predicting public perception onto future draft classes based on college or international play
- Principal Component Analysis with sentiment and popularity data
Contacts

Alana Willis
- Winston-Salem State University
- awillis117@rams.wssu.edu
- linkedin.com/in/alana-willis-7a8014188/

Fiona Dunn
- Kenyon College
- dunn2@kenyon.edu
- linkedin.com/in/fiona-dunn-706035138/

Sahana Rayan
- Purdue University
- srayan@purdue.edu
- linkedin.com/in/sahana-rayan/
References


Thank You!

- Carnegie Mellon University
  - Rebecca Nugent
  - Ron Yurko
  - Beomjo Park
  - Pratik Patil
  - 2020 Cohort

- Atlanta Hawks
  - Maksim Horowitz
Variables of Interest

1. **Week**: Mon-Sun of a week in the regular season of the NBA

2. **Game Score**: a rough measure of a player’s productivity for a single game

3. **Popularity** = \( \frac{\text{wiki_views} + \text{avg_web_hits} + \text{avg_image_hits} + \text{avg_news_hits} + \text{avg_yt_hits}}{5} \)

4. **Social Score** = \( \frac{\text{sentiment\_week} + 5 \times \text{Popularity}}{6} \)
General EDA

Rookie Season Sentiments

Average Sentiment Score vs. Number Round Pick

Rookie Season Popularity

Popularity Score vs. Number Round Pick
Interesting Player Trends

Trae Young's Average Game Score
Rookie Season

Trae Young's Popularity
Rookie Season
General EDA

Sentiment Score by 1-15 Picks

- Collin Sexton
- Deandre Ayton
- Jaren Jackson Jr.
- Jerome Robinson
- Luka Doncic
- Marvin Bagley III
- Michael Porter
- Mikal Bridges
- Miles Bridges
- Mo Bamba
- Shai Gilgeous-Alexander
- Trae Young
- Troy Brown
- Wendell Carter Jr.
General EDA

Sentiment Score by 16-30 Picks

- Aaron Holiday
- Anfernee Simons
- Chandler Hutchison
- Donte DiVincenzo
- Dzanan Musa
- Grayson Allen
- Jacob Evans
- Josh Okogie
- Kevin Huerter
- Landry Shamet
- Lonnie Walker
- Moritz Wagner
- Omari Spellman
- Robert Williams III
- Zhaire Smith

Weeks
General EDA

Average Game Score vs Social Score

Luka Doncic

Anfernee Simons
Interesting Player Trends

Social Score by Player

Average Game Score by Player

Player Name
Clustering for Popularity Metrics

Scatter plot of Average Web hits and Wiki views

Box Plot of Average game score for the 4 clusters
Before Modeling
Decision Tree: Predicting Avg Game Score
Decision Tree: Predicting Social Score

Yes: avg\_ast < 5.3
- 355 (100%)
- 1146 (10%)
  - avg\_dreb < 3.9
    - 265 (90%)
    - 409 (5%)
    - 1817 (5%)

No
- 265 (90%)
- 409 (5%)
- 1817 (5%)

Graph showing importance of variables: avg\_ast, avg\_dreb, avg\_pts, avg\_oreb.
In The Beginning

Initial Research Questions

How do players interact with their fan bases conditioned on the demographic of their home market?

Can we quantify "home town bias" among local media outlets for each NBA team?

Comparing public perception vs. actual player value (using player stats and social media data)

Packages

TidyCensus

NBAStatR

TwitteR
Stage 1: Reddit Data

- Scraped all Reddit posts within subreddit r/nba using RedditExtractor
  - Filtered the posts and comments to match the dates of the rookie season (~October 2018 to April 2019)
- Used bing sentiment analysis on each of the comments to create a sentiment score
  - Resulted in over 2 million rows
  - Grouped comments about each player by week to reduce rows to 296
Stage 4: Final Dataset

- The game data, Reddit data, and web data were combined to form a data set where each row’s unique identifiers were the player name and the week
  - 206 observations by 25 variables
- This data set comprises of 3 types of data; Sentiment data (Reddit), popularity data (Wikipedia and Google trends), and basketball performance data (NBA stat)
Stage 2: Web Data

- Used `wikipediatrend` package to scrape daily Wikipedia page views for each player
  - Filtered by the dates of the rookie season (~October 2018 to April 2019)
- Used `gtrends` package to scrape daily Google image hits, web hits, news hits and YouTube hits for each player
  - Filtered by the dates of the rookie season (~October 2018 to April 2019)
  - Google trends assigns scores from 0 to 100
Stage 3: Game Data

● Scraped box score data from the NBA stats page
  ○ Only looking at 2018-2019 season for players drafted in top 30
● Each row in this data showed the player’s performance for a game played in the season
● Game score: a measure of player performance in a given game
Random Forest: Predicting Game Score

- $N_{trees} = 40 - 80$ with increments of 5
- $M_{try} = 1$ to 6
- $Min_{node\_size} = 3, 5$
XGBoost: Predicting Avg Game Score

Best Tuning Parameters:
nrounds = 120, eta = 0.025, max tree depth = 1
Clustering for Sentiment Metrics

Scatterplot of Sentiment and comment score

Box plot of Average game score for the 4 clusters