NBA Analytics: Part 1 How Shots Miss in the NBA

By Jason Leung

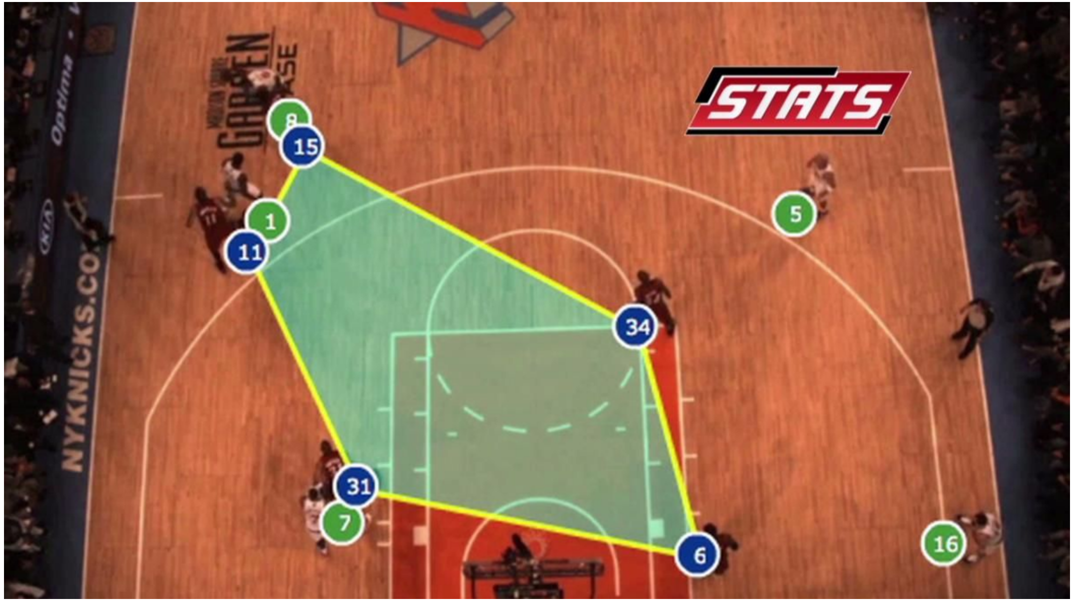
Data Science Immersive Capstone Project

Approaching NBA Analytics: Deep and Specific Instead of Wide

- I chose the NBA because it is a game with seemingly tractable well-structured data science problems.
- My goal: Focus on a problem that had not strictly been covered 100% before and add something a bit new to the sub-field
- Thus, use optical tracking data to engineer a hopefully new statistic. Shooting efficiency, scoring is most focused on so instead I chose rebounds/misses. ***What happens when a shot doesn't go in?***
- Finally, stay descriptive and not prescriptive. I have no experience coaching and playing and don't know cause and effect, what changing what action means inside the game.

Feature Engineering:
Existing SportVu Features

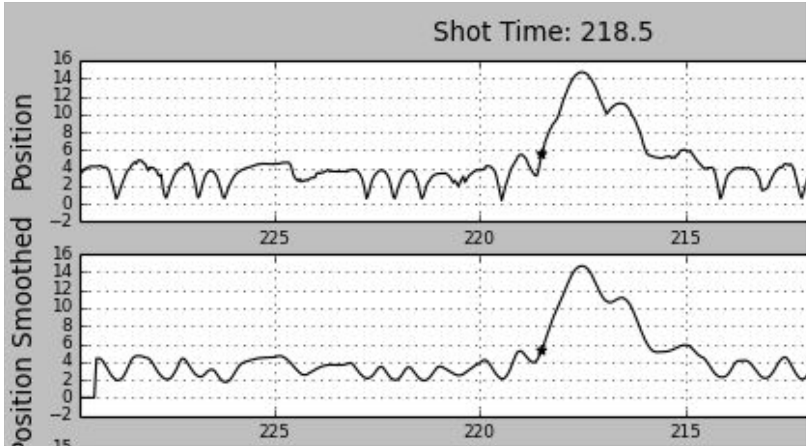
- SportVu data were at one point publicly accessible for the 2013 to 2016 seasons
- With only half a season, longitudinal analysis is not feasible, players average only 100 misses total over this time
- More popular stats based on video: during shot, nearest defender number of feet away
- Player max acceleration, speed comparisons
- Convex Hull / Voronoi Tessellation, dividing court into where each player is closest to that rebound



Tracking Stages from Shot to Rebound

- Shot release time from closest observation to shot location given by NBA API
- Shot peak time from max height after shot release
- Time hits rim when height nearest to 10 feet after peak time
- Rebounding peak time after ball hits rim to get rebound height
- Rebound time is <u>either</u> when ball drops to 8 feet (2 ft below rim) or when rebounder interrupts flight of ball


Ezra Shaw/Getty Images

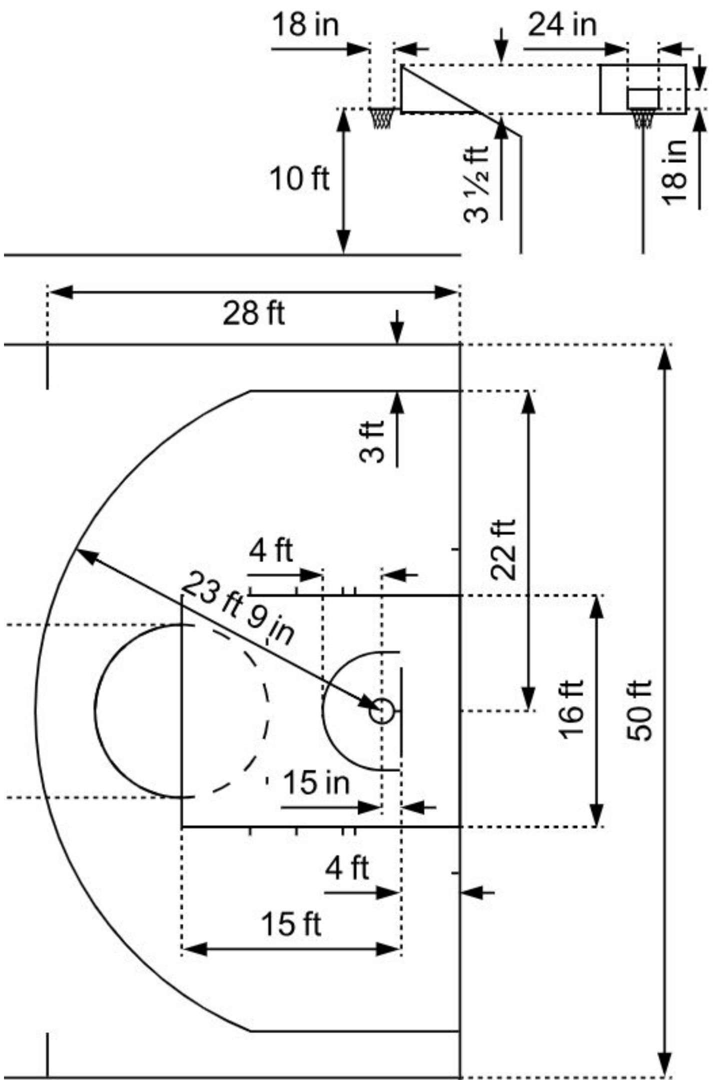


Feature Engineering: Realistic Useful Variables

- Shot Time
- Shot Angle oriented to basket
- Rebound Time
    - Timing not accurate beyond 1/10 seconds due to noise slotting each shot stage
- Rebound Angle
- Rebound Distance from basket
- Rebound Height
- No rebounding after initial ball’s arc of rim, extremely messy when ball hits multiple players’ hands
- No player movement during shot, during rebound opportunity

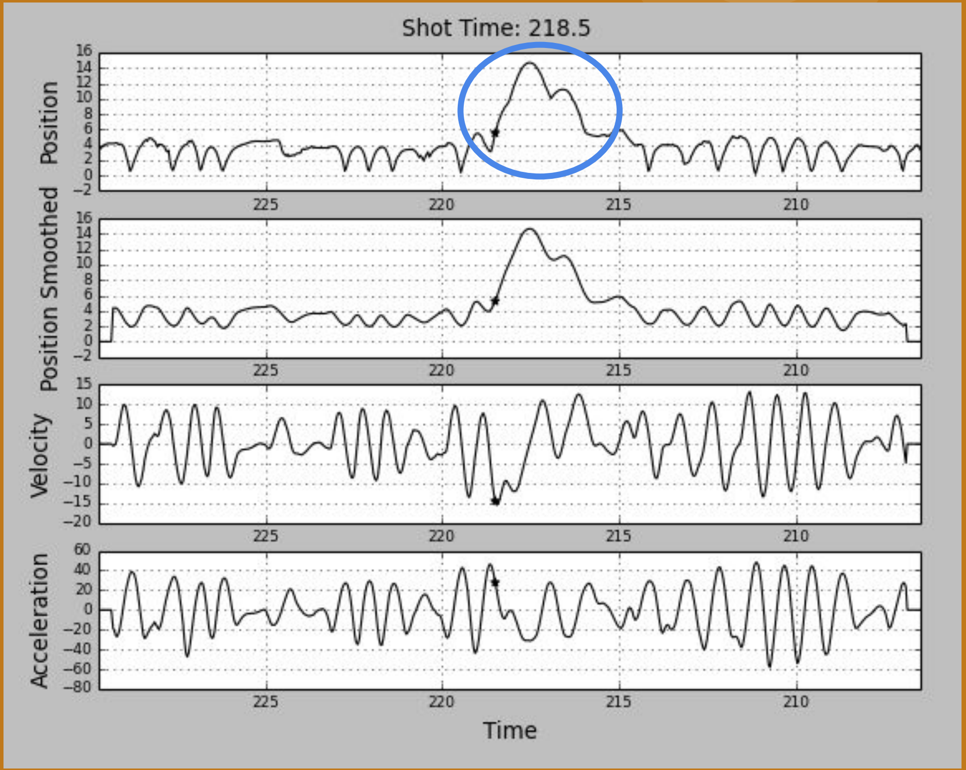
Scoping out SportVu Raw Data vs Court Dimensions

- SportVu produces one huge json file for a game
    - X and Y coordinates for all 10 players on court
    - X, Y and Z (height) for basketball
- Point in space model, from spaced out camera feeds SportVu infers centroid location for the 11 entities, 25 times every second
- Familiarize with important NBA dimensions on diagram on right
    - 10 ft basket, 4 ft behind hoop, 24 ft 3 point line



To the Rescue, NBA Movement Project on Github

- Example code extracting and processing raw movement data into usable form
- Identifies shot time just by position and acceleration curves (star in diagram)
- Either pass or shot when ball height rises above ~8 feet
- Characteristic shot curve, rebound curve, dribbling

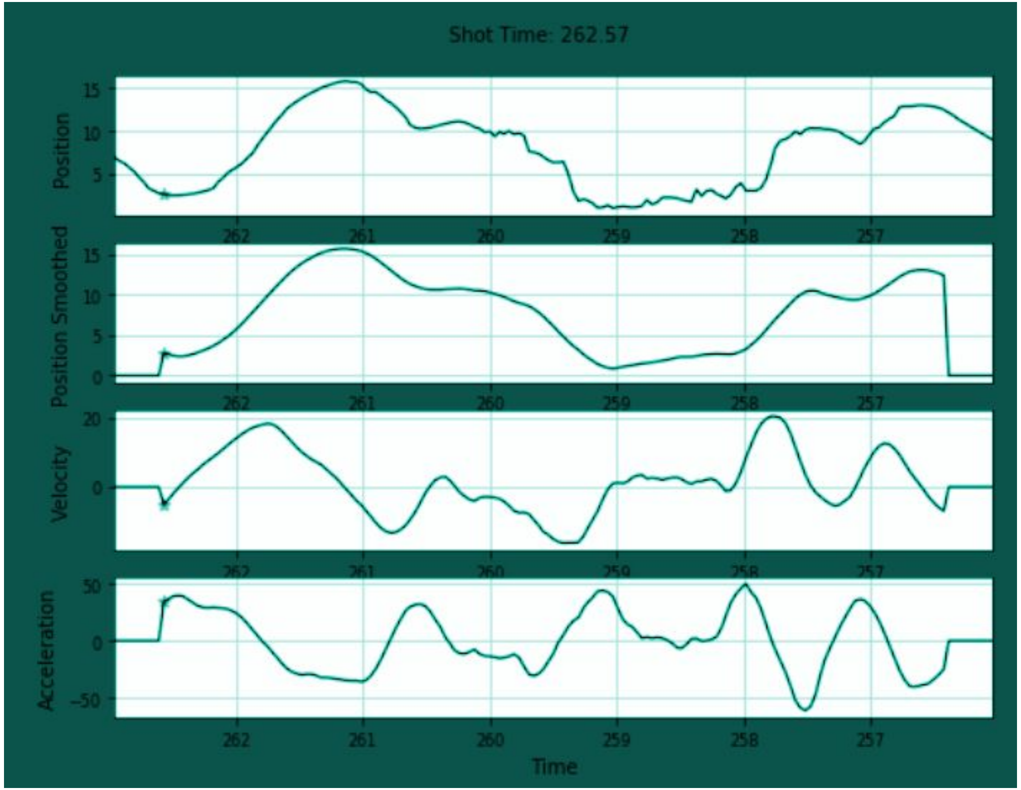


Primary Summary Statistics

shot range	**area**	**reb_rho**	**reb_angle**
24+ ft.	Center(C)	6.19	179.53
24+ ft.	(LC)	**6.72**	176.34
24+ ft.	Left Side(L)	5.95	**174.07**
24+ ft.	Right Side(R)	5.83	188.97
16-24 ft.	Left Side(L)	**6.29**	176.05
16-24 ft.	Right Side(R)	**5.46**	183.15
8-16 ft.	Right Side(R)	**5.16**	178.89
< 8 ft.	Center(C)	5.56	179.87

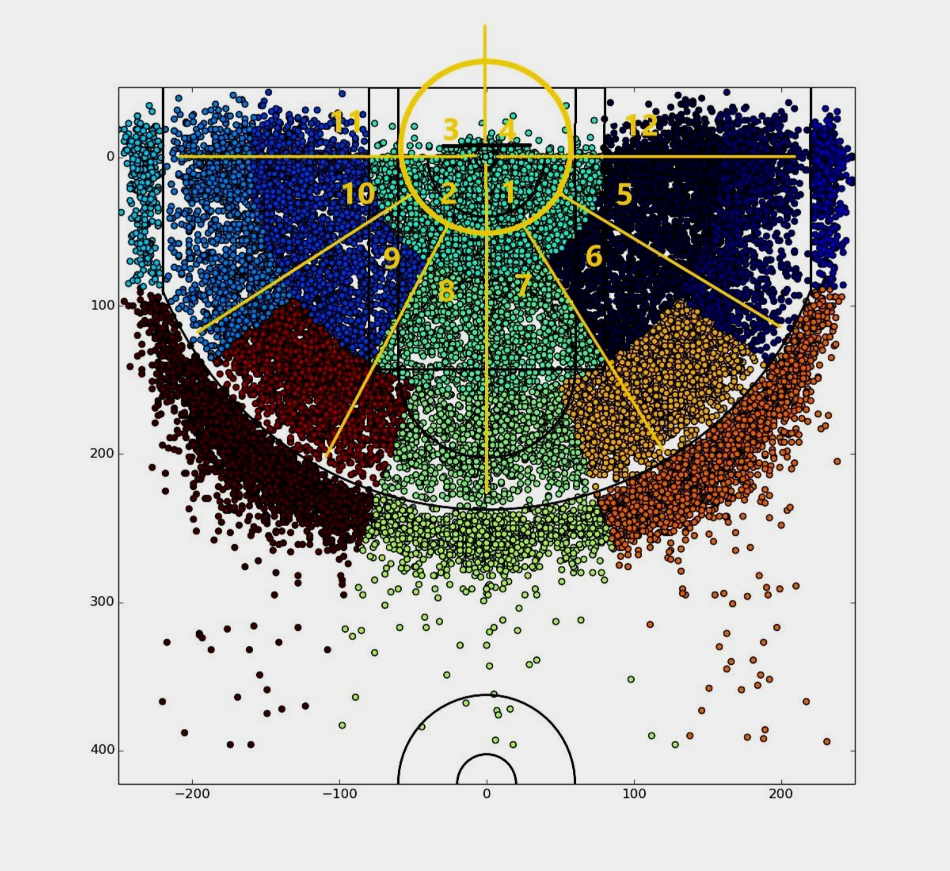
Engineering Train Dataset

- Example of contested rebound
- But no identifiable shooting motion from ball position data
- Movement data is messy! So NBA labeling of plays was used to anchor search of movement data wherever possible
- What about Y target variable?



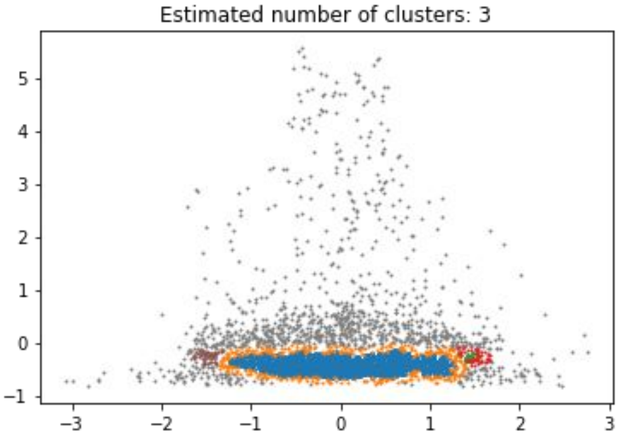
Y Target Rebound Location: Categorical Engineering

- 12 quadrants
  - Close rebounds quarter circles within 3.5 feet of basket
  - Long rebounds behind backboard or 30 degree slices along 180 degree span in front of backboard
- Strong relationship between rebound angle and shot angle can be used to predict quadrants
- Mediocre prediction scores, for 9 adjusted quadrants instead of 12 as well

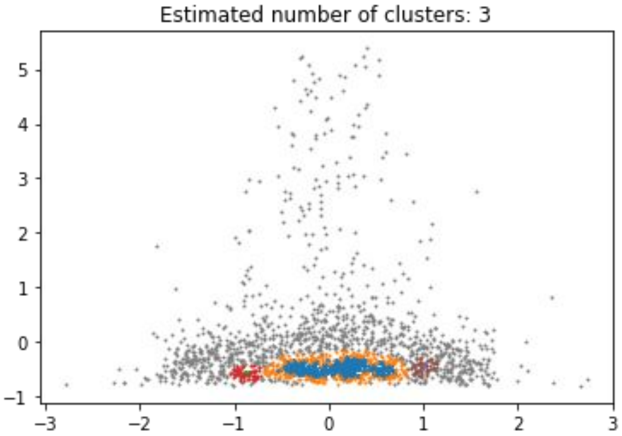


Unsupervised learning with DBSCAN

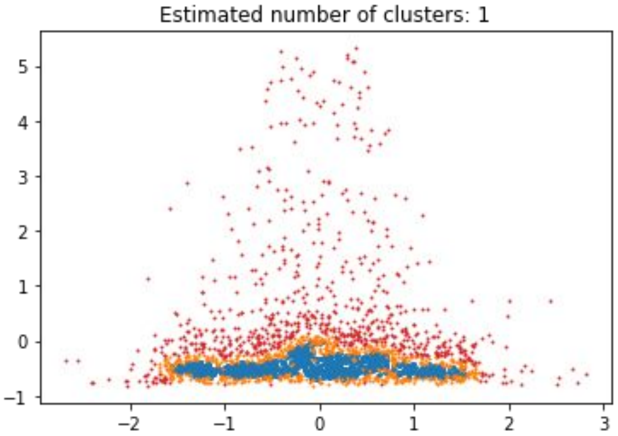
(Density-based spatial clustering of applications with noise)



Center(C) 24+ ft

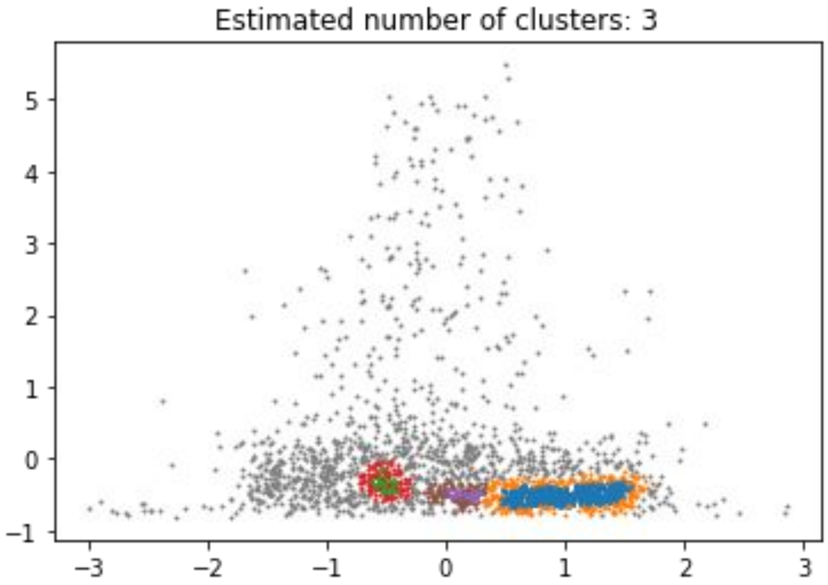


Center(C) 16-24 ft

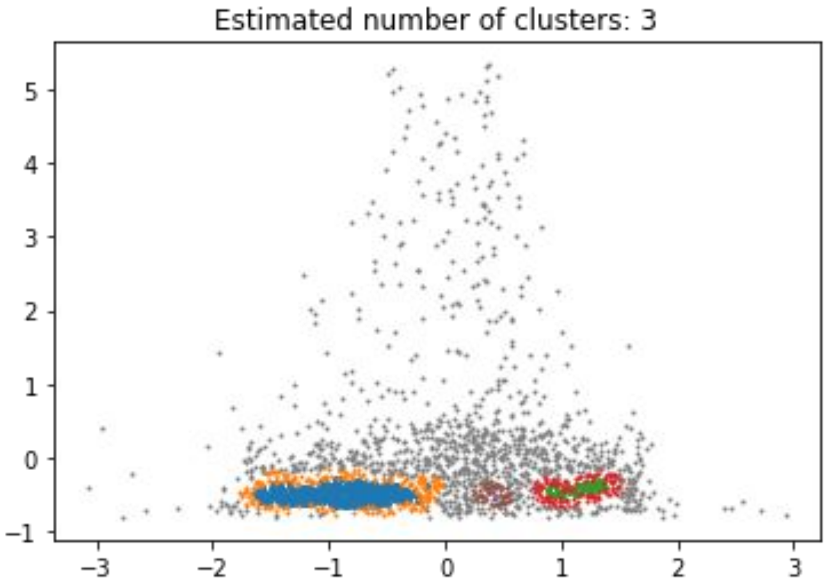


Center(C) 8-16 ft.

DBScan Clustering:
Comparing Shots from Each Side of Court



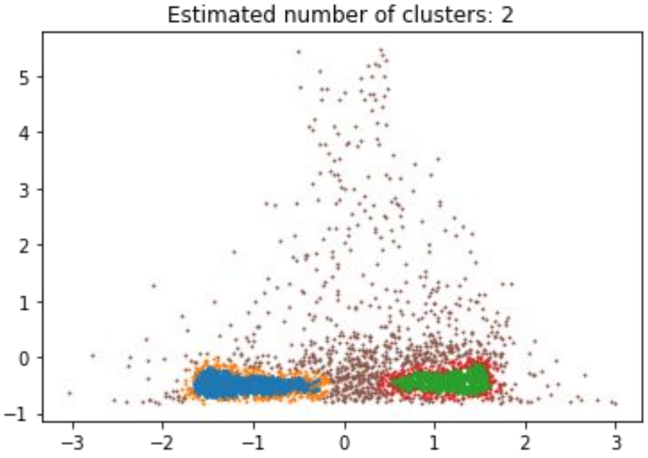
Left Side Center(LC) 16-24 ft.



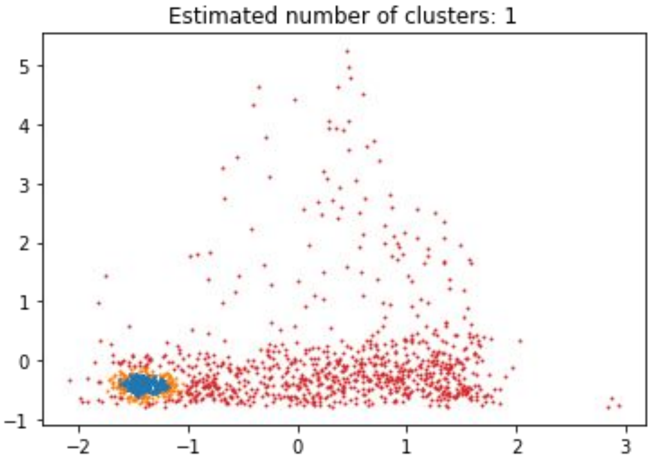
Right Side Center(RC) 16-24 ft.

DBScan Clustering: Comparing Across Shot Distance

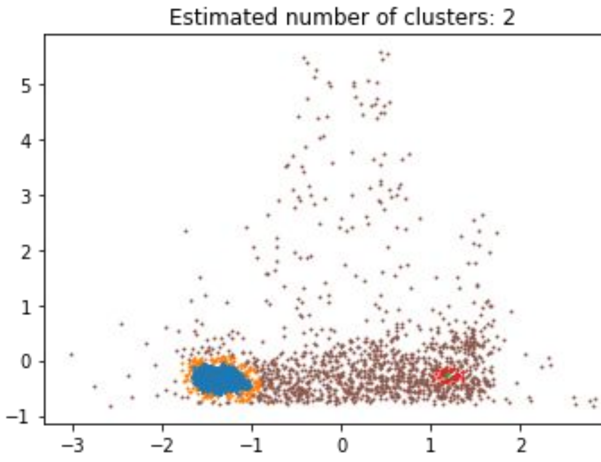
- Different high density rebound areas for each of the 3 shot distance ranges



Right Side(R) 8-16 ft.



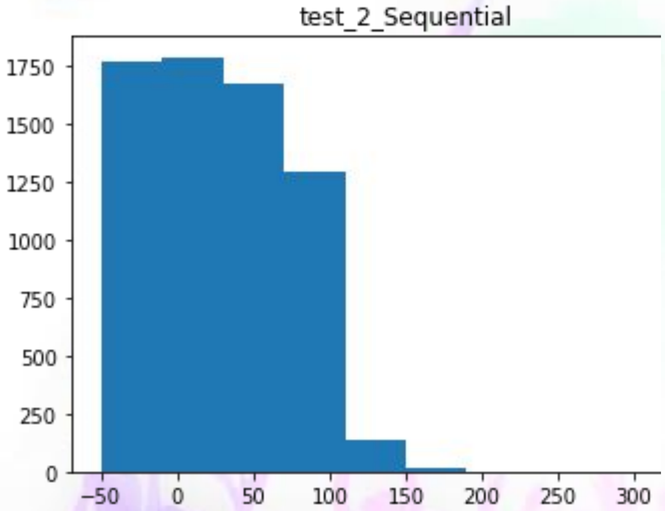
Right Side(R) 16-24 ft.



Right Side(R) 24+ ft.

Mediocre Model Predictions for Regression and Classification

- Works off of strong relationship between rebound angle and shot angle and medium relationship between rebound distance and shot distance
- Mostly quite bad prediction scores despite adjustments in predictor variables, by player, by shot zone, by shot type and adjusting Y variable (9 quadrants with more balanced frequencies)
    - Best r2 score: Adaboost (0.32 on test set)
- Regression scores predicting numerical rebound angle or distance were even worse
    - Best on neural network with dropout, see right
    - Still mediocre, residual error average +/- 50 degrees

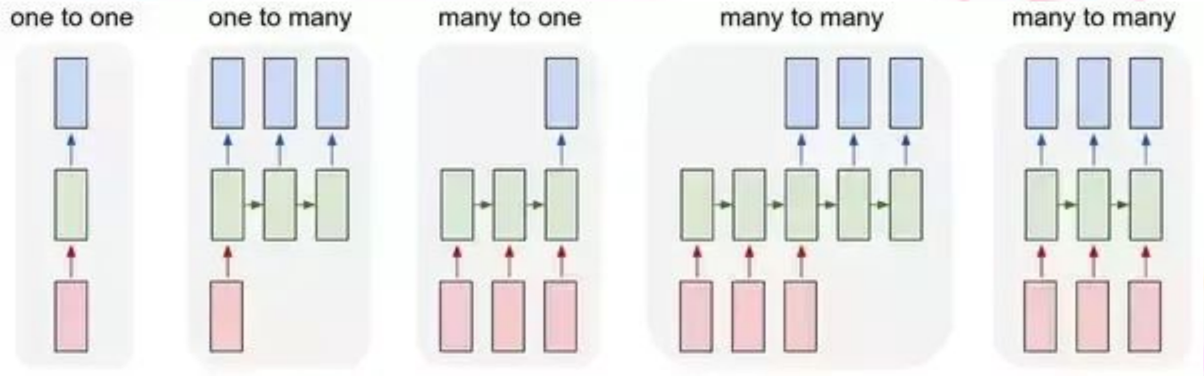


Mechanics of Rebounding and Limitations of Point-in-Space Data

- Harder to fully quantify with SportVu movement data
- However, can identify when offense is running to try for fast break, and measure separation to nearest defenders
- Analysis of any contested rebound could use video of body positioning of each contestant, timing and extension of arms, body orientation
- Sloan Sports Analytics Conference paper: “To Crash or Not To Crash: A quantitative look at the relationship between offensive rebounding and transition defense in the NBA”
- Change of possession definitely currently less examined in NBA Analytics

Bill Streicher-USA TODAY Sports

Machine Learning Model Types, Input vs Output Dimensions

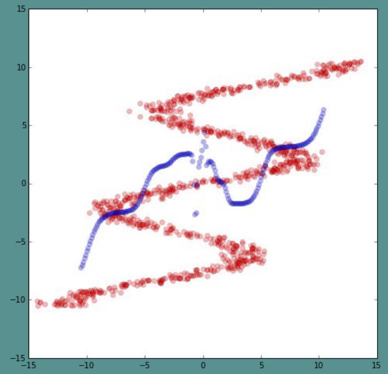


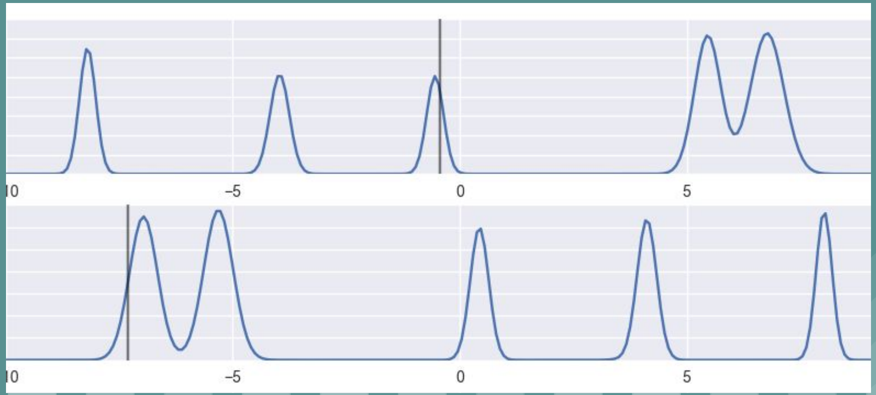
Inputs in red, output in blue, model processing stage in green. Left to right:

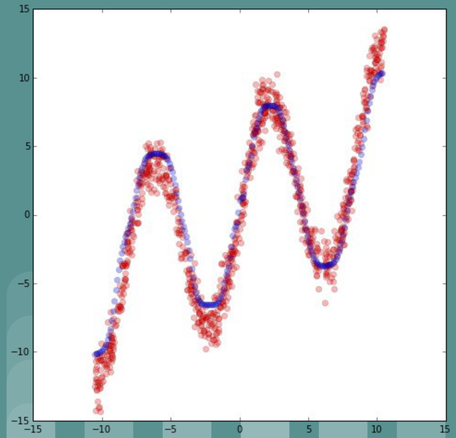
- Fixed-sized input to fixed-sized output: image classification
- Sequence output (e.g captioning from image to caption).
- Sequence input: sentiment analysis with sentence input classified to positive/negative
- Unmatched sequence input and output: translating sentences, not necessarily 1 : 1
- Synced sequence input and output

Exploring Mixture Density Networks to Address the Multi-Label Classification Problem

- Ideally, model predictions should be a probability distribution similar to DBScan density plots
- Output a sequence of probabilities that ball can go to each location (many to many model type)
- On right, classic models can't predict multiple Y clusters per one input X vector (X axis)
- Alternatives: ensemble model with new model for each DBScan cluster
- MDN or Mixture Density Networks with neural networks can output probability sequences (along single dimension below)







Takeaways and Lessons Learned

1. Movement data is hard, messy, and a fight to engineer new useful features, esp. without pre-labeled body positioning
    a. For the last seasons, NBA provides rich labeled play-by-play data but no longer any raw movement data at all
2. Target y variables with wide probability distributions, like rebound location, are inherently hard to model (without uncommon model types)
3. Next stage to analyze all player movement, who are going to which high probability rebound locations

4. Sports time series data is a rich to mine. Here with a few more variables I could accurately extract shooting motion time, time to reach rebound location etc.

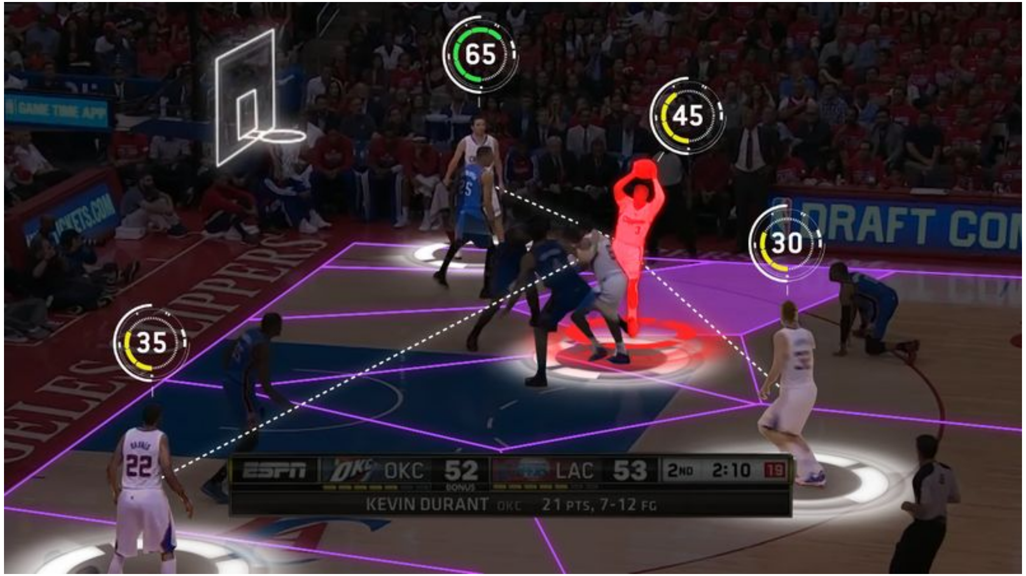
5. Defined, constrained Y variables make every stage easier. I will complete a smaller referee call project to get from start to end of a data science problem.

2nd Spectrum Tracking Extended Exclusive Deal with NBA

A ton of respect to 2nd Spectrum which does more complex analysis, in real-time, for every NBA team and is used by all NBA coaching staffs

Longitudinal analysis, shooting percentage by location for each player

To Be Continued...



NBA Analytics:
Part 2 Referee Types of Calls - False Positives vs False Negatives

TBD...

Thank you!!