NBA Shot Value Prediction

Ray Chen

Problem statement

- Being such a popular sport, basketball naturally has become very competitive
- So how can you edge out competitors?
  - Games can be decided by one team’s stronger shot selection, even if both teams have similar talent
- My model will try to predict which shots have the highest value
  - This will be calculated by percentage of shots made for this shot, and the points the shot is worth

Use case

- My model can be implemented by:
  - Players looking to better their own shot selection
  - Coaches taking stats in the game to see what value shots they are taking
  - Coaches scouting players by judging their shot selection

Dataset

- My dataset was taken off of Kaggle at https://www.kaggle.com/dansbecker/nba-shot-logs
    - 128 thousand by 21
    - 14 numeric, 7 non-numeric
    - Columns to determine shot value would be pts_type and shot_made

Data wrangling

- 3 Columns contain mistakes or N/A values
  - Shot_clock has n/a values
  - Touch_time has negative values
  - Pts_type doesn’t always match up shot_dist
- Deleted observations with 2 or more of these mistakes
- Imputed mean for shot_clock
- Changed negative values of touch_time to 0
- Changed pts_type to the correct points based on shot_dist

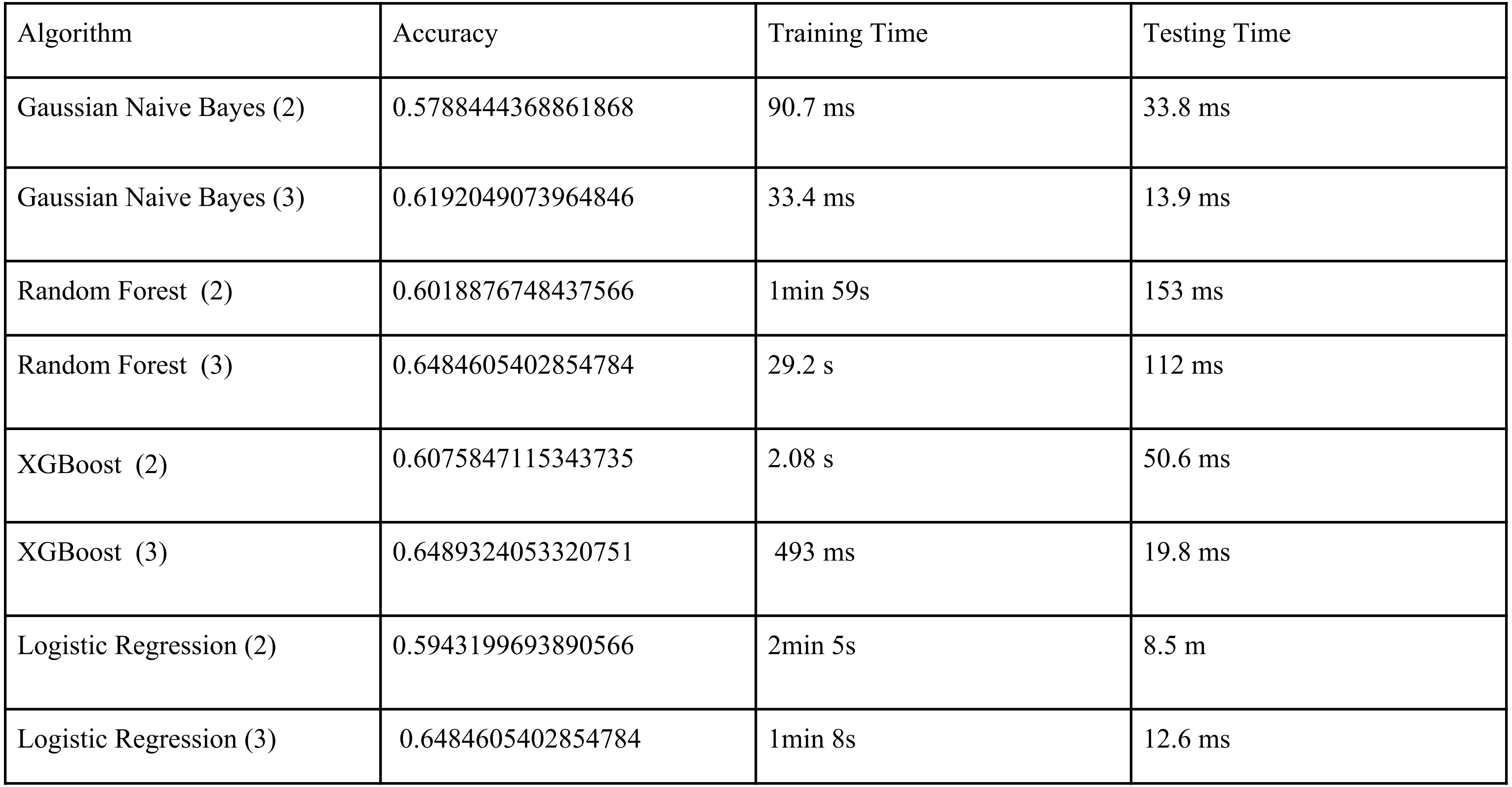
Initial findings

- 2 variables have an obvious impact on whether or not the shot was made:
  - points type
  - shot distance
- Others have an impact, though not as extreme
  - Shot number
  - Closest defender distance
  - Shot clock
  - Dribbles
  - Touch time
- These 6 variables will be used as predictor variables

Building the final model

- Split data set into 2, one for 2 pointers and one for 3 pointers
- Further split into testing and training data for both 2 and 3 pointers
- Looked at accuracy scores, training times, and testing times for XGBoost, random forest, naive Bayes, and logistic regression algorithms

Score Reports for Algorithms



Algorithm	Accuracy	Training Time	Testing Time
Gaussian Naive Bayes (2)	0.5788444368861868	90.7 ms	33.8 ms
Gaussian Naive Bayes (3)	0.6192049073964846	33.4 ms	13.9 ms
Random Forest (2)	0.6018876748437566	1min 59s	153 ms
Random Forest (3)	0.6484605402854784	29.2 s	112 ms
XGBoost (2)	0.6075847115343735	2.08 s	50.6 ms
XGBoost (3)	0.6489324053320751	493 ms	19.8 ms
Logistic Regression (2)	0.5943199693890566	2min 5s	8.5 m
Logistic Regression (3)	0.6484605402854784	1min 8s	12.6 ms

Final model

- XGBoost results in the highest accuracy for both kinds of shots
- Take the probability of shot being made and multiply it with the points type