Predicting Successful Shot Attempts in the NBA

By: Preethi Seshadri and Srinidhi Srinivasan

Motivation

- Predicting March Madness, Super Bowl, NBA finals, etc. are not trivial Machine Learning problems
- Sports Analytics in general can provide insights to both long-time fans and teams
- It would be interesting to analyze sports stats quantitatively

Problem

- We are predicting whether or not a basketball player will make a shot using NBA player shot log data from the 2014-2015 regular season.
- In addition to predicting whether or not a player will make a shot, we hope to determine which factors are most indicative of making a shot for the most well-known basketball players, and how these differ from player-to-player

Our Data

- Kaggle dataset collected from NBA api (from 2014-15 season)
- Original dataset: 128,069 instances and 21 features
- Features we used to predict:
  - SHOT_NUMBER
  - PERIOD
  - GAME_CLOCK
  - SHOT_CLOCK
  - DRIBBLES
  - TOUCH_TIME
  - SHOT_DIST
  - PTS_TYPE
  - CLOSE_DEF_DIST

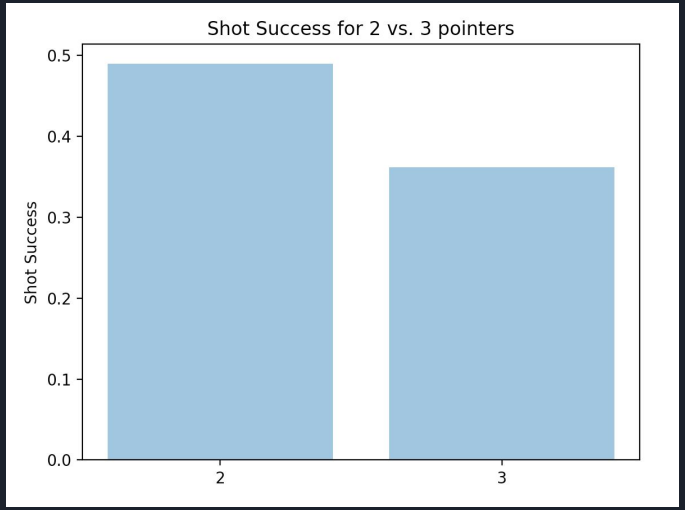
FEATURES NOT USED
GAME_ID
MATCHUP
LOCATION
W (WIN OR LOSE)
FINAL_MARGIN
CLOSEST_DEFENDER
CLOSEST_DEFENDER_PLAYER_ID
FGM
PTS
PLAYER_NAME
PLAYER_ID

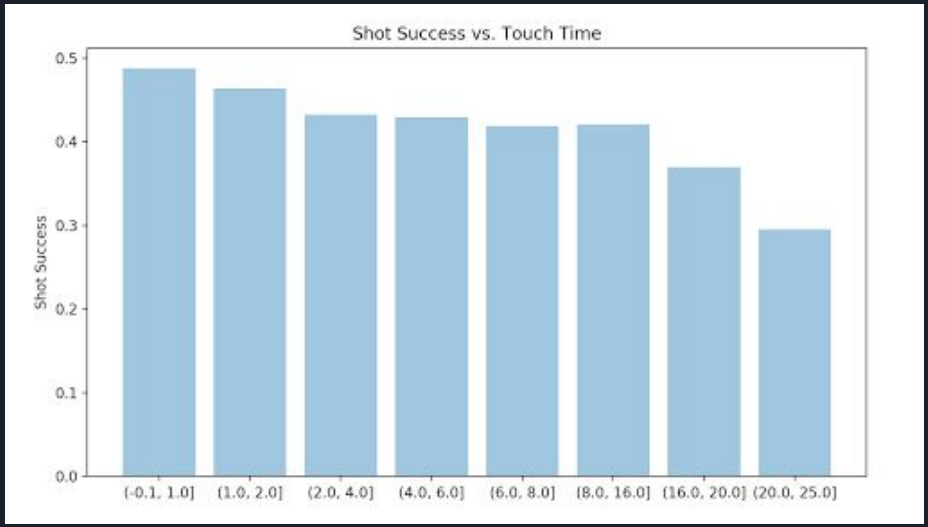
FEATURES WE USED
SHOT_NUMBER
PERIOD
GAME_CLOCK
SHOT_CLOCK
DRIBBLES
TOUCH_TIME
SHOT_DIST
PTS_TYPE
CLOSE_DEF_DIST

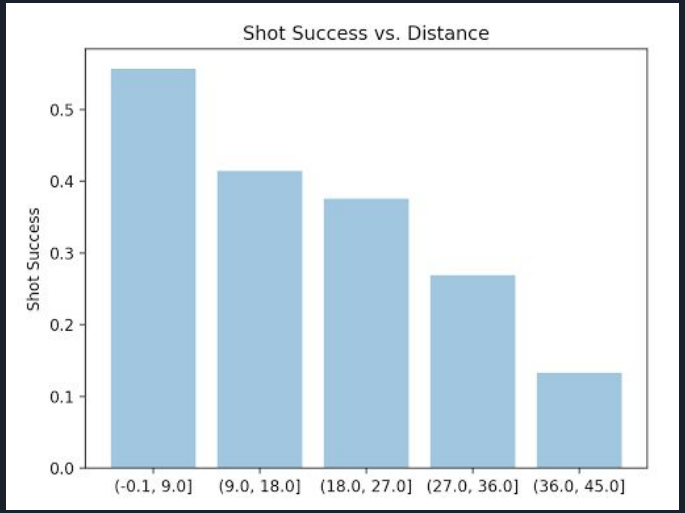
RESPONSE
SHOT_RESULT

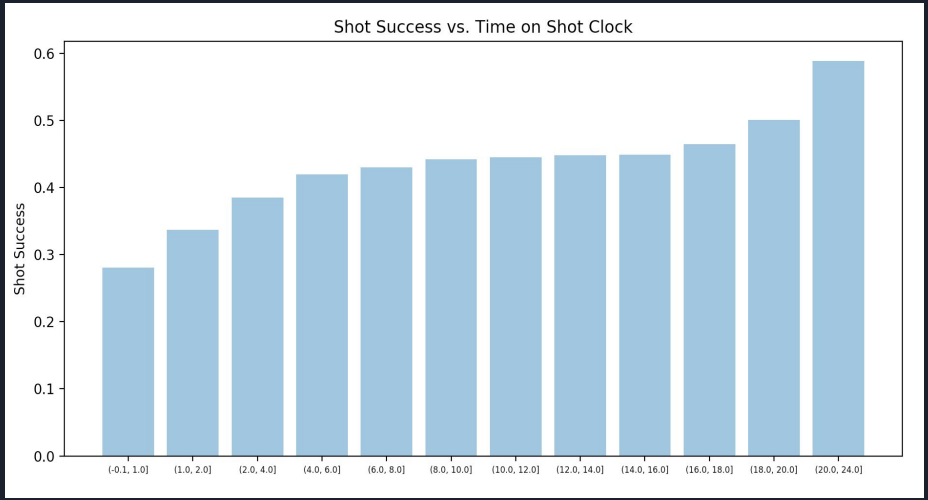
1) Removed features that were redundant:
   - CLOSEST_DEFENDER, CLOSEST_DEFENDER_PLAYER_ID

2) Removed features that were not generalizable to all players:
   - GAME_ID, PLAYER, MATCHUP, and LOCATION


Shot Success for 2 vs. 3 pointers
Shot Success
0.0
0.1
0.2
0.3
0.4
0.5
2
3


Shot Success vs. Touch Time
Shot Success
0.0
0.1
0.2
0.3
0.4
0.5
(-0.1, 1.0]
(1.0, 2.0]
(2.0, 4.0]
(4.0, 6.0]
(6.0, 8.0]
(8.0, 16.0]
(16.0, 20.0]
(20.0, 25.0]


Shot Success vs. Distance
Shot Success
0.0
0.1
0.2
0.3
0.4
0.5
(-0.1, 9.0]
(9.0, 18.0]
(18.0, 27.0]
(27.0, 36.0]
(36.0, 45.0]


Shot Success vs. Time on Shot Clock
Shot Success
0.0
0.1
0.2
0.3
0.4
0.5
0.6
(-0.1, 1.0]
(1.0, 2.0]
(2.0, 4.0]
(4.0, 6.0]
(6.0, 8.0]
(8.0, 10.0]
(10.0, 12.0]
(12.0, 14.0]
(14.0, 16.0]
(16.0, 18.0]
(18.0, 20.0]
(20.0, 24.0]

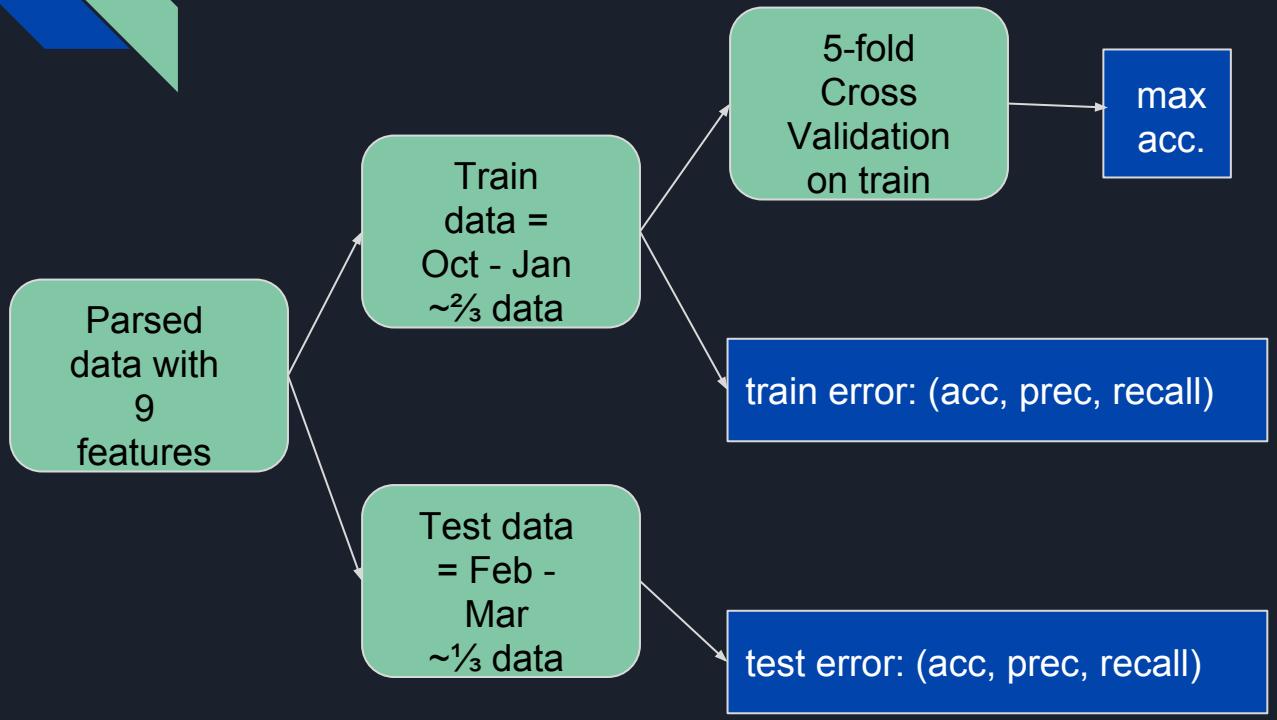
Dataset Information

- 116961 observations
- Close to having balanced classes
  - 45.6% of shots in the dataset are makes, and the remaining 54.5% are misses
  - 54.5% will be our baseline performance
- We looked at accuracy, precision, and recall when evaluating model performance

Filtering Data

1. Entire Data set
2. Average or above on both:
    a. number of shot attempts made
    b. percentage of successful shot attempts
3. 75% or above on both:
    a. number of shot attempts made
    b. percentage of successful shot attempts
4. Split Data by Point Type (2 pointer or 3 pointer)
    a. used all iterations of filtering (1-3 above)

Methods



MODELS:

Support Vector Machine (SVM)

Logistic Regression

Decision Trees

Random Forest

Results: Entire Data set

		Logistic Regression	SVM	Decision Trees
Training performance	accuracy	0.6038	0.6155	0.6175
	precision	0.5833	0.6389	0.6771
	recall	0.4711	0.3680	0.3143
Testing performance	accuracy	0.6072	0.6170	0.6202
	precision	0.5753	0.6248	0.6611
	recall	0.4781	0.3680	0.3161

Problems:

- Low training/testing performance for all 3 classifiers
- Indicates a high bias and underfitting for our models

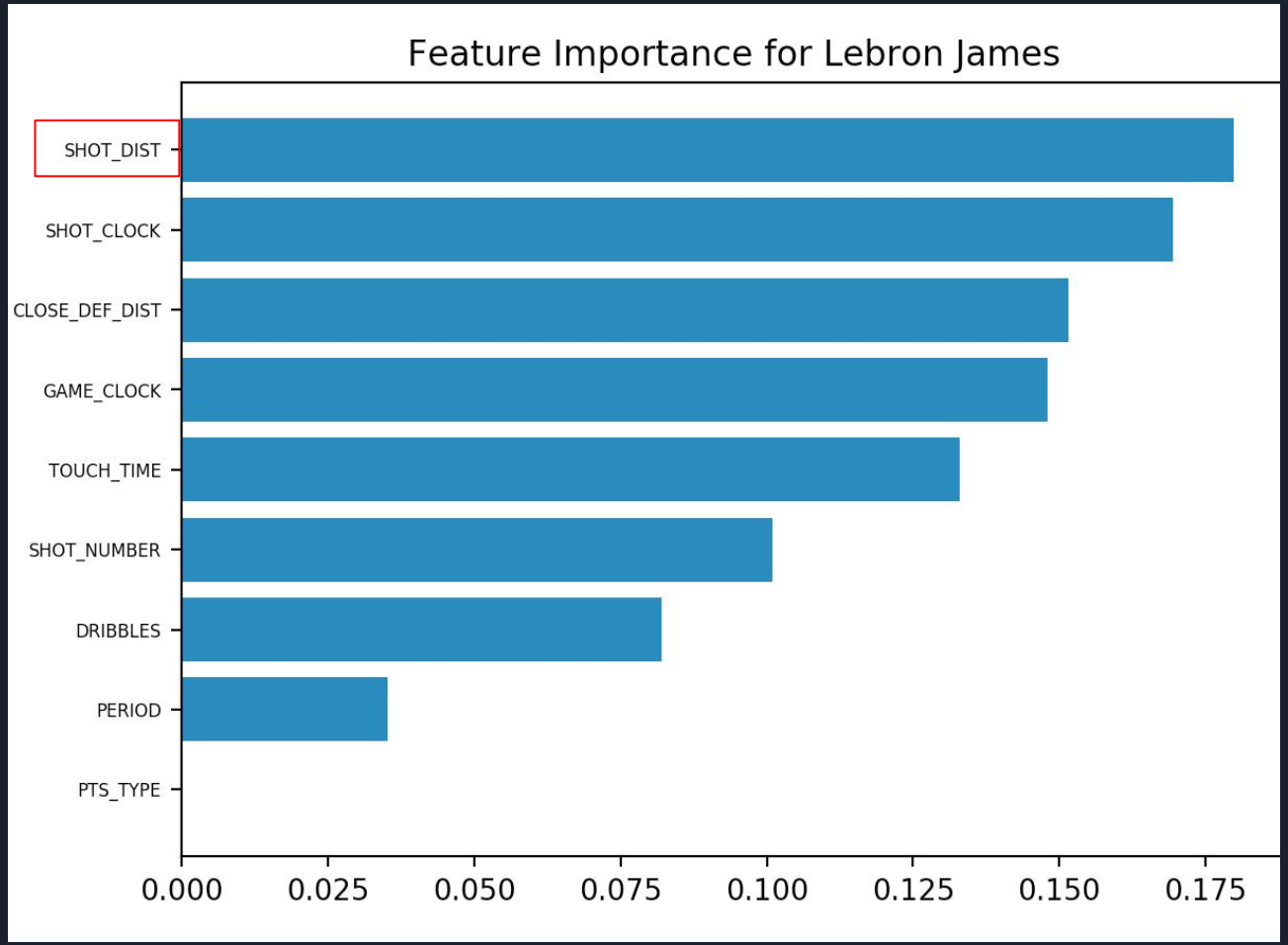
Results: Entire Data Set Feature Importance

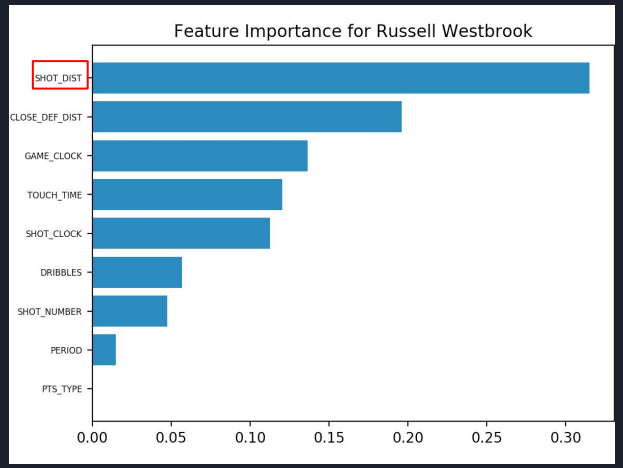
Testing performance

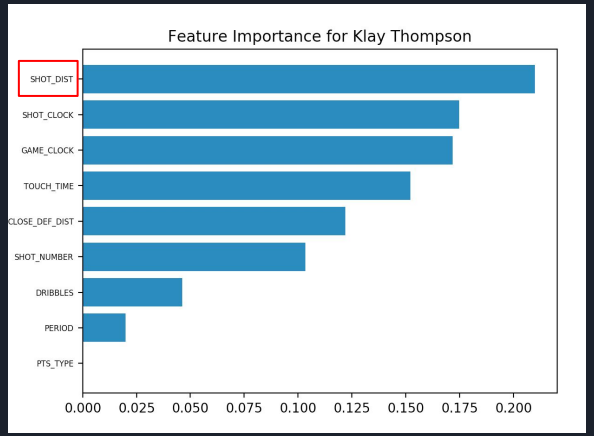
NBA player	accuracy	precision	recall
LeBron James	0.6611	0.7436	0.4874
Klay Thompson	0.5799	0.4651	0.2247
Kawhi Leonard	0.5643	0.5102	0.4032
James Harden	0.5751	0.5	0.3780
Steph Curry	0.5381	0.5325	0.4020
Russell Westbrook	0.5672	0.6410	0.1969

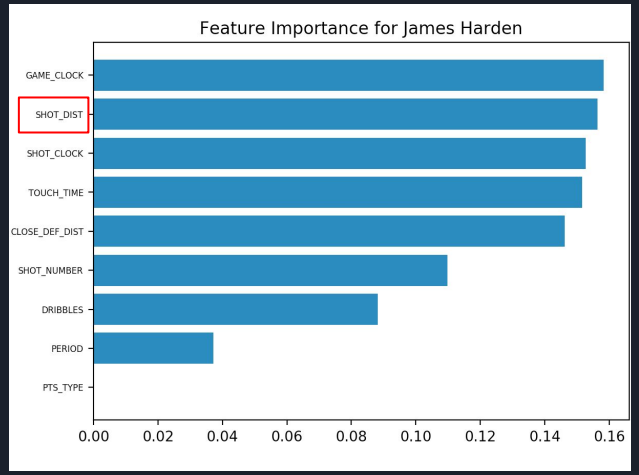
Problems:

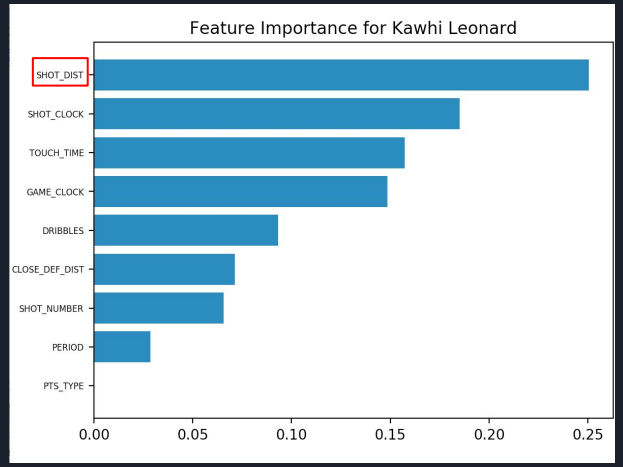
- Similarly, we see low training/testing performance even on a player level

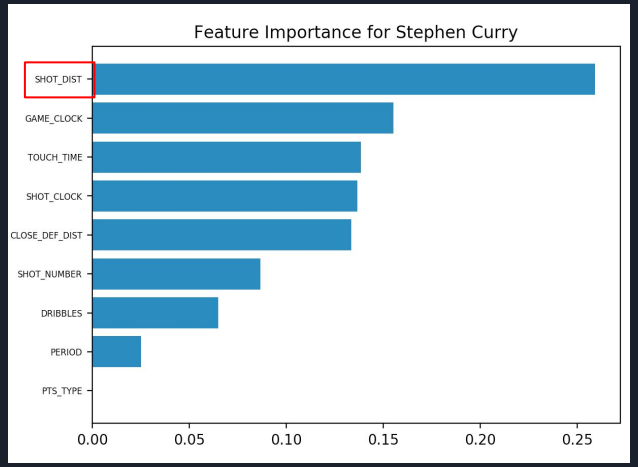

Feature Importance for Lebron James
SHOT_DIST
SHOT_CLOCK
CLOSE_DEF_DIST
GAME_CLOCK
TOUCH_TIME
SHOT_NUMBER
DRIBBLES
PERIOD
PTS_TYPE
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175



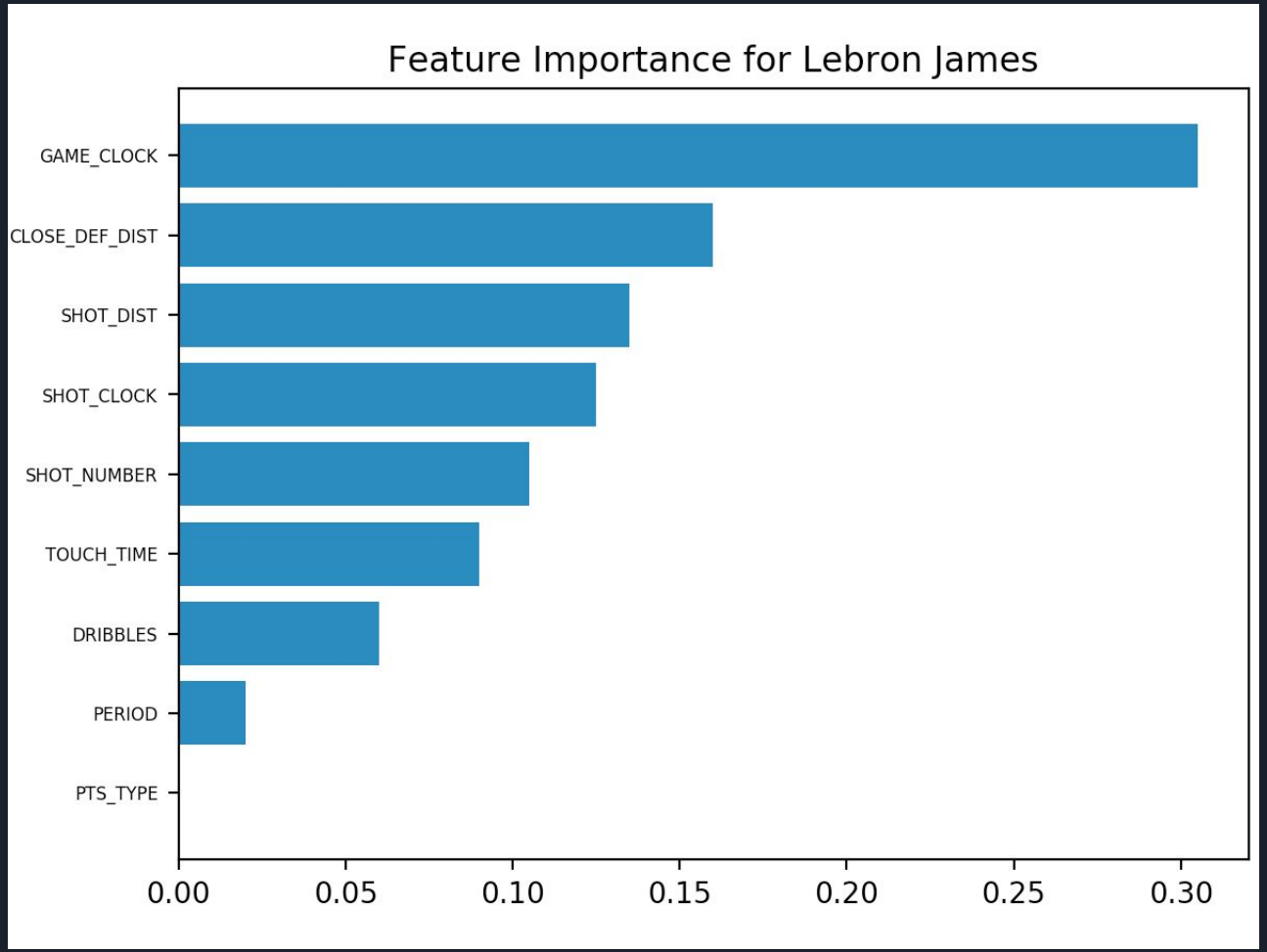


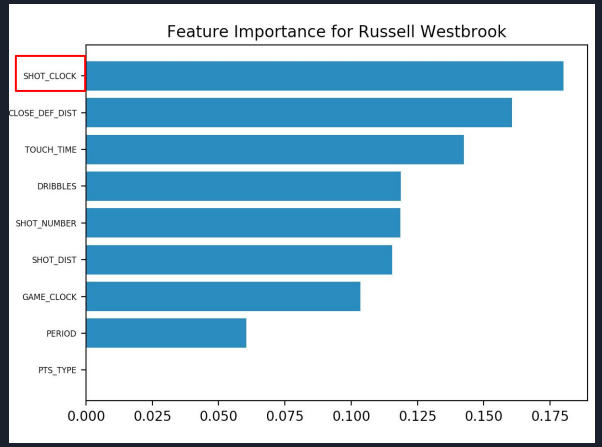


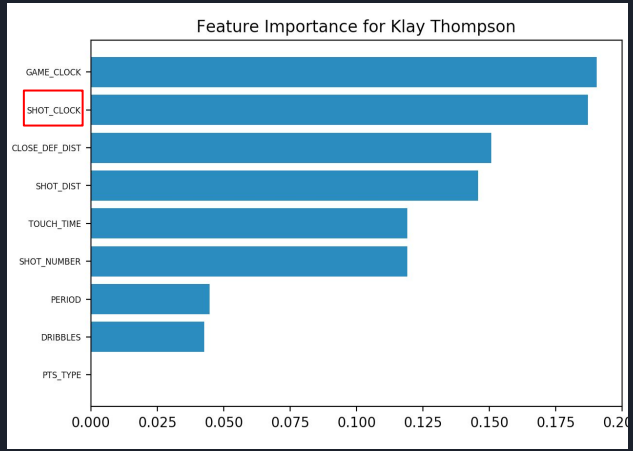


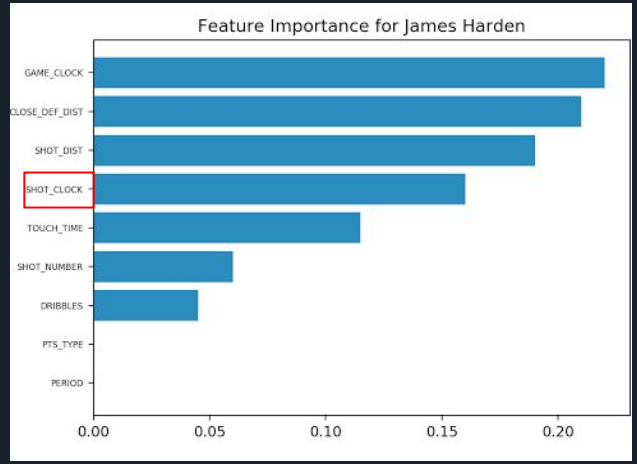


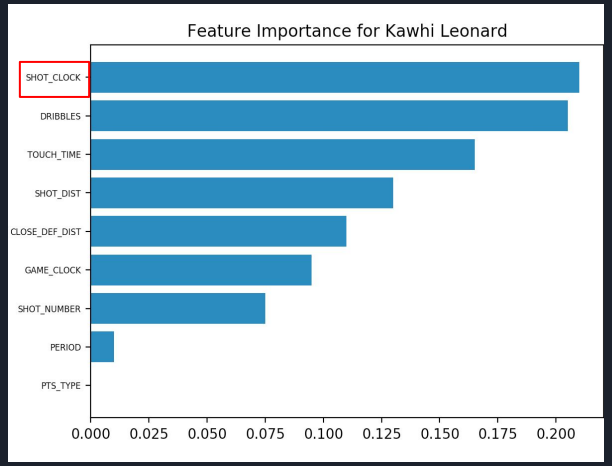
Results for Feature Importances using data set filtered by 2 pointers

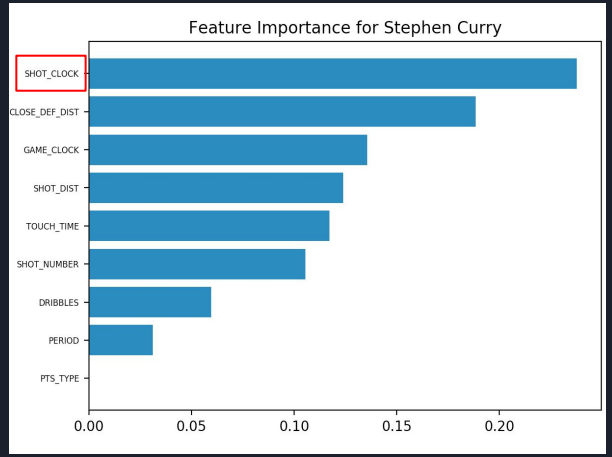

Feature Importance for Lebron James
GAME_CLOCK
CLOSE_DEF_DIST
SHOT_DIST
SHOT_CLOCK
SHOT_NUMBER
TOUCH_TIME
DRIBBLES
PERIOD
PTS_TYPE
0.00
0.05
0.10
0.15
0.20
0.25
0.30











Results for Feature Importances using data set filtered by 3 pointers

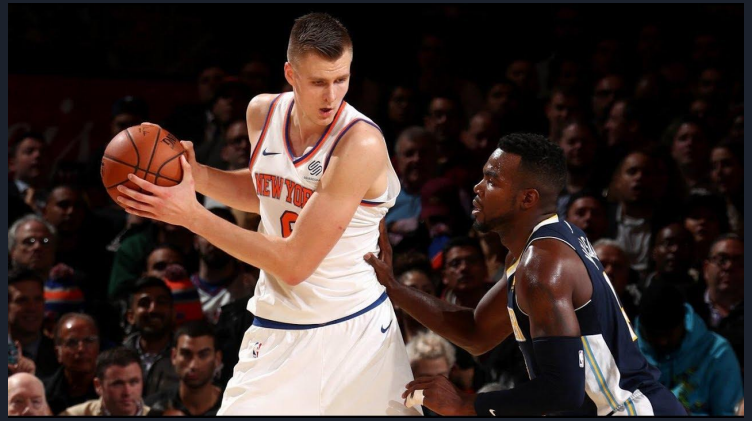
Conclusion

Insights gained from this project:

- All our models indicate underfitting, which highlights that we need more features in our dataset
- Predicting a successful shot attempt is not as clear-cut as we thought it would be
    - There is too much variation regarding the conditions in which players make and miss shots
- Sports analytics and sports betting are really challenging areas

Future Avenues

- Gather additional features
  - Height/weight of player and defender, angle and/or region of the court
- Explore unsupervised learning approaches
  - Clustering to group players and capture underlying structure in our data, and run our models on clustered data



Acknowledgements

We would like to thank Prof. Wu and all the grutors for their support and their helpful suggestions to our questions.

Thank You!

Questions?