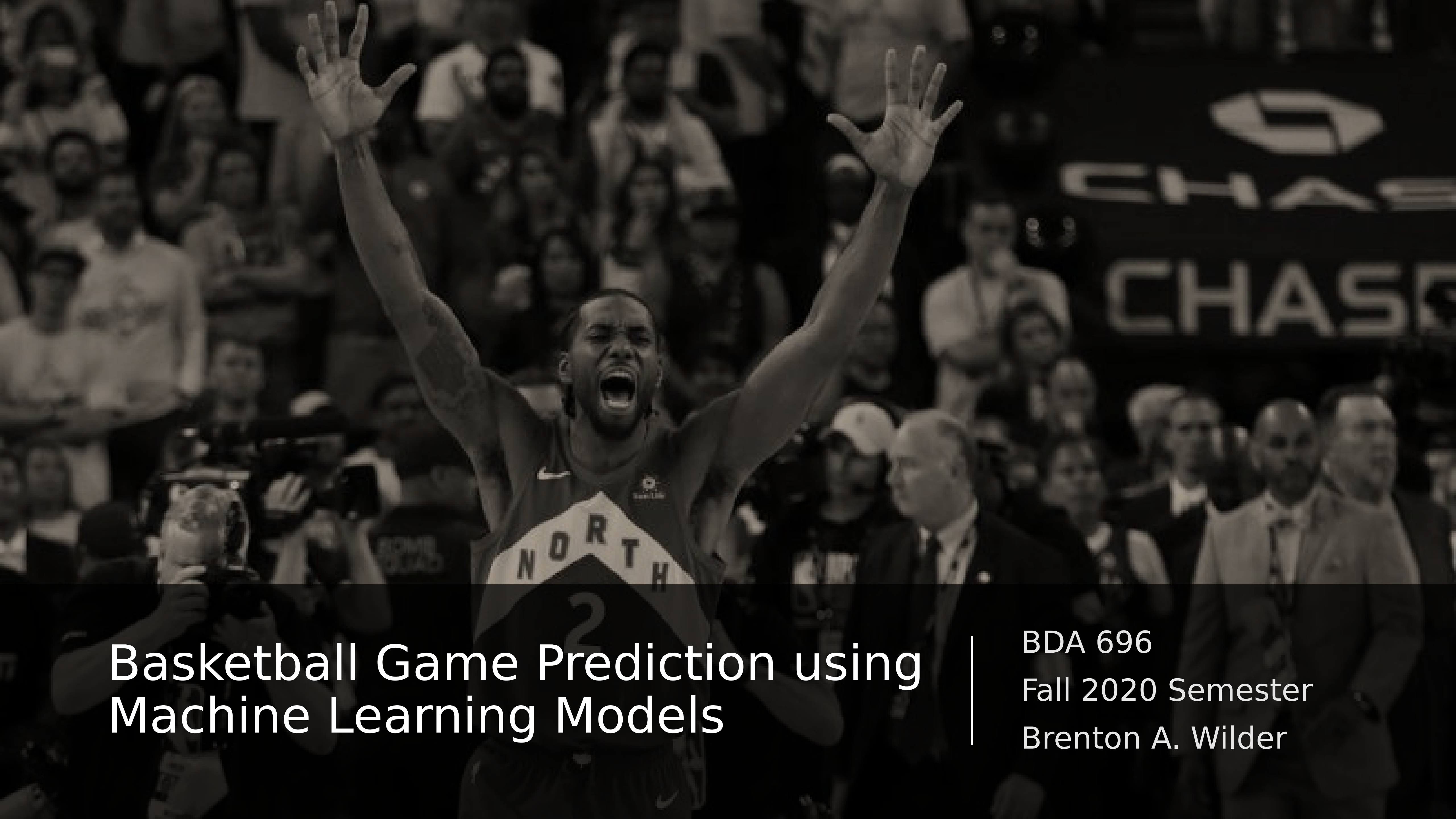
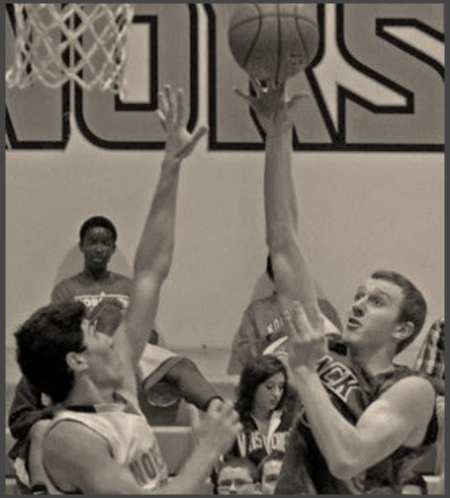


Basketball Game Prediction using Machine Learning Models

BDA 696
Fall 2020 Semester
Brenton A. Wilder

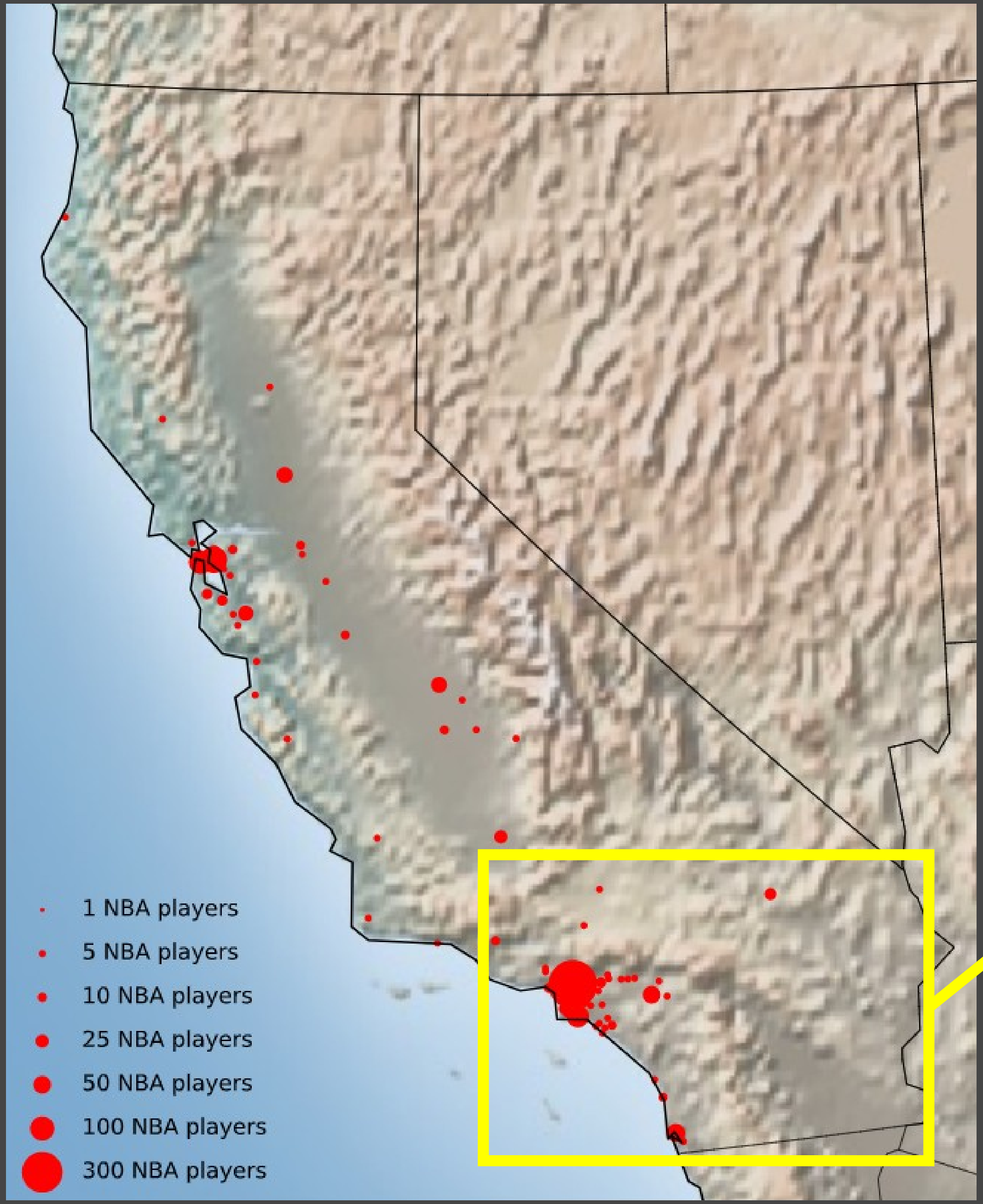
Windows Subsystem for Linux (WSL)

- Was a bit hard to get into, but eventually got it to work
- Had issues with my work mysteriously vanishing
- Overall was not bad!



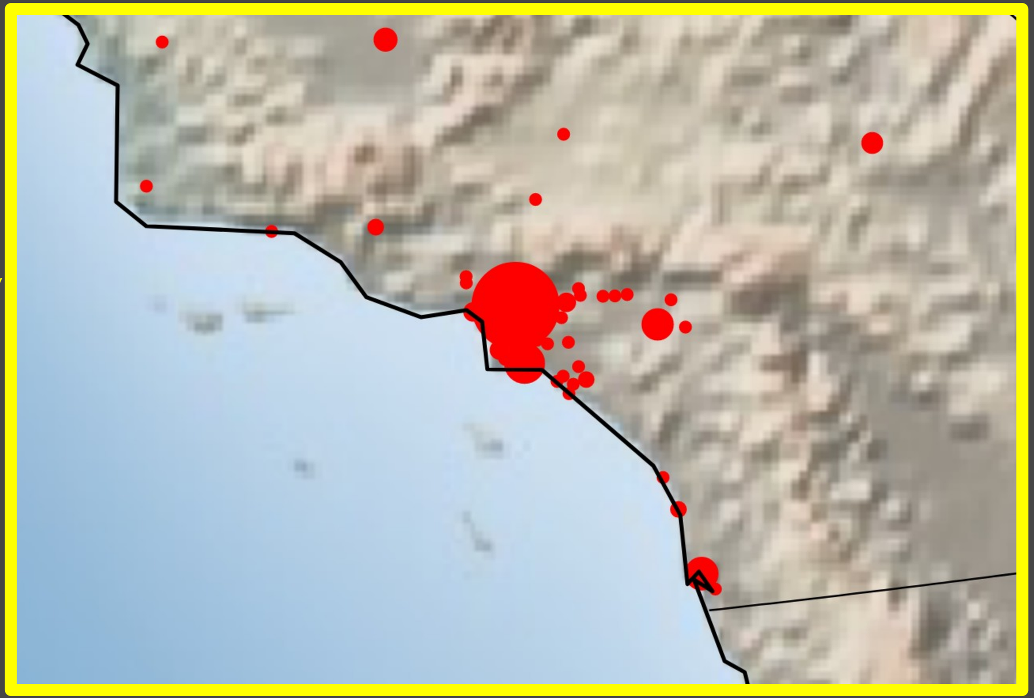
Motivation

- I am a huge basketball fan and has meant a lot to me growing up.
- It would be interesting to see if/ which features can be correlated to winning basketball games
- These kind of models could benefit players and coaches to analyze the game and understand trends



Motivation

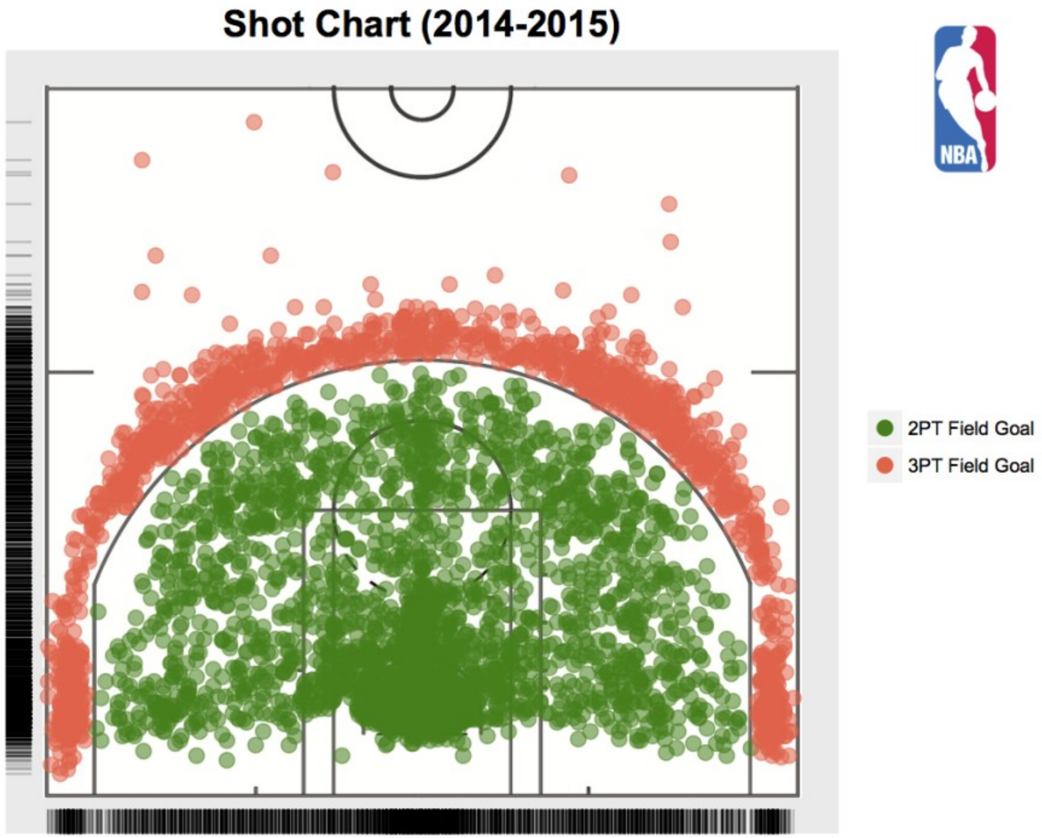
- Further, this is a locally important topic as the National Basketball Association (NBA) has most of its talent coming from right here in southern California



Including a lot of great players from SDSU!

https://www.youtube.com/watch?v=0-stx17fY-I

	ID	Player	height	weight	College
1314	1314	Joel Kramer	201.0	92.0	San Diego State University
1376	1376	Steve Malovic	208.0	104.0	San Diego State University
1657	1657	Michael Cage	206.0	101.0	San Diego State University
3077	3077	Randy Holcomb	206.0	102.0	San Diego State University
3491	3491	Kawhi Leonard	201.0	104.0	San Diego State University
3517	3517	Malcolm Thomas	206.0	102.0	San Diego State University
3635	3635	Jamaal Franklin	196.0	86.0	San Diego State University

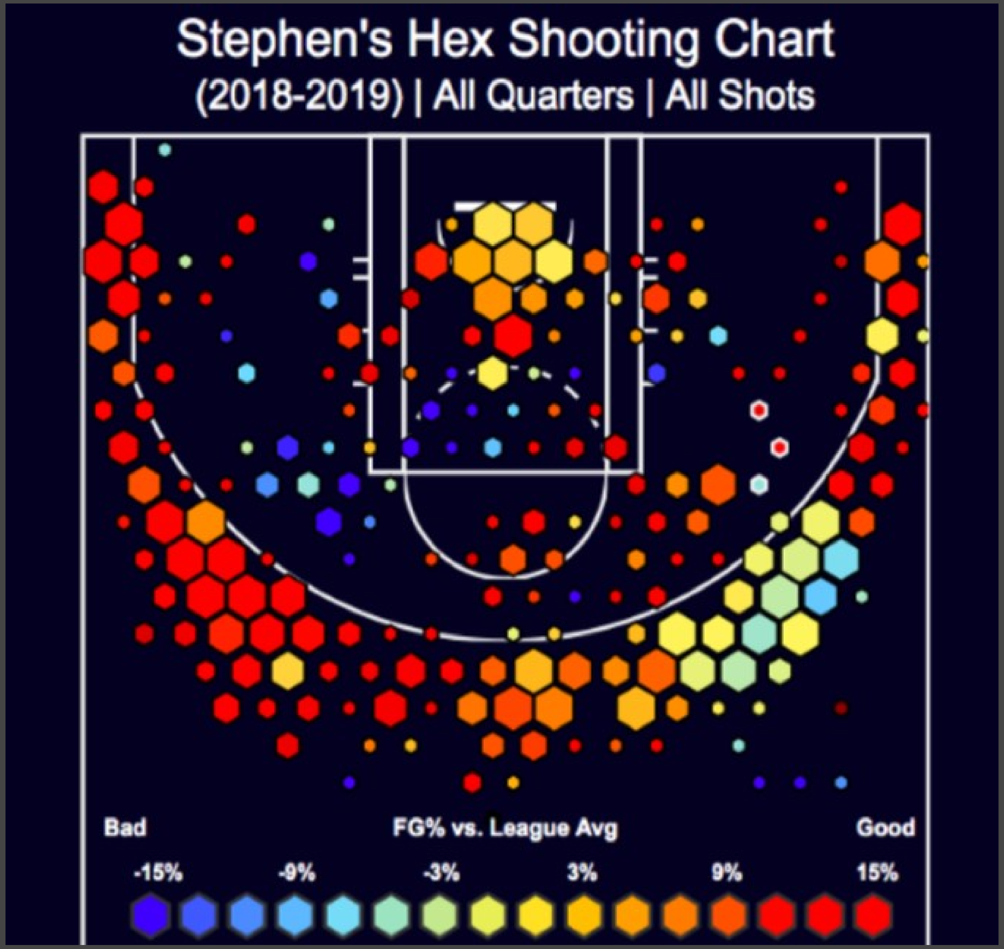


A little background on basketball

- Each team has 5 players on the court at one time
- Like soccer, players try to score the basket in the goal
  - Instead of dribbling with the feet, players dribble the ball with hands
- The players can either attempt a 2-point shot, which is located within the outer arc. Or they can attempt a 3-point shot, which is located outside the arc. This arc is called the “3-point line”
- If a player is fouled, they can also shoot free throws. These free throws are worth 1-point each.

https://412sportsanalytics.wordpress.com/2016/05/30/the-curious-case-of-the-3-point-line/

Why would basketball players need analytics?

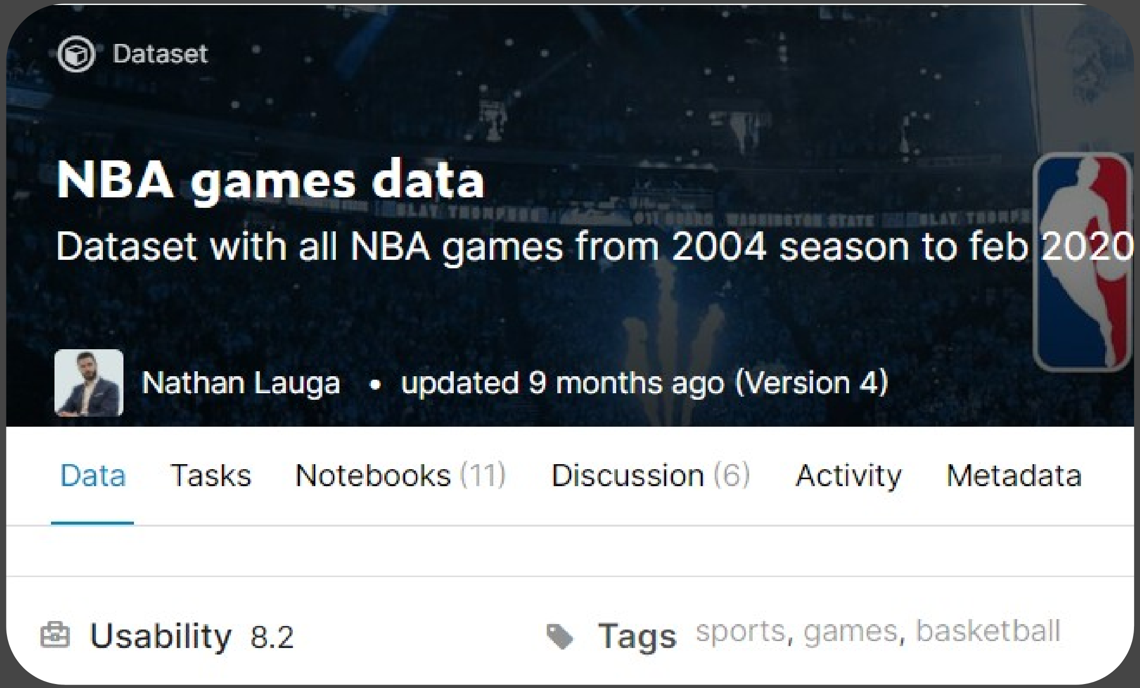


https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116

- **Game to game Strategy**
- **Player improvement**
- **Defensive schemes**
- **Team building**

How did we use data for this project?

https://www.kaggle.com/nathanlauga/nba-games



- Like the baseball dataset introduced by Julien, I analyzed games data from 2004 season to the most recent season (pre-COVID)
- **We wanted to predict if home team will <u>win</u> or <u>lose</u>**
- We attempted this using an ensemble of different machine learning classification models after extracting several features

Data preparation

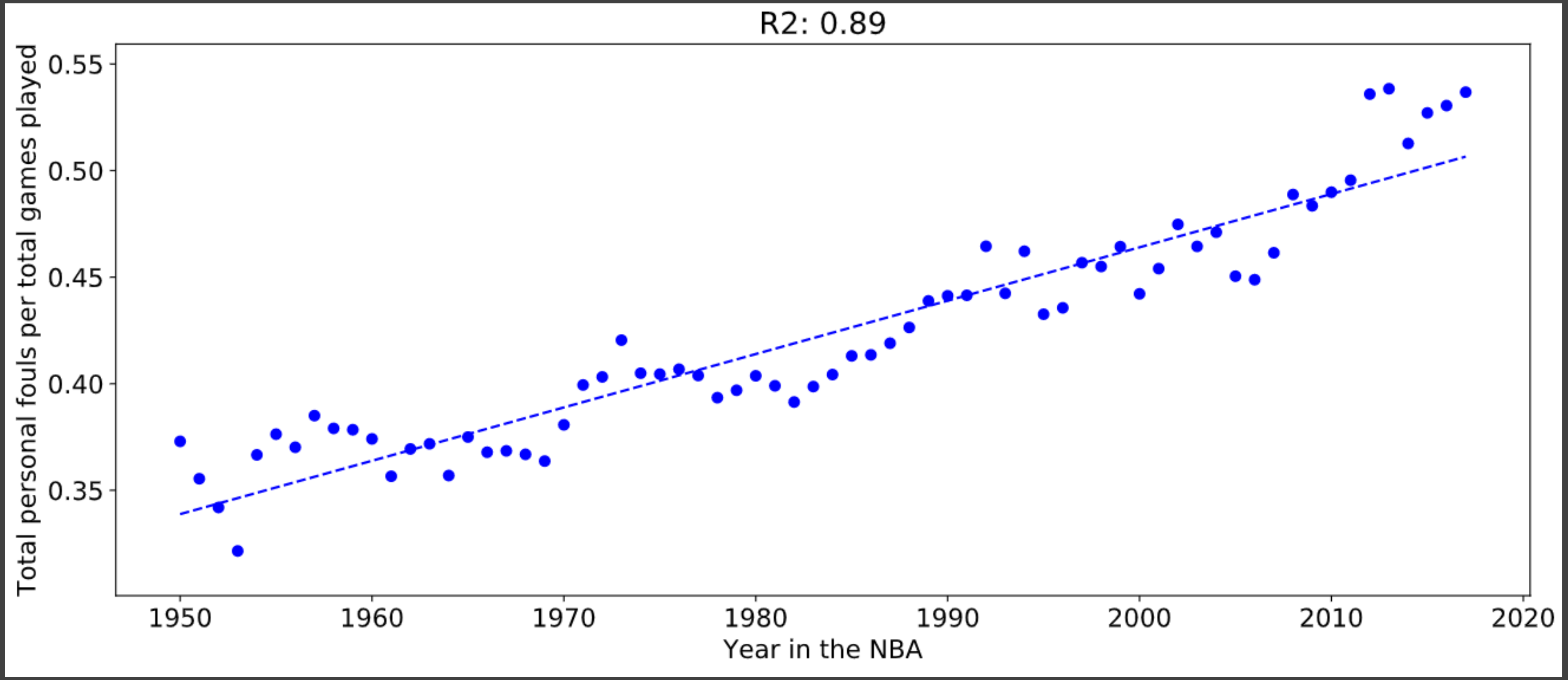
- Remove NaN values
- Convert all values to float values
- Remove variables that could be considered as target leakage
- Define target as home team win [1] or home team loses [0]

Feature Engineering

https://www.businessinsider.com/fans-favorite-nba-arenas-2018-10

- Tried several different features in the models, and I will overview of interesting ones created
- For example, one feature used the ***home team arena capacity*** to see if there was an impact from the fan noise or so-called “home court advantage”.

Feature Engineering



- In another example, I tried to capture the trends of the most recent season. This is important because NBA teams tend to do a lot of trading and team dynamics can change.
- I also took the difference of these recent season trends between the home team and away team of the game trying to be predicted.

Complete list of features

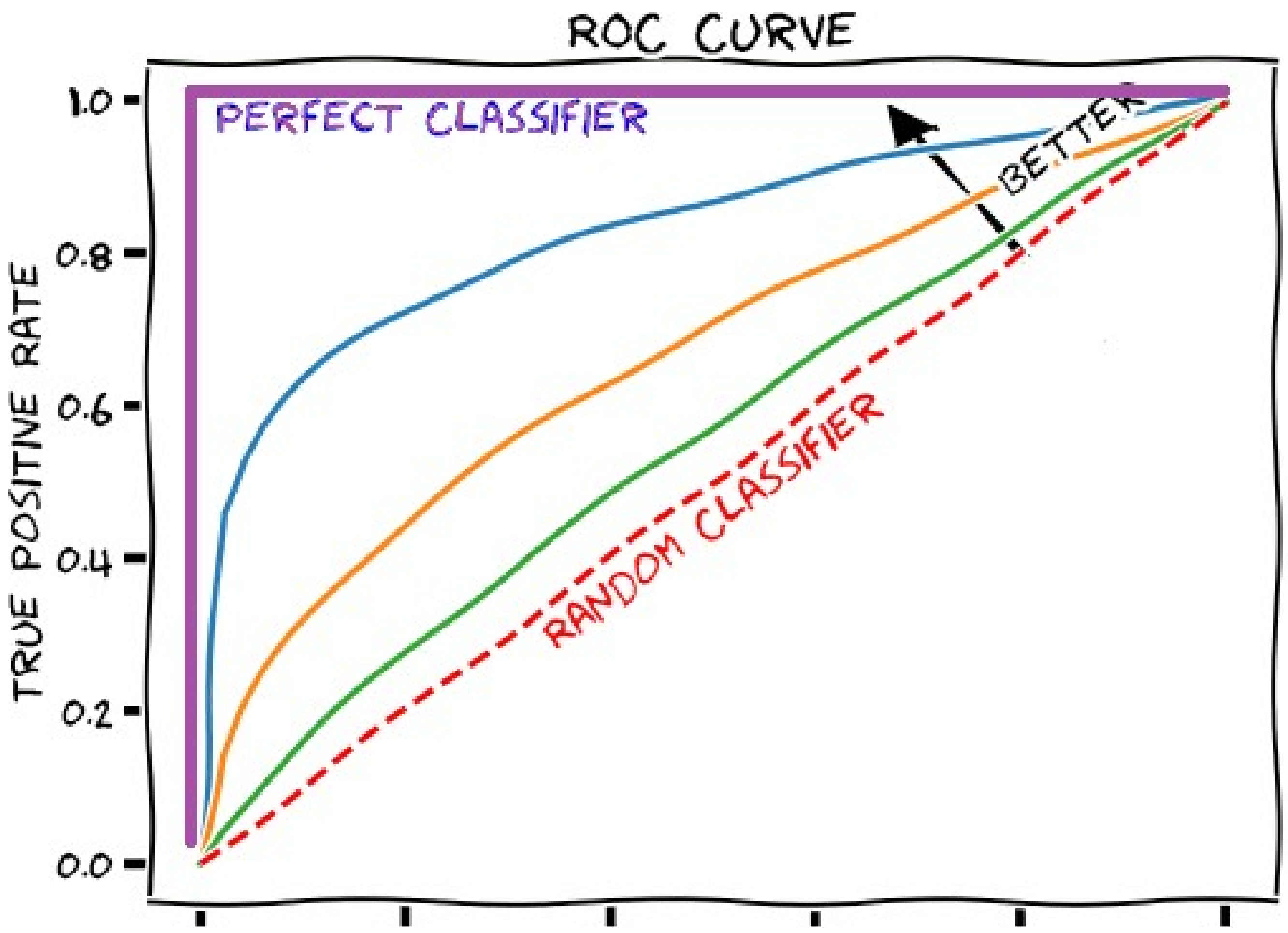
1. `GAME_DATE_EST` - the date the predicted game was played on (from original dataset)
2. `GAME_ID` - the unique game ID for the predicted game (from original dataset)
3. `HOME_TEAM_ID` - the unique team ID for the Home Team (from original dataset)
4. `VISITOR_TEAM_ID` - the unique team ID for the Away Team (from original dataset)
5. `SEASON` - the code for which NBA season (from original dataset)
6. `TEAM_ID_away` - the unique team ID for the Away Team (from original dataset)
7. `ARENACAPACITY_homeTeam` - approximate arena capacity or amount of fans that can be seated for Home Team
8. `ARENACAPACITY_awayTeam` - approximate arena capacity or amount of fans that can be seated for Away Team
9. `YEARFOUNDED_homeTeam` - the year the Home Team was founded
10. `YEARFOUNDED_awayTeam` - the year the Away Team was founded
11. `CONFERENCE_homeTeam` - the conference of the Home Team (West=0 and East=1)
12. `CONFERENCE_awayTeam` - the conference of the Away Team (West=0 and East=1)
13. `G_homeTeam` - the amount of games in the season the Home Team has played up until the predicted game (0-82 games)
14. `G_awayTeam` - the amount of games in the season the Away Team has played up until the predicted game (0-82 games)
15. `W_homeTeam` - the amount of wins in the season the Home Team has won up until the predicted game
16. `W_awayTeam` - the amount of wins in the season the Away Team has won up until the predicted game
17. `L_homeTeam` - the amount of losses in the season the Home Team has lost up until the predicted game
18. `L_awayTeam` - the amount of losses in the season the Away Team has lost up until the predicted game
19. `W_PCT_homeTeam` - the win percentage for the current season of the Home Team up until the predicted game
20. `W_PCT_awayTeam` - the win percentage for the current season of the Away Team up until the predicted game

Complete list of features

21. `WEEKDAY` - the weekday number for the predicted game
22. `WEEKEND_GAME` - is this a weekend game? (1=True and 0=False)
23. `MONTH_NUM` - the month number for the predicted game (1 to 12)
24. `PLAYOFF_GAME` - is this a playoff game? (1=True and 0=False)
25. `HIST_PPG_homeTeam` - Home Team long-term average for points per game (2004-2019)
26. `HIST_PPG_awayTeam` - Away Team long-term average for points per game (2004-2019)
27. `HIST_FGpercent_homeTeam` - Home Team long-term average for field goal percent (2004-2019)
28. `HIST_FGpercent_awayTeam` - Away Team long-term average for field goal percent (2004-2019)
29. `HIST_FTpercent_homeTeam` - Home Team long-term average for free throw percent (2004-2019)
30. `HIST_FTpercent_awayTeam` - Away Team long-term average for free throw percent (2004-2019)
31. `HIST_FG3percent_homeTeam` - Home Team long-term average for 3-PT field goal percent (2004-2019)
32. `HIST_FG3percent_awayTeam` - Away Team long-term average for 3-PT field goal percent (2004-2019)
33. `HIST_APG_homeTeam` - Home Team long-term average for assists per game (2004-2019)
34. `HIST_APG_awayTeam` - Away Team long-term average for assists per game (2004-2019)
35. `HIST_REB_homeTeam` - Home Team long-term average for rebounds per game (2004-2019)
36. `HIST_REB_awayTeam` - Away Team long-term average for rebounds per game (2004-2019)
37. `DIFF_HIST_PPG` - Difference of `HIST_PPG_homeTeam` and `HIST_PPG_awayTeam`
38. `DIFF_HIST_FG` - Difference of `HIST_FGpercent_homeTeam` and `HIST_FGpercent_awayTeam`
39. `DIFF_HIST_FT` - Difference of `HIST_FTpercent_homeTeam` and `HIST_FTpercent_awayTeam`
40. `DIFF_HIST_FG3` - Difference of `HIST_FG3percent_homeTeam` and `HIST_FG3percent_awayTeam`

Complete list of features

41. `DIFF_HIST_APG` - Difference of `HIST_APG_homeTeam` and `HIST_APG_awayTeam`
42. `DIFF_HIST_REB` - Difference of `HIST_REB_homeTeam` and `HIST_REB_awayTeam`
43. `PPG19_homeTeam` - Home Team short-term average for points per game (2019)
44. `PPG19_awayTeam` - Away Team short-term average for points per game (2019)
45. `FGper19_homeTeam` - Home Team short-term average for field goal percent (2019)
46. `FGper19_awayTeam` - Away Team short-term average for field goal percent (2019)
47. `FTper19_homeTeam` - Home Team short-term average for free throw percent (2019)
48. `FTper19_awayTeam` - Away Team short-term average for free throw percent (2019)
49. `FG3per19_homeTeam` - Home Team short-term average for 3-PT field goal percent (2019)
50. `FG3per19_awayTeam` - Away Team short-term average for 3-PT field goal percent (2019)
51. `APG19_homeTeam` - Home Team short-term average for assists per game (2019)
52. `APG19_awayTeam` - Away Team short-term average for assists per game (2019)
53. `REB19_homeTeam` - Home Team short-term average for rebounds per game (2019)
54. `REB19_awayTeam` - Away Team short-term average for rebounds per game (2019)
55. `DIFF_PPG19` - Difference of `PPG19_homeTeam` and `PPG19_awayTeam`
56. `DIFF_FG19` - Difference of `FGper19_homeTeam` and `FGper19_awayTeam`
57. `DIFF_FT19` - Difference of `FTper19_homeTeam` and `FTper19_awayTeam`
58. `DIFF_FG319` - Difference of `FG3per19_homeTeam` and `FG3per19_awayTeam`
59. `DIFF_APG19` - Difference of `APG19_homeTeam` and `APG19_awayTeam`
60. `DIFF_REB19` - Difference of `REB19_homeTeam` and `REB19_awayTeam`

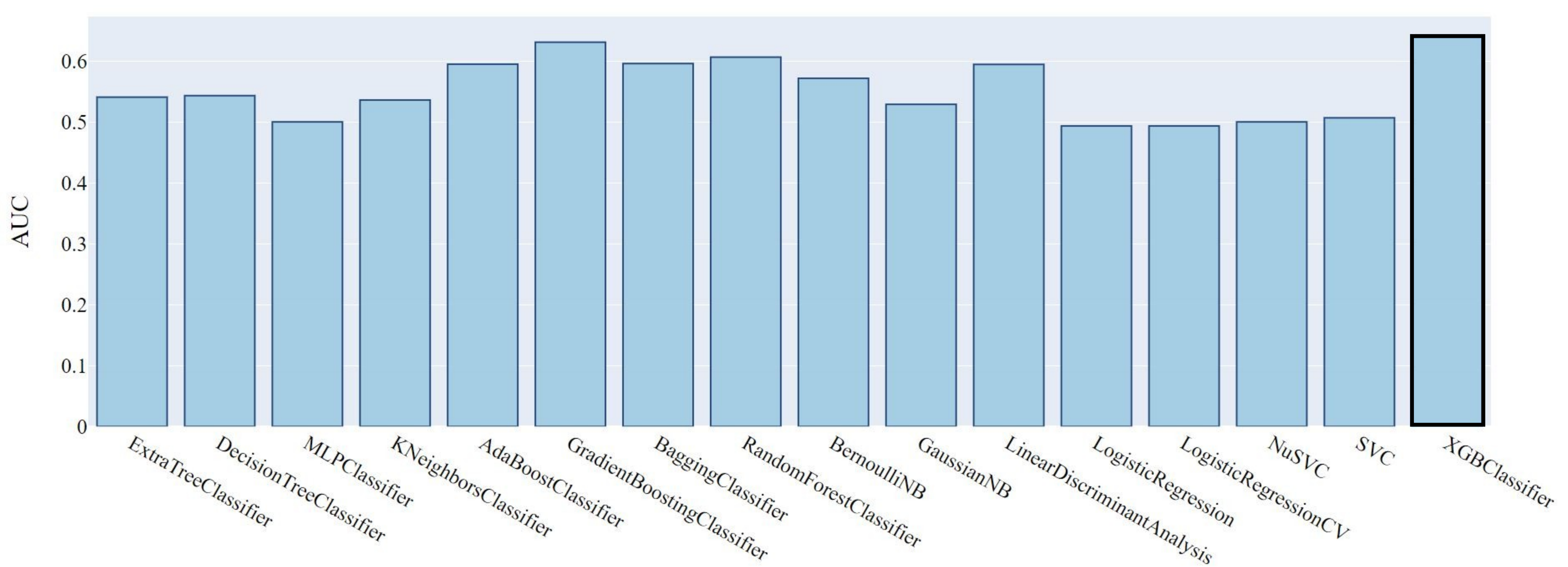


https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

ROC & AUC

Classification Models

- After engineering 60 features utilizing the available data, everything was inputted through 16 different classification schemes in Scikit learn.
- XGBClassifer was found to have the highest AUC ROC with 0.639
- The python code reads this output and selects XGBClassifer to be used later on



```
(.venv) brent@DESKTOP-896E3VE:~/NBA/NBAgames$ /home/brent/NBA/
NBA games project: Finished creating 60 features
NBA games project: Running all possible classification models
NBA games project: Finished run of all classification models
NBA games project: Select XGBClassifier as classifier
NBA games project: Beginning brute force......
```

Brute force combination for model selection

- Took so long!!
- At first, I ran Exhaustive Feature Selector function from mlxtend, however, upon some smaller testing, I calculated the code would take approximately 100-200 years to run on my pc.
- So, I found another tool Sequential Feature Selector, also from mlxtend, that took significantly less time

Sequential Feature Selector

$x^+ = \arg\max J(X_k + x), \text{ where } x \in Y - X_k$
$X_{k+1} = X_k + x^+$
$k = k + 1$

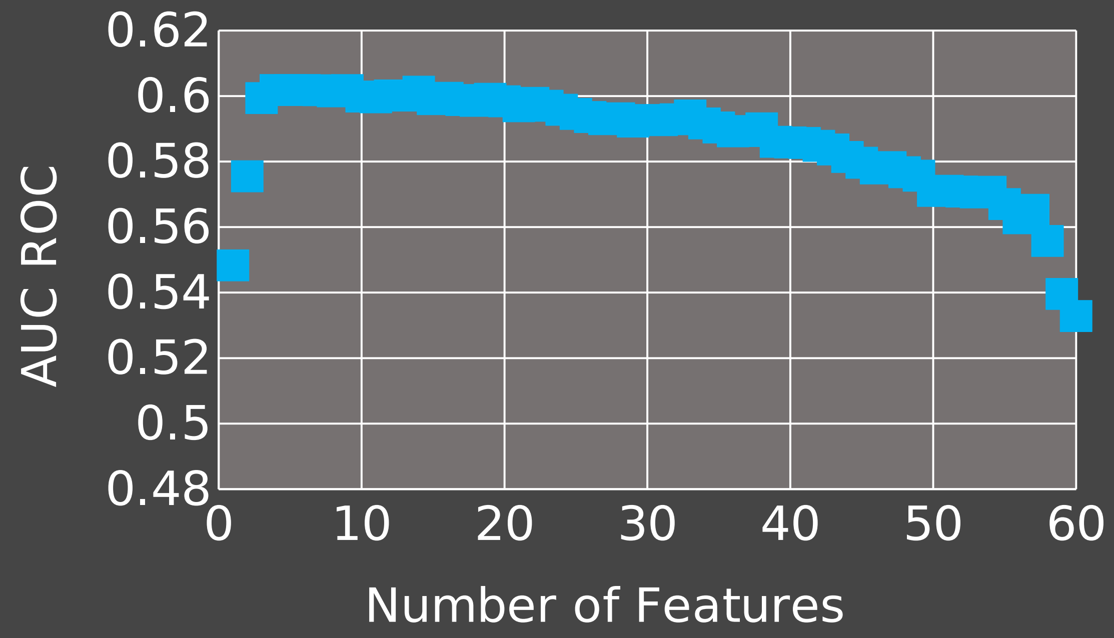
Go to Step 1

- in this step, we add an additional feature, x^+, to our feature subset X_k.
- x^+ is the feature that maximizes our criterion function, that is, the feature that is associated with the best classifier performance if it is added to X_k.
- We repeat this procedure until the termination criterion is satisfied.

Termination: $k = p$

- We add features from the feature subset X_k until the feature subset of size k contains the number of desired features p that we specified *a priori*.

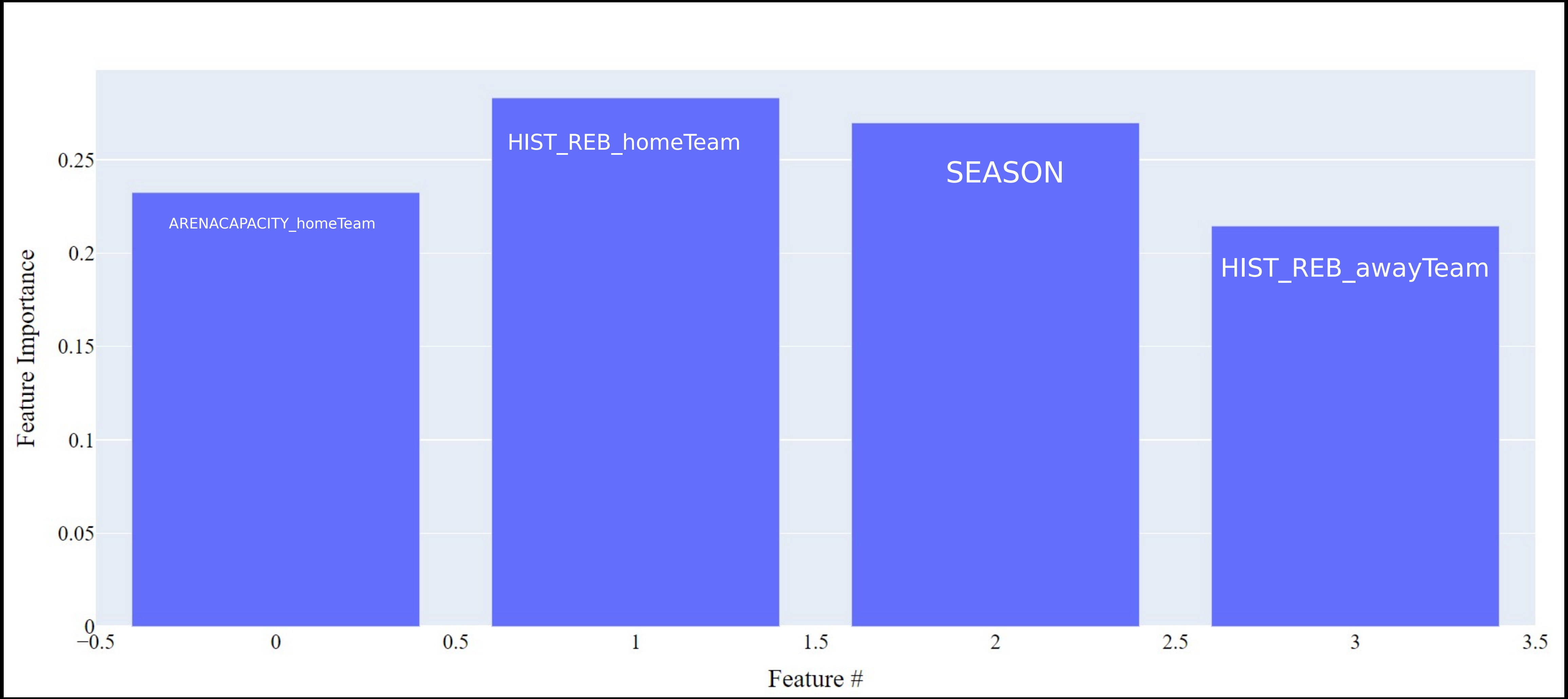
http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/



FEATURE_IDX	AVG_SCORE
(4, 6, 29, 35)	0.601848
(2, 4, 6, 29, 35)	0.601848
(2, 3, 4, 6, 29, 35)	0.601848
(2, 3, 4, 5, 6, 11, 27, 29, 35)	0.601834
(2, 3, 4, 5, 6, 29, 35)	0.60177

Selecting best combination of features

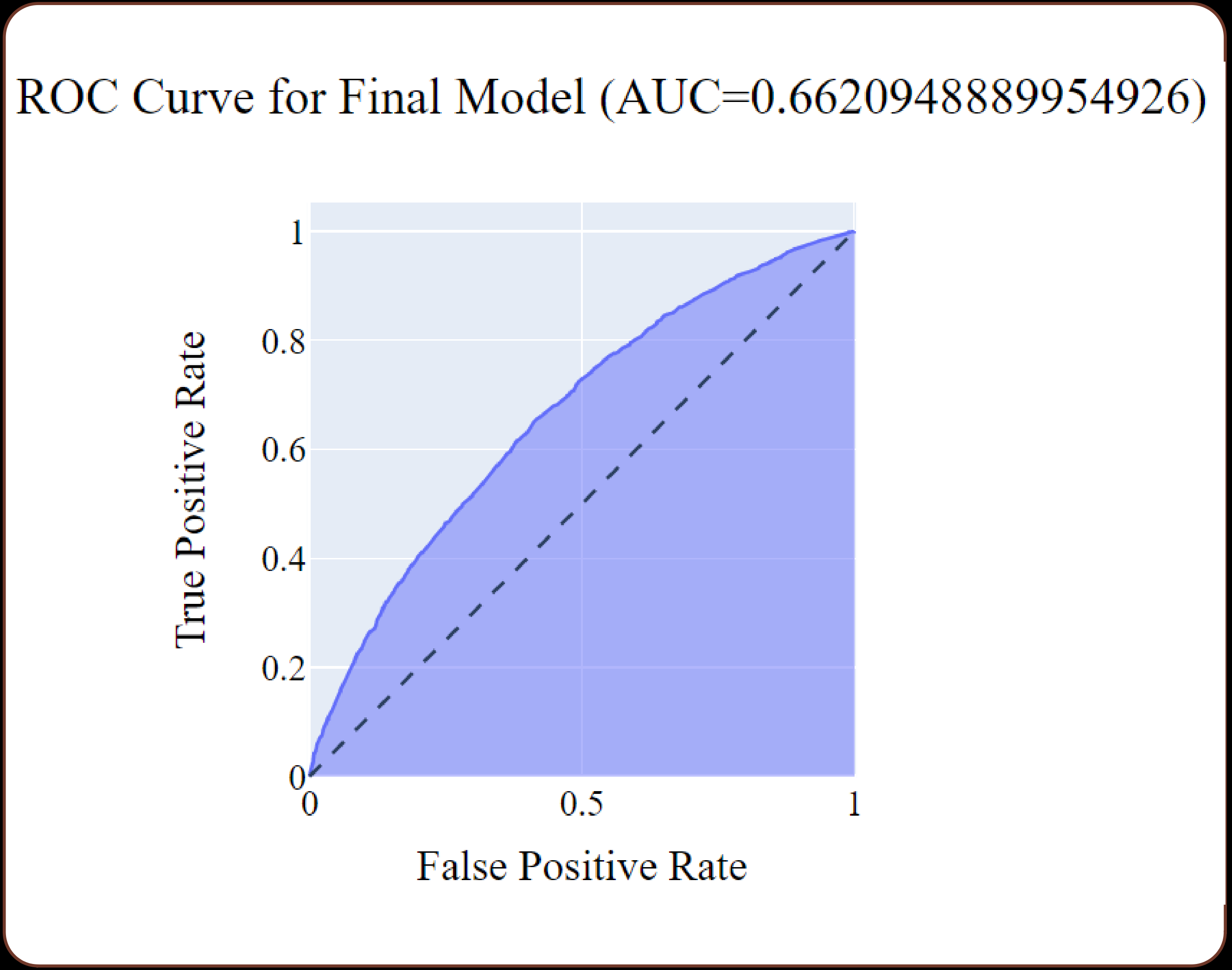
- Run time took about 30 minutes to successfully find best combination
- These indexes were then ran back through the code automatically to make plots of the final model with optimal features
- The advantage of this is, say I get brand new features for this dataset, I will not have to rewrite code as the algorithm will automatically pull the best model based on FEATURE_IDX



Final Feature importance

Final ROC curve

- Not much improvement. Still do not trust model.
- We were able to improve the **AUC score from 0.64 to 0.66** after using the optimal feature combination
- It is clear that more important features are missing from this model.
- **Still, we were able to show a repeatable methodology for finding optimal model to try!**



Discussion what worked what didn't

We were right about home court advantage helping determine the outcome, as well as the season-to-season difference being a key driver

However, a lot of our features were not very predictive, as the final model had 54 of our features removed.

How could we improve the model?

- Clearly this model needs better features to improve accuracy
- We would need to find an external database to tie into that could include other features. Some ideas could be:
  - Number of NBA all-stars on home and away team for the current season
  - Number of historical championships for the home and away teams
  - Number of players on max contract deals for current season
  - Financial cap space of teams for current season
  - Number of years with the same head coach

References

- Data = nba-games
- mlxtend = documentation
- NBA = https://www.nba.com/
- NBA analytics = https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116
- Measuring Performance: AUC (AUROC) = https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

Thank you!!