A STUDY ON THE NBA

CS 450 Final Project

ABSTRACT

The NBA is a 1.9 billion dollar industry, we want to shed some light on the possible outcomes of the ultimate prize the NBA championship and the playoffs

Sean Pinder, Brigham Tenney, Cory Swainston, Zach Newell

CS 450

Introduction:

https://www.basketball-reference.com

(Cory)

Some members of our team are avid NBA fans. We thought it would be interesting to use the wealth of available basketball statistics to predict something about player or team performance.

We considered several ideas including:

* Prediction of individual performance for second year players
* Prediction of NBA champions
* Prediction of MVP for each year

Finally we determined that we would try predicting the world champions given a team's stats for the season. Beyond just predicting whether or not a given team would win the title, we wanted to predict how far they would advance in the playoffs in a given year using a numeric ranking from 0-5, 0 being a lottery year and 5 being the championship.

This problem was interesting to us because with the playoffs coming up, we can use this year's data to predict the results with reasonable accuracy. More importantly, we hoped that we could find a way to isolate key statistics that lead to playoff success in NBA teams.

We eventually found....... To Be Continued

Data Preparation:

(Brigham)

The data set that we choose included teams per game data from multiple categories. This data set provided many different issues and advantages.

Because NBA statistics are so well documented, we just needed to grab the data from basketball-reference.com as a csv file. Also, the NBA records many different types of data from each game, so we were able to have 46 attributes to use in our analysis.

Unfortunately, there are only 72 years of NBA history and many of those years different attributes were recorded and the playoff system we were hoping to predict was changed over that time. As such we only had 34 years of reliable data to use. 34 instants of data were far too few to reliably use so we broke each year's data down by team. So, each team of each year was a new instant of data. We now had 983 instants of data which was a reasonable number to use for the various algorithms. But this meant that normal scaling methods wouldn't work as it we would want it too.

One of the issues with comparing things year to year is how trends in the NBA influence how the game is played and won therefore what is important one year is less important the next. As such normalizing all of our data would introduce all of those inconsistences as well, so we needed to normalize the data by year then randomize and split it. Because of the complexity of this problem we choose to normalize the data in excel rather than attempting to do so with python. We broke the data up by year and used z-score normalization on each attribute from team to team. This way every entry of data was scaled according to how dominate that team was at that attribute compared to the other teams in the league for that year.

This method wouldn't entirely remove the inconsistencies from year to year or the inherent noise of sports data, but it would help to contain that noise some.

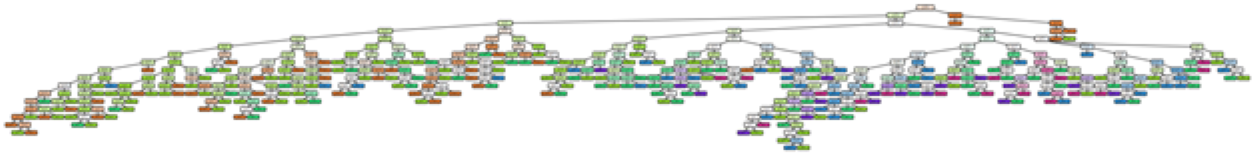
Mining / Learning from the data:

(Sean)

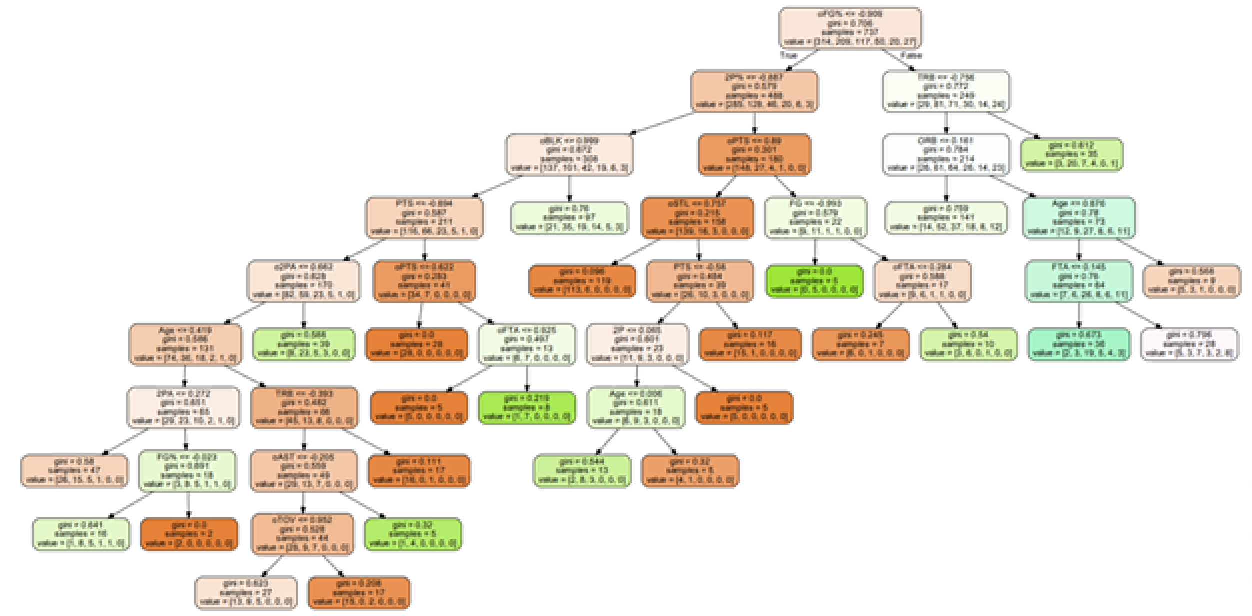
We tried a number of different algorithms with the aim of choosing one or two different models with a high enough accuracy to be respectable. Among the algorithms we tried were a neural network, naive bayes, K-nearest neighbors, support vector machine, decision tree, and multi-layer perceptron regression. Ultimately the one that performed the best was our decision tree; the other advantage of using a decision tree being that we were able to visualize our model in a way that might be more difficult with other algorithms.

Results:

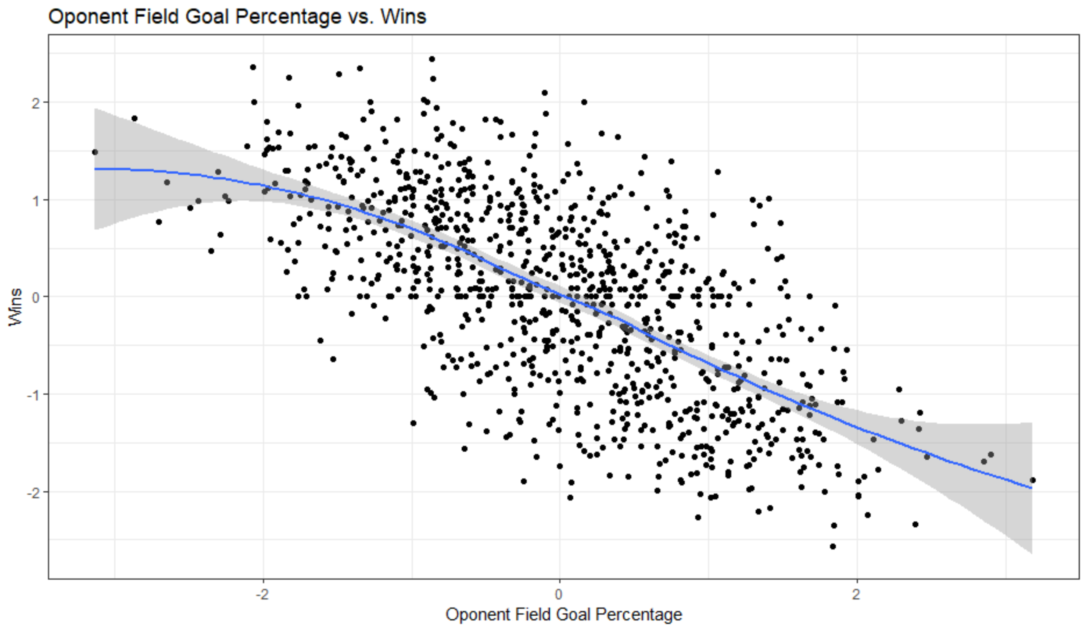
We went through several iterations of our decision tree. The first one was quite enormous; a visualization can be be found below.

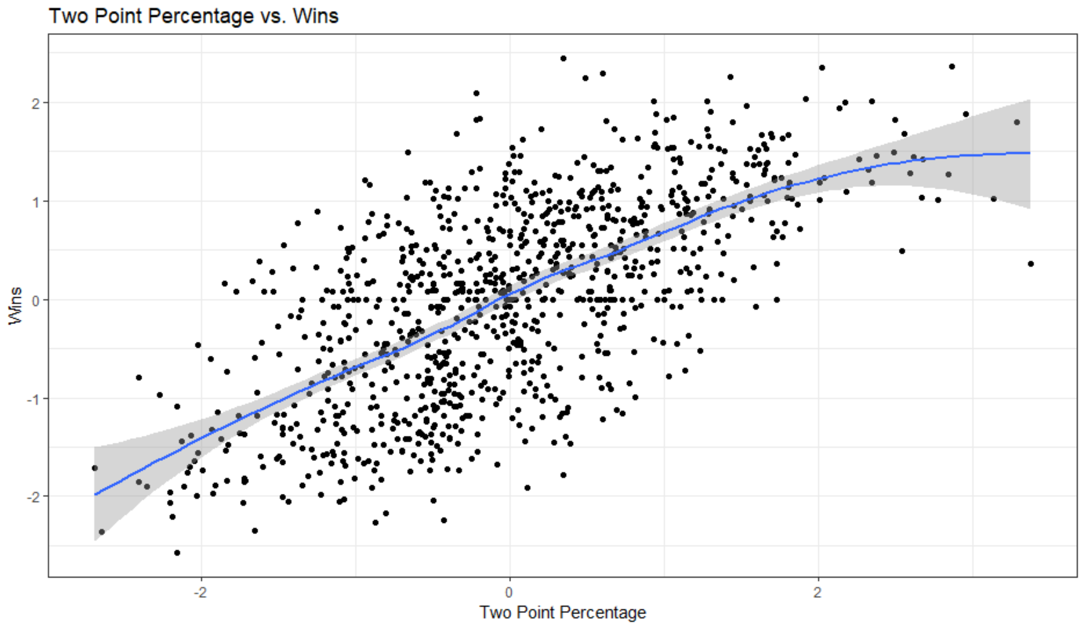


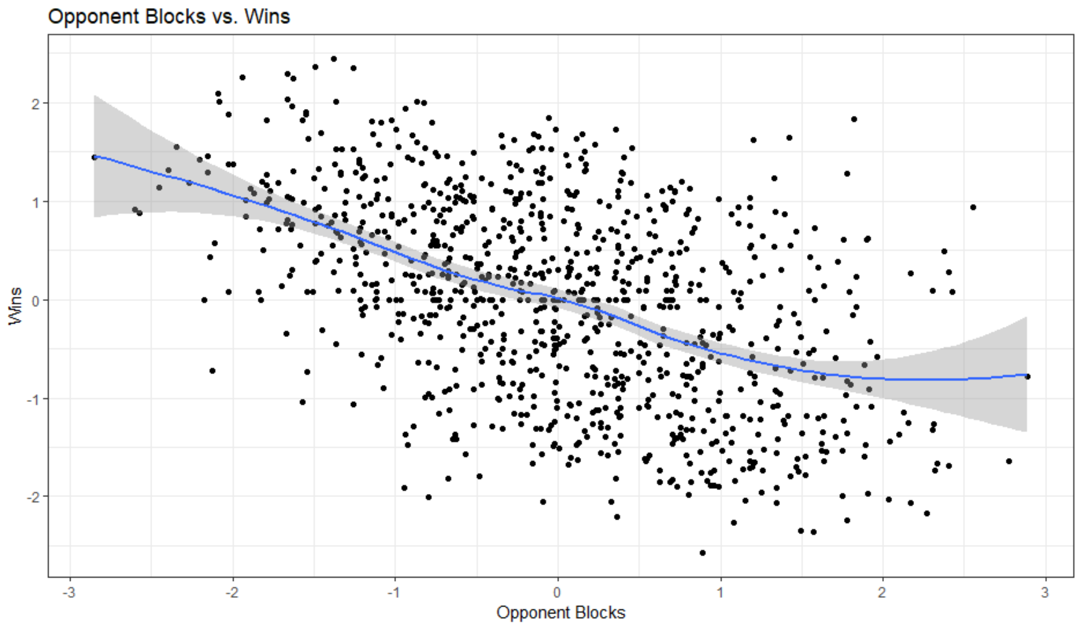
We found two problems with this model: first was that it was extremely tall so we suspected that it might be overfit to noise in our data, and the second was that the two most important attributes that the decision tree split on were wins and losses. We felt like those two attributes were too obvious for determining whether a team would win a championship or not so we pruned the height of the tree and removed wins and losses from our features. Our refined model is visualized below.



In this model we found that the three most important features for deciding whether a team would win championship were Opponent Field Goal Percentage, Two Point Percentage, and Opponent Blocks. We've also included with our results plots that show how each of these is correlated with a team's wins.


Oponent Field Goal Percentage vs. Wins


Two Point Percentage vs. Wins



(Include predictions on this year's playoff results including some sort of percentages)

Conclusions:

(NOT TILL LATER)

(Include which algorithm achieved the best results)

(Include how we determined percentages)

Lessons Learned:

(Zach)

We learned quite a few lessons regarding machine learning algorithms, types of data (and their uses), and learning strategies.

As mentioned above, we trained quite a few different algorithms in order to find the best fit for our application. We learned that this is actually a good strategy to finding the best algorithm for a given data set. Through our experimentations we found the best fit for our data.

Next we learned how different types of data yield different results. We trained and tested on raw numeric, normalized, and we even tried binned data for learning. This helped concretize our textbook knowledge with real life applications. Each type of data mentioned has their own use but we found that normalized data worked the best for our learning purposes.

Finally we picked up some new strategies when it comes to machine learning. Collaboration is first and foremost on this list. As a group we were able to pool our knowledge to accomplish our task.

Another strategy we found worked well has to do with the type of problem we were facing. It was important for us not to set unrealistic expectations for our predictions. As with any sport prediction there is no perfect recipe for winning. We definitely couldn’t expect that our predictions were going to get high accuracy scores, it all depends on the kind of prediction at hand.