ECO 395 Project: Taylor Neal

1) Abstract

The goal of this analysis is to determine what regular season NBA team statistics are most useful in determining which teams win in playoff match-ups. With these NBA team statistics identified, we seek to build a probabilistic model utilizing our machine learning toolbox. Leveraging a probabilistic model for playoff game outcomes will allow us to simulate a large number of potential brackets and estimate percentage chances for each NBA franchise winning the championship in a given year.

2) Introduction

In an age of increasing access to data of all kinds, sports analytics data is no exception. The primary challenge these days is figuring out what variables and measures to actually utilize when modeling an outcome of interest. There are many reasons we might have interest in using available statistics to predict outcomes in NBA playoff games. There could be incentive to place bets when Vegas odds differ substantially from a model one has confidence in. Or, it might just be a matter of ensuring your optimism levels for a favorite team are appropriately set (it is generally good to avoid crushing disappointments when possible).

Given the vast number of professional basketball advanced metrics, machine learning tools can be an effective way of sorting through what is most important to our outcome of interest (which team will win the game). This project will tackle the selection / regularization problem by settling on the best machine learning process based on cross validation. Cross validation of our model process will also allow us to determine an optimal number of years of historical NBA playoff game results to include in this analysis.

Attempts were made at utilizing principal components analysis (reducing the dimensional of our many available statistics) and random forests (collections of many tree models with each having access to a limited, random subset of the data). However, an approach utilizing cross validated logistic LASSO regression (a penalized version of a logistic regression at an optimized penalty value) was determined to be the best at modeling our probability outcome of interest, so the bulk of this report will focus on results reached through that analysis.

Upon constructing our probability model for win percentage in a given NBA playoff match-up, a Monte Carlo simulation will give us insight into what we would expect championship probabilities to be for a given year's bracket (we will show predictions of the current 2022 playoffs which are in process and 2021 to compare with the known result from last year).

For data and modeling simplicity, we will make use of team statistics from the relevant completed regular season in order to predict the probability outcomes for playoff games in a given year. While it would likely be more accurate to model team performance based on individual players to account for injuries and the constant flux of NBA rosters (similar to how fivethirtyeight.com does), that is beyond the scope of this project for the time being.

3) Methods

Our data was obtained from basketball-reference.com. The 15 most recent years of historical data were incorporated into this analysis (although only the 10 most recent years of data excluding the Covid/bubble playoff year were included in our final modeling approach). Separate data sets for team shooting, average per game stats, advanced metrics, detailed standings, team vs team records and additional opponent per game / shooting data were all considered. These datasets were available for the individual years in question. And, thus, a fair amount of data cleaning and preparation went into combining said data into a usable historical

record of results and accompanying statistics. In the final model formulation, advanced metrics and detailed standings carry the load of modeling our outcomes of interest.

Cross validation over 20 random iterations of our 10 years of historical data were utilized to test and determine the best modeling approaches. Two historical years were withheld from the training set in each case to test various approaches against each other. This project found that logistic LASSO regression outperformed random forests on average and offered better interpretability over the construction of a PCA analysis. Additionally, excluding the pandemic year when the playoffs were played in a bubble without any fans improved the stability of our approach.

Many engineered factors were attempted to incorporate data more efficiently into our modeling process. For instance, one engineered factor that remains in the final model is a difference in average age of the two teams matched up with each other. Other interactions such as various interaction terms between one team's offense and the other's defense were not found to be impactful. Our dependent variables were normalized prior to model builing in order to have more comparable coefficients.

In the upcoming section, we will dive into the results of the cross validated LASSO regression approach and see what it implies for both future and past predictions of overall champions. We will plot the LASSO solution as it varied with lambda and examine the ROC curve for the selected regression (chosen by maximizing the area under the ROC curve). We can then begin to look into our resulting coefficients to answer the question of which regular season statistics carry the most weight in predicting.

4) Results

Despite the vast number of statistics and advanced metrics available. We see in figure 1 (below) that only a small number of regression coefficients make it into our final model. At the maximum area under the ROC curve, we make use of 7 variables for predicting whether the home team will win their matchup (modeled this way with an intercept in order to accout for home field advantage).

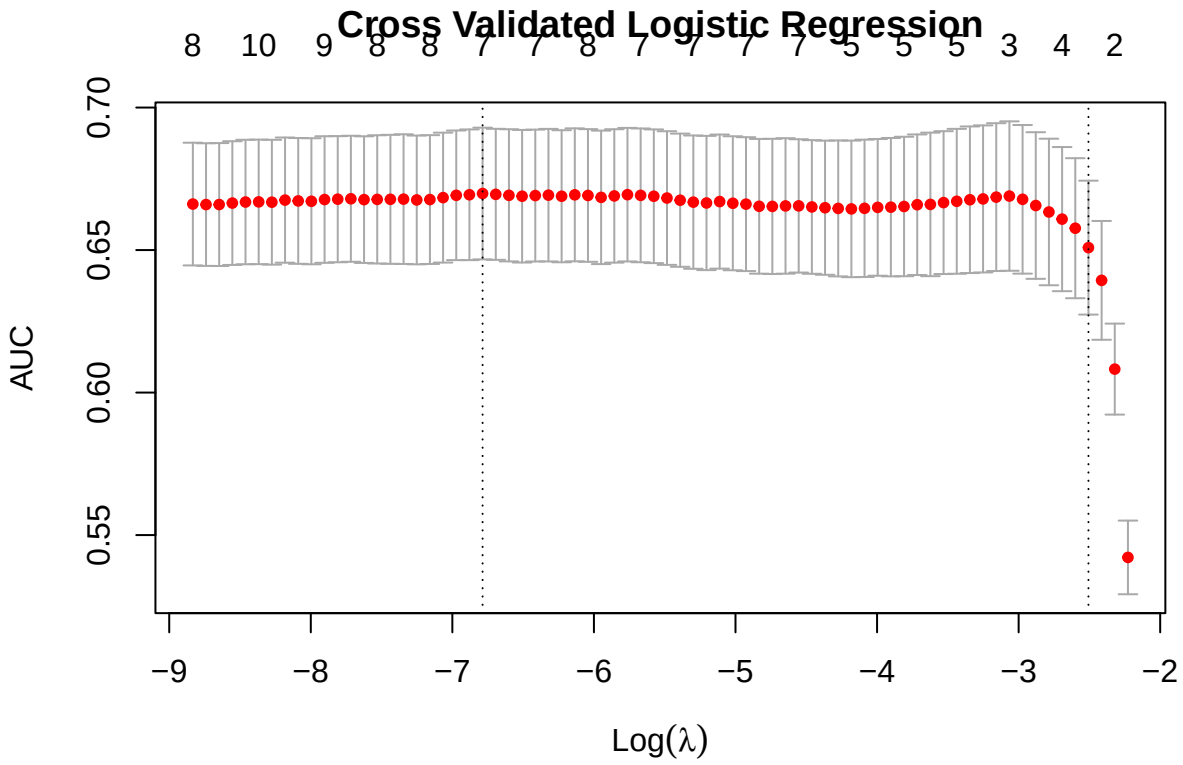

Figure 1: Cross validated logistic LASSO regression - optimizing for area under the curve.

Figure 2 (below), highlights the actual ROC curve of our optimized solution. Given the inherent randomness of a game like professional basketball, it is not entirely surprising that our best curve does not incorporate more area on the plot.

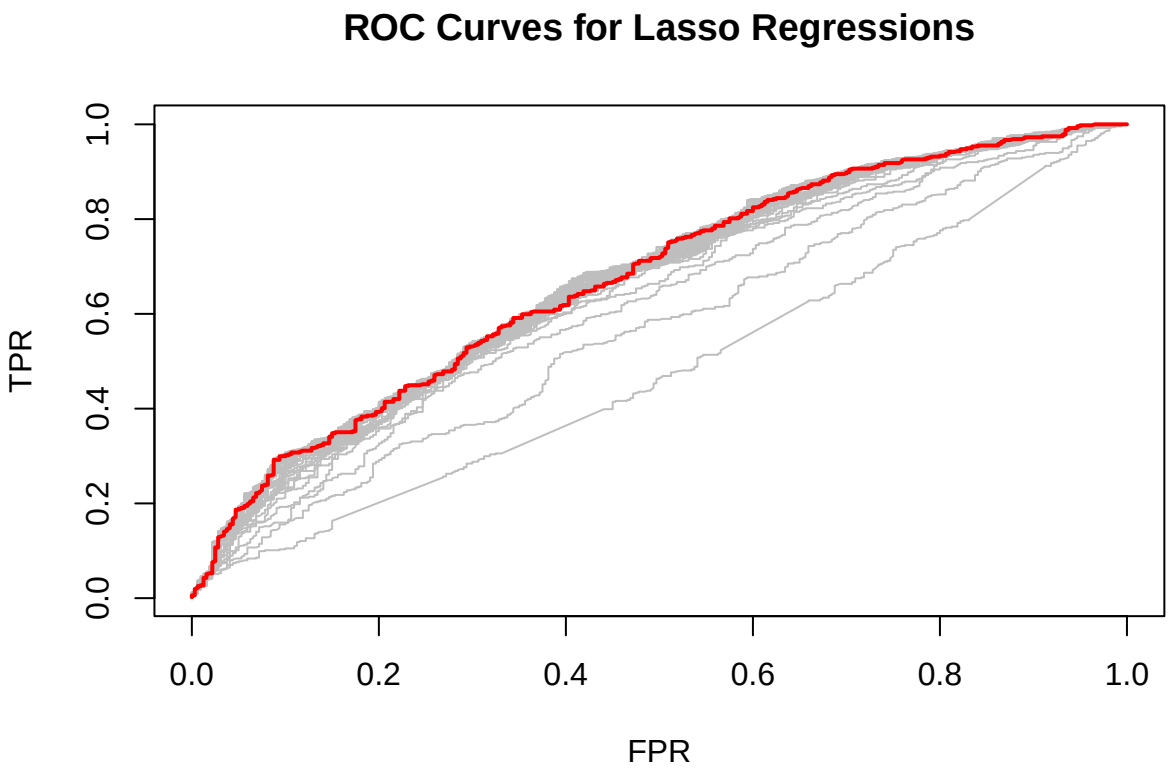

Figure 2: ROC curve for our optimized solution is highlighted in red.

In table 1 (below), we show the coefficients for our optimized solution (max AUC) and the most penalized solution within one standard error of the optimal. Note that here variables ending in .x are statistics for the home team and variables ending in .y are statistics for the away team. It is curious that there are different variables showing as important for the home vs away distinction. But this was consistent over our cross validation process. And the high correlation between many of this variables likely makes it a near equivalent choice between choosing what would be of most importance for one side to come out with a win. Note that our model is for the outcome of a home win so positive coeffients make a home win more likely and negative coefficients make a home win less likely.

Table 1: Coefficients for our Lasso regression where area under the ROC curve is maximized and the most penalized regression within one standard error of the optimum solution.

	Max AUC	One se
(Intercept)	0.532	0.476
SRS.x	0.378	0.000
ORtg.x	-0.247	0.000
DRtg.x	0.000	0.000
TS..x	0.370	0.000
MOV.x	0.000	0.091
SRS.y	0.000	0.000
ORtg.y	-0.022	0.000
DRtg.y	0.000	0.000
TS..y	-0.291	-0.020
MOV.y	-0.275	-0.089
age_diff	0.072	0.000

Out Monte Carlo simulations were run 1,000 times in order to get a sizable sample of potential tournament results. We find that the model estimates Phoenix (27.6%) and Golden State (22.4%), which are both teams in the western conference, to have the best shot at a championship this year.

Table 2: Monte Carlo results for 2022 playoffs.

NBA_Champion	Win_percentage
PHX-2022	0.276
GSW-2022	0.224
BOS-2022	0.179
MIA-2022	0.151
UTA-2022	0.084
MIL-2022	0.026
PHI-2022	0.025
DEN-2022	0.024
DAL-2022	0.006
BKN-2022	0.002
MIN-2022	0.002
MEM-2022	0.001

East_Champion	Win_percentage
MIA-2022	0.392
BOS-2022	0.360
PHI-2022	0.131
MIL-2022	0.090
BKN-2022	0.013
ATL-2022	0.008
CHI-2022	0.006

West_Champion	Win_percentage
PHX-2022	0.406
GSW-2022	0.331
UTA-2022	0.166
DEN-2022	0.068
DAL-2022	0.013
MIN-2022	0.009
MEM-2022	0.006
NOP-2022	0.001

The table below (table 3) allows us to look at the predicted results prior to the playoffs from last year (which notably featured many wild games and series upsets). Utah (which has a phenomenal statistical regular season) was a huge favorite to win the title (43%) based on this model but they ended up losing in the second round.

Table 3: Monte Carlo results for 2021 playoffs.

NBA_Champion	Win_percentage
UTA-2021	0.430
BKN-2021	0.199
PHX-2021	0.116
LAC-2021	0.102
PHI-2021	0.075
MIL-2021	0.058
DEN-2021	0.018
DAL-2021	0.001
LAL-2021	0.001

East_Champion	Win_percentage
BKN-2021	0.487
PHI-2021	0.285
MIL-2021	0.200
ATL-2021	0.016
MIA-2021	0.006
BOS-2021	0.003
NYK-2021	0.002
WAS-2021	0.001

West_Champion	Win_percentage
UTA-2021	0.554
PHX-2021	0.207
LAC-2021	0.177
DEN-2021	0.055
LAL-2021	0.005
DAL-2021	0.001
POR-2021	0.001

5) Conclusion

Based on our results above, it appears that the advanced stats best at predicting playoff match-ups are: SRS (simple rating system - which account for point differential and strength of opponents), ORtg (offensive rating - points per 100 possessions), true shooting percentage (TS), average margin of victory (MOV) and average age differential. Most of the signs appear to make logical sense (with the exception of the coefficient on ORtg of the home team being negative). And it is interesting to note that an older average team age is adding to the likelihood of victory (likely picking up on a component of playoff experience).

Upon inspection of our optimized ROC plot and the less than impressive predictions for the 2021 playoffs (with Utah being showing up as too much of a favorite that year to be reasonable). Utilizing regular season NBA stats for the same year does not appear to do a great job of modeling the outcomes in the NBA playoffs. This might be due to the inherent randomness in professional basketball, but it is likely that a more detailed player-centric approach to modeling this problem could yield greater results in the future.