ORIE 4740: Final Project

Kenneth Bogart (kmb295) Benjamin Hwang (bwh57) Emily Lutz (eel52)

Project Motivation

In this document, we discuss our findings based on NBA team data from 2003 to 2015 (basketball-reference.com). Using these twelve seasons of data and the help of RStudio, we attempt to answer a few different questions: what do teams that make the playoffs have in common, can we predict which teams will make the playoffs, and which team will win the championship? We will attempt to answer these questions primarily by using logistic regression, decision trees, and principle component analysis.

In the growing and popular field of sports analysis and predictions, studying basketball is an interesting application of our statistical knowledge in a field that is not typically covered in a classroom setting.

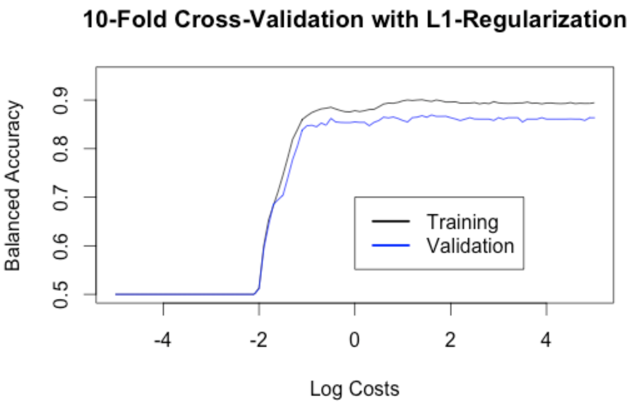
Playoff Classification: What does it take to make the playoffs in the NBA?

Using the features below, we wanted to determine if we could classify teams that make the playoffs. Features such as points scored per game and wins per game played were left out, as they do not allow for an interesting model. The teams with the greatest amount of wins make the playoffs automatically (ignoring the fact that making the playoffs or not depends on if other teams make the playoffs within each division), and typically these are teams with the most points per game. Rather than directly inputting points per game, field goals per game were used in the data.

Features:

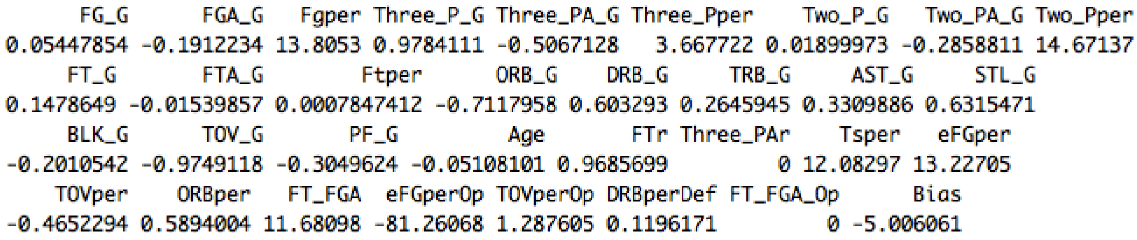
Field goals per game
Field goals attempted per game
Field goal percentage
3-point field goals per game
3-point field goals attempted per game
3-point field goal percentage
2-point field goals per game
2-point field goals attempted per game
2-point field goal percentage
3-point attempt rate
Free throw attempt rate
True shooting percentage (shooting efficiency)
Offensive rebound percentage
Free throws made per field goal attempt
Opponent free throws made per field goal attempt
Defensive rebound percentage
Free throws made per game
Free throws attempted per game
Free throw percentage
Offensive rebounds per game
Defensive rebounds per game
Total rebounds per game
Assists per game
Steals per game
Blocks per game
Turnovers per game
Personal fouls per game
Age
Effective field goal percentage
Turnover percentage
Opponent effective field goal percentage
Opponent turnover percentage

With the goal of using L1-regularization via LiblineaR, we selected the optimal model through 10-fold cross-validation on a sample of 300 of the 359 observations. We found the optimal cost to vary depending on the sample, but the resulting balanced accuracies were always very similar.



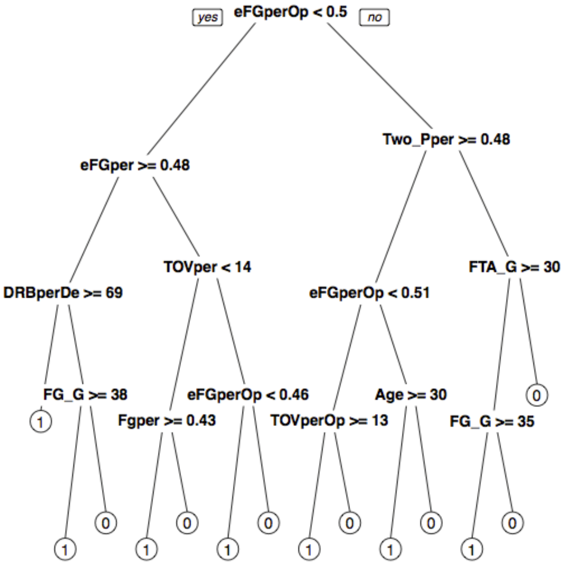
We then used the optimal cost to run the model on the remaining test data so that we could observe the selected features. Generally, the features that were weighted zero were the 3-point attempt rate (percentage of field goal attempts from the 3-point range) and opponent free throws per field goal attempt.

Running a generalized linear model on the selected features showed that the most significant (in terms of p-value: at the .001 level) feature in classifying playoff teams is opponent's effective field goal percentage. Effective field goal percentage is a statistic adjusted for the fact that a 3-point field goal is worth one more point than a 2-point field goal. Opponent turnover percentage was the next most significant feature (at the .05 level), where turnover percentage is an estimate of turnovers committed per 100 plays. Examining the weights of the L1-regularization selected features shows that opponent effective field goal percentage is the most heavily weighted (very negative) percentage-feature. It is significantly more heavily weighted than a team's own effective field goal percentage, which certainly says a lot about the importance of good defense in making the playoffs. Other notable significant attributes include steals per game and average age of the team.



```
      FG_G       FGA_G     Fgper Three_P_G Three_PA_G Three_Pper     Two_P_G    Two_PA_G Two_Pper
0.05447854 -0.1912234 13.8053 0.9784111 -0.5067128   3.667722 0.01899973 -0.2858811 14.67137
      FT_G        FTA_G         Ftper       ORB_G     DRB_G      TRB_G      AST_G      STL_G
0.1478649 -0.01539857 0.0007847412 -0.7117958 0.603293 0.2645945 0.3309886 0.6315471
     BLK_G       TOV_G        PF_G          Age        FTr Three_PAr     Tsper    eFGper
-0.2010542 -0.9749118 -0.3049624 -0.05108101 0.9685699         0 12.08297 13.22705
    TOVper     ORBper    FT_FGA   eFGperOp TOVperOp DRBperDef FT_FGA_Op       Bias
-0.4652294 0.5894004 11.68098 -81.26068 1.287605 0.1196171         0 -5.006061
```

After 100 resamples of the observations (without replacement), retraining and retesting, the average accuracy of the LiblineaR model came to approximately 88.5%. This is a strong classifier, because it performs well without even considering division and conference structure of the NBA. We also compared this accuracy to that of a generalized linear model using only the features opponent effective field goal percentage and opponent turnover percentage. It resulted in an average balanced accuracy of 70.8%.

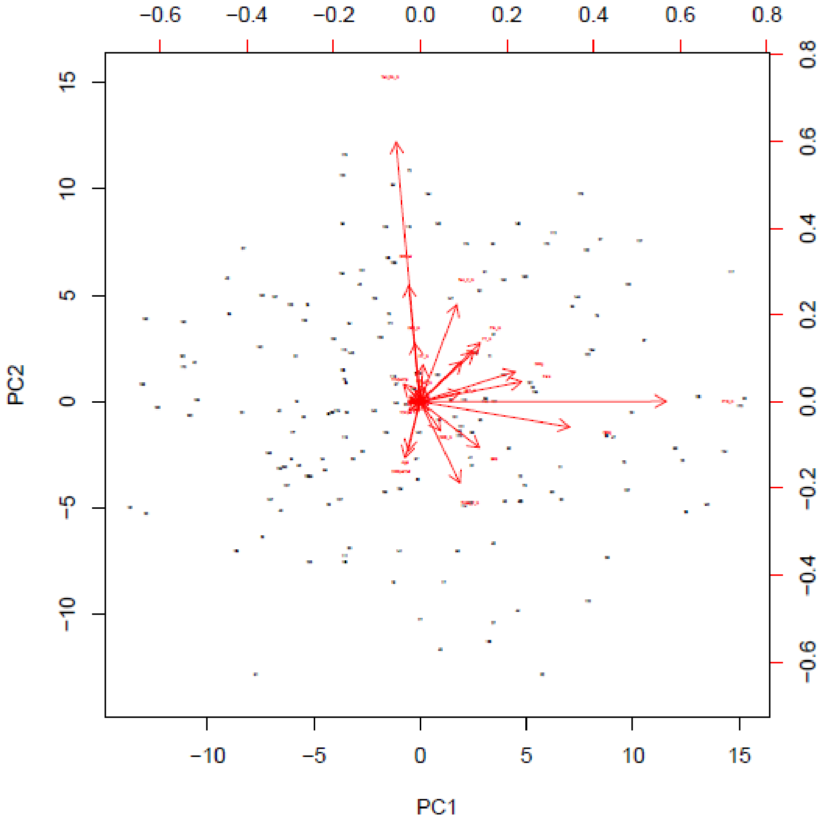


We next built a decision tree model, utilizing cross-validation once again to find an optimal depth of four. The average of the balanced accuracies of 100 of these trees was only 74.3%, but once again it is evident that opponent's effective field goal percentage is key to classifying playoff teams, as shown in the figure to the left. However, a model with a depth of one, sampled on the same exact data at the same time, was able to obtain an accuracy of 74.1%. The single attribute split on was of course still opponent's effective field goal percentage.

The randomForest package was also used to create an ensemble of decision trees. Doing so improved balanced accuracy to approximately 85%. This occurs because the variance of the forest is smaller than that of an individual tree by a factor of 500 (the number of trees in the forest).

We have shown that it is possible to classify playoff teams relatively well. We believe that we could predict with similar accuracy which teams will make the playoffs if we had data from incomplete seasons (perhaps with fewer than half of the games played).

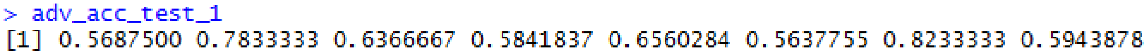
Championship Classification: What does it take to win the NBA championship?
Once playoff seeding is set (or predicted from our model or others) we wanted to see how we could then take that set of teams and predict playoff performance. We used ordered logistic regression and created a “finish” covariate based on how teams finished. If a team won the championship they would be assigned a value of 1, while runners up received a 2. If they lost in their conference semi-final they received a 3. Losing in earlier rounds results in a 4 or 5. We approached this modeling task by using the function polr within the MASS package. We quickly found issues with regards to the algorithm not being able to converge or find a suitable starting value. To alleviate this issue, we worked by first checking the PCA biplot of our data.



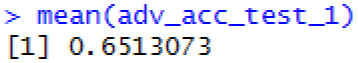
Here we can see that some of the covariates appear to be redundant, so we truncated several of those covariates to try and alleviate the issues with our model. For example three point field goal attempts was redundantly explaining variance associated with points per game.

Once we trimmed our model, the issue became less frequent, but issues still arose when trying to get an average accuracy. We found that the training set sampled made a huge difference in terms of accuracy and it also resulted in our model breaking on several iterations. We think that this high variation in data is most likely due to the small amount of data that we possess. Out of the 176 observations (teams that made the playoffs) we have in our data there are only 11 instances of teams that finished 1. Due to this small percentage, it is likely that when sampling randomly without replacement, very few instances of 1 and 2s made it into the training data. This heavily affects performance, and as a future test, finding a way to set the proportion of 1s in a sample would fix this issue.

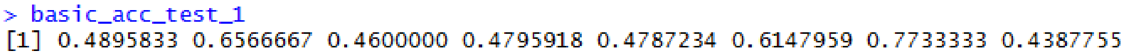
Regardless of these errors, we were able to at least collect some accuracies before the model would randomly break. The two main models which we tried consisted of testing on two separate sets of covariates. In our data, our features can be defined as simple statistics (observed phenomena that are tracked during a game) and advanced statistics (which uses simple statistics to create more complex features). Upon running both models for 10 trials (of which 8 values were able to be extracted) we found not surprisingly that the advanced statistics model on average did better at predicting a champion (finish of one).



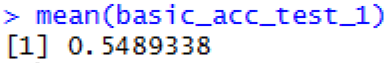
```
> adv_acc_test_1
[1] 0.5687500 0.7833333 0.6366667 0.5841837 0.6560284 0.5637755 0.8233333 0.5943878
```



```
> mean(adv_acc_test_1)
[1] 0.6513073
```

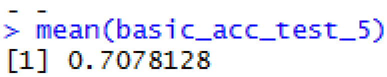


```
> basic_acc_test_1
[1] 0.4895833 0.6566667 0.4600000 0.4795918 0.4787234 0.6147959 0.7733333 0.4387755
```

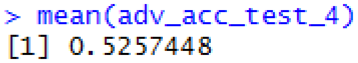


```
> mean(basic_acc_test_1)
[1] 0.5489338
```

Note that there is a significant amount of variability in the accuracies most likely due to the different training sets. Interestingly enough however, for the rest of the predicted playoff finishes, the models switch between which is better. For example the average accuracy for the 5th place finish is worse for the advanced model.



```
> mean(basic_acc_test_5)
[1] 0.7078128
```



```
> mean(adv_acc_test_4)
[1] 0.5257448
```

Using both models and the current NBA playoffs (which are still on-going) we decided to see which team might win the championship. Using our advanced model first, we actually got a pretty exciting result. It predicted the Golden State Warriors to be this season's champion -- they are not only currently in the Western Conference finals (equivalent of the semifinals of a tournament) but are considered heavy favorites to win the title! We decided to test the model without inputting the finishes of teams that have already been eliminated to see how well our model can predict other positions as well.

Team	Finish	Advanced Model Prediction	Basic Model Prediction
Atlanta Hawks	TBA	4	1
Boston Celtics	5	5	5
Brooklyn Nets	5	5	5
Chicago Bulls	4	4	5
Cleveland Cavaliers	TBA	5	4
Dallas Mavericks	5	5	5
Golden State Warriors	TBA	1	3
Houston Rockets	TBA	5	4
Los Angeles Clippers	4	4	4
Memphis Grizzlies	4	5	3
Milwaukee Bucks	5	5	5
New Orleans Pelicans	5	5	5
Portland Trail Blazers	5	5	5
San Antonio Spurs	5	3	4
Toronto Raptors	5	4	5
Washington Wizards	4	5	5

In the table above is a summary of what both models outputted. As mentioned before, the basic model seemed to do better determining which teams would lose earlier in the playoffs. It was fairly accurate at predicting the teams that would indeed go on to lose in the first round. Another cool prediction made by the basic model was that the Atlanta Hawks will win the championship. They aren't necessarily the favorites, but they are the number one seed in the Eastern conference and are also still in the running.

We think that these models have shown that there might be some predictive power to how the regular season team statistics might influence performance in the playoffs. Our accuracies were not incredible, but again we saw that with potentially more data to utilize there would be more championship instances to train on. Another improvement mentioned would be class balancing which could be explored further to gain an accuracy boost.

As a finishing comment, there is a lot of randomness that occurs in the NBA. Players get injured, or teams make in season-changes that alter their performance within a season, so just the fact that we were able to predict something feasible (although it's possible we are just getting lucky) makes us pretty happy.