Thomas Fratantoni, A15940388

Cristian Jarquin, A14489201

Prof. Mukamel

4 June 2021

Final Report: NBA Playoff Predictor

**Introduction**

The motivation for our project was inspired by our love of basketball. We initially wanted to be able to predict whether an NBA team would win or lose any particular game. We had originally planned to see which predictors would be most impactful when trying to predict a game’s outcome, and use them in a model to predict the winner. However, we made some changes to our project plan since the dataset we found had overall statistics of the season as a whole instead of statistics of individual games. Our new goal is to be able to predict this year’s NBA championship winner with the data from the regular season. Being able to predict the champions is significant for people placing bets since this will give a probability of a team’s chance to win which will give an edge over the house and other bettors. The “Team Per Game Stats'' data sets we used came from a website called basketball-reference.com (<https://www.basketball-reference.com/leagues/NBA_2018.html>) which had data on each team’s average predictor in one game. We used four datasets from the 2017-2020 regular seasons. Within each dataset was information on field goal percentage, 3-pt percentage, 2-pt percentage, rebounds, steals, etc. Each statistic had 30 observations, which is the total number of teams in the NBA. The datasets all had 25 predictors for the regular season which included a 3-pt field goal average, 2-pt field goal average, average number of assists, average number of rebounds, average number of field goals, and more. We hypothesized that using all of the predictors from four different years we would successfully be able to predict the current year's championship winner.

**Methods**

The initial preparation of the data consisted of removing predictors such as 3-pt percentage, 2-pt percentage, free throw percentage, minutes played, and total rebounds. We got rid of all the predictors that were percentages since there were already predictors that contained the same information. Having multiple similar predictors would be superfluous since they provide the same data. Another minor adjustment was to get rid of asterisks that were appended to team names to indicate that they made it to the playoffs. In addition, we added a column to indicate whether or not the team won a championship. Each team was assigned a 0 if they didn’t win and a 1 if they did.

We decided to use a logistic regression model for this project, using LogisticRegression from the sklearn package. Since the prediction we were looking for was a categorical variable, logistic regression was well-suited for this problem. Additionally, if we wanted to look at the relative probability estimates, logistic regression could provide those values as well. We knew that by using this type of model, we would be able to incorporate multiple predictor variables and get a parsimonious model with easily interpretable results. We were looking to make a prediction about which team would win the championship this year, which is an event that has not happened yet.

We used a similar approach to k-fold cross validation to test our data, leaving one regular season dataset out and used the three other datasets combined as the training set. We did this instead of randomly shuffling the data in order to make sure there was a relatively equal balance of successful and unsuccessful teams in each set. Although this allowed us to have one championship winner per dataset, it did make the process more code-intensive, as we had to manually create the training and testing sets for each iteration. This led us to use fewer datasets than we might have otherwise due to the time involved. In total, there were four iterations using each season as a testing set once.

The heatmap shown in Figure 1, shows all different combinations of two predictors. There is a color range on the right side of the graph from white to red. The darker the hue the stronger the correlation. When looking at “CHIP” (championship) on the Y axis, we can see that it correlates most with blocks and assists. The predictors that it correlates least with are turnovers and defensive rebounds. This gives us some intuition for which predictors will have the greatest effect on a team’s likelihood to win a championship. Our results indicate that the team with the greatest blocks and assists will have the greatest chance to win the championship.

Graphical user interface, application

Description automatically generated

Figure 1: Feature heat map displaying the correlation of our predictors.

**Results**

**I. Model Selection:**

We used forward stepwise selection to determine the selection of parameters with the least mean squared error on the testing set during the cross validation process. Unfortunately, this process resulted in a lot of singular matrix errors, forcing us to discard many potential predictor variables. Once we selected the logistic regression model with optimal parameters, we regularized the parameters using L2 (Ridge Regression) with different intensities. We tested out intensities ranging between 0.1 to 1.0, using the statsmodel library’s logit model. We chose to regularize the model with L2 because given our limited number of relatively unique predictors, we believed that it would be beneficial to include them in some limited capacity in the final model. When comparing our models we found that the optimal model had an intensity of 0.7 because it had the lowest mean squared error at 0.1. The Figure 2 below shows the mean squared error for each of the L2 values, ranging between 0.17 and 0.1. The model with the lowest MSE overfitted the least to the training sets. We wanted our model to not be overly flexible since a very flexible model would place too much value on anomalies and unlikely playoff runs. Models that did not regularize the parameters as much had much more mean squared error on the test set due to overfitting, as they placed too much emphasis on predictor variables that didn’t correlate greatly with playoff success on average.

Chart, scatter chart

Description automatically generated

Figure 2: Scatter plot of Mean Squared Error of each L2 value

**II. Model Estimation:**

The final parameter estimates are shown in Figure 3 below. As we can see, the coefficients with the highest magnitude would have the biggest influence on the prediction. In this case, blocks would have the most impact on a team winning championship. This did not align with our initial intuition, which was that 3-pt field goals would have the most influence. We cannot yet determine the final accuracy of our predictions because this year’s playoffs are still in progress.

|  |  |  |
| --- | --- | --- |
| Abbreviation | Name | Coefficient Value |
| PTS | Points | 0.09043647 |
| AST | Assists | 0.29864974 |
| P3 | 3-pt Field Goals | 0.00498689 |
| DRB | Defensive Rebounds | -0.28256269 |
| FG | Field Goals | 0.32538206 |
| TOV | Turnovers | 0.13590661 |
| BLK | Blocks | 1.03031267 |

Figure 3: Chart of Coefficient Values for each predictor

**Conclusions and Discussion**

Although it is too early to tell whether the model picked the most likely championship winner, the model has produced somewhat mixed results. Figure 4 below displays the 16 teams in the playoffs this year and their predicted chances at winning the championship (in percentage).

Table

Description automatically generated

Figure 4: Chart of Percentage of winning NBA championship per team

The model makes a few solid predictions that many basketball analysts and fans would agree with. For example, the model predicted that the Brooklyn Nets have the best chance at winning the championship, which many who are familiar with the team would agree with. They undoubtedly are the most star-powered team in the league, having two former MVPs in James Harden and Kevin Durant, as well as all-star Kyrie Irving and lethal shooter Joe Harris. Additionally, the model ranked the Nuggets, 76ers, Bucks, and Suns relatively high, which is fitting given their regular season success. However, there were a few clear errors. The model gave the Memphis Grizzlies the third best chance of winning the championship, a prediction that has not aged well given that the Grizzlies have already been eliminated by the Utah Jazz after five games. Additionally, the model ranked the Washington Wizards, a team that barely made the playoffs and has already been eliminated, higher than last year’s champions, the Los Angeles Lakers.

There are a few probable reasons why this model came up short in some of its predictions. Firstly, it only used basic stats, such as points and assists, as predictors instead of advanced stats. While this made these stats easier to access and work with, it lessened the predictive ability of the model because these stats were not adjusted for factors such as the pace of the game. For example, a fast team might score more points in a game than most other teams, but they could still have worse records than more skilled teams that play at a slower pace. Secondly, the model did not account for injuries, which can either hinder a team’s regular season performance or end their playoff run early. The Los Angeles Lakers are a prime example of this, as they have been plagued by injuries all year and in the playoffs.

Thirdly, the model did not take previous postseason success into account when making predictions for this year’s playoffs. As any basketball fan knows, more experienced teams tend to do better in the postseason. Additionally, some teams rest their stars frequently throughout the season in preparation for the playoffs, which may slightly hurt their regular season record but lets them play at full strength when it matters most. Another reason this model fell short is because it only looked at championship winners and non-championship winners from the past few seasons, giving no credit to teams that made deep playoff runs but were sent home before winning it all. A prime example of this is the Miami Heat, who made it to the finals last year only to lose to the Lakers. They were not rewarded for their run, however, with the model placing them 14th out of the sixteen playoff teams. However, this may have been more of a feature than a bug, as the Heat were the only team to be swept in the first round this year.

Below in Figure 5, we compared our model’s predictions to FiveThirtyEight’s predictions forecasted on May 26 (<https://projects.fivethirtyeight.com/2021-nba-predictions/>). Dots that are closer to the red line are teams with more similar predictions, while dots that are less close show teams with a greater difference in prediction.

Chart, scatter chart

Description automatically generated

Figure 5: Scatter plot comparing our predictions to FiveThirtyEight’s predictions.

If a researcher was interested in making a similar model, the best next steps would likely be to add more and better data to the model, as well as expand the range of predictors. Instead of just using the past four regular seasons as training data, the model could use the last 20 seasons. Additionally, the model could utilize statistics such as regular season win percentage, playoff win percentage, offensive rating, and more advanced metrics to create a more diverse range of predictors. Adding these features to the model should work to prevent overfitting and improve predictive power by a good deal.