CS230

Team Success: Predicting NBA Playoff Appearance based on the team's average statistices per game.

Manuel Rolon-Osuna*
Department of Computer Science
Stanford University
mrolonos@stanford.edu

Abstract

Data Analytics is becoming an integral part of modern-day sports, making teams adopt quantitative approaches to success rather than traditional qualitative strategies. In the National Basketball Association, nearly every action a player takes is represented by a statistic. Given all actions a team can take, there must be a relationship between success and average statistics. At the start of the season, each team aspires to qualify for the post-season tournament. Thus, modelling qualification for that tournament with a team's average statistic can be a new way to represent team succes. First alternate models with different parameters were fitted to find the best predicting logistic regression model. Given the high accuracy from the binary classifier, a shallow neural network was built to create premier accuracy.

1 Introduction

As technology develops so do the analytical processes applied to large data sets. Given statistics can be represented by data sets, sports have began implementing data science to coaching strategies. Basketball in particular documents individual statistics for nearly all the actions possible in a game. This information can be leveraged by team managers to improve their success. Some managers are dismissive of analytical approaches, and prefer a traditional coaching philosophy based on empirical observations. However, by utilizing data, team managers are able to evaluate their team's overall performance and determine benchmarks for success. [1] Using these benchmarks the managers can generate predictions for how successful their team might be and if they will meet their season goal. Given the increased popularity of neural networks, it is no surprise that there exists research on deep learning approaches to analyze analyze commonalities within the game of basketball. Reinforcement learning has been applied to predict player behavior, however, the learning is slow due to the large number of different. [2] Also inspiring my project, is past research that identified a defensive commonality with playoff teams. [3] However, these studies have player actions or searched for similarities, rather than monitoring progress towards a goal desired by all teams. In my project, success will be determined by a team qualifying for the playoffs.

2 Related work

My project will not be the first research into to applying deep learning concepts to basketball statistics. Previous research has taken a deep learning approach in order to make predictions about basketball.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Like my project, these projects were interested in predicting outcomes by making use of in game statistics. Previous studies were able to identify statistics that were more correlated to winning than others. [4]

A common approach is analyzing statistics as features to determine a probability of outcomes. There is an original paper by researchers that observed Serbian basketball and applied a neural network approach to understand the significance of statistics. Here the paper identifies rebounds as being the one of the most important statistics through the use of " feed-forward technique in neural networks." [5] This is similar to the idea of defensive importance proposed by Neiman and Loewenstein. [2] While these papers are separated by some years, they both touch on a common theme that there are some statistics that are crucial to positive outcomes in basketball. For instance, a German study identified key features and found that features of a model provide insight and are useful for NBA coaches to improve their team's capabilities. [1] The German study's weakness is that it analyzed strings along with numeric values thus, not providing consistent inputs. These papers provided valuable insight that motivates my project's usage of features to predict win probability.

Not only have academic researchers analyzed this topic, but so have previous CS230 students. A Winter 2020 project made use of a 3-layer neural network in order to evaluate the "Ideal basketball player" [7]. This paper also references the paper described above analyzing Serbian basketball. In finding the "ideal" basketball player, the former CS230 student analyzed statistics to determine how good a player was. This is similar to my research into how successful a team is. Focusing on team success is a more intuitive approach because basketball is a team sport in which five players are simultaneously playing. Thus, evaluating a player's significance would be limited to recruitment purposes for adding individuals to a team.

The existing research made significant steps in applying neural networks to game statistics, however, their approach did not apply neural networks to the right classification of success.

3 Dataset and Features

In order to represent the highest level of basketball competition, I will be using the tables that are found on the website Basketball Reference. This is an extensive data set formatted in a clean manner, however, a pre-processing stage was necessary to prepare the data for my project. I had to go over 40 CSV files to input the binary playoff data, that is whether a team made playoffs or not as well as change names for teams that had changed location. This was crucial to maintain consistency over the four decades between companies. The data came from after 1980, because in 1979 the three point line was introduced. Omitting the introductory year will avoid any likely outliers. Building a logistic regression will allow the statistics to act as an input for an output between 0 and 1. I then merged the data from four decades into one csv with season averages for all teams across time.

Figure 1: A dataset of a team's average stats for a season with added in Playoff binary each row is a team's season.

PLAYOFFS	FGp	THPp	ORB	WINp
1	0.464	0.324	15.1	0.524
1	0.512	0.315	13.3	0.683
0	0.444	0.275	17.1	0.317
0	0.467	0.314	12.5	0.317

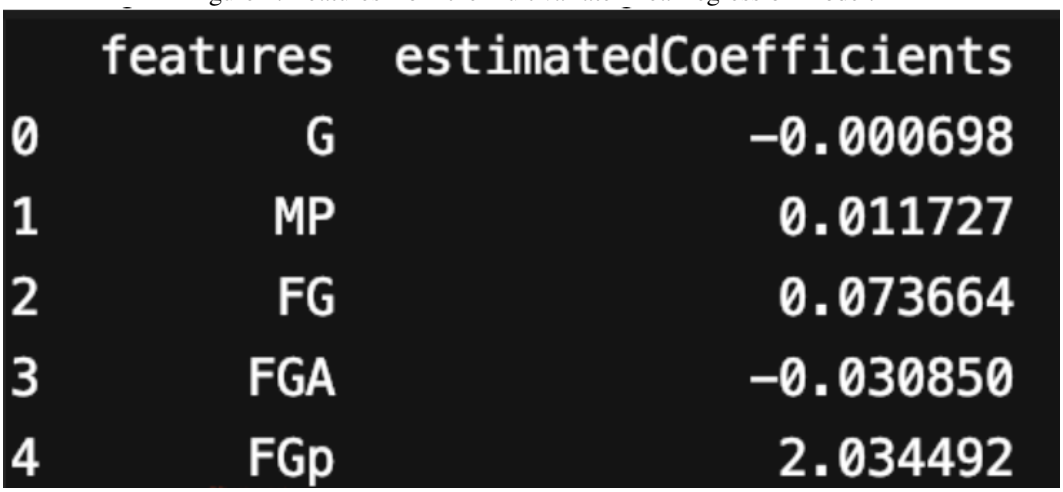
4 Method

After a season, a basketball team is left with one determinant for success, a championship. However, there is only one of those a year, therefore it would be better to model success as playoff appearance. This is an accolade only experienced by half of the league. A team's playoff appearance in a season can only be within 0 and 1, so it would make sense to use logistic regression. I kept the features that were recommended in past research such as defensive stats and rebounds. After coding a multivariate

linear regression model that correlates team statistics to playoff appearance percentage, I was able to see the weights of the different statistics. A sample of the features from my first model can be found in Figure 2.
Given the diversity amongst the statistics I found it necessary to reduce my parameters and focus only on the most important ones: the field goal percentages, win percentage, and a rebounding statistic.

5 Experiments/Results/Discussion

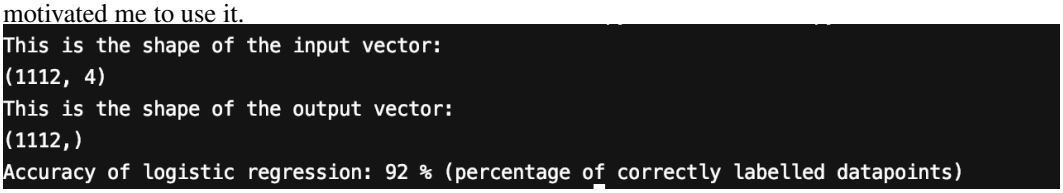
Figure 2: Features from the multivariate linear regression model.



	features	estimatedCoefficients
0	G	−0.000698
1	MP	0.011727
2	FG	0.073664
3	FGA	−0.030850
4	FGp	2.034492

These features are now modern updates to the issue of outdated features from the Serbian study and the omission of team names makes inputs more consistent avoiding mixing string and numbers. However, as mentioned earlier it was not effective to include all available statistics. This was because some statistics had a completely opposite impact as seen in Figure 2 where some coefficients were originally negative during my milestone. I fitted the model with three different types of rebounding statistics (Total Rebounds, Offensive Rebounds, or Defensive Rebounds) only including one with the field goal percentages. All rebounding statistics proved to be fairly accurate, but Offensive Rebounds combined with the field goal percentages made for the best accuracy with 92%.

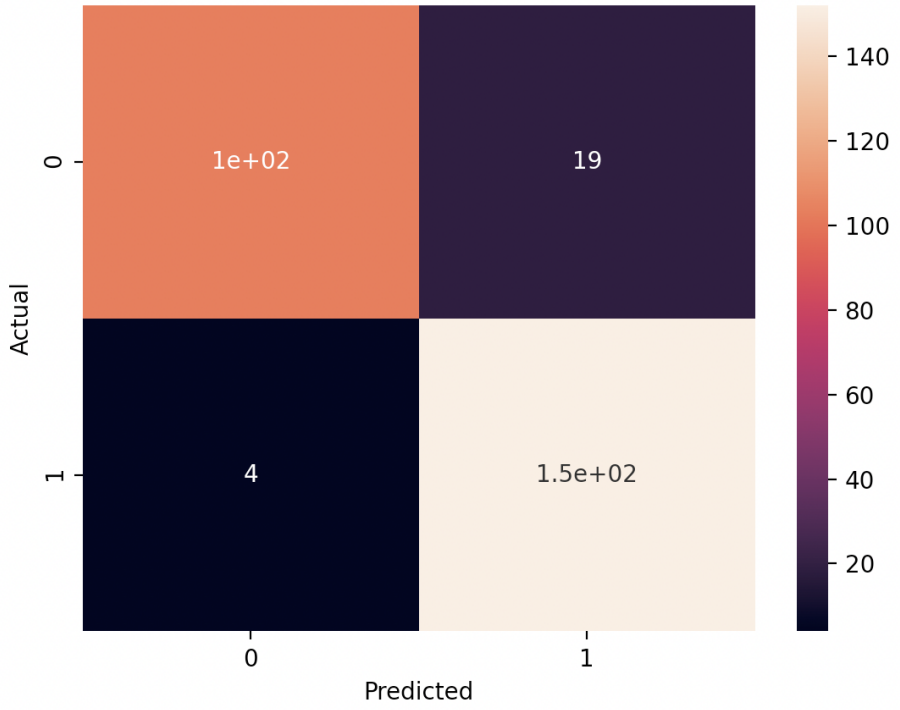
Figure 3: A 94% accuracy result from using offensive rebounds from a smaller training set that motivated me to use it.



```
This is the shape of the input vector:
(1112, 4)
This is the shape of the output vector:
(1112,)
Accuracy of logistic regression: 92 % (percentage of correctly labelled datapoints)
```

The results from my baseline model show certain stats are important to a team's success, notably the percentage related statistics. Given than a logistic regression model was well fitting, this suggests that a neural network would increase its accuracy. For this one, a shallow neural network will suffice. There is an input vector x of shape (1112, 4) and an output vector of shape (1112,). The logistic model will be used for the hidden layer, then it will be activated by relu. These activated vector will be inputs to a sigmoid function that will produce the probability between 0 and 1. I had trouble getting the neural network implemented in code, however, my baseline model is the key step in beginning the neural network to predict team success. It has promising accuracy as can be seen by the heat map of my model's accuracy.

Figure 4: The 94% accuracy result from using offensive rebounds



6 Conclusion/Future Work

My program sought to look at statistics in a vacuum and observe their accuracy in predicting playoff success. While my project had some highlights of finding strongly weighted statistics, and a rigid predictability rate.

The basketball games are not played in a vacuum and for this reason it is the case that perhaps looking at statistics alone will not provide enough information to make accurate predictions. Thus, future research should consider training a neural network with information about possible opponents. My next steps would be to get my neural network running properly, and then possibly had layers as I increased the amount of inputes. Additional inputs can be drawn from other statistics such as a ELO scores.

7 Contributions

I worked independently, which I thought to be an exciting task at first. However, the old phrase "two minds is better than one" was something I felt very often. While I was able to work on a project that I was excited about, this entailed much more obligations. I was responsible for finding all existing research and deciding the methodology I would use. I additionally had to learn about setting up all supplemental software such as GitHub and AWS. However, I am happy to say I was able to contribute on all aspects of the project life cycle from research, methodology, troubleshooting, and write-ups!

References

[1] Jiaxuan Wang & Ian Fox & Jonathan Skaza & Nick Linck & Satinder Singh & Jenna Wiens & Mozer (2018). "The Advantage of Doubling: A Deep Reinforcement Learning Approach to Studying the Double Team in the NBA." ArXiv, abs/1803.02940.

[2] Tal Neiman & Yonatan Loewenstein, (2011). "Reinforcement learning in professional basketball players," Discussion Paper Series dp593, The Federmann Center for the Study of Rationality, the Hebrew University, Jerusalem.

[3] Kohli, Ikjyot Singh. (2016). "Finding Common Characteristics Among NBA Playoff Teams: A Machine Learning Approach." SSRN Electronic Journal. 10.2139/ssrn.2764396.

[4] Thabtah, Fadi and Zhang, Li and Abdelhamid, Neda. (2019). "NBA Game Result Prediction Using Feature Analysis and Machine Learning." Annals of Data Science. 6. 10.1007/s40745-018-00189-x.

[5] Ivankovic, Zdravko , Rackovic, Miloš , Branko, Markoski , Dragica, Radosav Ivkovic, M..(2010). Appliance of Neural Networks in Basketball Scouting.Acta Polytechnica Hungarica

[6] Saladi, Vamsi. (2020) "DeepShot: A Deep Learning Approach To Predicting Basketball Success" CS230 Final Project. Stanford University Winter 2020.