DeepShot:

A Deep Learning Approach To Predicting Basketball Success

Vamsi Saladi (vamsi99@stanford.edu)

Link to Video Presentation: https://youtu.be/Hlxv3yIBVpo

Introduction

Social media has become an incredibly effective tool at not only helping people communicate but also helping organizations effectively publicize themselves. Starting off as a platform to primarily update others on your life and keep in contact with long lost friends, social media has now become a business of its own. Businesses and organizations have the ability to generate incredible leverage by having an effective social media presence. Not only can individuals optimize their social media for popularity, organizations and businesses have been trying to optimize their social media presence for views, comments, and reach. This is what makes optimizing social media such an interesting and relevant problem.

Data and Preprocessing

We are using three distinct data frames to draw from for the training data. First, we have a dataset that has the list of all NBA players who played from the year 1950 onwards, and several characteristics of them. I combine this dataframe with another list of all NBA players that made the hall of fame to be my true results. Next, we have a database that is a superset of the previous database, which has the information above along with the start and end years of their careers and the position in basketball that they played. This is the data frame that will serve as the training data in combination with your third dataset, which is a list where every row is individual season statistics for every NBA player since 1950. We have the points they scored per game, the rebounds they collected, etc. This third dataframe includes just about every measurable statistic available in sports analytics, including value over replacement player, player efficiency rating, etc. Thus, we can consolidate physical traits of the player with their season by season statistics to get our training data. After all of this, we have 3922 training examples.

Data Preprocessing

First, we address data preprocessing. We go through all of the data and take out any data before 1982, which was 3 years after the 3 point line was implemented. This was to ensure that the game could be applied to modern day basketball. Next, I had to make sure that for each of the columns in the seasons data, I had valid and usable data. First, I had to make sure that the player had data for some of the more eccentric metrics. Then, I had to go through and fill out some of the percentages which were just not given for some reason in the data file itself. For example, I had to go and fill in 3P%, 2P%, FT%, the true y-value of the player data, which is what I trained my logistic regression to predict.

Next, I organized the data with the fields that were deemed most important by research for the purpose of a simple baseline. The fields I choose were:

- Games Played
- Points Scored
- Free Throw Percentage, 3 point Percentage, 2 point Percentage
- Effective Field Goal Percentage
- Offensive Rebound Percentage
- Steal Percentage
- Turnover Percentage
- Assist Percentage
- Block Percentage

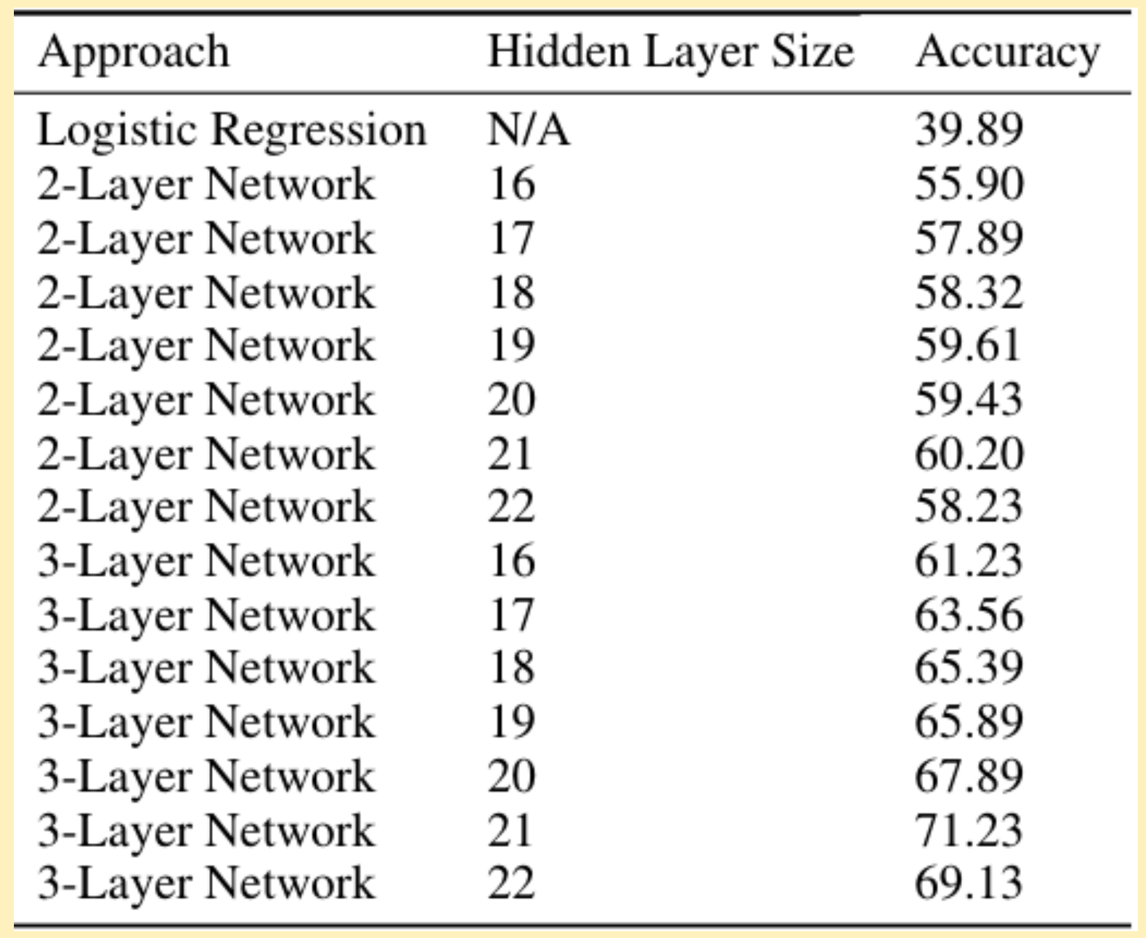
Baseline Model

The baseline model was trained to do one simple thing. Given a players' essential features gathered from his data, predict what level of success the player would achieve. More specifically, the network would predict which of the 5 categories of success the player will end up falling in: released within 4 years of data point, remained in the league as role player, became a starter, became a starter and all star, and became a perennial all star. In my implementation of Logistic Regression, we were able to achieve an overall accuracy of 39.89%.

Neural Networks with Varying Hidden Dimensions

Next, we designed a 2-layer and 3-layer neural network that had an input layer of size 11 (from the important features that we selected) and we tried different hidden layer sizes ranging from 16 to 22. The output layer was 5 units (since there are five categories of prediction) and we use a softmax function as activation function. For each 2-layer network, we also did hyper-parameter tuning. We first did a grid search of learning rates between 10^-8 and 0.1, and then fine tuned once we were more in the chosen range.

Results



Approach	Hidden Layer Size	Accuracy
Logistic Regression	N/A	39.89
2-Layer Network	16	55.90
2-Layer Network	17	57.89
2-Layer Network	18	58.32
2-Layer Network	19	59.61
2-Layer Network	20	59.43
2-Layer Network	21	60.20
2-Layer Network	22	58.23
3-Layer Network	16	61.23
3-Layer Network	17	63.56
3-Layer Network	18	65.39
3-Layer Network	19	65.89
3-Layer Network	20	67.89
3-Layer Network	21	71.23
3-Layer Network	22	69.13

Now, from this table we can see that the best overall accuracy we achieved was 71.23%, which was achieved using a 3-layer neural network with a hidden dimension size of 21. Now, it is interesting to look at the general trends of these results.

Additionally, we see that we are able to match the general accuracy level shown in the Barron study on predicting soccer player's success, which is promising since we are able to achieve state of the art (or comparable) results with just 2 and 3 layers.

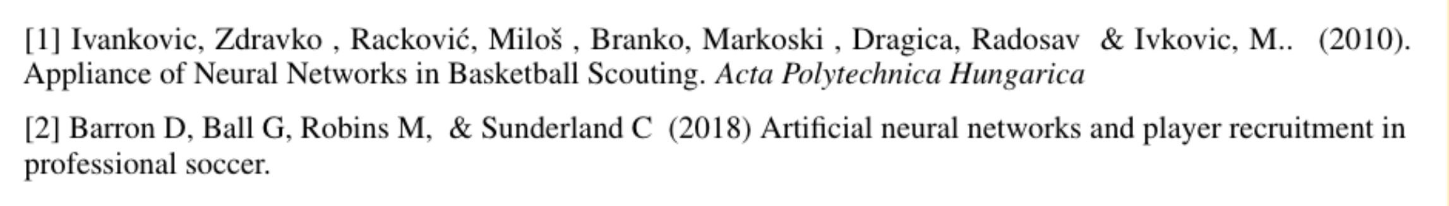
As we expect, the more layers we add the better the network seems to have performed. However, there seemed to be an interesting trend in terms of how accuracy is affected by the hidden dimension size. Though we would expect higher hidden dimension to relate to higher accuracy. But we see that in both the 2-layer and 3-layer networks, the hidden size of 22 both times results in a lower accuracy. This is an odd anomaly that is hard to explain since we have done hyper-parameter searching for all of the networks. Thus, we cannot attribute it to the a difference in tuning.

Future

The biggest conclusion we can draw is that we are able to perform at state of the art or comparable levels with our 3-layer neural network. The network performs far better than any regression method and is also significantly better at predicting success than a random guess amongst the categories, which is precisely what we want.

It is encouraging that we are able to achieve such success with deep learning since these methods can be incredibly useful for teams and organization, along with businesses. However, there are many improvements that can be made with these methods. First, we could always start with deeper networks, which would help us potentially achieve even higher accuracy.

References



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

[2] Barron D, Ball G, Robins M, & Sunderland C (2018) Artificial neural networks and player recruitment in professional soccer.

Acknowledgements

I would like to thank Professor Andrew Ng for his teaching in this class and for providing the tools necessary to conduct this project. I definitely learned a lot during the course of this class, and I plan to use these techniques in my future work.