Future Computing Technologies Lab: Creative Inquiry

SEMESTER REPORT

April 18, 2020

Jason Teets
Clemson University
Department of Electrical and Computer Engineering
jteets@clemson.edu

1 Introduction

My project was based on statistical data from the National Basketball Association. The data I worked with was from the 1956 season forward although as I will discuss in later sections that data can be subdivided due to the fact that more statistics have been kept as years progressed. The task was primarily classification although some visualization was also completed to look into specific aspects of the data. I attempted two classifications. One was to classify a player as MVP or not and the other was to classify a player as All-NBA or not based upon their yearly statistics. This task was chosen to explore two questions. The first is whether or not statistics can be an accurate predictor of award winners. The second is if statistics prove of any use what models will be the best predictors.

2 Materials and Methods

2.1 Acquiring the Dataset

My first step was to acquire the data. I used the website Kaggle.com to find the entry entitled "NBA Players stats since 1950". This data did not yet contain the award winner status of the players.

2.2 Adding Features

To add this to the data I created tables with the award winners names and years. I then used the excel VLookup function to look up the year given in the player data in my tables. If the name looked up at that year matched the player name the award column was assigned a value of one, otherwise it received a zero. I also added a position number to represent what position the player is listed as playing. This is important for the All NBA predictions as these teams are typically picked with two guards, two forwards, and a center.

2.3 Preparing the Data for Training

Once that was completed a few other preparations were made with the data. Two versions of the spreadsheet were created, one with only the statistics collected over the entirety of the data set and the other with empty or uncollected data filled in with a zero. In the intitial data if a player switched teams during the year there were entries for each team and a total entry. The individual team entries have been removed as it is unreasonable for any of these models as they are currently applied to connect the five games played with one team, where totals are low to an award won for an entire season's work. The discovery of this flaw in the data was made by visualizing the point spread for award winners and non award winners. Some data points seemed to be outliers with very low values for award winners, and upon examining the data these were identified as players

who switched teams and the above fix was made. You can see the scatter plots of the data for four specific statistics of interest on pages 6 and 7 of this report.

2.4 Preparing Models

Three types of models were prepared to utilize on the data: SVM, K Nearest Neighbors, and Neural Networks. The SVM model was constructed with the linear,sigmoid,polynomial, and rbf kernels. K Nearest Neighbors was attempted with a k value from 1 to 15. It was also tried for each k value with no pca and a pca 1 to 15. The Neural network was constructed with two layers of 1024 nodes and one output layer. These models were tested on one set of all players and another set from 1980 forward that contained vastly expanded statistics.

3 Results

Table 1: Results Table

Classification Task	Model Types	Model Details	Percent Accuracy
MVP since 1956	KNN	6 Neighbors	99.746
MVP since 1956	SVM	Linear Kernal	99.788
MVP since 1956	Neural Net	2 dense layers (1024 nodes)	99.746
MVP since 1980	KNN	8 Neighbors	99.789
MVP since 1980	SVM	Rbf, Sigmoid, Polynomial (2)	99.789
MVP since 1980	Neural Net	2 dense layers (1024 nodes)	99.789
All-NBA since 1956	KNN	9 Neighbors	98.153
All-NBA since 1956	SVM	Linear Kernel	98.453
All-NBA since 1956	Neural Net	2 dense layers (1024 nodes)	97.712
All-NBA since 1980	KNN	7 Neighbors PCA: 15	98.732
All-NBA since 1980	SVM	Polynomial Kernel (4)	98.970
All-NBA since 1980	Neural Net	2 dense layers (1024 nodes)	98.045

3.1 All-NBA Test

First we will examine the All-NBA classification models. An important fact to keep in mind here is that roughly 97.5 percent of all players represented do not win an All-NBA honor. For the set of projections using only 1980 forward data the number is closer to 98 percent This means that accuracy numbers which may appear high are not necessarily valuable. If an accuracy is not at least above that threshold a model that simply output a zero for every entry would be just as accurate.

3.1.1 SVM Models

For the data containing all players since 1956 all but the sigmoid kernel SVM model exceeded the threshold of 97.5. The linear kernel had 98.453 percent accuracy, rbf was 98.072 percent and simoid was 96.716 percent. The best polynomial models had a degree of two or three and reached 98.114 percent accuracy.

For the data containing all players since 1980 all SVM models performed rather well. Linear model accuracy was 98.943 percent, rbf was 98.494 percent and sigmoid was 98.124 percent. Polynomial models performed even better, with a fourth degree polynomial model reaching 98.970 percent. It is worth noting all models surpassed the assumed zero threshold, and the polynomial and linear models did so by almost a full percentage point.

3.1.2 K-Nearest Models

The K Nearest neighbor models were also reasonably good predictors. For the data since 1956, The peak accuracy was 98.153 percent using nine neighbors, although smaller neighbors could also perform decently with three neighbors reaching 98 percent accuracy.

For KNN models on the 1980 forward data seven neighbors performed the best with an accuracy of 98.705 percent. This model was additionally improved upon by using a pca of 15 and reached an accuracy of 98.732 percent.

3.1.3 The Neural Net

On the 1956 forward data the Neural Net managed only 97.712 percent accuracy, the worst of the different model types. On the 1980 forward data the Neural Net was also the least succesful model with only 98.045 percent accuracy.

3.2 MVP Test

Next we will evaluate our MVP classification models. Much like the All-NBA data it is important to note the vast majority of players will not be MVP. Roughly 99.75 percent of all players represented do not win the MVP. For the set of projections using only 1980 forward data the number is closer to 99.80 percent This means that accuracy numbers which may appear high are not necessarily valuable. If an accuracy is not at least above that threshold a model that simply output a zero for every entry would be just as accurate.

3.2.1 SVM Models

For the data set containing all players since 1956 the SVM model with a linear kernel outperformed the 99.75 percent threshold by reaching 99.788. All other SVM models with other kernels achieved no better than 99.746 percent.

For the 1980 forward data no SVM model outperformed the 99.8 percent thrshold of simply assuming a player was not the MVP. Models constructed with a C value of 1 and an rbf or sigmoid kernel both reached 99.789 percent accuracy. The best polynomial

kernel model had a C value of 1 and degree of 2 and also reached 99.789. Overall linear proved the least effective reaching only 99.762 percent accuracy.

3.2.2 K-Nearest Models

For the data set containing all players K Nearest Neighbors models with six or more neighbors could achieve accuracy of 99.746, just below the threshold we are looking for. PCA values were unable to improve this.

For the 1980 forward data no KNN model outperformed the 99.8 percent threshold, with the most successful being 8 neighbor and greater modeals reaching 99.789 percent. The addition of pca to models did not help in achieving a greater accuracy.

3.2.3 The Neural Net

The Neural net achieved at best 99.746 percent on the entire player data set, which is below the thresh hold.

On the 1980 forward data the neural net also failed to outperform the threshold. In the training phase it seemed to stall at roughly 99.789 percent accuracy, much like the best versions of other models used. The neural net did not seem best suited to this task as after only a couple epochs training didn't yield any considerable improvement.

4 Experience

This section is especially important.

Overall I thought the CI was very well organized. I had no problems meeting up with Ben every week to discuss the notebook for the week or my work on my project. I think I learned a lot about data visualization that I didn't know both from a conceptual and coding point of view. I had some notion of the different types of learning algorithms but I learned to implement them which I had not done in practice before. I would say the biggest challenge is that many of the methods used are library specific so you don't have any intuitive understanding of parameters or what is really going on at first. This is pretty easy to overcome though with the notebooks distributed by the lab and online documentation. I would say this lab exposed me to data science and machine learning in a way I would not be able to engage in during the first year of my undergraduate curriculum. It also provided me the experience of putting together a report and summarizing data which is not really a part of the early computer science curriculum. I wouldn't say the CI necessarily is going to cause me to pursue any specific opportunities or paths, but it did keep me open to the idea of doing more data science work.

5 Conclusions and Future Work

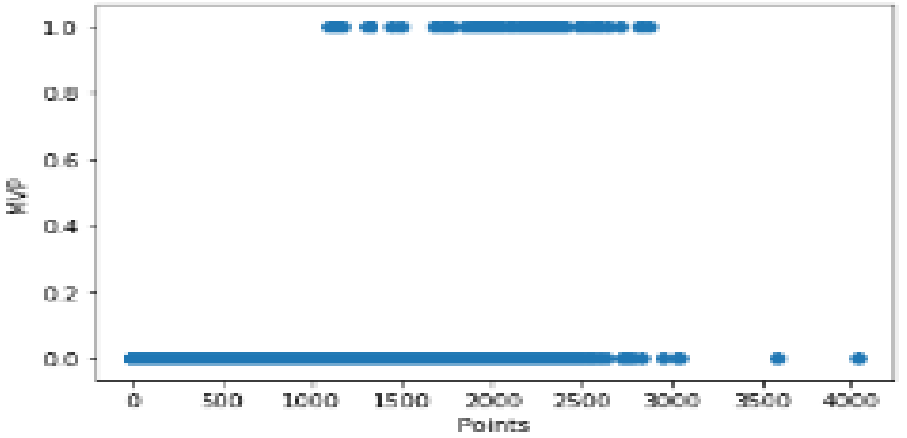
The conclusion to be drawn from this research is that pure statistics alone are a poor predictor of who will win MVP, but they can provide somewhat useful predictions about a

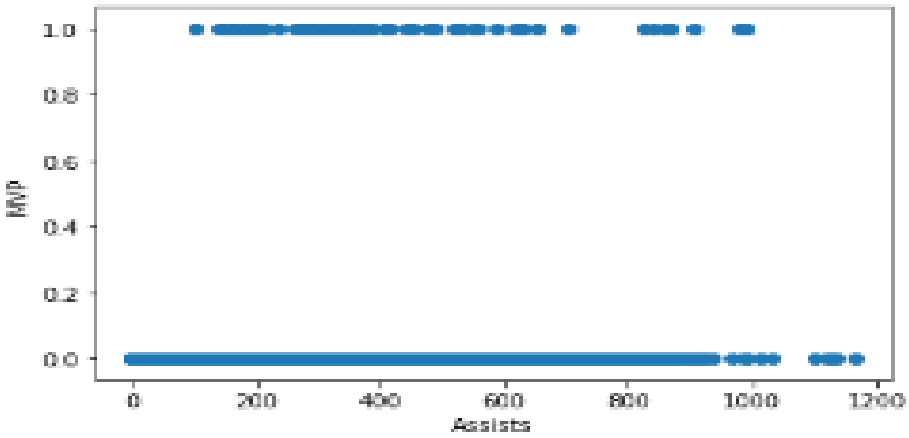
players All-NBA status. The All NBA Models clearly overcame the percentage threshold of just assuming all players did not win the award making it somewhat useful, however the MVP model could not accomplish this. In an honest analysis however it should be noted the All NBA models would likely be still a bit less effective than asking someone informed on the sport to make predictions.

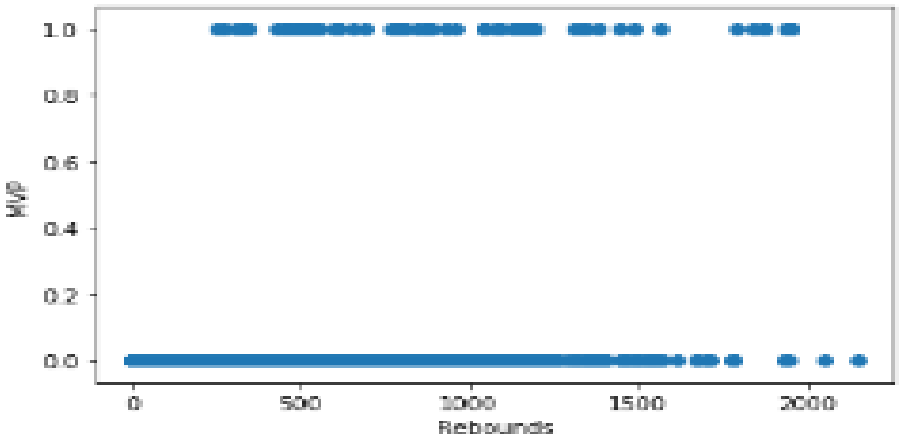
For the questions currently examined I believe the next step would be to accumulate a few other metrics to aid the model. One added statistic that might be helpful is the number of wins a players team had that year. If a player was on a team that was more successful they might get more attention and thus be more likely to win an award. A metric could also be added containing the number of times a player has previously won each award. Many sports analysts express a belief that the group which awards the MVP honor are less likely to give it to a player who has won it several times in the past, especially in recent years. The line of reasoning goes that if two players posses relatively equal resumes the one who has not won before will be given the award because fans prefer parity to predictability. By adding previous award wins this hypothesis could be tested. In terms of other projects I could see this methodology being used on other sports for predicting other awards and honors. Some similar data sets are on Kaggle for other sports which may be of some use.

6 Visualization Plots

Figure 1: MVP Scatter Plots







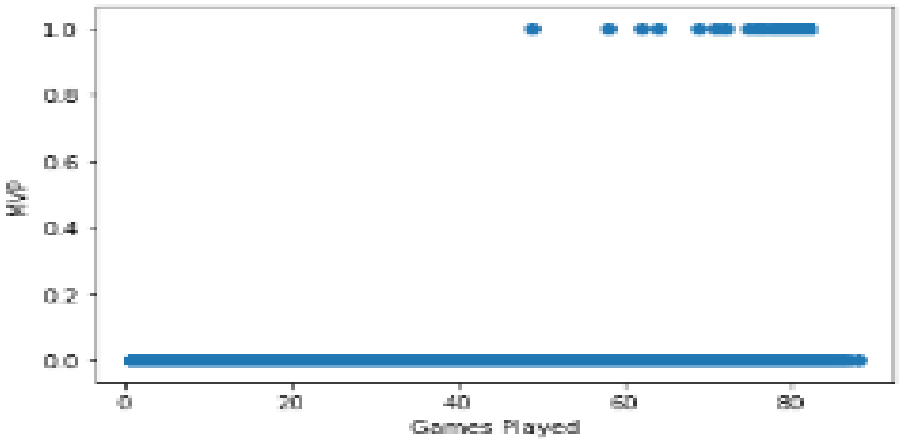


Figure 2: All-NBA Scatter Plots