**DSC 478 Final Project**

**Matt Watson**

**Executive Summary**

The goal of this Data Analysis project was to predict whether or not a player will be in the Hall of Fame. The data used came from SLAM Magazine's list of the top 500 greatest NBA players of all time, created in 2011. The data was obtained before the 2019 NBA season started. There was a lot of data cleaning applied before I could start my analyses. I had to manually obtain the position for each player as well as a list of all NBA players in the Hall of Fame. The position variable was needed to fill in missing values for Steals, "STL", Blocks, "BLK", and 3-point field goal percentage, "3P%". Steals and Blocks were not recorded until 1973 and the 3-point line was not introduced until 1979. Since these variables differ greatly among different position, I grouped each of these variables by position and then replaced the missing value by the median of each group.

Once the data was ready to use, I first performed some Exploratory Data Analysis (EDA). I looked at the medians/counts across the various variables. I also plotted some box plots that showed that players in the Hall of Fame tended to be ranked higher (closer to 1), had a higher average Points Per Game, Win Shares and Win Shares per 48 minutes.

Next, I performed a Logistic Regression on the data to attempt to predict "HoF," which has a value of "1" if the player is in the Hall of Fame, and a value of "0" if they are not. Logistic Regression is used to predict a Binary variable, has only two values, either 0 or 1, by using the remaining variable as the predictors. The players who are not currently eligible for the fame, whether it be that they are still active or have not been retired for 3 years, were removed from the data. I split the data into training in testing sets, with 80% being in the training data and 20% in the test. This is done so you can train the model on the training data and then test the model with testing data. The Logistic Regression model was able to accurately predict "HoF" for roughly 93.48% of the observations in the testing data. This is much higher than I expected, as going into this project I believed that certain accolades such as the number NBA Finals appearances or the Number of All-Star seasons would impact Hall of Fame voting. This does not appear to be the case, however. Lastly, I used the separate list of ineligible players to try to predict their odds of being voted into the Hall of Fame. The results from this prediction seem quite accurate, as players such as Kobe Bryant, Tim Duncan and Lebron James all had very high odds (greater than 90%) of being voted into the Hall of Fame.

Next, I ran a K-Nearest Neighbors (KNN) model on the data. KNN is a type of supervised machine learning, as we know the class ("HoF") we are trying to predict. KNN measures the distance between a new observation and all of the other observations in the data. It then assigns the class to the new variable based on what the classes of the "k" closest neighbors. Say we use a k of 5 and 3 out of the 5 have a value for "HoF" of 1. KNN would assign a "1" to the new observation. I first had to normalize the data, so the values ranged from 0 to 1 for all of the variables. This prevents one variable, such as "From", the starting year of the player's NBA career, being weighted too importantly. Since "From" ranged from 1947 to 2006 and many of the variables only ranged from 0 to 30, normalizing the data was important. I tested a variety of k's, from 1 to 10, and found that a k of 5 gave the best results. I then used Cross-Validation on a KNN model with a k of 5. Cross validation takes different train/test splits of the data and compares the results. This gave a score (accuracy) of 0.87 +/- 0.4, meaning that expected accuracy for the data on a KNN model with a k of 5 is 83% to 91%. I then predicted the probabilities for the ineligible players again and got similar results as those from the Logistic Regression model. The players I would expect to be in the Hall of Fame had high probabilities, but the rest of the players had very low probabilities. I would say that the Logistic Regression model performed better in prediction for the ineligible players.

I then ran 2 Principal Component Analysis models, one on the data used for Logistic Regression and one on the data used for KNN. PCA is type of unsupervised machine learning used to reduce dimensionality of the data. It groups the various variables into "components." For PCA, you want to select the number of variables that explain a high percentage of the variance, usually over 95%. I found for the Logistic Regression data that 2 components would work well and for the KNN data that 10 components would work well. I transformed both of the data sets so that instead of the 20 some variables they used to have, they now have 2 and 10 variables corresponding to the number of components selected.

The Logistic Regression model with the PCA data performed similarly to the original Logistic Regression model, as it gave the same accuracy of 93.48%. However, when it came to predicting probabilities for the ineligible players, it performed much worse as the results did not really make sense. The players you would expect to be in the Hall of Fame had low probabilities, while players you would not expect to be in the Hall of Fame had high probabilities. The KNN model with PCA data performed slightly worse than the original KNN model, as the score on the cross validation was 0.85 +/- 0.5, or 80% to 90%. Predicting probabilities for the ineligible players with this model also gave poor results.

In conclusion, both the original Logistic Regression and KNN models performed quite well, with accuracies ranging from 80% to roughly 94%. The predictions for ineligible players from these models also seemed quite accurate. Although the Logistic Regression and KNN models run on the PCA transformed data resulted in similar accuracies, the predictions were much worse. This leads me to believe that using PCA for dimensionality reduction on a data set of this size is not necessary. Going into this project, I only expected to obtain accuracies of around 60%. Since I obtained much higher accuracies, I can say that just looking at players game stats is good enough when it comes to predicting whether or not they will be in the Hall of Fame. Various player accolades, such as MVP and All-Star Games are not that important.