ISyE 6416 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Final Report

Team Member Names: Bella Smith & Betsie Last

Project Title: Baseball Predictions

Problem Statement

Using k-means clustering, does including the previous playoff history significantly better our prediction for which teams are likely to make playoffs? Additionally, is there a difference in the predictability of k-means clustering depending on the year (prior to 1994 or 1994 and after)?

Data Source

We collected the data from baseballreference.com. While there are many websites that provide baseball statistics, we needed a site where we could find the statistics on any date rather than just the data from the end of the season. As described in our data background we collected data from August 16th of each year and this site allowed us to input specific dates from each season which allowed us the flexibility we needed.

Data Background

We will use data from 1975 to 2015. Rather than using the data from the whole season, we will collect our data from August 16th of each year. This will allow us to predict whether or not a team is likely to make playoffs at a time most relevant to baseball fans as well as managers. We chose August 16th as our date to collect the data since July 31st is the current trade deadline and September 1st is the current roster expansion deadline. Using data from this date allows us to account for the finalized rosters while giving ample time before the roster expansion occurs for managers to make decisions on who to include in the expanded roster.

Additionally, we chose to compare the effectiveness of clustering prior to 1994 with the years after 1994 because after the 1994 season the playoff structure drastically changed. Prior to 1994 the playoffs consisted of two rounds, allowing only 4 teams to make the playoffs (except in 1981 when a different structure was tried), due to league expansion in 1995 the league and playoff structure changed drastically now allowing either 8 or more recently 10 teams to make the playoffs.

Finally, while compiling the data we ran across some missing data. For teams that were added to the league we did not have previous playoff history so for any missing previous years we entered that they did not make playoffs. Additionally, when a team moved to a different city or changed their name for the "new" team we used the

originating teams playoff history. The last issue that we ran into is that there were no playoffs in 1994 due to a strike. To handle this issue we worked under the assumption that playoff teams would be chosen under the new playoff structure and calculated who would have made playoffs and used this as the data.

We will look at the three variables: runs scored, runs against, and previous history making playoffs three years prior. The statistic for previous history will be computed as the percentage of times in the previous 3 years a team went to playoffs x 100. We chose to include runs scored and runs against because they are known to have a mostly linear relationship with winning percentage, which is the main predictor for whether a team will go to playoffs. Additionally, we chose to include the playoff history of the past three years because while teams who are performing well tend to stay that way, there are changing rosters that may affect whether or not they make playoffs from year to year. We thought that including only the previous three years would be able to account for both of these aspects.

Methodology

We will do two types of k-means clustering: two dimensional and three dimensional, adding the previous playoff history to as our third variable. We will see the impact of this on our model- does adding this variable increase the accuracy of the prediction model? Our first k-means clustering will be two-dimensional: including the variables runs scored and runs against. We will cluster in two groups: likely to make it to playoffs, and unlikely to make playoffs. Our second k-means clustering will be in three dimensions: including the same variables as the two-dimensional, runs scored and runs against, while adding previous playoff history. We will again cluster in two groups: likely to make it to playoffs, and unlikely to make playoffs.

Since the final clusters can be impacted by our choice of initial centers, we will run both the two and three dimensional k-means algorithms ten times. We will choose the centers for each of these ten replicates by randomly selecting two data points each time. Each data point will then be assigned to the closest initial center using the standard Euclidean distance for two dimensions and the standard Euclidean distance for three dimensions. This process will optimize the minimization of the total mean squared error between the data points and their initial clusters. Then, the new center of each cluster will be computed by averaging all of the data points assigned to the cluster. We will then again compute which cluster each point belongs to and alternate between these two steps for 100 iterations. We choose 100 iterations because the k-means algorithm is known to converge quickly. After we run the k-means algorithm ten times, we will choose the clustering that provides us with the smallest within-cluster variation.

Since whether or not the teams made playoffs is known, we used this information to compute an error statistic for each year of the two models. To determine the error statistic, we will look at two different aspects of the data. First, we will analyze the models ability to correctly identify playoff teams. This will be the percentage of playoff teams that it identified as unlikely to make playoffs. This part of the statistic will give us information about how accurate our model is at predicting playoff teams. For example, if the clustering was able to classify all four playoffs teams as likely to make playoffs then

this error would be 0. Next a different aspect of the data we will consider is the total number of teams the clustering process classified as likely to make it to the playoffs, which we will compute as the percentage of teams that the clustering predicted to make it out of all the teams. For example, if the same clustering predicted that 9 out of the 26 teams are likely to make playoffs, the error from this aspect would be 9/26. For our overall error statistic, we added these two percentages together in order to make a more representative statistic of the process. Using this statistic, we are able to evaluate the models prediction ability while accounting for the amount of teams predicted to be likely to make playoffs. For example, if the clustering predicted that all teams are likely to go to the playoffs, there will be no error in its prediction of how many were correctly clustered. However, there is still a great error here since this prediction is not exclusive at all and that is why the second aspect was added in.

In order to compare whether adding this third variable to our k-means clustering improves the accuracy of our predictions, we ran a paired t-test comparing the error statistics for the two models of each year. We chose to use a paired t-test as opposed to unpaired since we are comparing two different methods for analyzing the same data set. The null hypothesis for this paired t-test is that there is no difference when you add the previous year's playoff statistic into the k-means clustering algorithm. This would indicate that there is no change in the prediction accuracy and that the error statistic does not improve or worsen when this third variable is added. We will take the standard significance level alpha=0.05 to decide whether to reject or fail to reject the null hypothesis.

In addition to the above comparison, we will also compare the error statistic for both our 2 and 3 variable clustering pre and post 1994. Since we are no longer comparing different models for the same data we will use an unpaired t-test. The null hypothesis for these unpaired t-tests are that there is no difference of the prediction ability pre and post 1994. This would indicate that there is no change in the prediction accuracy and that the error statistic does not improve or worsen from prior 1994 to post 1994. We will take the standard significance level alpha=0.05 to decide whether to reject or fail to reject the null hypothesis.

Evaluation and Final Results

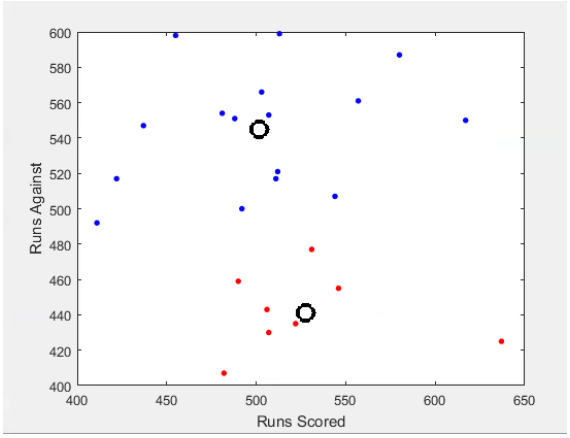
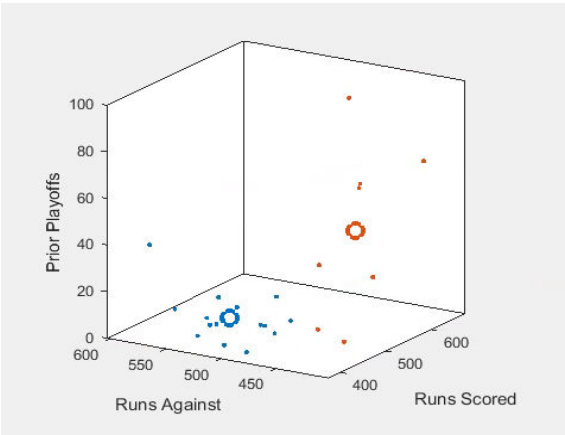


Figure 1. a) Clustering using 2 variables (1975)



b) Clustering using 3 variables (1975)

A sample of our results is shown above and one can visually see that the 2 models give similar results. Since k-means clustering is an unsupervised algorithm- it does not take into account training data to help it make predictions. We found that k-means clustering is not the best method for making predictions of who will make it to the playoffs. We originally hypothesized that adding another predictor variable to our algorithm would strengthen its ability, but we did not find this to be the case.
To test this hypothesis we first ran a paired t-test to compare the errors for the 2-dimensional and 3-dimensional case. Below is our output:

```
Paired T for 3D - 2D

             N    Mean   StDev  SE Mean
3D          41  0.7368  0.2015   0.0315
2D          41  0.7249  0.2011   0.0314
Difference  41  0.0118  0.0647   0.0101


95% CI for mean difference: (-0.0086, 0.0323)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = 1.17  P-Value = 0.248
```

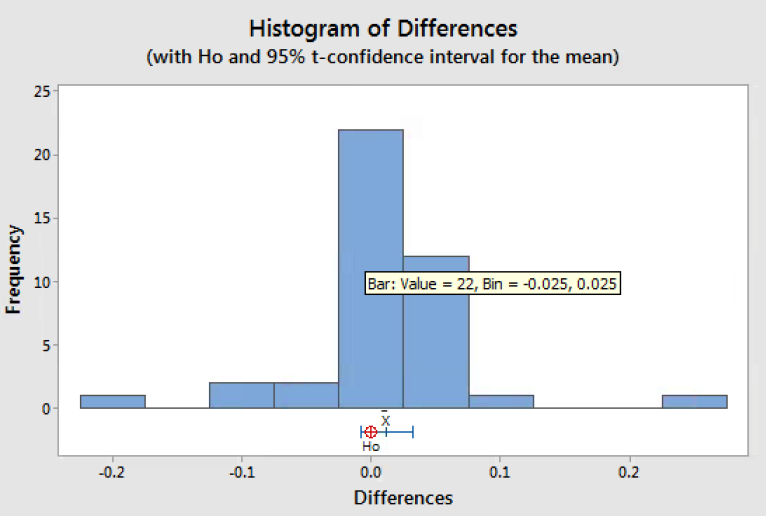


Figure 2. Output for paired t-test

The output for the paired t-test indicated that there is no significant difference between the prediction power of k-means clustering with our added variable of data versus without. The 95% confidence interval contained 0 and our p-value was 0.248, so at a level of 0.05 we fail to reject the null hypothesis indicating that the k-means clustering did not improve when we added the previous playoff statistic.

Since there was not a difference between the 2 models we were then interested if either model was significantly better (or worse) at predictions post 1994 (when compared to pre 1994), since this was when there was a major change in playoff structure. We ran 2 separate unpaired t-tests, one on the 3 variable and one on the 2 variable error statistic. Below is the output for both t-tests:

```
          N   Mean  StDev  SE Mean
3d pre   19  0.730  0.220    0.051
3d post  21  0.732  0.185    0.040


Difference = μ (3d pre) - μ (3d post)
Estimate for difference:  -0.0022
95% CI for difference:  (-0.1337, 0.1293)
T-Test of difference = 0 (vs ≠): T-Value = -0.03  P-Value = 0.973  DF = 35
```

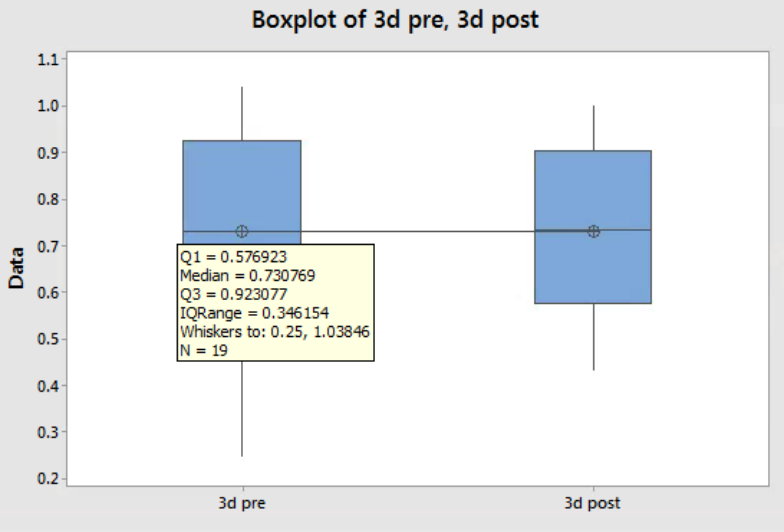


Figure 3. Output for unpaired t-test (3 variables)

```
          N   Mean  StDev  SE Mean
2d pre   19  0.723  0.224    0.051
2d post  21  0.718  0.184    0.040


Difference = μ (2d pre) - μ (2d post)
Estimate for difference:  0.0044
95% CI for difference:  (-0.1282, 0.1371)
T-Test of difference = 0 (vs ≠): T-Value = 0.07  P-Value = 0.946  DF = 34
```

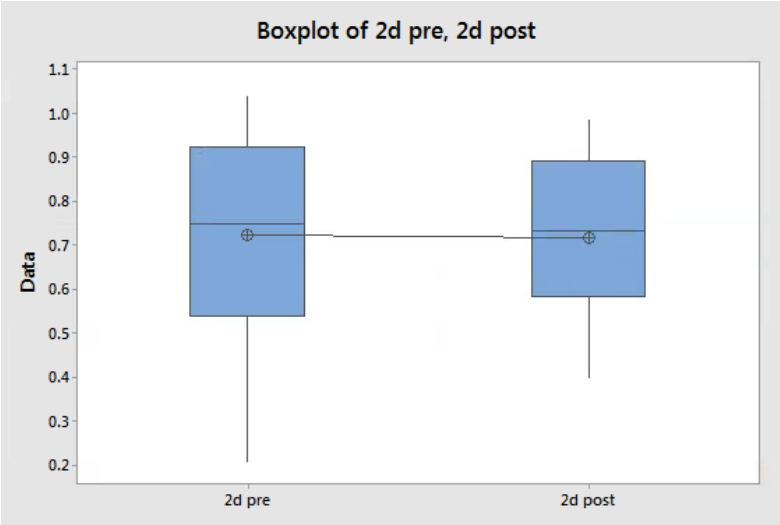


Figure 4. Output for unpaired t-test (2 variables)

The output for both of these t-tests indicated that there is no significant difference between the prediction power of either of the k-means clustering pre and post 1994. Both of the 95% confidence intervals contained 0, and the p-value for 3 variables was 0.973 and the p-value for 2 variables was 0.946, so at a level of 0.05 we fail to reject the null hypothesis indicating that both k-means clustering's did not differ between the different playoff structure.

Improvements

Since adding this previous playoff statistic did not improve our predictive ability, we considered various other wasy to improve the model. One possible improvement could be adding a different variable into the k-means algorithm that could possibly strengthen the outcome (possibly a different combination of playoff history). Another possible improvement would be to choose a different error statistic to analyze that would encompass more types of possible error that could occur in these predictions.

Additionally, changing the method altogether could also be beneficial in making predictions for who will make the playoffs. We could try a supervised model, such as LDA classification. Using this method would allow us to include training data, giving the algorithm some prior knowledge of playoff history. Another method that may improve the predictions is bootstrapping. Since we had a relatively small data set for each year, if we could create a statistic that we believe to be a good indicator of playoff likelihood, this method would allow us to "generate" more data by resampling and create a more informed estimation of this statistic.