Zach Johnson

Arizona State University

MAT494 – Mathematical Methods in Data Science

College Football Trends Analysis Using K-Means Clustering

Introduction

In the modern landscape of sports statistics, nearly every quantifiable occurrence on the field is tracked. This is especially true in college football, where you can find data on anything from first downs and field goals made to opponent fourth down conversion percentage or average yards per play. Each statistic measures a specific event that occurs within the game, but it is likely not obvious what impact all of this data has on the sport as a whole. In NCAA FBS football, what connects teams to one another? Which statistics can be used to make up a team's identity? In this analysis, I will attempt to answer the following questions. First, does a team's statistical output remain consistent, season after season? Second, does the conference a team plays in have any bearing on their statistical output? To answer these questions, I will be performing k-means clustering and analyzing performance by comparing the clustering results to expected groups.

Data

For this analysis, I will use the "College Football Team Stats Seasons 2013 to 2021" dataset by Jeff Gallini on *Kaggle.com*. The dataset compiles team statistics from the NCAA for every FBS level team in NCAA Division I football. The data represents accumulated season-long statistics for each team from the 2013 season to 2021. The 2021 season data was excluded from my analysis, as the data format was inconsistent with the rest of the years available. With 8 seasons worth of usable data and 110-130 teams present in the FBS each season, the combined data has a total of 970 observations. Only columns with available data for each season were kept, resulting in a total of 148 variables. The majority of variables are numerical, with the exception of time of possession, average time of possession, team, and conference. Only the numerical data was considered for clustering, while team and conference were used to classify the observations.

Methods

To answer the questions outlined in the introduction, I will be using k-means clustering of FBS football data in two dimensions. The k-means clustering algorithm will be performed for each combination of variables as dimensions (in order to reduce runtime, a random sample of variables from the dataset will be used, instead of the full set of 144). Data was normalized by scaling values between 0 and 1. K-means clustering was performed in Python, using sklearn.

To compare clustering results to expected labels, performance of each clustering model is measured using adjusted rand index (ARI). The best model is one that maximizes the ARI score, as a strong ARI score (close to 1) indicates that the majority of points fall within their expected cluster. The best model was plotted in a scatterplot using Seaborn and the Pyplot library. Data points were plotted along the axes of the two selected variables and styled according to their actual and expected clusters.

To explore the statistical output of teams over time, each expected cluster would contain all season observations from a single team. A model with over a hundred different clusters would be very inefficient, so in order to ensure significant results, this section of analysis focused specifically on teams within the Big 12 conference. There are 10 teams within the Big 12, so the resulting clustering models will contain 10 clusters. The expected labels of each cluster are the names of each team within the conference.

For the second half of this analysis, the created cluster models will contain 11 separate clusters, representing the 11 conference alignments recognized by the dataset.

Theory

K-means clustering works to minimize the sum of squares distance between the points and k cluster centers. Let $X = (x_1, x_2)$, where X is a 2-dimensional vector. If $S_1, S_2, \dots S_k$ are clusters, then the goal is to minimize the equation $\sum_{i=1}^{k} \sum_{X \in S_i} \left|\left|X - \mu(X)\right|\right|^2$, where $\mu(X)$ is the cluster center of S_k. The values of $\mu(X)$ are at first assigned at random, then adjusted with each step to the mean value of points within the cluster, $\frac{1}{|S|} \sum X \in S$. The mean of the cluster then has the lowest possible within-cluster sum of squares (WCSS), as proven below:

$$E[X] = \frac{1}{n} \sum X = \bar{X}, n = |S|$$

$$\delta WCSS = \delta \sum_{X \in S} \left|\left|X - \mu(X)\right|\right|^2 = \sum 2(X - \mu) = 0,$$

as WCSS is minimized when its derivative is 0.

$$E[\delta WCSS] = \sum 2(E[X] - \mu)$$

If $\mu = \frac{1}{|S|} \sum X$, then

$$E[\delta WCSS] = \sum 2(\bar{X} - \bar{X}) = 0.$$

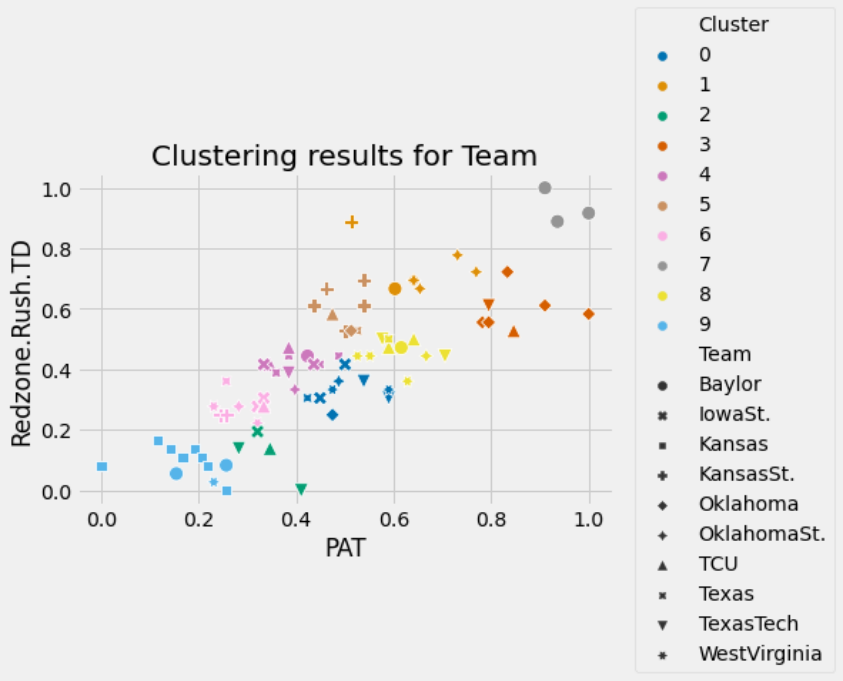
So, $\bar{X}$ is a critical value of μ(X), where WCSS is minimized.

As the centroid is reassigned to the mean of the cluster with each step, the WCSS will decrease and eventually converge.

The rand index, $R = \frac{TP+TN}{TP+TN+FP+FN}$, where TP are true positives, TN are true negatives, FP are false positives and FN are false negative classifications. The adjusted rand index is corrected for chance, using the formula $ARI = \frac{R - E[R]}{max(R) - E[R]}$.

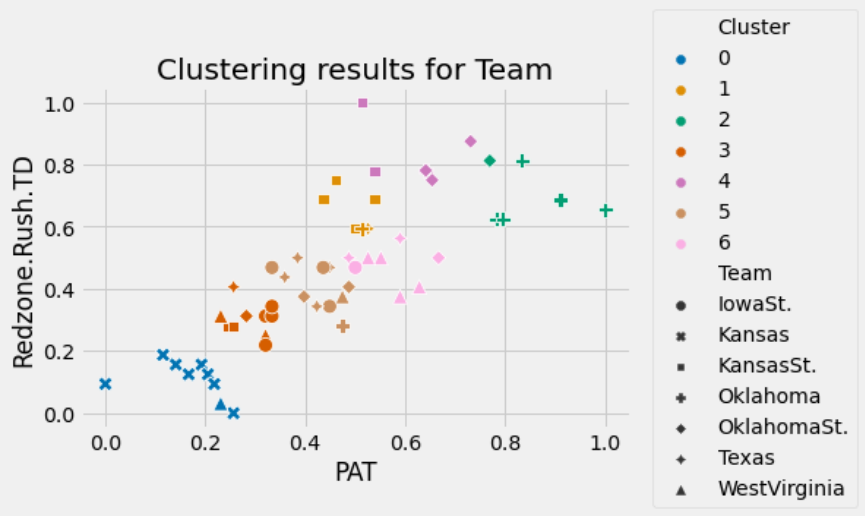
Results

The following plot shows the resulting best model relating observations based on team:



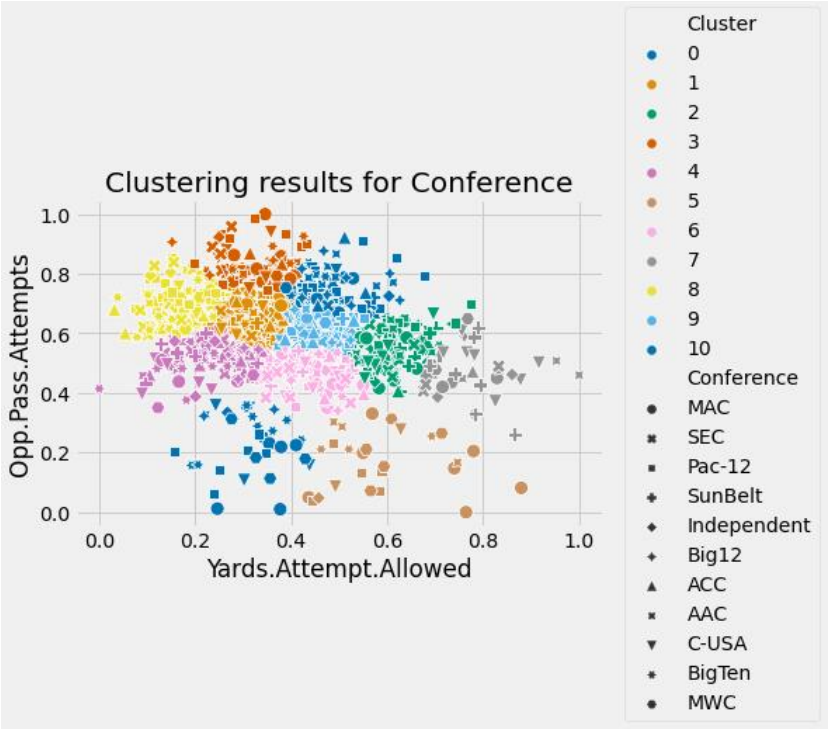
The model has an ARI of 0.1932 and a silhouette score of 0.4063. The silhouette score reflects how distinct the clusters are from each other by taking into account the intra-cluster distance and the distance between clusters. The variables chosen as dimensions for the best clustering model are PAT's and redzone rushing TD's.

From looking at the graph, there is clear distinctions between teams, especially around the outskirts of the chart. In particular, Kansas and Kansas St both have high cluster accuracy. The majority of data points for Oklahoma and Oklahoma St fall within a single cluster as well, but those data points that are not within the cluster are far apart, while the data for schools like TCU and Texas Tech are all over the place. To attempt to improve ARI even further, TCU, Texas Tech and Baylor, the three schools with the highest variance between points, were removed from the data. A new model was created using the updated data, and its results can be seen here:



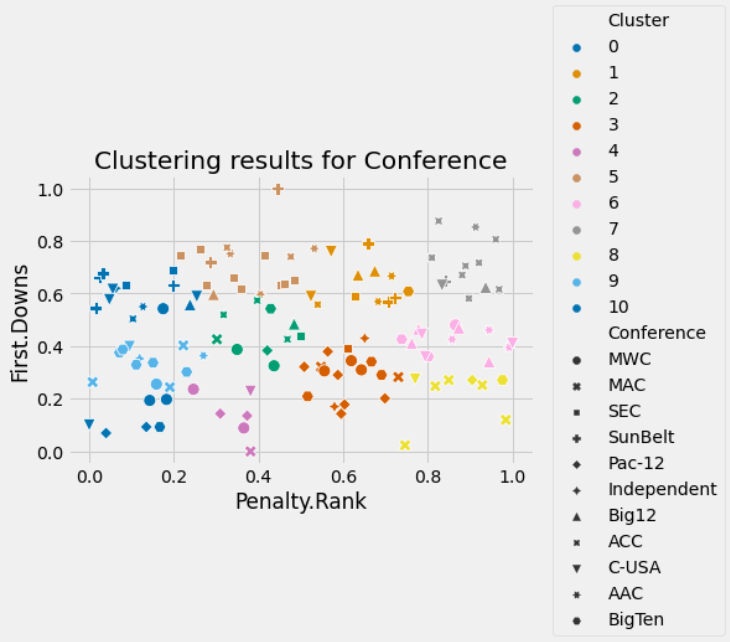
With an increased ARI score of 0.2638, removing high-variance schools from the data did help to improve performance. Notably, the clusters for Kansas and Kansas St contained less false positives. Still, it can be seen that while data points representing the same school are in many cases close together, they can end up in different clusters. This is indicative of the model's lower silhouette score (0.3975), meaning that the different clusters are not very distinguishable from each other. Also, the clustering process optimizes based on within-cluster sum of squares and not ARI, so the chosen clusters do not reflect the most accurate groupings possible. However, while the ARI score is still not very high, there is strong visual evidence that these teams performed similarly season after season.

Next, the results for clustering based on conference alignment can be seen here:



This model resulted in a silhouette score of 0.3279 and a rand score of 0.01959. If the selected best model only has an ARI score of less than 0.1, it does not bode well for the hypothesis that there is a statistical distinction between the different FBS conferences. Visually, it appears that there are certain areas that contain a higher density of data points from a certain conference, but there is very little evidence to suggest a clear trend. To potentially improve the model, the process was repeated using only data from a single season.

Using only data from the 2020 season, the following model was produced:



This model performed significantly better than the all-seasons data, with a silhouette score of 0.3688 and ARI score of 0.1296. There are multiple clusters with a high concentration of teams from a single conference, but each conference has too many outliers to declare a clear distinction.

Conclusion

From the results of my analysis, it is difficult to definitively answer the questions that this project has been investigating. Within the Big 12, there is strong evidence suggesting that some teams consistently produce similar numbers in redzone rushing TDs and PATs. From further analysis, it seems likely that similar trends could be discovered between different conferences and different statistics. For now, however, it is fair to conclude that there is at least some connection between seasons for some teams. When it comes to a statistical connection within conferences, the evidence is not as clear. My analysis suggests that the relationship is not entirely random, but more experimentation is needed to determine the extent of a conference's effect on team statistics. Based on the results of this project, a much more comprehensive analysis is recommended to further understand the trends that I have explored.