Grouping College Football Teams

Christian Stone

January 28, 2021

Main Objective:

The main objective of this analysis is to see how effective machine learning methods can be in grouping college football teams based on the results of their games through history. As a result, the model will be focused on creating clusters as this will be the way that similar teams will be grouped together.

There are a variety of different reasons why different groups may benefit from being able to group teams by their success. Schools can claim to have a team in a certain group which can help them with recruiting or with helping with their branding. Media companies can use groupings as a way to hype games. It can even help people unfamiliar with the sport understand a teams expected level of success.

Data Used:

The dataset used for this created from the ground up by taking data from the site Winsipedia. Winsipedia is a database showing every team in the FBS level of College Football as well as their record (wins, losses, and ties) against every other team that they have played in their history. A set of functions were developed in order to pull in and manipulate this data into a single dataframe for the overall data for each team. The dataframe includes each of the 130 FBS teams as the rows and the columns list the team's name, conference, wins, losses, ties, total games played, win percentage, differential, the amount of teams they have played, and the amount of teams they have a winning record against. This analysis will simply create models to group all 130 teams by their historical success based on this data.

Data Exploration and Feature Engineering:

Since the dataset was built with the intent to use it for similar types of data analysis, little work had to be done in order to make the data available for analysis. Since conferences are a categorical features, they had to be one-hot encoded to be usable for the model. A heatmap was created showing the correlation of the different features of the data:

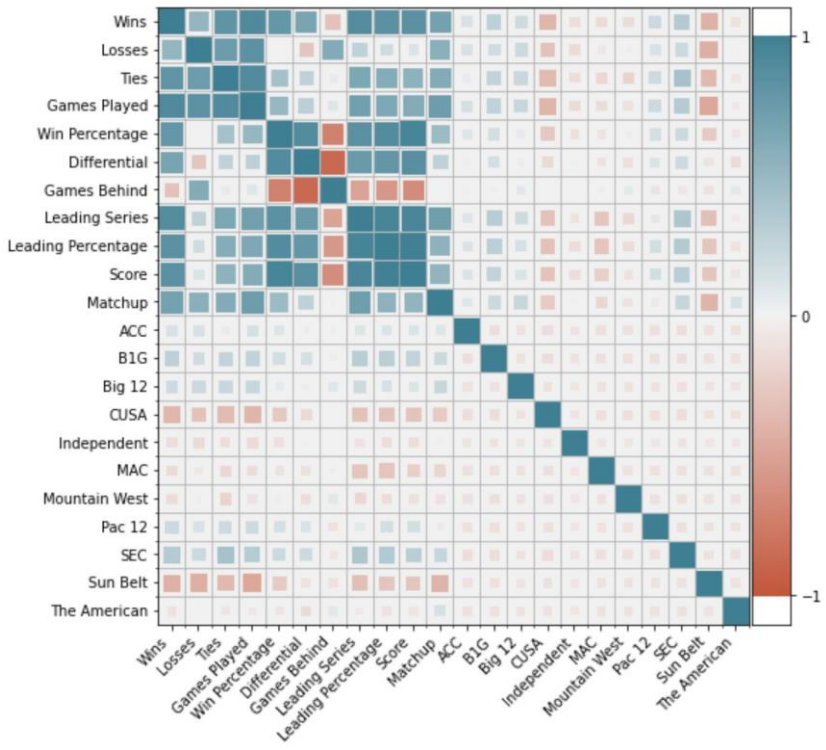


Figure 1: Heatplot of feature correlation

As can be seen above, most features either had little impact on other features or were extremely closely correlated. In preliminary model testing, it was found that having such a large amount of features made for unreliable models that were hard to visualize. In order to rectify this issue, the dataset was condensed down to just four features: wins, win percentage, matchups, and series a team was leading. To make sure that all the features had an equal amount of weight on the model, a min-max scaler was used to keep all data in the set between 0 and 1.

Lastly, the 4 final features were looked at to see whether or not they exhibited a large amount of skew.

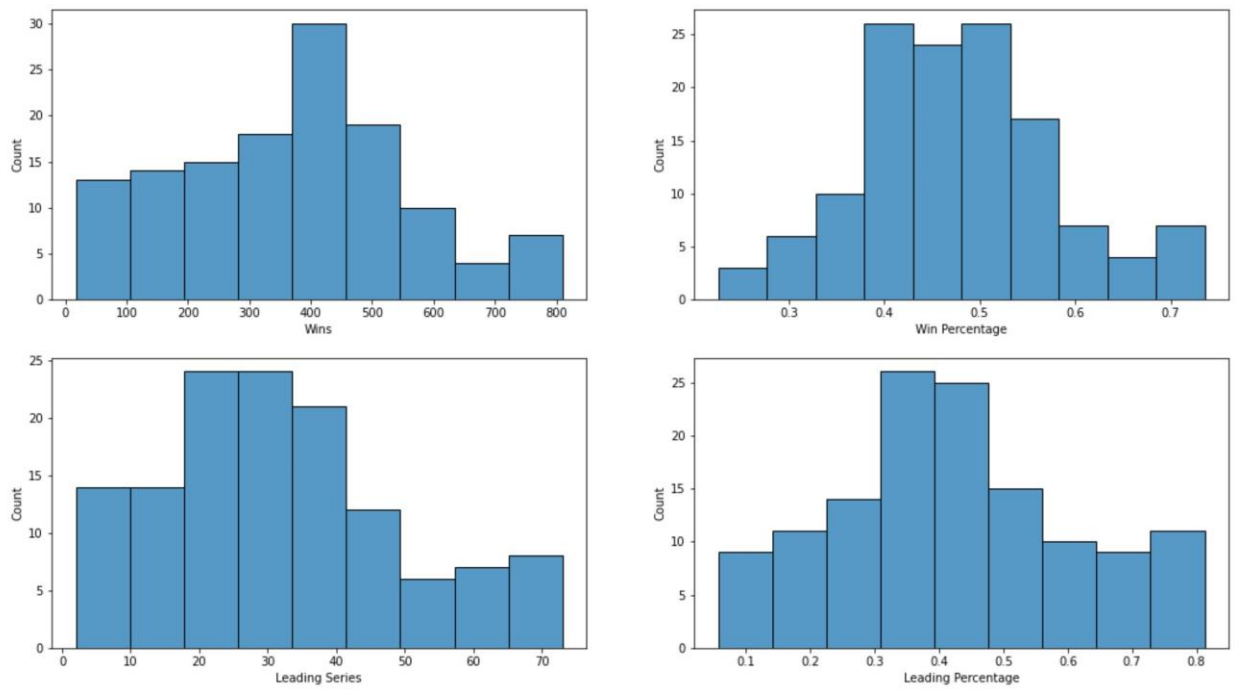


Figure 2-5: Histograms of model features

As can be seen, the plots exhibit some level of skew, however; the shapes are fairly normalized and so no further effort was made to correct the shapes.

Models Used:

Three different models were used in order to cluster the data through different means: K-Means, DBSCAN, and Agglomerative Clustering. The next step was to find the ideal hyperparameters for each of the methods.

The main hyperparameter necessary for the K-Means and Agglomerative Clustering models is the total number of clusters needed. Initially, an inertia plot was created to see if there was a clear “elbow point”, however; there was no clear point where the inertia started to level off.

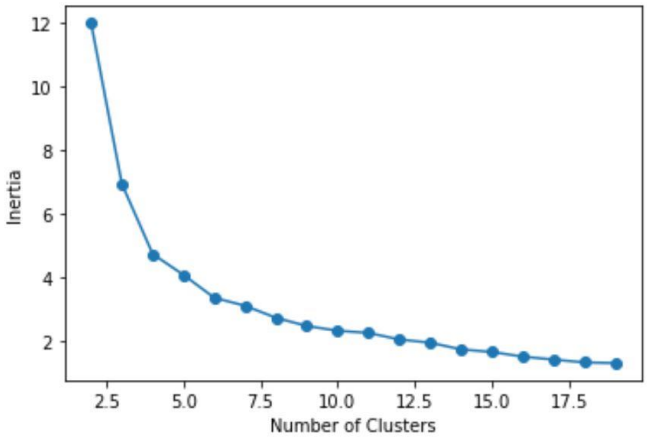
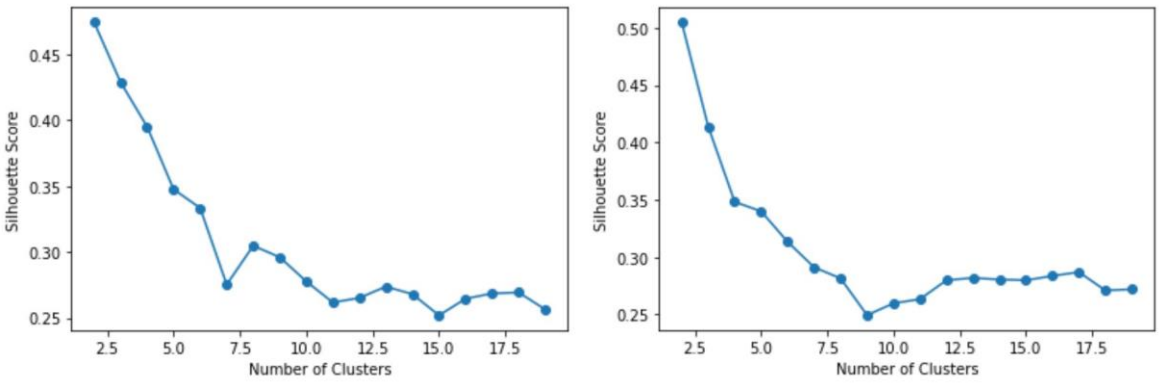


Figure 6: Inertia plot for different numbers of clusters

The backup plan was to plot the Silhouette Scores in order to see if there was an idea number of clusters. This proved more successful as the numbers 15 and 9 provided the best scores for the respective models.

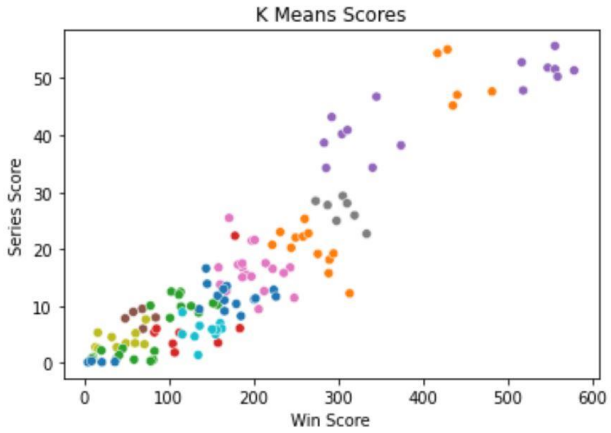


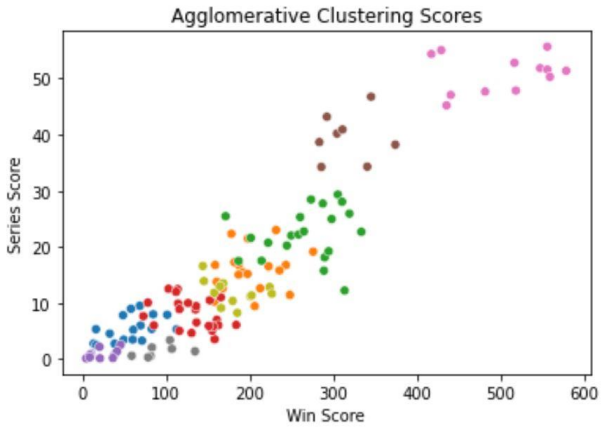
Figures 7 and 8: Silhouette scores for numbers of clusters for K-means (left) and Agglomerative Clustering (right)

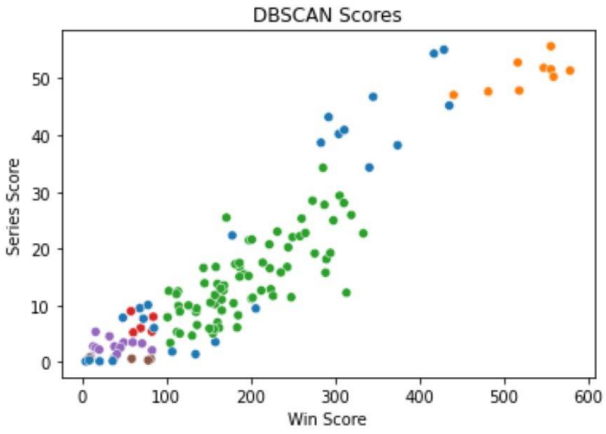
DBSCAN takes the different hyperperameters epsilon and minimum number of samples and so different analysis had to be done. Due to the fact that the teams in the data ended up being fairly evenly spaced to one another it was difficult to find an ideal model for DBSCAN. In the end, different epsilon values were tested in order to find a model that created the most unique groups (4), while keeping outlier values to a minimum.

Recommended Model:

Since unsupervised learning has no prior target feature with which to check how well the model performed, other analysis has to be done to do this. An easy way to do this is to simply plot the data in a way that show the clusters. Since the data used for the model had four features, ideally the visualization would need to be in 4 dimensions. Since this is not possible, the two sets of features (wins and series leading) had their two feature sets (total and percentage) multiplied together to create a score value. These two score values were used to plot the data in order to visualize the clusters.







Figures 9-11: Plots of the models' performances based on scores calculated from the features

In the end, two of the models performed very well while the DBSCAN model did not, given the needs of the model. Problems with the DBSCAN model included the fact that the data was fairly evenly spread out, meaning that there was no reason that any obvious densely packed clusters would form. Another issue is the fact that this type of model categorizes certain samples as outliers. While there are scenarios where designating outliers can be helpful or significant for what the model is being made for, one of the points of this analysis is to make sure that each team has a designated group.

Both the K-means and Agglomerative Clustering methods effectively mapped different groups of teams. There is overlap in the clusters in the figures above, however; this is due to the fact that the figures are 4 dimensions of data mapped onto a 2 dimensional figure. Both models work well and so the best model will come down to a preference of whether 9 or 15 clusters is ideal. Another way to decide is by looking at a commonly accepted grouping of college teams, the blue bloods. Blue bloods are a term designated to teams who have exhibited an amount of success in their history that sets them apart from all others. It is generally accepted that there are 8 blue bloods in college football and the most elite grouping in each of the models has 7 (K-Means), 9 (DBSCAN), and 12 (Agglomerative Clustering). By this metric, the K-Means appears to be the best at grouping the teams.

Key Findings:

One of the biggest takeaways from this analysis is that DBSCAN is a very ineffective way of clustering in cases such as this. This is likely because the data does not follow any sort of pattern and so there is no reason for the data to pack together into dense clusters. The data could be manipulated in order to create a greater sense of density between samples, however; this new method would come with its own issues as well. As an example, the data used in the model could have its values rounded to certain values in order to reduce noise and force density between the samples. Issues with this include the fact that the values to be rounded to will have to be decided, and in doing so would skew the data, especially for teams with values close to the boundary values.

Another important takeaway from this analysis is that it can be important to limit the amount of features used when trying to create a model. While in other forms of machine learning, it can be important to try to include as much relevant information as possible as it can help to predict or categorize the target variable depending on the type of analysis. With unsupervised learning, however; too many features may cause the model to become unfocused and unreliable as a result. Originally, the dataset used for this contained many different features, some of which were just manipulations of one or more of the other features. By only focusing on two factors (wins and series leading) and only using the totals and percentages for each, the model was able to be more focused. Additionally, this made it easier to visualize the results as a 2-dimensional plot consisting of all of the relevant data was able to be created.

Next Steps:

There are a variety of different ways in which this analysis can be changed or improved upon. While this analysis prioritized the amounts and percentages of wins and the series lead against other teams, other records such as total championships, rankings in polls, or postseason records could be useful in trying to compare or group teams. As noted before, one has to be careful when adding features to a model as the model can become unfocused or difficult to visualize.