Recent Graduate Marketing Data

Machine Learning

Final Exam – Ashley Bucolt

12/18/2022

The data being utilized in this analysis is part of the “College Majors and their Graduates” data set from the site FiveThirtyEight (no author listed) but was obtained from Kaggle. In particular, the recent-grads data set will be analyzed which includes data about various college majors, their popularity, and their employability after graduation.

The goal will be to use this data to discover some insights on which programs should be promoted in the future. That is, looking at employability, salary, gender composition, rank, etc. we will be able to rework the data to identify clusters and patters among the college majors. Once those patterns are identified, they can be summarized and potentially used in the future as promotional data. To do this, we will first need to process the data so that all variables have the same direction (ie. if a high value for one variable is good, then a high value in another is also good). Once the data is processed, we can determine the importance of each variable and create the weights for our cluster analysis.

The intended direction of the data is positive, so a high value is desirable and a low value is undesirable. To achieve this, we will be reverse ranking the majors and calculating rates for employment, non-low wage jobs, and college jobs. Doing this will allow the data to all be read in a positive way. Thus, a high score in non-low wage jobs means that the major does not have a lot of low wage jobs. Previously the rates for both low wage jobs and unemployment were written negatively, so it was important to rewrite these in the opposite direction.

To complete a cluster analysis using k means we need to determine how many clusters to use. We can figure this out two ways, using the elbow method or the silhouette method. Below you can see the results of both methods.

Chart, line chart

Description automatically generated

Now that we have determined the most optimal number of clusters, we can create a cluster plot using K-means to better illustrate the clustering. The graph below shows our three clusters obtained using K-means clustering.

Chart, scatter chart

Description automatically generated

The three clusters above separate the penguins into our three penguin species sampled. These clusters where created using all numerical data from the dataset, and it is clear that each of these three species have very clear statistical distinctions. The characteristics that separate the species into these three distinct clusters can be found in the cluster characteristics we obtain through Hierarchical clustering.

Because we are using a different method of clustering to show characteristics, we want to revalidate our optimal number of clusters for the new clustering algorithm. This was achieved by creating an Agnes dendogram that illustrates our clusters in a tree format. Shown below, our Agnes dendogram illustrates the largest jump in groups at a height of 30, indicating three clusters, affirming the findings from earlier.

Diagram

Description automatically generated with low confidence

Now that we have confirmed that three clusters will be optimal for Hierarchical clustering, we can make our characteristics. The chart below illustrates the values of the centroids of each cluster, for each value they were evaluated on.

Chart, bar chart

Description automatically generated

From the illustration above, we can gain much greater insight into the three clusters of penguins.

From this we can determine that Cluster 1 has the largest number of the Adelie penguin. Cluster 2 has a very large population the Chinstrap penguin. Lastly, cluster 3 has the Gentoo penguins. An interesting point to notice from this information, is that cluster two has very low amounts of the other two species, while the other two clusters have a small gradient as population of the minority species are denser.

Using this information, we can gleam insights into characteristics of the clusters and the species in them. Cluster two has a high density of Chinstrap penguins, as well as the largest Culmen (beak) length and depth and the largest density coming from Dream Island. This can inform us about the characteristics of the chinstrap penguin, such that they live mostly on Dream Island and have long and deep Culmens. Similarly, Gentoo that live on Biscoe Island have the largest flippers and body mass based on the graph.

This methodology allows us to easily cluster the data from our dataset and present them in an easy-to-read fashion. By doing this we can more easily make insights about the lives of these penguins through time. As more data is collected over time, we can compare the graphs then to the graphs we have made now to clearly see changes in these penguins.