GMM Clustering

Team SheCodes
Monica Dommaraju
Asha Balshiram Aher
Swati Ganesh Narkhede
Sri Sruthi Chilukuri

- **Comparison with K-Means**
    - With K-Means we got 3 as the optimized number of clusters using elbow method. GMM with AIC and BIC method also gave 3 as the optimized number of clusters.
    - We have also tried the Silhouette analysis method and found 2 as the optimum number of clusters. However, the mean average score that we got is 0.1 for 2 clusters which is not reliable.
    - So the number of total employees are divided into three clusters of sizes 1140 (Cluster 0), 182 (Cluster 1) and 148 (cluster 2) when compared to (645, 622 and 203) employees from KMeans
    - This shows that clustering of employees is not evenly distributed, with most of the employees in cluster 0.
    - Employees in Cluster 1 have the highest attrition rate of 19% when compared to 16% in Cluster 0 and 10% in cluster 1, where as it was (7%, 18% and 18%) in K-Means.
    - The total number of employees to analyze required to minimize the attrition rate is very high in GMM when compared to Kmeans.
    - GMM also supports the **predict_proba** method as it uses probabilistic models under the hood. This helped us to detect outliers by checking the probability of each point belonging to every cluster. If the probabilities are less than 0.5 for every cluster, then that point is considered as outlier. GMM has 148 outliers, whereas for KMeans it was 24.
    - Cluster models are only circular in K-Means. This results in a significant overlapping area of circles. In GMM, the cluster shape is flexible, it can be spherical, circular or elliptical. We have used covarience_type of full for modelling.
    - When applied K-Means and GMM clusters on PCA components, we found that the distribution of data points in clusters is different in GMM than K-Means clustering. In GMM Clustering, cluster 1 has 820 data points, cluster 2 has 226 data points and cluster 3 has 424 data points. Whereas, in K-Means clustering, cluster 1 has 820, cluster 2 has 424 and cluster 3 has 226 data points. In K-Means clustering, employee attrition rate is 15.60%, 16.37% and 16.98% for cluster 1, cluster 2 and cluster 3 respectively. Whereas in GMM clustering, employee attrition rate is 18.78%, 16.37% and 15.66% for cluster 1, cluster 2 and cluster 3 respectively. K-Means clustering performed better than GMM clustering as per evaluation by Silhouette Score. As per Elbow method, optimal number of clusters is 3 whereas as per AIC/BIC, optimal number of clusters is 6.

- Our main objective of performing clustering on the dataset, is to group employees with similar professional and personal characteristics. Then for each cluster, the percentage of employee attrition can be calculated from the class label. This helps the HR department to focus on employee groups with the highest attrition rate and take necessary actions (may be by adjusting salaries/ providing additional benefits/perks etc) to retain the employees.

- As the attrition rate doesn't differ much among 3 clusters in GMM, it is required by the HR department to check the employees of each cluster and identify the causes of attrition. KMeans performed better in a way that HR can focus more on only two clusters instead of 3.

- Another metric that we have used to measure the accuracy of the model is accuracy score. The Kmeans algorithm performed better over GMM when we ran it on train test split. The reason could be that GMM is a soft clustering method and is known to perform better on datasets like documents and images.

- On the other hand, Kmeans is a hard clustering method which means that a point wholly belongs to a single cluster with a 0 probability of being able to belong to another.

- The following deductions from KMeans still hold true for GMM
    - Higher the age, higher the monthly income.
    - Higher the number of working years, higher the monthly income.
    - Higher the performance rating, higher the percentage of salary hike.
    - Attrition rate of employees with higher income and older age is less than the the young employees with less income.