Data Science Asssignment

Sachin Negi
sachinnegi010997@gmail.com

June 11, 2021

1. Do the descriptive statistics and do the null value condition check, write inference on it.

1. Null value is checked in the dataset
No NaN value is obtained in the data.

2. Descriptive statistics

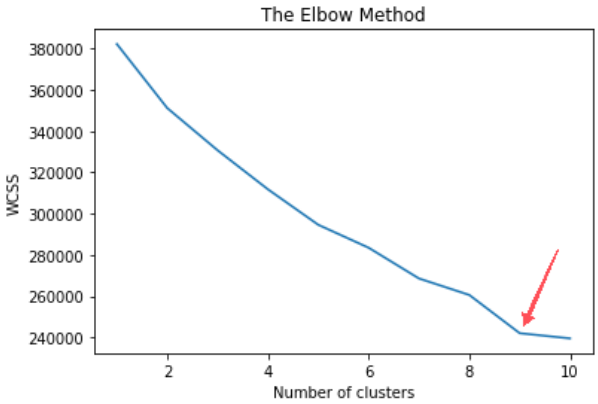
- **Measure of central tendency**
  Mean- Each column mean is calculated.
- **Measure of dispersion or variation**
  Standard Deviation- Standard deviation of each column is calculated.
- **Measure of position**
  Quartiles are values that divide your data into quarters.
  Q1(25 %), Q(50%), Q3(75%) is calculated for each column.
  Maximum and Minimum value in a column is calculated.

2. Apply the scaled data on the k-mean clustering algorithm, identify how many number of cluster is optimised cluster using the elbow graph and table which plot errors vs number of clusters

Preprocessing of Dataset

- The dataset consist of 18 features.
- Their are 14 numeric features and 4 categorical features
- The categorical features are converted to numeric.
- After the conversion the total number of features become 31.
- Standard scaling is done on the features.

Elbow graph is used to fix number of clusters

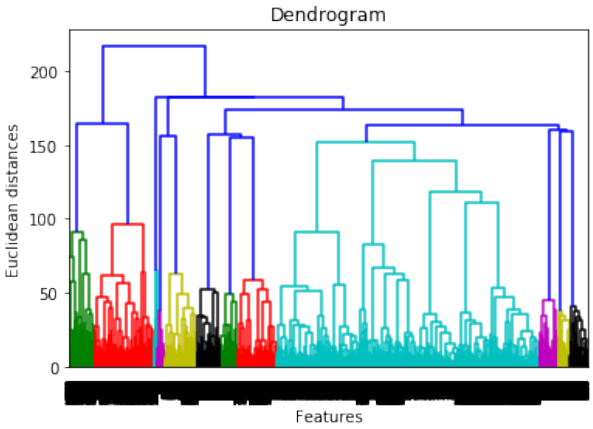


WCSS is the sum of squared distance between each point and the centroid in a cluster
Selecting the number of clusters where the change in WCSS begins to level off.
The value of k(number of cluster) =9

- The k-mean cluster is applied to obtain 9 different clusters
- The group are sorted based on distance of centroid from the centre. For this euclidean distance is calculated of the cluster's centroid.
- The lowest distance is given alphabet symbol A while the largest distance is given alphabet symbol H.

3. Apply the scaled data on the Hierarchial algorithm, identify how many number of cluster is optimized cluster using the dendrogram and table which plot error vs number of cluster, write a detailed inference.

The dendrogram is a visual representation of the compound correlation data. The individual compounds are arranged along the bottom of the dendrogram and referred to as leaf nodes. Compound clusters are formed by joining individual compounds or existing compound clusters with the join point referred to as a node.



- The dendrogram is used to decide number of cluster.
- The dendrogram diagram alone cannot tell optimum number of cluster to be chosen.

- One way to obtain optimised number of cluster is to look at changing Euclidean distance and corresponding number of cluster.
- We can observe increase in euclidean distance at each subsequent stage of clustering.
- By hit and trial method it can be seen that at 6-7 and 8-9 their is big difference in Euclidean distance.
- We can take number of cluster as 7 or 9.
- We will take 9 as number of cluster.

4. Selected the output of K-mean/hierarchial cluster using K-mean /Hierarchial either one, group the data based on cluster and give an infererence

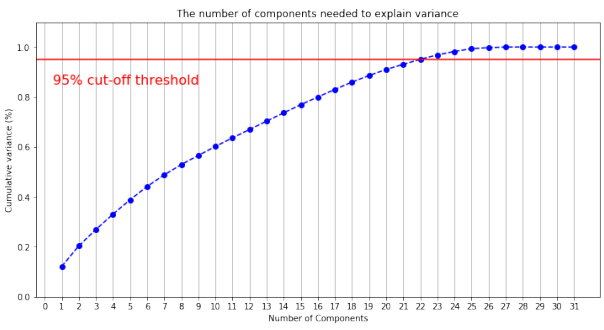
- Taken the k-mean clustering of the given dataset.
- All the entries in the month of Feb are in the same group.

5. Apply PCA with scaled data and identify the cluster using K-mean and Hierarchical, group the data based on the cluster and given an inference

Principal Component Analysis or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

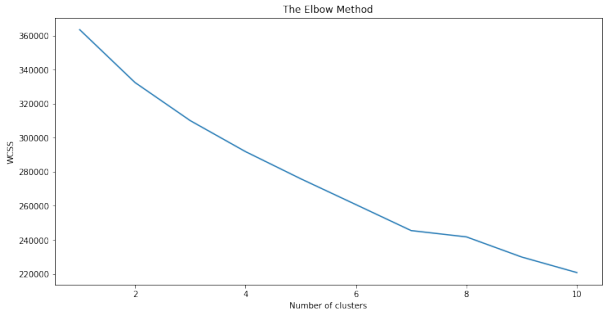
Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

- To select number of component for our PCA we select fixed variance.We typically use 95% or 99% variance explained.
- Graph is plotted between variance explained Vs number of components.
- Corresponding to 95% variance explained the number of component obtained is 22.

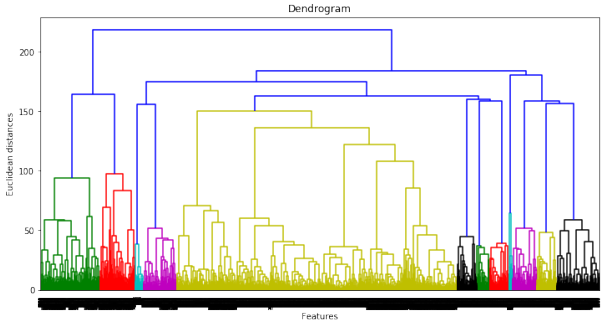


- PCA is applied on the dataset corresponding to 22 components.

- The elbow method is applied to obtain the number of clusters.



- The number of cluster obtained corresponding to PCA applied dataset is 7.
- K-mean clustering is applied to obtain the clusters.



- The dendrogram is plotted.
- The Euclidean distance is selected such that to obtain same number of cluster as that in K-mean clustering.
- Hierarchical Clustering is applied on the PCA applied dataset.
- One way to obtain optimised number of cluster is to look at changing Euclidean distance and corresponding number of cluster.
- We can observe increase in euclidean distance at each subsequent stage of clustering.
- After hit and trial We take number of cluster to be 7.

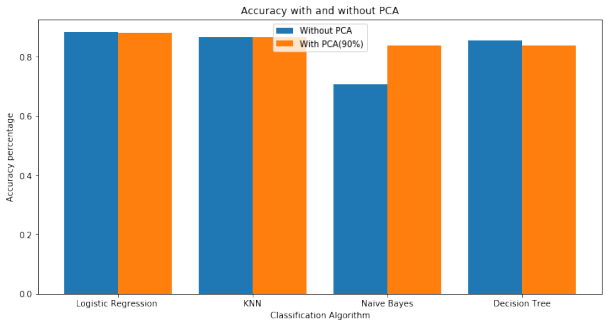
6. Split the data into train and test, apply data without PCA build classification model Logistic regression, KNN, Naives Bayes, Decision tree and write inference

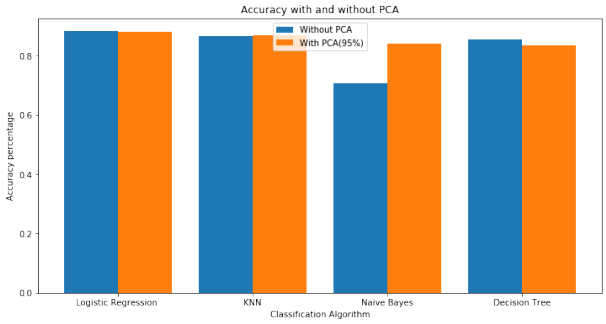
- As it is classification problem, the dataset last column corresponding to revenue column is taken as output.
- The dataset is spilt as train set and test set in the ratio of 80 : 20.
- Different classification algorithm is applied like logistic regression, KNN, Naive Bayes and Decision Tree.
- The confusion matrix is obtained for corresponding classification algorithm.
- Accuracy corresponding to each model is calculated.

7. Split the dataset into test and train, apply data after PCA build classification model Logistic regression, KNN, Naives Bayes, Decision tree and write inference

- The dataset is preprocessed as in previous case.
- PCA is applied taking 22 component as calculated before correspoding to 95 % variance explained.
- Different classification algorithm is applied like logistic regression, KNN, Naive Bayes and Decision Tree.
- The confusion matrix is obtained for corresponding classification algorithm.
- Accuracy corresponding to each model is calculated.
- Same study is done corresponding to 90 % variance explained.

8. Compare all the model and write an inference which model is best/optimized and give more insight about the final model





- Without PCA, the heighest accuracy is obtained for logistic regression, that is 88.28% and after applying PCA with (95%) marginally decrease.
- For KNN, the accuracy marginally decrease after applying PCA(95%) .
- For Naive Bayes their is significant improvement in accuracy after applying PCA(95%) .
- For Decision tree the accuracy decrease slightly after PCA(95%) .
- Similar result was obtained for PCA(90%).