Applied Machine Learning
Assignment 4

Sai Vishnu Kanisetty SXK175300

Objective of this exercise:
Experimentation of clustering, dimensionality reduction and using the reduced dimensions in neural networks on two datasets

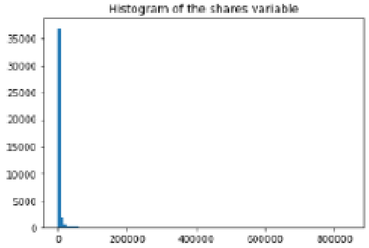
- Predict whether the articles published by Mashable will be shared more number of times or less
- Predict the Success of Bank Telemarketing

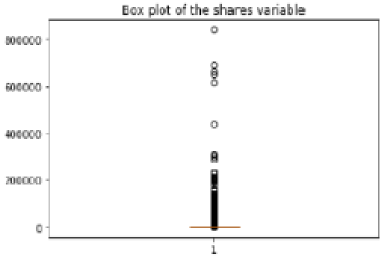
Step 1: Data Understanding for both the datasets

Mashable Dataset:

Mashable Dataset:

Data has 39,644 records of 61 variables and there are no missing values. Two non-predictive features (URL and Timedelta) have been dropped from further analysis. All the numerical predictor features have been normalized. Some information about the shares variable: Mean being more than the median and having values above the 2 standard deviations (box plot) shows that the data is positively skewed

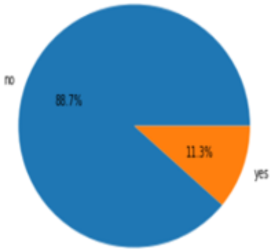




Median of the shares (1400) has been considered cutoff for the classification. The split of the articles to high shares (20,082) and low shares (19,562) will be ~50% because of considering median as the split point. Data has been divided into 70% train and 30% test in both the datasets.

Telemarketing Data:

Data published in UCI machine learning has been used [2]. Data has 20 variables that are likely to describe whether a customer has subscribed to the product in telemarketing or not. All the categorical variables have been converted using one-hot encoding. All the numerical variables have been normalized. Success rate of telemarketing is 11.3%. Data has been divided into 70% train and 30% test in both the datasets. This dataset is interesting because of the class imbalance present in predictor variable compared to Mashable dataset.

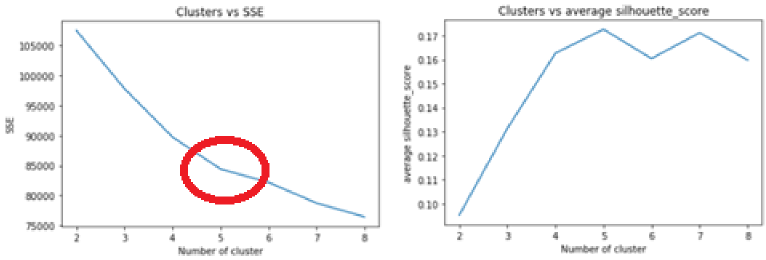


.

Step 2: Clustering

Mashable dataset

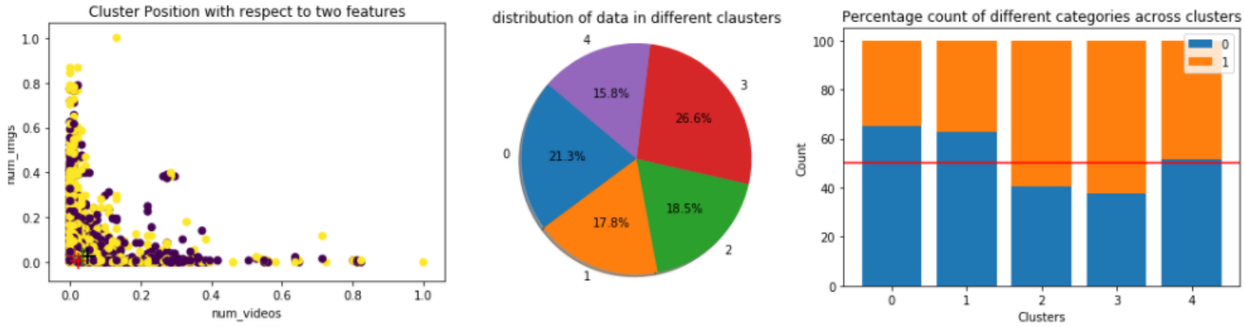
K-means: Identified number of clusters based on the SSE and the highest average silhouette value. Silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)



Based on the above plots, 5 clusters have been considered as optimal because of the elbow found in SSE plot and the highest average silhouette score

Cluster features:

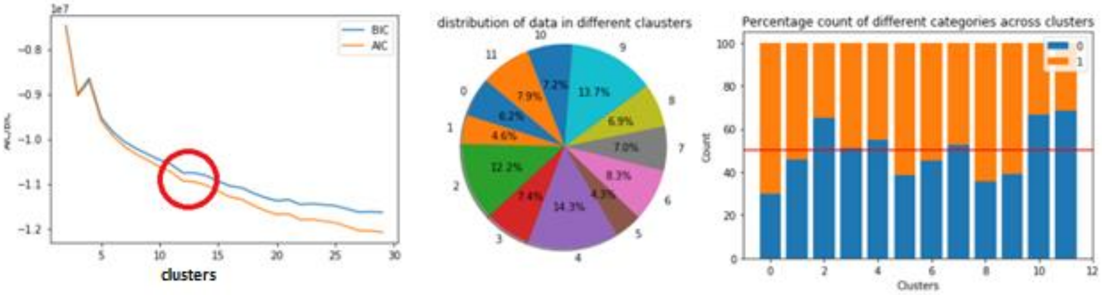
- Plot in the left of the below image, shows the cluster centers for two features “num_videos” and “num_images”. 5 cluster are very close to each other and they can be seen at the lower left edge of the plot with ‘+’ marks, and this is an indicator that they are not spread.Note: Yellow and brown are the colours of the class labels.
- Plot in the middle of the below image, shows the percentage of observations in each cluster and it has been observed to be approximately equal
- Plot in the right of the below image, shows that the percentage of two class labels in each clusters and it has been observed that none of the clusters are loaded on a single class label and distribution of class labels in all cluster is in the range of ~50%, which is the split of class labels in original data



Expectation Maximization: Number of cluster have been selected based on the BIC score at each cluster. BIC offers a relative estimate of the information lost when a given model is used to represent the process that generates the data.

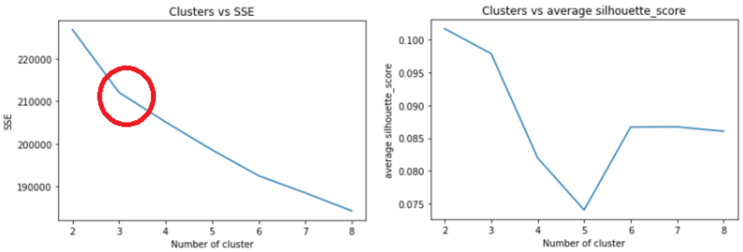
Cluster Features:

- Plot in the left of the below image, shows that after 12 clusters the decrease in BIC is marginal. So, 12 clusters have been considered as the optimal number of clusters
- Plot in the middle of the below image, shows the percentage of observations in each cluster and it has been observed that except two clusters (4 and 9), the percentage of observations are mostly the same
- Plot in the right of the below image, shows that the percentage of two class labels in each cluster and it has been observed that none of the clusters are loaded on a single class label and distribution of class labels in all cluster is in the range of ~50%, which is the split of class labels in original data



Telemarketing Data

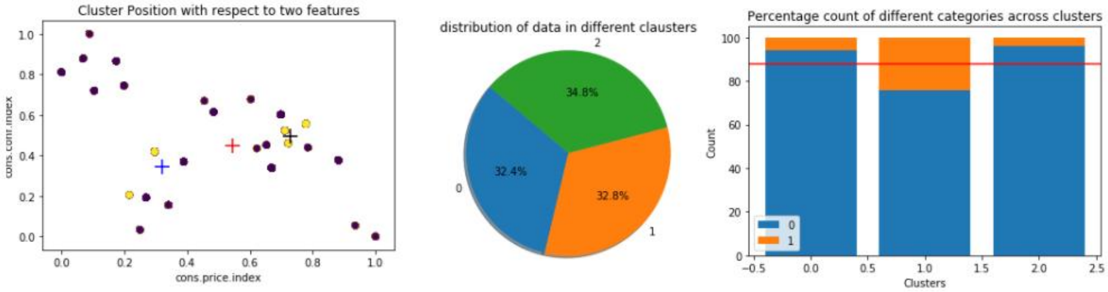
K-means: Identified number of clusters based on the SSE and the highest average silhouette value. Silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)



Based on the above plots, 3 clusters have been considered as optimal because of the elbow found in SSE plot and the 3 clusters also has the second highest average silhouette score

Cluster features:

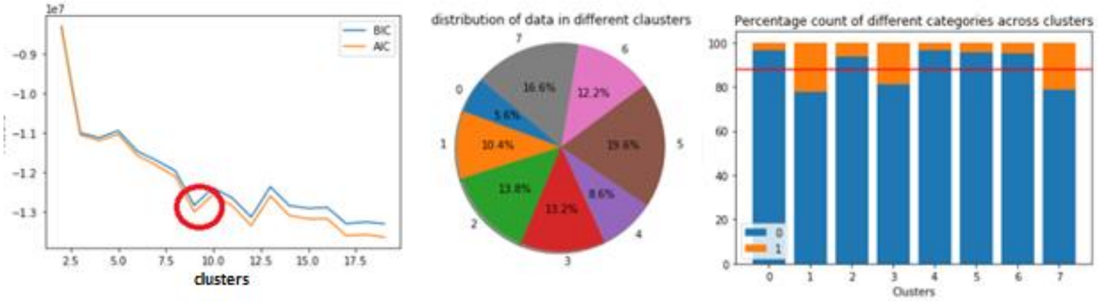
- Plot in the left of the below image, shows the cluster centers for two features “cons_price_index” and “cons_conf_index”. Clusters centers are marked by + and shows that the cluster centers are spread out
- Plot in the middle of the below image, shows the percentage of observations in each cluster and it has been observed to be approximately equal. Note: Yellow and brown are the colours of the class labels.
- Plot in the right of the below image, shows that the percentage of two class labels in each clusters and it has been observed that none of the clusters are loaded on a single class label and distribution of class labels in all cluster is in the range of ~88%, which is the split of class labels in original data



Expectation Maximization: Number of cluster have been selected based on the BIC score at each cluster. BIC offers a relative estimate of the information lost when a given model is used to represent the process that generates the data.

Cluster Features:

- Plot in the left of the below image, shows that after 8 clusters the decrease in BIC is marginal. So, 8 clusters have been considered as the optimal number of clusters
- Plot in the middle of the below image, shows the percentage of observations in each cluster and it has been observed that the distribution is varied
- Plot in the right of the below image, shows that the percentage of two class labels in each cluster and it has been observed that none of the clusters are loaded on a single class label and distribution of class labels in all cluster is in the range of ~88%, which is the split of class labels in original data

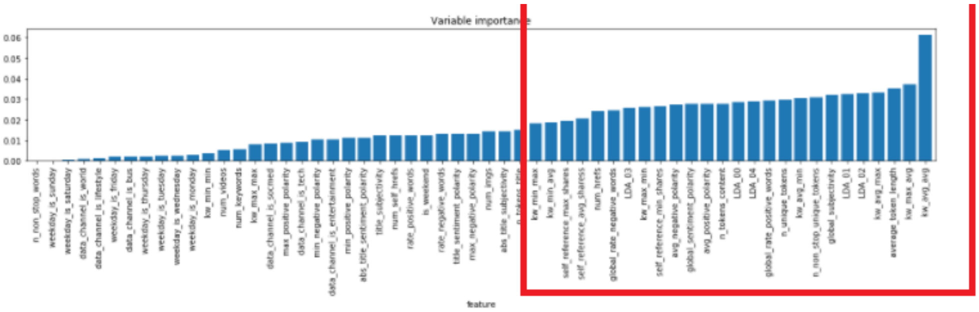


Step 3: Dimensionality Reduction

Feature Selection (Decision Trees): Decision tree classifier with entropy as the criteria has been used and features that have more than mean **importance in splitting the nodes** have been retained.

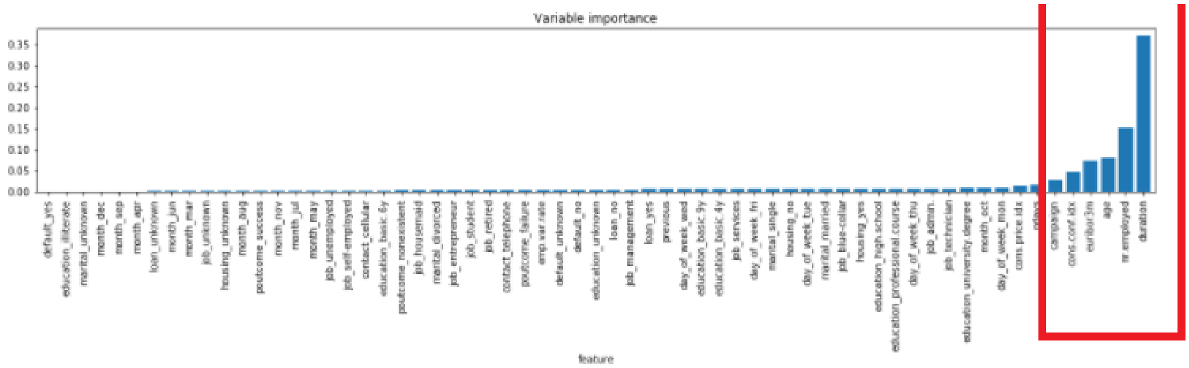
Mashable dataset

26 out of 58 features have been retained. Variables importance of all the features is represented in the below image and retained feature names are the ones marked in the red box.



Telemarketing dataset

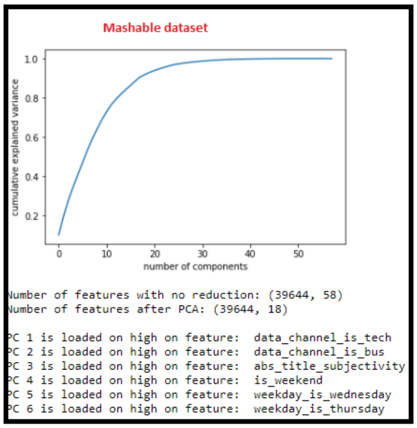
6 out of 63 features have been retained. Variables importance of all the features is represented in the below image and retained feature names are the ones marked in the red box.



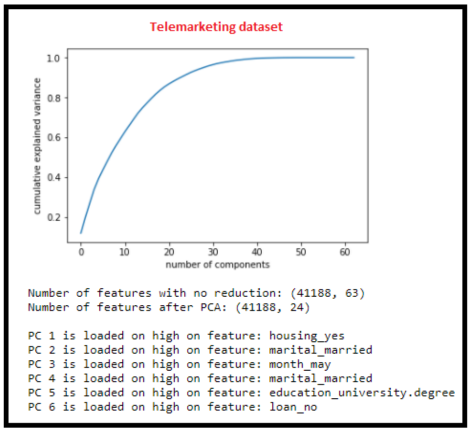
PCA: Retained variables that explain cumulative 90% of the variance in data.

Data in transformed domain: The new components will be orthogonal to each other.

The image below has the number of features retained and the features on which the principal components are loaded on for both the datasets. PCA has 18 and Telemarketing has 24 features in new domain.

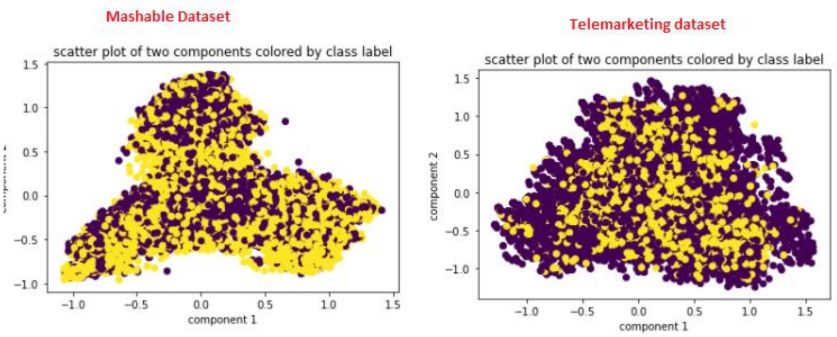


```
Number of features with no reduction: (39644, 58)
Number of features after PCA: (39644, 18)

PC 1 is loaded on high on feature:  data_channel_is_tech
PC 2 is loaded on high on feature:  data_channel_is_bus
PC 3 is loaded on high on feature:  abs_title_subjectivity
PC 4 is loaded on high on feature:  is_weekend
PC 5 is loaded on high on feature:  weekday_is_wednesday
PC 6 is loaded on high on feature:  weekday_is_thursday
```



```
Number of features with no reduction: (41188, 63)
Number of features after PCA: (41188, 24)

PC 1 is loaded on high on feature: housing_yes
PC 2 is loaded on high on feature: marital_married
PC 3 is loaded on high on feature: month_may
PC 4 is loaded on high on feature: marital_married
PC 5 is loaded on high on feature: education_university.degree
PC 6 is loaded on high on feature: loan_no
```

Plot of two components in both datasets colored by the class label.

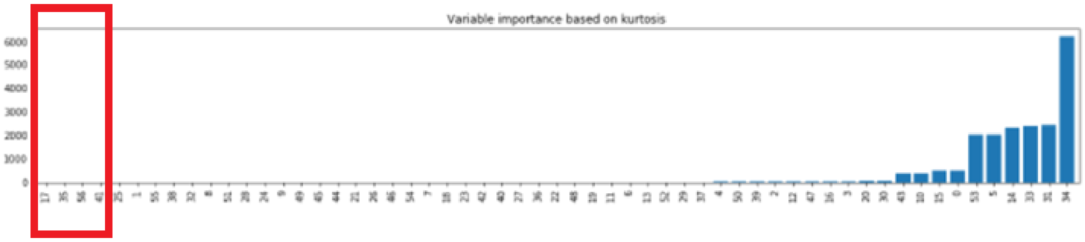


ICA: Components that have kurtosis in the range of (+1 to -1) have been retained. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution and using this measure can help in removing features that deviate from normal distribution. **Note:** Fisher's definition has been used for kurtosis, which means 3.0 is subtracted from the result to give 0.0 for a normal distribution.

Data in transformed domain: Similarity between new components will be very minimal.

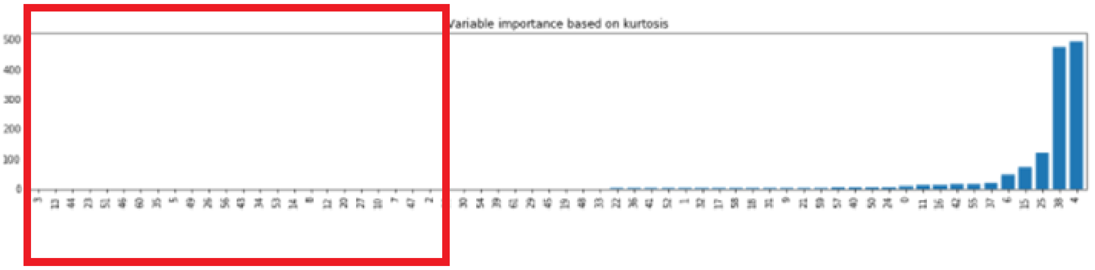
Mashable dataset

Four components have been retained and kurtosis values are represented in the below image. From the plot we can see that some components have very high kurtosis value. Components in the red box have been retained.



Telemarketing dataset

24 components have been retained and kurtosis values are represented in the below image. From the plot we can see that some components have very high kurtosis value. Components in the red box have been retained.



Randomized Projections For both the datasets, random weightage of the existing features has been done and half the number of existing features has been considered as the optimal number of components. So, the number of components for Mashable dataset is 29 and number of components for Telemarketing dataset is 31.

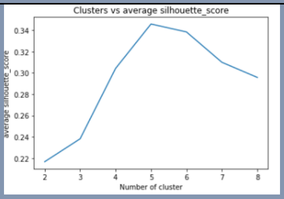
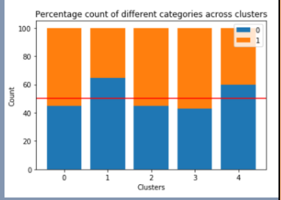
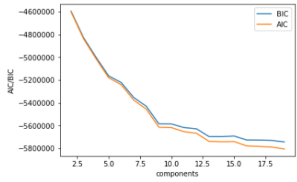
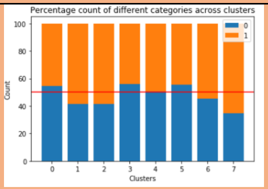
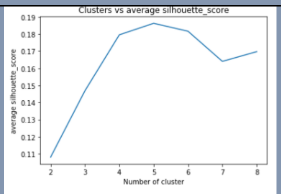
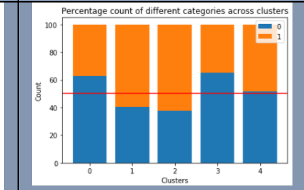
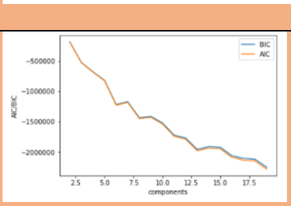
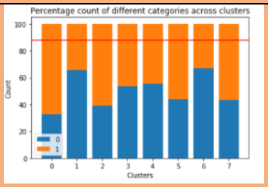
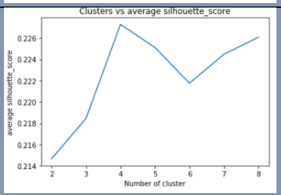
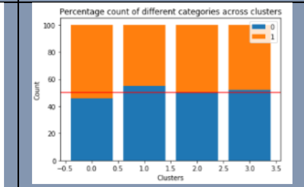
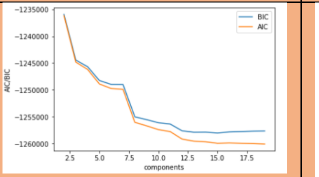
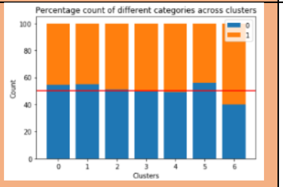
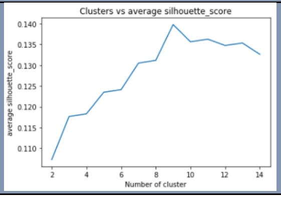
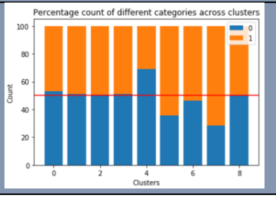
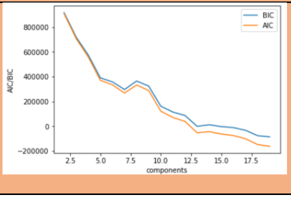
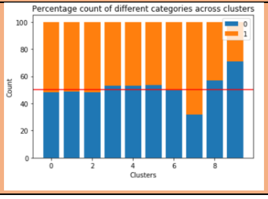
Data in transformed domain: Even though the linear combinations of features are random, RCA still manages to pick up correlation between features

Step 3: Dimensionality Reduction + Clustering

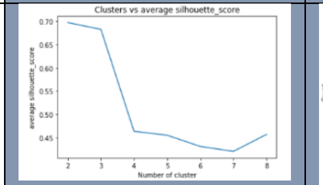
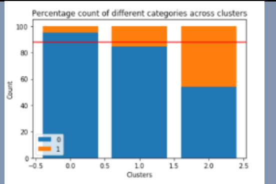
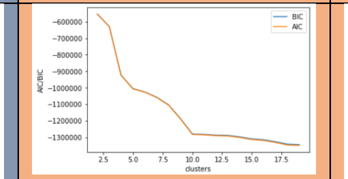
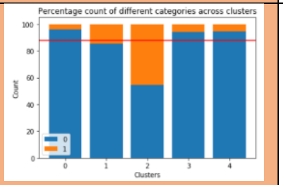
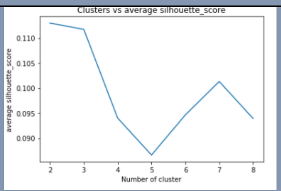
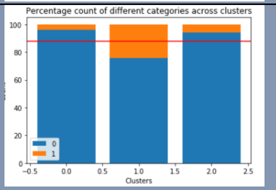
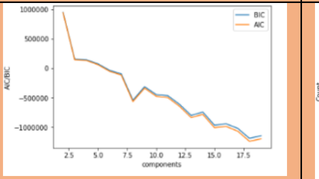
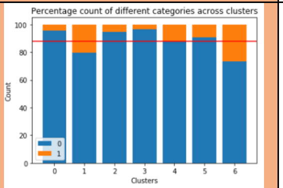
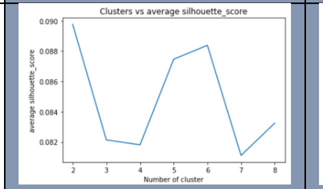
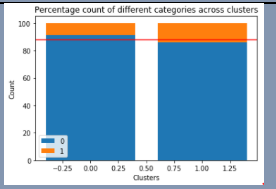
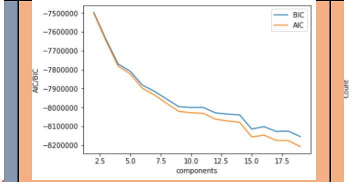
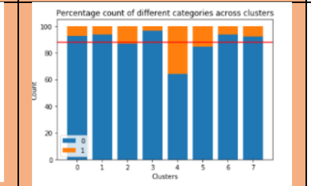
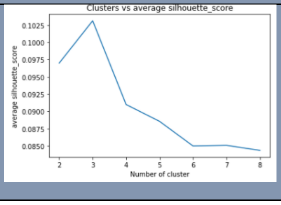
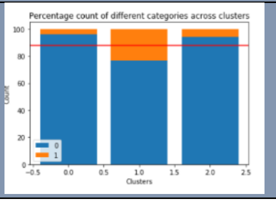
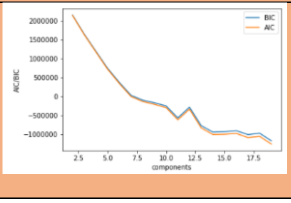
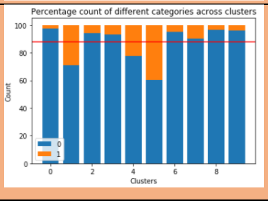
Please look at the ipython notebook for full images.

The images below represent the clustering (K Means and Expectation Maximization) results on 4 different datasets obtained by using different dimensionality reduction techniques.

Mashable Dataset (Fig 1)

Dataset	K means clusters (silhouette index)	Kmeans Cluster Features	Expectation Maximization Clusters (BIC)	Expectation Maximization cluster features	Optimal Number of clusters
Feature Selection					Kmeans: 5 EM: 8
PCA					Kmeans: 5 EM: 8
ICA					Kmeans: 4 EM: 7
RA					Kmeans: 9 EM: 10

Telemarketing Dataset (Fig 2)

Dataset	K means clusters (silhouette index)	Kmeans Cluster Features	Expectation Maximization Clusters (BIC)	Expectation Maximization cluster features	Optimal Number of clusters
Feature Selection					Kmeans: 3 EM: 5
PCA					Kmeans: 3 EM: 7
ICA					Kmeans: 2 EM: 8
RP					Kmeans: 3 EM: 10

Summary

Mashable			
data	features	Kmeans optimal clusters	EM optimal clusters
full dataset	58	5	12
Feature Selection Decision Tree	26	5	8
PCA	18	5	8
ICA	4	4	7
RA	29	9	10

TeleMarketing			
data	features	Kmeans optimal clusters	EM optimal clusters
full dataset	63	3	8
Feature Selection Decision Tree	6	3	5
PCA	24	3	7
ICA	24	2	8
RA	31	3	10

Mashable dataset: ICA has the least number of features (4) and the optimal clusters for both K Means and Expectation Maximization are also the least for ICA. Even the K-means and Expectation Maximization cluster features (mentioned in Fig 1) for ICA shows that all that the class labels are loaded equally across all the clusters and the percentage loading is ~50%, which is the class label distribution in original dataset.

Telemarketing dataset: Feature selection (decision tree) has the least number of features (6) and the optimal clusters for Expectation Maximization is also the least for Feature selection. Much variation has not been observed for number of optimal clusters using K-Means across different datasets.

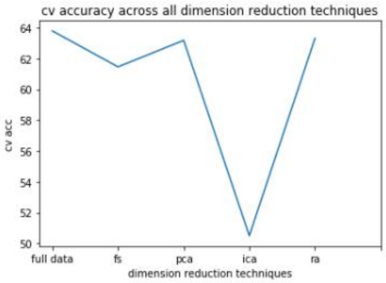
Step 4: Neural Networks

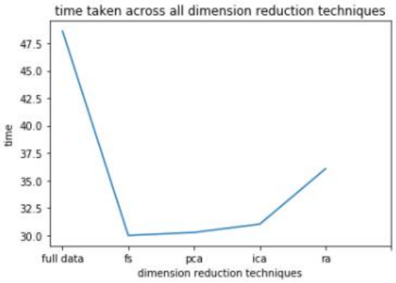
Using the best hyperparameters identified in last assignment, neural networks has been executed on full dataset, feature selection dataset (fs), pca dataset, ica dataset, random projection(RA) dataset. Because there is no experimentation in this exercise only the final cv accuracy obtained is displayed and the learning curves are not, because learning majorly help in identifying bias vs variance scenario, which is not a requirement for this assignment.

Mashable dataset

Hyperparameters used: Two layer neural networks architecture with below mentioned hyperparameters:

{'batch_size': [64],'epochs': [10],'optimizer': ['rmsprop'],'first_hidden_units':[20],'second_hidden_units':[10],'hidden_activation_fun':['relu'],'output_activation_func':['sigmoid']}





Summary:

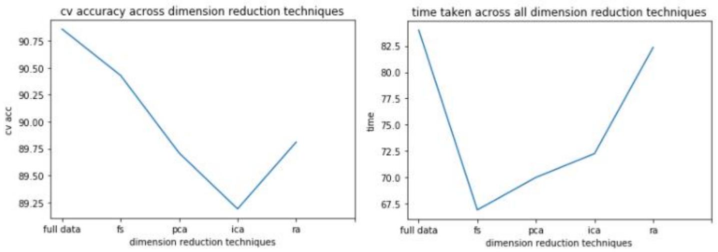
Three-fold CV accuracy: Three-fold cv is highest using the full data and the second highest is for the random projections. This shows Neural networks is not suffering from the curse of dimensionality because full data has the highest number of features compared to any other dimensionality reduction datasets.

Time taken: Time taken also shows that full data takes the highest time and the second highest is random projections (ra) and this can be related to the number of components in full data set (58) being the highest and second highest being the random projections dataset (29).

Telemarketing dataset

Hyperparameters used: Two layer neural networks architecture with below mentioned hyperparameters:

{'batch_size': [32],'epochs': [10],'optimizer': ['adam'],'first_hidden_units':[30],'second_hidden_units':[15],'hidden_activation_fun':['relu'],'output_activation_func':['sigmoid']}



Summary:

Three-fold CV accuracy: Three-fold cv is highest using the full data and the second highest is for the random projections. Relating the three-fold cv results with the number of features (mentioned in step 3 summary for telemarketing data), shows that, as the number of features increases, the cv accuracy also increases. This shows that Neural networks is not suffering from the curse of dimensionality.

Time taken: Time taken also shows that full data takes the highest time and the second highest is random projections (ra).

Step 5: Cluster Results + Neural Networks

Here we are trying the estimate the usage of identified clusters as reduced dimensions and implement Neural Networks using them.

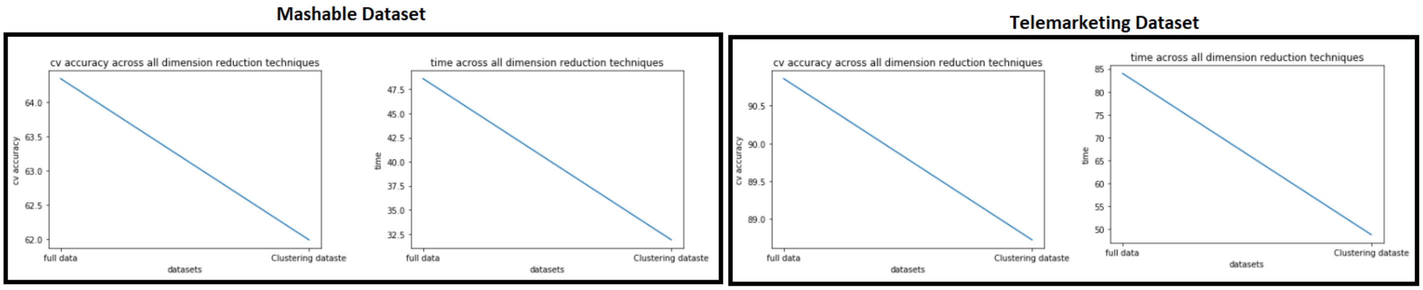


FIG 3

Mashable dataset: In step 2, we have identified the optimal number of clusters for k means as 5 and Expectation Maximization as 12. Created a dataset with 13 columns:

- One columns for k Means, because it being a hard cluster
- 12 columns for probability values assigned by each cluster to all observations

Summary Cv accuracy using clustering (62.03%) is close to Cv accuracy using the full data set (64%) and the time taken using clustering (27 sec) is nearly half the time taken using the full dataset (49 sec).

Telemarketing dataset: In step 2, we have identified the optimal number of clusters for k means as 3 and Expectation Maximization as 8. Created a dataset with 9 columns:

- One columns for k Means, because it being a hard cluster
- 8 columns for probability values assigned by each cluster to all observations

Summary Cv accuracy using clustering (88.72%) is close to Cv accuracy using the full data set (91%) and the time taken using clustering (47 sec) is nearly half the time taken using the full dataset (84 sec).

Next Steps: Observe the change in accuracy using neural networks after:

1. Increase the percentage of variance explained in PCA
2. Take less number of random projections