BUAN 6341.501 - Applied Machine Learning
Assignment 4 - Pavan Gorantla (PCG180000)

Objective

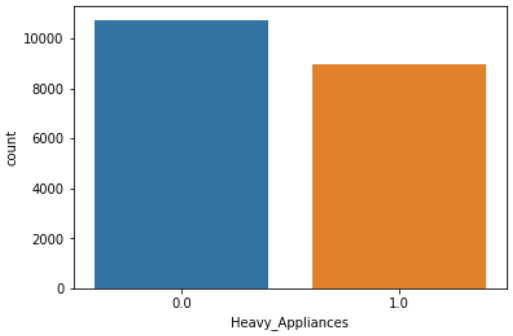
Experimentation of Clustering Algorithms (K-means and Expectation Maximization) and Dimensionality Reduction Algorithms (Feature Selection using Decision Tree, PCA, ICA and RP) to predict the class labels outputs of two datasets which have been downloaded from archives of UCI Machine Learning Repository.

Dataset 1 - Appliances Energy Prediction (Link: appliances+energy+prediction)
Dataset consists of 19735 observations on a total of 29 variables. More information about these variables can be found in the above given weblink. The response variable is created using Appliances variable where appliances greater than median are classified as Heavy Appliances and the goal is to predict this classifier.

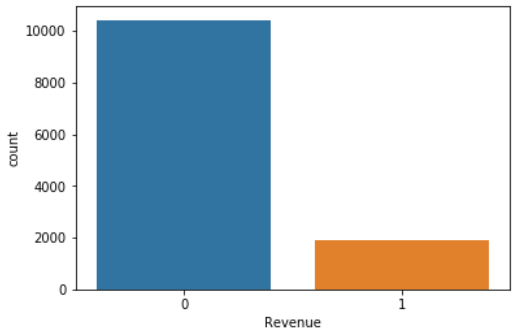
Dataset 2 - Online Shoppers Purchasing Intention (Link: online+shoppers+purchasing+intention)
Dataset consists of 12330 observations on a total of 18 variables. More information about these variables can be found in the above given weblink. The response variable Revenue specifies if an online user on the shopping website has purchase intent or not based on the pages visited and duration spent on each page.

Normalization and Train/Test

Both datasets features are normalized to a range of [0,1] before carrying out the dimension reduction algorithms. Target from Dataset 1 is classified as 1 for Heavy Appliances and 0 for the remaining. Response from Dataset 2 is classified into 1 for users with purchase intent and 0 for the remaining users. All features in Dataset 1 are numeric and hence they are just normalized for our clustering algorithms. There are few categorical features in Dataset 2 which are transformed into dummies before normalizing. Below are the classes distribution plots for both datasets. Train & Test split is done before doing neural network analysis.



1 - Appliances Energy Prediction



2 - Online Shoppers Purchase Intent Prediction

Experiments done on both Datasets

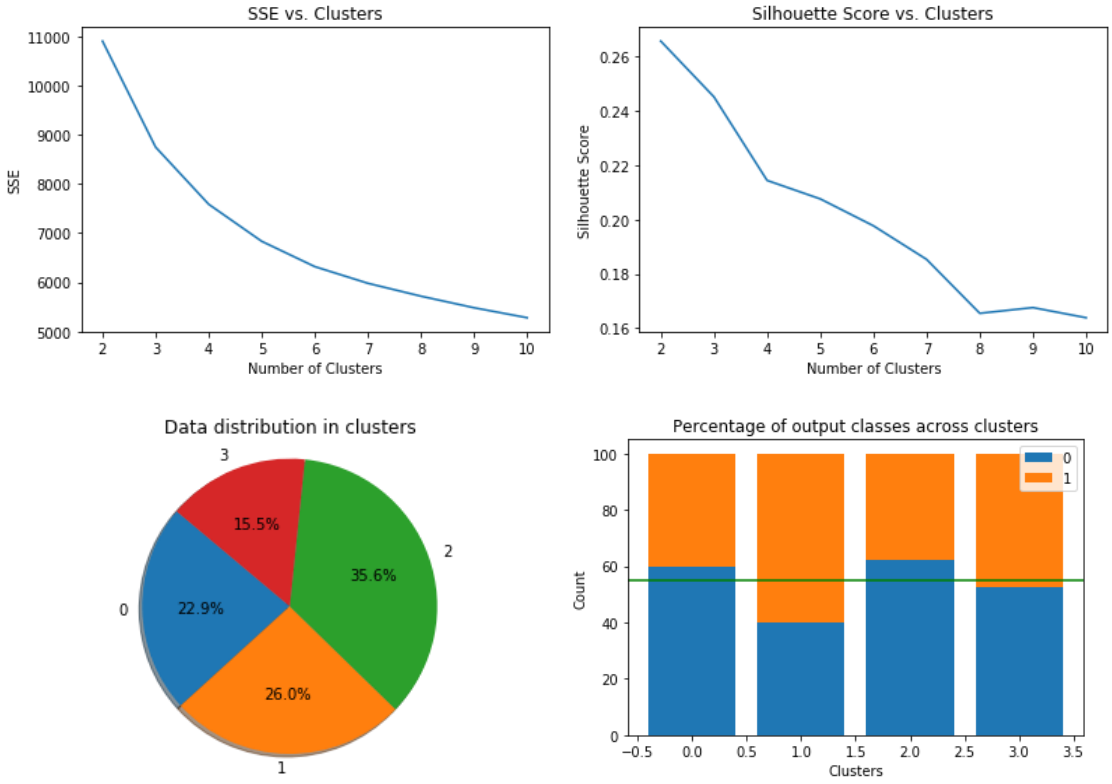
Task 1: K-Means and Expectation Maximization using all available features before Dimensions Reduction
Task 2: Dimensionality Reduction Algorithms like feature selection using Decision Trees, PCA, ICA and RP
Task 3: Comparing clustering algorithms like K-Means and EM on reduced features from DT, PCA, ICA, RP
Task 4: Neural Networks with inputs as reduced features from various algorithms (Tree, PCA, ICA and RP)
Task 5: Neural Networks with inputs as the clustering results from K-Means and Expectation Maximization

Summary and Results

- Independent Components Analysis is found to be best reduction technique for both the datasets
- Using reduced dimensions for neural networks did not improve accuracy but improved the runtime
- Using cluster results as inputs for neural networks did not improve accuracy but improved runtime

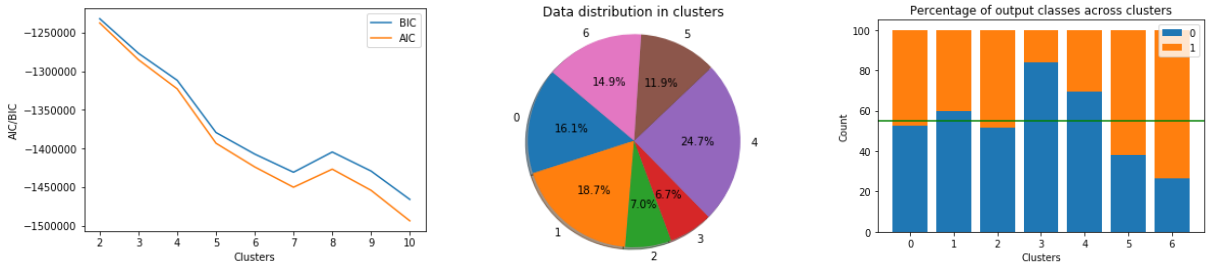
Task 1 - Clustering Algorithms (Dataset 1 - Heavy Appliances Prediction)

K-Means: Based on SSE and highest average silhouette value which is a measure of similarity of an object to its own cluster compared to others. From below plots, we can say that optimal number of clusters is 4.



Left plot shows percentage of observations in each cluster and we can see that all the clusters are almost similar in size. Right plot shows percentage of two output classes in each cluster and we can see that none of them are loaded on single class label and distribution of classes is in the range of 55%, like the original.

Expectation Maximization: Identified 7 as the optimum number of clusters for Expectation Maximation using BIC score which is an estimate of information lost when a given model is used represent the process.



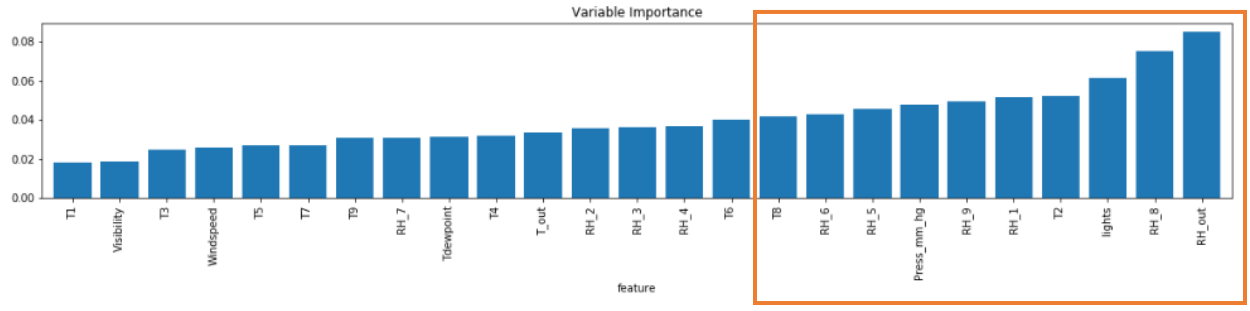
Left plot shows that after 7 clusters, the decrease in BIC is marginal. So, we have considered the optimal number of clusters to be 7. Middle plot shows percentage of observations in each cluster and we can see that except clusters 2 and 3, the percentage of observations are almost similar. Right plot shows percent split of two classes in each cluster and none of the clusters are loaded on single class label and distribution of output class labels in all clusters is in the range of 55%, which is the split of class labels in original data.

Task 2 - Dimensionality Reduction (Dataset 1 - Heavy Appliances Prediction)

Feature Selection using Decision Tree

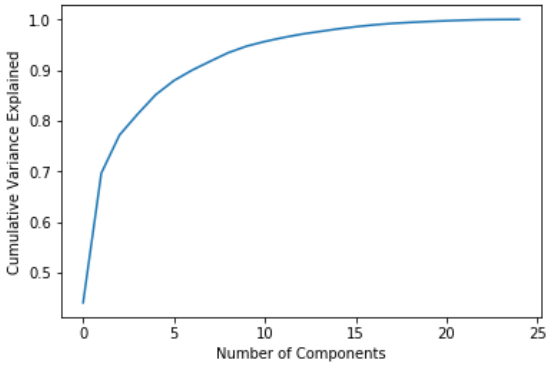
10 out of 25 features have been retained and importance of all features is represented in the below graph

```
Number of features before transformation: (19735, 25)
Number of features after transformation: (19735, 10)
Columns retained: Index(['lights', 'RH_1', 'T2', 'RH_5', 'RH_6', 'T8', 'RH_8', 'RH_9',
       'Press_mm_hg', 'RH_out'],
      dtype='object')
```



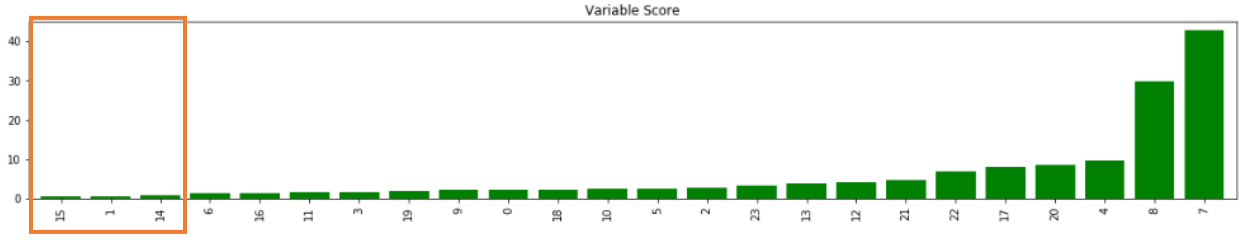
Principal Components Analysis

Retained variables that explain cumulative 90% of the variance in data. Below plot shows the number of features retained. Using Principle Components Analysis, we retained 8 new orthogonal features out of 25.




```
Number of features with no reduction: (19735, 25)
Number of features after PCA:
 (19735, 8)
```

Independent Components Analysis



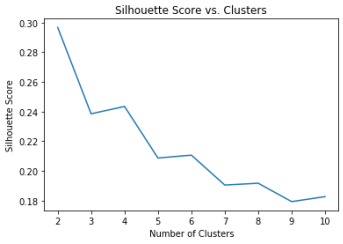
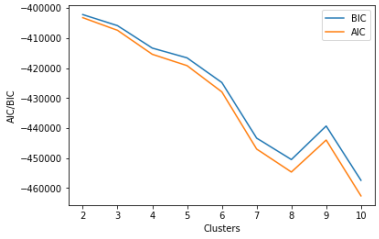
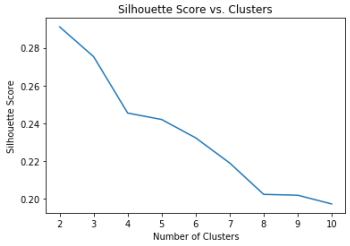
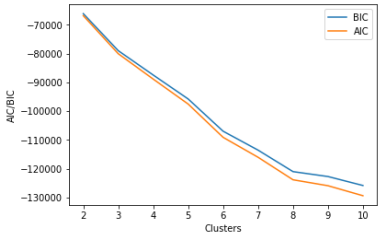
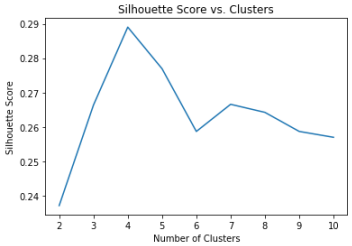
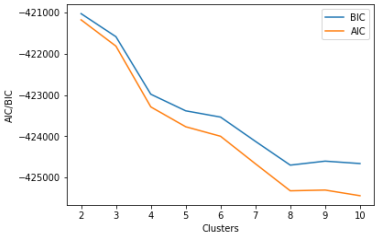
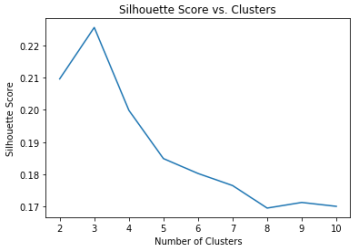
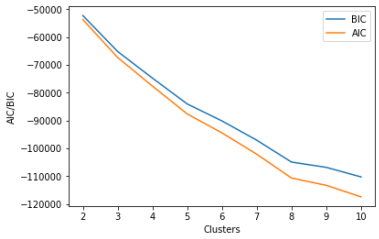

```
Number of features after ICA (19735, 3)
```

Only 3 components have been retained based on kurtosis values shown above. From the plot we can see that some components have very high kurtosis value and components in the red box have been retained.

Randomized Projections: Random weightage of existing features is done and 12 of the existing features are considered as optimal number of components. RCA manages to pick up correlation between features.

Task 3 - Clustering after Dimensions Reduction (Dataset 1 - Heavy Appliances Prediction)

Following plots show the clustering results after using four different dimensionality reduction techniques.

Method	K-means (Silhouette)	Expectation Maximization (BIC)	Optimal Clusters
Tree			**K-Means - 4** **EM - 8**
PCA			**K-Means - 4** **EM - 8**
ICA			**K-Means - 4** **EM - 8**
RP			**K-Means - 3** **EM - 8**

Following table gives summary of all the above observations regarding the dimension reduction methods.

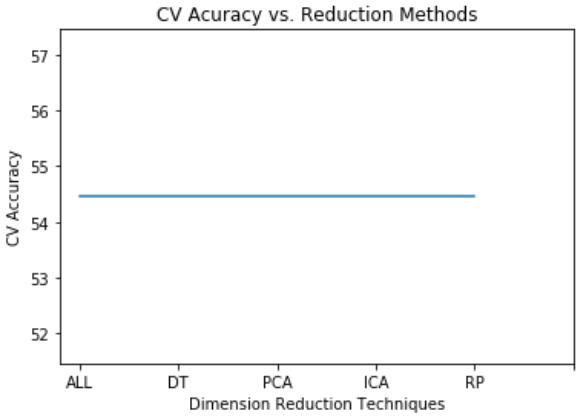
Method	K-Means Clusters	Expectation Maximize Clusters	Retained Features
Feature Selection with Tree	**4**	**8**	**10**
Principal Components	**4**	**8**	**8**
Independent Components	**4**	**8**	**3**
Randomized Projections	**3**	**8**	**12**

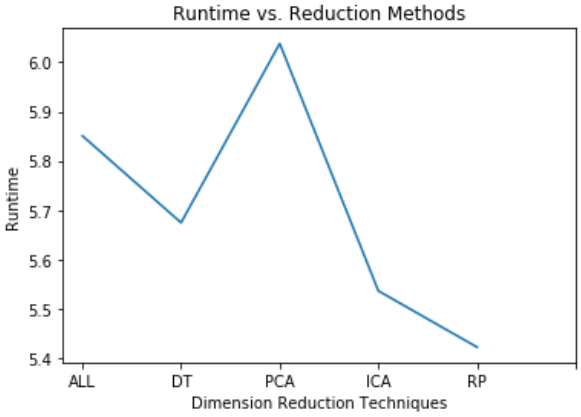
We can observe that numbers of clusters are uniform for all reduction techniques but ICA has successfully reduced the features into just 3 independent components. The next best method is PCA with the features reduced to 8 orthogonal components. Feature Selection with decision tree gave out 10 retained features.

Task 4 - Neural Networks after Dimensions Reduction (Dataset 1 - Heavy Appliances Prediction)

Using the best parameters found in the neural networks experiments from previous assignments, the Cross-Validation error and Runtime are found for different features retained from the reduction methods.

```
{'batch_size': [256], 'epochs': [20], 'optimizer': ['sgd'], 'first_hidden_units': [10], 'second_hidden_units': [10], 'hidden_ac
tivation_fun': ['tanh'], 'output_activation_func': ['sigmoid']}
```

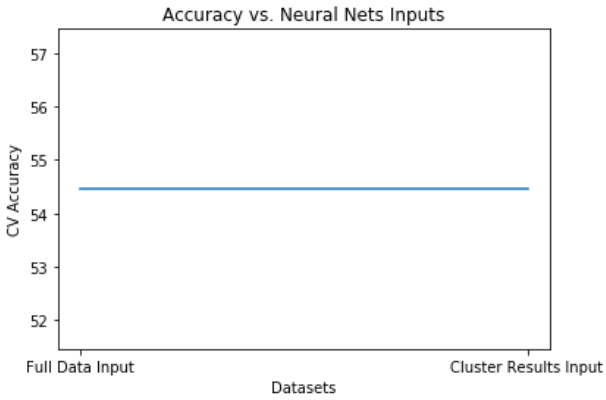


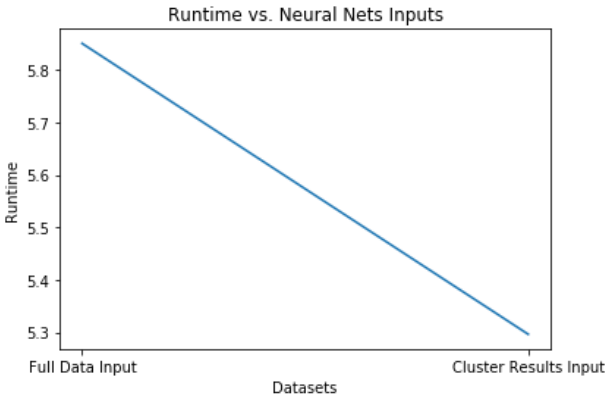


We don't see variation in Cross-Validation error but RP is efficient and takes very less time to run. Using PCA takes more computation time, but results in the features being reduced to 8 orthogonal components.

Task 5 - Neural Networks using Clustering Results (Dataset 1 - Heavy Appliances Prediction)

Using the best parameters found in the neural networks experiments from the previous assignments, the CV error and Runtime are found with original features and cluster results obtained from Task 1 as the inputs for neural networks with the best parameters found from the previous assignment.

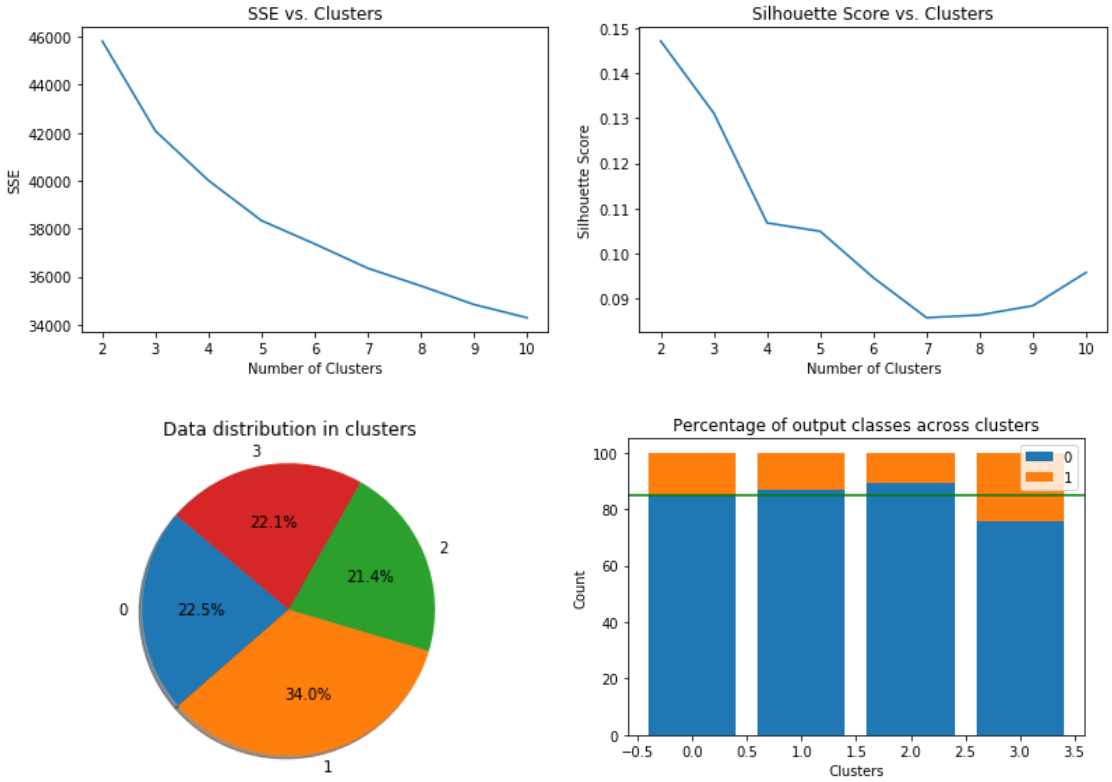




We don't see any variation in Cross-Validation error but we can clearly find that the neural network with cluster results as inputs is more efficient taking less runtime than the original features. Since, we don't see any kind of variation in the accuracy we may use cluster results and save computation cost and time.

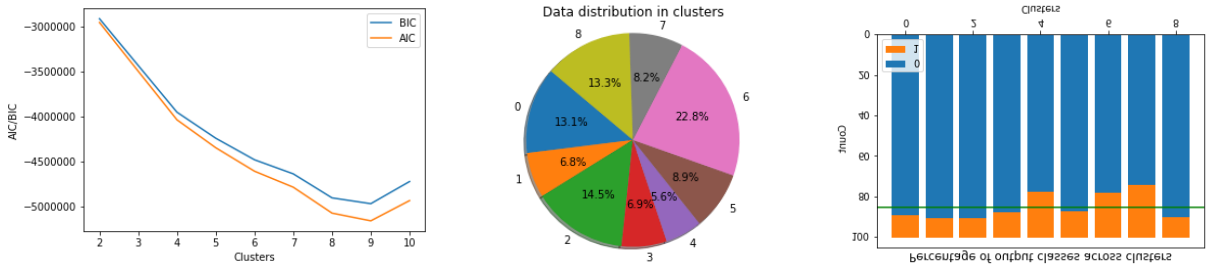
Task 1 - Clustering Algorithms (Dataset 2 - Online Shopping Intention Prediction)

K-Means: Based on SSE and highest average silhouette value which is a measure of similarity of an object to its own cluster compared to others. From below plots, we can say that optimal number of clusters is 4.



Left plot shows percentage of observations in each cluster and we can see that all the clusters are almost similar in size. Right plot shows percentage of two output classes in each cluster and we can see that none of them are loaded on single class label and distribution of classes is in the range of 85%, like the original.

Expectation Maximization: Identified 9 as the optimum number of clusters for Expectation Maximation using BIC score which is an estimate of information lost when a given model is used represent the process.



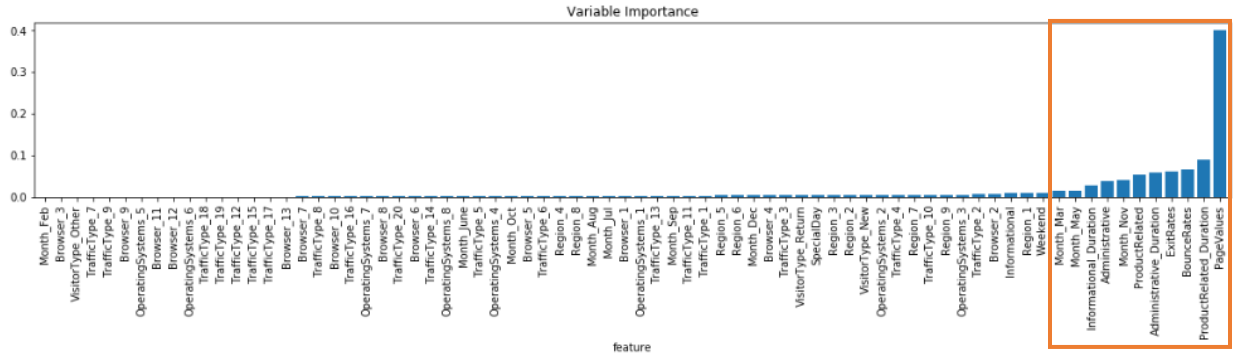
Left plot shows that after 9 clusters, the decrease in BIC is marginal. So, we have considered the optimal number of clusters to be 9. Middle plot shows percentage of observations in each cluster and we can see that 1,3,4,5,7 clusters size is similar and 0,2,6,8 cluster size is similar. Right most plot shows percent split of two classes in each cluster and none of the clusters are loaded on single class label and distribution of output class labels in all the clusters is in the range of 85%, which is the split of class labels in original data.

Task 2 - Dimensionality Reduction (Dataset 2 - Online Shopping Intention Prediction)

Feature Selection using Decision Tree

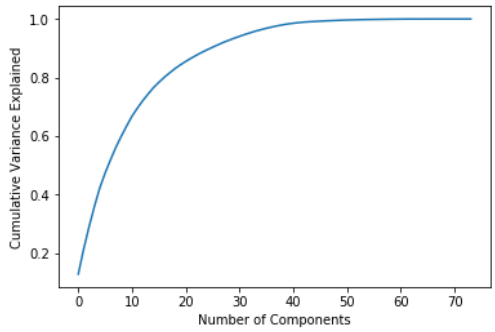
11 out of 74 features have been retained and importance of all features is represented in the below graph.

```
Number of features before transformation: (12330, 74)
Number of features after transformation: (12330, 11)
Columns retained: Index(['Administrative', 'Administrative_Duration', 'Informational_Duration',
       'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates',
       'PageValues', 'Month_Mar', 'Month_May', 'Month_Nov'],
      dtype='object')
```



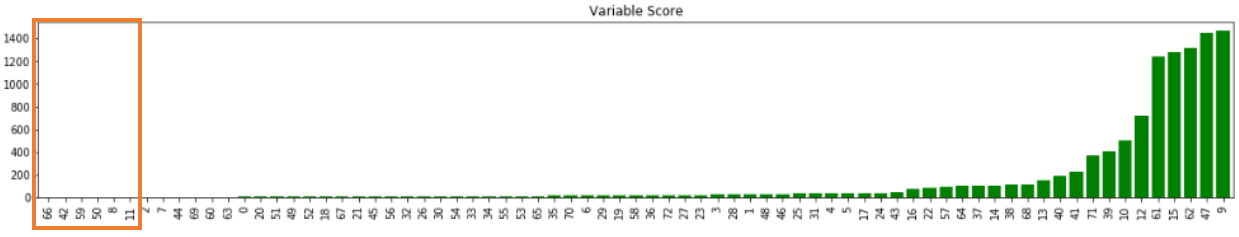
Principal Components Analysis

Retained variables that explain cumulative 90% of variance in data. Below plot shows number of features retained. Using Principle Components Analysis, we have retained 26 new orthogonal features out of 74.




```
Number of features with no reduction: (12330, 74)
Number of features after PCA:
 (12330, 26)
```

Independent Components Analysis



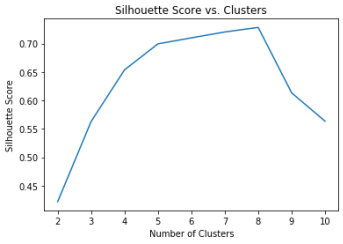
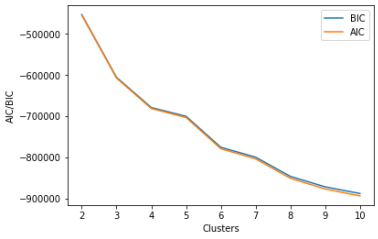
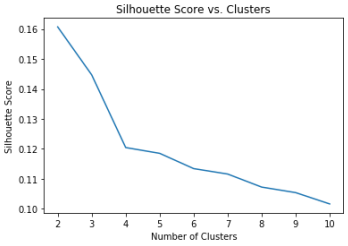
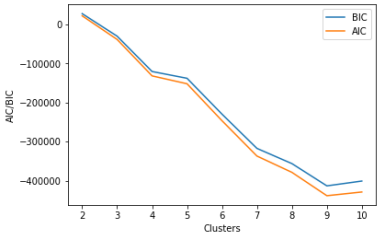
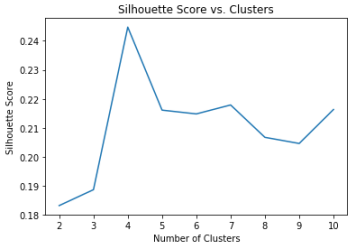
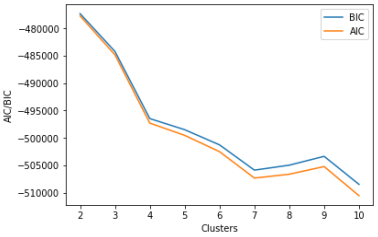
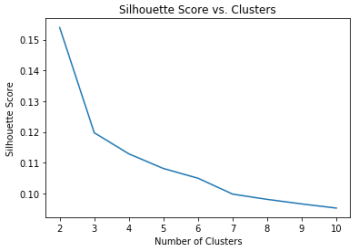
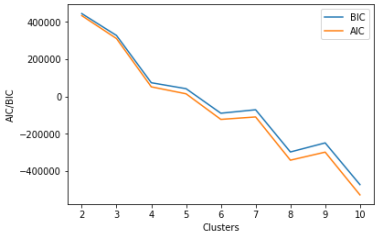

```
Number of features after ICA (12330, 6)
```

Only 6 components have been retained based on kurtosis values shown above. From the plot we can see that some components have very high kurtosis value and components in the red box have been retained.

Randomized Projections: Random weightage of existing features is done and 37 of the existing features are considered as optimal number of components. RCA manages to pick up correlation between features.

Task 3 - Clustering after Dimensions Reduction (Dataset 2 - Online Shopping Intention Prediction)

Following plots show the clustering results after using four different dimensionality reduction techniques.

Method	K-means (Silhouette)	Expectation Maximization (BIC)	Optimal Clusters
Tree			**K-Means - 5** **EM - 6**
PCA			**K-Means - 4** **EM - 9**
ICA			**K-Means - 4** **EM - 7**
RP			**K-Means - 3** **EM - 8**

Following table gives summary of all the above observations regarding the dimension reduction methods.

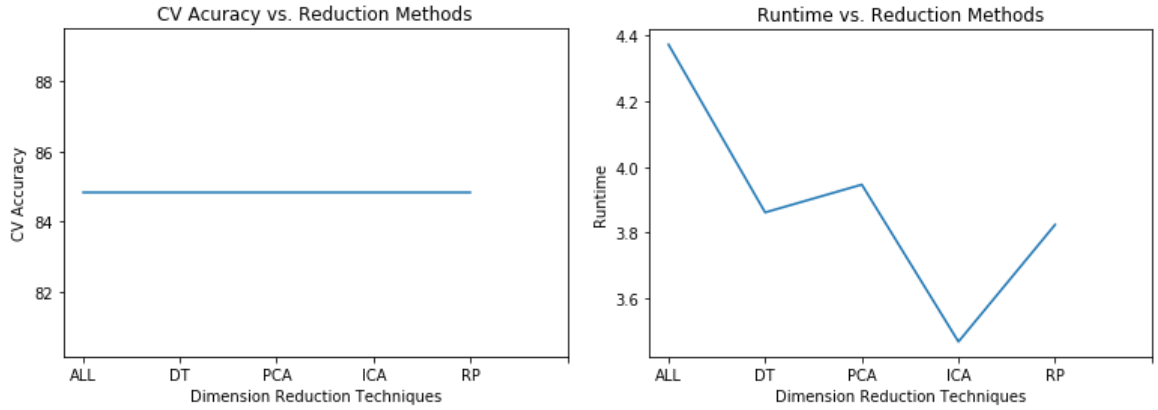
Method	K-Means Clusters	Expectation Maximize Clusters	Retained Features
Feature Selection with Tree	**5**	**6**	**11**
Principal Components	**4**	**9**	**26**
Independent Components	**4**	**7**	**6**
Randomized Projections	**3**	**8**	**37**

We can observe that numbers of clusters are uniform for all reduction techniques but ICA has successfully reduced the features into just 6 independent components. The next best method is Tree with the features reduced to 11 features along with their importance. PCA gave out 11 orthogonal components as retained.

Task 4 - Neural Networks after Dimensions Reduction (Dataset 2 - Online Shopping Intention Prediction)

Using the best parameters found in the neural networks experiments from the previous assignments, the CV error and Runtime are found for different features retained from different reduction algorithms.

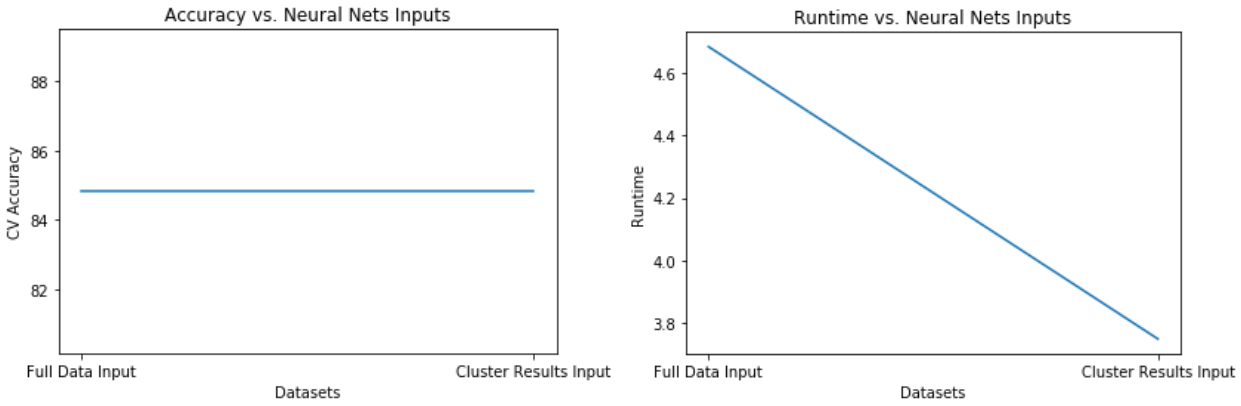
```
{'batch_size': [256], 'epochs': [20], 'optimizer': ['sgd'], 'first_hidden_units': [10], 'second_hidden_units': [10], 'hidden_ac
tivation_fun': ['sigmoid'], 'output_activation_func': ['sigmoid']}
```



We don't see variation in Cross-Validation error but ICA is efficient and takes very less time to run. Without reducing any features, the computation time is inefficient and takes more time when compared to others.

Task 5 - Neural Networks using Clustering Results (Dataset 2 - Online Shopping Intention Prediction)

Using the best parameters found in the neural networks experiments from the previous assignments, the CV error and Runtime are found with original features and cluster results obtained from Task 1 as the inputs for neural networks with the best parameters found from the previous assignment.



We don't see any variation in Cross-Validation error but we can clearly find that the neural network with cluster results as inputs is more efficient taking less runtime than the original features. Since, we don't see any kind of variation in the accuracy we may use cluster results and save computation cost and time.