BUAN 6341 APPLIED MACHINE LEARNING ASSIGNMENT 4

Project Report

K-Means, Expectation Maximization, feature selection, PCA, ICA, Randomized Projections

By Animesh Kansal (axk169531)

Datasets Used:

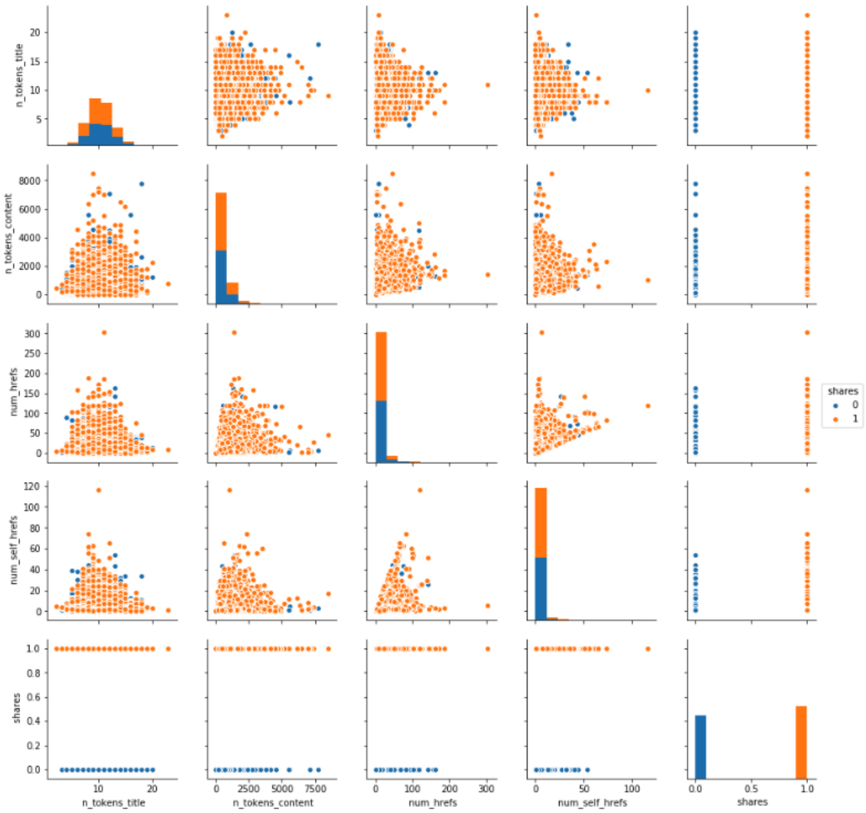
1. Online news popularity data set (39644, 61)
2. Telecom Churn data set (7043, 21)

Scripting Language: Python as scripting language

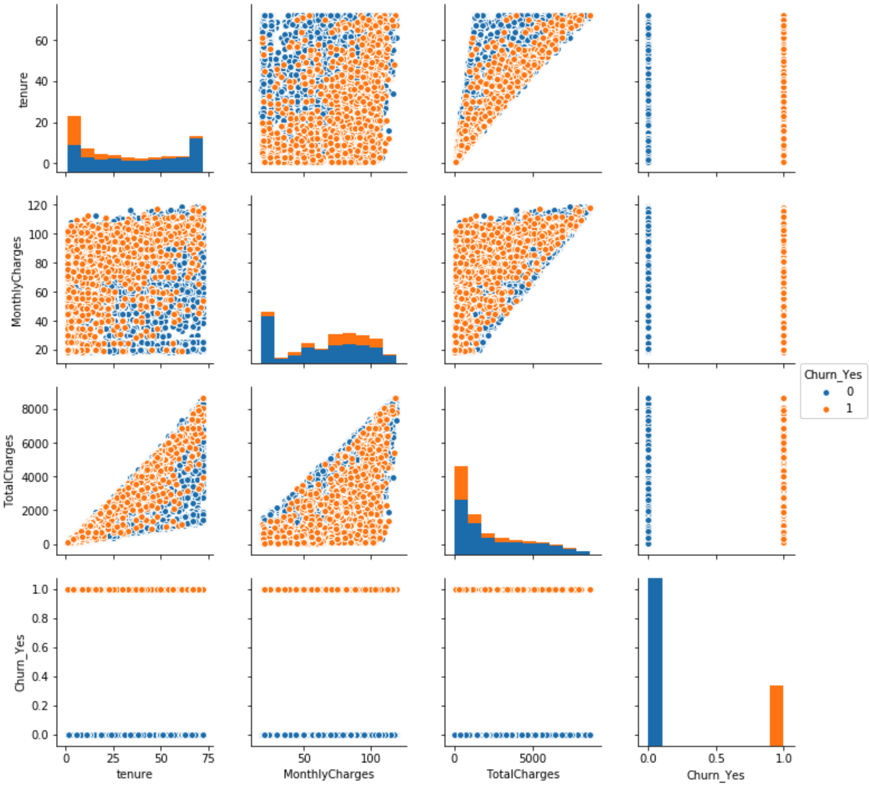
Tasks:

1. Run the clustering algorithms on your datasets and describe your observations (with plots).

For dataset 1, According to initial pair plots, I am unable to see any clusters. I have drawn randomly 5-6 features and drawn pair plots. But not able to identify any cluster.



Simmilarly, for dataset 2, According to Initial pair plots , I am not able to see any clusters. I have drawn the numeric features and not able to identify any cluster



```
from sklearn.cluster import KMeans
from sklearn.mixture import GaussianMixture
```

Then, I Created K- Means object with n_clusters = 2, Because we have divided "number of shares" into small and large based on the median value

Every time, if the accuracy is less than 50%, I will predict the opposite results, i.e. 1 to 0 and vice versa, Because prediction cannot go below 50% for a binary output

K-Means Accuracy:

- Churn Analysis = 55.02%
- Online News Share = 52.87%

Expectation Maximization Accuracy:

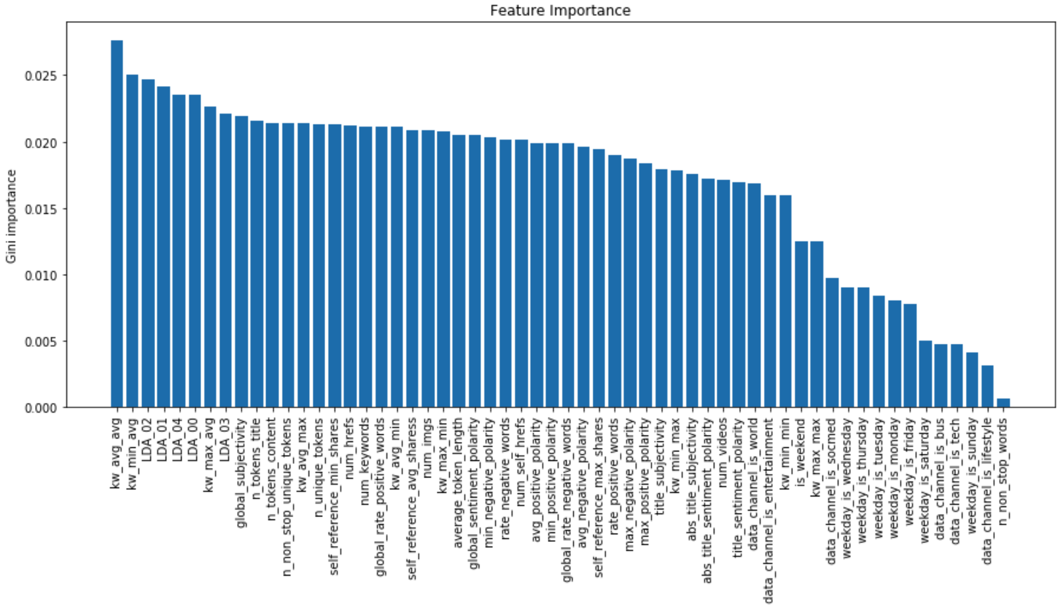
- Churn Analysis = 55.02%
- Online News Share = 57.64 %

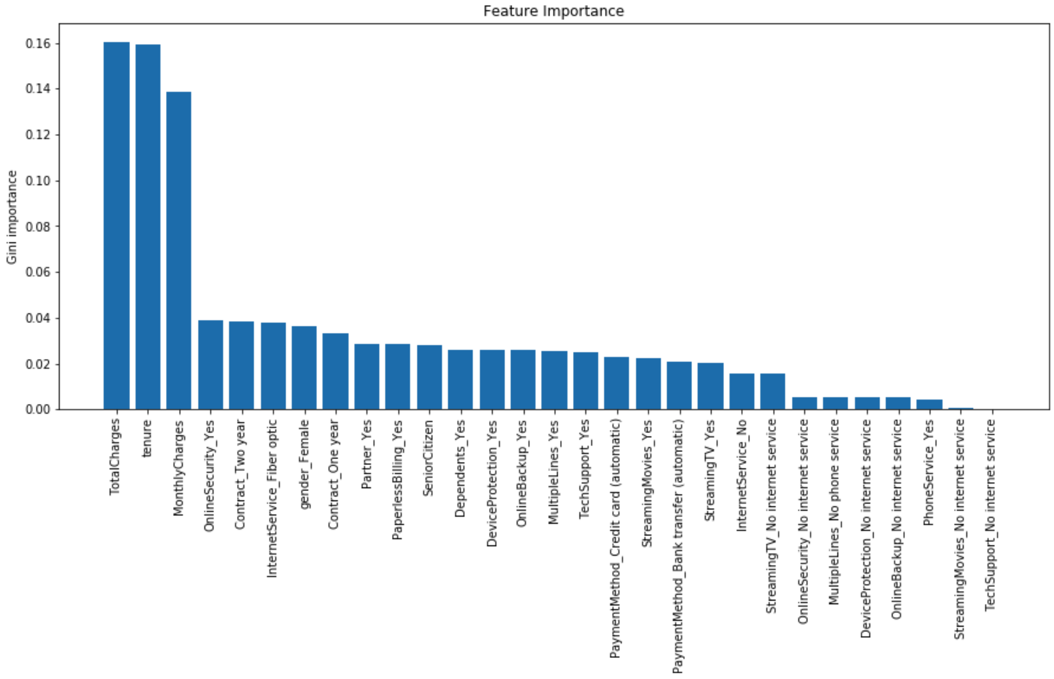
2. Apply the dimensionality reduction algorithms on your datasets and describe your observations
3. Run the clustering algorithms again, this time after applying dimensionality reduction. Describe the difference compared to previous experimentation (with plots).

Feature selection:

```
from sklearn.ensemble import ExtraTreesClassifier
```

I am using Decision Tree to get the top 20 features





<u>After Feature selection, Selecting top 10 and 20 Attributes from corresponding Data sets</u>

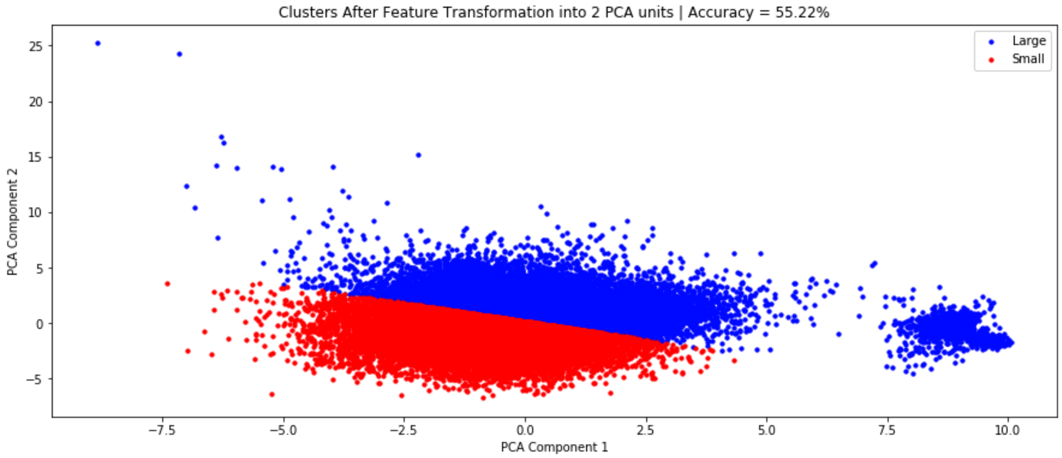
<u>K-Means Accuracy:</u>

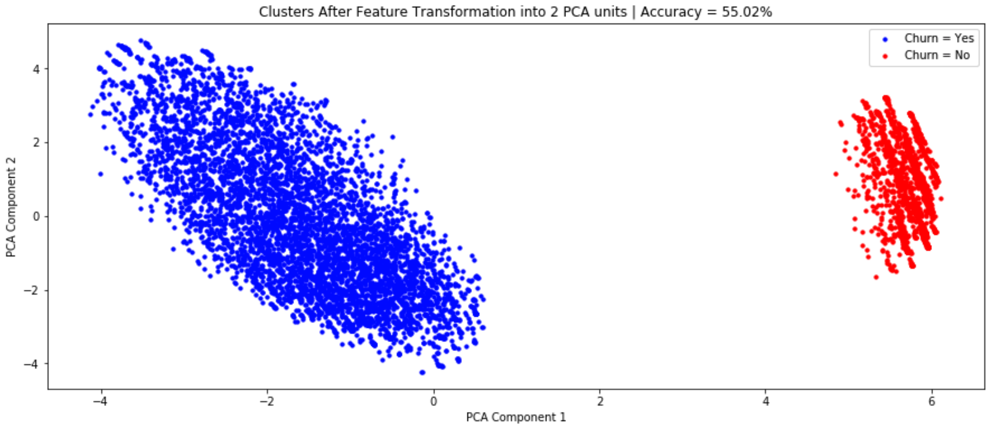
- Churn Analysis = 52.83%
- Online News Share = Accuracy = 50.86 %

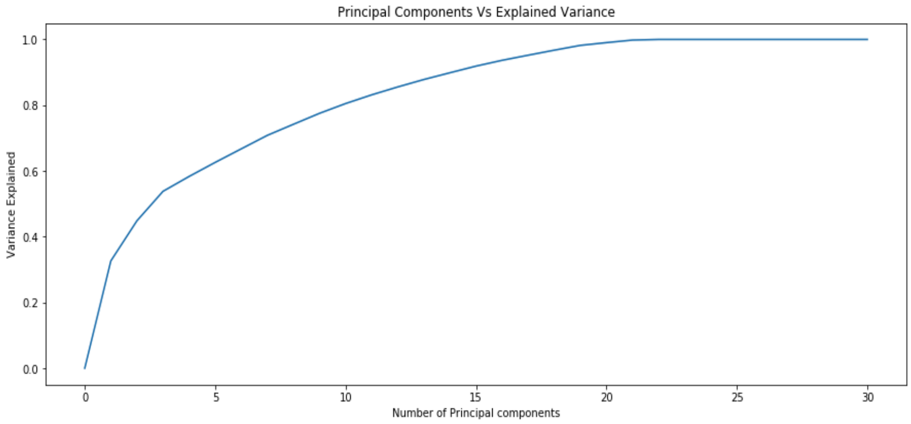
Feature Transformation:

1. PCA

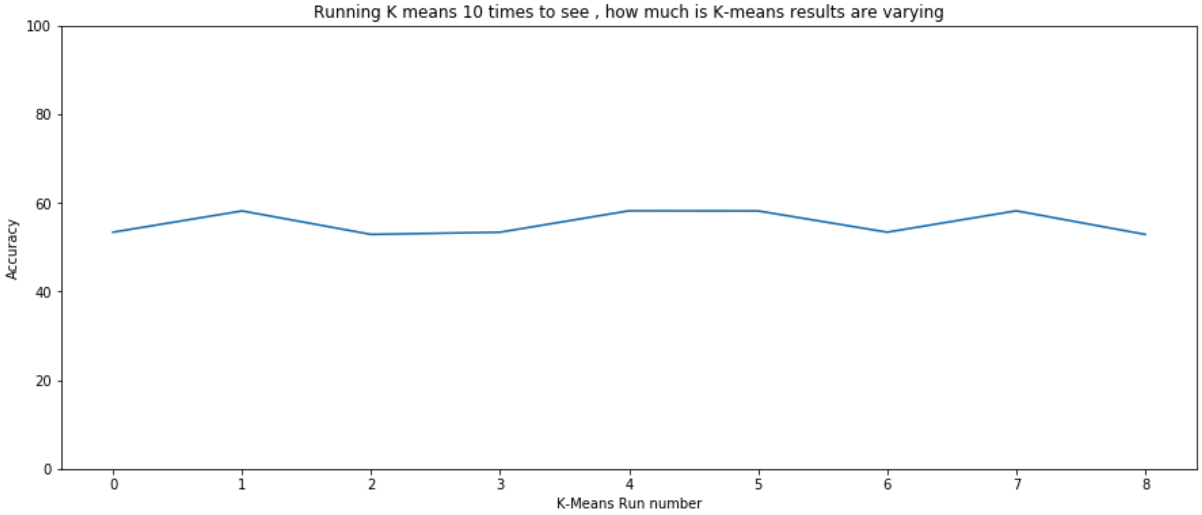
For Visualizing Clusters, taking PCA = 2, Clearly shows, no overlapping in Clusters, Clusters are Exclusive



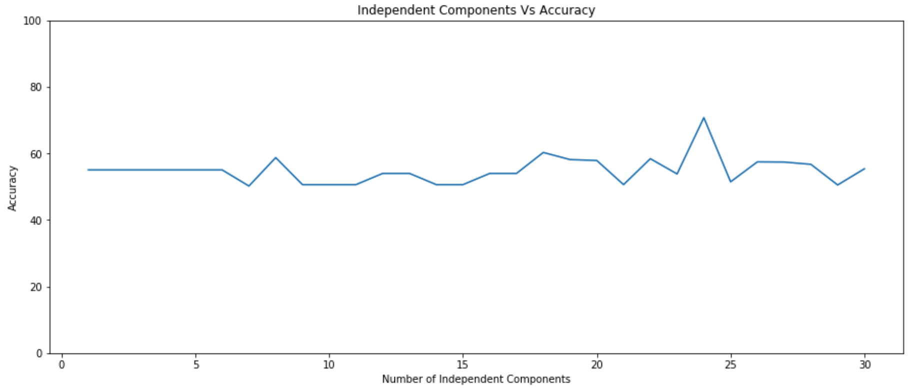




Almost similar curve for both data sets, So I selected 10 and 20 features in both so that variance explained ~ 70%



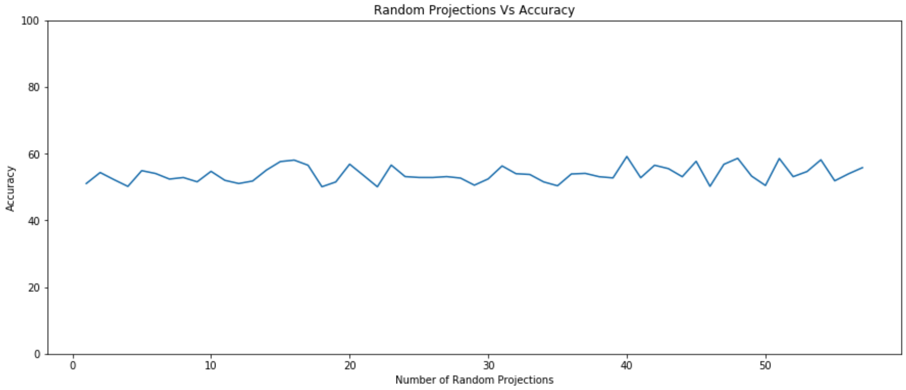
2.) ICA



1) As observed, ICA didn't result in different K-Means accuracies for given number of components

2)Experimenting with Number of ICA components Vs Accuracy

3) RP



1) As observed, RP also didn't result in different K-Means accuracies for given number of components

2)Experimenting with Number of RA components Vs Accuracy

4. Run your neural network learner from assignment 3 on the data after dimensionality reduction (from task 2).

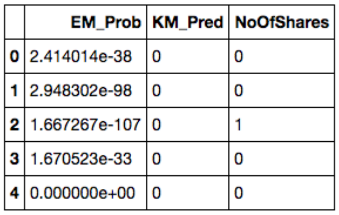
I used Keras on top of TensorFlow and created a 2 layer ANN (1 input layer, 2 hidden, 1 output layer)

ANN Accuracy After Dimensionality Reduction Using PCA:

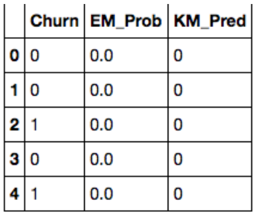
- Churn Analysis = 79.55%
- Online News Share = 61.1%

5. Use the clustering results from task 1 as the new features and apply neural network learner on this new data consisting of only clustering results as features and class label as the output.

I used Cluster labels from k-means and probabilities from EM and Run into ANN
New data looks like: No of Shares (0: Small , 1 : large)



	EM_Prob	KM_Pred	NoOfShares
0	2.414014e-38	0	0
1	2.948302e-98	0	0
2	1.667267e-107	0	1
3	1.670523e-33	0	0
4	0.000000e+00	0	0



	Churn	EM_Prob	KM_Pred
0	0	0.0	0
1	0	0.0	0
2	1	0.0	0
3	0	0.0	0
4	1	0.0	0

ANN Accuracy After this:

- Churn Analysis = 73.42%
- Online News Share = 53.3%

In real life, we are not able to evaluate a K – Means Cluster, but we are evaluating using class Labels.

Algorithm	K- Means	EM Gaussian Mixture	K-Means Feature selection	K-Means + PCA	K-Means +ICA	K-Means +RP	NN	NN+ PCA	EM+ KM feed to NN
Telecom Churn Analysis Dataset	52.87 %	57.64 %	50.86 %	58.17 %	53%	59.14%	69.31%	61.61%	53.36%
Online News Sharing Dataset	55.02 %	55.02 %	52.83 %	55.02 %	70.27%	67.62%	79.55%	79.72%	73.42%

Questions:

How did you choose k?

A. Small or Large shares, Churn or no churn: Hence 2 Clusters

What type of clusters did you get?

A. Exclusive clusters, no overlap in clusters, as shown through PCA

ICA and RP, Compare and contrast the different algorithms.

A. ICA performed best in Online News Sharing Dataset, while RP Performed best in Churn Analysis Dataset.

When you re-ran your neural network algorithms were there any differences in performance? Speed? Anything at all?

A. With Less dimensions, I almost halved the number of inputs. also, I observed that Run reduced to half of normal run will all the attributes