Dataset 1

Energy consumption data set

In Part-1 I have applied various classification algorithms on the Energy consumption data set and classified the houses as “high energy” consumers or “low energy” consumers.

I have dropped the lights and the date column from the dataset as they are not useful for us when doing the classification. I have split the dataset into training set and testing set with a ratio of 30% testing set and 70% training set. And standard student scaling is done on all the feature variables. Student scaled variable has a mean of 0 and standard distribution of 1.

Dataset-2

In Part-2, I have tried to predict if a customer is going to quit the services of a Telecom operator based on various features.
I selected this dataset as Churn is a one of the biggest problems in the telecom industry. Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is
1.9% - 2%. If we are able to predict the churn of a customer this will be worth a lot of money as Telecom company can try to save the customer who may quit their services by offering him better plans and offers.

> To better understand this problem we can define Churn as Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.
>
> Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.
>
> Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which companies control, such as how billing interactions are handled or how after-sales help is provided.

predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at

focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

Summary statistics of the dataset used:

• Dataset consists of 7043 observations on 21 variables.

• The dependent variable for the Classification model is Churn variable which is Binary column with values "Yes" if the customer quit the services and "No" if the customer is still with the telecom operator.

We then proceed to convert the predictor variable in a binary numeric variable with values 1 for "Yes" and 0 for "No". And as Machine Algorithms only understand numbers we convert all the categorical variables into dummy variables.
We now have 46 columns.

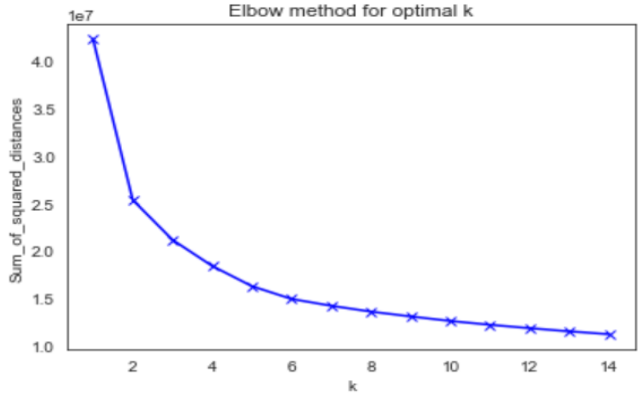
I have split the dataset into training set and testing set with a ratio of 30% testing set and 70% training set. And standard student scaling is done on all the feature variables. Student scaled variable has a mean of 0 and standard distribution of 1.

K MEANS

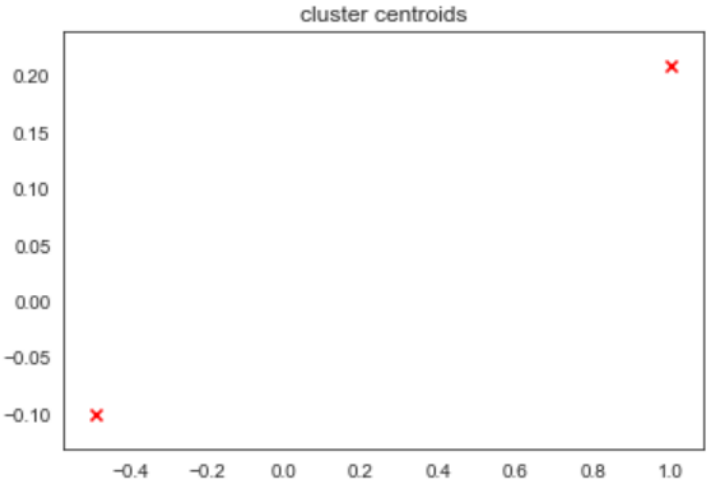
Clustering is the task of grouping together a set of objects in a way that objects in the same cluster are more similar to each other than to objects in other clusters. Similarity is a metric that reflects the strength of relationship between two data objects.

Energy Appliances Dataset

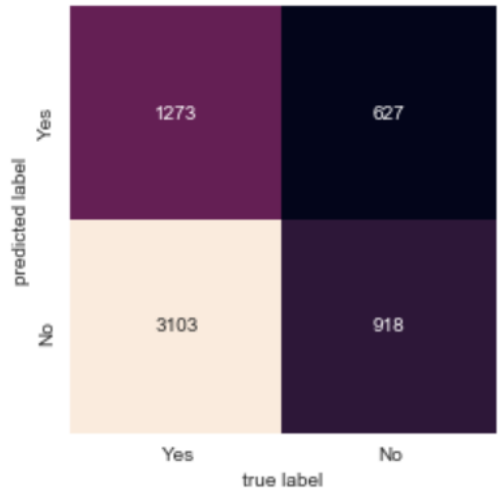
The dataset's label is 'consumption level' which denotes the level of energy consumption of a customer. Our task is to cluster the records into two: the customers who have consumed low energy and consumers who have consumed high energy. Since k means is an unsupervised learning algorithm, it doesn't require a labelled dataset. Hence, we would drop the 'consumption level' column to make the dataset unlabeled. It's the task of k means to cluster the records of the datasets into the two classes.
We use the elbow method to determine the optimal number of clusters. In this method, we run the k-means clustering on the dataset for a range of values of k, and calculate the SSE (sum of squared errors) for each value of k. Our goal is to choose a small value of *k* that has a low SSE and marks the 'elbow' of an arm.



As from the figure above, the optimum value of k is 2. Now, we implement our k means algorithm to cluster the records into two: High consumption and Low Consumption.
After implementing k means on our dataset, we find out the cluster centers and plot it. Since our dataset contains more than 2 features, it is impossible to visualize this dataset on the computer, hence I have only plotted the cluster centers, i.e cluster centroids. In reality, all the data points would be clustered around these 2 clusters.



Now, we see how model is doing by looking at the percentage of records that were clustered correctly by comparing it with the class labels in our dataset. We see that almost **37%** records have been correctly clustered. We can call this the accuracy of the mode.

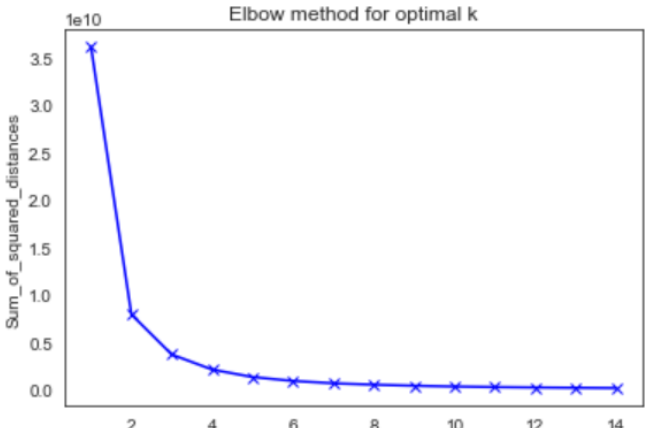


Here is a confusion matrix for the result we see that, there are a lot of wrong clustered records.

Telecom Churn Dataset

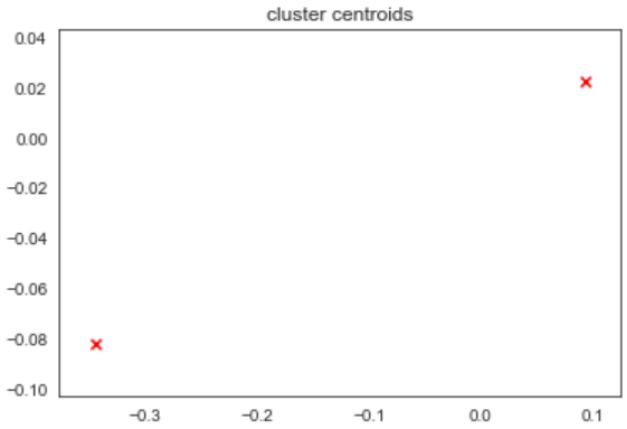
The dataset's label is 'Churn' which denotes the churn status of a telecom customer. Our task is to cluster the records into two: the customers who have churned and the ones who haven't. Since k means is an unsupervised learning algorithm, it doesn't require a labelled dataset. Hence, we would drop the 'churn' column to make the dataset unlabeled. It's the task of k means to cluster the records of the datasets if they churned or not.

The Elbow graph for selecting the appropriate number of clusters.

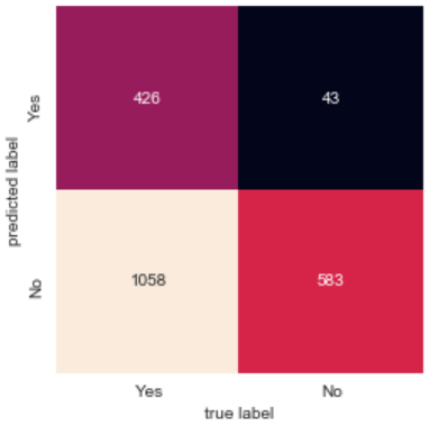


As from the figure above, the optimum value of k is 2. Now, we implement our k means algorithm to cluster the records into two: Churned or not churned.

After implementing k means on our dataset, we find out the cluster centers and plot it. Since our dataset contains more than 2 features, it is impossible to visualize this dataset on the computer, hence I have only plotted the cluster centers, i.e cluster centroids. In reality, all the data points would be clustered around these 2 clusters.



Now, we see how model is doing by looking at the percentage of customer records that were clustered correctly by comparing it with the class labels in our dataset. We see that almost **47.81%** records have been correctly clustered. We can call this the accuracy of the model.



Here is the Confusion matrix associated with the above K-means procedure and comparing with the original class label.

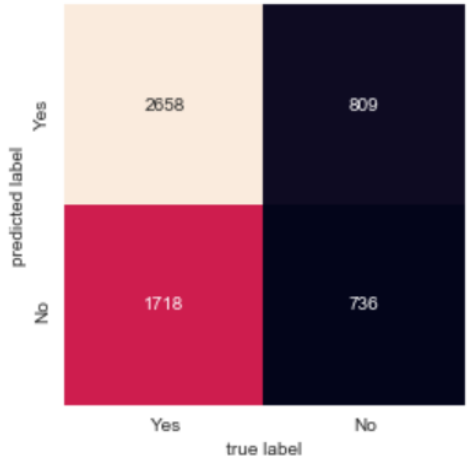
EXPECTATION MAXIMIZATION

Expectation Maximization comes under Gaussian models which are more of a way of thinking and modeling rather than a particular algorithm. Clusters are modeled as Gaussian distributions and not by their means. There is a correspondence between all data points and all clusters rather than correspondence between each data point to its own cluster, as is the case in K-means clustering.

Expectation Maximization works the same way as K-means except that the data is assigned to each cluster with the weights being soft probabilities instead of distances. The advantage is that the model becomes generative as we define the probability distribution for each model.

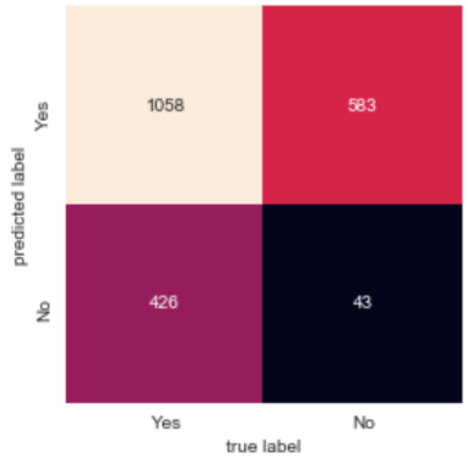
Energy Appliances Dataset

The accuracy of the expectation maximization algorithm on the dataset is **57.32%**. The EM algorithm can incorporate underlying assumptions about how the data was generated.



Telecom Churn Dataset

The accuracy of the expectation maximization algorithm on the dataset is **52.18%**.

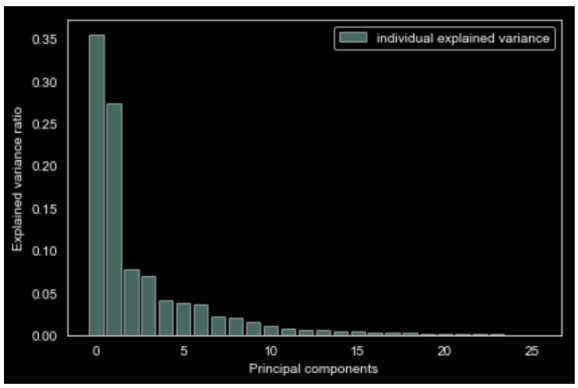


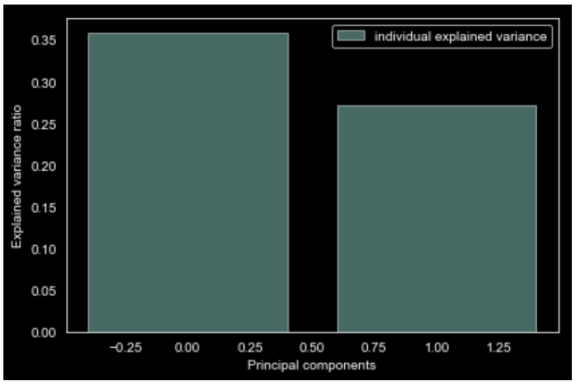
PRINCIPAL COMPONENT ANALYSIS

PCA is a dimensionality reduction technique used to transform high-dimensional datasets into a dataset with fewer variables, where the set of resulting variables explains the maximum variance within the dataset. PCA is used prior to unsupervised and supervised machine learning steps to reduce the number of features used in the analysis, thereby reducing the likelihood of error.

Energy Appliances DATASET

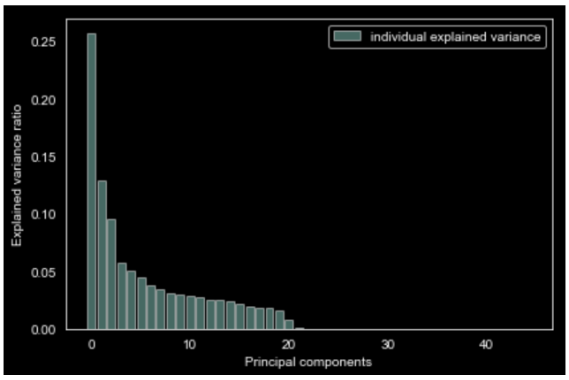
After performing PCA on the dataset, I retrieved the explained variance ratios to get a better idea of how principal components describe the variance in the data.

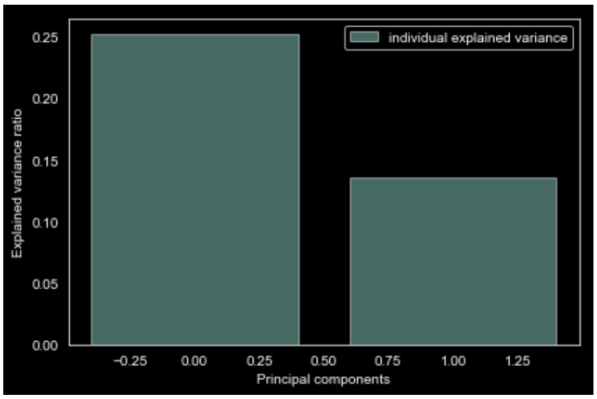




It can be seen that 2 principal components **PC1 and PC2 describe approximately 63%** of the variance in the data. The first two components explain the maximum variance as compared to the other components.

Telecom Churn Dataset





It can be seen that 3 principal components describe approximately 48% of the variance in the data. The first two components explain the maximum variance as compared to the other components.

PC1 and PC2 explain 38% of the variance in the dataset.

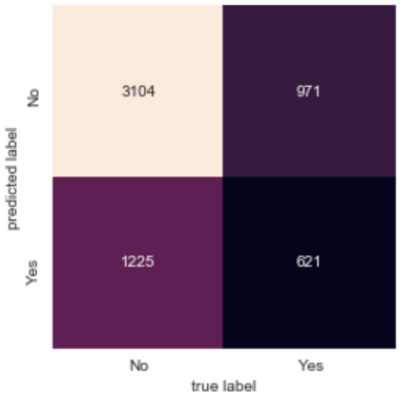
Performing K-Means on the dataset after feature transformation using PCA

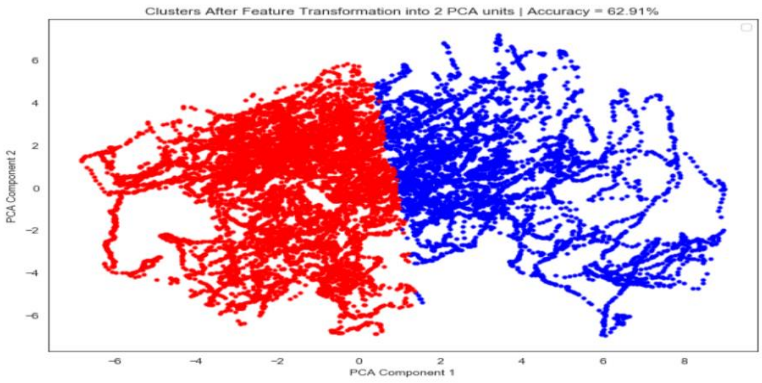
Energy Appliances DATASET

After transforming the features into 2 dimensions, I performed k means on the dataset and obtained the following results:

Accuracy of the model is 62.91%

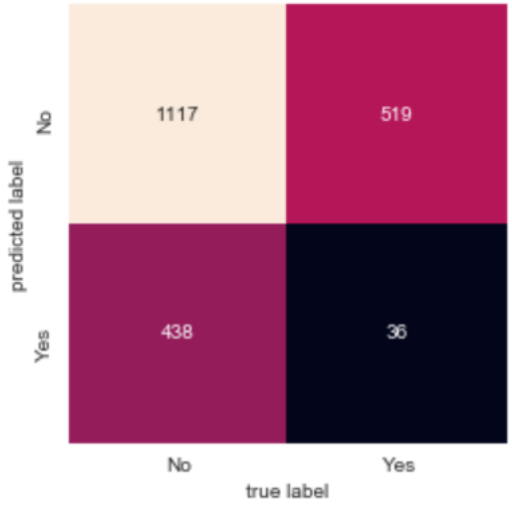
Confusion Matrix:





The plot above shows the distribution of the data points in 2 clusters after transforming the features in 2D space. It can be concluded that after performing PCA, the accuracy of the k means clustering was improved significantly indicating that feature transformation plays an important role in unsupervised learning.

Telecom Churn Dataset

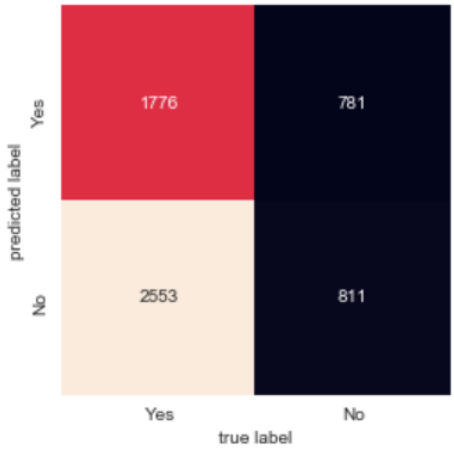


Accuracy of the model is 54.64%

The plot above shows the distribution of the data points in 2 clusters after transforming the features in 2D space. It can be concluded that after performing PCA, the accuracy of the k means clustering was improved significantly indicating that feature transformation plays an important role in unsupervised learning.

Performing EM on the dataset after feature transformation using PCA

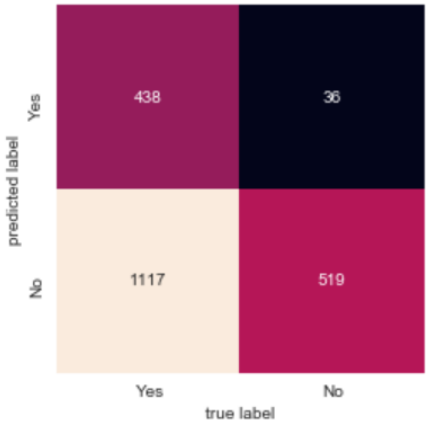
Energy Appliances DATASET



After transforming the features into 2 dimensions, I performed EM on the dataset and obtained the following results:
Accuracy of the model is **43.69%**
The accuracy has dropped after using feature transformation for Expected Maximization Algorithm. The reason for this could be that the 2 principal components that we selected according to the explained variance ratio, contribute only for about 63% of the variation in the dataset.

Telecom Churn Dataset

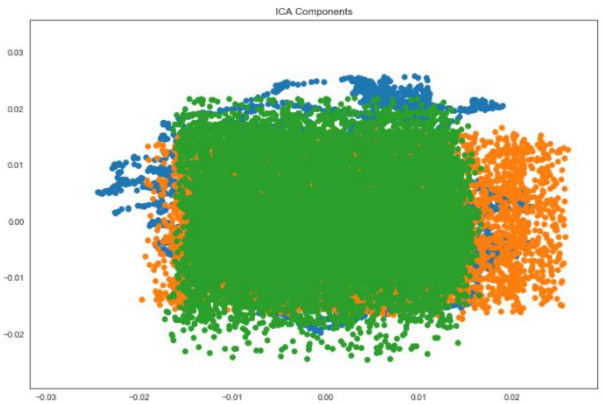
Accuracy of the model is **45.36%**.
The accuracy has dropped after using feature transformation for Expected Maximization Algorithm. The reason for this could be that the 2 principal components that we selected according to the explained variance ratio, contribute only for about **48%** of the variation in the dataset.

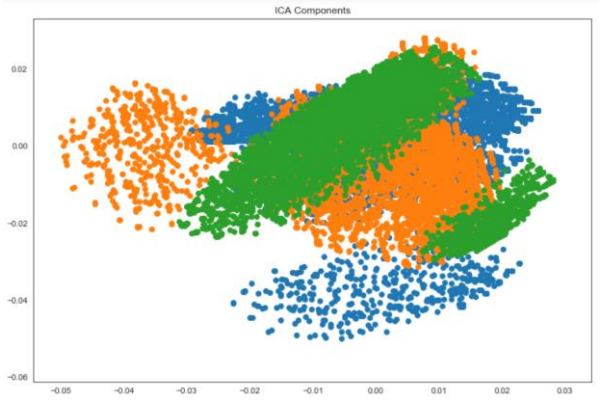


INDEPENDENT COMPONENT ANALYSIS

Independent Component Analysis (ICA) is based on information-theory and is also one of the most widely used dimensionality reduction techniques. The major difference between PCA and ICA is that PCA looks for uncorrelated factors while ICA looks for independent factors.

Here, n_components will decide the number of components in the transformed data. We have transformed the data into 3 components using ICA.





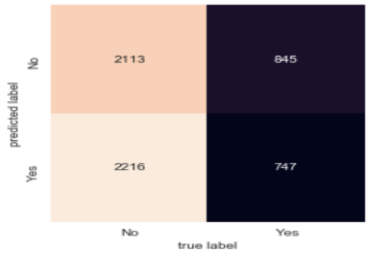
Energy Appliances Dataset **Telecom Churn Dataset**

The data has been separated into different independent components which can be seen very clearly in the above image. X-axis and Y-axis represent the value of decomposed independent components.

Performing K-Means on the dataset after feature transformation using ICA

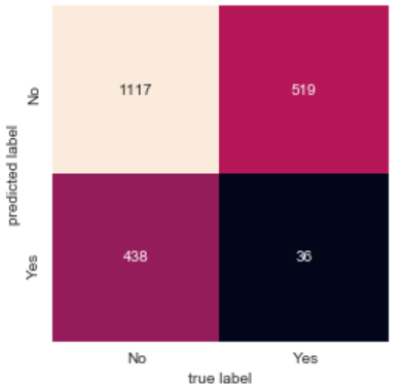
Energy Appliances Dataset

The accuracy of the model after feature transformation using ICA is **48.3%.** One reason for the reduction in the accuracy can be that, since ICA looks for independent factors and does not take into account correlation, it is possible that highly correlated features might have been incorporated in the model.



Telecom Churn Dataset

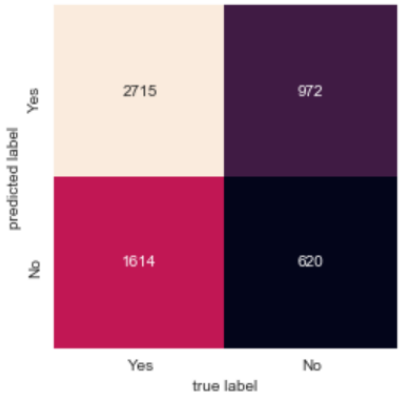
The accuracy of the model after feature transformation using ICA is **54.64%.**



Performing EM on the dataset after feature transformation using ICA

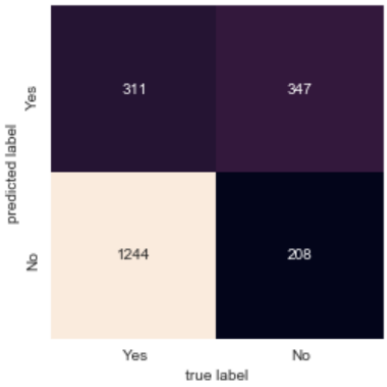
Energy Appliances Dataset

The accuracy of the model after transforming features using ICA and performing EM is **56.32%.** This model is performing better as compared to the K means model after using ICA.



Telecom Churn Dataset

The accuracy of the model after transforming features using ICA and performing EM is **24.6%.**



RANDOMIZED PROJECTIONS

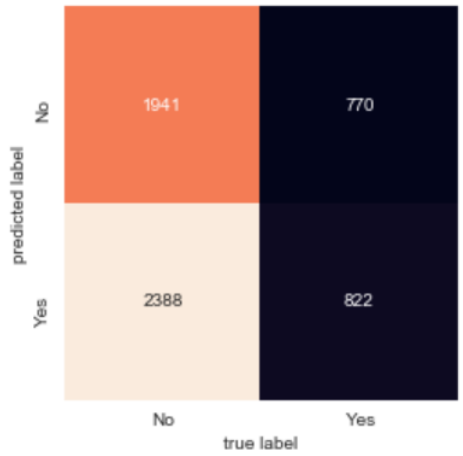
Random projection is a technique used to reduce the dimensionality of a set of points which lie in Euclidean space. The sklearn.random_projection.GaussianRandomProjection reduces the dimensionality by projecting the original

input space on a randomly generated matrix where components are drawn from the following distribution N(0,1/ncomponents).

Performing K-Means on the dataset after feature reduction using Randomized Projections

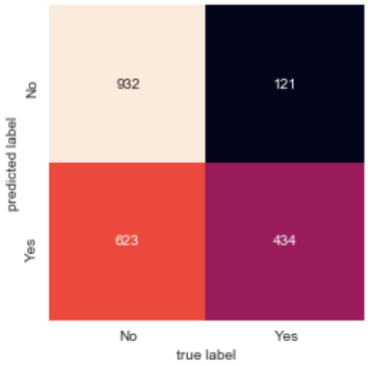
Energy Appliances Dataset

The accuracy of the model is **46.66%.** This accuracy is more than the accuracy of the model which used all the features.



Telecom Churn Dataset

The accuracy of the model is **64.74%**

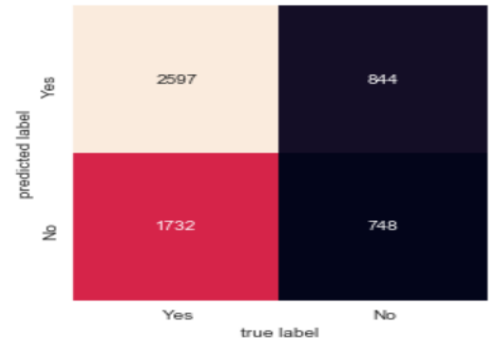


Performing EM on the dataset after feature reduction using Randomized Projections

Energy Appliances Dataset

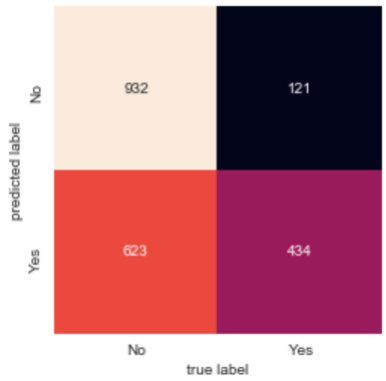
The accuracy of the model is **56.49%.** This is more than the accuracy of the model which used all the features. We can conclude that feature reduction using randomized projections has improved the accuracy of the model.

Confusion Matrix



Telecom Churn Dataset

The accuracy of the model is **64.74%.**



STEP FORWARD FEATURE SELECTION

Step forward feature selection starts with the evaluation of each individual feature and selects that which results in the best performing selected algorithm model. Next, all possible combinations of the that selected feature and a subsequent feature are evaluated, and a second feature is selected, and so on, until the required predefined number of features is selected.

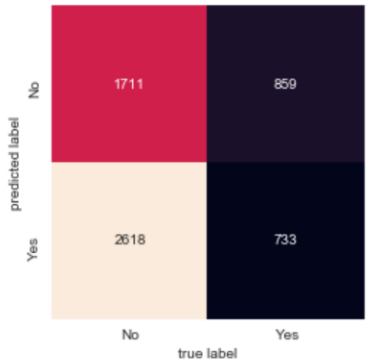
After performing feature selection on the original dataset, I obtained the best 5 features: Phone Service, Internet Service, Online Security, Contract, Payment Method.

Performing K-Means on the dataset after feature reduction using Step Forward Feature Selection

Energy Appliances Dataset

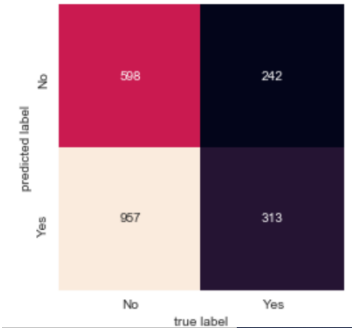
The accuracy of the model is **41.28**%. This accuracy is least as compared to all other models indicating that we might have not selected the best features.

Confusion Matrix



Telecom Churn Dataset

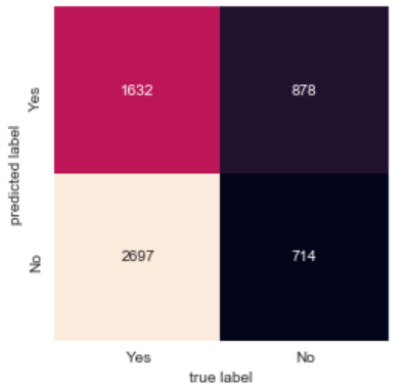
The accuracy of the model is **43.18%.**



Performing EM on the dataset after feature reduction using Step Forward Feature Selection

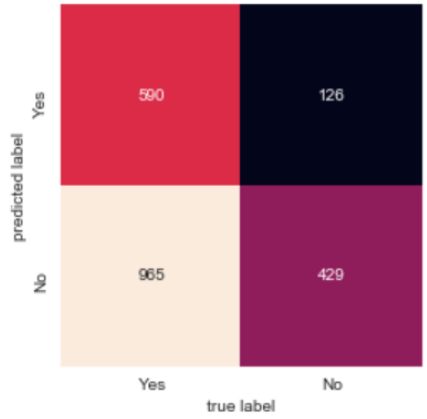
Energy Appliances Dataset

The accuracy of the model is **41.28%.** This accuracy is least as compared to all other models indicating that we might have not selected the best features.



Telecom Churn Dataset

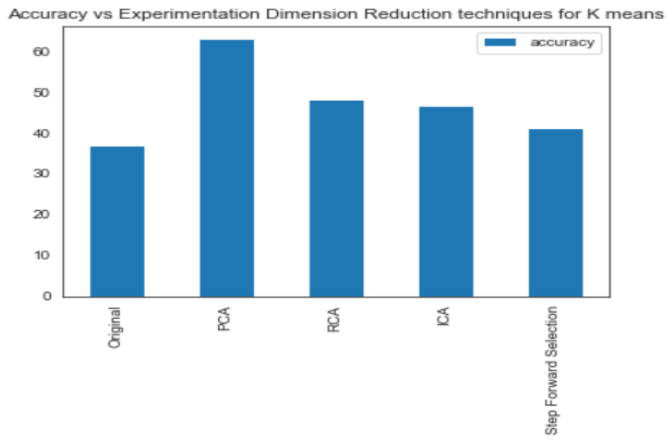
The accuracy of the model is **43.18%.**



Even though the accuracy of this and K-means using Step forward feature selection is same. We can see that the misclassifications are different.

Comparison of the different dimension reduction/transformation techniques on the 2 clustering algorithms

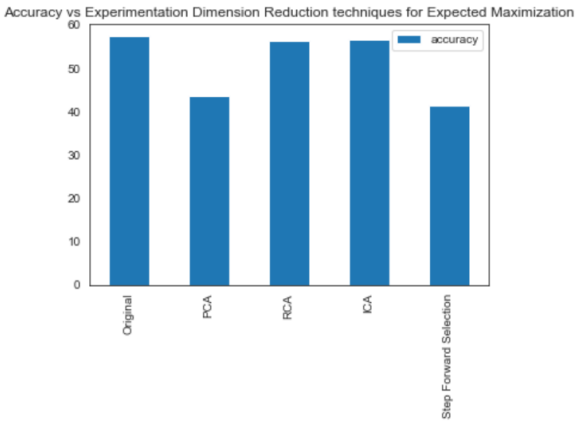
Energy Appliances Dataset



For K means clustering, the best accuracy is given by the model with **Principal component analysis** performed on it.

	accuracy
Original	37.003884
PCA	62.910000
RCA	48.300000
ICA	46.660000
Step Forward Selection	41.280000

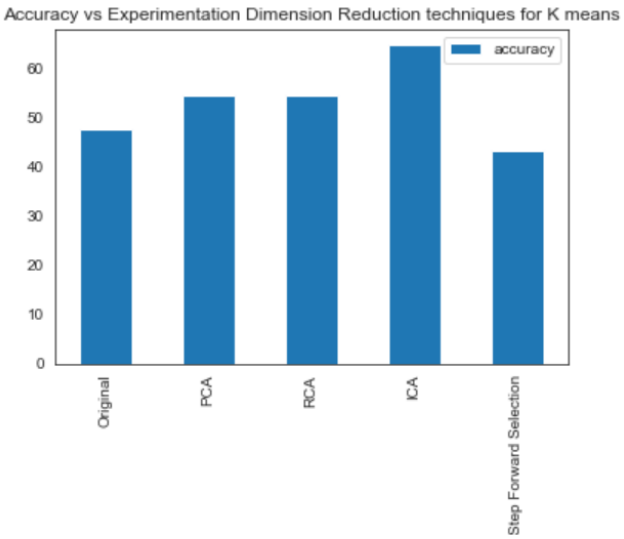
For Expectation Maximization, the best accuracy is given by the model which uses all the features but the model with RCA and ICA come very close. In a real world setting this can be considered as a trade off and the model using one of the dimensionality reduction technique may be used as it will be faster and more efficient.



	accuracy
Original	57.32
PCA	43.69
RCA	56.32
ICA	56.49
Step Forward Selection	41.28

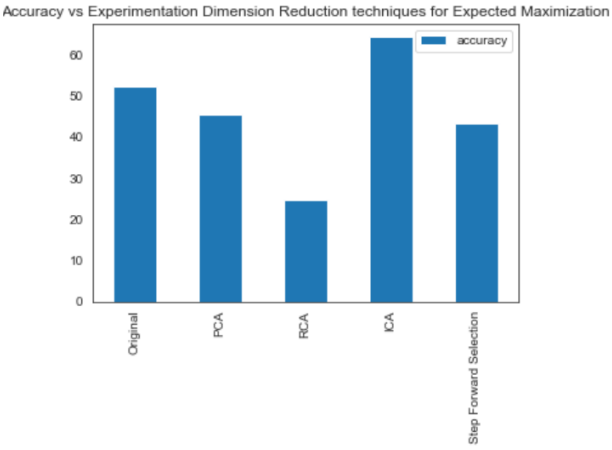
Telecom Churn Dataset

For K means clustering, the best accuracy is given by the model with Independent component analysis performed on it.



	accuracy
Original	47.819905
PCA	54.640000
RCA	54.640000
ICA	64.740000
Step Forward Selection	43.180000

For **Expectation Maximization**, the best accuracy is given by the model with Independent component analysis used as dimension reduction technique

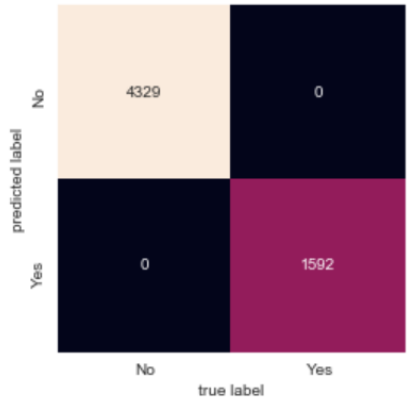


	accuracy
Original	52.18
PCA	45.36
RCA	24.60
ICA	64.27
Step Forward Selection	43.18

NEURAL NETWORK IMPLENTATION

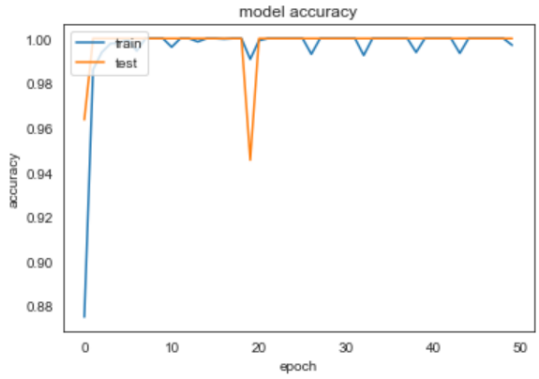
I implemented the neural network classifier from the previous assignment to the dataset which has been transformed using PCA.

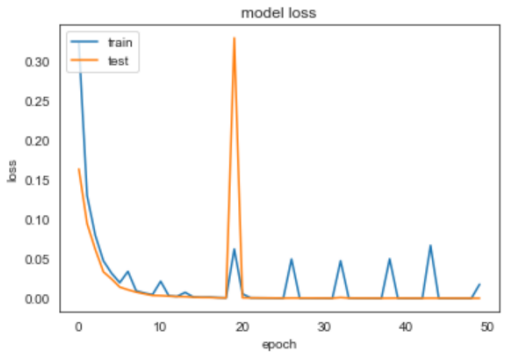
Energy Appliances Dataset



The accuracy I am getting is 100% which seems to be wrong and the neural network seems to have overfitted the data

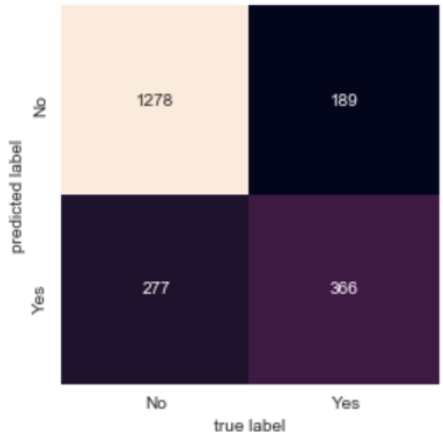
From the plots below we see that model has overfitted the data.

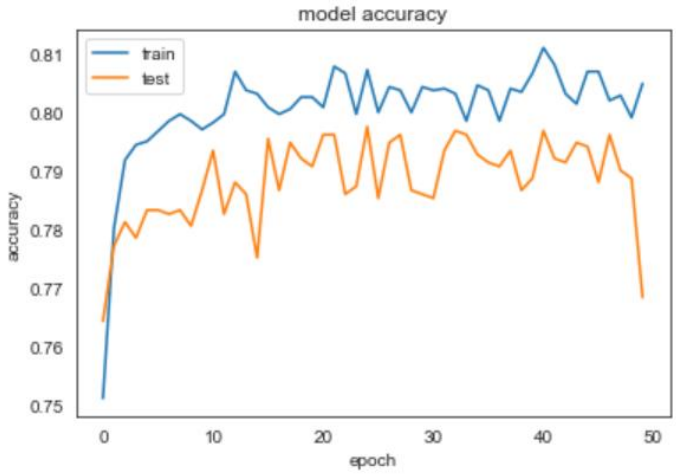


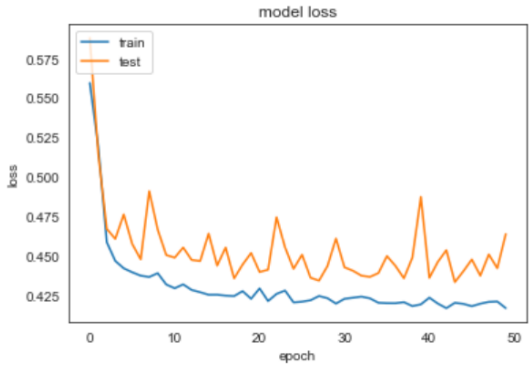


Telecom Churn Dataset

The accuracy of the model is **77.91%** which is much lower than the original one This indicates that the neural network is not classifying the dataset as well with the feature reduction techniques.







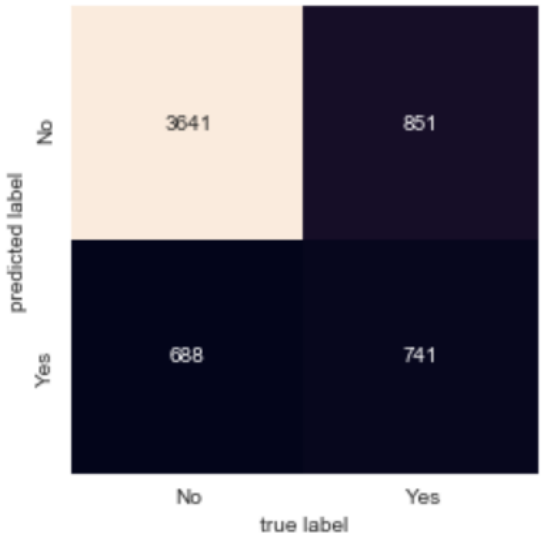
From the plot of accuracy and plot of loss, we see that the model could be trained a little more, and that the model has comparable performance on both train and test datasets. If these parallel plots start to depart consistently, it might be a sign to stop training at an earlier epoch.

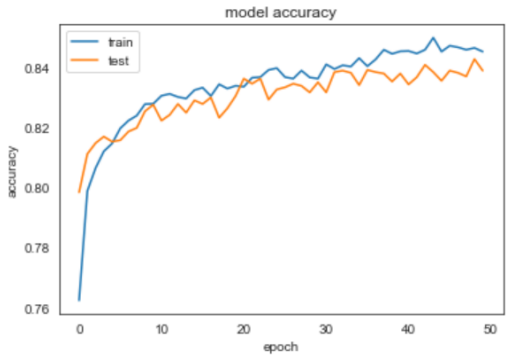
NEURAL NETWORK IMPLEMENTATION USING CLUSTERS AS FEATURES

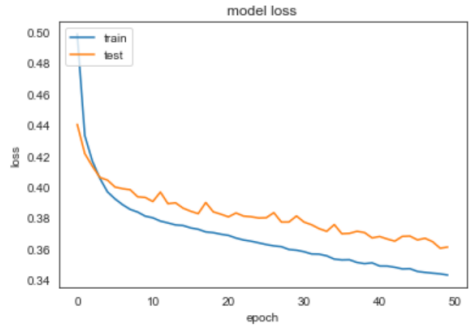
In this task, I have added a column of cluster labels to the dataset as a feature. I will be using the class labels to perform Neural Network Classification.

Energy Appliances Dataset

The accuracy of the model is 74.01% which is almost equivalent to the accuracy of the original model.
Confusion Matrix –

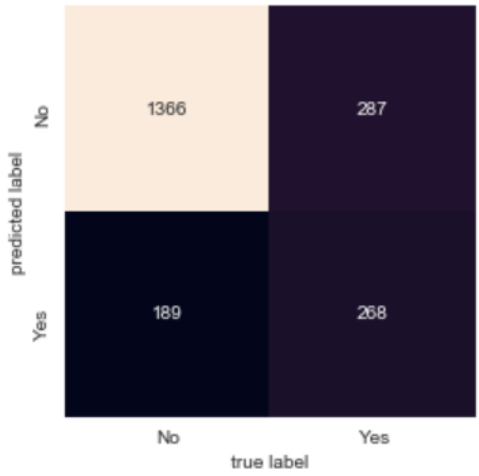


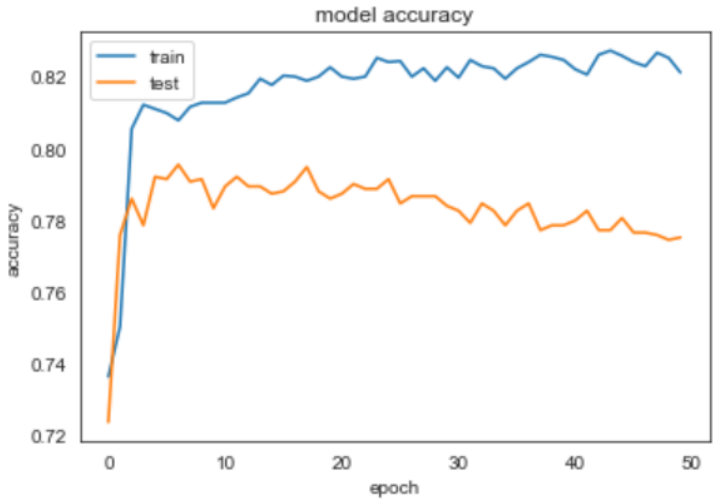


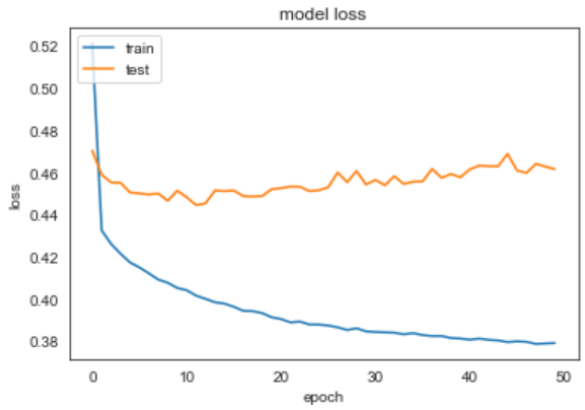


Telecom Churn Dataset

The accuracy of the model is **77.44%**







which is much lower than the original one (This indicates that the neural network is not classifying the dataset as well with the feature reduction techniques.