CS7641 Assignment 3: Unsupervised Learning

Jack Chai Jie Feng
jchai41@gatech.edu

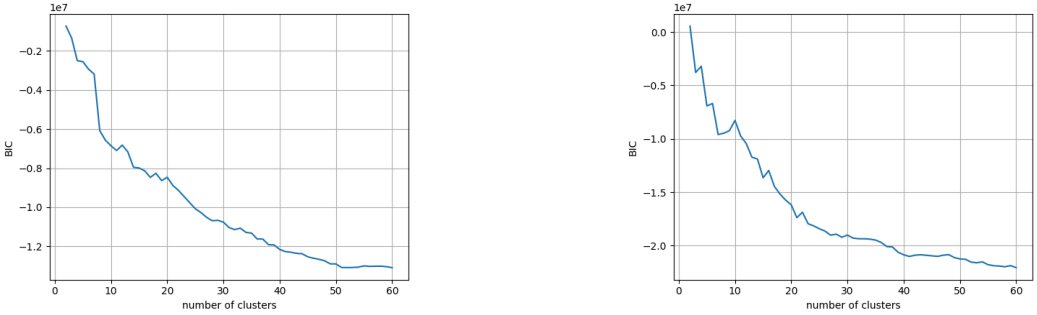
DATASET

The models used in this report are scale variant; all datasets will have to be transformed by a standard scaler to achieve zero mean and unit variance. The 2 datasets used in this report are the same one as in assignment 1, namely income dataset and bank churn dataset. Both datasets have an imbalance target label class and f1 scores are used as evaluation where necessary. The income dataset has a lot of categorical features which we will discuss how it affects the experiment in a later section.

CLUSTERING

The clustering algorithm chosen for this report is Gaussian Mixture Model (GMM) and K-Means clustering (KMeans). These 2 algorithms are chosen due to the difference in the cluster shape and hard/soft clustering; Gaussian Mixture Model is a soft clustering algorithm and does not assume any shape and KMeans on the other hand, is a hard clustering algorithm that only allows spherical cluster shape.

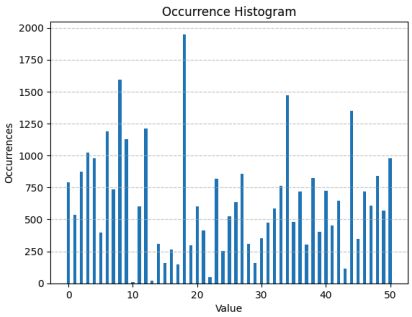
GMM



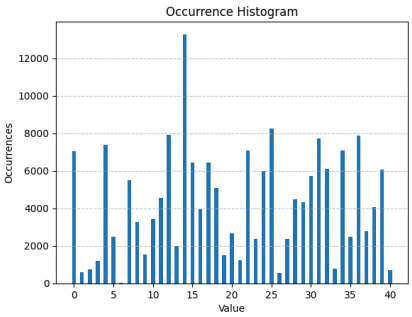
Income Dataset Bank Churn Dataset

Figure 1.1.a: GMM BIC curve for both datasets

The Bayesian Information Criterion (BIC) is commonly used for model selection in Gaussian Mixture Models (GMMs). It balances model complexity with goodness of fit to the data and helps in determining the optimal number of components (clusters) for the GMM. There is a clearer elbow in the bank churn dataset compared to the income dataset (figure 1.1.a), possibly due to the high number of categorical features in the latter. We decided to use 51 clusters for the income dataset and 41 clusters for the bank churn dataset.



Income Dataset

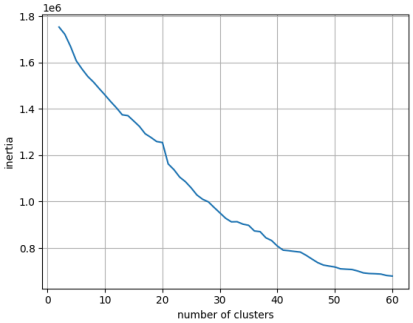


Bank Churn Dataset

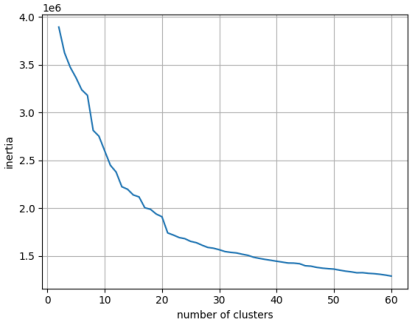
Figure 1.1.b: GMM data points distribution for both datasets

After selecting the number of clusters for each dataset, we can analyze the distribution of data in each cluster in figure 1.1.b. In both datasets, there is an uneven distribution of data points in each cluster. Clusters with sparse data points might be less informative and may be challenging for us to interpret its characteristics.

KMeans

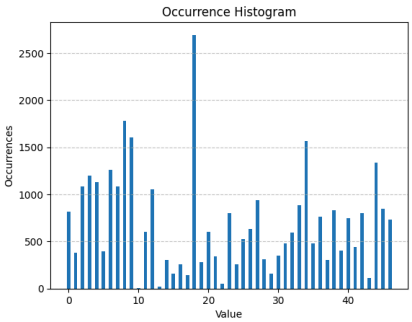


Income Dataset

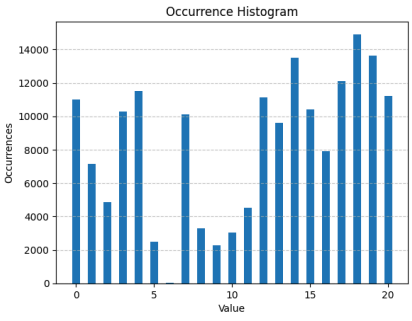


Bank Churn Dataset

Figure 1.2.a: KMeans inertia curve for both datasets



Income Dataset



Bank Churn Dataset

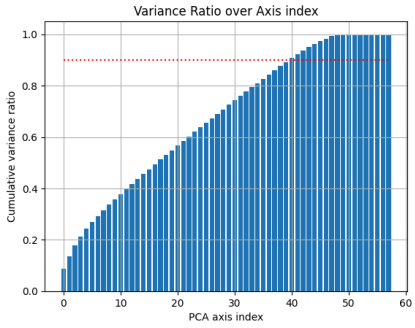
Figure 1.2.b: KMeans data points distribution for both datasets

Inertia measures the sum of squared distances of samples to their closest cluster center and it is only suitable for spherical clusters. We use the same elbow method on figure 1.2.a to determine the ideal number of clusters, which is 47 and 21 for income dataset and bank churn dataset respectively. Same trend of elbow clarity is also seen in the 2 inertia curve. As per before, figure 1.2.b analyzes the distribution of data points in each cluster and notes that it has uneven distribution too.

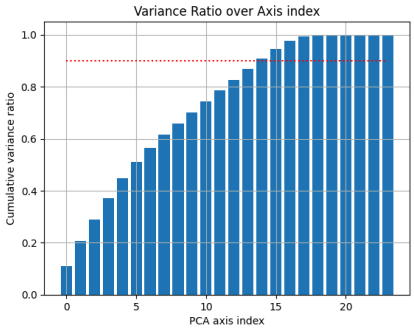
DIMENSIONALITY REDUCTION

PCA

By setting a threshold of 0.9 variance ratio, we can systematically select the number of axes after PCA transformation. Figure 2.1 shows the cumulative sum of dataset variance and the 0.9 cut off point. After dropping the least variant axes until we meet the 0.9 threshold, we are left with 40 and 14 features for income dataset and bank churn dataset respectively.



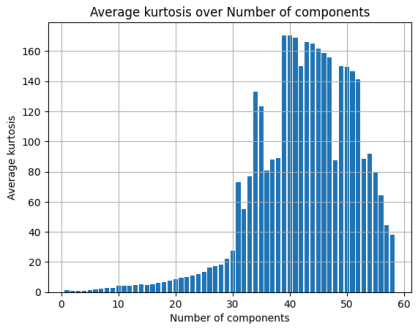
Income Dataset



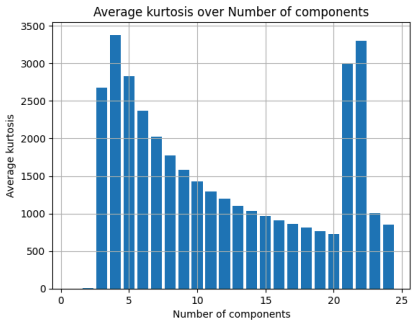
Bank Churn Dataset

Figure 2.1: Cumulative sum of axes explained variance ratio

ICA



Income Dataset



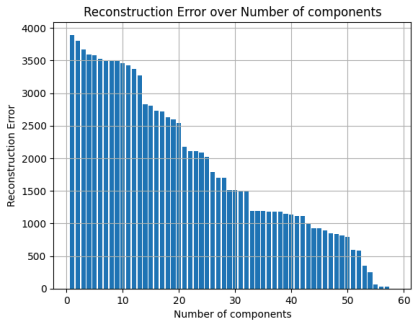
Bank Churn Dataset

Figure 2.2: Average kurtosis with varied number of components

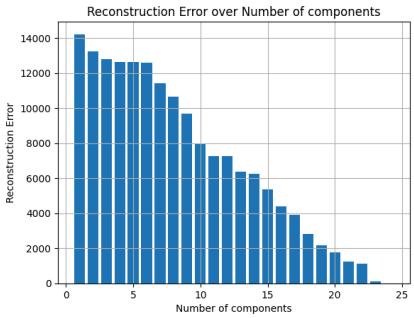
To determine the number of components in ICA, we decided to analyze the mean absolute kurtosis as we change the number of components. Kurtosis measures the "tailness" of probability distribution and a value close to 0 indicates that the distribution of the data is similar to that of a normal distribution. By selecting the number of components that produced the highest mean absolute kurtosis, we would have curated a set of features that behave least like a normal distribution. From figure 2.2, the most suitable number of components for income dataset and bank churn dataset is 39 and 4 respectively.

Random Projection

For random projection, we would like to select the number of components which will be able to reconstruct the dataset. As more components equal lower reconstruction error, we will use the elbow method again to determine the ideal number of components for random projection. Figure 2.3 illustrates the reconstruction error curve as we increase the number of components and we have chosen 33 and 21 for income dataset and bank churn dataset respectively. Note that it is quite unclear where the elbow is for the bank churn dataset.



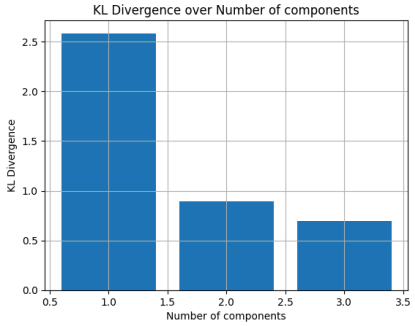
Income Dataset



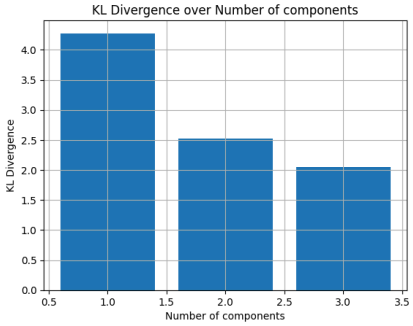
Bank Churn Dataset

Figure 2.3: Reconstruction error with varied number of components

t-distributed Stochastic Neighbor Embedding (t-SNE)



Income Dataset



Bank Churn Dataset

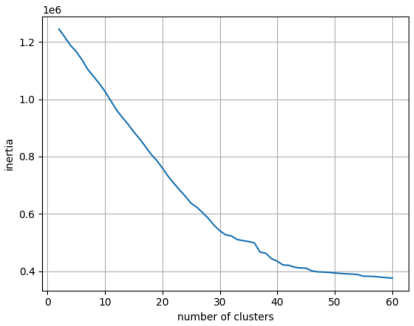
Figure 2.4: KL Divergence with varied number of components

t-SNE is more commonly used for data visualization. Only 1, 2 and 3 dimensions embeddings were generated due to computational complexity. We can observe that KL divergence is an inverse of the number of dimension embeddings.

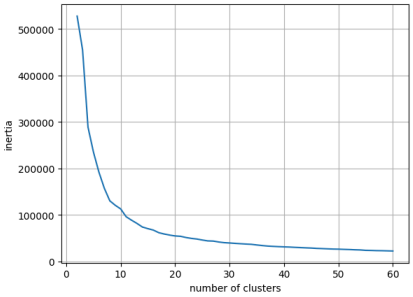
COMBINING DIMENSIONALITY REDUCTION & CLUSTERING

ICA + KMeans

The most intriguing combination can be attributed to a good number of components based on mean absolute kurtosis, which resulted in highly independent features from the original dataset. As observed before, the difference in elbow clarity between both curves remains the same due to the high number of categorical features in the bank income dataset.

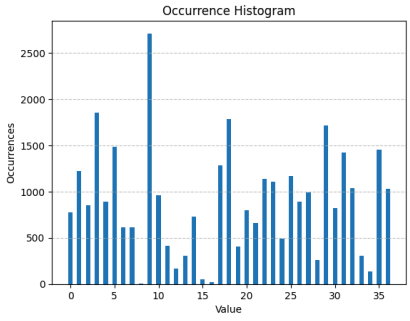


Income Dataset

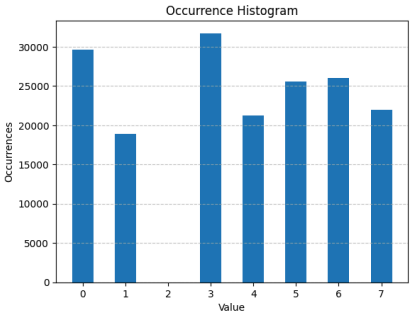


Bank Churn Dataset

Figure 3.1.a: KMeans inertia curve after ICA for both datasets



Income Dataset



Bank Churn Dataset

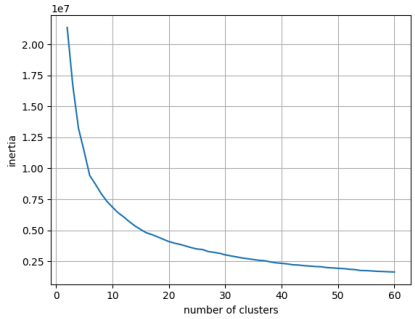
Figure 3.1.b: KMeans data points distribution after ICA for both datasets

Similarly, after deciding on an elbow value (37 for the income dataset and 8 for the bank churn dataset), we can take a look at the distribution of data points in each cluster. The income dataset

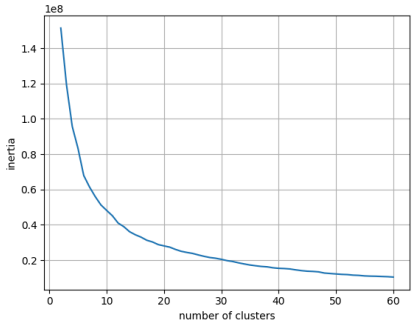
still has uneven distribution of data points while the bank churn dataset shows a better distribution, apart from a single outlier: cluster 2.

t-SNE + KMeans

Between the GMM BIC curve and the KMeans inertia curve after t-SNE, the latter has a clearer elbow pattern. From the KMeans inertia curve in figure 3.2.a, we set the number of clusters to be 16 and 22 for the income dataset and bank churn dataset respectively. Both datasets show the most evenly distributed data points in each cluster among every demonstration.

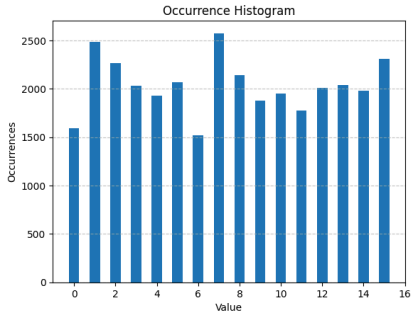


Income Dataset

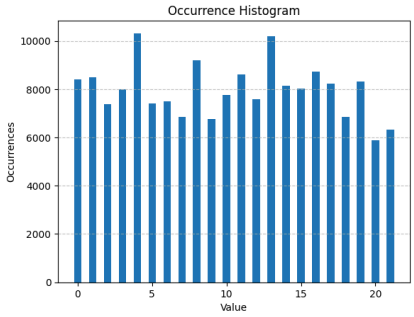


Bank Churn Dataset

Figure 3.2.a: KMeans inertia curve after t-SNE for both datasets



Income Dataset



Bank Churn Dataset

Figure 3.2.b: KMeans data points distribution after t-SNE for both datasets

DIMENSIONALITY REDUCTION NEURAL NETWORK

The income dataset is used in the neural network sections below.

PCA

While PCA allows us to find a new set of features with the highest variance, it is only limited to capturing linear relationships in the data. Combine this with the high number of categorical features in the dataset that will be one hot encoded, getting the top k features to meet a variance

ratio threshold might not be effective in capturing the underlying non linearity structure of the data as seen in figure 4.1.a. The learning curve seems to be converging but as the training example increases, the curves' instability signals that the model is sensitive to small changes in the dataset. This instability was not observed in assignment 1 and thus, it is due to the model generating vastly different PCA features with every minor change in the dataset.

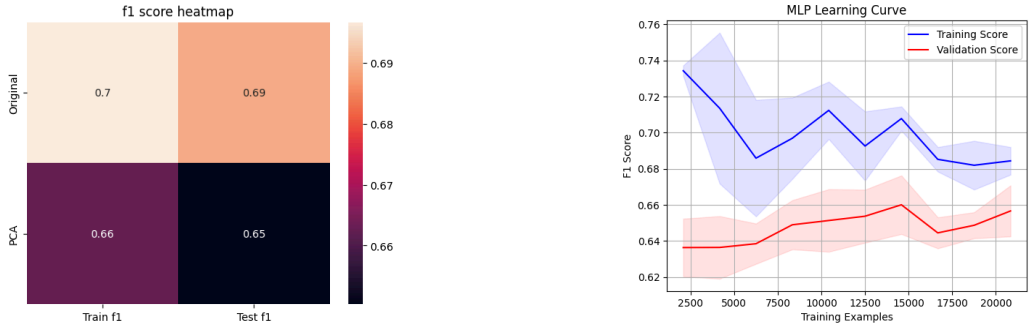


Figure 4.1: PCA NN f1 score heatmap and learning curve

t-SNE

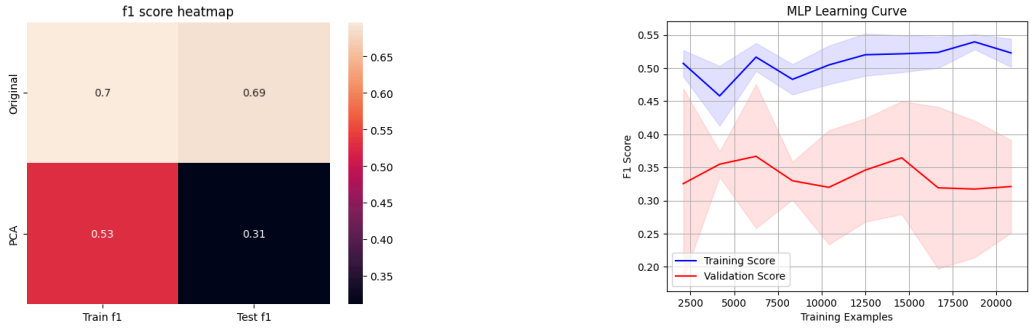


Figure 4.2: t-SNE NN f1 score heatmap and learning curve

t-SNE produced disappointing results for a few reasons. The 3 dimensions generated are not sufficient in representing the whole dataset so the f1 of the training dataset is only 0.53. Next, every t-SNE transformation of the test dataset is a refitting and transformation on the isolated test dataset input, creating heterogeneous inputs for the neural network which results in a very low 0.31 f1 score. The learning curve also shows that the testing curve remains relatively flat.

DIMENSIONALITY REDUCTION + CLUSTERING NEURAL NETWORK

The income dataset in this section is curated by combining the reduced dimension and both GMM and KMeans clustering labels from the original normalized dataset. The aim is to reduce

the dimensions of the original dataset to mitigate the curse of dimensionality and use cluster labels to provide data points similarity observed from the original dataset.

PCA

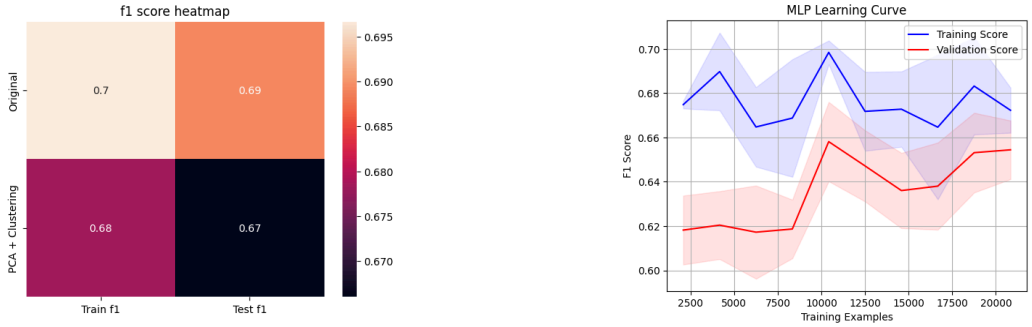


Figure 5.1: PCA + clustering NN f1 score heatmap and learning curve

Adding clustering labels shows a slight improvement in f1 score compared to the previous section. However, it still did not produce a better result than the original dataset. The flat training curve combined with an improving validation curve suggest that the model is suffering underfitting, possibly due to loss of important non linear relationships in the PCA reduced dimensions. With the added modeling complexity, the addition of both training and prediction time makes this method slower than the original method (165s vs 65s). Both scores and duration factors suggest that this new method is not ideal for the given dataset.

t-SNE

Despite the addition of cluster labels as dimensions, t-SNE exhibits consistent performance issues. The inclusion of cluster labels led to a slight improvement in f1 scores, but the poor performance compared to the original dataset remains unchanged.

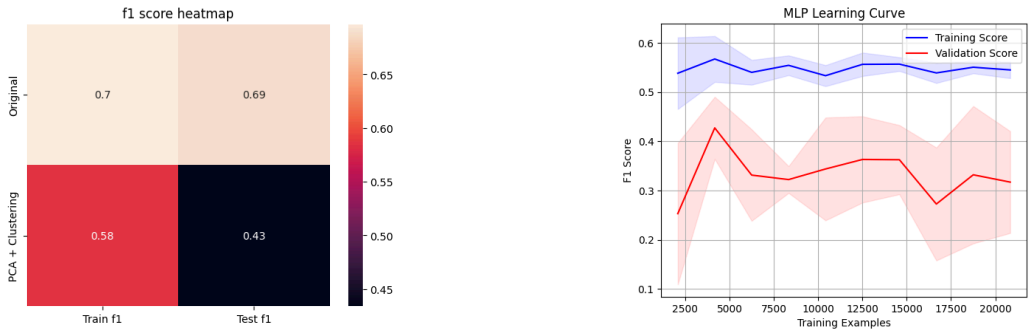


Figure 4.2: t-SNE + clustering NN f1 score heatmap and learning curve