Studying the Effects of Clustering and Dimension Reduction on Neural Networks Performance

Nico Medellin[1]

I. Abstract

The purpose of this study is to investigate the effects of clustering and dimension reduction on neural networks performance. We are evaluating two clustering algorithms, K-Means and Expectation Maximization (EM), along with three linear dimensionality reduction techniques: Principal Component Analysis (PCA), Independent Component Analyses (ICA), and Randomized Projections (RP). We apply these algorithms to two small, multi-featured datasets: the Customer Personality Dataset and the Spotify 2023 Dataset. The neural net trained on 2 k-mean clusters performed the best out of all the neural networks, beating the baseline neural net by 3 percent on accuracy and F1.

II. Introduction

This project attempts to explore the application of clustering and dimensionality reduction algorithms on real-world datasets. Building our prior supervised learning projects, the objective of this study is to investigate how clustering methods and dimension reduction techniques can be used to improve neural networks.

Specifically, we are evaluating two clustering algorithms—K-Means and Expectation Maximization (EM)—alongside three linear dimensionality reduction techniques: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Randomized Projections (RP). These algorithms are applied to two small, multi-featured datasets: the Customer Personality Dataset and the Spotify 2023 Dataset.

We hypothesize that dimensions reduction techniques and clustering algorithms should improve the overall performance of neural networks. However, due to the simple nature of our datasets, we believe that the improvement may be marginal.

III. Methods

The purpose of dimension reduction is to ideally reduce high dimension datasets to lower dimensional spaces by mapping the features to a few primary components. This is often used as an approach to deal with the curse of dimensionality that is often, which is important given that the more features a dataset contains, the more sample data is needed. Dimension reduction is also often used as an approach to reduce the amount of noise in a dataset.

Dimensionality reduction (DR) is typically useful as it often improves model performance and reduces overfitting on training data by eliminating redundant features. For very large training sets, DR can speed up training of models.

We will be using three different forms of DR for this report: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Random Projection (RP). A brief overview of the various forms of DR can be found below:

- PCA
  - Computes components that are orthogonal (uncorrelated) to one another
  - The first component aims to captures the most variance possible, while the second aims to capture the most variance orthogonal to the first.
- ICA
  - Attempts to break down the data into additive and independent non-Gaussian components.
  - ICA attempts to break components down into statistically independent components
  - Aims to maximize non-Gaussianity (i.e. kurtosis)
- RP
  - Projects data to a lower-dimensional space using a randomly generated matrix while trying to preserve pairwise distances
  - Does not use eigenvectors, it only use random linear projection
  - RP is known to be computationally fast and memory-efficient.

Throughout this report we will be using various metrics to measure clustering performance and dimension reduction performance, and neural net performance. Below are high level definitions of the metrics we will be using:

Clustering:

- BIC Score (Bayesian Information Criterion)
  - A model selection criterion that balances goodness of fit with model complexity by penalizing the number of parameters.
  - This metric penalizes model complexity.
  - A lower BIC value typically indicates a better model fit.
- AIC Score (Akaike Information Criterion)
  - An estimate of the information lost by a model
  - **Lower is better**
- Log-Likelihood
  - Quantifies how probable the observed data is under a particular model
  - It **does not** publish overfitting of data.

[1]Georgia Institute of Technology, Department of Computer Science

  - It typically always increases in value as the number of clusters increases
  - **Higher is better.**
- Silhouette Score
  - A clustering quality metric based on both cohesion (how close points are within the same cluster) and separation (how far they are from other clusters)
  - This evaluates how well-separated and internally cohesive clusters are; often used for estimating the number of clusters in unsupervised learning
  - **Higher score is better**
- **Davies-Bouldin Score**
  - A clustering metric that compares the overlap between clusters based on scatter and separation.
  - This score is often associated with penalizing overlapping or poorly sepereated clusters
  - A lower score is better.

IV. Results

A. Base Clustering Results

Analyzing the initial results of our dataset, we try to determine the optimal number of clusters for for both our spotify and customer preference dataset. We see that across the board for both K-Means and GMM clustering, that we get the highest silhouette scores for both datasets when clusters are equal to 2 which is surprising given that the underlying datasets are vastly different (characteristics of Spotify data vs. customer spending habits).

K-means and GMM both support two clusters for the for the spotify dataset. For the CPA dataset, GMM BIC, AIC, and Davies-Bouldin support closer to 25 to 30 clusters. K-means for CPA shows a clear preference for a cluster size of 2 given the high silhouette score (0.20) and the ralatively low Davies Bouldin score (2.2, for context the lowest DB score is ~1.93 for the CPA dataset)

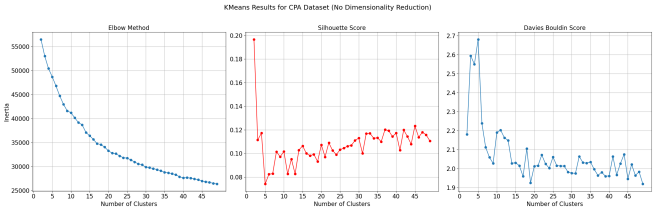

Figure 11 *Base K-Means Clustering on CPA Dataset*

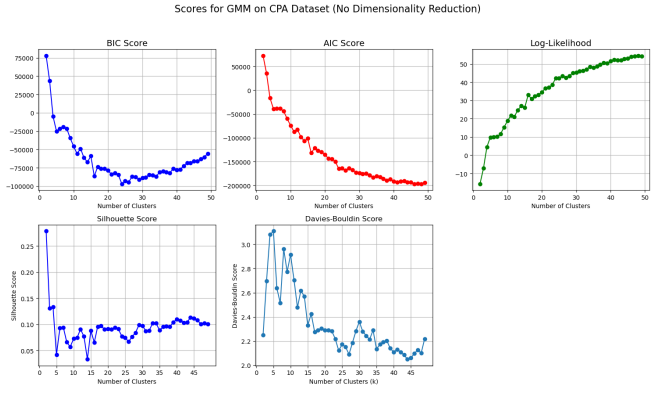

Figure 11 *Base GMM Clustering on CPA Dataset*

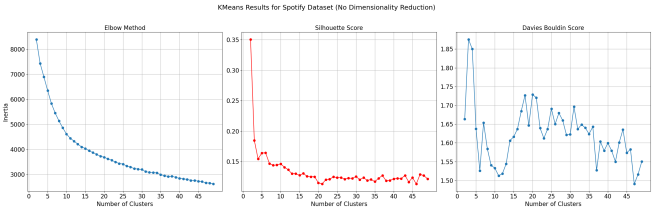

Figure 11 *Base K-Means Clustering on Spotify Dataset*

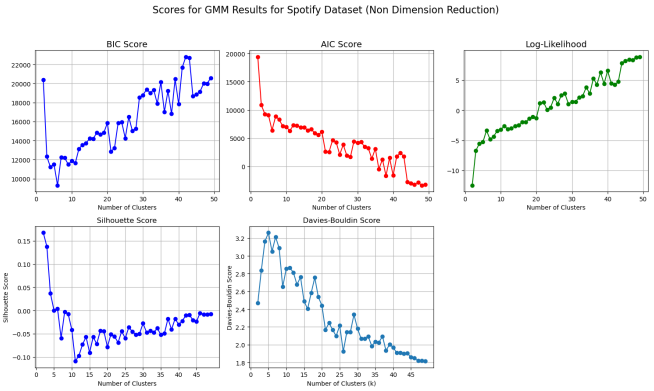

Figure 11 *Base GMM Clustering on Spotify Dataset*

B. Dimension Reduction - CPA

For the CPA dataset, we see that for PCA the optimal cutoff for number of components is around 20 components as it maintains 90% of all variance. A majority of variance is accounted for in a single PCA component, with a steep drop off with the following components. For ICA, the cutoff is around 15 components where 90% of kurtosis is maintained.

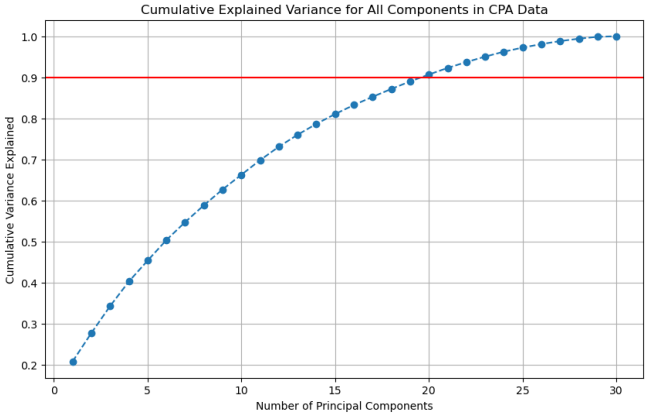

Figure X *Scree Plot CPA*

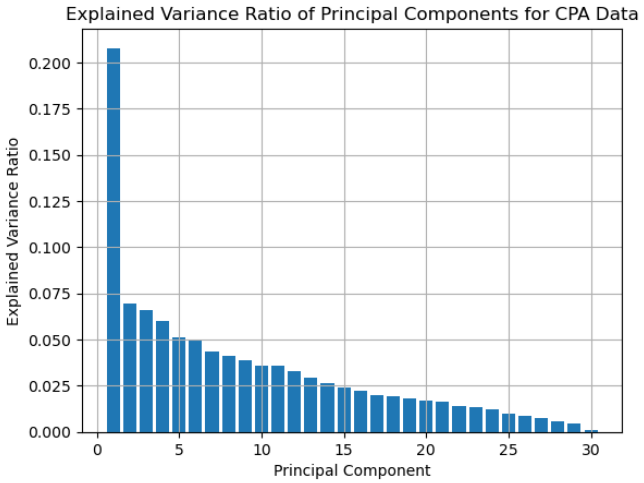

Figure X *Order of Importance PCA Dimensions for CPA*

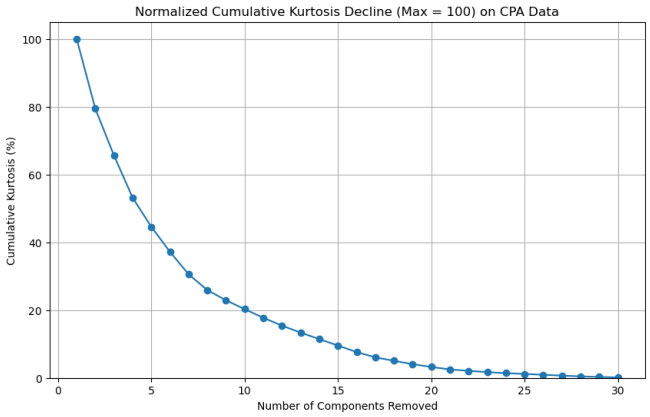

Figure X *Normalized Kurtosis CPA*

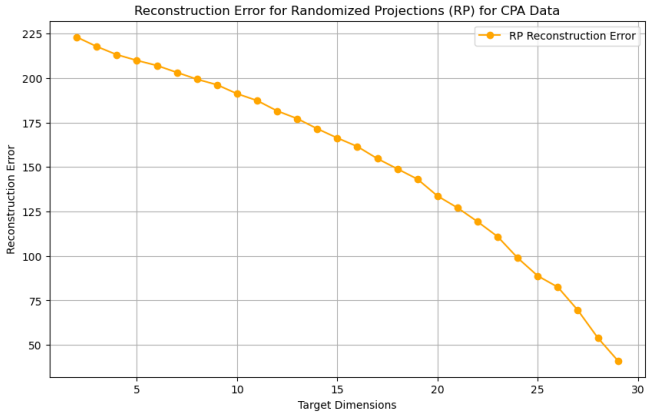

Figure X *Reconstruction Error by CPA*

C. Dimension Reduction - Spotify

D. Clustering on Dimensionally Reduced Datasets

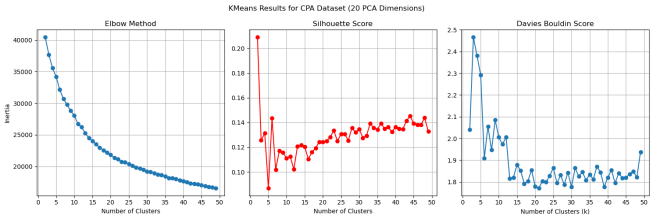

Figure X *KMeans on PCA Reduced CPA Data*

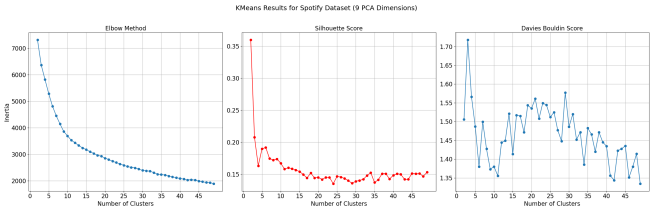

Figure X *KMeans on PCA Reduced Spotify Data*

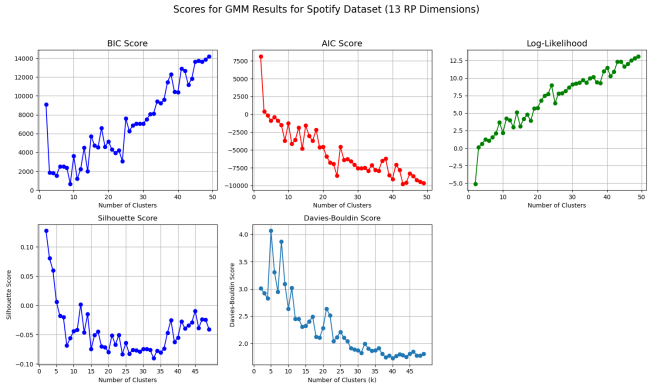

Figure X *GMM on RP Reduced Spotify Data*

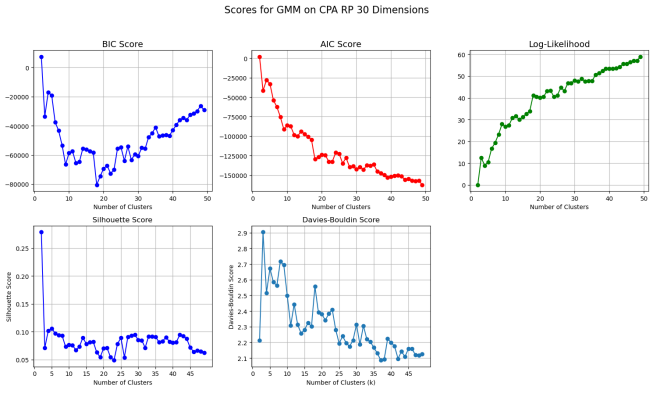

Figure X *GMM on RP Reduced CPA Data*

E. Training Neural Networks on Dimensionally Reduced Data and Clustering Labels

We see that the neural net trained on the 7 dimension ICA model may not have the highest accuracy or F1 as training size increases, however we do see that it has the smallest amount of overfitting shown by the small difference

between test and training results. The random project model performed the worst out of all the models when looking at accuracy and F1. The neural net trained on 2 k-mean clusters performed the best out of all the neural networks, beating the baseline neural net by ~3% on accuracy and F1. We also see that the K-means trained neural net retains a higher level of accuracy and F1 as training size increases, where we see the base neural net has a drop of both of those metrics as training size increases on the training set. It will be interesting to see if the base neural net training set would continue to perform worse if the training size of the model were to increase.

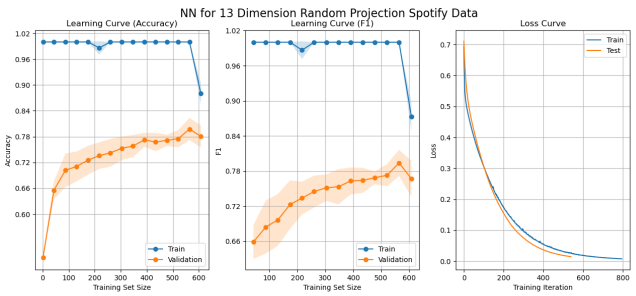

Figure X *NN for 13 Dimension Random Projection Spotify Data*

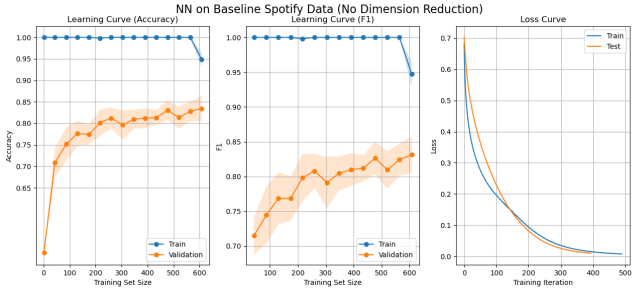

Figure X *NN on Baseline Spotify Data (No Dimension Reduction)*

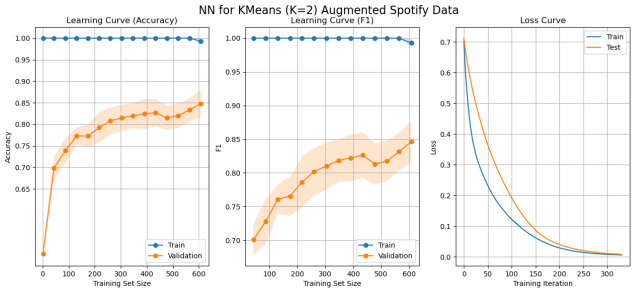

Figure X *NN for KMeans (K=2) Augmented Spotify Data*

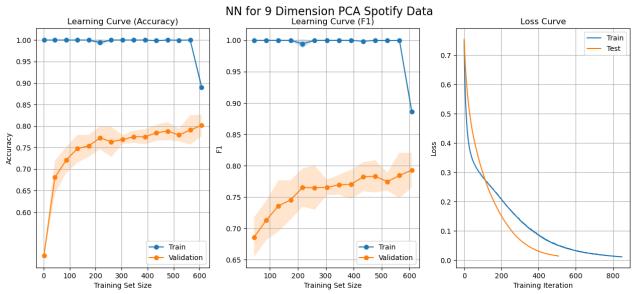

Figure X *NN for 9 Dimension PCA Spotify Data*

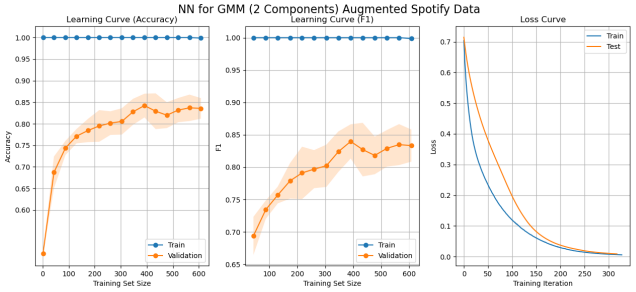

Figure X *NN for GMM (2 Components) Augmented Spotify Data*

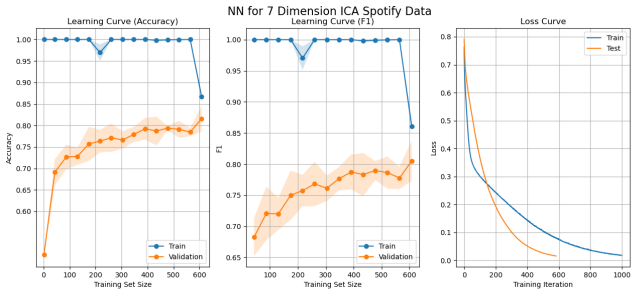

Figure X *NN for 7 Dimension ICA Spotify Data*

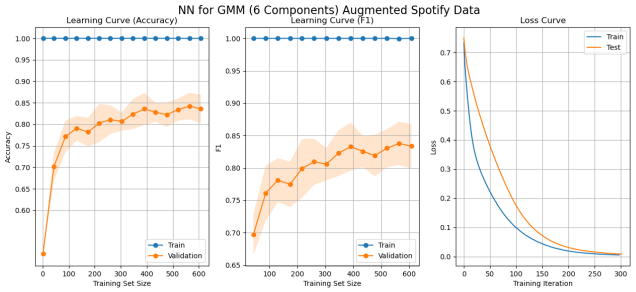

Figure X *NN for GMM (6 Components) Augmented Spotify Data*