Assignment 4 – sxb180048 – Sai Pratheek Banda

Loan Status Classification

Classification Problem:

The aim of the problem is to predict the classification of the loan status to **'Fully Paid'** and '**Charged OFF'**. The interesting part of this classification is that it will help us understand the patterns for customers if they are going to get "**Charged OFF'**. **Class 0(Fully Paid)**: 28972 Observations| **Class 1 (Charged OFF)**: 7451 Observations

EDA:

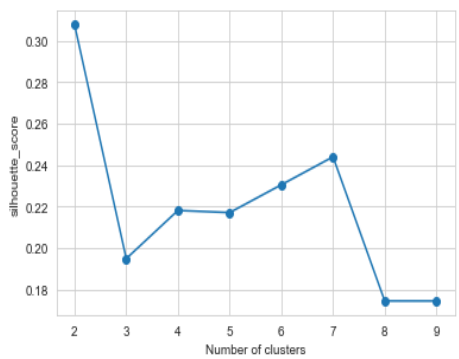
1) Features 'Customer ID', 'Loan ID' were insignificant due to its inability to explain the target variable therefore dropped
2) Though the Dataset contains 100000 Records, after dropping the not available rows, we get consistent data of 26423.
3) The Features Term, Home Ownership, Purpose and Target Variable were converted into dummy variables.
4) In the process I have chosen to use F-1 Score as metric because it's of utmost importance that we get the Class 1 classification better than other measures or accuracy.

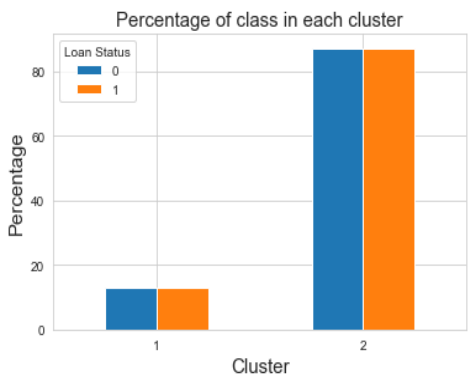
Data Sampling:

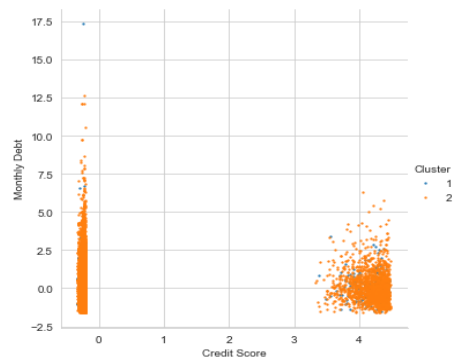
Since the data is heavily unbalanced in terms of the target classification '**Charged OFF',** The consistent data was split into train and test. Due to heavy imbalance and for the quest of a better model I have under sampled the train data to a **Class 0(Fully Paid)**: 7000 Observations| **Class 1 (Charged OFF)**: 5194 Observations.
It is a tradeoff I have chosen to for training the model to the information lost in the process.

CLUSTERING*:*

K Means:

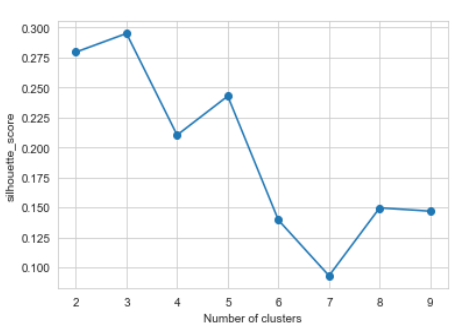


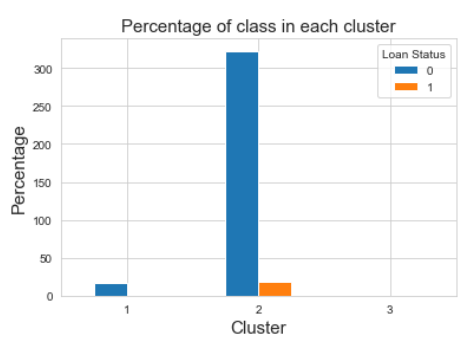


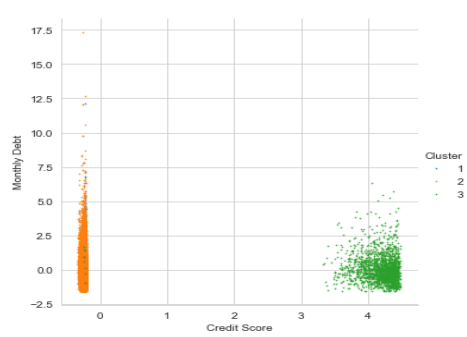


K means clustering done with multiple Clusters along which Silhoutte score was calculated for each Cluster. The optimal number of clusters found **(K)** was two.	The graph shows us the distribution of the observations in each cluster. The data is highly skewed in nature due to the high imbalance in the obervations but uniform in nature.	The seperation between the Clusters is not clear as you see from the graph. The 1st cluster is visualy distributed around the other and there clear overlap.

Expectation Maximization:



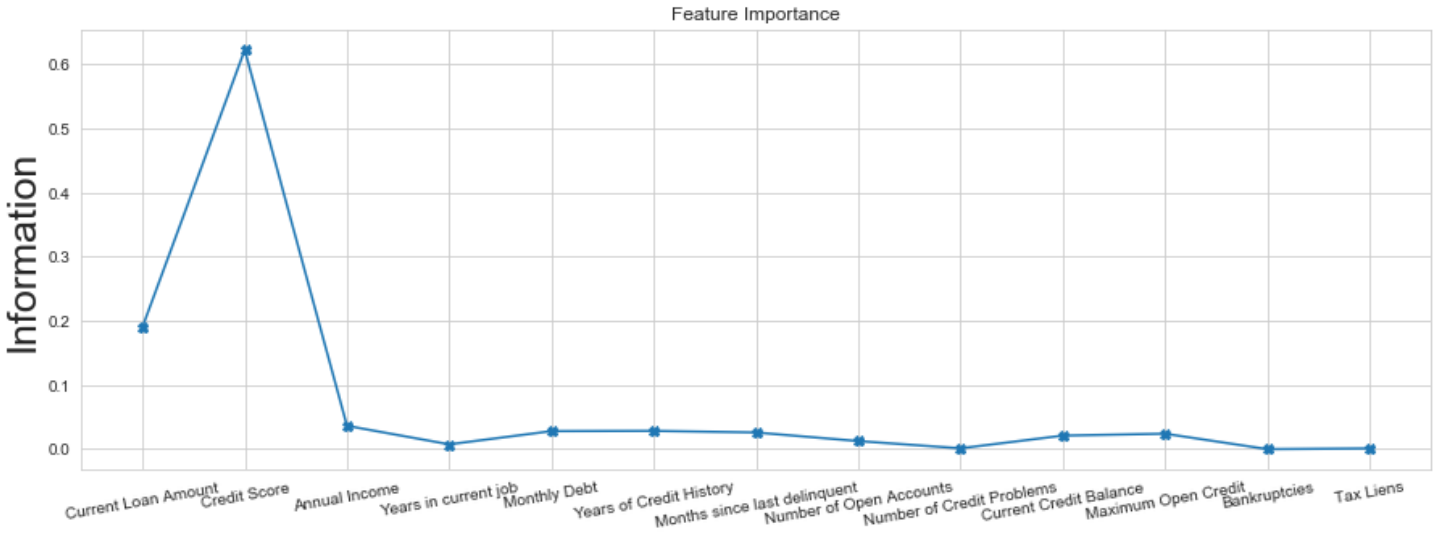




Expectation Maximization was done with Gaussian Mixture model, along with silhouette score was calculated, the **optimum number of clusters to be found was three.**	The Observation distribution is further divided, with the cluster having only 9 observations. It unevenly distributed across and within.	The Clustering isn't clearly separable. It clusters the **same way K means does. The clusters are not compact but spread out.** This is due to the clusters being dependent on various features.

Feature Selection:

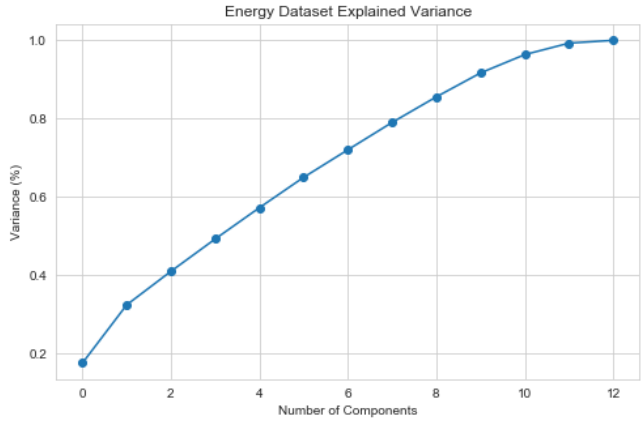
Decision Tree:



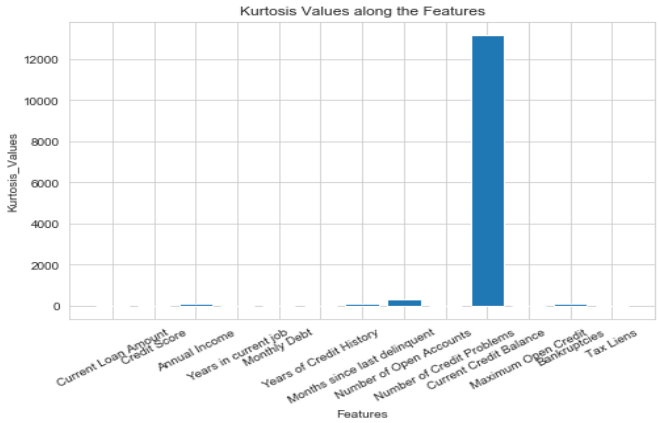
The features picked through feature Selection 'Credit Score', 'Current Loan Amount', 'Annual Income', 'Maximum Open Credit', 'Current Credit Balance', 'Monthly Debt'

Feature Reduction:

Principal Component Analysis:



Independent Component Analysis:



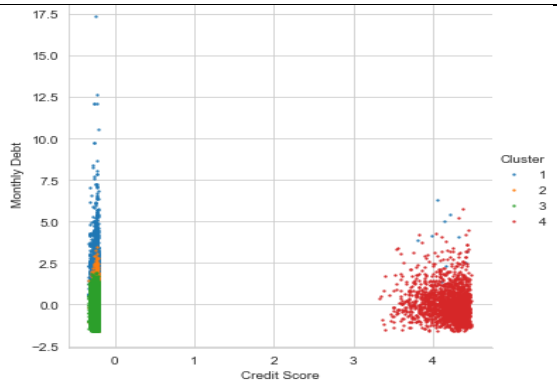
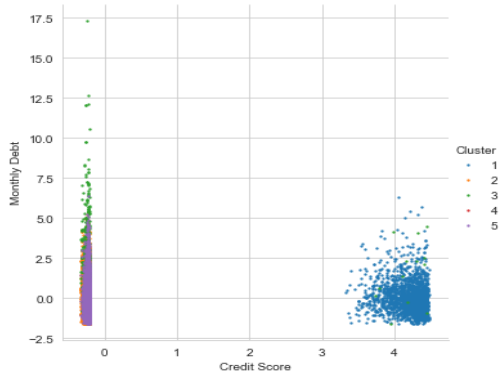
Principal Component Analysis: The graph is **the Explained Variance** across components, we pick 10 Components as it explains **90%** of the data.	To pick the best features for ICA, we use the kurtosis values, and pick the features with value above 100. which are 'Annual Income', 'Months since last delinquent', 'Number of Open Accounts', 'Current Credit Balance', 'Bankruptcies'

Random Optimization:

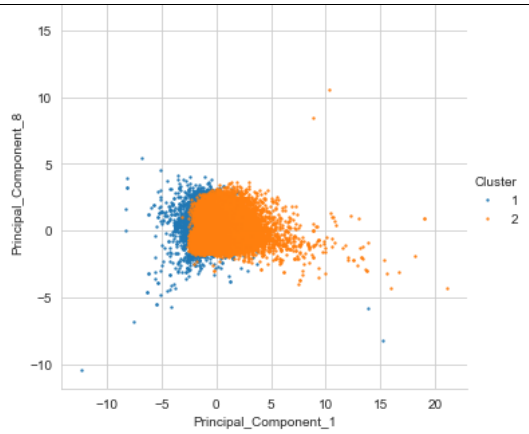
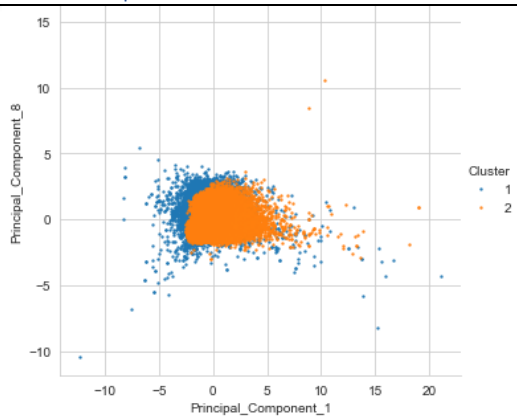
We use Gaussian Random Projection to pick Random Components. We pick 4 components which has previously proven to be the number best suited for understanding the data.

Clustering After Dimension Reduction:

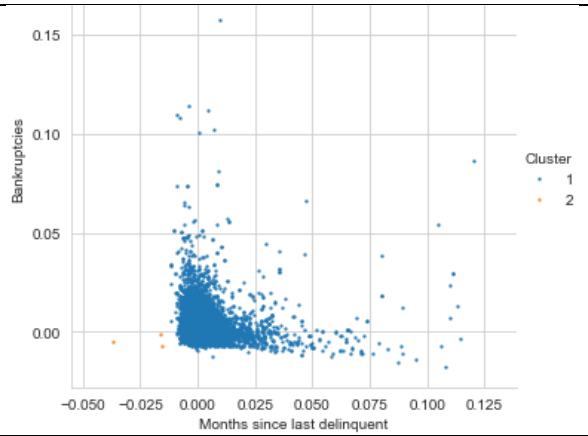
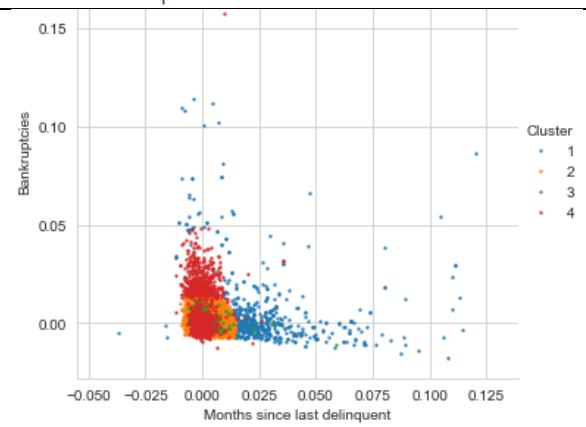
Decision tree:

K Means	Expectation Maximization
	
Explanation: Optimum Clusters using Silhoutte Score is 4 More clusterd formed did not help the alignment of class labels.They did not naturally line up either. This explains that the features could not distingush the dependent variable well enough.	*Explanation:* Optimum Clusters using Silhoutte Score is 5 The clusters are compact in nature but are skewed and are not aligned themselves in terms of class labels or the clutser itself. Overlay in the distribution though distinguishable is not clearly seperated.
Difference: K Means on the Feature Selection with descion tree did not change the clustering pattern nor was it effeictive for us to learn anything meanigful from the data.	*Difference:* Expectation maximization on Feature Selection with decision tree was also redundant to classify the observations and into the target classes woth more clusters.

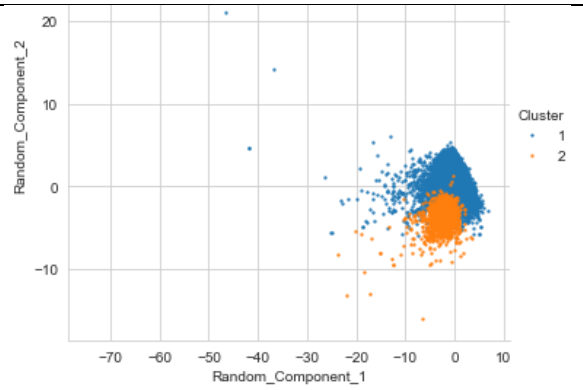
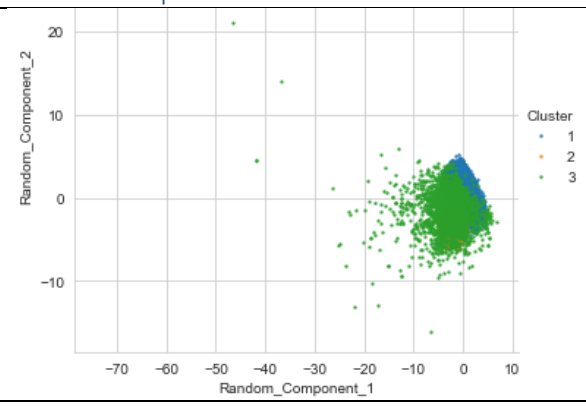
Principal Component Analysis:

K Means	Expectation Maximization
	
Explanation: Optimal Clusters using Silhoutte core is 2 The Clustering though compact does a good job with PC1 and PC8. The compononets help distinguish the Class lables well thus understanding the variance within.	Explanation: Optimal Clusters using Silhoutte core is 2 Clusters are compact in nature but there is a clear overlay as it superimposes itself over each other.
Difference: Does a better job when compares to the decision tree but still does not converge to make it distinguishable	Difference: The Clusters converges but not in way to help the class labels, they do a better job than Decision tree to classify the class labels.

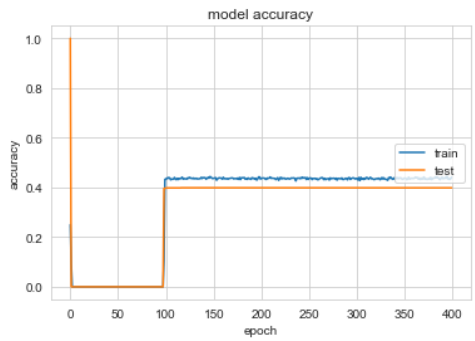
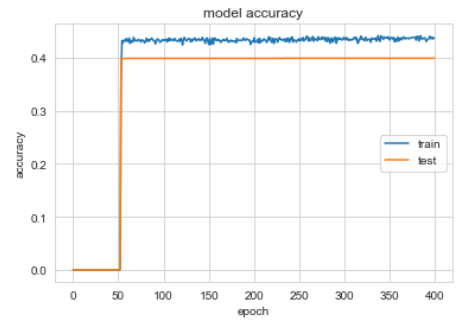
Independent Component Analysis:

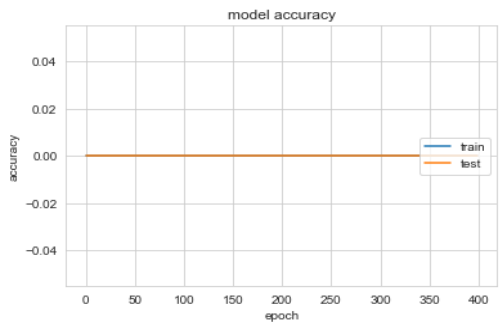
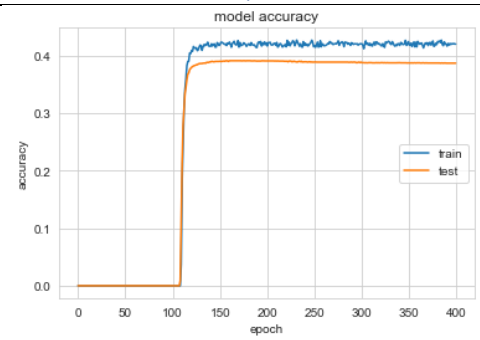
K Means	Expectation Maximization
	
Explanation: Optimal Clusters using Silhoutte core is 2 The Clustering compact, does a good job with being distinguishable . Due to extremely imbalanced data, is has a hard time to disguish within the target labels	Explanation: Optimal Clusters using Silhoutte core is 4 Cluster is spread out and is overalapped into each other and is clearly a bad combination to work on.
Difference: ICA does not do abetter job than Decision tree and PCA and the simple K means, ICA does better to understand the data.	Difference: Expectaion maximization does not improve with the original , Decision tree, PCA or ICA.

Random Optimization:

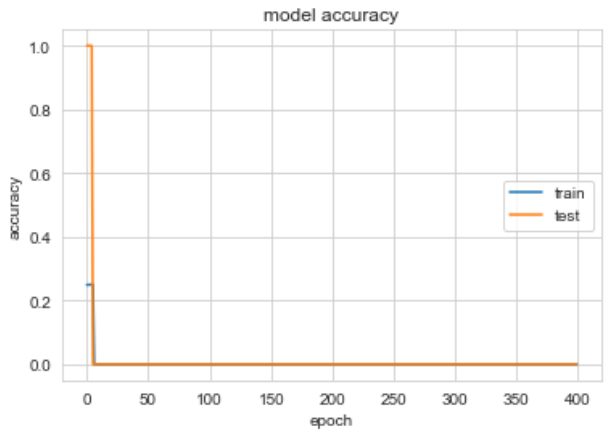
K Means	Expectation Maximization
	
Explanation: Optimal Clusters with Silhouttee Score is 2 Highly skewed clusters,with superimposition on each other, resulting in bad result to exemplify the class labels.	Explanation: EM with random optimization gives a better result in understanding the data , but the overlap is still huge for it to be considered a good clustering algorithm for this data.
Difference: Definitely better than K means and PCA. Thus can be considered a competent clustering for the random optimization features.	Difference: EM does well when compared with the other Dimension reduction techniques but still isnt operable to consider.

Artificial Neural Networks: Under-sampling on train data was done for the model to learn better.

Decision Tree	Principal Component Analysis
	
Optimal Neurons and Layers: (6,2,2,6)	Optimal Neurons and layers: (8,10,4)
F1 Score: 40.09%	**F1 Score:** 41.66%
Feature Selection is the worst when compared to the all the algorithms previously ran including LOGIT, KNN, DT.	Though there is an increase in the F1 Score and comparing the previous algorithms run, this is still a very low score.

Independent Component Analysis	Random Optimization
	
Optimal Neurons: -	Optimal Neurons: (2,3,1)
F1 Score: 0	**F1 Score:** 41.06
ICA tries to separate the data based on signals independent of each other, this dataset is not independent of each other thus does not work in dimension reduction to give a result.	Using Random Components worked to find a good F1 Score but is still less than a flip coin. Which does not work well to find a good model.

Artificial Neural Network on Clustered Data



- Creating a separate dataset with just the clusters of K means and Expectation Maximization as features.
- These features did not enough information to carry out any kind of classification
- The high imbalance in the original dataset contributed to the results obtained.
- Under sampling did not help as repeating the same cluster results still does not explain why, rather just gives a random set of numbers.
- Thus, running ANN on this dataset with clusters as features is not a adaptable model.