Prepared by: Jyotsna Nagaprasad Ambekar Date: 08/07/2022

BUAN 6341 Applied Machine Learning

ASSIGNMENT NO 4
Seoul Bike Data

Executive Summary

- **Classification problem with data classified based on median.**
- **Features without high correlation are used to find clusters through K-Means and Expectation Maximization algorithms**
- **The silhouette scores are high when more features are included and dimensionality reduction has advantage on lowering the complexity but there is a tradeoff with accuracy.**
- **Prediction accuracy can be further improved by implementing XGBoost or Random Forest Algorithm that is apt when data has multiple discrete features.**

Context

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. The goal is to help commuting in the cosmopolitan cities and reduce the pollution amounts caused by cars, individual motorcycles etc. and create green cities altogether. It also is a healthiest way for travelling to work or school. It is important to make rental bike available and accessible to the public at the right time lessening the waiting time. Eventually, providing the city with stable supply of rental bikes becomes a major concern as it can be used as alternative to the commuters in cities. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Introduction

In this project, the objectives were to implement Neural Network algorithm to predict the Rented Bikes count on a given day with given attributes. This report details various experiments conducted using the dataset.

About the Data

The dataset consists of 14 features and 8760 records with no missing values. The data contains information regarding weather and how much solar radiation is observed on the day, seasons, holiday, functional hours etc. The target variable is the rented bike count per hour.

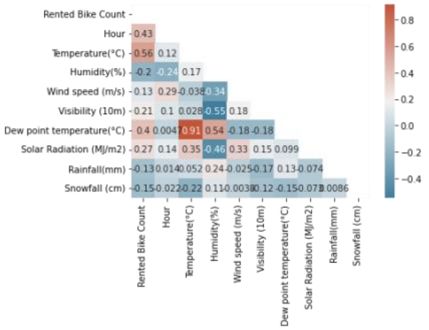
Attribute information

Date : year-month-day
Rented Bike count - Count of bikes rented at each hour
Hour - Hour of the day
Temperature-Temperature in Celsius
Humidity - %
Windspeed - m/s
Visibility - 10m
Dew point temperature - Celsius
Solar radiation - MJ/m2
Rainfall - mm
Snowfall - cm
Seasons - Winter, Spring, Summer, Autumn
Holiday - Holiday/No holiday
Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

The independent variables used for training the model are Hour, Temperature(°C), Humidity(%), Wind speed(m/s), Visibility (10m), Solar radiation (MJ/m2), Rainfall(mm), Snowfall, Seasons & Holidays one hot encoded columns. The season column encoded as 0 is Autumn, 1 is spring, 2 is summer, 3 is winter. The Holiday column encoded as 4 is Holiday and 5 is No-Holiday

Data Preprocessing

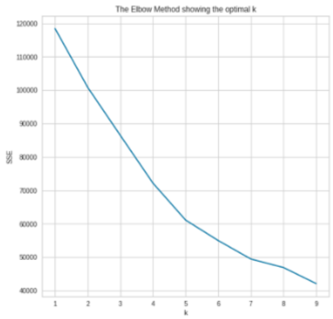
It is the technique in which data is processed so that it is appropriate to feed the model. Preprocessing steps applied were One hot encoding to encode categorical variables like Seasons & Holiday columns, Splitting the dataset into training and test sets, Feature scaling for standardizing data. Firstly, data exploration was done by understanding the types of data present and description of data. The Date column which was initially string was converted to DateTime format.



The temperature and dew point temperature have correlation value of 0.91. Due to this high correlation, multi collinearity arises if both the features are used to develop the model. Hence dew point temperature is dropped from the independent variables.

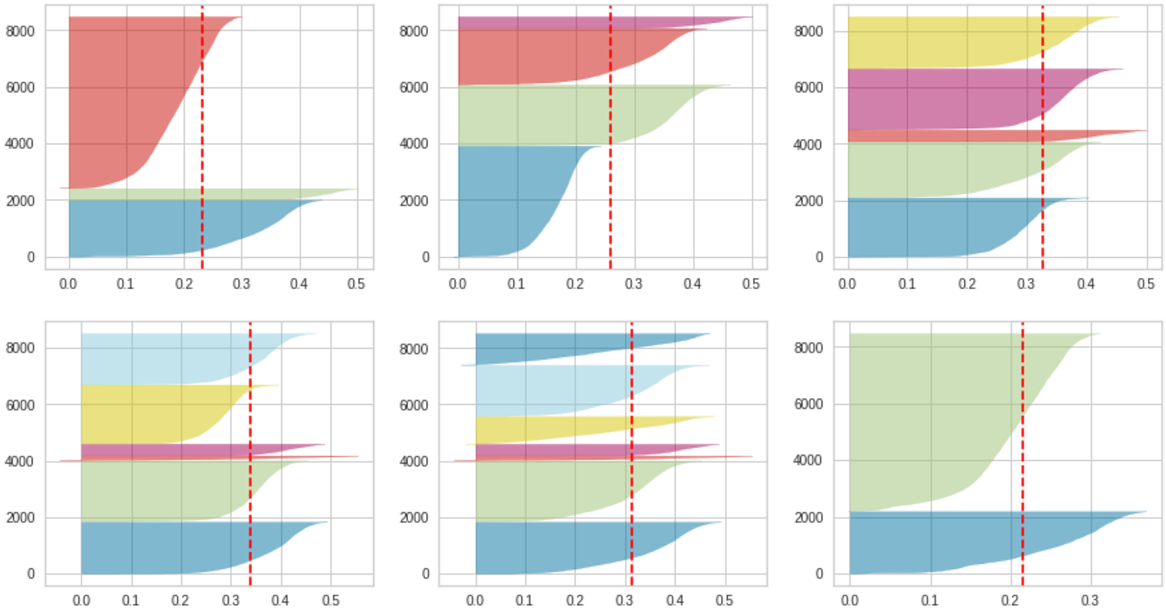
Experimentation

Task – 1: K-Means & Expectation Maximization clustering algorithms were implemented. Clustering is an unsupervised Machine Learning technique. Before implementing the clustering algorithms, optimal number of clusters that are possible in this dataset was found by 'Elbow method'. Elbow method is based on distortion & inertia which is used to find optimal number of clusters by fitting model with a range of k values. It is a Line plot between 'Sum of squared errors or Inertia'. Distortion is the average of squared distances from cluster centers of respective clusters. Inertia is the sum of squared distances of samples to their closest cluster center. The graph represents change in inertia corresponding to the increase in number of clusters. Then the clusters number is decided looking at the 'elbow' of the graph which is 5 for this dataset.



With all the features, KMeans algorithm was implemented with 5 clusters. K-Means clustering is a simple unsupervised machine learning algorithm which aggregates the data points based on some Euclidean or non-Euclidean similarity between the data points. The data points are allocated to each of the clusters through reducing the in-cluster sum of squares between the imaginary cluster centroid and the individual data points.

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The value of Silhouette score varies from -1 to 1. If the score is 1, the cluster is dense and well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters. A negative score [-1, 0] indicate that the samples might have got assigned to the wrong clusters. 'silhouette_score' for the data set is used for measuring the mean of the Silhouette Coefficient for each sample belonging to different clusters. The silhouette_score for the K-Means algorithm developed when all features are used to develop is 0.326. This shows that the clusters are overlapped with data points very near to decision boundary. Silhoutte plot for different number of clusters -



This graph also shows 5 is the optimum number of clusters with almost equal fluctuations. There are 2087 data points in cluster 1, 1972 in cluster 2, 407 in cluster 3, 2159 in cluster 4, 1835 in cluster 5.

One important characteristic of K-Means is that it is a ‘hard clustering method’ i.e., it will allocate each point to one cluster. A limitation to this approach is that there is no uncertainty measure or probability that tells us how much a data point is associated with a specific cluster. Gaussian Mixture model – Expectation Maximization is one such method which tells us the probability of one data point belonging to different clusters. EM algorithm is an iterative approach that cycles between two modes. The first mode attempts to estimate the missing or latent variables, called the estimation-step or E-step. The second mode attempts to optimize the parameters of the model to best explain the data, called the maximization-step or M-step. The algorithm tries to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset. Since we saw there is lot of overlap between clusters in K-Means, by knowing the probability we can understand the probability of a distribution belonging to the cluster. This implies the class labels almost align with the clusters.

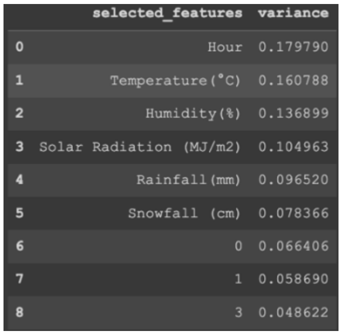
Dimensionality Reduction

Task – 2: Four dimensionality reduction techniques were implemented. Namely, Forward selection, Principal Component Analysis, Independent Component Analysis, Random Projections. Forward selection was done taking the decision tree classifier and giving different values for the ‘k-features’ parameter. The maximum accuracy is obtained when there are 9 features. There is variation of score when we include more than 9 features which can be ignored for reduction in complexity. The maximum score obtained during forward selection is 90.83% with 9 features. These 9 features selected are **‘Hour’, ‘Temperature(°C)’, ‘Humidity (%)’, ‘Solar Radiation (MJ/m2)’, ‘Rainfall(mm)’,**

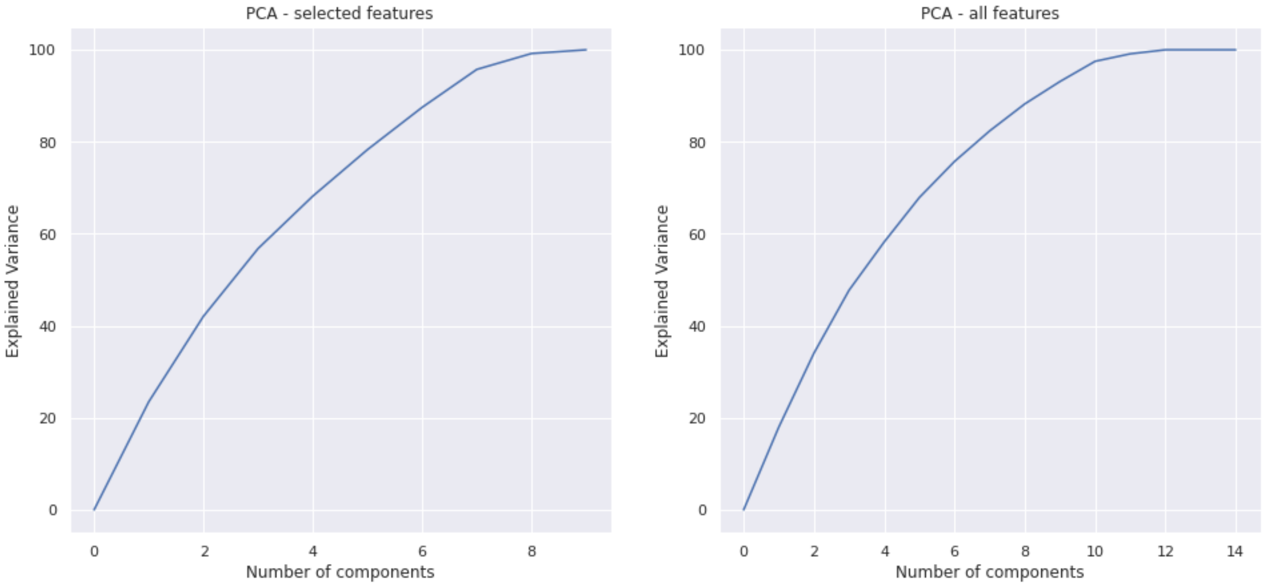
'Snowfall (cm)', 'Autumn', 'Spring', 'Winter'. Being a holiday or not surprisingly has considerably very low effect on the count of bikes rented. This might indicate that people use rented bikes not only to commute to office or school but also on holidays to visit places. May be on the holidays they visit stores to buy groceries or tourists are interested in riding bikes. Comparing the accuracy scores of the models developed with all the features and selected features varies from 90.9% to 90.3%. Hence, we can absolutely use smaller number of features since there is very little decrease in the decision tree model. On selecting 9 features, increase in scores is on addition of one feature is shown by the table -

1	0.7472
2	0.8149
3	0.8743
4	0.8973
5	0.9013
6	0.9042
7	0. 9045
8	0.9071
9	0.9083

Principal Component Analysis, which is a dimensionality-reduction method is used to transform large set of variables into smaller one that still contains most of the information of the large dataset. The reduction in dimensionality always comes at the expense of accuracy. The trick is to use PCA with very little trade off from accuracy for simplicity. Here, PCA was done using these selected features as well as all the features. The following table shows how much variance in the dataset is explained by the selected features. The same table for all features shows how little variance is explained by adding the remaining features. Hence PCA helps in selecting the features which are essential in building the model thus reducing the time spent and complexity. This is explained by the plot below which states that using only the 9 features, we got 100% variance explained. For ease of visualization, all dimensionality reduction algorithms were developed with 2 components and 3 components.



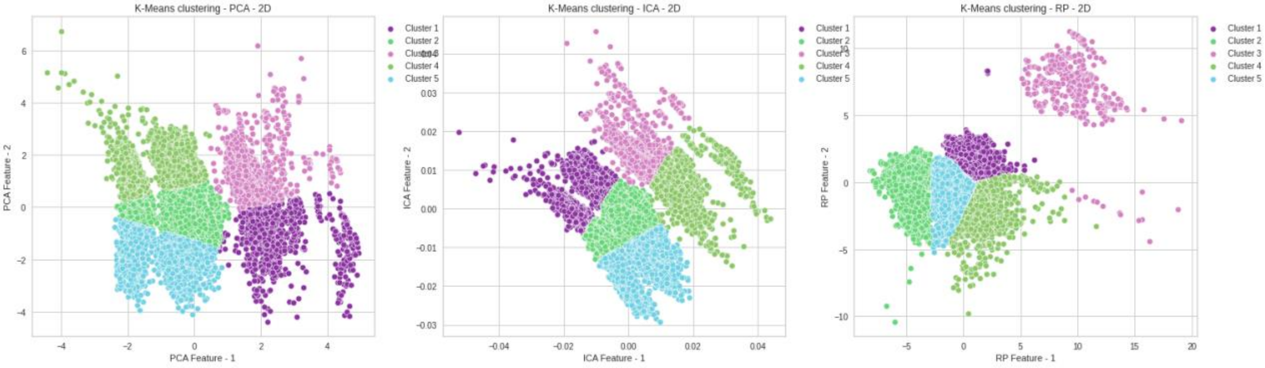
	selected_features	variance
0	Hour	0.179790
1	Temperature(°C)	0.160788
2	Humidity(%)	0.136899
3	Solar Radiation (MJ/m2)	0.104963
4	Rainfall(mm)	0.096520
5	Snowfall (cm)	0.078366
6	0	0.066406
7	1	0.058690
8	3	0.048622



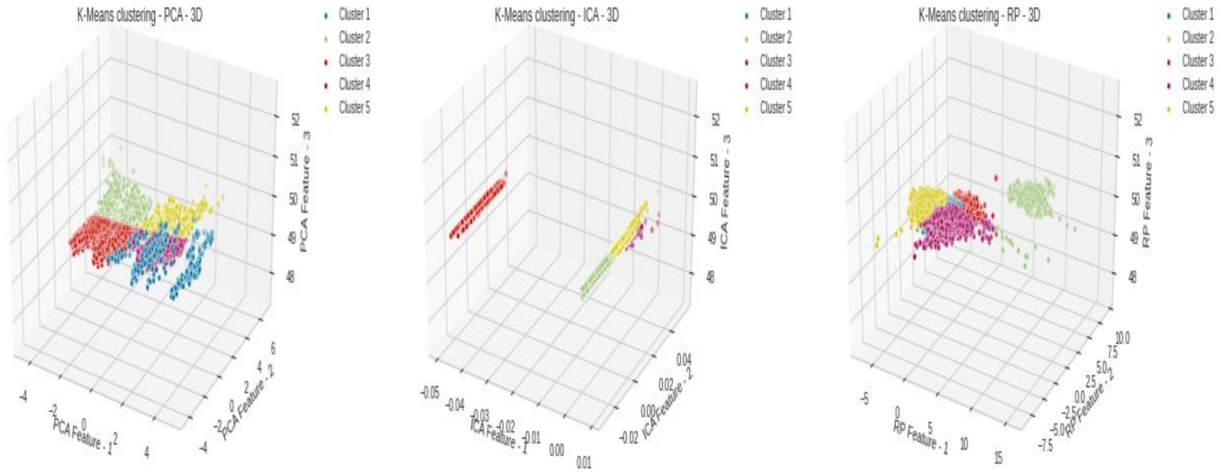
Another dimensionality reduction technique implemented was Independent Component Analysis which is a computational method for separating multi-variate signal into its underlying components. Random Projections was also implemented. It is very fast & robust to outliers. One of the big differences between PCA and ICA is, PCA compresses the information and takes into consideration only those variables which have high variance compared to all others. It doesn't consider if there is variable which has high collinearity with target variable and less variance. ICA aims to separate information based on maximally independent basis.

Random Projection is another method of dimensionality reduction and data visualization that simplifies the complexity of high-dimensional datasets. The method generates a new dataset by taking the projection of each data point along a randomly chosen set of directions. The projection of a single data point onto a vector is mathematically equivalent to taking the dot product of the point with the vector.

Task – 3: K-Means clustering algorithm was implemented using the 2 and 3 dimensions of PCA, ICA and RP. The scatter plots of the clusters –



We can see that there is no perfect boundary between these clusters which is shown by the low silhouette score. Also K-Means is sensitive to outliers. One cluster in RP has comparatively far away compared to other dimensionality reduction algorithms. The scatter plots when 3 features are considered in PCA, ICA, RP are –

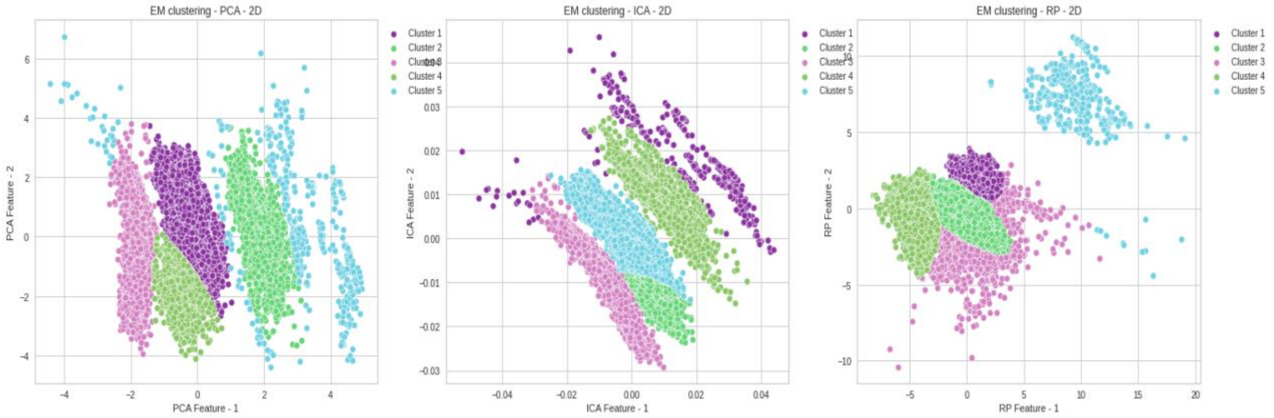


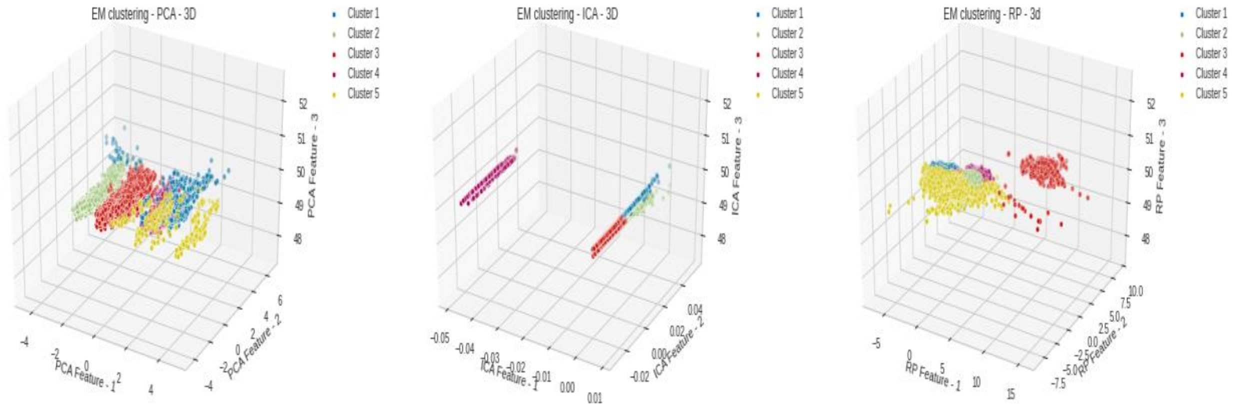
5 clusters are distributed along the 3 axes. The features generated by each dimensionality reduction algorithm is plotted on the 3 axes of the plot. In PCA and RP, we can see some overlap among the clusters. In ICA, as the algorithm states, the features are separated based on mutual

independence. That can be clearly seen in the clusters. The silhouette scores of the K-Means algorithms after dimensionality reduction is applied for 2 and 3 features are –
All the silhouette scores are less than the ones compared to K-Means when all the features are used to develop the algorithm i.e., Hour, Temperature and humidity variables are not sufficient for clustering. There is a tradeoff between features and accuracies. When the number of features is decreased there is definite decrease in accuracy. Hence though visualization is not possible, we should include more features in the model to get better scores.

PCA 2D	0.1195
ICA 2D	0.1157
RP 2D	0.1415
PCA 3D	0.1757
ICA 3D	0.1771
RP 3D	0.1814

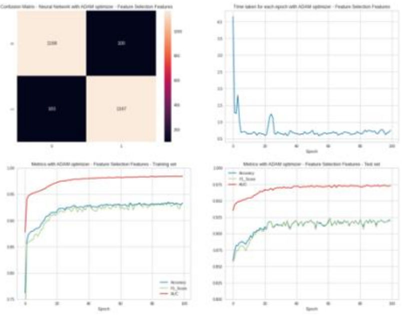
For Expectation Maximization algorithm, 2D, 3D graphs are –



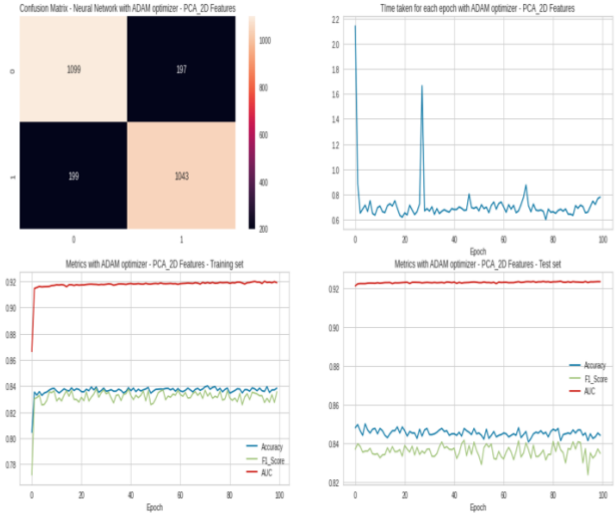


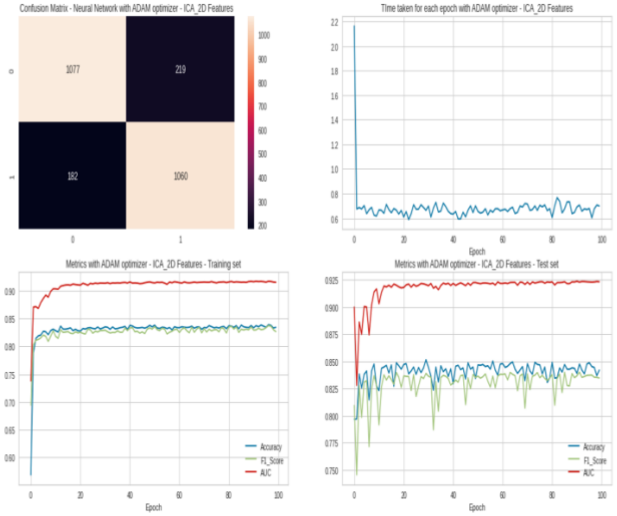
In expectation maximization algorithm is run with PCA of 2 dimensions cluster 5 is spread all over the scatter plot. The similar behavior is seen in an ICA plot with cluster 1. In the RP plot cluster 1, 2, 3, 4 are very tightly overlapped and there is very small difference between the clusters compared to PCA and ICA. The accuracies of EM clustering

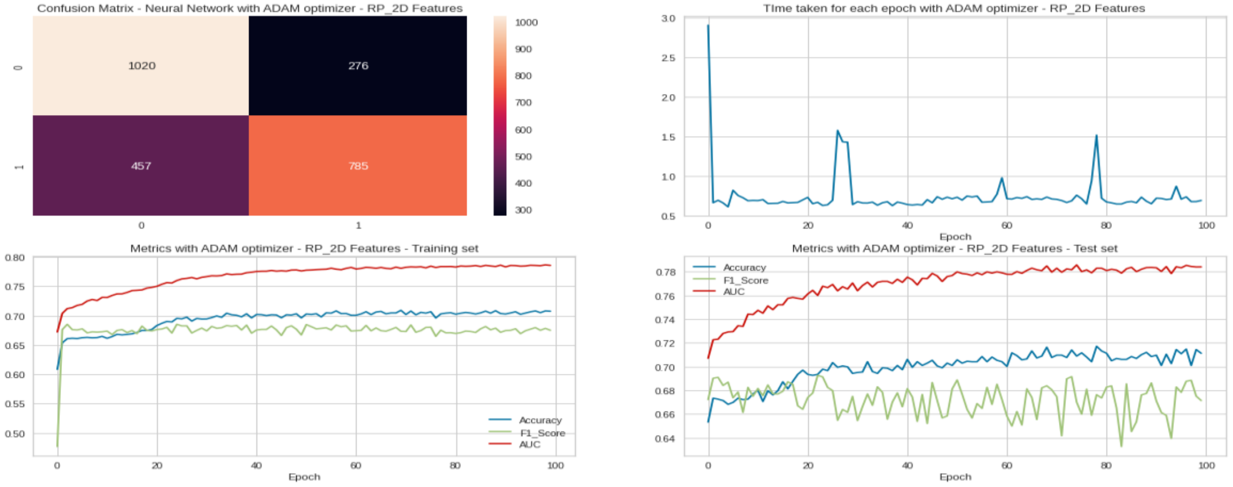
A similar trend of distribution of clusters is observed in the plots where dimensionality reduction algorithm has 3 features. In PCA the cluster 1 is spread all over whereas in ICA we can see the information separated.



Task – 4: When feature selection is done and neural network is developed on the selected features maximum accuracy is obtained – 92%. The time taken is almost same as that of the neural network developed in Assignment 3. But as the algorithm converges the time taken for each epoch is 0.5 sec compared to the 1 sec in assignment 3. The training and test accuracies reach 98% and 97% respectively for neural network developed with the results of feature selection. The performance is better when 9 features are selected though feature selection method compared to the one in the assignment 3. This is because almost all the information is retained and the variance in the data is explained by the 9 features. In the case of neural network with PCA with 2 dimensions the time taken varies from 0.6 to 2.2 seconds. The test and training accuracies almost reach 92%. In the case of neural network with ICA with 2 dimensions the time taken varies from 0.6 to 2.2 seconds. The test and training accuracies almost reach 92%. In the case of neural network with RP with 2 dimensions the time taken varies from 0.6 to 3 seconds. The test and training accuracies almost reach 78%.

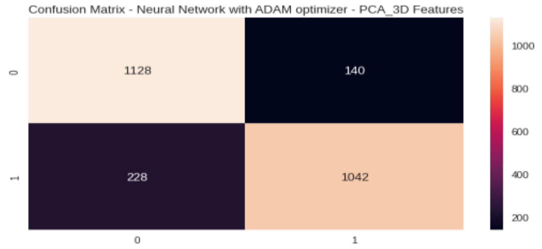


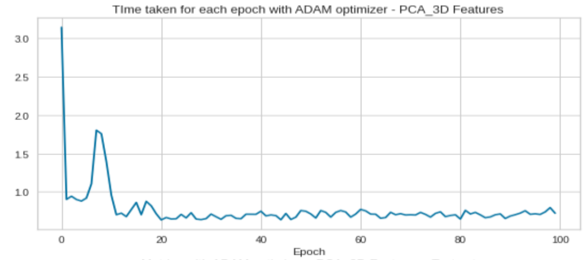


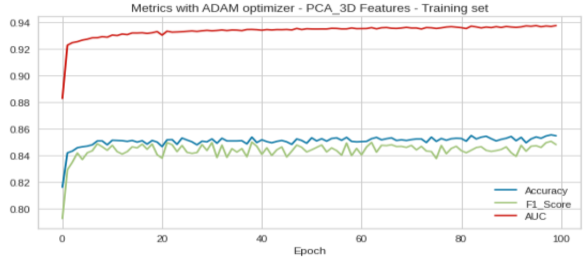


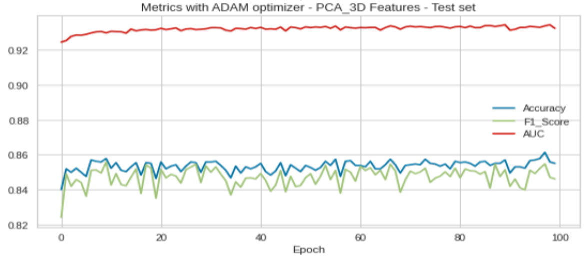
In the case of neural network with PCA with 3 dimensions the time taken varies from 0.6 to 2.2 seconds. The test and training accuracies almost reach around 93%. In the case of neural network with ICA with 3 dimensions the time taken varies from 0.5 to 2 seconds. The test and training accuracies almost reach

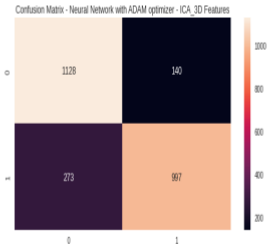
90%. In the case of neural network with RP with 3 dimensions the time taken varies from 0.6 to 2.2 seconds. The test and training accuracies almost reach 88%.

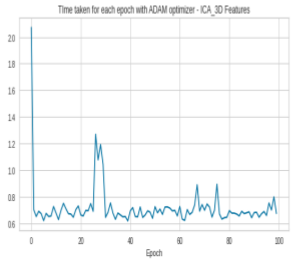


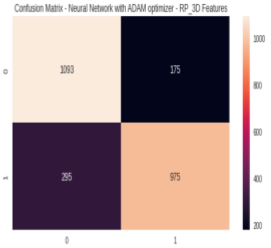


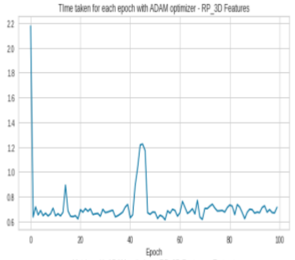


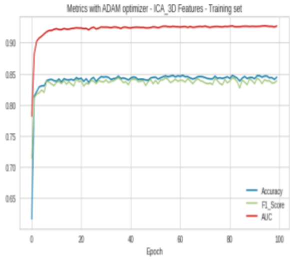


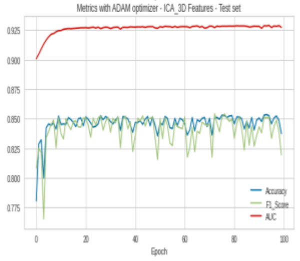


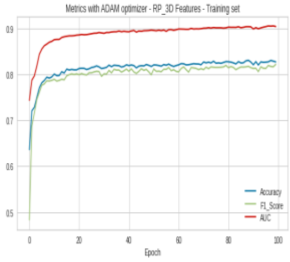


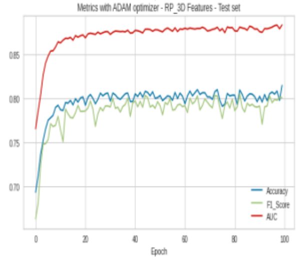




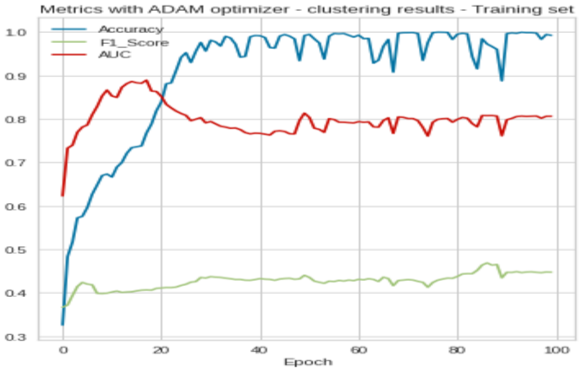


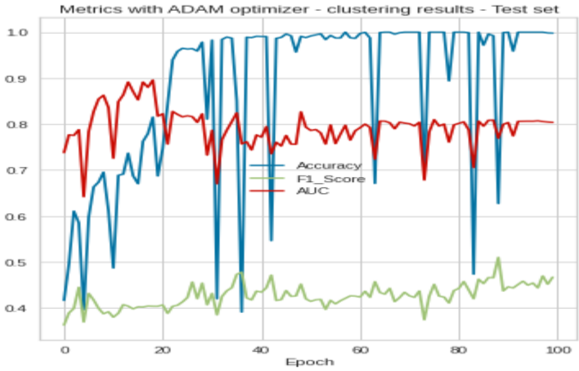






Task – 5: In the task 5, x is taken as features and class labels from clustering are taken as output variable and Neural network with multi class classification is developed with the same nodes and layers as that of Assignment 3. The plots of that model are –





The plot of test set shows the accuracy is varying a lot through the epochs. AUC is very low which shows the model isn't performing very well.

Conclusion

The neural network is the best model developed on the dataset. Through more experimentation on the activation functions and number of layers, we can get the best neural network which can have upto 97% accuracy without overfitting. The K-Means clustering methods had very less silhouette scores which show that the model was not clustered well. This can be better achieved by taking more information that can affect the bikes rented count. Instead of including only the weather conditions, information about the challenges the company might go through while making the bikes available can be included in the dataset. The losses and profits of the company affect the bike manufacturing capacity of the company.