Introduction

The given data set pertains to the historical demand for bike-sharing in Seoul. The data contains the count of rented bikes on every hour of a given day from December 2017 to November 2018. This data also provides different predictors that could explain the variance in demand of bikes for rent such as temperature, humidity, wind speed, Visibility, Dew point, Solar Radiation, Rainfall, snowfall, season, and whether the day is a holiday or functioning day. The objective is to predict how the bike renting going to be give the features or conditions such as climatic and date and season etc., so that proper planning can be done by the business teams.

Date pre-processing:

Basic primary preprocessing of data is done such as scaling and hot encoding to improve the data structure and eliminating correlated features to improve accuracy and performance.

Binary classification threshold:

We can treat this problem as binary classification problem by converting the output variable ie., count of bikes into high or low. As the data is right skewed, we can take the median of the count of bikes rented(504.5) as the threshold and consider all the observations with count of bikes more than median as high and observations with count of bikes less than median as low. High is encoded as 1 and low is encoded as 0. We can use clustering, decision tree algorithm(for feature selection) and ANN algorithms on binary classification easily as well.

Clustering

We can run the k means clustering algorithm with Euclidian distance measure soft clustering with expectation maximization to find meaningful separations in the data set.

To find the number of clusters, elbow plot. Elbow plot captures the within sum of squares of all the clusters with different K's. the optimum K would be number of K at which total sum of squares is least.

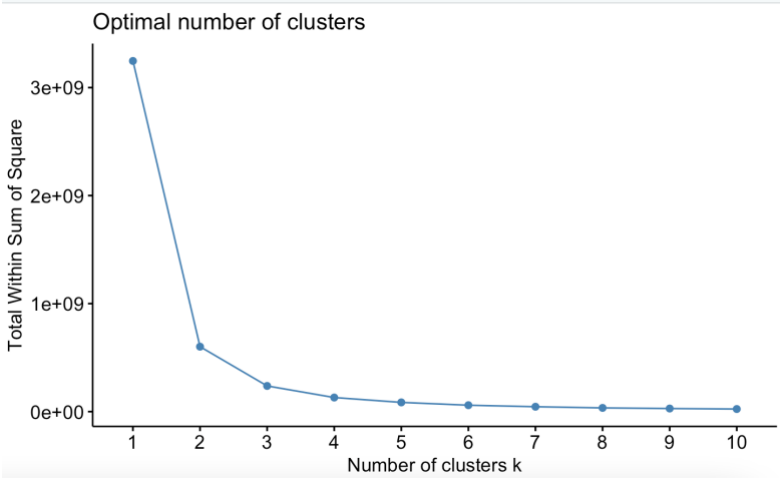

Fig 1. Elbow plot for optimum K

From the above figure, we can observe that 3 number of clusters are optimum for this dataset. After increasing the clusters from 3 to 4 the reduction in sum of squares is not that significant. So we can consider 3 as the optimum number of clusters for this dataset.

K Means clustering:

K means clustering algorithm is used from "stats" library by giving 2 as number of clusters. Below summary statistics and plots for the cluster labels helps to better understand the labels to know their characters. This will also help to predict and understand the characters of new data points.

```
  k.label Rented Bike Count     Hour Temperature(∞C) Humidity(%) Wind speed (m/s)
1       1         0.4493727 11.91659        12.96268    51.00230         1.836791
2       2         0.3157046 10.73976        12.73738    71.40922         1.520735
  Visibility (10m) Solar Radiation (MJ/m2) Rainfall(mm) Snowfall (cm) Seasons_Spring
1        1843.5324               0.6569518   0.03394593    0.04373564      0.2016257
2         694.6288               0.4088101   0.35807804    0.13224766      0.3440826
  Seasons_Summer Seasons_Winter Holiday_No Holiday Functioning Day_Yes
1      0.2627673      0.2551688          0.9459268           0.9641279
2      0.2325056      0.2308933          0.9593679           0.9703322
>
```

Fig 2. Summary statistics for K labels.

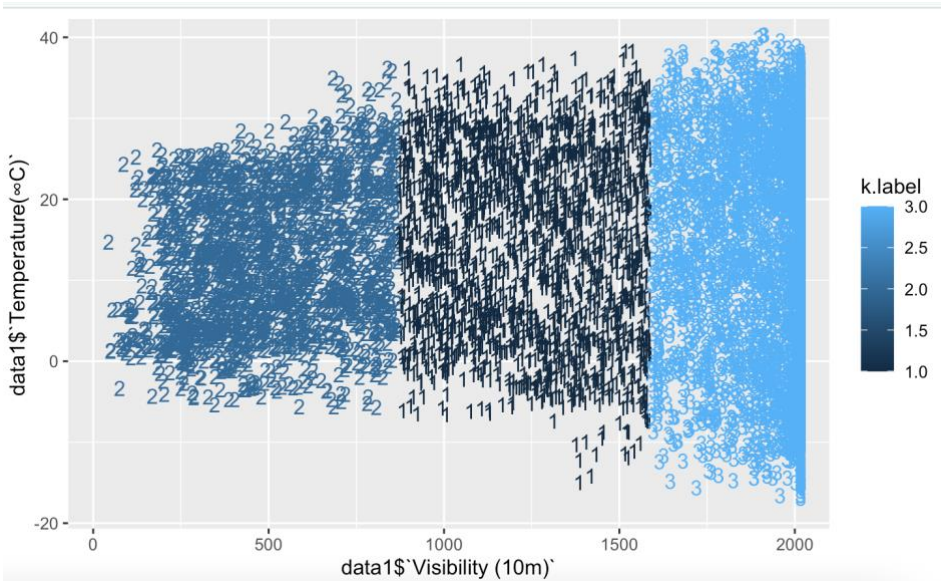

Fig 3. K means cluster plot.

From the above summary statistics and plot from fig 2 and fig 3, we can say that cluster 3 is having the observations with high visibility and cluster2 is having observations with low visibility and cluster 3 is having observations with medium visibility with respect to different temperatures.

Expectation Maximization:

We can use soft clustering technique with expectation maximization algorithm to cluster the dataset

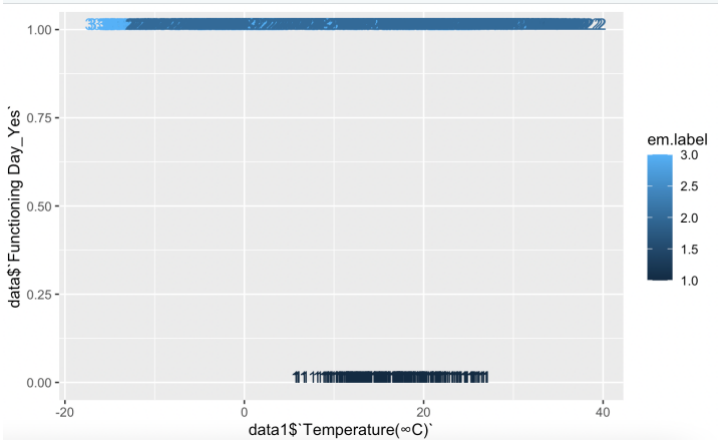

Fig 4. EM cluster plot.

From the above fig.4 we can see that EM clustering is not as clear and specific as K means since we see some overlapping with cluster 2 and 3 when temperature is between 0 to 20 and it's a functioning day, the data point can belong to either cluster 2 or cluster 3.

Dimension reduction

Different types of dimension reduction techniques are used such as decision tree, Principal component analysis(PCA), Independent component analysis(ICA) and random projection on the dataset to identify the most important features that captures the structure of the data and play most important role in determining the true hypothesis.

Decision tree:

Decision tree algorithm can be used as a feature selection technique. As the top nodes of the decision tree are most important in determining the hypothesis, we can consider those features are important.

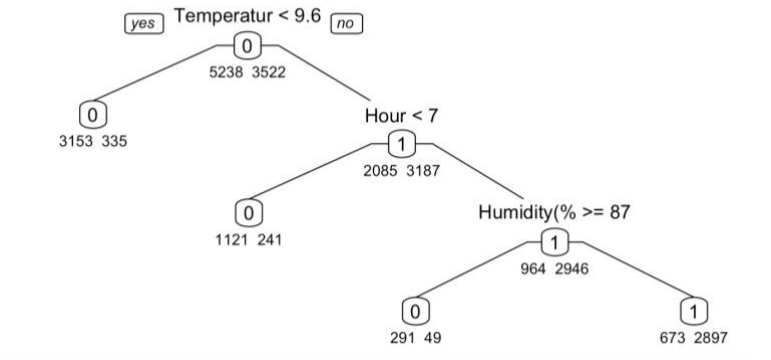

Fig 5. Simple decision tree.

A simple decision tree is drawn on the dataset with minsplit = 1000. The fig.5 shows that temperature, Hour and Humidity features are most important features in the dataset. We can consider this features for the reduced data set for applying further machine leaning algorithms.

Principal component analysis(PCA):

PCA linearly transforms the dataset into different dataset that have same features as the original one. The component that captures the most variance in data set is the first component and so on.

```
Importance of components:
                            PC1      PC2     PC3   PC4   PC5    PC6    PC7    PC8
Standard deviation     608.2993 11.99591 6.82119 1.118 1.053 0.6991 0.5053 0.4111
Proportion of Variance   0.9995  0.00039 0.00013 0.000 0.000 0.0000 0.0000 0.0000
Cumulative Proportion    0.9995  0.99986 0.99999 1.000 1.000 1.0000 1.0000 1.0000
                           PC9   PC10   PC11   PC12   PC13
Standard deviation     0.3772 0.2674 0.2275 0.2076 0.1579
Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000
Cumulative Proportion  1.0000 1.0000 1.0000 1.0000 1.0000
```

Fig 6. Principal component analysis.

The above figure shows the individual components of the PCA output. We can see that PCA1 captures 99.95% of variance in the data set and cumulative variance of PCA1 and PCA2 is 99.86%. Since only these two features captures almost 100% of the variance in the data, we can consider these two components as features for further machine learning algorithms.

Independent component analysis(ICA):

ICA is similar to PCA but instead of reducing the corelating in among the components, ICA transforms the features into new features that are mutually independent. After transforming the features into new feature set, to reduce the dimensions, decision tree algorithm is used to select the most important features that determine the true hypothesis.

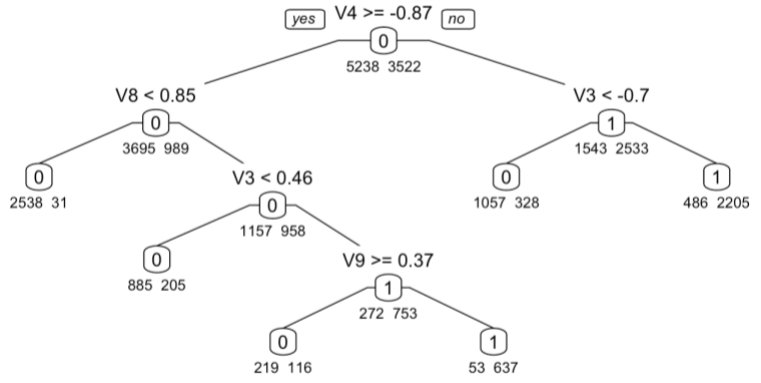

Fig 7. Decision tree on ICA output

From the above figure we can observe that V4, V8, V3 features are selected by decision tree as most important for determining the hypothesis. Dimension can be reduced for the data set by considering these features for further machine learning models.

Random Projections:

Random component analysis projects the dataset on random directions and generates new set of features that can capture the correlation between the features. After creating new feature set with random directions, Decision tree is used as dimension reduction technique to select the important features.

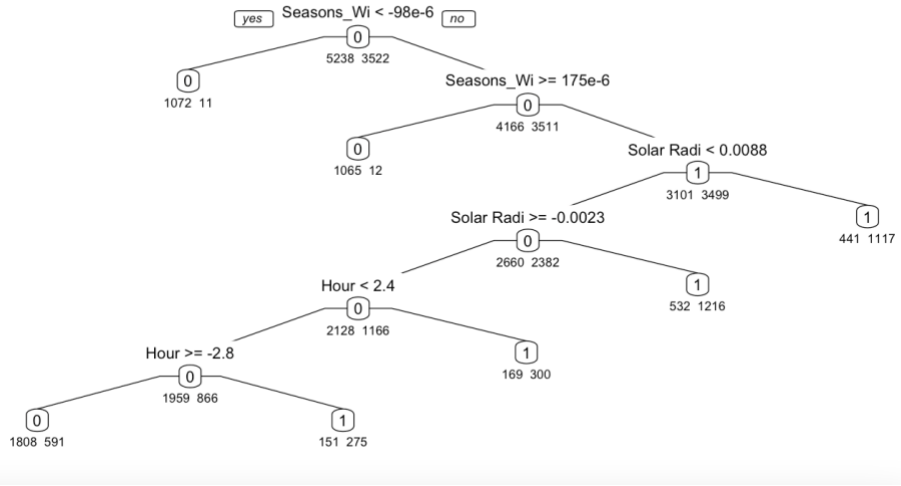

Fig 8. Decision tree on Random projections output

Above decision tree is drawn after projecting the data set in random directions to reduce the features for dimension reduction. Apparently Seasons_winter, Solar radiation, and hour are important features in the new dataset.

Dimension reduction and clustering

Dimension reduction & K means:

K means clustering is re used on the new dataset with reduced dimensions.

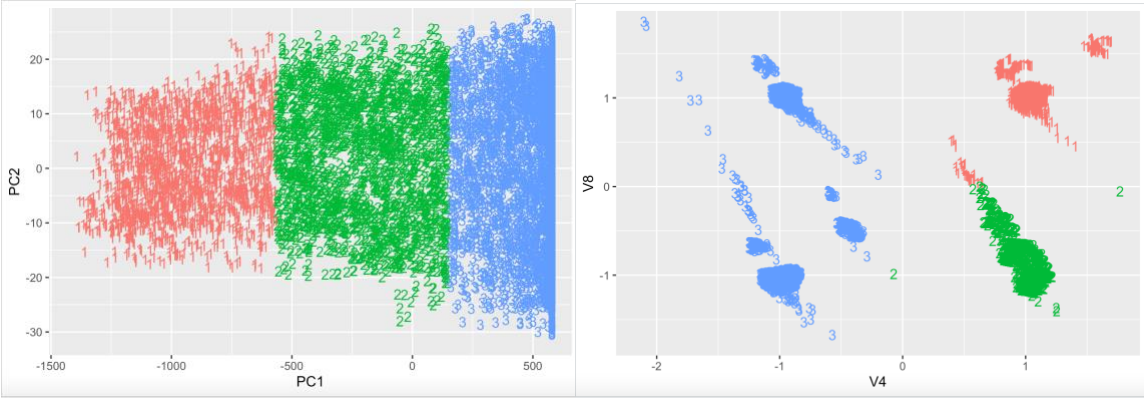

Fig 9. Kmeans on PCA output(left) and on ICA output(Right)

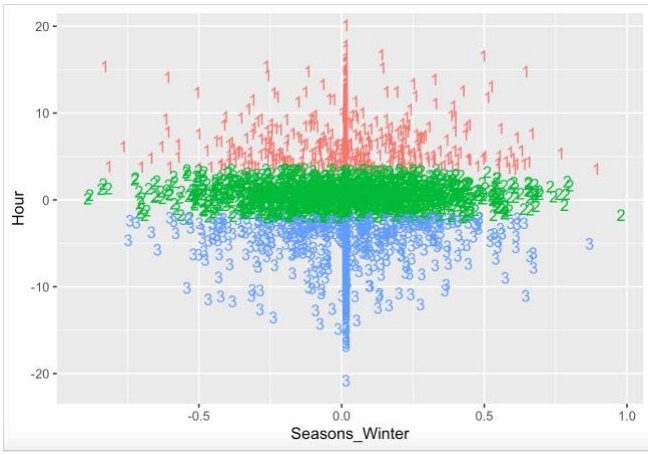

Fig 10. Clustering on output of randomized projections

Above two figures.9,10 shows the cluster results after applying the above dimension reduction and feature transformation techniques. Only the important features that are selected by decision tree algorithm are used for clustering. It is interesting to note that clusters formed on the original dataset and clusters formed on the PCA output are almost same. This tells that PCA works best on this data set to reduce the dimensions. Only 2 components/ features are able to capture the variance in all the features and gave same output as original data.

Dimension reduction & Expectation Maximization:

Expectation maximization clustering Is used in on the above reduced feature datasets.

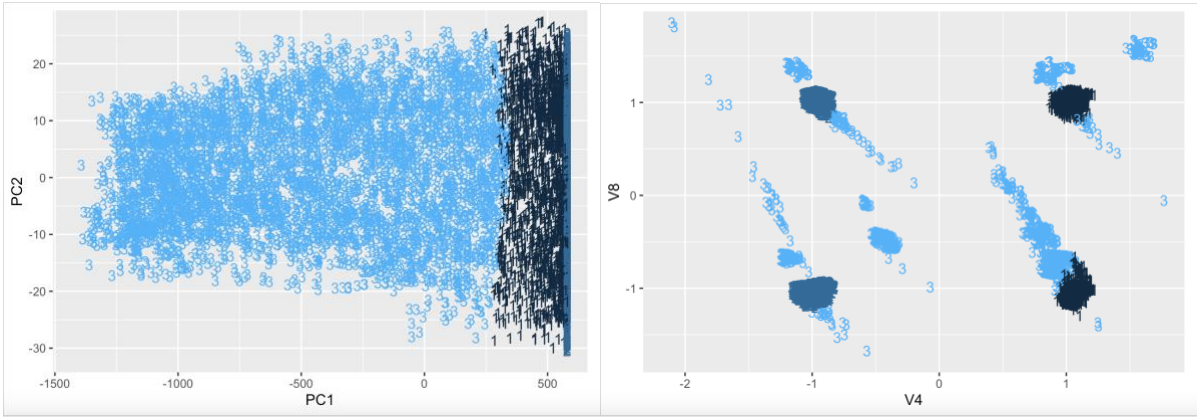

Fig 11. EM Clustering on PCA output(left) and on ICA output(Right)

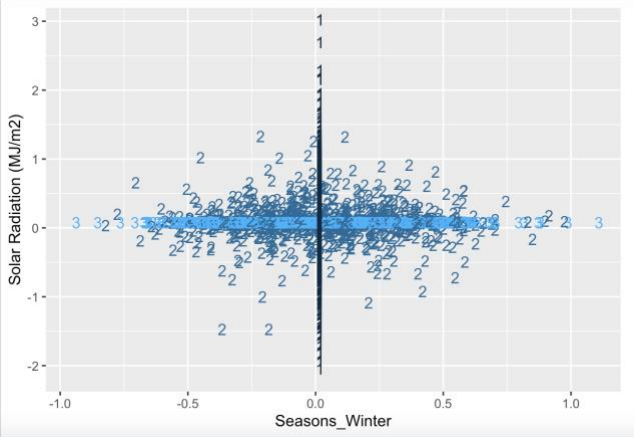

Fig 12. EM Clustering on random projections.

Above two figures.9,10 shows the cluster results after applying the above dimension reduction and feature transformation techniques. Only the important features that are selected by decision tree algorithm are used for clustering.

It is interesting to note that PCA cluster results with K means and EM are very similar and showing the similar cluster results as the original data set. Hence we can conclude that PCA dimension reduction worked best on this data set as 2 components/features are representing the variation in the entire dataset. ICA and RP clusters are similar to K means clustering but not as distinguished as K means. For example, EM clustering with RP have over laps in the center where as there are no overlapping for k means clustering

Dimension reduction and Neural Network

Neural network algorithm is used to predict the high or low bike rentals on the reduced feature sets after above dimension reduction algorithms.

	ANN on PCA output		ANN on ICA output		ANN on Random proj output	
	Train	Test	Train	Test	Train	Test
Accuracy :	0.8301	0.8314	0.8578	0.855	0.7764	0.7725
95% CI :	(0.8204, 0.8394)	(0.8166, 0.8456)	(0.8488, 0.8664)	(0.841, 0.8683)	(0.7658, 0.7868)	(0.7559, 0.7884)
No Information Rate :	0.5993	0.5947	0.6024	0.5875	0.5993	0.5947
P-Value [Acc > NIR] :	<2e-16	< 2e-16	<2e-16	<2e-16	< 2.2e-16	< 2.2e-16
Kappa :	0.645	0.6477	0.7029	0.6998	0.5422	0.5331
Mcnemar's Test P-Value :	0.129	0.02258	0.7095	0.1829	1.46e-11	0.002834
Sensitivity :	0.8650	0.8740	0.8836	0.8854	0.7793	0.7850
Specificity :	0.7778	0.7690	0.8187	0.8118	0.7721	0.7540
Pos Pred Value :	0.8534	0.8474	0.8807	0.8701	0.8364	0.8240
Neg Pred Value :	0.7939	0.8061	0.8228	0.8325	0.7005	0.7050
Prevalence :	0.5993	0.5947	0.6024	0.5875	0.5993	0.5947
Detection Rate :	0.5184	0.5198	0.5323	0.5202	0.4671	0.4669
Detection Prevalence :	0.6075	0.6134	0.6044	0.5978	0.5584	0.5666
Balanced Accuracy :	0.8214	0.8215	0.8511	0.8486	0.7757	0.7695
'Positive' Class :	0	0	0	0	0	0

Fig 13. Train and test accuracies of neural network on dimension reduction datasets.

From the above table we can see that PCA and ICA outputs have better accuracies than random projections output. PCA did very well with K means clustering, EM clustering and supervised machine learning(ANN). ICA did well with K means clustering and supervised machine learning. Random projections did well In K means clustering but not so well with EM clustering and ANN.

It is also important to note that the time taken to run the reduced data sets is significantly lower that running on the complete data set. Also it took least time of around 25 seconds to run the ANN on PCA output where it took over 45 seconds when run on ICA and random projections.

Artificial Neural networks on Cluster labels as features

The class labels generated by K means and EM clustering are taken as new features and ANN model is implemented to predict the Count of bike rentals.

	Train	Test
Accuracy	0.5993	0.5947
95% CI	(0.5869, 0.6116)	(0.5757, 0.6136)
No Information Rate	0.5993	0.5947
P-Value [Acc > NIR]	0.5055	0.5084
Kappa	0	0
Mcnemar's Test P-Value	<2e-16	<2e-16
Sensitivity	1.0000	1.0000
Specificity	0.0000	0.0000
Pos Pred Value	0.5993	0.5947
Neg Pred Value	NaN	NaN
Prevalence	0.5993	0.5947
Detection Rate	0.5993	0.5947
Detection Prevalence	1.0000	1.0000
Balanced Accuracy	0.5000	0.5000
'Positive' Class	0	0

Fig 14. ANN model on cluster labels.

We can use stacking ensemble method to take the outputs of other learners and predict the output using different learner. Apparently, the accuracy of the model is low compared to learner on the reduced data set. This could be due to using too less features.