BANA 7047 – Prof. Yan Yu

Individual Case II

Last Name: BADRE

First Name: SHASHANK

UCID: M12383328

European Employment Data

Summary

European Employment Data shows the percentage employed in different industries in Europe countries during 1979. In European Employment Data set there are 26 observations and 10 variables. The purpose of examining this data is to get insight into patterns of employment amongst European countries in 1970s.

I have sampled a European Employment data set that contains 90% of original data. The dimensions for the sampled data set is 23 observations and 10 variables. I have scaled all variables except Country and for country variable I have created dummy variables by using MLR library in R.

Using Average silhouette width and Elbow method I found that K=3 is best to apply Kmeans clustering. After applying K=3 for K means clustering I have got 3 clusters and summary statistics for variables belonging to each of these clusters can be seen below.

cluster	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	25.77	2.07	28.90	0.93	8.30	8.51	1.13	17.13	7.29
2	12.23	0.89	26.86	0.94	8.65	16.04	5.11	22.70	6.61
3	57.75	1.10	12.35	0.60	3.85	5.80	6.20	8.60	3.60

Table: Summary statistics for K=3 Kmeans clustering

I have also done hierarchical clustering on this data set. The distance method used is “ward” to calculate the distance between records in the data set for Hierarchical clustering. I have cut dendrogram in such a way that we get 3 clusters. The summary statistics for variables belonging to each of these clusters can be seen below.

cluster	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1.00	14.17	0.87	26.24	0.91	8.61	15.74	4.93	21.92	6.61
2.00	23.17	2.32	30.78	0.98	8.33	8.02	0.92	18.15	7.38
3.00	57.75	1.10	12.35	0.60	3.85	5.80	6.20	8.60	3.60

Table: Summary statistics for Hierarchical clustering

Cincinnati Zoo Data

Summary

Data set in excel sheet “qry_Food_by_Month.xls” is taken from Cincinnati Zoo data. It consists of 55 observations and 7 columns. I have used clustering technique to get insight into patterns in this data set. Using Average silhouette width and Elbow method I found that K=3 is best to apply for Kmeans clustering. After applying K=3 for K means clustering I have got 2 clusters and summary statistics for variables belonging to each of these clusters can be seen below.

cluster	Oct, 10	Nov, 10	Dec, 10	Jan, 11	Feb, 11	Mar, 11
1	158.2558	75.97674	110.5581	17.53488	28.7907	67.25581
2	1133.4167	398.66667	466.3333	70.08333	146.8333	449.0833

I have also done hierarchical clustering on this data set. The distance method used is “ward” to calculate the distance between records in the data set for Hierarchical clustering. I have cut dendrogram in such a way that we get 2 clusters. The summary statistics for variables belonging to each of these clusters can be seen below.

cluster	Oct..10	Nov..10	Dec..10	Jan..11	Feb..11	Mar..11
1	101.5294	43.23529	49.35294	8.529412	14.73529	40.26471
2	807.3333	313.38095	412.95238	62.142857	119	329.14286

Association:

Food_4_association data set contains 19076 observations and 118 variables. We will use this data set to understand association i.e. for market basket analysis. This analysis will allow us to know which items are frequently brought together.

```
    lhs                                rhs                              support confidence     lift count
[1] {Small.Pink.LemonadeFood}       => {Chicken.Nugget.BasketFood}  0.003355001  0.5925926 16.03446    64
[2] {Side.of.CheeseFood}            => {Cheese.ConeyFood}           0.004665548  0.6846154 25.91215    89
[3] {Side.of.CheeseFood}            => {Hot.DogFood}                0.006290627  0.9230769 21.60566   120
[4] {Hot.Chocolate.Souvenir.RefillFood} => {Hot.Chocolate.SouvenirFood} 0.014992661  0.5596869 13.18097   286
[5] {Cheese.ConeyFood,
     Side.of.CheeseFood}            => {Hot.DogFood}                0.004351017  0.9325843 21.82819    83
[6] {Hot.DogFood,
     Side.of.CheeseFood}            => {Cheese.ConeyFood}           0.004351017  0.6916667 26.17903    83
```

- Above rules shows items sets wherein lift is greater than 10
- More the lift then more is the strong association between items

Classification using SPSS Modeler

Summary

German credit score data has records that are classified as good credit and bad credit. We will use this data set to train our model to predict whether a test record is a good credit or bad credit. German credit score data set consists of 1000 observations and 21 features. Variable “Response” is the dependent variable in this data set. Response variable 0 indicate “good credit risk” and 1 indicates “bad credit risk”.

German credit score data set was split randomly in 80 % train data set and 20 % test data set. Predictive models were built using train data set and then measures were noted for IN sample data set. I have predicted values for test data set and then measured Out of sample measures. Following table gives the result obtained from Logistic Regression method, Classification Tree method, General Additive Model method, Neural Network and Linear Discriminant Analysis method.

Method	Misclassification rate		Area under Curve	
	In Sample	Out of Sample	In Sample	Out of Sample
Classification Tree	0.42	0.44	0.694	0.688
Logistic Regression	0.21	0.28	0.836	0.799
Neural Network	0.23	0.29	0.813	0.753

Best model for predicting good credit and bad credit for German Credit Score data is found to be Logistic Regression Model. Final LDA model has misclassification rate of 0.21 for In sample data and 0.28 for Out of sample data, which is lowest as compared to all other models.

1. European Employment Data

European Employment Data shows the percentage employed in different industries in Europe countries during 1979. In European Employment Data set there are 26 observations and 10 variables. The structure for this data set is given below.

- Country: Name of country. This variable is factor data type.
- Agr: Percentage employed in agriculture. This variable is numeric in data type.
- Min: Percentage employed in mining. This variable is numeric in data type.
- Man: Percentage employed in manufacturing. This variable is numeric in data type.
- PS: Percentage employed in power supply industries. This variable is numeric in data type.
- Con: Percentage employed in construction. This variable is numeric in data type.
- SI: Percentage employed in service industries. This variable is numeric in data type.
- Fin: Percentage employed in finance. This variable is numeric in data type.
- SPS: Percentage employed in social and personal services. This variable is numeric in data type.
- TC: Percentage employed in transport and communications. This variable is numeric in data type.

Now let us look at the summary of the data set.

Country	Agr	Min	Man
Austria : 1	Min. : 2.70	Min. :0.100	Min. : 7.90
Belgium : 1	1st Qu.: 7.70	1st Qu.:0.525	1st Qu.:23.00
Bulgaria : 1	Median :14.45	Median :0.950	Median :27.55
Czechoslovakia: 1	Mean :19.13	Mean :1.254	Mean :27.01
Denmark : 1	3rd Qu.:23.68	3rd Qu.:1.800	3rd Qu.:30.20
EGermany : 1	Max. :66.80	Max. :3.100	Max. :41.20
(Other) :20			

PS	Con	SI	Fin	SPS
Min. :0.1000	Min. : 2.800	Min. : 5.20	Min. : 0.500	Min. : 5.30
1st Qu.:0.6000	1st Qu.: 7.525	1st Qu.: 9.25	1st Qu.: 1.225	1st Qu.:16.25
Median :0.8500	Median : 8.350	Median :14.40	Median : 4.650	Median :19.65
Mean :0.9077	Mean : 8.165	Mean :12.96	Mean : 4.000	Mean :20.02
3rd Qu.:1.1750	3rd Qu.: 8.975	3rd Qu.:16.88	3rd Qu.: 5.925	3rd Qu.:24.12
Max. :1.9000	Max. :11.500	Max. :19.10	Max. :11.300	Max. :32.40

TC
Min. :3.200
1st Qu.:5.700
Median :6.700
Mean :6.546
3rd Qu.:7.075
Max. :9.400

Now, I have sampled a European Employment data set that contains 90% of original data. The dimensions for the sampled data set is 23 observations and 10 variables. I have scaled all variables except Country and for country variable I have created dummy variables by using MLR library in R.

I have started K means clustering by specifying initially 5 clusters. The number of observations in each cluster is given below.

For K = 5:

Cluster Number	1	2	3	4	5
Number of records	5	5	6	2	5

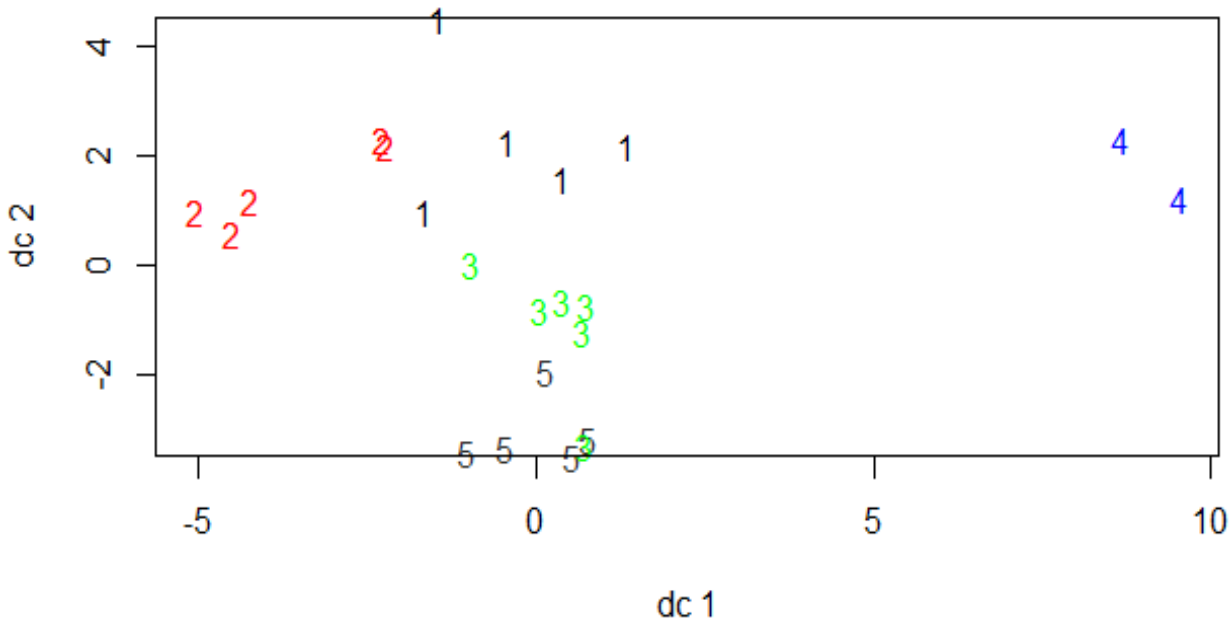


Figure: Clusters for k = 5

To understand the best possible K for clustering, I have applied elbow method and silhouette analysis. It has been found from these plots that best possible K for clustering is K =2.

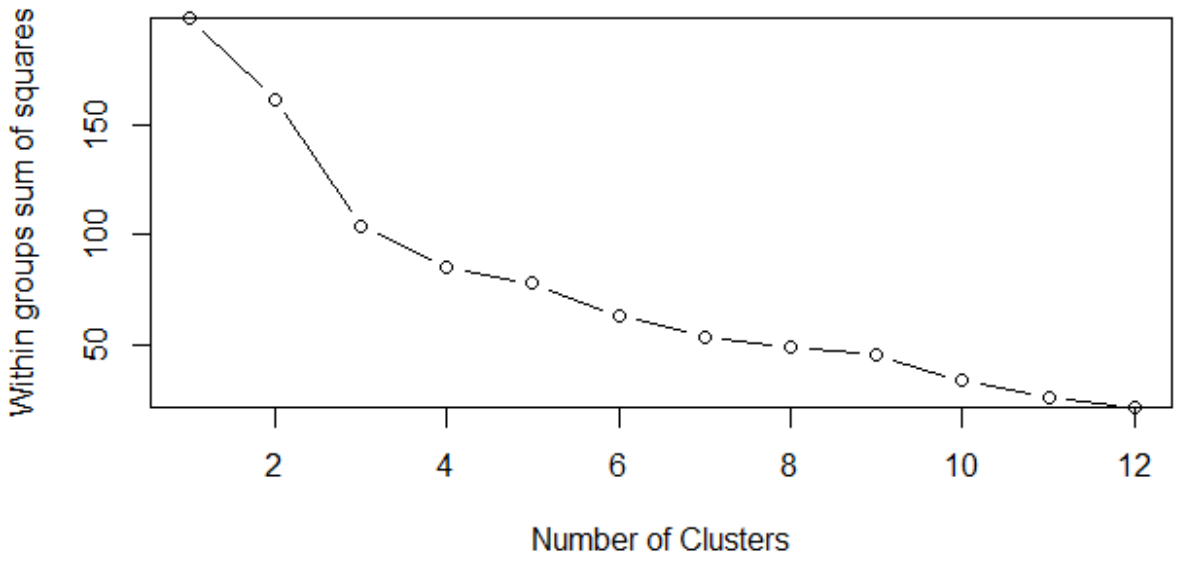


Fig: Elbow Method on European Employment Data set

Fig: Silhoutte Analysis on European Employment Data set

As silhouette width is highest for K =3. Hence we use K = 3 for our final analysis.

Cluster Number	1	2	3
Number of records	7	14	2

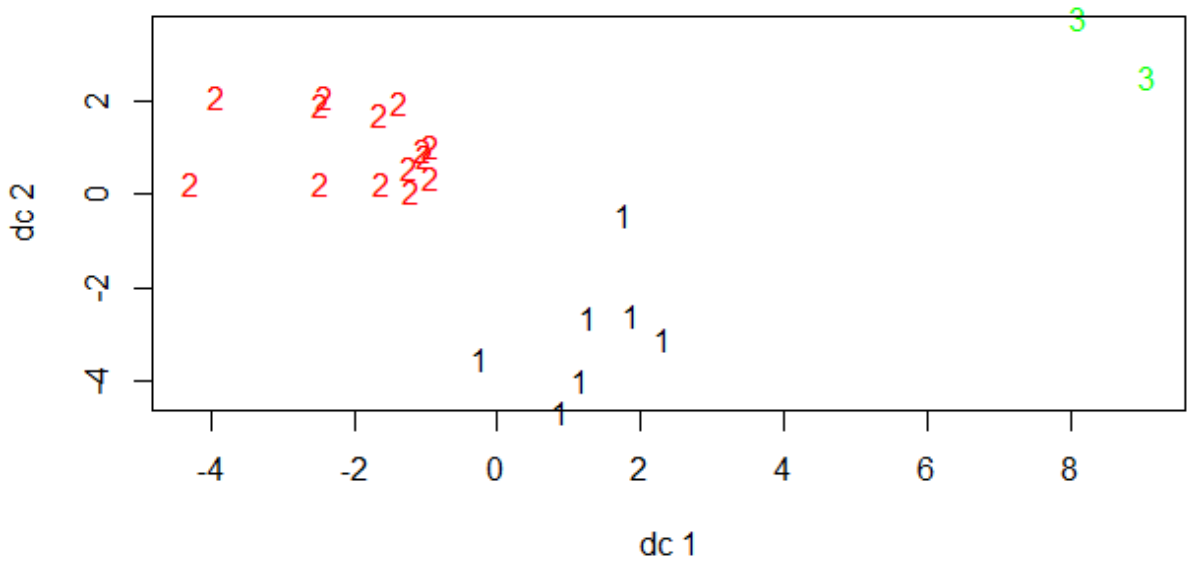


Fig: K=3 Kmeans clustering

The summary statistics for K=3 Kmeans clustering is given below.

cluster	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	25.77	2.07	28.90	0.93	8.30	8.51	1.13	17.13	7.29
2	12.23	0.89	26.86	0.94	8.65	16.04	5.11	22.70	6.61
3	57.75	1.10	12.35	0.60	3.85	5.80	6.20	8.60	3.60

Hierarchical Clustering on European Employment Data set

The distance method used is “ward” to calculate the distance between records in the data set for Hierarchical clustering.

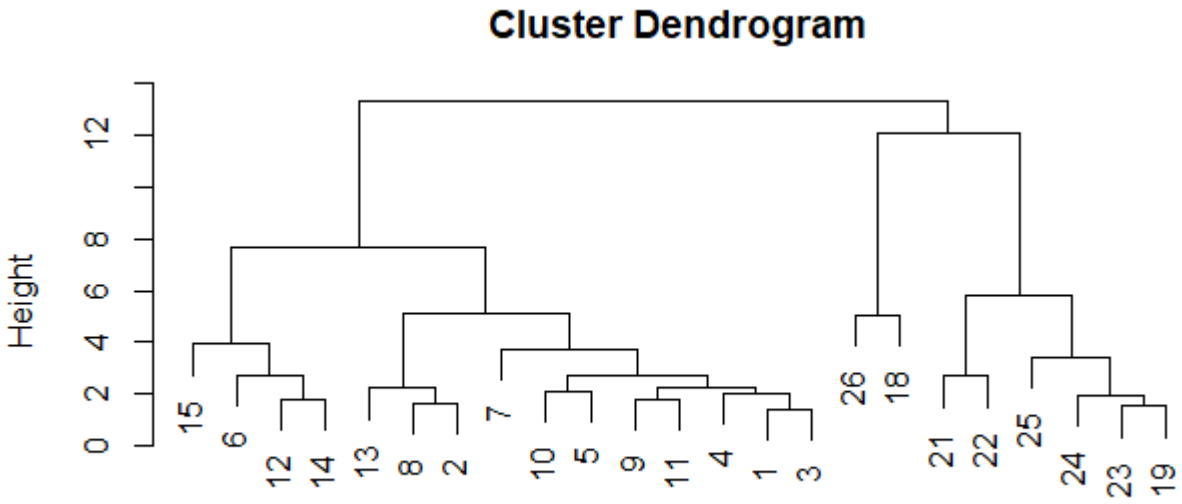


Fig: Dendrogram using ward method.

We cut dendrogram in such a way that we get 3 clusters. The number of records for each cluster is given below.

Cluster Number	1	2	3
Number of records	15	6	2

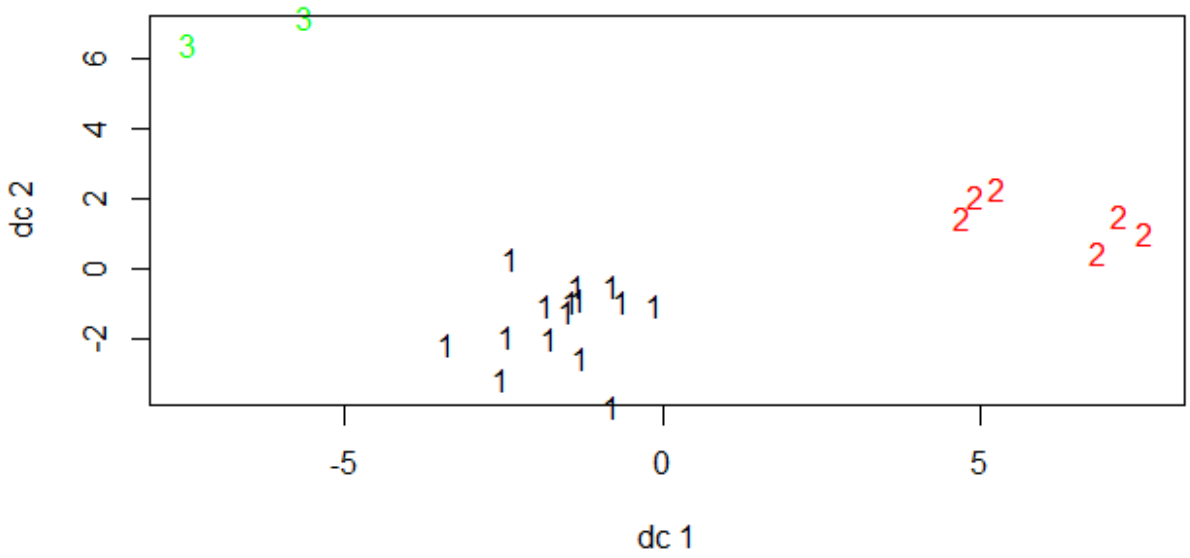


Fig: Clusters obtained using Hierarchical clustering

Summary statistics for 3 clusters are given below.

cluster	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1.00	14.17	0.87	26.24	0.91	8.61	15.74	4.93	21.92	6.61
2.00	23.17	2.32	30.78	0.98	8.33	8.02	0.92	18.15	7.38
3.00	57.75	1.10	12.35	0.60	3.85	5.80	6.20	8.60	3.60

2. Cincinnati Zoo Data

Data set in excel sheet “qry_Food_by_Month.xls” is taken from Cincinnati Zoo data. It consists of 55 observations and 7 columns. The structure for this data set is given below.

- NickName: This variable is of character data type
- Oct, 10 : This variable is of numeric data type
- Nov, 10 : This variable is of numeric data type
- Dec, 10 : This variable is of numeric data type
- Jan, 11 : This variable is of numeric data type
- Feb, 11 : This variable is of numeric data type
- Mar, 11 : This variable is of numeric data type

This data set does not need scaling as the numeric variables have the same scale.

Let us look at the summary for this data set.

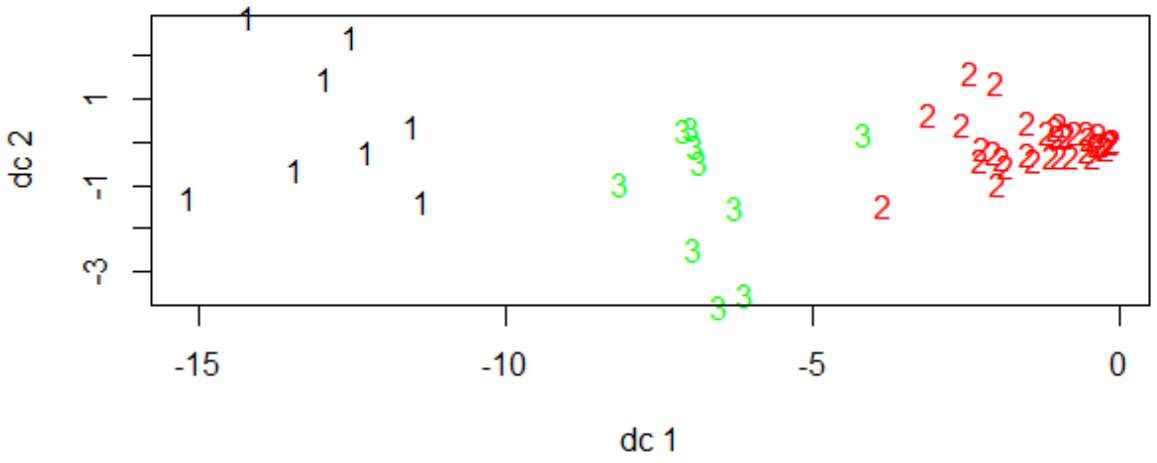
```
  NickName            Oct, 10        Nov, 10        Dec, 10
Length:55          Min.   :   0   Min.   :  2.0   Min.   :   0.0
Class :character   1st Qu.:  39   1st Qu.: 26.0   1st Qu.:   5.5
Mode  :character   Median : 154   Median : 66.0   Median :  56.0
                   Mean   : 371   Mean   :146.4   Mean   : 188.2
                   3rd Qu.: 524   3rd Qu.:265.5   3rd Qu.: 275.0
                   Max.   :2002   Max.   :597.0   Max.   :1089.0
   Jan, 11          Feb, 11          Mar, 11
Min.   :  0.0   Min.   :  0.00   Min.   :  0.0
1st Qu.:  0.0   1st Qu.:  0.00   1st Qu.: 15.0
Median :  8.0   Median : 11.00   Median : 48.0
Mean   : 29.0   Mean   : 54.55   Mean   :150.6
3rd Qu.: 50.5   3rd Qu.: 93.50   3rd Qu.:232.5
Max.   :186.0   Max.   :279.00   Max.   :785.0
```

K-means Clustering

I have applied K means clustering on this data set by specifying K=3 initially.

Cluster Number	1	2	3
Number of records	8	37	10

Table: Number observations for each cluster



Thus, above table shows that there are 8 records in cluster in 1, 37 records in cluster number 2 and 10 records in cluster number 3.

I have used elbow method and silhouette analysis to understand the optimal length of K for clustering.

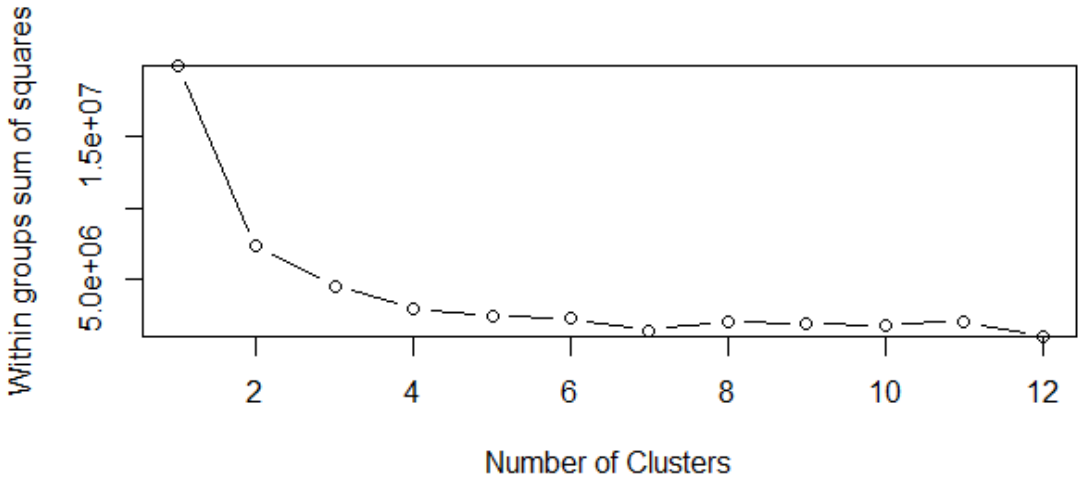


Fig: Elbow method

Fig: Silhouette Analysis

From elbow method and silhouette analysis, we can conclude the optimal K for clustering is 2.

I have used K=2 for K means clustering and the results of clusters can be seen below.

Cluster Number	1	2
Number of records	43	12

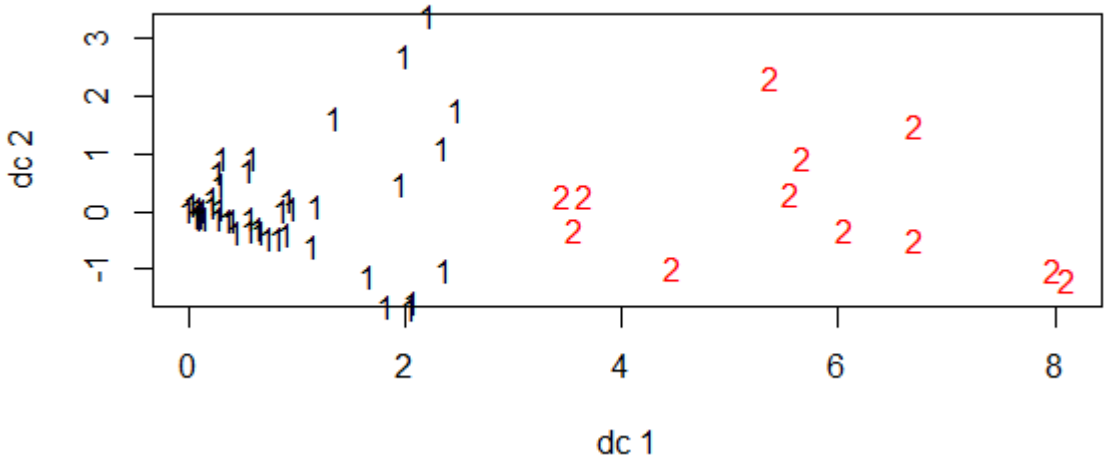


Fig: K=2 Kmeans clustering

The summary statistics for K=2 Kmeans clustering is given below.

cluster	Oct, 10	Nov, 10	Dec, 10	Jan, 11	Feb, 11	Mar, 11
1	158.2558	75.97674	110.5581	17.53488	28.7907	67.25581
2	1133.4167	398.66667	466.3333	70.08333	146.8333	449.0833

Hierarchical Clustering

The distance method used is “ward” to calculate the distance between records in the data set for Hierarchical clustering.

We cut dendrogram in such a way that we get 2 clusters. The number of records for each cluster is given below.

Cluster Number	1	2
Number of records	34	21

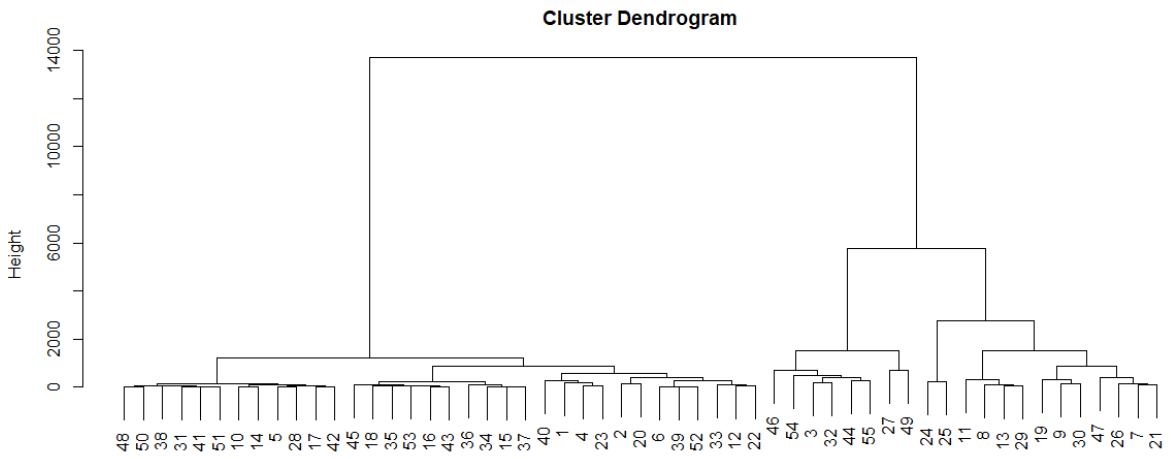


Fig: Dendrogram

The summary statistics for Hierarchical clustering is given below.

cluster	Oct..10	Nov..10	Dec..10	Jan..11	Feb..11	Mar..11
1	101.5294	43.23529	49.35294	8.529412	14.73529	40.26471
2	807.3333	313.38095	412.95238	62.142857	119	329.14286

Association:

Food_4_association data set contains 19076 observations and 118 variables. We will use this data set to understand association i.e. for market basket analysis. This analysis will allow us to know which items are frequently brought together.

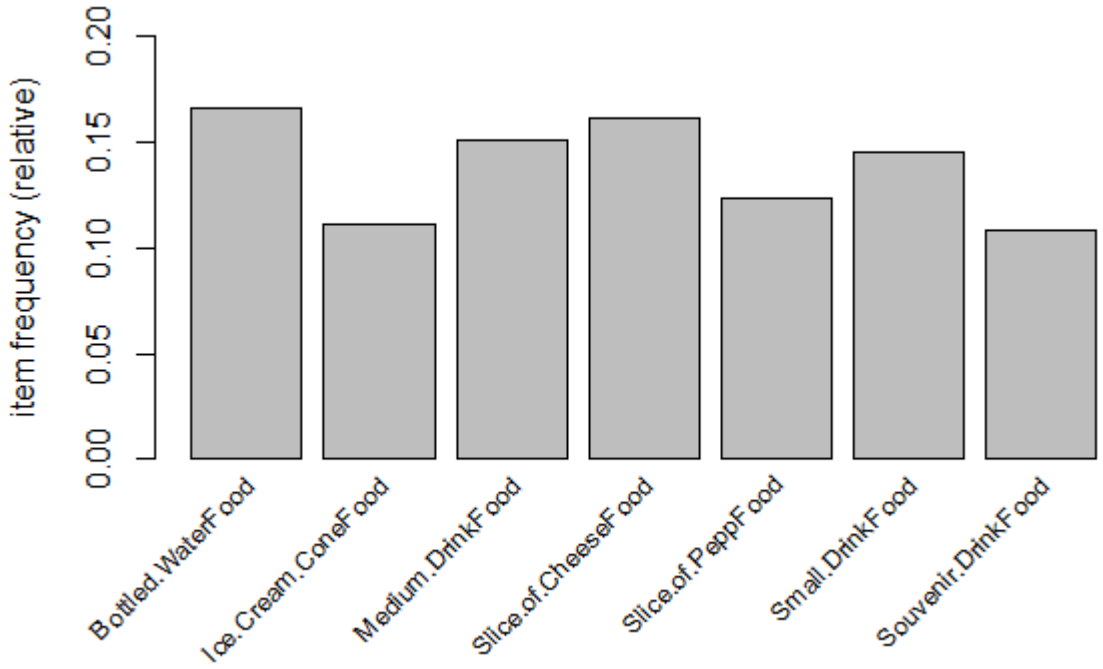


Figure: Item Frequency Plot

- Above plot shows frequently purchased items i.e. items that are present in more than 10% of transactions in the data set

After using apriori algorithm for market basket analysis on the data set at support level of 0.3% and confidence level of 50%, we get a set of 40 rules.

```
> inspect(subset(food_rules, size(food_rules)>3))
    lhs                                rhs                          support confidence     lift count
[1] {Krazy.KritterFood,
     Medium.DrinkFood,
     Slice.of.PeppFood}    => {Slice.of.CheeseFood} 0.003250157  0.5535714 3.437477    62
[2] {Medium.DrinkFood,
     Slice.of.PeppFood,
     Small.DrinkFood}      => {Slice.of.CheeseFood} 0.003145313  0.6000000 3.725781    60
[3] {Medium.DrinkFood,
     Slice.of.CheeseFood,
     Small.DrinkFood}      => {Slice.of.PeppFood}   0.003145313  0.5172414 4.191545    60
```

- Above snapshot shows association of items having size more than 3
- Support is number of percentage of transactions that include both antecedent and consequent item sets
- Confidence is number of transactions with all items in both item sets divided by number of transactions with the antecedent item set
- It can be inferred that when kritter food, drink food and pepp food are bought together then there is 55% chance of buying slice of cheese food along with these items

```
    lhs                                  rhs                                support confidence      lift count
[1] {Hot.Chocolate.Souvenir.RefillFood} => {Hot.Chocolate.SouvenirFood} 0.01499266  0.5596869 13.180972   286
[2] {ToppingFood}                       => {Ice.Cream.ConeFood}         0.02856993  0.9981685  8.947868   545
[3] {Chicken.TendersFood}               => {French.Fries.BasketFood}    0.01729922  0.7586207  7.771992   330
[4] {CheeseburgerFood}                  => {French.Fries.BasketFood}    0.01687985  0.7931034  8.125264   322
[5] {GatoradeFood,
     Slice.of.PeppFood}                 => {Slice.of.CheeseFood}        0.01011743  0.5830816  3.620724   193
[6] {Medium.DrinkFood,
     Slice.of.PeppFood}                 => {Slice.of.CheeseFood}        0.01362969  0.5273834  3.274858   260
[7] {Bottled.WaterFood,
     Slice.of.PeppFood}                 => {Slice.of.CheeseFood}        0.01069407  0.5151515  3.198903   204
```

- Above rules specify items wherein support is greater than 10%

```
    lhs                                  rhs                                support confidence     lift count
[1] {Small.Pink.LemonadeFood}           => {Chicken.Nugget.BasketFood}  0.003355001  0.5925926 16.03446    64
[2] {Side.of.CheeseFood}                => {Cheese.ConeyFood}           0.004665548  0.6846154 25.91215    89
[3] {Side.of.CheeseFood}                => {Hot.DogFood}                0.006290627  0.9230769 21.60566   120
[4] {Hot.Chocolate.Souvenir.RefillFood} => {Hot.Chocolate.SouvenirFood} 0.014992661  0.5596869 13.18097   286
[5] {Cheese.ConeyFood,
     Side.of.CheeseFood}                => {Hot.DogFood}                0.004351017  0.9325843 21.82819    83
[6] {Hot.DogFood,
     Side.of.CheeseFood}                => {Cheese.ConeyFood}           0.004351017  0.6916667 26.17903    83
```

- Above rules shows items sets wherein lift is greater than 10
- More the lift then more is the strong association between items

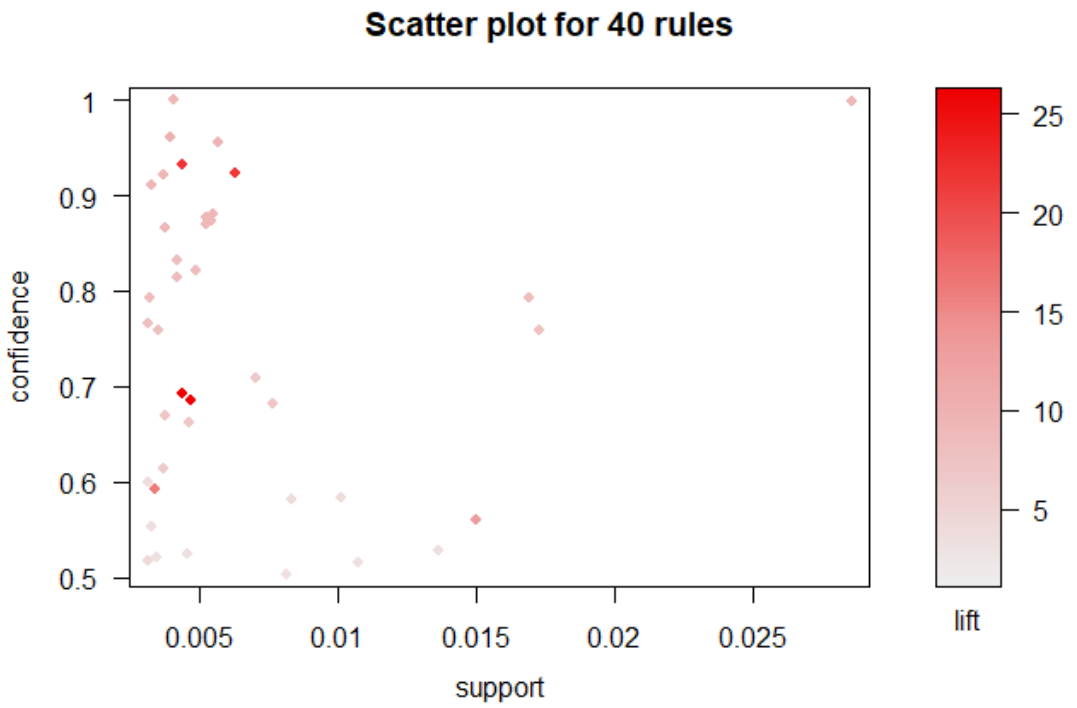


Fig: Scatter plot for association rules

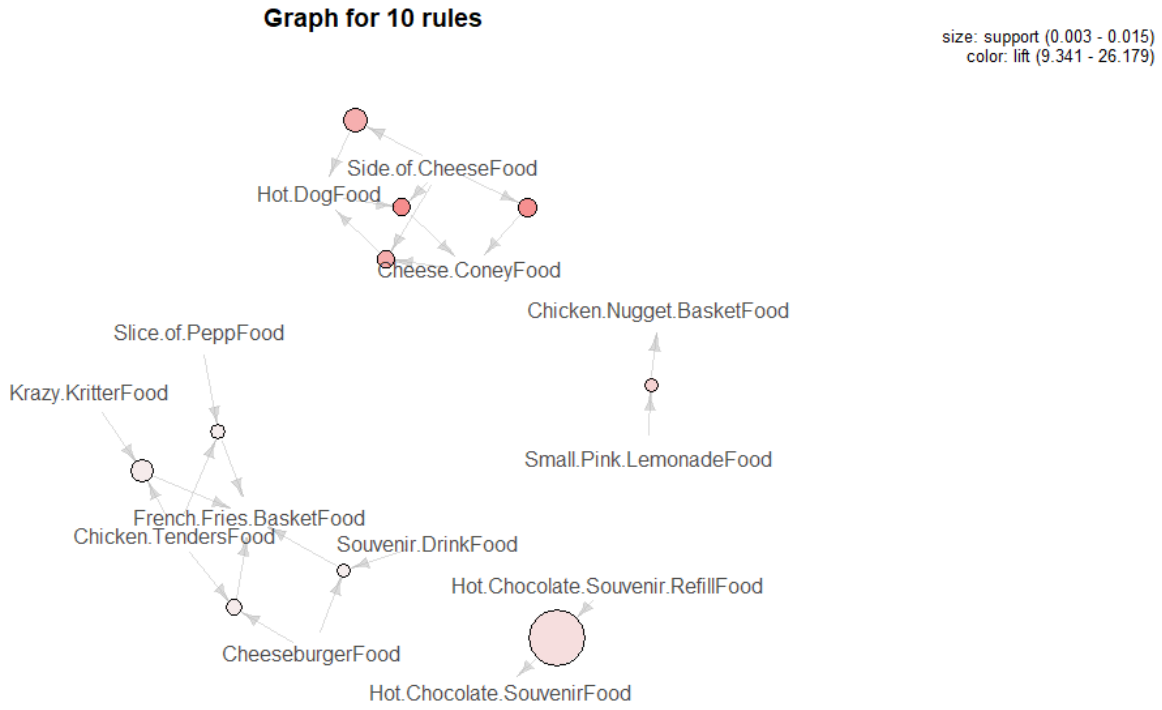
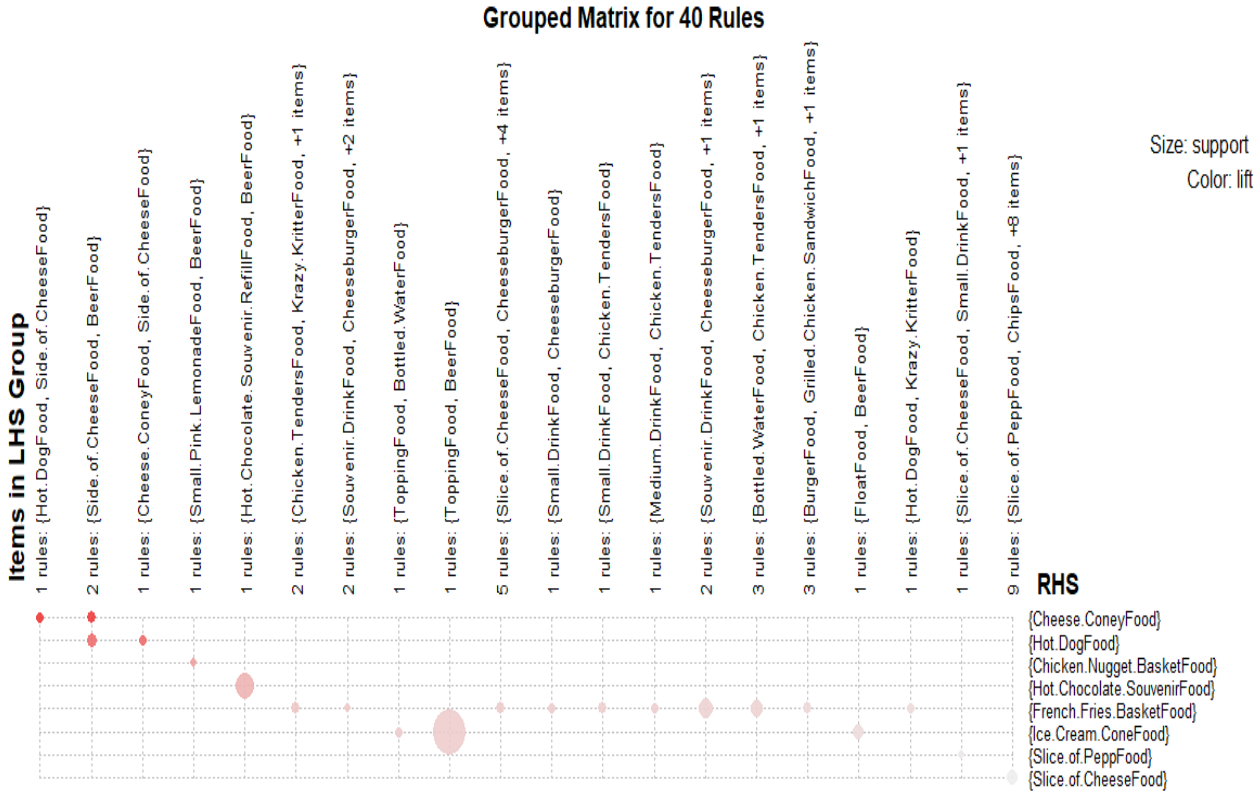


Fig: Graph for 10 rules

Below plot shows association rules grouped in a matrix form.



3. Classification using SPSS Modeler

German credit score data has records that are classified as good credit and bad credit. We will use this data set to train our model to predict whether a test record is a good credit or bad credit. German credit score data set consists of 1000 observations and 21 features. Variable “Response” is the dependent variable in this data set. Response variable 0 indicate “good credit risk” and 1 indicates “bad credit risk”.

Classification Tree Model

Results for output field Response1

Individual Models

Comparing $R-Response1 with Response1

'Partition'	1_Training		2_Testing	
Correct	477	58.03%	99	55.62%
Wrong	345	41.97%	79	44.38%
Total	822		178	

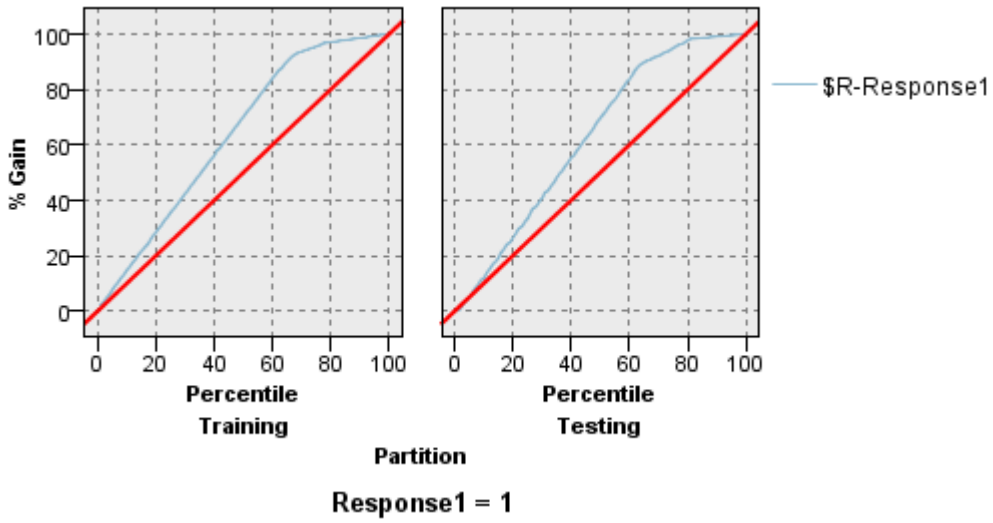
Performance Evaluation

'Partition' = 1_Training	
0	0.284
1	0.318

'Partition' = 2_Testing	
0	0.29
1	0.27

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$R-Response1	0.694	0.389	0.688	0.376



Logistic Regression Model

- Results for output field Response1
  - Individual Models
    - Comparing $L-Response1 with Response1

'Partition'	1_Training		2_Testing	
Correct	648	78.83%	129	72.47%
Wrong	174	21.17%	49	27.53%
Total	822		178	

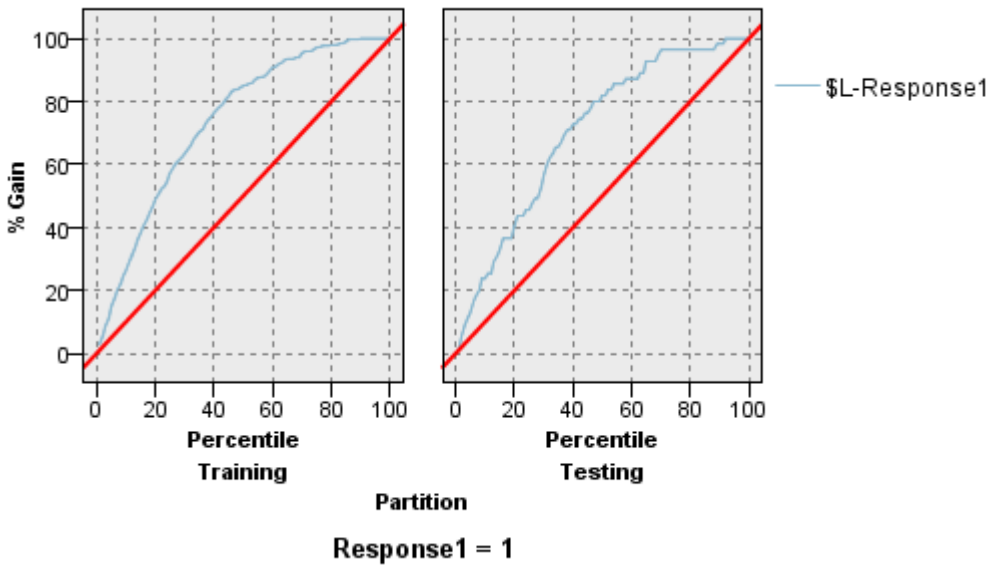
    - Performance Evaluation

'Partition' = 1_Training	
0	0.157
1	0.828

'Partition' = 2_Testing	
0	0.122
1	0.604

  - Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$L-Response1	0.836	0.673	0.799	0.599



Neural Network Model

Results for output field Response1

Individual Models

Comparing $N-Response1 with Response1

'Partition'	1_Training		2_Testing	
Correct	635	77.25%	126	70.79%
Wrong	187	22.75%	52	29.21%
Total	822		178	

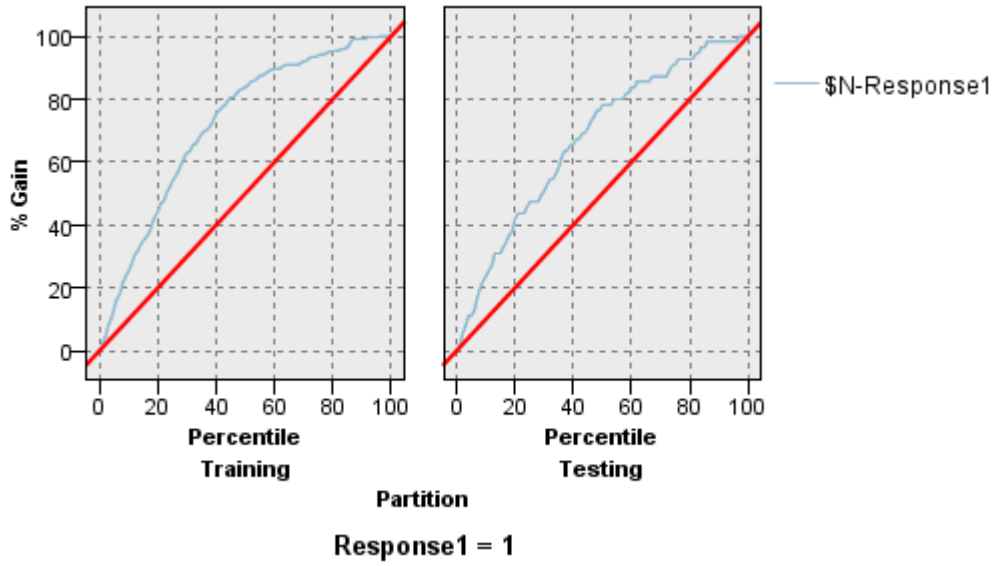
Performance Evaluation

'Partition' = 1_Training	
0	0.141
1	0.784

'Partition' = 2_Testing	
0	0.121
1	0.538

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$N-Response1	0.813	0.625	0.753	0.505



SPSS Diagram