**Codes used to analyze the “mortDefault2000” Dataset**

> sampleDataDir <- rxGetOption("sampleDataDir")

> getwd()

[1] "/home/remoteuser"

> inputFile <- file.path(sampleDataDir, "mortDefaultSmall2000.csv")

> mortDS <- rxImport(inData = inputFile, outFile = "ADS.xdf", missingValueString = "M", stringsAsFactors = TRUE, overwrite = TRUE)

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.039 seconds

> rxGetInfo(mortDS, numRows=5)

File name: /home/remoteuser/ADS.xdf

Number of observations: 10000

Number of variables: 6

Number of blocks: 1

Compression type: zlib

Data (5 rows starting with row 1):

creditScore houseAge yearsEmploy ccDebt year default

1 691 16 9 6725 2000 0

2 691 4 4 5077 2000 0

3 743 18 3 3080 2000 0

4 728 22 1 4345 2000 0

5 745 17 3 2969 2000 0

|  |
| --- |
| > rxGetVarInfo(mortDS)  Var 1: creditScore, Type: integer, Low/High: (486, 895)  Var 2: houseAge, Type: integer, Low/High: (0, 40)  Var 3: yearsEmploy, Type: integer, Low/High: (0, 14)  Var 4: ccDebt, Type: integer, Low/High: (0, 12275)  Var 5: year, Type: integer, Low/High: (2000, 2000)  Var 6: default, Type: integer, Low/High: (0, 1) |
|  |
| |  | | --- | | >rxSummary(~., data = mortDS, blocksPerRead = 2)  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds  Computation time: 0.006 seconds.  Call:  rxSummary(formula = ~., data = mortDS, blocksPerRead = 2)  Summary Statistics Results for: ~.  Data: mortDS (RxXdfData Data Source)  File name: ADS.xdf  Number of valid observations: 10000    Name Mean StdDev Min Max ValidObs MissingObs  creditScore 699.6437 5.061007e+01 486 895 10000 0  houseAge 19.9890 7.616776e+00 0 40 10000 0  yearsEmploy 4.9711 2.012826e+00 0 14 10000 0  ccDebt 4973.8959 1.993051e+03 0 12275 10000 0  year 2000.0000 0.000000e+00 2000 2000 10000 0  default 0.0011 3.314966e-02 0 1 10000 0  > rxHistogram(~creditScore, data = mortDS )  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.006 seconds  Computation time: 0.011 seconds.  head(mortDS)  creditScore houseAge yearsEmploy ccDebt year default  1 691 16 9 6725 2000 0  2 691 4 4 5077 2000 0  3 743 18 3 3080 2000 0  4 728 22 1 4345 2000 0  5 745 17 3 2969 2000 0  6 539 15 3 4588 2000 0  > rxGetInfo(mortDS, getVarInfo = TRUE, numRows=3)  File name: /home/remoteuser/ADS.xdf  Number of observations: 10000  Number of variables: 6  Number of blocks: 1  Compression type: zlib  Variable information:  Var 1: creditScore, Type: integer, Low/High: (486, 895)  Var 2: houseAge, Type: integer, Low/High: (0, 40)  Var 3: yearsEmploy, Type: integer, Low/High: (0, 14)  Var 4: ccDebt, Type: integer, Low/High: (0, 12275)  Var 5: year, Type: integer, Low/High: (2000, 2000)  Var 6: default, Type: integer, Low/High: (0, 1)  Data (3 rows starting with row 1):  creditScore houseAge yearsEmploy ccDebt year default  1 691 16 9 6725 2000 0  2 691 4 4 5077 2000 0  3 743 18 3 3080 2000 0  C:\Users\Yenem\AppData\Local\Microsoft\Windows\INetCacheContent.Word\plot (3).png  > myCube = rxCube(~F(creditScore):catDebt, data = mortDataNew)  Rows Read: 9982, Total Rows Processed: 9982, Total Chunk Time: 0.002 seconds  Computation time: 0.004 seconds.  > rxLinePlot(Counts~creditScore|catDebt, data=rxResultsDF(myCube))  C:\Users\Yenem\AppData\Local\Microsoft\Windows\INetCacheContent.Word\plot (4).png | |

> logitModel<-rxLogit(default~ccDebt+yearsEmploy+houseAge+creditScore, data=mortDS)

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Starting values (iteration 1) time: 0.012 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.006 seconds

Iteration 2 time: 0.013 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Iteration 3 time: 0.010 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Iteration 4 time: 0.010 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Iteration 5 time: 0.011 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Iteration 6 time: 0.011 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Iteration 7 time: 0.011 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Iteration 8 time: 0.011 secs.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Iteration 9 time: 0.010 secs.

Elapsed computation time: 0.101 secs.

> summary(logitModel)

Call:

rxLogit(formula = default ~ ccDebt + yearsEmploy + houseAge +

creditScore, data = mortDS)

Logistic Regression Results for: default ~ ccDebt + yearsEmploy + houseAge +

creditScore

Data: mortDS (RxXdfData Data Source)

File name: ADS.xdf

Dependent variable(s): default

Total independent variables: 5

Number of valid observations: 10000

Number of missing observations: 0

-2\*LogLikelihood: 98.0903 (Residual deviance on 9995 degrees of freedom)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.727707 5.023113 -1.738 0.0823 .

ccDebt 0.001445 0.000222 6.511 7.45e-11 \*\*\*

yearsEmploy -0.321303 0.157559 -2.039 0.0414 \*

houseAge 0.013048 0.041306 0.316 0.7521

creditScore -0.011633 0.007161 -1.625 0.1042

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Condition number of final variance-covariance matrix: 2.075

Number of iterations: 9

> predictDF <- data.frame(

+ creditScore = rep(c(700, 800),4),

+ ccDebt = rep(c(1000, 10000), 4),

+ houseAge = rep(c(1,5,10,20), 2),

+ yearsEmploy = 7)

> predictDF

creditScore ccDebt houseAge yearsEmploy

1 700 1000 1 7

2 800 10000 5 7

3 700 1000 10 7

4 800 10000 20 7

5 700 1000 1 7

6 800 10000 5 7

7 700 1000 10 7

8 800 10000 20 7

> Predition<-rxPredict(modelObject = logitModel,data=predictDF,outData = predictDF,type = "response")

Rows Read: 8, Total Rows Processed: 8, Total Chunk Time: 0.002 seconds

> Predition

creditScore ccDebt houseAge yearsEmploy default\_Pred

1 700 1000 1 7 2.135591e-08

2 800 10000 5 7 3.123352e-03

3 700 1000 10 7 2.401691e-08

4 800 10000 20 7 3.796000e-03

5 700 1000 1 7 2.135591e-08

6 800 10000 5 7 3.123352e-03

7 700 1000 10 7 2.401691e-08

8 800 10000 20 7 3.796000e-03

> predictDF <- rxPredict(modelObject = logitModel, data = predictDF,

+ outData = predictDF)

Rows Read: 8, Total Rows Processed: 8, Total Chunk Time: 0.002 seconds

> predictDF[order(predictDF$default\_Pred, decreasing = TRUE),]

creditScore ccDebt houseAge yearsEmploy default\_Pred

4 800 10000 20 7 3.796000e-03

8 800 10000 20 7 3.796000e-03

2 800 10000 5 7 3.123352e-03

6 800 10000 5 7 3.123352e-03

3 700 1000 10 7 2.401691e-08

7 700 1000 10 7 2.401691e-08

1 700 1000 1 7 2.135591e-08

5 700 1000 1 7 2.135591e-08

**K MEANS**

rxDataStep(inData = inputFile, outFile = "mortDefault2000.xdf", varsToKeep=c("creditScore", "houseAge", "yearsEmploy", "ccDebt","default","year"))

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.034 seconds

> kclusts1 <- rxKmeans(formula= ~creditScore + houseAge +yearsEmploy + ccDebt +year,

+ data = "mortDefault2000.xdf",

+ outFile = "mortDefault2000.xdf", numClusters=5)

blocksPerRead must be 1 if predictions or residuals are be being written to the input data source.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.002 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.006 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

> kclusts1

Call:

rxKmeans(formula = ~creditScore + houseAge + yearsEmploy + ccDebt +

year, data = "mortDefault2000.xdf", outFile = "mortDefault2000.xdf",

numClusters = 5)

Data: "mortDefault2000.xdf"

Number of valid observations: 10000

Number of missing observations: 0

Clustering algorithm:

K-means clustering with 5 clusters of sizes 1062, 1041, 3025, 2577, 2295

Cluster means:

creditScore houseAge yearsEmploy ccDebt year

1 700.5226 20.18173 4.870998 8435.572 2000

2 700.2152 20.09894 4.894332 1513.306 2000

3 700.2215 19.93884 4.987769 4898.183 2000

4 699.5390 20.03104 4.939464 6455.495 2000

5 698.3338 19.86885 5.065795 3377.872 2000

Within cluster sum of squares by cluster:

1 2 3 4 5

747197478 523060378 593616372 636310897 531494789

Available components:

[1] "centers" "size" "withinss" "valid.obs" "missing.obs"

[6] "numIterations" "tot.withinss" "totss" "betweenss" "cluster"

[11] "params" "formula" "call"

>

**TO CHEK IF IT WORKS IN ANY CLUSTER NUMBER**

|  |
| --- |
| > clust1Lm <- rxLinMod(default ~ creditScore + houseAge + yearsEmploy + ccDebt + year ,  "mortDefault2000.xdf",rowSelection = .rxCluster == 1 )  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds  Computation time: 0.014 seconds.  > clust5Lm <- rxLinMod(default ~ creditScore + houseAge + yearsEmploy + ccDebt + year,  "mortDefault2000.xdf",rowSelection = .rxCluster == 5)  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds  Computation time: 0.049 seconds.  > clust1Lm  Call:  rxLinMod(formula = default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year, data = "mortDefault2000.xdf", rowSelection = .rxCluster ==  1)  Linear Regression Results for: default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year  Data: "mortDefault2000.xdf" (RxXdfData Data Source)  File name: mortDefault2000.xdf  Dependent variable(s): default  Total independent variables: 6 (Including number dropped: 1)  Number of valid observations: 1062  Number of missing observations: 0    Coefficients:  default  (Intercept) -1.294048e-01  creditScore -9.587951e-05  houseAge 3.954167e-05  yearsEmploy -1.214970e-03  ccDebt 2.491417e-05  year Dropped  > summary(clust1Lm)  Call:  rxLinMod(formula = default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year, data = "mortDefault2000.xdf", rowSelection = .rxCluster ==  1)  Linear Regression Results for: default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year  Data: "mortDefault2000.xdf" (RxXdfData Data Source)  File name: mortDefault2000.xdf  Dependent variable(s): default  Total independent variables: 6 (Including number dropped: 1)  Number of valid observations: 1062  Number of missing observations: 0    Coefficients: (1 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) -1.294e-01 4.899e-02 -2.641 0.00838 \*\*  creditScore -9.588e-05 5.669e-05 -1.691 0.09107 .  houseAge 3.954e-05 3.575e-04 0.111 0.91194  yearsEmploy -1.215e-03 1.326e-03 -0.916 0.35981  ccDebt 2.491e-05 3.274e-06 7.609 2.22e-16 \*\*\*  year Dropped Dropped Dropped Dropped  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.08933 on 1057 degrees of freedom  Multiple R-squared: 0.05483  Adjusted R-squared: 0.05125  F-statistic: 15.33 on 4 and 1057 DF, p-value: 3.427e-12  Condition number: 1.1365  > summary( clust5Lm)  Call:  rxLinMod(formula = default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year, data = "mortDefault2000.xdf", rowSelection = .rxCluster ==  5)  Linear Regression Results for: default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year  Data: "mortDefault2000.xdf" (RxXdfData Data Source)  File name: mortDefault2000.xdf  Dependent variable(s): default  Total independent variables: 6 (Including number dropped: 1)  Number of valid observations: 2295  Number of missing observations: 0    Coefficients: (1 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) 0 0 Dropped 2.22e-16 \*\*\*  creditScore 0 0 Dropped 2.22e-16 \*\*\*  houseAge 0 0 Dropped 2.22e-16 \*\*\*  yearsEmploy 0 0 Dropped 2.22e-16 \*\*\*  ccDebt 0 0 Dropped 2.22e-16 \*\*\*  year Dropped Dropped Dropped Dropped  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0 on 2290 degrees of freedom  Multiple R-squared: NaN  Adjusted R-squared: NaN  F-statistic: NaN on 4 and 2290 DF, p-value: NA  Condition number: 3491114492  > clust5Lm  Call:  rxLinMod(formula = default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year, data = "mortDefault2000.xdf", rowSelection = .rxCluster ==  5)  Linear Regression Results for: default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year  Data: "mortDefault2000.xdf" (RxXdfData Data Source)  File name: mortDefault2000.xdf  Dependent variable(s): default  Total independent variables: 6 (Including number dropped: 1)  Number of valid observations: 2295  Number of missing observations: 0    Coefficients:  default  (Intercept) 0  creditScore 0  houseAge 0  yearsEmploy 0  ccDebt 0  year Dropped  > clust3Lm <- rxLinMod(default ~ creditScore + houseAge + yearsEmploy + ccDebt + year, "mortDefault2000.xdf",rowSelection = .rxCluster == 3)  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.005 seconds  Computation time: 0.112 seconds.  > clust3Lm  Call:  rxLinMod(formula = default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year, data = "mortDefault2000.xdf", rowSelection = .rxCluster ==  3)  Linear Regression Results for: default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year  Data: "mortDefault2000.xdf" (RxXdfData Data Source)  File name: mortDefault2000.xdf  Dependent variable(s): default  Total independent variables: 6 (Including number dropped: 1)  Number of valid observations: 3025  Number of missing observations: 0    Coefficients:  default  (Intercept) 0  creditScore 0  houseAge 0  yearsEmploy 0  ccDebt 0  year Dropped  > summary(clust3Lm)  Call:  rxLinMod(formula = default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year, data = "mortDefault2000.xdf", rowSelection = .rxCluster ==  3)  Linear Regression Results for: default ~ creditScore + houseAge + yearsEmploy +  ccDebt + year  Data: "mortDefault2000.xdf" (RxXdfData Data Source)  File name: mortDefault2000.xdf  Dependent variable(s): default  Total independent variables: 6 (Including number dropped: 1)  Number of valid observations: 3025  Number of missing observations: 0    Coefficients: (1 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) 0 0 Dropped 2.22e-16 \*\*\*  creditScore 0 0 Dropped 2.22e-16 \*\*\*  houseAge 0 0 Dropped 2.22e-16 \*\*\*  yearsEmploy 0 0 Dropped 2.22e-16 \*\*\*  ccDebt 0 0 Dropped 2.22e-16 \*\*\*  year Dropped Dropped Dropped Dropped  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0 on 3020 degrees of freedom  Multiple R-squared: NaN  Adjusted R-squared: NaN  F-statistic: NaN on 4 and 3020 DF, p-value: NA  Condition number: 9182786855 |
|  |
| |  | | --- | |  | |

**K MEANS 2**

Kmeanscluster<-rxKmeans(formula = ~creditScore + houseAge + yearsEmploy + ccDebt + year, data=mortDS, numClusters = 3, outFile = mortDS,algorithm = "lloyd",overwrite = TRUE)

blocksPerRead must be 1 if predictions or residuals are be being written to the input data source.

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds

> Kmeanscluster

Call:

rxKmeans(formula = ~creditScore + houseAge + yearsEmploy + ccDebt +

year, data = mortDS, outFile = mortDS, overwrite = TRUE,

numClusters = 3, algorithm = "lloyd")

Data: mortDS

Number of valid observations: 10000

Number of missing observations: 0

Clustering algorithm:

K-means clustering with 3 clusters of sizes 2685, 2725, 4590

Cluster means:

creditScore houseAge yearsEmploy ccDebt year

1 698.1572 19.96834 5.008194 2517.291 2000

2 700.7828 19.94826 4.920367 7404.522 2000

3 699.8370 20.02527 4.979521 4967.911 2000

Within cluster sum of squares by cluster:

1 2 3

2465661786 2815003925 2161055991

Available components:

[1] "centers" "size" "withinss" "valid.obs" "missing.obs"

[6] "numIterations" "tot.withinss" "totss" "betweenss" "cluster"

[11] "params" "formula" "call"

**Decision TREE Mortgage**

|  |
| --- |
| sampleDataDir <- rxGetOption("sampleDataDir")  > inputFile <- file.path(sampleDataDir, "mortDefaultSmall2000.csv")  > mortDS <- rxImport(inData = inputFile, outFile = "ADS.xdf", missingValueString = "M", stringsAsFactors = TRUE, overwrite = TRUE)  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.032 seconds  > mortgageTree <- rxDTree(default ~ houseAge + year + creditScore + yearsEmploy +ccDebt, data = mortDS, cp=0.01)  The variance is zero for the variable(s): "year"  which are eliminated for further analysis.  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.011 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.008 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: 0.003 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 4, Total Rows Processed: 4, Total Chunk Time: 0.004 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 4, Total Rows Processed: 4, Total Chunk Time: 0.002 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 4, Total Rows Processed: 4, Total Chunk Time: 0.002 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: Less than .001 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.006 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: 0.002 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: 0.002 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: Less than .001 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.004 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.015 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.006 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: 0.002 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: 0.002 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.007 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: 0.002 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.006 seconds  Rows Read: 2, Total Rows Processed: 2, Total Chunk Time: Less than .001 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.003 seconds  Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.002 seconds  Elapsed time for DTreeEstimation: 0.412 secs.  Elapsed time for BxDTreeBase: 0.414 secs.    > mortgageTree  Call:  rxDTree(formula = default ~ houseAge + year + creditScore + yearsEmploy +  ccDebt, data = mortDS, cp = 0.01)  File: /home/remoteuser/ADS.xdf  Number of valid observations: 10000  Number of missing observations: 0  Tree representation:  n= 10000  node), split, n, deviance, yval  \* denotes terminal node  1) root 10000 10.9879000 0.0011000000  2) ccDebt< 9340 9899 2.9990910 0.0003030609 \*  3) ccDebt>=9340 101 7.3663370 0.0792079200  6) creditScore>=706.5 43 0.9767442 0.0232558100 \*  7) creditScore< 706.5 58 6.1551720 0.1206897000 \*  >  > plot(rxAddInheritance(mortgageTree))  >  > text(rxAddInheritance(mortgageTree))  > plot(rxAddInheritance(mortgageTree),uniform = TRUE,margin = 0.1)  > text(rxAddInheritance(mortgageTree),digits = 2,cex = 0.6)  > title(main = "A Simple Classification Tree for Mortgage Data") |
|  |
| |  | | --- | |  | |











