Reglas de asociación (Basket Market Analysis)

Transformamos las variables numéricas en categóricas aplicando la función `discretizeDF`.

Cabe resaltar que ahora la base de datos que se utilizará es “dcat” con las variables numéricas como categoricas.

Seguidamente, se transformará “dcat” en un data transactions para poder aplicar el Basket Market Analysis.

```
transactions in sparse format with
 5000 transactions (rows) and
 113 items (columns)
```

Con el siguiente summary, se puede ver con más detalle lo que se tiene:

```
transactions as itemMatrix in sparse format with
 5000 rows (elements/itemsets/transactions) and
 113 columns (items) and a density of 0.1415929

most frequent items:
NAME_EDUCATION_TYPE=Secondary / secondary special
                                             3746
                           REGION_RATING_CLIENT=2
                                             3641
                                    CODE_GENDER=F
                                             3098
                       NAME_FAMILY_STATUS=Married
                                             3095
                                         TARGET=0
                                             2865
                                          (Other)
                                            63555

element (itemset/transaction) length distribution:
sizes
  16
5000

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     16      16      16      16      16      16

includes extended item information - examples:
                        labels        variables      levels
1                CODE_GENDER=F      CODE_GENDER           F
2                CODE_GENDER=M      CODE_GENDER           M
3 NAME_INCOME_TYPE=Businessman NAME_INCOME_TYPE Businessman

includes extended transaction information - examples:
  transactionID
1             1
2             2
3             3
```

Apriori

El primer paso consiste en especificar los parámetros:

El siguiente paso es crear las reglas de asociación:

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5   0.002      1
 maxlen target  ext
     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 10

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[110 item(s), 5000 transaction(s)] done [0.00s].
sorting and recoding items ... [98 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [1.76s].
writing ... [2499418 rule(s)] done [0.55s].
creating S4 object  ... done [1.13s].
```

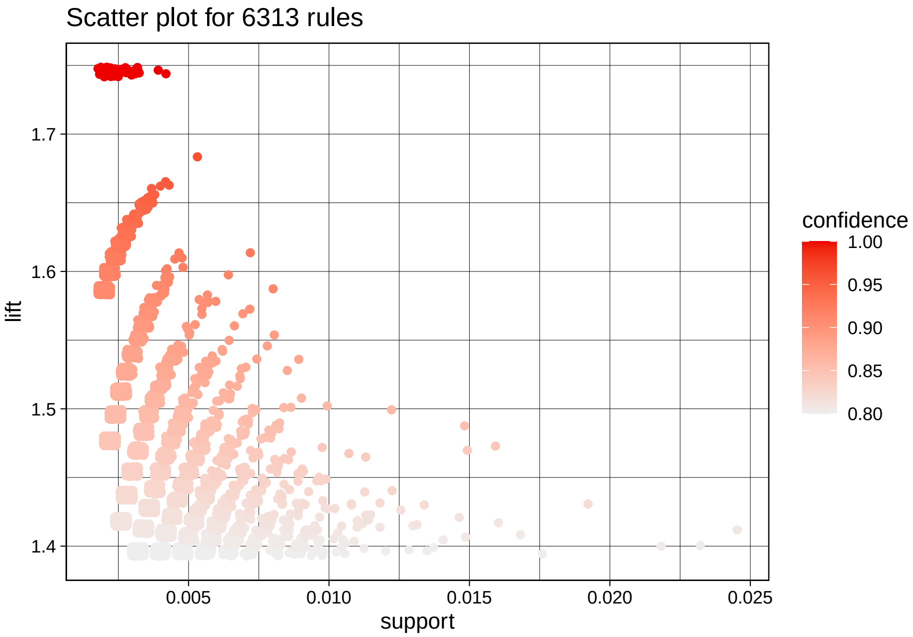
Dividimos las reglas de asociación obtenidas según lo consecuente que es la variable respuesta. La variable respuesta es `TARGET`, que toma valores de 1 o 0.

Se eliminan las reglas redundantes en ambos casos:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.01065 0.02847 0.02959 0.05815 0.06071
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
2.000e-09 4.957e-04 1.916e-03 2.117e-03 3.702e-03 4.951e-03
```

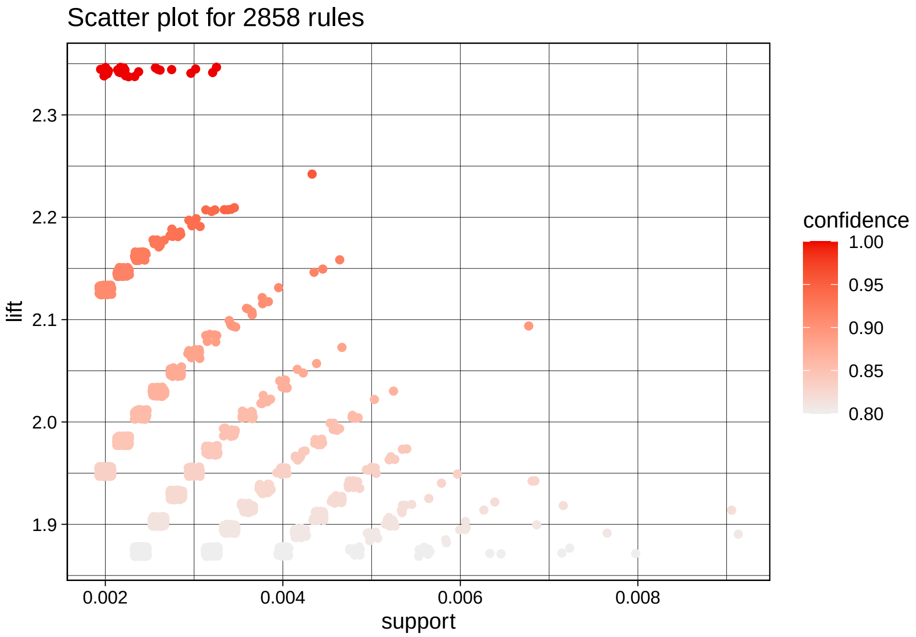
Figura 52: Scatter plot for 6313 rules



Como se puede ver, el primer gráfico muestra la matriz de puntos de las reglas de asociación filtrada respecto la métrica lift. La reglas de asociación de interés corresponden a los puntos con un color rojo de mayor intensidad (confianza que supere la mínima, 0.8) y se aprecia, estas reglas se situan en el gráfico con un soporte mayor al mínimo (0.002).

En el último gráfico se ve algo parecido, aquí las reglas de asociación que interesan corresponden a los puntos con una intensidad roja mayor y los puntos más grandes, que corresponderan a las reglas que tienen un soporte superior al mínimo (0.002).

Figura 53: Scatter plot for 2858 rules



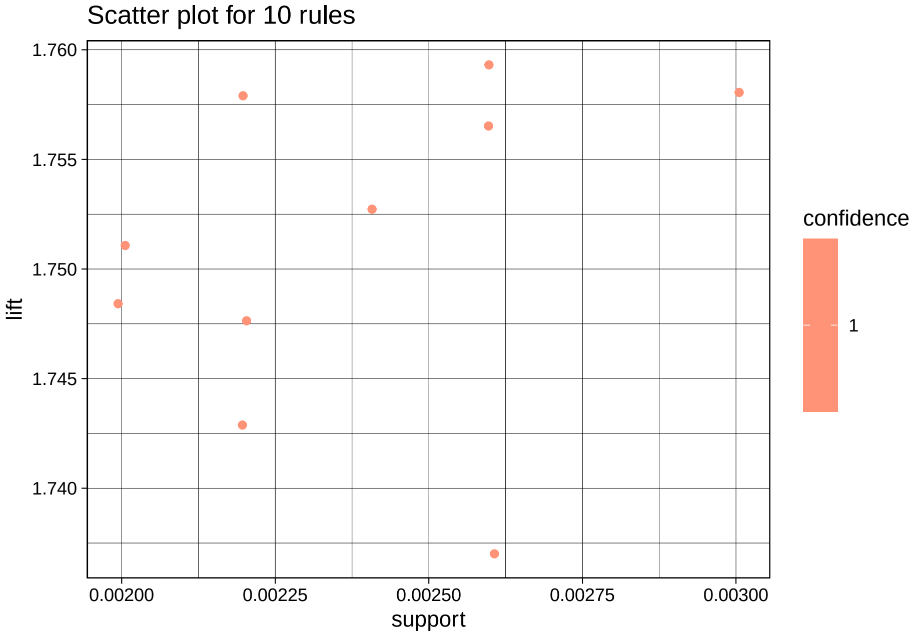
Estos gráficos se interpretan de manera igual a los anteriores vistos.

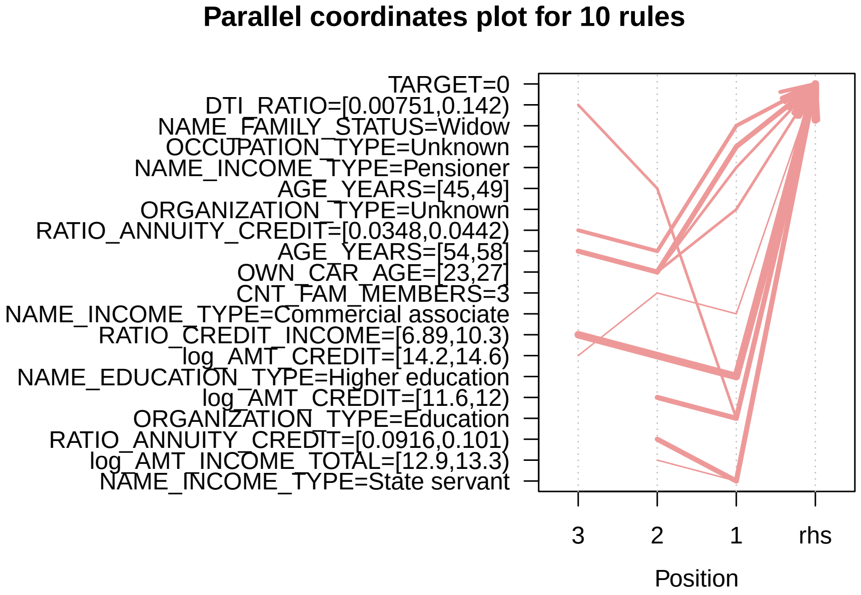
Con Target = 0 se obtienen 6313 reglas y con Tagret = 1 2858 reglas. Con la gran cantidad de reglas, la atención se centra en las 10 primeras reglas en cada caso con mayor lift.

Por tanto, se ven las 10 primeras reglas en cada caso con mayor lift, es decir, van ordenadas de forma decreciente siendo la primera la que tiene una mayor asociación encontrada con la variable respuesta, y se grafican en cada caso.

```
     lhs                                         rhs        support confidence coverage     lift count
[1]  {NAME_INCOME_TYPE=State servant,
      log_AMT_INCOME_TOTAL=[12.9,13.3)}       => {TARGET=0}  0.0020          1   0.0020 1.745201    10
[2]  {NAME_INCOME_TYPE=State servant,
      RATIO_ANNUITY_CREDIT=[0.0916,0.101)}    => {TARGET=0}  0.0026          1   0.0026 1.745201    13
[3]  {ORGANIZATION_TYPE=Education,
      log_AMT_CREDIT=[11.6,12)}               => {TARGET=0}  0.0026          1   0.0026 1.745201    13
[4]  {NAME_EDUCATION_TYPE=Higher education,
      log_AMT_CREDIT=[14.2,14.6),
      RATIO_CREDIT_INCOME=[6.89,10.3)}        => {TARGET=0}  0.0030          1   0.0030 1.745201    15
[5]  {NAME_INCOME_TYPE=Commercial associate,
      CNT_FAM_MEMBERS=3,
      log_AMT_CREDIT=[14.2,14.6)}             => {TARGET=0}  0.0020          1   0.0020 1.745201    10
[6]  {NAME_FAMILY_STATUS=Widow,
      AGE_YEARS=[54,58],
      RATIO_ANNUITY_CREDIT=[0.0348,0.0442)}   => {TARGET=0}  0.0024          1   0.0024 1.745201    12
[7]  {ORGANIZATION_TYPE=Unknown,
      OWN_CAR_AGE=[23,27],
      AGE_YEARS=[54,58]}                      => {TARGET=0}  0.0022          1   0.0022 1.745201    11
[8]  {NAME_INCOME_TYPE=Pensioner,
      OWN_CAR_AGE=[23,27],
      AGE_YEARS=[54,58]}                  => {TARGET=0}  0.0022          1    0.0022 1.745201     11
[9]  {OCCUPATION_TYPE=Unknown,
      OWN_CAR_AGE=[23,27],
      AGE_YEARS=[54,58]}                  => {TARGET=0}  0.0026          1    0.0026 1.745201     13
[10] {ORGANIZATION_TYPE=Education,
      AGE_YEARS=[45,49],
      DTI_RATIO=[0.00751,0.142)}          => {TARGET=0}  0.0022          1    0.0022 1.745201     11
```

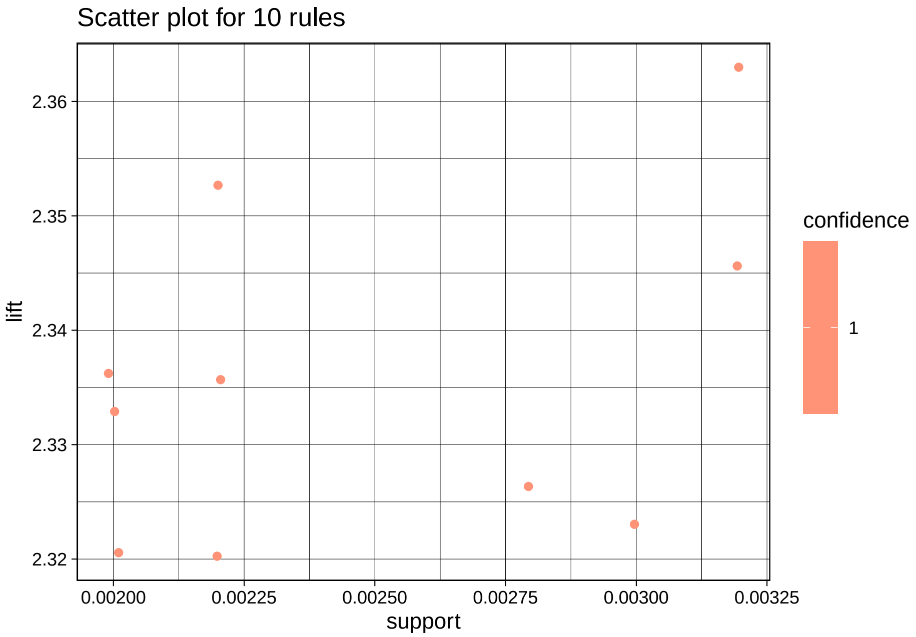
Figura 54: Scatter plot for 10 rules

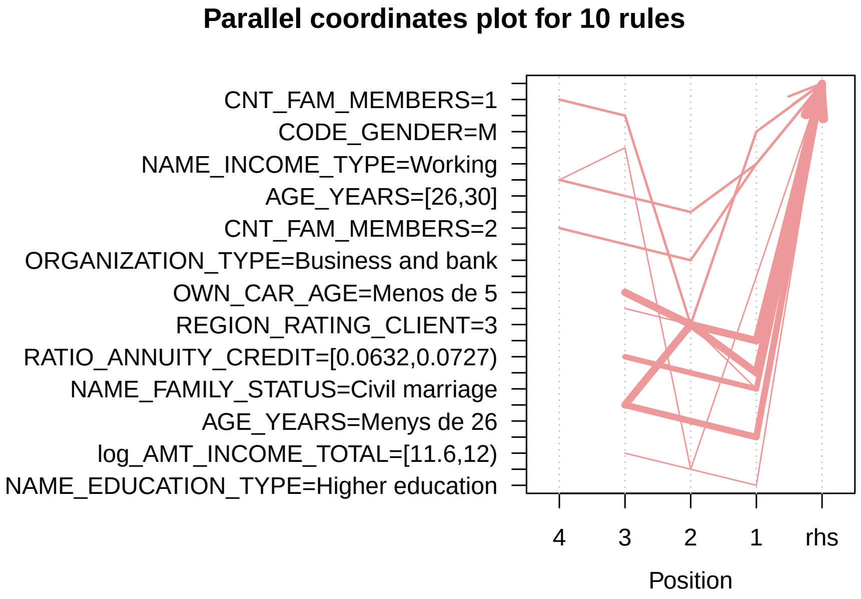




```
     lhs                                          rhs        support confidence coverage    lift count
[1]  {NAME_EDUCATION_TYPE=Higher education,
      OWN_CAR_AGE=[23,27],
      log_AMT_INCOME_TOTAL=[11.6,12)}          => {TARGET=1}  0.0020          1   0.0020 2.34192    10
[2]  {log_AMT_CREDIT=[12.4,12.9),
      AGE_YEARS=Menys de 26,
      RATIO_ANNUITY_CREDIT=[0.0727,0.0821)}    => {TARGET=1}  0.0030          1   0.0030 2.34192    15
[3]  {NAME_FAMILY_STATUS=Civil marriage,
      OCCUPATION_TYPE=Low-mid skill laborers,
      RATIO_ANNUITY_CREDIT=[0.0632,0.0727)}    => {TARGET=1}  0.0028          1   0.0028 2.34192    14
[4]  {OCCUPATION_TYPE=Low skill laborers,
      REGION_RATING_CLIENT=3,
      RATIO_ANNUITY_CREDIT=[0.0727,0.0821)}    => {TARGET=1}  0.0032          1   0.0032 2.34192    16
[5]  {NAME_FAMILY_STATUS=Civil marriage,
      REGION_RATING_CLIENT=3,
      AGE_YEARS=[45,49]}                       => {TARGET=1}  0.0020          1   0.0020 2.34192    10
[6]  {OCCUPATION_TYPE=Low-mid skill laborers,
      REGION_RATING_CLIENT=3,
      OWN_CAR_AGE=Menos de 5}                  => {TARGET=1}  0.0032          1   0.0032 2.34192    16
[7]  {NAME_INCOME_TYPE=Working,
      ORGANIZATION_TYPE=Business and bank,
      OWN_CAR_AGE=[28,32],
      CNT_FAM_MEMBERS=2}                       => {TARGET=1}  0.0022          1   0.0022 2.34192    11
[8]  {NAME_INCOME_TYPE=Working,
      ORGANIZATION_TYPE=Trade and telecom,
      AGE_YEARS=[26,30],
      DTI_RATIO=[0.142,0.276)}                 => {TARGET=1}  0.0022          1   0.0022 2.34192    11
[9]  {OCCUPATION_TYPE=Unknown,
      OWN_CAR_AGE=[23,27],
      log_AMT_INCOME_TOTAL=[11.2,11.6),
      DTI_RATIO=[0.142,0.276)}                 => {TARGET=1}  0.0020          1   0.0020 2.34192    10
[10] {CODE_GENDER=M,
      REGION_RATING_CLIENT=3,
      OWN_CAR_AGE=[19,22],
      CNT_FAM_MEMBERS=1}                       => {TARGET=1}  0.0022          1   0.0022 2.34192    11
```

Figura 55: Scatter plot for 10 rules





ECLAT

Para este apartado, se crearán las reglas de asocioación con ECLAT.

```
Eclat

parameter specification:
 tidLists support minlen maxlen            target  ext
    FALSE   0.002      1     10 frequent itemsets TRUE

algorithmic control:
 sparse sort verbose
      7   -2    TRUE

Absolute minimum support count: 10

create itemset ...
set transactions ...[110 item(s), 5000 transaction(s)] done [0.00s].
sorting and recoding items ... [98 item(s)] done [0.00s].
creating bit matrix ... [98 row(s), 5000 column(s)] done [0.00s].
writing  ... [1992604 set(s)] done [0.66s].
Creating S4 object  ... done [0.55s].


set of 51080 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9    10
    3   111  1756  8211 15442 14643  7758  2596   560

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    6.00    7.00    6.55    7.00   10.00

summary of quality measures:
    support           confidence          lift          itemset
 Min.   :0.002000   Min.   :0.8000   Min.   :1.396   Min.   :      2
 1st Qu.:0.002200   1st Qu.:0.8235   1st Qu.:1.437   1st Qu.: 271311
 Median :0.002400   Median :0.8462   Median :1.477   Median : 670530
 Mean   :0.002764   Mean   :0.8626   Mean   :1.505   Mean   : 767225
 3rd Qu.:0.003000   3rd Qu.:0.9091   3rd Qu.:1.587   3rd Qu.:1289119
 Max.   :0.024600   Max.   :1.0000   Max.   :1.745   Max.   :1987637

mining info:
 data ntransactions support
   tr          5000   0.002
                                                                                                   call
 eclat(data = tr, parameter = list(support = soporte_minimo, minlen = 1, maxlen = tamanyo_conjunto))
 confidence
        0.8


set of 10520 rules

rule length distribution (lhs + rhs):sizes
   3    4    5    6    7    8    9   10
   4  206 1324 3155 3432 1864  479   56

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   6.000   7.000   6.672   7.000  10.000
```

```
summary of quality measures:
    support           confidence          lift          itemset       
 Min.   :0.002000   Min.   :0.8000   Min.   :1.874   Min.   :    694  
 1st Qu.:0.002000   1st Qu.:0.8235   1st Qu.:1.929   1st Qu.: 418570  
 Median :0.002400   Median :0.8462   Median :1.982   Median : 917806  
 Mean   :0.002532   Mean   :0.8607   Mean   :2.016   Mean   : 882066  
 3rd Qu.:0.002800   3rd Qu.:0.9091   3rd Qu.:2.129   3rd Qu.:1238898  
 Max.   :0.009200   Max.   :1.0000   Max.   :2.342   Max.   :1976927  

mining info:
 data ntransactions support
   tr          5000   0.002
                                                                                                call
 eclat(data = tr, parameter = list(support = soporte_minimo, minlen = 1, maxlen = tamanyo_conjunto))
 confidence
        0.8

     lhs                                          rhs        support confidence     lift itemset
[1]  {NAME_INCOME_TYPE=State servant,                                                            
      log_AMT_INCOME_TOTAL=[12.9,13.3)}        => {TARGET=0}  0.0020          1 1.745201   10346
[2]  {NAME_EDUCATION_TYPE=Higher education,                                                      
      log_AMT_CREDIT=[14.2,14.6),                                                                
      RATIO_CREDIT_INCOME=[6.89,10.3)}         => {TARGET=0}  0.0030          1 1.745201   13160
[3]  {NAME_INCOME_TYPE=Commercial associate,                                                     
      CNT_FAM_MEMBERS=3,                                                                         
      log_AMT_CREDIT=[14.2,14.6)}              => {TARGET=0}  0.0020          1 1.745201   14725
[4]  {NAME_INCOME_TYPE=Commercial associate,                                                     
      NAME_FAMILY_STATUS=Married,                                                                
      ORGANIZATION_TYPE=Personal services,                                                       
      RATIO_CREDIT_INCOME=[0.125,3.51)}        => {TARGET=0}  0.0030          1 1.745201   23611
[5]  {CODE_GENDER=F,                                                                             
      OCCUPATION_TYPE=Low-mid skill laborers,                                                    
      REGION_RATING_CLIENT=2,                                                                    
      RATIO_ANNUITY_CREDIT=[0.101,0.111)}      => {TARGET=0}  0.0028          1 1.745201   62972
[6]  {NAME_INCOME_TYPE=State servant,                                                            
      RATIO_ANNUITY_CREDIT=[0.0916,0.101)}     => {TARGET=0}  0.0026          1 1.745201   66662
[7]  {NAME_FAMILY_STATUS=Widow,                                                                  
      AGE_YEARS=[54,58],                                                                         
      RATIO_ANNUITY_CREDIT=[0.0348,0.0442)}    => {TARGET=0}  0.0024          1 1.745201   95632
[8]  {ORGANIZATION_TYPE=Education,                                                               
      log_AMT_CREDIT=[11.6,12)}                => {TARGET=0}  0.0026          1 1.745201  118213
[9]  {NAME_INCOME_TYPE=Pensioner,                                                                
      OWN_CAR_AGE=[23,27],                                                                       
      AGE_YEARS=[54,58]}                       => {TARGET=0}  0.0022          1 1.745201  132450
[10] {ORGANIZATION_TYPE=Unknown,                                                                 
      OWN_CAR_AGE=[23,27],                                                                       
      AGE_YEARS=[54,58]}                       => {TARGET=0}  0.0022          1 1.745201  132466

     lhs                                           rhs        support confidence    lift itemset
[1]  {NAME_INCOME_TYPE=Working,                                                                  
      ORGANIZATION_TYPE=Business and bank,                                                       
      OWN_CAR_AGE=[28,32],
      CNT_FAM_MEMBERS=2}                    => {TARGET=1}  0.0022          1 2.34192   20786
[2]  {NAME_INCOME_TYPE=Working,
      ORGANIZATION_TYPE=Trade and telecom,
      AGE_YEARS=[26,30],
      DTI_RATIO=[0.142,0.276)}              => {TARGET=1}  0.0022          1 2.34192   78683
[3]  {OCCUPATION_TYPE=Unknown,
      OWN_CAR_AGE=[23,27],
      log_AMT_INCOME_TOTAL=[11.2,11.6),
      DTI_RATIO=[0.142,0.276)}              => {TARGET=1}  0.0020          1 2.34192  135824
[4]  {NAME_EDUCATION_TYPE=Higher education,
      OWN_CAR_AGE=[23,27],
      log_AMT_INCOME_TOTAL=[11.6,12)}       => {TARGET=1}  0.0020          1 2.34192  137658
[5]  {CODE_GENDER=M,
      REGION_RATING_CLIENT=3,
      OWN_CAR_AGE=[19,22],
      CNT_FAM_MEMBERS=1}                    => {TARGET=1}  0.0022          1 2.34192  204659
[6]  {log_AMT_CREDIT=[12.4,12.9),
      AGE_YEARS=Menys de 26,
      RATIO_ANNUITY_CREDIT=[0.0727,0.0821)} => {TARGET=1}  0.0030          1 2.34192  213819
[7]  {NAME_FAMILY_STATUS=Single / not married,
      OCCUPATION_TYPE=Low skill laborers,
      log_AMT_INCOME_TOTAL=[11.6,12),
      AGE_YEARS=Menys de 26}                => {TARGET=1}  0.0020          1 2.34192  217788
[8]  {NAME_FAMILY_STATUS=Separated,
      OCCUPATION_TYPE=Low-mid skill laborers,
      log_AMT_INCOME_TOTAL=[12,12.5),
      RATIO_CREDIT_INCOME=[0.125,3.51)}     => {TARGET=1}  0.0024          1 2.34192  244609
[9]  {NAME_FAMILY_STATUS=Civil marriage,
      OCCUPATION_TYPE=Low-mid skill laborers,
      RATIO_ANNUITY_CREDIT=[0.0632,0.0727)} => {TARGET=1}  0.0028          1 2.34192  303233
[10] {OCCUPATION_TYPE=Low-mid skill laborers,
      CNT_FAM_MEMBERS=2,
      RATIO_ANNUITY_CREDIT=[0.0632,0.0727),
      DTI_RATIO=[0.00751,0.142)}            => {TARGET=1}  0.0026          1 2.34192  319187
```