Building a Robust Geodemographic Segmentation Model

P12-Churn-Modelling : Lequel des clients est susceptible de quitter la banque ?

On a dummify les variables Gender et Geography vu qu'elles sont des variables catégorielles.

```
Model 1: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                    coefficient    std. error       z        p-value
  -----------------------------------------------------------------
  const             -3.92076       0.245354      -15.98     1.76e-057 ***
  CreditScore       -0.000668329   0.000280345    -2.384    0.0171    **
  Age                0.0727060     0.00257551     28.23     2.52e-175 ***
  Tenure            -0.0159491     0.00935487     -1.705    0.0882    *
  Balance            2.63707e-06   5.14213e-07     5.128    2.92e-07  ***
  NumOfProducts     -0.101523      0.0471342      -2.154    0.0312    **
  HasCrCard         -0.0446764     0.0593395      -0.7529   0.4515
  IsActiveMember    -1.07544       0.0576856     -18.64     1.43e-077 ***
  EstimatedSalary    4.80699e-07   4.73663e-07     1.015    0.3102
  Female             0.528483      0.0544884       9.699    3.04e-022 ***
  Spain              0.0352178     0.0706379       0.4986   0.6181
  Germany            0.774714      0.0676740      11.45     2.41e-030 ***

Mean dependent var   0.203700   S.D. dependent var   0.402769
McFadden R-squared   0.153161   Adjusted R-squared   0.150787
Log-likelihood      -4280.678   Akaike criterion     8585.355
Schwarz criterion    8671.879   Hannan-Quinn         8614.643

Number of cases 'correctly predicted' = 8103 (81.0%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(11) = 1548.43 [0.0000]

           Predicted
              0      1
  Actual 0  7666    297
         1  1600    437

Excluding the constant, p-value was highest for variable 18 (Spain)
```

Apply Backward Elimination (see MLR)

P-value of Spain was Highest and (>5%), so I am deleting it. We get :

```
Model 2: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                    coefficient    std. error       z        p-value
  -----------------------------------------------------------------
  const             -3.91097       0.244526      -15.99     1.41e-057 ***
  CreditScore       -0.000666615   0.000280294    -2.378    0.0174    **
  Age                0.0727230     0.00257536     28.24     2.00e-175 ***
  Tenure            -0.0159766     0.00935423     -1.708    0.0876    *
  Balance            2.63733e-06   5.14201e-07     5.129    2.91e-07  ***
  NumOfProducts     -0.101288      0.0471276      -2.149    0.0316    **
  HasCrCard         -0.0449303     0.0593378      -0.7572   0.4489
  IsActiveMember    -1.07519       0.0576828     -18.64     1.53e-077 ***
  EstimatedSalary    4.81342e-07   4.73649e-07     1.016    0.3095
  Female             0.528343      0.0544870       9.697    3.11e-022 ***
  Germany            0.762937      0.0633614      12.04     2.16e-033 ***

Mean dependent var   0.203700   S.D. dependent var   0.402769
McFadden R-squared   0.153137   Adjusted R-squared   0.150961
Log-likelihood      -4280.802   Akaike criterion     8583.603
Schwarz criterion    8662.917   Hannan-Quinn         8610.451

Number of cases 'correctly predicted' = 8100 (81.0%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(10) = 1548.18 [0.0000]

           Predicted
              0      1
  Actual 0  7665    298
         1  1602    435

Excluding the constant, p-value was highest for variable 11 (HasCrCard)
```

P-value of HasCrCard was Highest and (>5%), so I am deleting it. We get:

```
Model 3: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                    coefficient    std. error       z       p-value
  ------------------------------------------------------------------
  const            -3.94435       0.240579     -16.40    2.07e-060 ***
  CreditScore      -0.000664033   0.000280270   -2.369   0.0178    **
  Age               0.0727303     0.00257516    28.24    1.73e-175 ***
  Tenure           -0.0161505     0.00935127    -1.727   0.0842    *
  Balance           2.64543e-06   5.14070e-07    5.146   2.66e-07  ***
  NumOfProducts    -0.101333      0.0471228     -2.150   0.0315    **
  IsActiveMember   -1.07438       0.0576668    -18.63    1.81e-077 ***
  EstimatedSalary   4.81783e-07   4.73661e-07    1.017   0.3091
  Female            0.528489      0.0544853      9.700   3.02e-022 ***
  Germany           0.761879      0.0633445     12.03    2.55e-033 ***

Mean dependent var    0.203700   S.D. dependent var   0.402769
McFadden R-squared    0.153080   Adjusted R-squared   0.151102
Log-likelihood       -4281.088   Akaike criterion     8582.175
Schwarz criterion     8654.279   Hannan-Quinn         8606.582

Number of cases 'correctly predicted' = 8111 (81.1%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(9) = 1547.61 [0.0000]

           Predicted
              0      1
  Actual 0  7673    290
         1  1599    438

Excluding the constant, p-value was highest for variable 13 (EstimatedSalary)
```

P-value of EstimatedSalary was Highest and (>5%), so I am deleting it. We get:

```
Model 4: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                    coefficient    std. error       z       p-value
  ------------------------------------------------------------------
  const            -3.89591       0.235717     -16.53    2.31e-061 ***
  CreditScore      -0.000666426   0.000280263   -2.378   0.0174    **
  Age               0.0727016     0.00257462    28.24    2.01e-175 ***
  Tenure           -0.0159836     0.00934933    -1.710   0.0873    *
  Balance           2.65326e-06   5.13979e-07    5.162   2.44e-07  ***
  NumOfProducts    -0.100475      0.0471176     -2.132   0.0330    **
  IsActiveMember   -1.07509       0.0576636    -18.64    1.41e-077 ***
  Female            0.528981      0.0544804      9.710   2.74e-022 ***
  Germany           0.762059      0.0633400     12.03    2.43e-033 ***

Mean dependent var    0.203700   S.D. dependent var   0.402769
McFadden R-squared    0.152978   Adjusted R-squared   0.151197
Log-likelihood       -4281.605   Akaike criterion     8581.210
Schwarz criterion     8646.103   Hannan-Quinn         8603.176

Number of cases 'correctly predicted' = 8115 (81.2%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(8) = 1546.57 [0.0000]

           Predicted
              0      1
  Actual 0  7676    287
         1  1598    439
```

<u>Remarque:</u> la (R^2) augmente au fur a mesure, ce qui est prouve que les modèles prédictifs s'améliorent à fur à mesure. Cela signifie qu'on n'a pas retiré à tort une variable.

→We have decided to stop at this step because all independent variables are significant.

Transformer les variables indépendantes .

On remplace la variable Balance par log_Balance= ln(Balance+1). We get :

```
Model 5: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                     coefficient    std. error       z       p-value
  ---------------------------------------------------------------
  const              -3.91258       0.237164      -16.50    3.84e-061 ***
  CreditScore        -0.000674866   0.000280272    -2.408   0.0160    **
  Age                 0.0726550     0.00257451     28.22    3.24e-175 ***
  Tenure             -0.0158791     0.00934627     -1.699   0.0893    *
  NumOfProducts      -0.0950198     0.0475374      -1.999   0.0456    **
  IsActiveMember     -1.07578       0.0576458     -18.66    1.01e-077 ***
  Female              0.526721      0.0544591       9.672   3.97e-022 ***
  Germany             0.747595      0.0650515      11.49    1.44e-030 ***
  log_Balance         0.0690263     0.0139592       4.945   7.62e-07  ***

Mean dependent var    0.203700   S.D. dependent var    0.402769
McFadden R-squared    0.152787   Adjusted R-squared    0.151006
Log-likelihood       -4282.570   Akaike criterion      8583.141
Schwarz criterion     8648.034   Hannan-Quinn          8605.107

Number of cases 'correctly predicted' = 8127 (81.3%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(8) = 1544.64 [0.0000]

           Predicted
               0      1
  Actual 0  7687    276
         1  1597    440
```

On ajoute la variable WealthAccumulation= Balance / Age au modèle précèdent, we get :

```
Model 6: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                      coefficient    std. error      z      p-value
  ---------------------------------------------------------------
  const               -3.82758       0.248202     -15.42   1.18e-053 ***
  CreditScore         -0.000675560   0.000280329   -2.410  0.0160    **
  Age                  0.0706681     0.00309455    22.84   2.00e-115 ***
  Tenure              -0.0159252     0.00934677    -1.704  0.0884    *
  NumOfProducts       -0.0955301     0.0475596     -2.009  0.0446    **
  IsActiveMember      -1.07339       0.0576722    -18.61   2.57e-077 ***
  Female               0.525712      0.0544733      9.651  4.88e-022 ***
  Germany              0.746337      0.0651330     11.46   2.13e-030 ***
  log_Balance          0.0950938     0.0266187      3.572  0.0004    ***
  WealthAccumulati~   -4.33552e-05   3.77862e-05   -1.147  0.2512

Mean dependent var    0.203700   S.D. dependent var    0.402769
McFadden R-squared    0.152918   Adjusted R-squared    0.150940
Log-likelihood       -4281.908   Akaike criterion      8583.815
Schwarz criterion     8655.919   Hannan-Quinn          8608.222

Number of cases 'correctly predicted' = 8123 (81.2%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(9) = 1545.97 [0.0000]

           Predicted
               0      1
  Actual 0  7684    279
         1  1598    439

Excluding the constant, p-value was highest for variable 21 (WealthAccumulation)
```

Remarque : WeilthAccumulation n'est pas significative et il n'y pas d'amélioration du modèle (la diminution de R^2). Cela est peut-être dû au fait que weilthAccumulation(Balance / Age) est lié a Age et Balance(i.e. existence de colinéarité entre weilthAccumulation, Age, log_Balance).

Vérification de multi colinéarité en utilisant VIF.

```
Variance Inflation Factors

Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

       CreditScore    1.001
               Age    1.450
     NumOfProducts    1.152
    IsActiveMember    1.011
            Female    1.003
           Germany    1.271
            Tenure    1.001
       Log_Balance    5.860
WealthAccumulation    5.722

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation coefficient
between variable j and the other independent variables
```

Le VIF de log_Balance et WealthAccumulation est supérieur aux autres variables, on décide alors de retirer log_Balance. We get :

```
Model 7: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                      coefficient    std. error       z       p-value
  -----------------------------------------------------------------
  const               -3.93393       0.246385      -15.97    2.18e-057 ***
  CreditScore         -0.000671869   0.000280046    -2.399   0.0164    **
  Age                  0.0758300     0.00274955     27.58    1.98e-167 ***
  Tenure              -0.0158045     0.00934123     -1.692   0.0907    *
  NumOfProducts       -0.121038      0.0471074      -2.569   0.0102    **
  IsActiveMember      -1.07881       0.0576645     -18.71    4.23e-078 ***
  Female               0.526299      0.0544270       9.670   4.05e-022 ***
  Germany              0.808180      0.0629285      12.84    9.44e-038 ***
  WealthAccumulati~    7.07501e-05   1.94600e-05     3.636   0.0003    ***

Mean dependent var   0.203700   S.D. dependent var   0.402769
McFadden R-squared   0.151650   Adjusted R-squared   0.149869
Log-likelihood      -4288.318   Akaike criterion     8594.635
Schwarz criterion    8659.528   Hannan-Quinn         8616.601

Number of cases 'correctly predicted' = 8121 (81.2%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(8) = 1533.15 [0.0000]

           Predicted
              0      1
  Actual 0  7685    278
         1  1601    436
```

En enlevant log_Balance, On obtient que WeilthAccumulation devient très significative et son coefficient est même devenu positif. Revérifions la multi colinéarité :

```
Variance Inflation Factors

Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

       CreditScore    1.001
               Age    1.115
     NumOfProducts    1.118
    IsActiveMember    1.011
            Female    1.003
           Germany    1.187
            Tenure    1.001
WealthAccumulation    1.387

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation coefficient
between variable j and the other independent variables
```

Aucun signe de multi colinéarité flagrant.

On décide finalement pour être plus cohérant avec notre étude, nous allons enlever WeilthAccumulation et garder log_Balance. We get :

```
Model 9: Logit, using observations 1-10000
Dependent variable: Exited
Standard errors based on Hessian

                   coefficient    std. error       z       p-value
  -----------------------------------------------------------------
  const            -3.91258       0.237164      -16.50    3.84e-061 ***
  CreditScore      -0.000674866   0.000280272    -2.408   0.0160    **
  Age               0.0726550     0.00257451     28.22    3.24e-175 ***
  Tenure           -0.0158791     0.00934627     -1.699   0.0893    *
  NumOfProducts    -0.0950198     0.0475374      -1.999   0.0456    **
  IsActiveMember   -1.07578       0.0576458     -18.66    1.01e-077 ***
  Female            0.526721      0.0544591       9.672   3.97e-022 ***
  Germany           0.747595      0.0650515      11.49    1.44e-030 ***
  log_Balance       0.0690263     0.0139592       4.945   7.62e-07  ***

Mean dependent var   0.203700   S.D. dependent var   0.402769
McFadden R-squared   0.152787   Adjusted R-squared   0.151006
Log-likelihood      -4282.570   Akaike criterion     8583.141
Schwarz criterion    8648.034   Hannan-Quinn         8605.107

Number of cases 'correctly predicted' = 8127 (81.3%)
f(beta'x) at mean of independent vars = 0.135
Likelihood ratio test: Chi-square(8) = 1544.64 [0.0000]

           Predicted
               0      1
  Actual 0  7687    276
         1  1597    440
```

<u>Vérification de multi colinéarité par la matrice de corrélation :</u>

Regardons la matrice de corrélation des variables Age, log_Balance, WealthAccumulation, log_WA :

```
Correlation Coefficients, using the observations 1 - 10000
Two-tailed critical values for n = 10000: 5% 0.0196, 1% 0.0258

             Age     log_Balance WealthAccumula~          log_WA
          1.0000          0.0345         -0.2463         -0.0075 Age
                          1.0000          0.8651          0.9984 log_Balance
                                          1.0000          0.8889 WealthAccumula~
                                                          1.0000 log_WA
```

Astuce : Plus la valeur absolue des coefficients de la matrice(coeff) se rapproche de 1, plus cela montre la colinéarité entre les variables.

- Si coeff > 0.9 => Très corrélées (Doit retirer une variable)
- Si coeff>0.7 => Très corrélées (recommande de faire qq chose)
- Si 0.3<=coeff<0.5 => corrélation modérée (essayer d'enlever une variable pour voir)
- Si 0<coeff<0.3 => Faible corrélation (on laisse les variables)

Par exemple dans notre cas, on a :

- log_WA et log_Balance sont très corrélées(on retire une variable).
- WealthAccumulation et log_Balance sont très corrélées(on retire une variable).
- log_WA et WealthAccumulation sont très corrélées(on retire une variable).
- log_WA et Age ne sont pas très corrélées.
- log_Balance et Age ne sont pas très corrélées.
- WealthAccumulation et Age ne sont pas très corrélées.

Dans cette section nous avons appris:

1. Ce qu'est la segmentation géo-démographique
2. Comment construire un VRAI modèle de segmentation
3. Comment transformer des variables indépendantes
4. Comment créer des variables dérivées (nouvelles VIs)
5. L'intuition derrière les colinéarités
6. Comment vérifier les colinéarités en utilisant les VIFs
7. Comment lire une matrice de corrélation