**Building a Robust Geodemographic Segmentation Model**

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

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquementOn a dummify les variables Gender et Geography vu qu’elles sont des variables catégorielles.

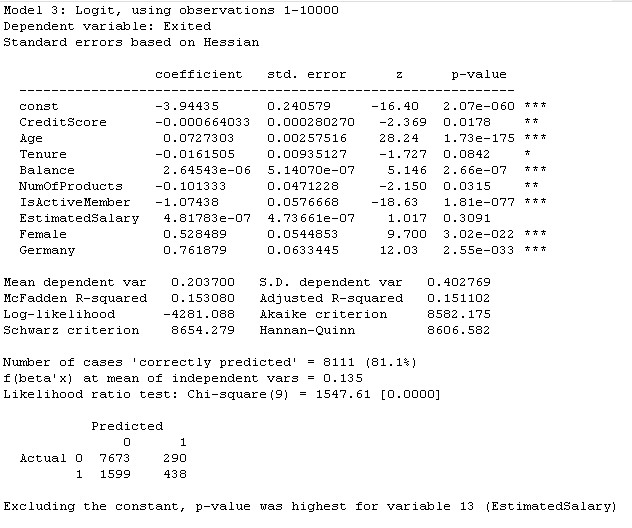
**Apply Backward Elimination ( see MLR)**

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

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

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



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

Une image contenant texte, capture d’écran, Police, nombre

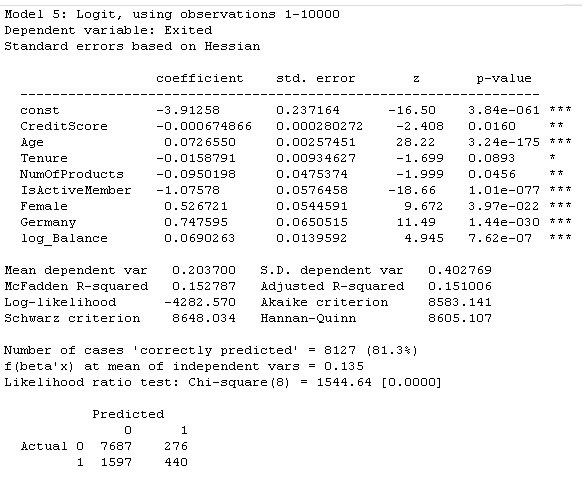
Description générée automatiquement

Remarque: 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 :



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

Une image contenant texte, capture d’écran, Police, document

Description générée automatiquement

**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.**

Une image contenant texte, reçu, capture d’écran

Description générée automatiquementLe VIF de log\_Balance et WealthAccumulation est supérieur aux autres variables, on décide alors de retirer log\_Balance. We get :

Une image contenant texte, capture d’écran, Police, menu

Description générée automatiquement

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

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

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 :

Une image contenant texte, capture d’écran, Police, document

Description générée automatiquement

**Vérification de multi colinéarité par la matrice de corrélation :**

Regardons la matrice de corrélation des variables Age, log\_Balance, WealthAccumulation, log\_WA :

Une image contenant texte, capture d’écran, reçu, Police

Description générée automatiquement

**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.

Une image contenant texte, Police, capture d’écran, blanc

Description générée automatiquement