
# Modélisation prédictive du défaut de paiement en gestion du risque de crédit

Lamarana DIALLO

# Sommaire

# Introduction

Le projet consiste à modéliser des données sur des détenteurs de cartes de crédit d'une institution financière Nord- Américaine. L'objectif principal de la modélisation consiste à estimer la probabilité qu'un détenteur d'une carte de crédit ne parvient pas à payer son solde dû (défaut de paiement) dans un futur horizon de 12 mois après la date de collecte des différentes caractéristiques du modèle. Un client est considéré être en défaut de paiement si l'une au moins des conditions suivantes est vérifiée pendant la période visée (12 mois après une date donnée qui a servi à collecter les variables):

- Retard de paiement de 90 jours ou plus ;
- Faillite
- Radiation ;

Il faut mentionner que c'est une modélisation d'une clientèle déjà détentrice de cartes de crédit (l'institution financière dispose assez d'informations en termes de comportement : ce sont les différentes variables décrivant cette clientèle). Ainsi, la modélisation consistera à prédire le comportement d'un détenteur dans un futur horizon de 12 mois en se basant sur son comportement actuel et passé. Le modèle résultant est donc un modèle comportemental qui permettra de suivre, dans le temps, le comportement des détenteurs de cartes de crédit en termes de capacités de remboursement du solde dû à la date exigée par l'institution financière.

# Présentation des données et échantillonage

## Présentation des données

Dans cette section, nous présentons une vue d'ensemble des données. Notre table de données des détenteurs de cartes de crédit contient au total 24877 observations (détenteurs) décrits par 33 variables (30 variables quantitatives et 3 variables qualitatives) représentant :

- les montants des achats et des avances de fonds ;
- les montants des soldes ;
- les montants des paiements effectués sur le solde de la carte ;
- les ratios d'utilisation de la carte ;
- les retards de paiements du solde de la carte ;
- quelques informations sociodémographiques (âge, statut matrimonial, ...);
- etc.

Les variables présentant les informations sociodémographiques sont extraites en date d'observation (informations en date d'extraction des données). Toutes les autres variables ont été considérées suivant plusieurs historiques : mois courant (date d'extraction des données), 6 derniers mois et 12 derniers mois avant la date d'extraction des données. Ainsi, cela permet d'avoir des informations du « Présent » (mois courant) et du « Passé » (historiques).

Dans le cadre de ce travail, l'horizon de prédiction considérée est de 12 mois après la date d'observation des différentes caractéristiques d'un client. Ainsi, la variable à modéliser est construite en tenant compte de cet horizon de prédiction. La variable à modéliser est une variable binaire ayant deux(2) modalités:

- **0** : le client n'a jamais été en défaut de paiement du solde de sa carte de crédit pendant toute la période des 12 mois après la date d'extraction des données ;
- **1** : le client a été en défaut de paiement au moins une fois dans la période des 12 mois après la date d'extraction des données.

| **Valeurs réelles** | **Nombre** | **Pourcentage** |
|---|---|---|
| 0 | 24167 | 97.15% |
| 1 | 710 | 2.85% |
| **Total** | 24877 | 100% |

Table 1.1: Variable `defaut_paiement`

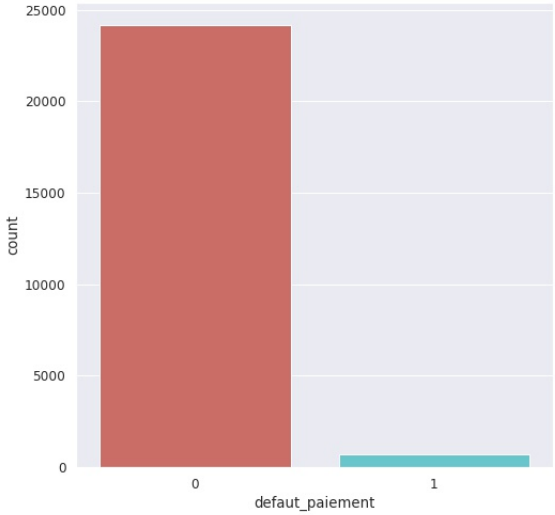


Figure 1.1: Distribution variable `défaut de paiement`

Ces résultats montrent que cette clientèle présente un taux de défaut de paiement de **2.85%** (évènement rare !).

## Exploration des variables qualitatives

### Statut compte

| **Valeurs réelles** | **Nombre** | **Pourcentage** |
|---|---|---|
| **O** | 23501 | 94.5% |
| **P** | 1376 | 5.5% |

Table 1.2: Variable `statut_compte`

**Statut matrimoniale client**

| Valeurs réelles | Nombre | Pourcentage |
|---|---|---|
| **0** | 1227 | 4.93% |
| **1** | 4194 | 16.86% |
| **2** | 2035 | 8.18% |
| **3** | 191 | 0.77% |
| **4** | 450 | 1.81% |
| **5** | 1430 | 5.75 |
| **Manquantes** | 15350 | 61.7% |

Table 1.3: Variable `statut_matrimonial_client`

**Type de résidence**

| Valeurs réelles | Nombre | Pourcentage |
|---|---|---|
| **O** | 3912 | 15.73% |
| **P** | 4282 | 17.21% |
| **R** | 2269 | 9.12% |
| **Manquantes** | 14414 | 57.94% |

Table 1.4: Variable `type_residence`

## Exploration des variables quantitatives

| Variable | Manquantes | Minimum | Médiane | Moyenne | Maximum |
|---|---|---|---|---|---|
| Anciennete_compte | 0 | 0 | 3 | 3.14 | 6 |
| retard_courant | 0 | 0 | 0 | 0.05 | 3 |
| retard_min_6derniers_mois | 20 | -997 | 0 | -137.01 | 1 |
| retard_max_6derniers_mois | 20 | -997 | 0 | -136.94 | 3 |
| retard_min_12derniers_mois | 20 | -997 | 0 | -137.01 | 1 |
| retard_max_12derniers_mois | 20 | -997 | 0 | -136.94 | 3 |
| solde_courant | 0 | -5121.71 | 0 | 287.38 | 19944.64 |
| solde_courant_pct_max1to6m | 0 | -999 | -999 | -625.161 | 342448.2 |
| solde_courant_pct_max1to12m | 0 | -999 | -999 | -625.16 | 342448.2 |
| solde_courant_pct_moy1to6m | 0 | $-6.10^4$ | $-9.10^2$ | $4.10^2$ | $2.10^7$ |
| solde_courant_pct_moy1to12m | 0 | $-6.10^4$ | $-9.10^2$ | $4.10^2$ | $2.10^7$ |

| | | | | | |
|---|---|---|---|---|---|
| pct_utilisation_courant | 0 | -506.57 | 0 | 14.93 | 549.5 |
| pct_utilisation_min_1to6m | 20 | -9988 | 0 | -135.3 | 218.06 |
| pct_utilisation_min_1to12m | 20 | -9988 | 0 | -135.3 | 218.06 |
| pct_utilisation_max_1to6m | 20 | -997 | 0 | -120.5 | 365.75 |
| pct_utilisation_max_1to12m | 20 | -997 | 0 | -120.5 | 365.75 |
| pct_utilisation_moy_1to6m | 0 | -2497 | 0 | -128.25 | 218.06 |
| pct_utilisation_moy_1to12m | 0 | -2497 | 0 | -128.25 | 218.06 |
| achats_courant | 0 | 0 | 0 | 167.14 | 21801.65 |
| achats_courant_pct_moy_1to6m | 0 | $-9.10^2$ | $-9.10^2$ | $-5.10^2$ | $1.10^6$ |
| achats_courant_pct_moy_1to12m | 0 | $-9.10^2$ | $-9.10^2$ | $-5.10^2$ | $1.37.10^6$ |
| avances_courant | 0 | 0 | 0 | 23.66 | 18807.94 |
| paiements_courant | 0 | 0 | 0 | 182 | 15107.59 |
| solde_courant_autres | 0 | 0 | 286.41 | 1272.38 | 48828.19 |
| solde_total_courant | 0 | -2708.39 | 755.91 | 1559.76 | 62843.81 |
| solde_courant_autres_pct_limite | 0 | -999 | 6.57 | -159.29 | 110672 |
| age_client | 14414 | 16 | 36 | 38.8 | 90 |
| score_actuel_compte | 0 | 34.26 | 990.37 | 980.64 | 999.3 |

Table 1.6: Exploration des variables quantitatives

La proportion des données manquantes est relativement marginale à l'exception de la variable `age_client`. Cette forte proportion de données manquantes au niveau de cette variable est probablement due à une lacune au niveau du croisement des données (en effet, le système de données contenant les informations sociodémographiques du client est différent de celui contenant les informations financières). Le traitement des données manquantes sera présenté ci-dessous selon la variable.

# Sélection de variables

Dans cette partie, nous allons faire une étude de corrélation entre variables. Cette analyse sera effectuée sur les différentes familles de variable qui sont :

- Famille de variables présentant les retards de paiement du solde de la carte de crédit;
- Famille de variables présentant les soldes;

- Famille de variables présentant les montants des achats effectués à l'aide de la carte de crédit;
- Famille de variables présentant les ratios d'utilisation de la carte de crédit.

## Famille des retards de paiement

L'analyse de corrélation de la famille de variables présentant les retards de paiement du solde de la carte de crédit permet d'avoir le graphiques suivants :

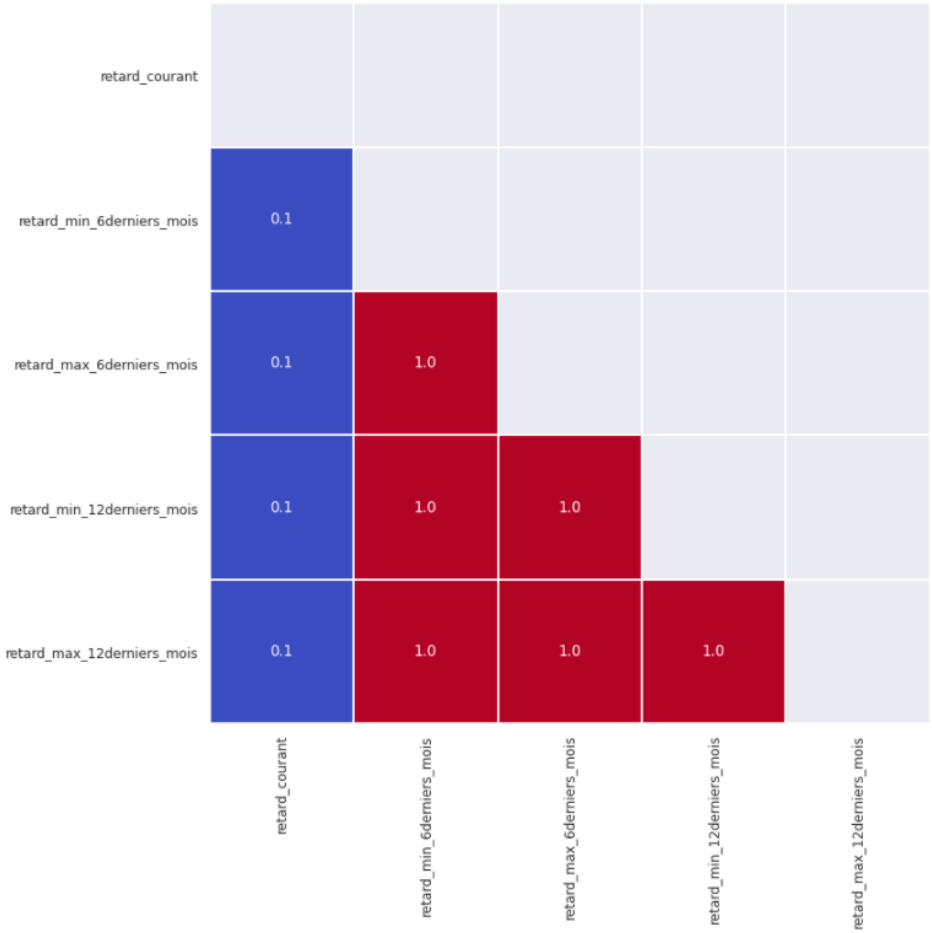


Figure 1.2: Corrélation famille des retards de paiement

Sur le graphique de corrélation, nous pouvons remarqué quatre variables qui parfaitement corrélées:

- retard_min_6derniers_mois ;
- retard_max_6derniers_mois ;
- retard_min_12derniers_mois ;
- retard_max_12derniers_mois.

Après analyse, nous avons choisi de sélectionner seulement la variable **retard_min_6derniers_mois** présentant le nombre minimum de jours de retard de paiement du solde de la carte de crédit au cours des 6 derniers mois précédant la date d'extraction des différentes variables.

## Famille des soldes

L'analyse de corrélation de cette famille de variables permet d'obtenir le graphique suivant :

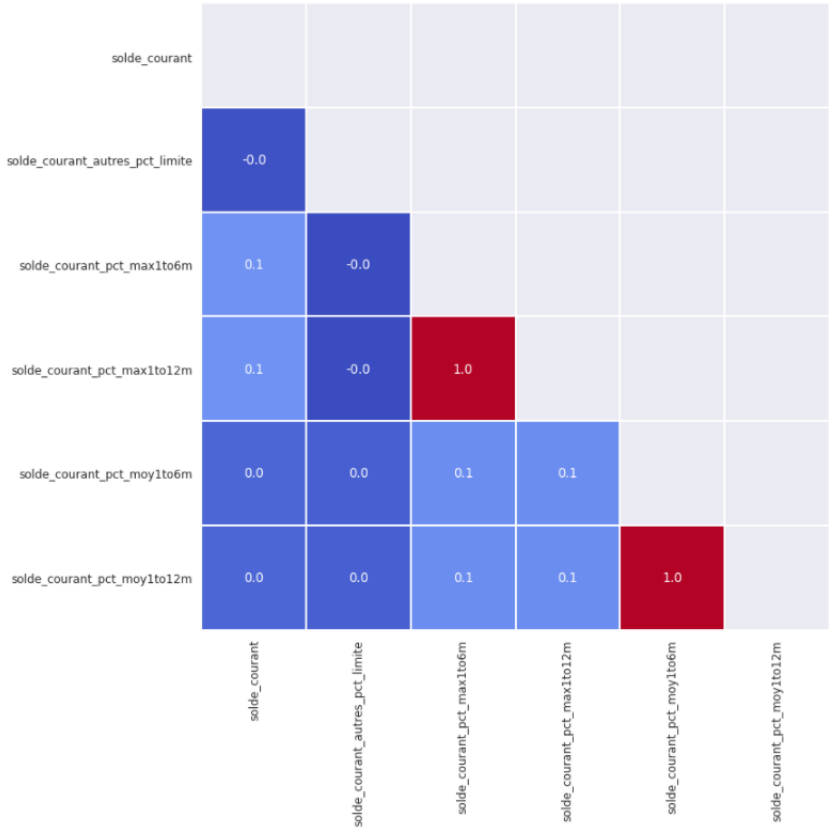


Figure 1.3: Corrélation famille des soldes

Le résultat montre un forte corrélation entre les variables **solde_courant_pct_max1to12m** et **solde_courant_pct_max1to6m**, mais entre les variables **solde_courant_pct_moy1to6m** et **solde_courant_pct_moy1to12m**. Dans cette famille de variable, nous allons conserver les variables `solde_courant_pct_moy1to6m` et `solde_courant_pct_max1to12m` ainsi que les autres variables non corrélées.

## Famille des achats

L'analyse de cette famille donne les résultats suivants :

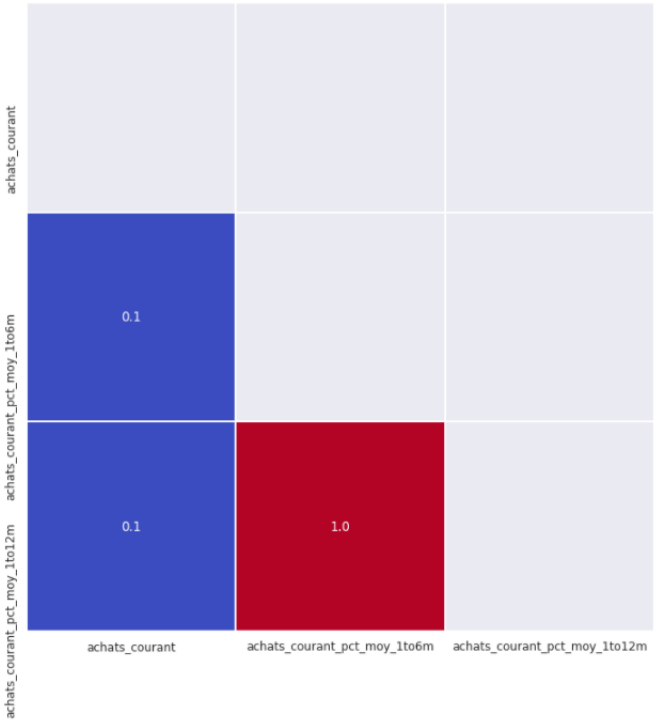


Figure 1.4: Corrélation famille des achats

Dans ce graphique de corrélation, on voit que les variables **`achats_courant_pct_moy_1to6m`** et **`achats_courant_pct_moy_1to12m`**. Nous allons garder la variable `achats_courant_pct_moy_1to6m` en plus de la variable `achats_courant` qui n'est pas en corrélation avec les autres variables de la famille.

## Famille des ratios d'utilisations de la carte

Lors de l'analyse de corrélation de la famille des ratios d'utilisations de la carte de crédit, nous avons remarqué que les variables sont parfaitement corrélées comme on peut le voir sur le graphique suivant.

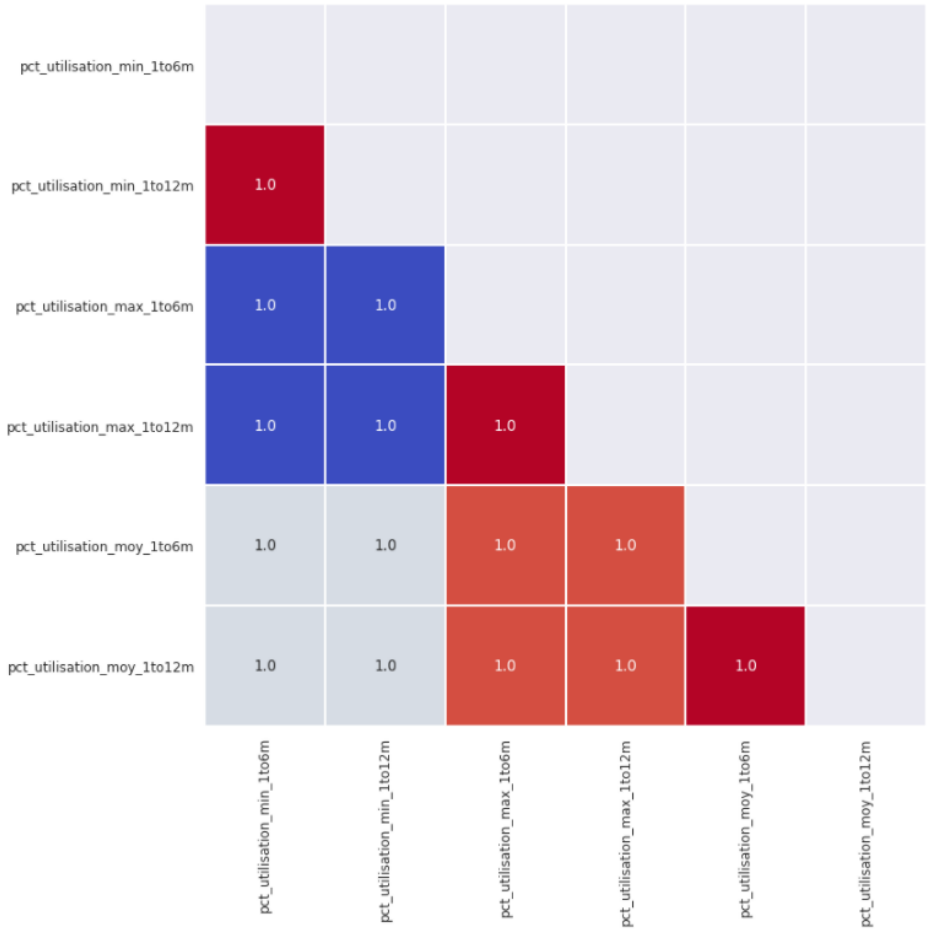


Figure 1.5: Corrélation famille des ratio d'utilisations de la carte

En premier lieu, nous avons décidé d'enlever toute les variables portant sur un historique de 12 mois. Puis, après une seconde analyse de corrélation, nous avons décidé d'enlever les variables `pct_utilisation_min_1to6m` et `pct_utilisation_max_1to6m` et de sélectionner la variable `pct_utilisation_moy_1to6m` (moyenne sur le ratio d'utilisation au cours des 6 derniers mois) car celle-ci est plus informative (elle intègre les variables `pct_utilisation_min_1to6m` et `pct_utilisation_max_1to6m`).

## Corrélation entre variables sélectionnées

Une nouvelle étude de corrélation entre les différentes variables sélectionnées permet d'avoir le graphique ci-dessous.

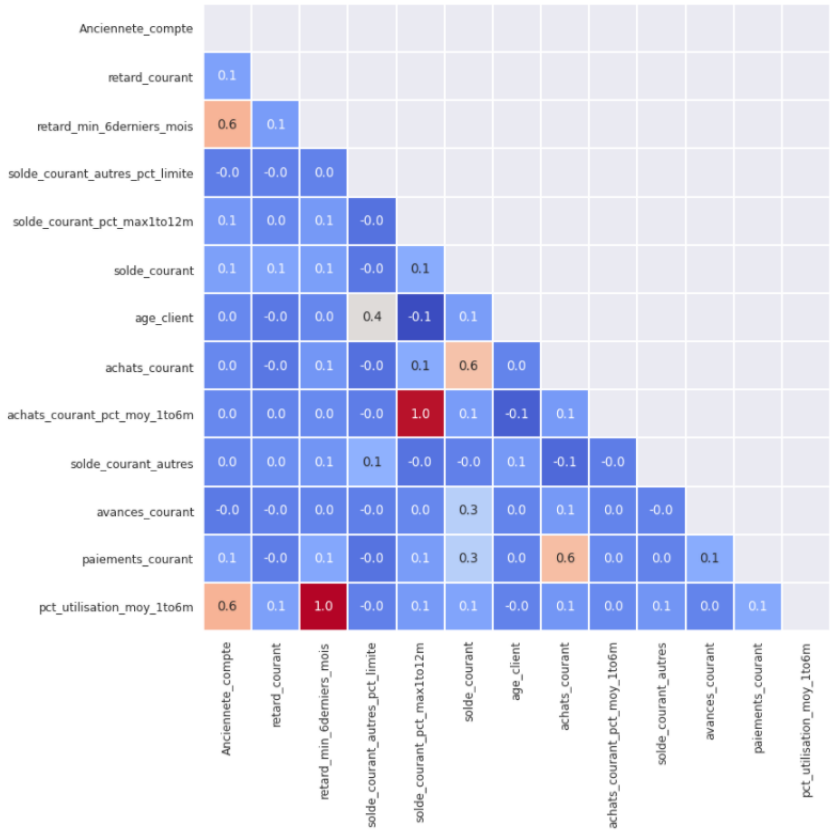


Figure 1.6: Corrélation des variables sélectionnées

Le graphique montre que les variables `achats_courant_pct_moy_1to6m` et `solde_courant_pct_max1to12m` sont parfaitement corrélées. De même, les variables `retard_min_6derniers_mois` et `pct_utilisation_moy_1to6m` sont parfaitement corrélées aussi. Nous allons conserver les variables *`achats_courant_pct_moy_1to6m`* et *`pct_utilisation_moy_1to6m`* pour la suite du projet.

## Catégorisation des prédicteurs

Certains algorithmes de modélisation (comme la régression logistique) recommandent fortement de procéder à une catégorisation des variables candidates à l'estimation du modèle. La catégorisation consiste à :

- regrouper des modalités d'une variable qualitative : le regroupement peut se faire selon le profil des différentes modalités (ressemblance du taux de la variable à modéliser au niveau de ces modalités) ou pour des raisons d'affaires (regroupement selon des stratégies d'affaires d'une organisation) ;

- discrétiser une variable quantitative (selon la cible ou raisons d'affaires, pour les mêmes raisons que ci-dessus).

Pour la discrétisation des variables quantitatives, une relation monotone croissante ou décroissante avec la variable à modéliser est généralement recherchée pour une *logique d'interprétation*. Des relations du type $\bigcap$ ou $\bigcup$ sont aussi acceptées car elles indiquent une monotonie décroissante ou croissante jusqu'à un certain niveau (jusqu'à une certaine modalité) et la situation s'inverse pour la suite. Il faut préciser que les modalités obtenues après la discrétisation d'une variable quantitative respectent un certain ordre. Cependant, ce type de relation n'est pas recherché pour une variable qualitative car il n'y a pas d'ordre au niveau des modalités.

## Traitement de données manquantes

Dans le cadre de ce projet, les données manquantes ont été analysées cas par cas. Selon leur profil relativement à la variable à modéliser(ici `défaut de paiement`) et leur proportion, nous les avons isolé dans une catégorie à part ou regroupement avec une autre catégorie si la ressemblance du profil de défaut de paiement (taux de défaut) le justifie.

## Regroupement de modalités des variables qualitatives

### Variable **statut_compte**

Cette variable présente deux modalités ayant un profil distinct (voir tableau ci-dessous avec le croisement avec la variable à modéliser). Elle présente le statut du compte du client (**O** : *compte opérationnel* et **P** *compte potentiellement opérationnel*). Elle n'a aucune donnée manquante. Cette catégorisation a été maintenue (`catégorie 1`= **O** et `catégorie 2` = **P**).

| | | **Défaut paiement** | | **Total** |
|---|---|---|---|---|
| | | 0 | 1 | |
| **statut compte** | O | 22990 | 511 | 23501 |
| | P | 1177 | 199 | 1376 |
| **Total** | | 24167 | 710 | 24877 |

Table 1.7: croisement des variables `statut_compte` et `defaut_paiement`

Sur ce tableau, on observe que le taux de défaut est plus élevé au niveau des comptes potentiellement opérationnels (`14.46%`) contre `2.17%` pour les comptes opérationnels.

**Variable `type_residence_client`**

Cette variable a trois modalités ayant un profil distinct (voir tableau ci-dessous avec le croisement avec la variable à modéliser). Elle présente le type de résidence du client. Elle enregistre une proportion de données manquantes dépassant les 50% (le client refuse de déclarer son type de résidence et l'institution financière n'exige pas cette information). Ainsi, les données manquantes ont été isolées dans une catégorie à part.

| | | **Défaut paiement** | | **Total** |
|---|---|---|---|---|
| | | 0 | 1 | Total |
| **type résidence client** | O | 3828 | 84 | 3912 |
| | P | 4174 | 108 | 4282 |
| | R | 2154 | 115 | 2269 |
| | Manquantes | 14011 | 403 | 14414 |
| **Total** | | 24167 | 710 | 24877 |

Table 1.8: croisement des variables `type_residence_client` et `defaut_paiement`

**Variable `statut_matrimoniale_client`**

C'est une variable qualitative qui a plusieurs modalités ayant un profil distinct (voir tableau ci-dessous avec le croisement avec la variable à modéliser). Elle présente le statut matrimonial du client (Marié, Célibataire, Conjoint de fait, ...). Elle enregistre une proportion de données manquantes dépassant les *60%* (le client refuse de déclarer son statut matrimonial et l'institution financière n'exige pas cette information). Ainsi, les données manquantes ont été isolées dans une catégorie à part.

| | | **Défaut paiement** | | **Total** |
|---|---|---|---|---|
| | | **0** | **1** | |
| **statut matrimoniale** | 0 | 1186 | 41 | 1227 |
| | 1 | 4041 | 153 | 4194 |
| | 2 | 2002 | 33 | 2035 |
| | 3 | 188 | 3 | 191 |
| | 4 | 438 | 12 | 450 |
| | 5 | 1387 | 43 | 1430 |
| | Manquantes | 14925 | 425 | 15350 |
| **Total** | | 24167 | 710 | 24877 |

Table 1.9: croisement des variables `statut_matrimonial_client` et `defaut_paiement`

## Discrétisation des variables quantitatives

Dans cette partie, nous présentons la discrétisation des variables quantitatives.

### Variable `retard_courant`

L'analyse du taux de défaut des modalités de la variable permet de les regrouper en deux classe : ceux qui n'ont jamais étaient en retard et ceux qui ont accusé au moins un retard dans leurs paiements.

| | | **Défaut paiement** | | **Total** |
|---|---|---|---|---|
| | | 0 | 1 | Total |
| **retard courant** | Catégorie 1 : $\leq 0$ | 24089 | 633 | 24722 |
| | Catégorie 2 : $> 0$ | 78 | 77 | 155 |
| **Total** | | 24167 | 710 | 24877 |

Table 1.10: croisement des variables `retard_courant` et `defaut_paiement`

### Variable `achats_courant`

L'analyse de la distribution de cette variable avec le *taux de défaut* permet de la catégoriser comme suit:

| | | **Défaut paiement** | | **Total** |
|---|---|---|---|---|
| | | 0 | 1 | |
| **achats courant** | Catégorie 1 : $\leq 98.58$ | 18831 | 551 | 19382 |
| | Catégorie 2 :]98.58 , 255.29] | 1773 | 59 | 1832 |
| | Catégorie 3 :]255.59 , 593.78] | 1779 | 52 | 1831 |
| | Catégorie 4 : $> 593.78$ | 1784 | 48 | 1832 |
| **Total** | | 24167 | 710 | 24877 |

Table 1.11: croisement des variables `achats_courant` et `defaut_paiement`

### Variable `paiements_courant`

L'analyse de la distribution de cette variable avec le taux de défaut donne le tableau suivant:

| | | Défaut paiement | | Total |
|---|---|---|---|---|
| | | 0 | 1 | |
| **paiements courant** | Catégorie 1 : $\leq 80$ | 16019 | 481 | 16500 |
| | Catégorie 2 :]80 , 169.21] | 2658 | 81 | 2739 |
| | Catégorie 3 :]169.21 , 399.91] | 2727 | 92 | 2819 |
| | Catégorie 4 : $> 399.91$ | 2763 | 56 | 2819 |
| **Total** | | 24167 | 710 | 24877 |

Nous avons procédé de la même manière avec les autres variables.

# Modélisation et analyse des résultats

## Méthode de modélisation utilisée

La régression logistique est très appropriée pour modéliser des réponses binaires du type «**Oui/Non**». Elle présente plusieurs avantages, parmi lesquels :

- presque pas d'hypothèses sur les données (ce qui est très pratique à cause du volume de données de plus en plus important et donc de plus en plus exigeant en termes de vérification de certaines hypothèses statistiques) ;
- coefficients estimés aisément interprétables;
- la discrétisation des variables quantitatives permet de contourner la problématiques des données extrêmes (les données extrêmes sont intégrées dans une catégorie et du coup leur effet devient moins sensible au modèle) et manquantes (les données manquantes peuvent être regroupées dans une catégorie à part si leur proportion est assez acceptable et si leur profil relativement à la variable réponse est très différent de celui des autres modalités de la variable considérée) ;
- l'équation du modèle est aisément transférable d'un logiciel à un autre pour les besoins de scoring sur de nouvelles données ;
- Etc.

## Échantillonnage

En modélisation prédictive, il est primordial d'effectuer l'échantillonnage. Ainsi nous séparons notre table principale en deux échantillons :

- **70%** pour l'échantillon d'apprentissage
- **30%** pour l'échantillon de test

Nous allons estimer le modèle sur les données d'apprentissage, puis évaluer la performance avec l'échantillon de test.

## Estimation du modèle

Comme indiqué précédemment, nous utilisons la régression logistique pour modéliser notre modèle. Nous avons *14 variables* candidates à l'estimation du modèle. L'estimation du modèle permet d'avoir les résultats résumés dans le tableau suivant:

| **Variables** | **catégories** | **Coefficients** |
|---|---|---|
| **statut_compte** | statut_compte_1 | -0.518511 |
| | statut_compte_2 | 0.518738 |
| **type_residence** | type_residence_1 | -0.270888 |
| | type_residence_2 | -0.103081 |
| | type_residence_4 | 0.374197 |
| **statut_matrimoniale** | statut_matrimonial_1 | 0.065567 |
| | statut_matrimonial_3 | -0.158473 |
| | statut_matrimonial_5 | 0.093133 |
| **retard_courant** | retard_courant_1 | -0.813538 |
| | retard_courant_2 | 0.813766 |
| **age_client** | age_client_1 | 0.234252 |
| | age_client_2 | -0.168252 |
| | age_client_3 | -0.451871 |
| | age_client_9 | 0.386099 |
| **anciennete_compte** | Anciennete_compte_1 | 0.243254 |
| | Anciennete_compte_2 | -0.085427 |
| | Anciennete_compte_4 | -0.157599 |
| **solde_courant_autres_pct_limites** | solde_cour_autres_pct_lmte_1 | -0.326695 |
| | solde_cour_autres_pct_lmte_2 | -0.355608 |
| | solde_cour_autres_pct_lmte_3 | 0.097321 |
| | solde_cour_autres_pct_lmte_4 | 0.585211 |
| **solde_courant** | solde_courant_1 | -0.809558 |
| | solde_courant_2 | -0.312736 |
| | solde_courant_3 | 0.393824 |

| | | |
|---|---|---|
| | solde_courant_4 | 0.728697 |
| **achats_courant** | achats_courant_1 | 0.026454 |
| | achats_courant_2 | 0.026120 |
| | achats_courant_3 | -0.052346 |
| **achats_courant_pct_moy_1to6m** | achats_cour_pct_moy_1 | 0.256414 |
| | achats_cour_pct_moy_2 | -0.276224 |
| | achats_cour_pct_moy_4 | 0.020038 |
| **solde_courant_autres** | solde_courant_autres_1 | 0.111851 |
| | solde_courant_autres_2 | 0.058708 |
| | solde_courant_autres_4 | -0.170331 |
| **avances_courant** | avances_courant_1 | -0.554083 |
| | avances_courant_2 | 0.283039 |
| | avances_courant_3 | 0.135634 |
| | avances_courant_4 | 0.135638 |
| **paiements_courant** | paiements_courant_1 | 0.416840 |
| | paiements_courant_3 | 0.014987 |
| | paiements_courant_4 | -0.431599 |
| **pct_utilisation_moy_1to6m** | pct_util_moy_1 | -0.431599 |
| | pct_util_moy_3 | -0.081084 |
| | pct_util_moy_4 | 0.462883 |

Table 1.13: Résultats estimation du modèle

La valeur de la constante(`intercept`) est égale à **-2.752**.

## Evaluation du modèle & sélection de variables

L'évaluation du modèle sur les données de test a permis d'avoir une précision (*`accuracy`*) de **97%**.

La sélection de variable avec la fonction **`RFECV`** par validation croisée de *`scikit learn`* permet d'avoir le graphique suivant.

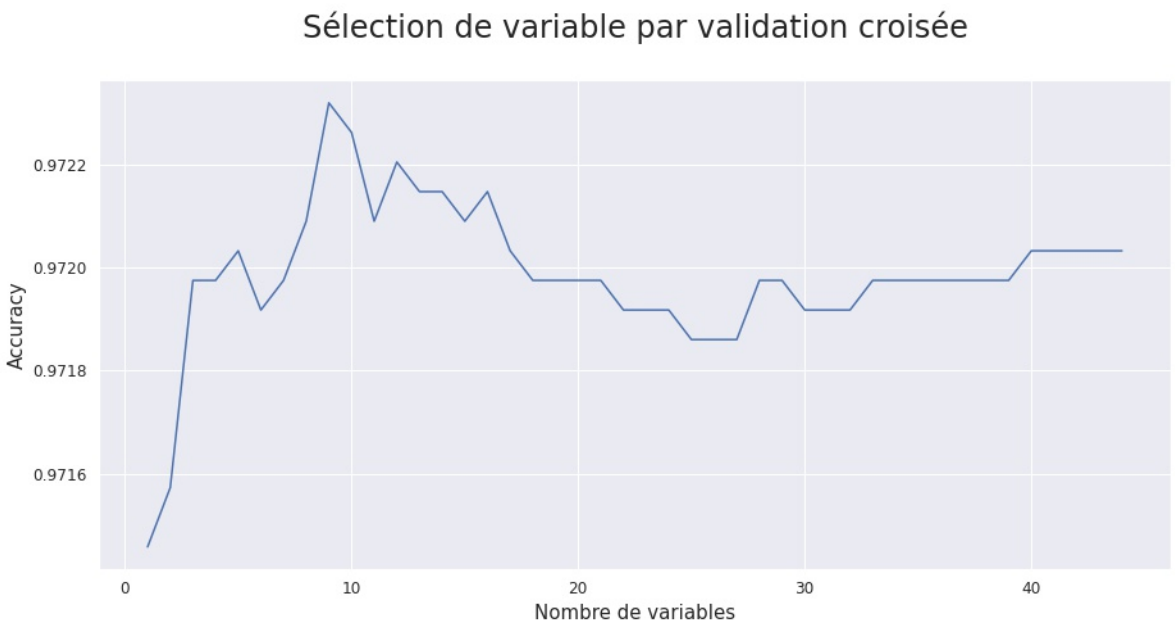


Figure 1.7: Sélection de variable par validation croisée

Comme le montre le graphique, l'algorithme de la sélection de variable retient **9 variables**. Une nouvelle estimation sur ces variables sélectionnées donne les résultats suivants.

| **Variables** | **catégories** | **coefficients** |
|---|---|---|
| statut_compte | statut_compte_2 | 1.234280 |
| retard_courant | retard_courant_1 | -0.877050 |
| | retard_courant_2 | 0.876939 |
| solde_courant_autres_pct_limite | solde_cour_autres_pct_lmte_4 | 0.762232 |
| solde_courant | solde_courant_3 | 0.934162 |
| | solde_courant_4 | 1.144191 |
| avances_courant | avances_courant_1 | -0.753689 |
| paiements_courant | paiements_courant_4 | -0.904899 |
| pct_util_moy | pct_util_moy_4 | 0.743853 |

Table 1.15: Estimation des variables sélectionnées

## Equation du modèle

À partir des coefficients estimés au niveau de chacune des modalités des différentes variables retenues du modèle final, il est aisé d'écrire l'équation du modèle. C'est cette équation qui va être

utilisée pour les besoins de *scoring* sur de nouvelles données afin de prédire la probabilité de faire défaut de paiement dans un futur horizon de 12 mois après la date d'observation des différentes caractéristiques (variables) du modèle.
En considérant $\beta_i$ le paramètre de la variable $x_i$, alors le score se calcul comme suit :

$$score = \beta_0 + \sum \beta_i x_i$$

où $\beta_0$ est la constante à l'origine.
La probabilité prédite est donnée par l'équation suivante.

$$proba = \frac{\exp^{score}}{1 + \exp^{score}}$$

Cette équation pourra être intégrée dans une application pour automatiser le calcul du score de chaque client et prédire la probabilité de défaut de paiement pour un nouveau client.

# Conclusion

Ce présent projet avait pour objectif de développer un modèle de Credit Scoring permettant de prédire la probabilité de défaut de paiement des nouveaux demandeurs de crédit. Pour cela, nous avons commencé par présenter les données sur lesquelles nous avons travaillé dans ce projet. L'étude de corrélation pour les différentes familles de variables a permis de faire des sélections. L'estimation du modèle a montré une précision intéressante sur l'échantillon de test, ce qui laisse présager que notre modèle répond bien sur de nouvelles données. L'équation du modèle permet de calculer le score de chaque client et prédire la probabilité du défaut de paiement pour un nouveau client. En somme, ce projet m'a permis de mettre en oeuvre mes connaissances sur la régression logistique avec la librairie *scikit-learn* de python.