# Proposition de sujet à la SEME : définition du risque de "défaut" des clients au crédit à la consommation

Adrien Ehrhardt

3 janvier 2018

**Résumé**

Le *Credit Scoring* est devenu une composante essentielle des institutions financières. Il s'agit d'évaluer la probabilité d'un client de faire défaut à un prêt. Plusieurs approches ont été proposées dans la littérature parmi lesquelles la régression logistique associée à une définition des bons et mauvais clients et qui a été adoptée par Crédit Agricole Consumer Finance (CA CF dans la suite). Cette méthode souffre de plusieurs défauts, parmi lesquelles une performance relativement médiocre, sans doute partiellement due à la définition bons/mauvais clients retenue. C'est sur cette définition que nous proposons d'orienter la réflexion dans ce sujet à destination de la Semaine d'Etude Maths-Entreprises.

## 1 Contexte

CA CF commercialise des crédits à la consommation sur différents pays grâce à ses filiales. Le crédit à la consommation couvre des biens de nature et de prix très variés : de l'automobile de luxe à la location de smartphones. En quelques mots, l'activité peut se découper :

- en canaux de distribution :
  - Circuit "long" : à travers des partenaires pour l'équipement de la maison (électroménager notamment) et l'automobile,
  - Courtage : autres établissements bancaires, enseignes de grande distribution, constructeurs automobiles, courtiers,
  - Circuit "court" : vente en marque propre (Finaref, Sofinco) par téléphone ou sur internet.
- en produits :
  - Prêt bancaire "classique" ou affecté à un bien,
  - Location avec option d'achat ou à longue durée,
  - Crédit renouvelable éventuellement associé à une carte de crédit,
  - Rachat de crédits.

| Travail | Logement | Durée d'emploi | Enfants | Status familial | Salaires | Score | Remboursement |
|---|---|---|---|---|---|---|---|
| Craftsman | Owner | 20 | 3 | Widower | 2000 | 225 | 0 |
| Technician | Renter | 10 | 1 | Common-law | 1700 | 190 | 0 |
| Licensed professional | Starter | 5 | 0 | Divorced | 4000 | 218 | 1 |
| Executive | By work | 8 | 2 | Single | 2700 | 202 | 1 |
| Office employee | Renter | 12 | 2 | Married | 1400 | 205 | 0 |
| Worker | By family | 2 | 0 | Single | 1200 | 192 | 0 |

TABLE 1 – Exemple simplifié d'une base de données de demandeurs de crédit à la consommation.

Pour chaque demandeur de crédit sur tous ces canaux et produits, nous collectons des données. La plupart de ces données sont demandées au client à travers un formulaire de souscription. Elles sont donc limitées (une vingtaine d'informations), déclaratives pour la plupart (bien que, suivant les montants, vérifiées grâce à des pièces justificatives) et socio-démographiques comme on peut le voir dans le tableau 1.

Dans la suite, ces informations, en nombre $d$, sont représentées par le vecteur aléatoire $X = (X^1, \ldots, X^d)$ de $\mathcal{X}$. On fera également référence au "remboursement" du client, que l'on définit plus tard dans le document, par la variable aléatoire $Y$ de $\mathcal{Y}$.

# 2 Modélisation actuelle

On décrit ici l'ensemble des étapes menant à la construction d'un score de crédit, c'est-à-dire que l'on suppose l'existence d'une fonction $f : \mathcal{X} \to \mathcal{Y}$ t.q. $f(X) = Y$ que l'on se propose d'estimer.

## 2.1 Collecte des données

Les données $X$ sont collectées une première fois à la saisie de la demande de crédit et peuvent être modifiées jusqu'à la production du dossier (le dossier a été accepté et les fonds sont disponibles pour le client). Toutefois, comme le score est calculé au moment de la demande, on ne considère pas cette dépendance temporelle de $X$.

Concernant les crédits renouvelables, on considère uniquement les clients ayant réalisé une utilisation à crédit, et on extrait leurs informations à la date de cette première utilisation. Là encore, on ne considère pas cette dépendance temporelle de $X$.

| Impayés consécutifs | Amélioration | Stabilité | Dégradation |
|---|---|---|---|
| 0 | 0 % | 95 % | 5 % |
| 1 | 60 % | 10 % | 30 % |
| 2 | 10% | 30 % | 60 % |
| 3 | 5% | 25 % | 70 % |
| 4 | 5% | 15 % | 80 % |
| 5 | 5% | 5 % | 90 % |

TABLE 2 – Exemple d'évolution de dossiers à différents niveaux d'impayés.

## 2.2 Définition Bons / Mauvais

Une fois les données $X$ collectées, on cherche à déterminer la variable à prédire $Y$. On explique ci-après la procédure actuelle de constitution de cette variable que l'on vous propose de remettre en question librement au cours de cette Semaine d'Etude Maths-Entreprises.

### 2.2.1 Impayés successifs

A compter de la date de production du dossier ou de la première utilisation à crédit (voir partie 2.1), les clients doivent rembourser une mensualité (une somme variable, tous les mois, à des dates différentes selon les clients). On observe $N$ mensualités pour chaque client. On considère généralement qu'un client est "mauvais" ($Y = 0$) si celui-ci n'a pas remboursé $K$ mensualités successives. A l'inverse, on le considère "bon" ($Y = 1$) si les $N$ mensualités ont été payées (à la date à laquelle elles devaient l'être). On supprime tous les clients "indéterminés" ayant entre une et $K - 1$ mensualité(s) non remboursée(s). On se place donc ici dans le cadre de la classification, i.e. $\mathcal{Y} = \{0, 1\}$.

Afin de déterminer $K$ à $N$ fixé, on cherche pour chaque niveau d'impayés la proportion de dossiers se dégradant, c'est-à-dire parmi tous les dossiers qui ont un impayé par exemple, le nombre de dossiers qui, le mois suivant, ont deux impayés. On construit ainsi le tableau 2 et on choisit $K$ tel que la proportion de dossiers se dégradant dépasse 50 %. On choisit généralement $K = 2$.

### 2.2.2 Détermination de l'horizon de prédiction

Il reste cependant à déterminer l'horizon $N$. En effet, un client mauvais ($Y = 0$) à l'horizon $N$ est également mauvais à l'horizon $N + 1$ (il a toujours au moins $K$ mensualités consécutives non remboursées). Pour autant, on ne peut augmenter infiniment $N$ : on est limité par les observations les plus récentes ; à l'inverse, si $N$ est grand alors la période de production des dossiers est d'autant plus lointaine et on se heurte au problème de pertinence d'une population "ancienne" dans la détermination du risque de défaut des nouveaux clients. Ce problème est connu sous le terme de *Population Drift* [3]. En pratique, on détermine $N$ par l'utilisation de la courbe de risque : on trace le taux de mauvais dans la population considérée en fonction de $N$ et on observe généralement une

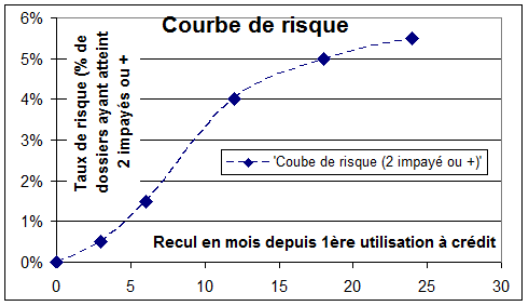


FIGURE 1 – Courbe "risque" pour déterminer l'horizon de prédiction.

stabilisation ou une inflexion. Dans la grande majorité des cas, cette stabilisation s'observe après $N = 12$ mois d'observation comme sur la figure 1.

Une fois $K$ et $N$ déterminés, nous disposons donc d'un $n$-échantillon $(\boldsymbol{x}, \boldsymbol{y}) = (x_i, y_i)_1^n$ qui va nous permettre d'estimer la fonction $f$.

## 2.3 Segmentation éventuelle

Dans certains cas, pour la population $\mathcal{X}$ visée, il peut être intéressant de développer deux scores sur des sous-populations distinctes $\mathcal{X}_1$ et $\mathcal{X}_2$ dont les performances seraient meilleures qu'un seul score sur $\mathcal{X}$. En pratique, ce travail relève d'une structure *a priori* : on va par exemple considérer qu'il faut un score dédié à chaque partenaire, ou à chaque marché, ...

Dans le cadre de la Semaine d'Etude Maths-Entreprises, on ignorera également cette éventualité et on se concentrera sur un modèle de données pour toute la population $\mathcal{X}$ pour laquelle on fournit des données.

## 2.4 *Feature Engineering*

Généralement, deux étapes de pré-traitement des variables explicatives sont effectuées :

- Sélection des variables explicatives les plus pertinentes,
- Discrétisation des variables continues et regroupement des modalités des variables qualitatives.

Cela permet d'une part d'éliminer les variables inutiles (qui n'apporteraient uniquement de la variance à l'estimation) et d'autre part de rendre le modèle plus flexible (dans le cas d'une relation non linéaire des prédicteurs continus à la variable à prédire).

Dans le cadre de ce sujet et par simplicité, on ignorera également ces deux étapes.

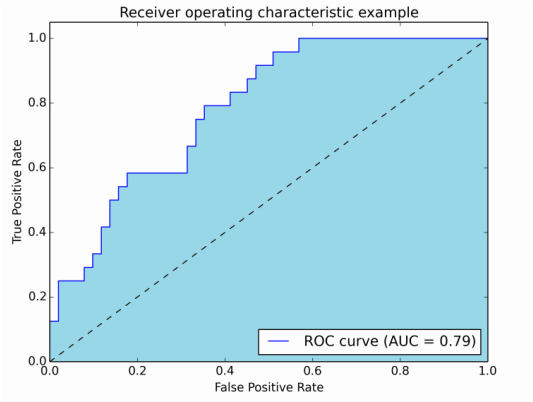


FIGURE 2 – Courbe ROC et AUC correspondante.

### 2.5 Régression logistique

Afin d'estimer $f$ à l'aide du $n$-échantillon obtenu plus haut, il faut se limiter à une classe de fonctions parmi lesquelles on souhaite choisir la plus adaptée. On se limite jusqu'à présent à la régression logistique :

$$\hat{f}(x) = p_{\hat{\theta}}(y|x) = \frac{1}{1 + e^{-\hat{\theta_0} - \sum_{j=1}^{d} \hat{\theta_j} x^j}}$$

**Intégration des refusés** La variable à prédire $Y$ ne peut être collectée que pour les clients financés (et, pour les clients au crédit renouvelable, ayant fait un achat à crédit) ce qui exclut *de facto* les clients refusés. Le modèle de régression logistique $p_\theta(y|x)$, bien qu'appris sur une population de clients financés et sous certaines hypothèses, se généralise naturellement à l'ensemble des demandes de crédit. Pour plus de détails, voir notamment [1].

On considère en première approximation qu'il n'y a pas de "biais de sélection" dans le $n$-échantillon proposé.

### 2.6 Mesure de la performance

Dans ce cadre supervisé, on peut imaginer plusieurs critères de performance pour mesurer la qualité de $\hat{f}$. Pour des raisons historiques, on s'intéresse à l'indice de Gini, directement lié à l'Aire Sous la Courbe ROC donnée en figure 2.

## 3 Problème(s)

Plusieurs problèmes se posent dans le cadre de la modélisation actuelle décrite brièvement en partie 2 dont certaines font actuellement l'objet d'une thèse CIFRE.

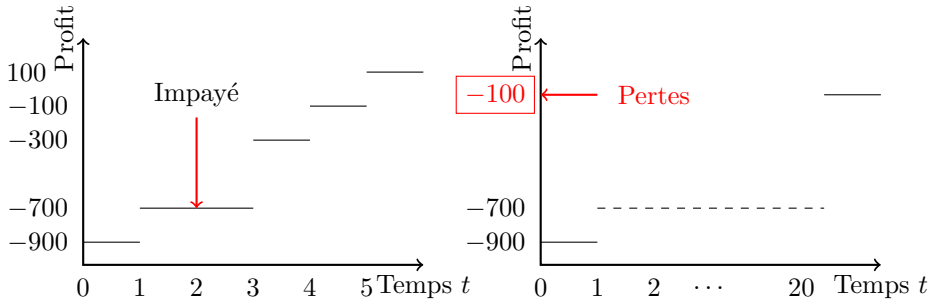


FIGURE 3 – Exemples de flux financiers.

Le problème que l'on soumet à la réflexion de la Semaine d'Etude Maths-Entreprises se concentre sur la partie 2.2. Pour se placer dans un cadre de classification en apprentissage supervisé, on applique cette "recette de cuisine" pour obtenir une variable à prédire $Y$ dans $\{0, 1\}$ sans pour autant que l'on justifie rigoureusement les raisons d'une telle approche.

On vous propose donc de réfléchir à une nouvelle définition argumentée de variable cible $Y$.

# 4 Pistes de résolution

Sans vouloir limiter le fruit de la réflexion de cette Semaine d'Etude Maths-Entreprises, deux axes de réflexion ont déjà été envisagés en interne, sans pour autant qu'ils aient été investigués.

## 4.1 Prédiction du profit généré par le client

L'objectif de l'entreprise est de maximiser son profit, raison pour laquelle il peut être plus intéressant d'accepter un client plus risqué (du point de vue du défaut) mais dont le taux sera plus élevé qu'un client peu risqué. Dans cette optique, on peut par exemple suivre les flux financiers du client dans le temps. En figure 3 à gauche, on voit à l'instant $t = 0$ le montant emprunté par le client qui commence ses remboursements mensuels. Il a un impayé qu'il rembourse plus tard et qui génère des agios. A $t = 6$, le crédit est terminé et il a généré un profit positif égal au cumul des intérêts perçus ainsi que des agios. A l'inverse en figure 3 à droite, le client cesse de payer et une partie de son crédit est passé en pertes.

## 4.2 Analyse de survie des dossiers

Une approche déjà étudiée dans [2] est liée à l'application de modèles historiquement liés à la bio-statistique : on étudie la survie des patients à une maladie en fonction de leurs caractéristiques (bilans sanguins, ...). Fort heureusement,

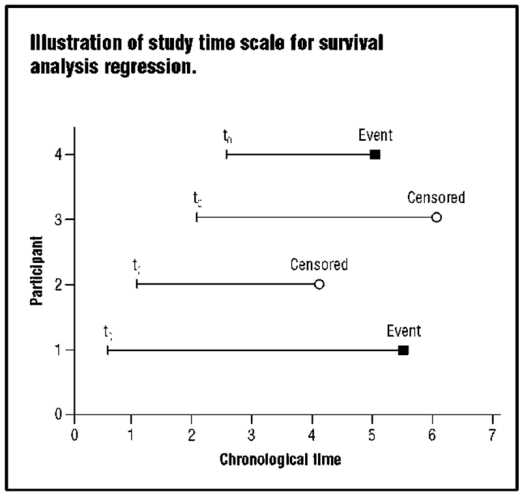


FIGURE 4 – Différences d'observation temporelle entre les clients.

au moment de l'étude tous les patients ne sont pas décédés ; on n'observe donc pas leur temps de décés. Certains peuvent par ailleurs changer d'établissement et sortir de l'étude. Ces cas sont présentés en figure 4. Par le même raisonnement, nous avons des clients "sains" au moment de l'observation qui sont censurés (on n'observe pas la fin du contrat), d'autre pour lesquels on observe le temps de fin de contrat et qui sont soit bons soit mauvais !

# 5 Données fournies

Afin de pouvoir éventuellement tester les pistes de résolution proposées lors de la Semaine d'Etude Maths-Entreprises, nous vous proposons un jeu de données anonymisées issues d'un partenaire de CA CF. Ces données concernent des prêts automobile (prêt bancaire "classique").

## 5.1 Période temporelle

On distingue généralement parmi l'ensemble des demandes trois sous-populations : les refusés (jugés trop risqués / instables), les sans-suite (acceptés par le système, ils n'ont finalement pas souhaité contracté de prêt) et les clients financés.

On ne sélectionne ici que les clients financés pour lesquels nous avons des informations de paiement exploitables.

La période de financement des crédits sélectionnée est de janvier à juin (inclus) 2015 de façon à avoir suffisamment de recul pour observer tout ou partie des remboursements jusqu'à aujourd'hui (la majorité des crédits automobile sont contractés pour 36 à 72 mois).

## 5.2 Les différents fichiers fournis et leur contenu

### 5.2.1 Données d'octroi

Les données dites de *financement* du client se trouvent dans la table **CLIENTS.CSV** dont les variables peuvent être regroupées en :

- ID : identifiant du dossier ;
- Informations socio-démographiques (préfixées par "CLIENT_") : ce sont ces variables qui servent classiquement au *Credit Scoring* ;
- Informations liées au crédit (préfixées par "CREDIT_") : montant accordé, nombre d'échéances, mensualité, ... ;
- Classe à prédire (préfixées par "PERF_") : la classe affectée à chaque dossier dans le cadre classique du *Credit Scoring*.
- Informations financières (préfixées par "FIN_") : les commissions données aux apporteurs ainsi que des estimations (sufixées par "_estim") des intérêts, agios, ...

### 5.2.2 Données comportementales

Les données dites d'*évènements* sur chaque dossier se trouvent dans la table **EVENEMENTS.CSV** dont les variables peuvent être regroupées en :

- ID : identifiant du dossier ;
- Type d'évènement : paiements, recouvrement, dossier terminé ;
- Evènement : remboursement total, partiel, contentieux, ...
- Les montants associés à chacun de ces évènements.

### 5.2.3 Données mensuelles

Les données mensuelles sur chaque dossier se trouvent dans la table **BILAN.CSV** qui fournissent une vision **très** simplifiée de l'évolution de l'encours de chaque dossier à chaque mois. En revanche, la "sortie" de cette table antérieure à décembre 2017 ne peut être expliquée qu'en se référant à la table d'*évènements*, plus complète.

### 5.2.4 Dictionnaire de données

Enfin le fichier **DICTIONNAIRE.XLSX** fournit des explications supplémentaires sur les données de chaque table (chacune séparée des autres dans un onglet spécifique) : libellé plus précis, signification des modalités, unités éventuelles, ...

# Références

[1] A. Feelders. Credit scoring and reject inference with mixture models. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 9(1) :1–8, 2000.

[2] Maria Stepanova and Lyn Thomas. Survival analysis methods for personal loan data. *Operations Research*, 50(2) :277–289, 2002.

[3] Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1) :69–101, 1996.