

# Projet SINF2275

## « Data mining and decision making »

### Projet classification et credit scoring

Année académique 2006-2007

---

**Professeurs :** Marco Saerens

**Adresse :** Université catholique de Louvain  
Information Systems Research Unit (ISYS)  
Institut d'Administration et de Gestion  
Place des Doyens 1  
B-1348 Louvain-la-Neuve  
Belgique

**Téléphone :** 010 47.92.46.

**Fax :** 010 47.83.24.

**Courriel :** [saerens@isys.ucl.ac.be](mailto:saerens@isys.ucl.ac.be)

---

#### **Objectif**

L'objectif de ce travail est la mise en pratique concrète d'un certain nombre de techniques d'analyse de données quantitatives, à travers l'étude d'un cas pratique nécessitant l'utilisation de logiciels de traitement statistique de données (SAS/JMP, SPSS, R, S-Plus, Matlab). L'application visée est le scoring d'une base de données de « credit scoring ». En résumé, il s'agit de détecter les « mauvais payeurs » à partir d'une série de caractéristiques mesurées sur des individus, en vue d'accorder ou refuser un prêt financier. L'impact financier de ce système de scoring sera évalué sur des données réelles.

#### **Contexte**

L'étude de cas concerne un organisme de prêts. Cet organisme désire automatiser sa règle de décision d'octroi d'un prêt. Pour ce faire, il s'adresse à vous. Vous serez donc responsable d'une étude visant à vérifier la faisabilité d'un tel système, sur base de données recueillies dans le passé. En effet, la société a pris soin de mettre en place une base de données des caractéristiques des personnes qui ont demandé un prêt à tempérament, et d'enregistrer si le remboursement de ce prêt s'est bien déroulé ou non. Il vous faudra donc mettre au point un **modèle de classification** en bon payeur/mauvais payeur, sur base des caractéristiques personnelles, et **évaluer ses performances** à partir de données nouvelles, afin de vérifier si cette automatisation est réaliste.

Ces données réelles sont consignées dans un fichier, *training.txt*, qui sera mis à votre disposition dès le début du projet. Ce fichier contient 700 observations et 21 variables discrètes et continues: des variables explicatives (caractéristiques des personnes) et une variable dépendante représentant le fait que le prêt a bien été remboursé ou non. Un document de description des variables, *credit.pdf*, sera également fourni.

A partir de cet échantillon de données représentant l'octroi de crédits, vous devrez construire un certain nombre de modèles permettant de détecter les « mauvais payeurs » à partir des variables explicatives (caractéristiques). Chose importante, les différents types d'erreurs n'ont pas le même poids. En effet, l'organisme de prêt considère qu'il est bien plus grave d'octroyer un prêt à une personne qui ne le remboursera pas (faux positif – erreur de type 1) que de ne pas octroyer un prêt à

une personne fiable (faux négatif – erreur de type 2). En réalité, l'organisme de prêt estime, qu'en moyenne, accepter une personne qui ne rembourse pas correctement son prêt de valeur  $x_5$  ( $x_5$  est le montant emprunté) occasionnera une perte approximative de  $(0.4 * x_5)$  et donc ( $gain = -0.4 * x_5$  €), ce qui se révèle catastrophique. L'organisme financier a en effet sous-estimé très largement la proportion de personnes qui ne remboursent pas leur emprunt. En revanche, une personne qui rembourse correctement son prêt fournit un bénéfice de  $(0.14 * x_5)$ , et donc ( $gain = +0.14 * x_5$  €). Bien entendu, les personnes qui ne sont pas acceptées pour le prêt fournissent un bénéfice nul ( $gain = 0$ ).

Une fois les modèles au point, l'organisme de prêt vous fournira un ensemble de **données réelles** indépendantes (300 données, *test.txt*) qu'il vous demandera de scorer à l'aide de votre meilleur modèle. Cet ensemble de données « réelles » ne vous sera fourni que deux semaines avant la remise du rapport.

Vous devrez calculer, sur ces données réelles,

1. Le taux d'erreurs de classification de votre meilleur modèle.
2. Le bénéfice obtenu sur ces données, *en sélectionnant les individus à l'aide de votre meilleur modèle*, et sachant qu'un mauvais payeur occasionne une perte de  $(0.4 * x_5)$  alors qu'un bon payeur fournit un bénéfice de  $(0.14 * x_5)$ .
3. A titre de comparaison, le bénéfice obtenu sur ces données, *en ne sélectionnant pas les individus* (donc en considérant la situation qui s'est produite jusqu'à présent, à savoir aucun filtrage des individus n'a été effectué : nous acceptons toutes les demandes de prêt). Il faudra donc répondre à la question suivante : cela vaut-il la peine d'effectuer une sélection à l'aide de votre modèle de scoring (quel est l'impact financier de votre système) ?

Sur base de ces évaluations, le groupe qui obtiendra les meilleurs résultats en terme de bénéfice sera sélectionné par l'organisme de prêt pour déployer sa solution au sein de l'entreprise.

## **Méthodologie**

Vous devrez remettre un rapport circonstancié, qui évalue clairement la possibilité de détection de « mauvais payeurs ». A cette fin, voici les différentes étapes que vous devrez franchir.

- **Rappel théorique:** Chaque méthode d'analyse utilisée devra être comprise: les hypothèses seront clairement énoncées et les principes théoriques seront esquissés.
- **Exploration des données:** Première étape dans le processus de modélisation, la phase d'exploration est particulièrement importante. Afin de se familiariser avec les données et de repérer d'éventuelles erreurs, chaque distribution de variable sera visualisée à l'aide d'un ou plusieurs graphiques. Certaines corrélations sont également calculées; par exemple, il est intéressant de calculer les corrélations entre la variable dépendante et les variables explicatives.
- **Sélection, transformation et recodage des variables:** L'information contenue dans les données n'est pas toujours sous une forme facilement interprétable. Très souvent, l'on est amené à transformer des variables de manière à extraire certains indicateurs ou à satisfaire les hypothèses du modèle. D'autre part, dans certains cas, certaines variables fournissent des informations redondantes ou n'apportent aucune information. Dans ce cas, celles-ci peuvent être éliminées ou transformées. De plus, certaines variables peuvent éventuellement être recodées et remplacées par un ensemble réduit de nouvelles variables sans perte significative d'information. Enfin, il est parfois utile d'éliminer des observations présentant des valeurs abérantes de l'échantillon.
- **Modélisation prédictive:** Il s'agit de la phase de modélisation. L'on estime les paramètres d'un modèle (dans notre cas, un modèle de classification) à partir d'un échantillon de données appelé « training set ». En général, l'on utilise plusieurs modèles différents afin de pouvoir

comparer leurs performances et choisir le « meilleur » selon un critère de performance bien défini.

- **Evaluation des modèles:** Les modèles sont comparés sur base d'une méthode de type « leave-n-out » ou de type « bootstrap », de manière à évaluer les modèles sur des données qui n'ont pas été utilisées pour l'estimation des paramètres. L'on calcule ainsi le taux d'erreur de classification et l'espérance du coût de mauvaise classification. En effet, il arrive que certains types d'erreur sont plus coûteux que d'autres. Dans ce cas, il faut calculer une estimation du coût (ou des bénéfices) occasionné par les erreurs de notre modèle.
- **Evaluation des performances sur données réelles:** Les performances du meilleur modèle sont évaluées sur des données réelles, indépendantes, comme si le modèle était mis en production. Cette base de données réelles est appelée « test set ».

Toutes les comparaisons de modèles seront effectuées sur base d'une procédure de type « leave-n-out » ou de type « bootstrap ». Vous choisirez, sur cette base, votre meilleur modèle qui entrera en compétition avec les modèles proposés par les autres groupes. En effet, l'organisme de prêt aura entre-temps compilé un ensemble de « données réelles » (qui serviront de test set et seront compilées dans le fichier : *test.txt*) qui seront scorées à l'aide de votre meilleur modèle. Les résultats (bénéfices) obtenus sur ces données réelles (test set) lui permettront de comparer les modèles proposés par les différents groupes.

Afin de faciliter le travail en groupe, nous avons rédigé, à titre d'exemple, quelques scénarios mettant chacun en œuvre un modèle différent. Nous vous encourageons à faire preuve d'originalité : vous pouvez très bien utiliser des modèles qui n'ont pas été présentés au cours !

### **Outils logiciels**

Nous demandons l'utilisation d'un outil logiciel spécialisé d'analyse des données.

Cinq outils (*SAS, SPSS, Matlab, R, S-Plus*) sont disponibles en salle informatique. Notons également que SAS/JMP et SPSS peuvent s'acheter à un prix modique (environ 25€, si vous souhaitez travailler à domicile). R, quant à lui, est gratuit, et peut être téléchargé à partir du site : <http://www.R-project.org>. L'utilisation de R, SAS ou Matlab est conseillée.

Notons que la version étudiant de S-Plus, l'équivalent commercial de R, est gratuite et peut être téléchargée à partir de [www.insightful.com](http://www.insightful.com).

Nous recommandons par ailleurs, dans le cadre de la modélisation, l'utilisation de R, Matlab, SAS ou S-Plus. Notons que toute la documentation de ces logiciels est disponible en format PDF.

### **Rapport**

Nous demandons :

1. Seul un rapport final sera demandé pour les trois projets. Il sera recommandé de rédiger le rapport sous forme d'article scientifique. Il sera recommandé de rédiger le rapport sous forme d'article scientifique (des templates pour Latex ou Word peuvent être trouvés sur des sites de publication scientifique tel que Springer ou des sites de conférence). Rappelons que le rapport d'analyse devra comprendre un bref **rappel théorique** reprenant l'analyse détaillée du mécanisme de prise de décision (règle de décision utilisée pour maximiser le bénéfice). L'étudiant pourra se baser sur la documentation des outils logiciels, sur des photocopies de chapitres de livres fournies, ou sur des ouvrages disponibles en bibliothèque.
2. Ce rapport imprimé devra être remis le vendredi de la dernière semaine de cours, en main propre, à l'un des membres de l'unité ISYS. Tout retard sera sanctionné de 2 points sur 20 par semaine de retard.

## **Organisation du travail en groupe**

Il est vivement conseillé de s'organiser et de se répartir le travail dès la remise de l'énoncé. Le rapport comportera le rappel théorique ainsi que les expérimentations à partir des modèles mis en oeuvre. Chaque groupe d'étudiants devra traiter un nombre minimum de scénarios égale au nombre de membres du groupe et comparer les résultats obtenus par ces modèles. Une dernière partie comprendra l'**application aux données réelles** (test set) et les résultats correspondants.

Notons également que, pour chaque modèle, **deux règles de décision** seront développées :

- La première règle cherche à **minimiser le pourcentage de classifications incorrectes** (règle de décision de Bayes : on affecte l'observation à la classe pour laquelle la probabilité a posteriori d'appartenance est la plus élevée). Le critère de performance du modèle est, dans ce cas, le pourcentage de classifications correctes en leave-n-out.
- La seconde règle de décision, plus appropriée dans notre cas, cherche à **minimiser le coût total** ou maximiser le bénéfice. Dans ce cas, il faut utiliser la théorie de la décision afin de déterminer un seuil de décision optimal, dépendant des coûts associés à chaque cas de figure (vrai mauvais payeur classé bon payeur; vrai bon payeur classé bon payeur, etc), calculé à partir des probabilités a posteriori fournies par le modèle (voir cours ou chapitres de livres). Le critère de performance du modèle est, dans ce cas, le bénéfice obtenu en leave-n-out ou en bootstrap en filtrant les individus à partir de la règle de décision mise au point (l'on n'octroie pas de prêt aux individus classés mauvais payeurs par la règle de décision).

**Scénario 1. Phase exploratoire.** A partir du « training set », l'étudiant effectuera (en SAS/JMP par exemple) des **statistiques univariées** pour chacune des variables (graphique de la distribution, moyenne, médiane, minimum, maximum), ainsi que **bivariées** (tests d'association) entre la variable dépendante (catégorielle) et les variables explicatives. Sur base de ces statistiques bivariées, il faudra **sélectionner** les variables qui sont fortement associées à la variable dépendante, sur base de tests de signification, et éliminer les autres. A partir de ces variables sélectionnées, il faudra estimer les paramètres d'un modèle de type **régression logistique**. Par ailleurs, l'on estimera un deuxième modèle de type régression logistique en sélectionnant, cette fois-ci, les variables par une procédure automatique « **stepwise** ». Les variables éliminées manuellement et par procédure stepwise sont-elles les mêmes ? Enfin, l'on estimera les **performances** des deux régressions logistiques à partir d'une procédure « **leave-n-out** » : il faudra évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu, en sélectionnant les individus à l'aide de votre modèle. A titre de comparaison, vous calculerez également le bénéfice obtenu si aucune sélection n'est effectuée, ce qui permettra de calculer l'impact financier de votre système. Les résultats sont-ils similaires à ceux obtenus sur le training set ? Par ailleurs, l'on comparera les résultats obtenus à ceux obtenus à partir d'un modèle équivalent (de type régression logistique), cette fois-ci estimé à partir de l'**ensemble des variables originales**.

**Scénario 2. Réduction des données.** A partir du « training set », l'étudiant devra effectuer une **réduction de dimensionnalité** des données, et ensuite estimer un modèle de classification qui se base sur les données réduites (en sélectionnant un nombre de facteurs réduit). En ce qui concerne la réduction des données, l'on utilisera, pour les variables continues, une **analyse en composantes principales** ainsi qu'une **analyse discriminante**. Pour les variables discrètes, l'on utilisera l'**analyse des correspondances multiples**. Il faudra étudier l'influence du nombre de facteurs conservés sur les performances en tâche de classification (si vous avez assez de temps). En ce qui concerne la classification à partir de ces variables réduites, il faudra estimer les paramètres d'un modèle de type **régression logistique** et comparer ses performances à celles obtenues à partir d'un modèle équivalent (également de type régression logistique), cette fois-ci estimé à partir de l'ensemble des variables originales. Les résultats obtenus à l'aide de la régression logistique simple sur l'ensemble des

variables serviront de point de comparaison à toutes les méthodes expérimentées. Cela signifie que vous estimerez une régression logistique sur base des variables réduites par analyse en composantes principales (variables explicatives continues) + les variables réduites par analyse des correspondances (variables explicatives discrètes). Il faudra évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu à partir d'une procédure « **leave-n-out** », en sélectionnant les individus à l'aide de votre modèle.

**Scénario 3. Clustering.** A partir du « training set », l'on s'arrangera pour préalablement résumer les variables discrètes en un ensemble réduit de variables continues en utilisant une **analyse des correspondances multiples**. Ensuite, l'on effectuera un **clustering** (hiérarchique ou k-means ou fuzzy clustering) des données. L'opération de clustering consiste à remplacer l'ensemble des variables explicatives par un ensemble de groupes (ou clusters).

Afin d'effectuer ce clustering, il faut

- Réduire les variables discrètes significatives en un ensemble restreint de variables numériques par analyse des correspondances.
- Effectuer le clustering sur l'ensemble des variables explicatives numériques + les variables obtenues par analyse des correspondances.

Voici encore quelques précisions :

- Lorsque vous utilisez le k-means, certains clusters peuvent être vides lorsqu'on demande un nombre de groupes important (> 5). Ce problème est inhérent au k-means qui est très sensible au fait qu'il y a des valeurs entières dans le data set. Pour réduire cet effet indésirable, vous pouvez, par exemple, effectuer préalablement une PCA sur les variables numériques. Ceci permet d'éviter dans une certaine mesure les régions "vides" dues à la présence de variables entières.
- Le modèle obtenu par clustering est assez rudimentaire (résumer les caractéristiques des observations à un seul cluster): il devrait a priori fournir des résultats plutôt médiocres. Nous vous conseillons dès lors d'effectuer un clustering **séparé par classe** (un clustering des "mauvais payeurs" et un clustering des "bons payeurs"). Ainsi, l'on obtient des clusters plus spécifiquement "bons payeurs" et "mauvais payeurs".

A des fins d'interprétation des groupes, un cluster ou groupe obtenu sera caractérisé à l'aide d'un **arbre de décision** qui discriminera ce groupe par rapport à tous les autres sur base des variables originales.

Ensuite, l'on calculera les proportions de classe « bon payeur » et « mauvais payeur » pour chacun des groupes, afin d'observer si certains groupes sont plus spécifiquement de type « bon payeur » ou « mauvais payeur ». Ces proportions de classe « bon payeur » et « mauvais payeur » fournissent une estimation des probabilités a posteriori des classes, conditionnellement au cluster d'appartenance de l'observation. Si un groupe est majoritairement « bon payeur » (« mauvais payeur »), tout élément qui appartient à ce groupe sera considéré comme « bon payeur » (« mauvais payeur »). On évaluera les performances de ce modèle simple en leave-n-out (il faudra évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu, en sélectionnant les individus à l'aide de votre modèle. A titre de comparaison, vous calculerez également le bénéfice obtenu si aucune sélection n'est effectuée, ce qui permettra de calculer l'impact financier de votre système).

**Scénario 4. Neural networks.** A partir du « training set » et de l'ensemble des variables, discrètes et continues, l'étudiant estimera un modèle de type « **artificial neural network** » en faisant varier le nombre d'« unités cachées » du réseau de neurones multicouche. Il faudra également estimer un modèle de type **régression logistique** à partir de l'ensemble des variables. L'on pourra également

inclure des effets d'**interaction** dans le modèle logistique. Pour tous les modèles estimés, il faudra évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu, en sélectionnant les individus à l'aide de votre modèle. A titre de comparaison, vous calculerez également le bénéfice obtenu si aucune sélection n'est effectuée, ce qui permettra de calculer l'impact financier de votre système. Les résultats sont-ils similaires à ceux obtenus sur le training set ?

**Scénario 5. Decision trees.** A partir du « training set » et de l'ensemble des variables, discrètes et continues, l'étudiant estimera un modèle de type « **decision tree** » (SAS/JMP ou R) en faisant varier la taille de l'arbre ainsi que le critère utilisé pour construire l'arbre. L'étudiant comparera les résultats obtenus et les variables sélectionnées aux résultats obtenus par une régression logistique avec sélection de variables « stepwise ». Les variables sélectionnées sont-elles les mêmes ? Enfin, pour tous les modèles estimés, il faudra évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu, en sélectionnant les individus à l'aide de votre modèle. A titre de comparaison, vous calculerez également le bénéfice obtenu si aucune sélection n'est effectuée, ce qui permettra de calculer l'impact financier de votre système. Les résultats sont-ils similaires à ceux obtenus sur le training set ?

**Scénario 6. Classifieur Bayésien naïf.** A partir du « training set », l'étudiant effectuera des **statistiques univariées** pour chacune des variables (graphique de la distribution, moyenne, médiane, minimum, maximum), ainsi que **bivariées** (corrélations et tests d'association) entre la variable dépendante et les variables explicatives. Sur base de ces statistiques bivariées, il faudra **sélectionner** les variables qui sont fortement associées à la variable dépendante, sur base de tests de signification, et éliminer les autres. Sur base des variables sélectionnées uniquement, il faudra estimer les paramètres d'un modèle de type **classifieur Bayésien naïf**. A cette fin, il faudra estimer les densités de probabilité des caractéristiques relevantes (aussi bien catégorielles que numériques), conditionnellement à la classe d'appartenance. Pour comparaison, l'on estimera également un modèle de type **régression logistique** à partir des variables sélectionnées. Enfin, pour tous les modèles estimés, il faudra évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu en leave-n-out (ou bootstrap), en sélectionnant les individus à l'aide de votre modèle. A titre de comparaison, vous calculerez également le bénéfice obtenu si aucune sélection n'est effectuée, ce qui permettra de calculer l'impact financier de votre système (toujours en leave-n-out). Les résultats sont-ils similaires à ceux obtenus sur le training set ?

**Scénario 7. Combinaison de classifieurs.** A partir de différents modèles de classification estimés sur le « training set », l'on effectuera une combinaison de classifieurs qui définit dès lors un nouveau modèle de classification. Pour ce dernier modèle, il faudra évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu en leave-n-out (ou bootstrap) en sélectionnant les individus à l'aide de votre modèle. A titre de comparaison, vous calculerez également le bénéfice obtenu si aucune sélection n'est effectuée, ce qui permettra de calculer l'impact financier de votre système obtenu en leave-n-out (ou bootstrap). Les résultats sont-ils similaires à ceux obtenus sur le training set ?

Finalement, quel modèle fournit les meilleurs résultats en terme de taux de classification correcte et de bénéfice, en sélectionnant les individus à l'aide de votre modèle ? Est-ce le même modèle qui obtient les meilleurs résultats sur le training set ? Effectuer une sélection sur base de votre meilleur modèle s'avère-t-il profitable ?

**Evaluation sur les données réelles.** Comme déjà mentionné, l'organisme de prêt fournira un ensemble de données réelles (300 observations – le **test set**) qui devra être scoré par le meilleur modèle obtenu (celui qui a obtenu les meilleures performances en leave-n-out). Ce meilleur modèle sera cependant **ré-estimé** sur l'ensemble des données disponibles (ensemble du training set) afin de tirer profit de toutes les données disponibles. Il sera ensuite appliqué tel quel aux données réelles. Il faudra donc, finalement, évaluer le taux d'erreur, la matrice de confusion, ainsi que le bénéfice obtenu, sur les données du test set, en sélectionnant les individus à l'aide de ce meilleur modèle. A titre de comparaison, vous calculerez également le bénéfice obtenu si aucune sélection n'est effectuée, ce qui

permettra de calculer l'impact financier de votre système sur les bénéfices de la société (toujours à partir du test set).

Bien sûr, ces scénarios sont indicatifs : vous pouvez (c'est même apprécié) **imaginer d'autres scénarios** et les tester. En effet, vous pouvez, pour chaque méthode testée

- Sélectionner ou non des variables;
- Transformer ou non des variables;
- Eliminer ou non les valeurs abérantes;
- Etc...

Par ailleurs, un grand nombre de méthodes de classification alternatives sont disponibles en R ou S-Plus.

### **Cotation du projet**

Chaque cote individuelle sera un compromis entre les cotes obtenues aux projets et la cote de l'examen oral.

