Revue d'Economie et de Statistique Appliquée
Numéro 21 Juin 2014
ISSN : 1112-234X

METHODES DES ARBRES DE DECISION POUR LE SCORING BANCAIRE

LOUNICI Nora[1],
KHERCHI MEDJDEN Hanya,[2]
SADI Khadidja[3]

RESUME :

Cet article présente l'automatisation du traitement de demande de prêt bancaire, offert aux jeunes chômeurs pour les encourager à créer leurs propres entreprises, par le biais d dispositif ANSEJ[4], selon une approche de Datamining : Les arbres de décisions. Les modèles de scoring pour la tâche d'apprentissage supervisé chosie exposent les possibilités pratiques des trois méthodes suivantes : *CHAID, Cart et C4.5.* une présentation empirique comparative de ces trois algorithmes d'Arbres de décisions permettra de sélectionner le modèle «emprunteur», le plus performant. Nous nous intéressons plus particulièrement à l'étape cruciale de valider les classements obtenus à travers l'étude d'un cas pratique nécessitant l'utilisation de plusieurs logiciels « open source » de Data Mining (tanagra, Sipina et Weka). Il s'agit de détecter automatiquement à partir d'une série de

Mots clés : classement, scoring bancaire, Data Mining, arbres de décision, logiciels open source Data Mining

I. INTRODUCTION

Plusieurs types de crédits peuvent être offerts par une banque. Les principaux crédits proposés sont l'octroi de crédits aux particuliers et aux entreprises. Cette opération peut engendrer des pertes majeures en cas de non remboursement. A ce propos, plusieurs travaux de recherche ont été réalisés pour détecter à l'avance les emprunteurs qui seront défaillants de ceux qui ne seront pas. Ces travaux sont basés essentiellement sur l'analyse des comptes annuels des emprunteurs.

Le risque de crédit fut pendant longtemps un problème majeur pour les banques, car les mesures de contrôle qu'elles entreprennent pour faire face aux différents risques étaient relativement peu développées. L'analyse et la gestion du risque de crédit a pris une telle ampleur, ce qui a contribué au développement de nouveaux outils afin de répondre au plus vite aux attentes de décideurs et se donner les moyens de minimiser les pertes.

La gestion classique de fonctionnement des institutions financières Algériennes consiste à déléguer la décision d'octroi de prêts aux agents de crédits et se base donc exclusivement sur leur avis. Au niveau de la BADR, les dossiers de prêts subissent un traitement manuel qui nécessite beaucoup de temps et de l'expérience. L'évaluation fortement empreinte de subjectivité humaine effectuée par le comité de crédit présente des insuffisances en termes de temps nécessaire pour le traitement des données et d'objectivité dans la prise de décision. Pour palier ces insuffisances et réduire le risque des impayés, il existe actuellement de multiples techniques de datamining (scoring, classement, association de produits et services, etc…).

L'élaboration de modèles prédictifs via le data-mining (fouille de données) est de plus en plus utilisée dans la finance, surtout aux Etats-Unis où les crédits à la consommation ont explosé ces dernières années. Désormais les méthodes de data mining envahissent de nombreux domaine et tout

[1] Maitre de conférences classe B à l'ENSSEA
[2] Maitre de conférences classe A à l'ENSSEA
[3] Maitre de conférences classe A à l'ENSSEA
[4] Agence nationale de soutien à l'emploi des jeunes

ISSN : 1112-234X

particulièrement les banques de détail, du fait de l'importante quantité de données accumulées. Les ordinateurs sont capables de construire des modèles prédictifs très précis, capables d'anticiper les risques financiers et faire gagner au décideur un temps considérable. De nombreux outils de fouille de données permettant d'effectuer des calculs de scoring, existent et sont arrivés à maturité : réseaux neuronaux, régression logistique, arbres de décision, etc.

Pour mettre en application une procédure de scoring, il faut au préalable disposer de données de bonne qualité. De ce fait, un travail conséquent doit être réalisé en amont pour mener à bien la modélisation des données. Il est également nécessaire de suivre une démarche méthodique pour arriver à la construction d'un modèle efficace. Par cette étude, on cherche à déterminer quelles sont les variables les plus pertinentes qui permettent de comprendre pourquoi la banque accorde un crédit à tel client et pas à un autre et par la même occasion construire un modèle prédictif permettant de classer un nouveau client demandeur de prêt dans l'une des 2 classes (accord ou refus de crédit).

Les principales méthodes de scoring sont *l'Analyse Discriminante, la régression logistique, les réseaux de neurones et les arbres de décision*. Dans cet article, nous nous concentrons sur l'utilisation de la technique *des arbres de décision* [Morg & Sonq 1963][Kass 1980][Brei & al 1984]. Cette approche peut être appliquée pour la construction de modèles explicatifs, dont le principal but est d'identifier les variables qui caractérisent le mieux la variable cible Y généralement binaire. On peut aussi s'en servir pour faire de la prédiction. Dans ce cadre là, le problème de l'apprentissage se résume à rechercher l'arbre optimal qui offre une bonne généralisation du modèle.

Le modèle décrit sous forme d'arbre peut être très simplement transcrit par une succession de règles disjointes, facilement interprétables. La structure arborescente intègre les attributs les plus marquants selon leur degré d'influence avec la variable à discriminer. Les attributs les plus pertinents sont ceux qui sont proches de la racine.

Cette méthode est très appréciée en Data Mining [Hang & al 01] car elle est très intuitive et visuelle. De plus, elle supporte les données hétérogènes, ainsi que les effets non linéaires. Elle peut être assimilée à une classification hiérarchique descendante à double objectif supervisé et non supervisé. Parmi les algorithmes les plus référencés, on cite : C4.5 [Quin 1993], ID3 (Quin 1986), CART [Brei & al 1984] et CHAID [Hart 1975].

Dans ce papier, nous définissons en section 2 l'aspect théorique des arbres de décision. La section4 est consacrée à la construction de la base d'analyse. La section5 introduit le contexte applicatif des données de l'ANSEJ. La section6 présente les expérimentations et analyses des résultats obtenus par les trois méthodes d'arbres de décision. Finalement la section 7 conclut l'article et évoque les évolutions futures, après une discussion des résultats obtenus.

2 Quelques notions théoriques sur les méthodes d'arbres de décision

Un arbre de décision est un graphe orienté acyclique dont les nœuds correspondent aux variables choisies sur la base de critères de qualité, quant aux arcs, ils représentent les modalités[5] d'une variable prédictive. Les nœuds terminaux sont appelés feuilles et évoquent les classes. La construction de l'arbre consiste à partitionner les données selon la variable explicative la plus discriminante. Ce processus est répété localement sur chaque nœud de l'arbre jusqu'à l'obtention de feuilles pures (ie. correspondent à des feuilles constituées d'individus d'une même classe), ou sur ordre d'un arrêt volontaire de la progression de l'arbre. Les différents nœuds de l'arbre sont caractérisés par la distribution des effectifs de la variable cible.

[5] On procède à une discrétisation an cas de variables numériques

ISSN : 1112-234X

Les performances de prédiction dépendent directement de la taille de l'arbre appris. Une première difficulté concerne le choix des variables pertinentes. Ce choix est fait sur la base d'un critère de séparation. Parmi les critères les plus fréquemment utilisés figurent :

– ***L'entropie de Shannon***, applicable à tout type de variables explicatives. Cette mesure est notamment utilisée par Quillan dans C4.5 et C5.0 pour mesurer l'incertitude :

Entropie ($nœud_t$) = $\sum_{i=1}^{k} - f_i \, log_2 f_i$ où f_i ,(i = 1, …, p) sont les fréquences relatives dans le nœud t des k classes à prédire.

– ***L'algorithme CART*** produit des arbres de décision binaires et applique l'indice de Gini appelé entropie quadratique pour sélectionner les variables explicatives de tout type.

Gini ($nœud_t$) = $\sum_{i=1}^{k} f_i \, (1 - f_i) = 1 - \sum_{i=1}^{k} {f_i}^2$

– ***La méthode CHAID*** s'appuie sur le test du khi-2 comme écart à l'indépendance pour choisir le meilleur attribut de segmentation, afin de construire des arbres de décision non-binaires.

La formule du khi-2, bien maitrisée par la communauté des chercheurs, découle du croisement des modalités de la variable cible avec une variable prédictive de type qualitative ou discrète. Les variables continues sont automatiquement discrétisées par les logiciels de data mining [Tuf 2010]. A un niveau donné de l'arbre, la découverte de la variable la plus significative est basée sur le test du khi-2. La valeur de ce test est calculée par la formule :

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - previj)}{previj} \qquad \text{où} \quad previj = \frac{n_{i.} \times n_{.j}}{n}$$

Le test du ${\chi^2}_T$ théorique est ensuite appliqué avec un niveau de significativité α, choisi par l'utilisateur pour vérifier s'il y'a indépendance entre le croisement des deux variables. Selon les cas :

- Si $\chi^2 > {\chi^2}_T$ l'hypothèse d'indépendance est rejetée dans le cas contraire, l'hypothèse d'indépendance est acceptée.

Dans cette méthode la détermination des seuils de significativité est une opération un peu délicate et nécessite l'avis de l'expert métier.

2.1 Le gain d'information

Les algorithmes d'arbres de décision procèdent généralement au calcul du gain en information. Cette mesure fournit l'information gagnée après la séparation du nœud parent en nœuds fils. L'attribut discriminant sélectionné est celui qui dispose du gain informationnel maximal et donc d'un besoin d'information minimal. Il est donné par la différence :

$$\textbf{\textit{Gain}} = critère\ du\ parent - \sum_i critère\ des\ fils$$

Une fois l'arbre de segmentation construit, on le défait progressivement pour générer les règles de décision. Le modèle résultant correspond à l'ensemble des chemins menant de la racine à une feuille. Ainsi, pour prédire la classe associée à un nouvel objet, il suffit de lui faire parcourir l'arbre de la racine jusqu'à l'une de ses feuilles, en prenant soin de vérifier toutes les conditions sur les meilleures variables au sens des critères évoqués ci-dessus, pour finalement lui attribuer la classe associée à la feuille terminale dans laquelle elle se trouve.

ISSN : 1112-234X

Un arbre de décision est déclaré correct et complet si tous les individus étiquetés sont correctement classifiés. Cette configuration idéale n'est jamais atteinte sur des applications réelles. Afin de s'approcher de cette solution, des opérations de *pré ou post-élagage* de l'arbre sont souvent nécessaires. Elles consistent à supprimer les feuilles les moins significatives et à les remplacer par un nœud terminal qui représente la classe majoritaire des individus classés par cette partie de l'arbre

2.2 Elagage

Le pré-élagage adopté par les algorithmes CHAID et ID3 consiste à imposer des règles d'arrêt lors du développement de l'arbre. Ce qui revient à fixer une condition d'arrêt pour bloquer la construction.

C'est le cas par exemple, de la méthode CHAID [Kass 1980] qui applique la valeur de la p-value du test du Khi-2 comme règle d'arrêt. Toute la difficulté de ces algorithmes est de trouver un juste milieu entre un arrêt précoce : sous-apprentissage, qui empêcherait le modèle d'apprendre toute l'information contenue dans les données, et un apprentissage par cœur des données: c'est ce qu'on appelle le sur-apprentissage[6].

Le post-élagage préconisé par CART [Brei & al.1984] puis repris par Quillan avec l'algorithme C4.5 consiste d'abord à construire récursivement un arbre en le laissant croître jusqu'à atteindre son maximum (growing phase) au risque de sur-apprendre. Le modèle obtenu n'est pas optimal, il reflète presque fidèlement les exemples de la base d'entrainement. Ensuite, dans une seconde phase au moyen d'un critère pénalisé, une comparaison est effectuée sur la succession de sous-arbres emboités afin de sélectionner l'arbre minimal qui offre le meilleur compromis entre l'erreur d'apprentissage et l'erreur de généralisation.

Pour extraire cet arbre optimal, on calcule le taux d'erreur en test ou par validation croisée des différents sous arbres et on retient le plus bas possible. C'est précisément la phase d'élagage (pruning phase), le mécanisme employé durant cette étape est appelé le coût complexité minimal[7].

Le temps d'apprentissage est certes plus long mais les performances de l'arbre sont meilleures. Ces méthodes utilisent un critère d'élagage basé sur l'estimation du taux d'erreur de classification.

2.3 Apports et limites des arbres de décisions

Les arbres de décision ont de nombreux avantages et se comptent parmi les méthodes de fouille de données les plus appréciées [Wux & al 08]. C4.5 est l'algorithme de référence par excellence en apprentissage supervisé. Parmi les avantages on cite :

- Elles permettent de traiter d'importants volumes de données hétérogènes (catégorielles ou numériques), pour des temps de calculs faibles.
- La méthode est de nature exploratoire, elle nécessite peu d'hypothèses sur les données.
- Les chemins résumant des décisions transcrites sous forme de règles (Si…Alors) sont compréhensibles et donc facilement interprétables par un utilisateur non initié.
- Les arbres de décision ne souffrent pas d'outliers. Ces valeurs extrêmes sont isolées dans des feuilles et sont facilement détectables.

Cependant, les arbres de décision souffrent de quelques inconvénients :

[6] Un phénomène de sur-spécialisation qui produit de bonnes performances pendant la phase d'apprentissage mais qui va fournir de mauvaises prédictions.
[7] MCCP (*Minimal Cost-Complexity Pruning*) en Anglais

ISSN : 1112-234X

- Le manque de précision dans les prédictions.
- Face à un nombre trop important de classes, les arbres de décisions ont tendance à devenir très complexes et complètement illisibles.
- Impossible de revenir sur les affectations des niveaux antérieurs.
- la recherche d'un arbre de décision optimal est un problème difficile.
- Le problème de généralisation confronté au sur-apprentissage est omniprésent avec les arbres de décision.

3. Construction de la base d'analyse

Les modèles que l'on construit ne représentent qu'un reflet approché de la réalité. Pour apprécier la capacité d'un modèle à bien représenter les données, la démarche classiquement utilisée consiste à séparer les instances en deux groupes : les données d'apprentissage (ou d'entrainement), et les données tests.

L'algorithme commence par apprendre et construire son modèle, type arbres de décision à partir du jeu de données d'apprentissage. Une fois l'apprentissage terminé, on mesure la fiabilité du modèle obtenu sur l'échantillon test qui doit être différent de l'échantillon d'apprentissage. Généralement on utilise 70% des individus pour apprendre le modèle et le reste pour mesurer la performance.

La mesure de la performance du modèle repose sur la détermination du taux d'erreur, donné par le rapport du nombre d'instances mal classées sur le nombre d'instances du jeu de données test. Il est également possible de calculer le taux d'erreur pour chacune des classes. L'inconvénient de cette approche est qu'elle nécessite beaucoup d'individus dans le jeu d'apprentissage, que l'on ne retrouve pas forcément dans son complément test

Pour palier à ce problème, La technique de la validation croisée peut être appliquée. Elle consiste à découper la population en k sous-populations d'effectif égal, ensuite à appliquer par effet de boucle le procédé suivant : le modèle est construit sur les k-1 subdivisions, quant à l'évaluation du modèle, elle est testée sur le sous ensemble restant. Par ce procédé tous les individus sont visités et se retrouvent au moins une fois dans le jeu de données test. Le taux d'erreur du modèle, est obtenu par le calcul de la moyenne des différentes validations. L'inconvénient de cette approche est le nombre de modèles à construire et le temps d'exécution.

4. Contexte applicatif sur les données de l'ANSEJ:

L'Etat a impliqué l'Agence nationale de soutien à l'emploi des jeunes (ANSEJ) dans son dispositif de lutte contre le chômage et a incité les banques à accorder des prêts bancaires sans intérêt aux jeunes pour les encourager à créer leurs propres micro-entreprises par le biais de cette agence. Le délai de traitement des dossiers de demande de crédit est relativement long et sans analyse objective. Nous souhaitons appliquer une des méthodes de fouille de données, à savoir les arbres de décision à des données fournies par la banque BADR. Nos objectifs sont multiples. Nous cherchons à analyser les données selon plusieurs méthodes d'arbres de décisions à travers différents logiciels de data mining et en juger les performances.

Le jeu de données mis à notre disposition porte sur 635 dossiers clients. Parmi ces clients, demandeurs de crédits pour financer des projets d'investissements, 522 demandeurs ont reçu l'accord et 113 ont essuyé un refus. Les dossiers en question, « collectés sur une période de 12 ans » de 1997 à 2008 ne comportent aucune valeur manquante. La base de données est caractérisée par 11 variables listées dans le tableau ***Tab 1.***

ISSN : 1112-234X

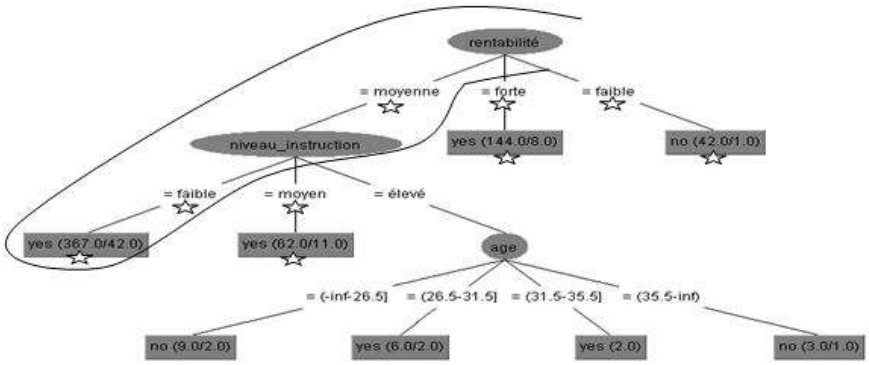
Afin d'analyser ces données nous avons mis en concurrence les trois méthodes de classements sur la base de données ANSEJ : les méthodes CART, C4.5, CHAID, de manière à sélectionner celui qui maximise la performance. Cette démarche est courante en DataMining. Plusieurs modèles comparables sont mis à l'épreuve pour finalement retenir le meilleur modèle.

5. Expérimentations et analyse des résultats

Dans cette section nous comparons les résultats obtenus en induction par arbres de décision sur nos données. Etant donnée la taille des données relativement faible, nous avons choisi de subdiviser dans un premier temps la base de données dans les proportions 2/3~~1/3, ensuite appliquer la techniques de Cross-validation (k=10). Nous savons que le taux d'erreur calculé sur l'échantillon d'apprentissage est trop optimiste, l'opération de validation croisée mise en œuvre permettra de valider le modèle de façon objective.

Pour réaliser les classements, nous avons fait appel aux algorithmes de WEKA, TANAGRA [Zig & Rako 00] et SIPINA [Rako 00]. L'algorithme J48 de WEKA est équivalent à la méthode C4.5. Les paramètres ont été fixés à 10 exemples au minimum par feuille, avec un facteur de confiance de 0,40 avec élagage.

5.1 Logiciel WEKA (méthode C4.5): L'arbre de décision délivré par WEKA se présente ainsi :



Parmi les règles que l'on peut soutirer de cet arbre, trois d'entre-elles concluent sur une acceptation du prêt. Les chemins relatifs à ces règles nous les avons indiqués par une étoile au niveau de l'arbre de décision. Exemple : le chemin délimité par un trait continu sur le graphe correspond à la règle suivante : ***si*** *« rentabilité = moyenne et niveau_instruction = faible »* ***alors «*** *367 clients ont vu leur prêt accepté et 42 seulement refusé »*. On conclut dans ce cas sur la classe majoritaire qui correspond à une demande de prêt acceptée. Par ailleurs, la variable Age est peu discriminante, car toutes les feuilles qui lui sont rattachées comportent peu d'individus. Ce nœud peut être supprimé.

Par conséquent, les variables les plus importantes qui influent sur la décision sont la *rentabilité du projet et le niveau d'instruction*. Parmi les modalités, 5 valeurs sont significatives et

ISSN : 1112-234X

interviennent dans la prise de décision. Il s'agit des trois modalités de la *variable rentabilité, ainsi que du niveau d'instruction, moyen et faible*. Ce qui prouve que les orientations sociopolitique de l'état visent à soutenir l'emploi des jeunes chômeurs.

5.2 Logiciel Tanagra (méthode CART)

L'apprentissage a été lancé dans les mêmes conditions que sur weka, pour que les résultats soient comparables. Tanagra a généré les règles suivantes :

Decision tree

- **Rentabilité estimée** in [moyenne,forte]
  - **Montant** < 2154475,0000
    - **valeurs des garanties** < 1039760,6250
      - **Montant** < 761299,0000 then Décision = **oui** (93,02 % of 43 examples)
      - **Montant** >= 761299,0000 then Décision = **non** (83,33 % of 12 examples)
    - **valeurs des garanties** >= 1050760,6250 then Décision = **oui** (92,53 % of 308 examples)
  - **Montant** >= 2154475,0000
    - **Rentabilité estimée** in [moyenne] then Décision = **non** (70,00 % of 20 examples)
    - **Rentabilité estimée** in [forte] then Décision = **oui** (80,00 % of 10 examples)
- **Rentabilité estimée** in [faible] then Décision = **non** (96,77 % of 31 examples)

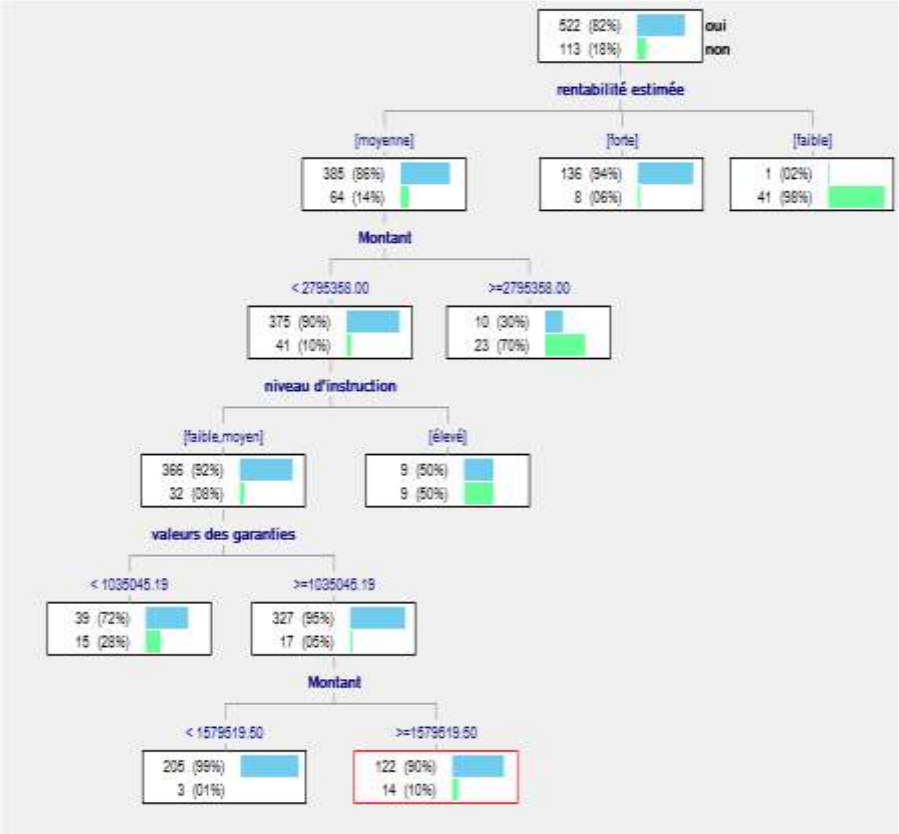
Computation time : 0 ms.
Created at 26/02/2014 15:52:02

Les variables les plus discriminantes mises en évidence par ce logiciel sont : *la rentabilité estimée, le montant de crédit et les valeurs de garanties.* Les modalités de ces variables qui agissent sur la décision sont : *Rentabilité faible, moyenne et forte, montant de crédit entre 76 et 215 millions et ainsi qu'une somme de garantie supérieure à 105 millions*. De ce fait, un montant de crédit inférieur à 76 millions a plus de chance d'être accepté qu'un montant plus élevé. De plus les projets à rentabilité moyenne et forte sont susceptibles d'être acceptés.

5.3 Logiciel Sipina (méthode C4.5) :

Les variables qui possèdent un bon pouvoir discriminant selon Sipina sont : *La rentabilité estimée du projet, le montant de crédit et les valeurs de garanties ainsi que le niveau d'instruction.*

ISSN : 1112-234X



La capacité prédictive est déduite de la matrice de confusion générée à partir de l'échantillon test. L'estimation du taux d'erreur moyen de classification en test obtenue avec les trois méthodes d'induction utilisées selon les deux techniques d'évaluation est résumée dans le tableau **Tab2**. Il faut remarquer que sous Sipina la procédure n'est pas totalement automatisée. Nous avons réalisé une partie des calculs manuellement pour retrouver le taux moyen de mauvaise classification en validation croisée. Cependant, le logiciel Tanagra ne fournit pas l'arbre de décision en validation croisée, mais on peut l'utiliser pour contrôler expérimentalement le risque de sur-apprentissage.

On remarque cependant à partir des matrices de confusion qu'en moyenne 90% des emprunteurs de l'échantillon test ont été bien classés. Le taux d'erreurs en apprentissage/test est plus optimiste qu'on validation croisé, il est estimé en moyenne respectivement à 9% en apprentissage et 10,5% en validation croisée.

Les résultats sont très satisfaisants, ce qui confirme un pouvoir discriminant assez important. De plus, les modèles fournissent des résultats comparables. On peut donc, considérer qu'il y'a quatre variables qui ont influencés et conditionnés la prise de décision :

- Niveau d'instruction

- Rentabilité : Les projets à rentabilité faible sont refusés d'office, ceux dont la rentabilité est moyenne ont moins de chance d'être acceptés que les projets à forte rentabilité.

- Montant du crédit : Une demande de crédit à montant relativement faible, a plus de chance d'être accepté qu'une demande de crédit à montant élevé.

ISSN : 1112-234X

- La garantie relative au crédit : Des valeurs de garanties suffisantes sont exigées.

CONCLUSION ET PERSPECTIVES

En conclusion, Nous avons opté pour le modèle d'arbre de décision pour ses caractéristiques intéressantes. D'abord, il hiérarchise les variables par ordre d'importance dans l'arbre, celles qui sont situées au plus prés de la racine sont les plus influentes. Ensuite, il synthétise de manière intelligible et visuelle le résultat de l'analyse. Enfin, Une branche de l'arbre (associée à un vecteur de modalités) représente une règle de décision et elle est définie par les modalités les plus efficientes.

Lors de cette étude, nous avons cherché à extraire les variables les plus pertinentes qui ont orienté le décideur dans sa décision de prêt. A cette fin, nous avons décidé de comparer trois méthodes d'arbres de décision, les méthodes CART, C4.5 et CHAID. L'application des diverses méthodes d'arbre de décision, nous l'avons réalisée dans le but de tester leur robustesse vis-à-vis des données de l'étude. De plus, afin de fournir un élément de comparaison supplémentaire, nous avons mené des tests pratiques sur trois logiciels libres. Le protocole d'évaluation adopté a consisté à comparer leurs résultats.

Avec les trois méthodes, l'ensemble des règles obtenues ne présentent pas d'incohérence et concordent dans leur globalité à quelques exceptions insignifiantes et qui ne concernent que très peu d'individus. En effet, les taux de mauvais classement découlant des différentes méthodes sont très proches. Le lien entre la variable cible et les différentes variables explicatives, a confirmé que la décision de crédit dépend principalement du niveau d'instruction du client, des crédits demandés ainsi que de la rentabilité du projet à réaliser. Enfin, pour vérifier le bien-fondé des modalités induites par le modèle, il serait intéressant d'interroger les experts métiers afin de confirmer la pertinence des variables d'influence sur la décision de prêt. De plus pour que les arbres et les règles soient plus significatifs, il est préconisé d'utiliser une base de données plus importante.

RÉFÉRENCES

[Brei & al 1984] Breiman L. & Friedman J.H. & Olshen R.A. & Stone C.J. " Classification and Regression Trees " Wadsworth Publishing Company, 1984.

[Hang & al 01] Hand D., Manilla H., Smyth P., Principles of data mining, Bardford Books, 2001.

[Hart 1975] Hartigan J. A. " Clustering algorithms ", Editions John Wiley & Sons. 1975.

[Kass 1980] Kass G. V. " An exploratory technique for investigating large quantities of categorical data ". Applied Statistics, pp.119–127. 1980.

[Morg & Sonq 1963] Morgan J.J. & Sonquist J.A. " Problems in the Analysis of Survey data and proposal. ". Journal of the American Statistical Association, pp.415-435. 1963.

[Quin 1986] Induction of decision trees. Machine Learning 1, 81–106.

[Quin 1993] Quinlan J.R. "C4.5 : Programs for Machine Learning" Morgan Kaufmann, 1993.

ISSN : 1112-234X

[Tuf 2010] Tufféry S. " Data Mining et statistiques décisionnelle.

L'intelligence des données. " TECHNIP, France, 2010.

[Wux & al 08] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang

Yang, Hiroshi Motoda, Geoffrey McLachlan, Angus Ng, Bing Liu, Philipu, Zhi-Hua Zhou, Michael Steinbach, David Hand, and Dan Steinberg. Top 10 algorithms in data mining. Knowledge and Information Systems,:1–37, Jan 2008.

[Zig & Rako 00] Zighed, D. et R. Rakotomalala (2000). Graphes d'induction. France: Hermes.

[Zig & al 1992] Zighed D., Auray J., Duru G., SIPINA : Méthode et Logiciel, Lacassagne, 1992.