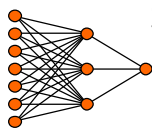


Stéphane Tufféry

Statisticien - Data Miner - Formateur



DATA MINING - SCORING



STATISTIQUE APPLIQUÉE

APPLICATION AU CRM



04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



1

Plan du cours

- Qu'est-ce que le data mining ?
- A quoi sert le data mining ?
- Les 2 grandes familles de techniques
- *Le déroulement d'un projet de data mining*
- *Coûts et gains du data mining*
- *Facteurs de succès - Erreurs à éviter*
- *Informatique décisionnelle et de gestion*
- La préparation des données
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- Logiciels et consultants
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



2

Le déroulement d'un projet de data mining

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



3

Les 10 étapes d'un projet

- Choix du sujet - Définition des objectifs
- Inventaire des données existantes
- Collecte, nettoyage et mise en forme des données
- Étude statistique de la base d'analyse
- Mise en œuvre des algorithmes (classification, scoring...)
- Élaboration des modèles
- Validation et choix d'un modèle
- Déclaration à la CNIL
- Déploiement du modèle
- Formation des utilisateurs
- Suivi des modèles

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



4

Définition des objectifs

- Le sujet retenu doit bien sûr requérir des outils de data mining (et pas de simples statistiques descriptives).
- Le sujet, la population ciblée et les objectifs doivent être précisément définis.
- Les objectifs doivent être réalistes (tenir compte des actions passées et de la saturation du marché).
- L'entreprise doit avoir au moins un minimum de connaissance du sujet.
- Le sujet doit faire partie des objectifs de l'entreprise et lui apporter un avantage réel.
- L'entreprise doit avoir la volonté et la possibilité de mettre en œuvre les solutions qui seront proposées par le data mining (vérifier les possibilités de la production).

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



5

Définition de la cible

- Définir la population cible
 - tous les clients, les clients « actifs », les clients « actifs » et sans risque...
 - unité statistique : individu, famille...
- Définir certains critères essentiels (variable cible)
 - tels que « client à risque » et « client sans risque »
- Prévoir l'utilisation opérationnelle des modèles produits
- Spécifier les résultats attendus
 - sous quelle forme faut-il fournir les résultats ? (cela dépend de leur utilisation et de leurs utilisateurs)
 - confidentialité de la restitution du score (un commercial peut-il voir les scores de tous les clients ou seulement des siens ?)
 - périodicité de mise à jour des données

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



6

Recensement des données utiles

- Il faut recenser, avec les spécialistes métier et les informaticiens, les données utiles :
 - accessibles (internes ou externes à l'entreprise)
 - fiables
 - suffisamment à jour
 - historisées, si besoin est
 - légalement utilisables
- Il y a les données :
 - du système d'information (SI) de l'entreprise
 - stockées dans l'entreprise, hors du SI (fichiers Excel...)
 - achetées ou récupérées à l'extérieur de l'entreprise
 - calculées à partir des données précédentes (indicateurs, ratios, évolutions au cours du temps)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



7

Quand on manque de données

- Enquêtes auprès d'échantillons de clients
 - en les incitant à répondre à des questionnaires en leur proposant des cadeaux
- Utilisation des mégabases de données (Consodata, Claritas)
- Géomarketing (type d'habitat en fonction de l'adresse)
 - données moins précises que des données nominatives
 - mais disponibles pour des prospects
- « Scoring prénom »
- Recours à des modèles standards pré-établis par des sociétés spécialisées (ex : scores génériques)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



8

Géomarketing

- Données économiques
 - nb entreprises, population active, chômage, commerces et services de proximité, habitudes de consommation...
- Données sociodémographiques
 - population, richesse, âge et nombre d'enfants moyens, structures familiales, niveau socioprofessionnel...
- Données résidentielles
 - ancienneté, type et confort des logements, proportion de locataires et propriétaires...
- Données concurrentielles
 - implantation de l'entreprise, implantation de ses concurrents, parts de marché, taux de pénétration...
- TMÎlotypes : beaux quartiers, classe moyenne, classe ouvrière, centre ville et quartiers commerçants...

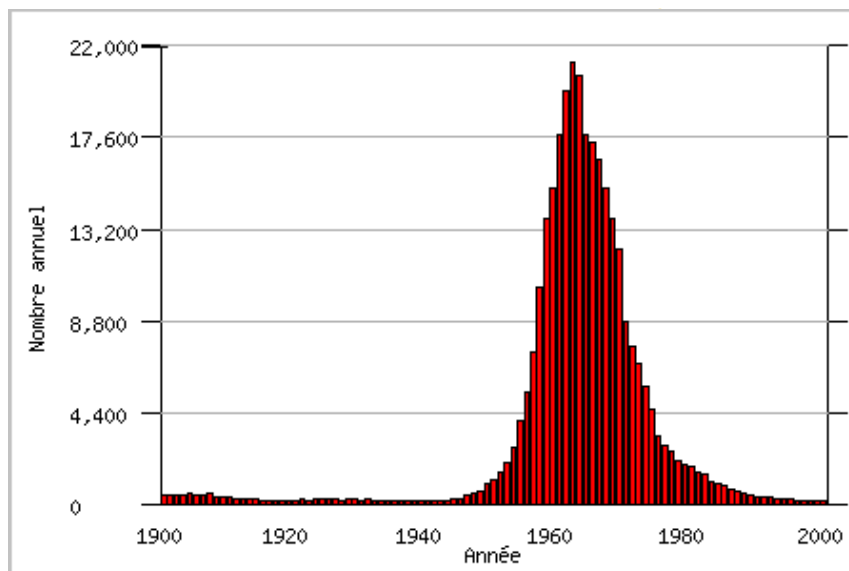
04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



9

Scoring prénom (ex : Pascal)



04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



10

Construction de la base d'analyse

no client	variable "cible" : souscripteur (O/N)	âge	CSP	situation famille	ilotype	revenus	...	var. explicative m	échantillon
client 1	O	58	cadre	marié	A	50 000	apprentissage
client 2	N	27	ouvrier	célibataire	A	30 000	test
...
client k	O	46	technicien	célibataire	B	40 000	test
...
client 2000	N	32	employé	marié	C	25 000	apprentissage
...

au moins 2000 enregistrements

données année $n+1$ données à fin année n répartition aléatoire des clients entre les 2 échantillons

↑
O : au moins 1000 clients ciblés dans l'année $n+1$ et acheteurs
N : au moins 1000 clients ciblés dans l'année $n+1$ et non acheteurs

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

11

Types de données 1/3

- Données de transaction et RFM
 - « où » (lieux des transactions, Internet...), « quand » (fréquence/récence des transactions), « comment » (mode de paiement), « combien » (nombre et montants des transactions), « quoi » (ce qui est acheté)
- Données sur les produits et contrats
 - nb, types, options, prix, date d'achat ou de souscription, date et motif de résiliation ou de retour du produit, durée moyenne de vie ou date d'échéance, délai et mode de paiement, remise accordée au client, marge de l'entreprise
- Anciennetés
 - âge, ancienneté comme client, ancienneté à l'adresse actuelle, ancienneté dans l'emploi, ancienneté du dernier sinistre (en assurance)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

12

Types de données 2/3

- Données sur les canaux
 - canal de prise de contact (parrainage, annonce presse, appel téléphonique, réponse à un mailing...)
 - canal privilégié de contact et communication (courrier, téléphone, Internet, magasin/agence...)
 - canal privilégié de commande (courrier, téléphone, Minitel, Internet, magasin/agence...)
 - canal privilégié de livraison (magasin/agence, domicile...)
- Données relationnelles et attitudinales
 - réactions aux propositions commerciales, réponses aux questionnaires, réponses aux enquêtes de satisfaction, appels au service clientèle, réclamations
 - image de la marque auprès du client, attractivité des concurrents, propension ou inertie du client au changement

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



13

Importance des retours

- Le data mining ne devine pas le profil des clients à cibler, il l'extrapole à partir des données fournies.



- Pour les études d'appétence, les retours des actions commerciales précédentes (refus d'achat) permettent de dégager les profils positifs et négatifs
 - > Il est capital de mémoriser cette information.

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



14

Types de données 3/3

- Données sociodémographiques
 - familiales (situation de famille, nb d'enfants et leur âge, nombre de personnes à charge)
 - professionnelles (salaire, PCS, nb d'actifs dans le ménage)
 - patrimoniales (patrimoine mobilier et immobilier, statut de propriétaire/locataire, valeur du logement, possession d'une résidence secondaire...)
 - géographiques (ancienneté à l'adresse, code INSEE de la commune, IRIS et ilot INSEE, type d'habitat déduit de l'IRIS ou de l'ilot)
 - environnementales et géomarketing (concurrence, population, population active, population cliente, taux de chômage, potentiel économique, taux de détention de produit... dans la zone d'habitation du client)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



15

Données à ne pas utiliser

- Non légalement utilisables
- Non fiables
 - trop de valeurs fausses ou inconnues
- Disponibles sur une durée trop courte
 - soumises aux variations saisonnières
- Redondantes
 - dont le poids est artificiellement augmenté, ou dont la colinéarité rend instable les résultats de certaines méthodes
- Non pertinentes
 - qu'il faut remplacer par de nouveaux indicateurs
- Trop corrélées à l'objectif de l'étude
 - qui entraînent un «sur-apprentissage» dans les prédictions
- Trop peu corrélées à l'objectif de l'étude
 - qui créent du « bruit », des fluctuations aléatoires

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



16

Sélection des données à utiliser

- Choix des variables les plus discriminantes
 - test de la variance (ANOVA, ANOVA Welch, Kruskal-Wallis)
 - indicateur η^2 ou ρ^2
 - utilisation d'un arbre CHAID ou CART
- Choix des discrétisations
 - utilisation d'un arbre CHAID
- Choix des interactions
 - utilisation d'un arbre CART ou CHAID
- Choix des variables les moins corrélées entre elles
 - test de multicolinéarité
- Transformation des variables (recodage, normalisation par un log ou une racine carrée)
 - permet de se rapprocher d'une loi normale

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



17

Sélection des variables (variable cible qualitative)

- Test du chi-2 (var. explicative qualitative)
- Test de la variance (var. explicative quantitative)
 - ANOVA (normalité - homoscédasticité)
 - Welch ANOVA (normalité - hétéroscédasticité)
 - Kruskal-Wallis (non normalité - hétéroscédasticité)
- Indicateur η^2 (var. explicative quantitative)
 - somme des carrés entre classes / somme totale des carrés
 - représente la proportion de la variance de la variable indépendante expliquée, linéairement ou non, par la variable cible
 - généralise le ρ^2 (ρ = coefficient de corrélation de Pearson) qui ne prend en compte que la part linéaire
- Arbre de décision

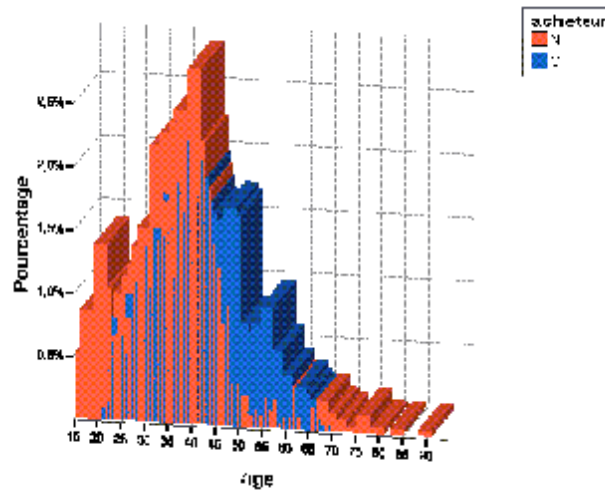
04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



18

Présence de 3 classes

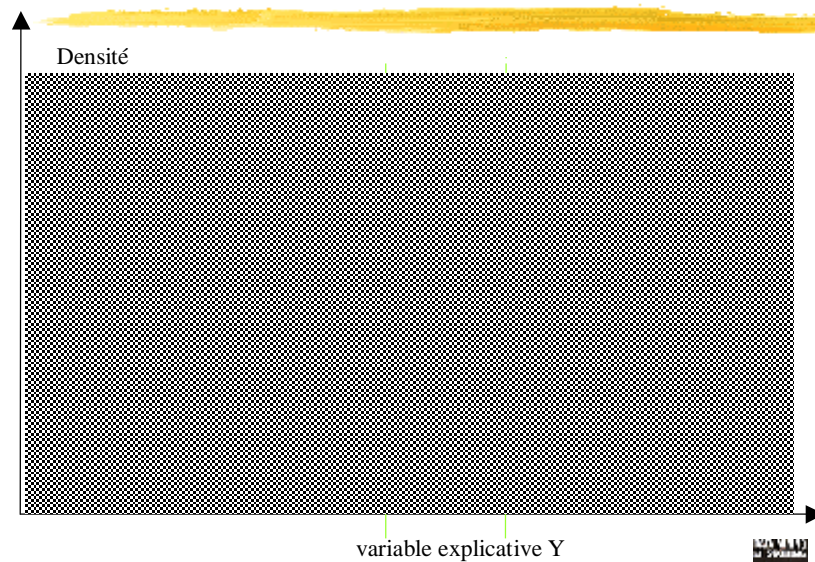


04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

19

Discrétisation en 3 tranches naturelles



04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

20

Méthodes inductives : 4 étapes

- Apprentissage : construction du modèle sur un 1^{er} échantillon pour lequel on connaît la valeur de la variable cible
- Test : vérification du modèle sur un 2^d échantillon pour lequel on connaît la valeur de la variable cible, que l'on compare à la valeur prédite par le modèle
 - Si le résultat du test est insuffisant (d'après la *matrice de confusion* ou la courbe *ROC*), on recommence l'apprentissage.
- Validation du modèle sur un 3^e échantillon, pour avoir une idée du taux d'erreur non biaisé du modèle
- Application du modèle à l'ensemble de la population



valeur prédite \hat{e}	A	B	TOTAL
valeur réelle e			
A	1800	200	
B	300	1700	
TOTAL			4000

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

23

Exemples de modèles prédictifs

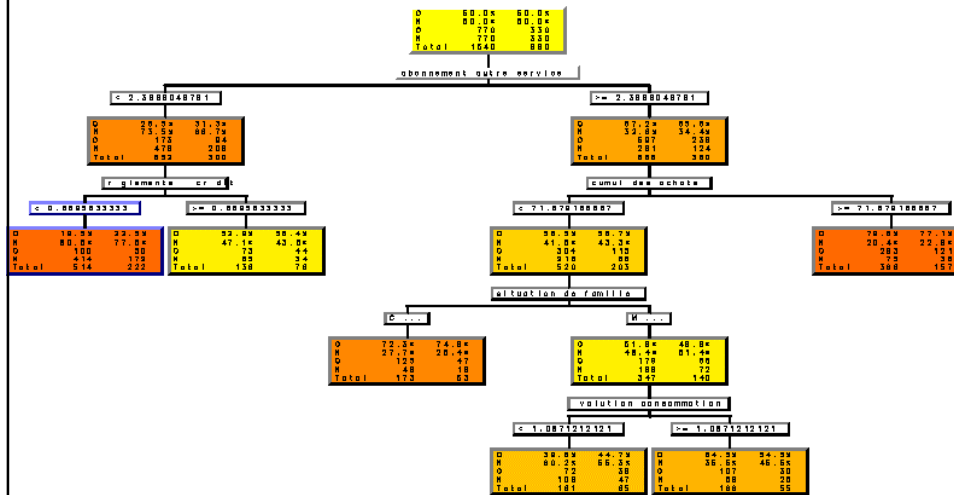
- Arbres de décision
 - Règles complètement explicites
 - Traitent les données hétérogènes, éventuellement manquantes, sans hypothèses de distribution
 - Détection de phénomènes non linéaires
 - Moindre robustesse
- Analyse discriminante linéaire
 - Résultat explicite $P(Y | X_1, \dots, X_p)$ sous forme d'une formule
 - Requier des X_i continues, sans colinéarité, et des lois $\{X_i\}/Y$ multinormales et homoscédastiques (NB aux « outliers »)
 - Optimale si les hypothèses sont remplies
- Régression logistique
 - Comme l'analyse discriminante, sans hypothèse sur les lois X_i/Y , X_i peut être discret, avec une précision parfois un peu inférieure

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

24

Algorithme d'arbre de décision

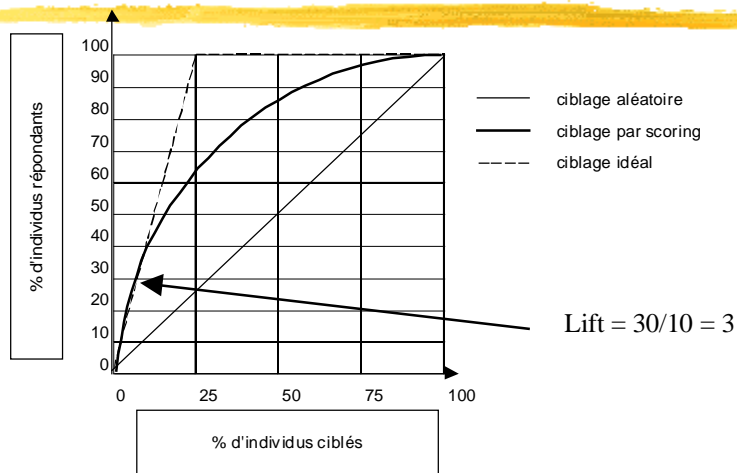


04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

25

Validation d'un modèle de score



COURBE DE CONCENTRATION

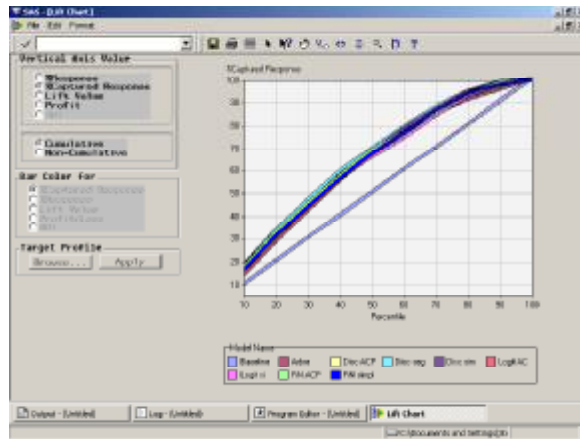
04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

26

Comparaison de modèles de score

- Les indicateurs statistiques de 2 modèles de nature différentes (exemple : R^2 d'une analyse discriminante et d'une régression logistique) sont rarement comparables
- On compare les modèles à l'aide des courbes de concentration, de lift ou ROC

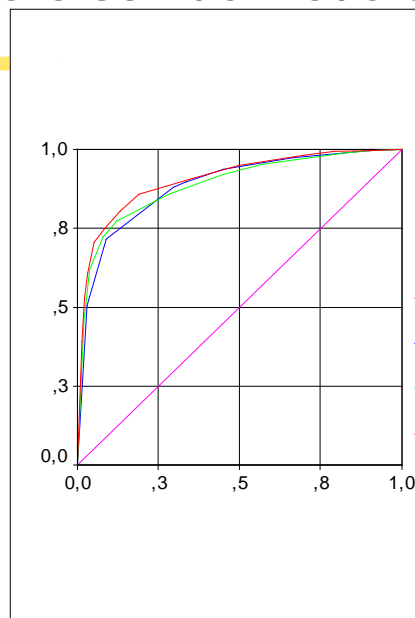


04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

27

Comparaison de modèles : courbe ROC



Zone sous la courbe

Variable(s) de	Zone
régression logistique	,906
analyse discriminante	,889
arbre de décision	,887

Source de la courbe

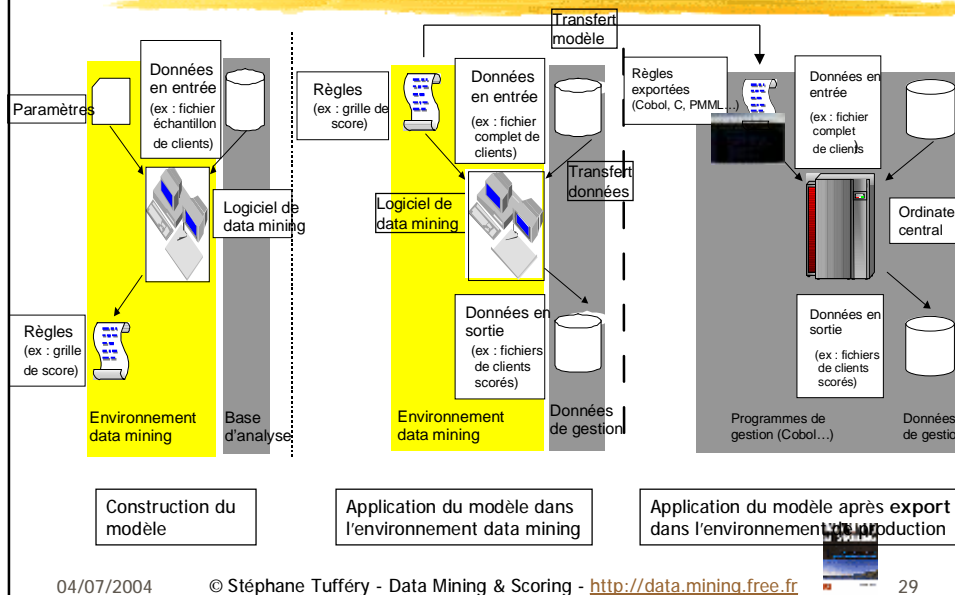
- Ligne de référence
- arbre de décision
- analys discriminante
- régress. logistique

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

28

Déploiement des modèles



Intégration des modèles dans le SI

- Plusieurs possibilités ne s'excluant pas :
 - intégration dans les fichiers clients de production et sur le poste de travail des commerciaux
 - intégration dans les fichiers clients de production ou d'infocentre
 - utilisation d'un tableur sur un PC pour réaliser un publipostage (marketing direct)
- Différents niveaux de finesse de l'information
 - notes fines dans les fichiers de production (par ex : de 1 à 1000)
 - restituées agrégées sur le poste de travail (par ex : de 1 à 10)
 - et regroupées en tranches (par ex : faible / moyen / fort)

Utilisation opérationnelle d'un score

revenus

couleur	% populat.	0 - 1000 €	1001 - 2 000 €	2 001 - 3 000 €	> 3 000 €
rouge	20 %	rien	rien	rien	rien
orange	20 %	rien	découvert = 150 € carte = débit imméd. prêt pers. = 0 €	découvert = 450 € carte = débit imméd. prêt pers. = 3000 €	découvert = 750 € carte = débit diff. prêt pers. = 4500 €
vert (« foncé »)	20 %	découvert = 150 € carte = retrait prêt pers. = 0 €	découvert = 450 € carte = débit imméd. prêt pers. = 3000 €	découvert = 1500 € carte = débit diff. prêt pers. = 9000 €	découvert = 1500 € carte = débit diff. prêt pers. = 12 k€
vert (« clair »)	40 %	découvert = 150 € carte = débit immédiat prêt pers. = 0 €	découvert = 750 € carte = débit diff. prêt pers. = 7500 €	découvert = 1500 € carte = débit diff. prêt pers. = 12 k€	découvert = 1500 € carte = <i>Gold</i> prêt pers. = 15 k€

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

31

Formation des utilisateurs

- Présenter l'objectif recherché avec les nouveaux outils
- Principe et fonctionnement des outils de data mining
 - sans entrer dans les détails techniques
- Limites des outils
 - il ne s'agit que d'outils statistiques
- Mode d'utilisation
 - *aide à la décision* et non pas prise automatique de décision
- Apport des outils (c'est le point le plus important)
- Ce qui change dans le travail des utilisateurs
 - du point de vue opérationnel
 - du point de vue organisationnel (adaptation des procédures, des délégations de pouvoir...)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

32

Cycle de vie d'un score

- Les outils de data mining (scores surtout) ont une phase d'expérimentation
 - sur une petite échelle
 - destinée à les ajuster et valider, et tester leur utilisation
- Quand les outils sont en production, ils doivent être appliqués régulièrement à des données rafraîchies
- Les outils en production doivent être revus régulièrement (tous les 1 à 3 ans)
 - évolution de l'environnement concurrentiel, économique, sociodémographique, réglementaire
 - apparition, disparition, modification de produits

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



33

Suivi du score

- Suivi ponctuel pour une campagne marketing
 - pour analyser les résultats et améliorer le score suivant
 - comparer les résultats des individus ciblés à ceux d'un échantillon témoin (cible aléatoire ou traditionnelle)
- Suivi permanent pour l'utilisation commerciale
 - en comparant les résultats des \neq niveaux de score
 - vérifier la bonne utilisation du score
 - s'assurer de la pertinence des « infractions » au score
 - vérifier le bon fonctionnement du score
 - pour un score de risque, le taux de défaillance dans chaque tranche de score doit rester à l'intérieur d'une fourchette fixée

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



34

Suivi du score

mois score	M-1	M-2	M-3	M-4	M-5	M-6	M-7	M-8	M-9	M-10	M-11	M-12
1												
2												
3												
...												
TOTAL												

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



35

Suivi du score : matrice de transition

score M-1 è	1	2	3	...
è score M-2				
1				
2				
3				
...				

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



36

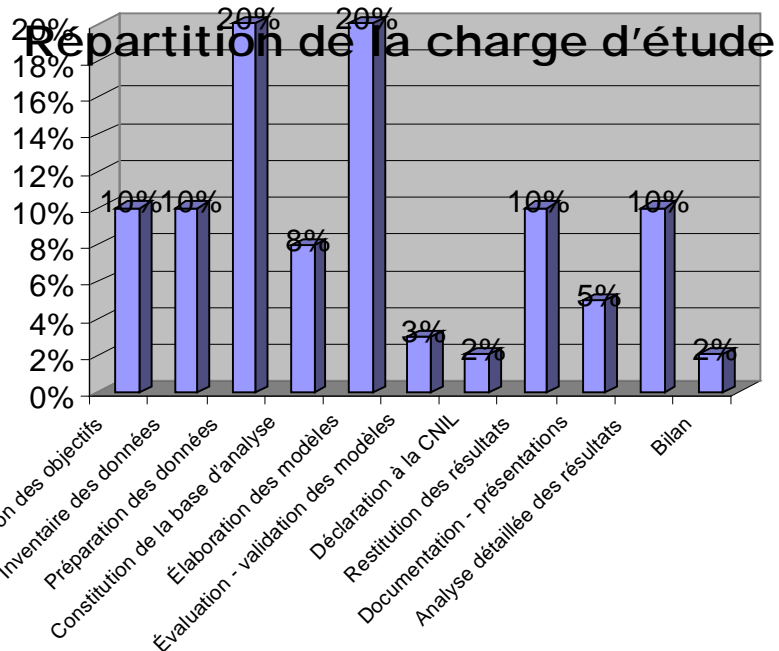
Coûts et gains du data mining

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



37



04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



38

Coûts

- Distinguer : coûts du data warehouse et du data mining
- Investissement initial dans un DW : environ 1 M€
- Dépend des choix techniques :
 - machine dédiée au décisionnel ou non
 - logiciels d'alimentation de l'entrepôt ou développements « maison »
 - modèle de données acheté ou élaboré en interne
- Coûts du data mining très inférieurs :
 - coûts humains inférieurs car équipe dédiée au DM + petite
 - coût des logiciels entre 2 k€ (sur PC mais avec plusieurs algorithmes) et 150 k€ (sur gros systèmes, intégrés à l'informatique de production)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



39

Le retour sur investissement

- Le RSI est difficile à évaluer :
 - les gains proviennent du data mining mais aussi d'une bonne communication, d'un marketing efficace, de commerciaux motivés
 - le DM n'est qu'une brique dans le marketing de bases de données (exemple du crédit préaccordé)
- Le RSI vient de :
 - l'augmentation des taux de réponse des actions marketing
 - l'augmentation de la productivité des commerciaux
 - la meilleure utilisation des canaux
 - la fidélisation des clients
 - la réduction des impayés...
- On peut tenter de l'estimer avec un échantillon témoin

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



40

Exemple de calcul de RSI

		ciblage tradit.	ciblage DM
A	nombre de clients ciblés	30 000	15 000
B	coût de chaque mailing	1 €	1 €
C	coût de chaque relance téléphonique	5 €	5 €
D	coût total (= A x (B + C))	180 000 €	90 000 €
E	nombre de nouvelles souscriptions	1 000	1 500
F	taux de souscription (= E / A)	3,33 %	10 %
G	coût par souscription (= D / E)	180 €	60 €
H	chiffre d'affaire annuel par souscription	150 €	175 € (montants souscrits supérieurs)
I	CA total annuel (= H x E)	150 000 €	262 500 €
	RSI (= I / D)	83 %	292 %

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



41

RSI d'un score d'attrition


A	coût d'acquisition d'un nouveau client	150 €
B	rentabilité annuelle des partants	450 €
C	temps d'activation d'un client	0,5 an
D	perte occasionnée par un départ (= A + (B x C))	375 €
E	coût de fidélisation d'un « partant » détecté	50 €
F	nombre total de clients	1 000 000
G	nombre de partants annuels	80 000
H	taux d'attrition = (G / F)	8 %
I	nombre de « partants » détectés (à tort ou à raison)	40 000
J	coût total de la fidélisation (= E x I)	2 000 000 €
K	nombre de vrais partants retenus	8 000
L	pertes évitées (= D x K)	3 000 000 €
	gain total net (= L - J)	1 000 000 €

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



42



Les facteurs de succès et les erreurs à éviter

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



43

Les facteurs de succès d'un projet



- Des objectifs précis, stratégiques et réalistes
- La qualité et la richesse des informations collectées
- Le stockage des informations relationnelles sur les clients (réponses aux sollicitations commerciales, aux enquêtes de satisfaction, canaux de prédilection...)
- La collaboration des compétences métiers et statistiques
- La maîtrise des techniques de data mining utilisées
- Une bonne restitution des résultats et l'implication de tous les partenaires chargés de leur mise en œuvre
- L'analyse des retours de chaque action pour la suivante

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



44

Le DM dans la culture d'entreprise

- L'entreprise doit veiller :
 - à ses compétences en data mining
 - à la qualité des données recueillies
 - à une mise en œuvre et un suivi rigoureux des actions s'appuyant sur le data mining
 - à une éventuelle adaptation de ses processus marketing
 - passer du marketing « produit » au marketing « client »
 - à une éventuelle adaptation de ses processus de décision
 - adaptation des délégations de pouvoir
- Le data mining est un processus itératif, chaque action préparant la suivante par l'exploitation de ses résultats.

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



45

« Vendre » le data mining

- Les commerciaux et analystes peuvent voir une mise en cause de leur savoir-faire
- Il faut les convaincre que le scoring ne fournit qu'une aide à la décision, et non la décision elle-même
 - laquelle reste toujours leur prérogative, comme l'exige la loi « Informatique et libertés »
- Il faut aussi les convaincre de bien alimenter les bases de données marketing
 - notamment en ce qui concerne les retours des campagnes : refus
- Ils doivent être sensibilisés au gain de productivité et de sécurité qu'ils peuvent attendre du scoring

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



46

Les idées fausses sur le DM

- Aucun a priori n'est nécessaire
- On n'a plus besoin de spécialistes du métier
- On n'a plus besoin de statisticiens (« Il suffit d'appuyer sur un bouton »)
- Le data mining permet de faire des découvertes incroyables
- Le data mining est révolutionnaire
- Il faut utiliser toutes les données disponibles
- Il faut toujours échantillonner
- Il ne faut jamais échantillonner

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



47

« Aucun a priori n'est nécessaire »

- Les **techniques prédictives** requièrent un *a priori* :
 - puisqu'il faut choisir une variable cible, soigneusement définie
- Certaines **techniques descriptives**, telle la classification, peuvent être mises en œuvre sans savoir quelles seront les classes obtenues, ni même quel est le nombre pertinent de classes
- Mais :
- Le résultat de la classification est influencé par le choix des données et de leur codage en entrée de l'algorithme :
 - > il est donc impossible d'être totalement neutre même dans une classification

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



48

« On n'a plus besoin de spécialistes du métier »

- Le concours des spécialistes (métier, marketing) est indispensable dans plusieurs phases :
 - la **définition des objectifs**
 - par exemple, avant l'élaboration d'un score de risque, il convient de s'entendre sur la définition précise d'un *risque*
 - le **recensement des données** utiles et légalement utilisables, données brutes et données composées
 - il est intéressant de connaître les données considérées comme pertinentes par les spécialistes
 - l'**analyse des résultats**
 - le spécialiste métier peut, au vu des 1ers résultats, dire s'ils paraissent triviaux, nouveaux et intéressants à creuser, ou surprenants et très suspects, auquel cas il faudra vérifier la validité des données et des méthodes utilisées

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



49

« On n'a plus besoin de statisticiens »

- Dans une étude de DM, la partie la + longue et la + déterminante est le **travail des données**.
 - Elle ne peut être effectuée qu'au vu d'analyses statistiques permettant de vérifier la fiabilité des données, leurs distributions, leurs corrélations... et de réaliser les mises en forme de données préalables. Ces opérations ne seront pas réalisées à l'identique pour tous les algorithmes de DM
 - Certains algos nécessitent un **échantillonnage** préalable
- Dans les algorithmes prédictifs, il faut prendre garde de ne pas inclure parmi les variables explicatives des variables corrélées *par définition* à la variable cible. Il faut se méfier du phénomène de **sur-apprentissage**
- Le **paramétrage fin** des algorithmes peut avoir une grande incidence sur les résultats obtenus

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



50

« Le data mining permet de faire des découvertes incroyables »

- Les règles mises à jour par le data mining sont rarement incroyables : elles font souvent intervenir des variables considérées comme discriminantes par les spécialistes, d'une façon conforme au bon sens.
- **Où réside donc l'apport du data mining ?**
 - > Dans le fait qu'il existe des milliers de combinaisons, conformes au bon sens, de variables *a priori* discriminantes dans une problématique donnée...
 - > ... et que le data mining permet de détecter LA meilleure combinaison possible (ou l'une des meilleures), avec, pour chacune de ces variables X , la meilleure valeur précise n à tester (« si $X \leq n$, alors... »)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



51

« Le data mining est révolutionnaire »

- Le data mining englobe l'analyse des données traditionnelle, dont il ne diffère que par les points suivants :
 - certaines techniques de DM n'appartiennent qu'à lui, comme les réseaux de neurones et les arbres de décision
 - le nombre d'individus étudiés et le nombre de variables est souvent beaucoup plus important en DM, où l'optimisation des algorithmes est importante
 - les modèles en DM sont plus souvent des ensembles de règles locales que des modèles globaux et cohérents
 - le DM recherche parfois plus la compréhensibilité des modèles que leur précision.

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



52

« Il faut utiliser toutes les données disponibles »

- Un algorithme de DM est-il d'autant + efficace qu'il a + de données en entrée ? **NON !**
- Les données non fiables ou mal renseignées perturbent tous les algorithmes
- Les données redondantes peuvent affecter une classification
- La présence d'individus hors-norme perturbe les modèles linéaires
- Les données avec des modalités aux effectifs irréguliers affectent les analyses factorielles
- Les données peu discriminantes ou colinéaires diminuent le pouvoir prédictif d'une analyse discriminante
- Les données trop nombreuses affectent les réseaux de neurones

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



53

« Il faut toujours échantillonner »

- Un bon échantillonnage est toujours délicat à réaliser, et nécessite une bonne connaissance de la population
 - difficile à avoir, surtout avec les populations instables que sont les clientèles
- Exemple d'inconvénient induit par l'échantillonnage :
 - un écart de distribution d'une variable dans l'échantillon d'apprentissage par rapport à la population totale, peut produire des écarts importants dans les résultats
- Autre contre-indication au recours à l'échantillonnage : la recherche de phénomènes rares (typologies de fraude) ou de segments étroits de clientèle

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



54

« Il ne faut jamais échantillonner »

- Certaines techniques de data mining, les **techniques de prédiction inductives** (arbres de décision, réseaux de neurones à rétropropagation supervisée...), **imposent le recours à l'échantillonnage**
 - puisqu'elles procèdent par élaboration d'un modèle à partir d'une partie de la population,
 - modèle ensuite testé sur une autre partie de la population
- Il peut aussi être souhaitable de travailler sur un échantillon de la population, si celle-ci est très grande, afin de **limiter des temps de traitement informatique prohibitifs**

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



55

7 pistes d'amélioration sur le MBDD

- Mémoriser les résultats des campagnes commerciales
- Sensibiliser les commerciaux à l'importance des données saisies
- Acheter des données externes (INSEE, Consodata...)
- Compléter ces données par des enquêtes auprès de clients et de commerciaux
- Préciser la définition et améliorer le calcul de données stratégiques : rentabilité, fidélité...
- Créer de nouvelles variables synthétiques pertinentes
- Augmenter la profondeur de l'historique des données.

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



56

Les spécificités de l'informatique décisionnelle par rapport à informatique de gestion

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



57

Les développements informatiques

- Le décisionnel est un processus exploratoire et itératif
 - mise en place de l'environnement technique du DM en parallèle des études de DM elles-mêmes
- Les projets ne sont pas parfaitement définis dès le début et strictement encadrés par un cahier des charges
 - seuls sont définis : objectifs, délais, acteurs, restitution
 - modèles statistiques inconnus au départ
 - variables discriminantes inconnues au départ
 - techniques de modélisation pas toujours choisies au départ
- > Nécessité d'une informatique réactive et souple
 - solution : équipe informatique dédiée au décisionnel, avec des ressources propres (serveur, espace disque...)
 - tout en conservant un souci d'industrialisation

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



58

Les applications informatiques

- Une application de *data mining* est pour l'essentiel un *modèle*, c.a.d. un ensemble d'instructions permettant :
 - soit de fournir en sortie une donnée qui résume des données en entrée
 - soit de fournir en sortie une donnée d'un nouveau type, utile pour la prise de décision et déduite des données en entrée
- Le 1^{er} cas de figure correspond au *data mining* descriptif, dont l'archétype est la *classification* : l'appartenance d'un individu à une classe résume l'ensemble de ses caractéristiques
- Le 2^d cas de figure correspond au *data mining* prédictif, dont l'archétype est le *scoring* : la nouvelle variable est une probabilité que le client ait un certain comportement (de risque, de consommation...)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



59

Les applications informatiques

- Comme une application informatique, une application de *data mining* connaît plusieurs phases :
 - développement (élaboration du modèle)
 - test (vérification des performances du modèle)
 - exploitation dans l'environnement de Production (application du modèle aux données de production pour obtenir la donnée prévue en sortie)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



60

Particularités des applications de DM

1/2

- La phase de développement s'effectue conformément à des données, contrairement à un développement informatique qui s'effectue conformément à un cahier des charges
- Développement et tests se déroulent dans le même environnement, seuls différant les jeux de données
- Un modèle optimal nécessite des va-et-vient entre tests et développement, ces va-et-vient étant pris en charge de façon largement automatique par certains logiciels pour éviter des pertes de temps inutiles et fastidieuses

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



61

Particularités des applications de DM

2/2

- L'analyse des données pour le développement et les tests est réalisée à l'aide d'un logiciel spécifique
- Ces logiciels proposent aussi souvent d'effectuer l'exploitation du modèle
- La concision des modèles de *data mining* : contrairement aux instructions d'un programme informatique (souvent relativement nombreuses), les instructions d'un modèle de *data mining* sont presque toujours peu nombreuses
 - la concision figure parmi les qualités recherchées d'un modèle (car elle va de pair avec sa lisibilité et sa robustesse)

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



62

Export des modèles

- Solution 1 : exploitation des modèles dans le même environnement et avec le même logiciel qu'en développement
 - « les données vont aux modèles »
 - prévoir transfert de données de l'environnement de production vers le décisionnel
- Solution 2 : implémentation des modèles dans un environnement d'exploitation distinct de celui de développement
 - « les modèles vont aux données »
 - prévoir export des modèles de l'environnement décisionnel vers la production (en Cobol, C, PMML (Predictive Modeling Markup Language, 1998) dérivé du XML)
 - solution plus performante pour traiter périodiquement de grandes masses de données

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

63



La gestion des données

- Mettre en cohérence et consolider dans une base de données marketing unique
 - des données provenant de multiples canaux de contact et de distribution
 - pour en tirer une vision unique et complète du client
- Conserver les données des actions commerciales passées
 - qui a été contacté, par quel canal, qui a répondu, au bout de quel laps de temps, qui a été relancé, combien de fois, qui a accepté, refusé, et quel fut le coût du processus
- Historiser les données qui évoluent dans le temps et dont l'évolution doit être connue
 - parfois plusieurs années de données en ligne
 - besoin de stockage important
 - Wal-mart : 24 téraoctets

04/07/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

64



Projets de data warehouse et DM

- Projets de DW et DM à mener en parallèle
 - ne pas séquentialiser l'entreposage des données et leur utilisation
 - car c'est souvent le DM qui dit ce qui doit être fiabilisé et historisé et comment il faut le faire
- 4 bonnes raisons à l'utilisation rapide du data mining :
 - générer un retour sur investissement rapide qui convaincra de l'intérêt du data mining
 - pointer les données et les indicateurs les plus importants pour construire des modèles pertinents
 - commencer à engranger des historiques sur les actions commerciales
 - développer et entretenir les compétences des personnes

