

Table des matières

Préface, de Gilbert Saporta.....	V
Avant-propos	VII
Table des matières	XI
1. Panorama du data mining	1
1.1. Qu'est-ce que le data mining ?.....	1
1.2. À quoi sert le data mining ?	5
1.2.1. Le data mining par secteur d'activité	6
1.2.2. Le data mining par type d'application	9
1.3. Data mining et statistique	13
1.4. Data mining et informatique	15
1.5. Mise en œuvre du data mining.....	21
2. Le déroulement d'une étude de data mining	23
2.1. Définition des objectifs	23
2.2. Inventaire des données existantes.....	24
2.3. Collecte des données.....	25
2.4. Exploration et préparation des données.....	28
2.5. Segmentation de la population.....	30
2.6. Élaboration et validation des modèles prédictifs	32
2.7. Synthèse des modèles prédictifs des différents segments	33
2.8. Itération des étapes précédentes	34
2.9. Déclaration des traitements à la CNIL.....	35
2.10. Déploiement des modèles.....	35
2.11. Formation des utilisateurs des modèles.....	35
2.12. Suivi des modèles	36
2.13. Enrichissement des modèles	38
2.14. Remarques.....	39
2.15. Cycle de vie d'un modèle	39
2.16. Charges pour un projet pilote	39
3. L'exploration et la préparation des données	41
3.1. Les différents types de données	41
3.2. L'examen de la distribution des variables	42
3.3. La détection des valeurs rares ou manquantes	43

3.4.	La détection des valeurs aberrantes	46
3.5.	La détection des valeurs extrêmes.....	48
3.6.	Les tests de normalité	50
3.7.	Homoscédasticité et hétéroscédasticité	54
3.8.	La détection des variables les plus discriminantes	55
3.8.1.	Variables explicatives qualitatives, discrètes ou découpées en classes.....	56
3.8.2.	Variables explicatives continues.....	59
3.8.3.	Précisions sur les tests non-paramétriques à 1 facteur	63
3.8.4.	ODS et automatisation de la sélection des variables discriminantes.....	67
3.9.	La transformation des variables	70
3.10.	Le choix des tranches de valeurs des variables continues	74
3.11.	La création de nouvelles variables	78
3.12.	La détection des interactions.....	80
3.13.	La sélection automatique des variables	82
3.14.	La détection de la colinéarité.....	83
3.15.	L'échantillonnage	86
3.15.1.	L'utilisation de l'échantillonnage	86
3.15.2.	Les méthodes d'échantillonnage aléatoire	87
4.	L'utilisation des données commerciales	89
4.1.	Les données utilisées dans les applications commerciales	89
4.1.1.	Données sur les transactions et données RFM.....	89
4.1.2.	Données sur les produits et contrats	90
4.1.3.	Anciennetés	90
4.1.4.	Données sur les canaux	92
4.1.5.	Données relationnelles, attitudinales et psychographiques	92
4.1.6.	Données sociodémographiques	93
4.1.7.	Quand on manque de données	93
4.1.8.	Données techniques	94
4.2.	Des données particulières.....	94
4.2.1.	Le géomarketing	94
4.2.2.	La rentabilité	99
4.3.	Les données utilisées par secteur d'activité.....	100
4.3.1.	Les données utilisées dans la banque	100
4.3.2.	Les données utilisées dans l'assurance	101
4.3.3.	Les données utilisées dans la téléphonie.....	102
4.3.4.	Les données utilisées dans la VPC	103
5.	Aperçu sur les techniques de data mining	105
5.1.	Un point de terminologie	105
5.2.	Classification des techniques	106
5.3.	Comparatif des techniques.....	108
5.4.	Utilisations de ces techniques dans le domaine commercial	110
6.	L'analyse factorielle	111
6.1.	L'analyse en composantes principales.....	111
6.1.1.	Principe de l'ACP	111

6.1.2.	Représentation des variables	117
6.1.3.	Représentation des individus.....	121
6.1.4.	Utilisation de l'ACP	122
6.1.5.	Choix du nombre d'axes factoriels.....	124
6.1.6.	En bref.....	126
6.2.	Les variantes de l'analyse en composantes principales	126
6.2.1.	ACP avec rotation	126
6.2.2.	ACP des rangs.....	128
6.3.	L'analyse factorielle des correspondances	128
6.4.	L'analyse des correspondances multiples.....	131
6.4.1.	Principe de l'ACM	131
6.4.2.	Récapitulatif sur l'AFC et l'ACM.....	136
6.4.3.	Mise en œuvre de l'ACM avec le logiciel SAS	137
7.	Les réseaux de neurones	147
7.1.	Généralités sur les réseaux de neurones.....	147
7.2.	Structure d'un réseau de neurones	150
7.3.	Choix de l'échantillon d'apprentissage	152
7.4.	Quelques règles empiriques pour le dimensionnement d'un réseau	152
7.5.	Normalisation des données	153
7.5.1.	Variables continues.....	153
7.5.2.	Variables discrètes.....	154
7.5.3.	Variables qualitatives	154
7.6.	Réseaux de neurones et séries temporelles	155
7.7.	Les algorithmes d'apprentissage.....	156
7.8.	Les principaux réseaux de neurones	157
7.8.1.	Le perceptron multicouches.....	157
7.8.2.	Le réseau à fonction radiale de base.....	160
7.8.3.	Le réseau de Kohonen.....	163
8.	Les techniques de classification automatique	167
8.1.	Définition de la classification	167
8.2.	Applications de la classification.....	168
8.3.	Complexité de la classification	169
8.4.	Structures de classification.....	169
8.4.1.	Structure des données à classer	169
8.4.2.	Structure des classes obtenues.....	169
8.5.	Quelques points méthodologiques	170
8.5.1.	Le nombre optimum de classes.....	170
8.5.2.	L'utilisation de certains types de variables	171
8.5.3.	L'utilisation de variables illustratives.....	172
8.5.4.	L'évaluation de la qualité de la classification	173
8.5.5.	L'interprétation des classes obtenues	173
8.5.6.	Les critères de bonne classification.....	175
8.6.	Comparaison de l'analyse factorielle et de la classification	175

8.7.	Inerties intraclasse et interclasse	176
8.8.	Mesures de qualité d'une classification	177
8.8.1.	Tous types de classification	178
8.8.2.	Classifications hiérarchiques ascendantes	179
8.9.	Les méthodes de partitionnement	180
8.9.1.	La méthode des centres mobiles	180
8.9.2.	Les <i>k-means</i> et les nuées dynamiques	182
8.9.3.	Traitement des données qualitatives	182
8.9.4.	Les <i>k-medoids</i> et leurs variantes	182
8.9.5.	Avantages des méthodes de partitionnement	183
8.9.6.	Inconvénients des méthodes de partitionnement	184
8.9.7.	Sensibilité au choix des centres initiaux	185
8.10.	La classification ascendante hiérarchique	187
8.10.1.	Principe de la classification ascendante hiérarchique	187
8.10.2.	Les principales distances utilisées	188
8.10.3.	Les méthodes par estimation de densité	192
8.10.4.	Avantages de la classification ascendante hiérarchique	195
8.10.5.	Inconvénients de la classification ascendante hiérarchique	195
8.11.	Les méthodes mixtes de classification	196
8.11.1.	Principe	196
8.11.2.	Illustration avec le logiciel SAS	196
8.12.	La classification neuronale	207
8.12.1.	Avantages de la classification neuronale	207
8.12.2.	Inconvénients de la classification neuronale	207
8.13.	La classification relationnelle	208
8.13.1.	Principe de l'analyse relationnelle	208
8.13.2.	Mise en œuvre de la classification relationnelle	209
8.13.3.	Avantages de la classification relationnelle	210
8.13.4.	Inconvénients de la classification relationnelle	210
8.14.	La classification de variables numériques	211
8.15.	Vue d'ensemble des méthodes de classification	217
9.	La recherche d'associations	221
9.1.	Principes	221
9.2.	Utilisation de la taxinomie	223
9.3.	Utilisation de variables supplémentaires	223
9.4.	Applications	224
10.	Les techniques de classement et de prédiction	227
10.1.	Introduction	227
10.2.	Techniques inductives et transductives	228
10.3.	Vue d'ensemble des techniques de classement et de prédiction	230
10.3.1.	Les qualités attendues d'une technique de classement et prédiction	230
10.3.2.	Pouvoir de généralisation	231
10.3.3.	Sur-apprentissage	232
10.4.	Le classement par arbre de décision	236
10.4.1.	Principe de l'arbre de décision	236

10.4.2.	Définitions – première étape de la construction de l'arbre	236
10.4.3.	Critère de séparation	240
10.4.4.	Répartition des individus dans les nœuds – deuxième étape de la construction de l'arbre	242
10.4.5.	Élagage – troisième étape de la construction de l'arbre	243
10.4.6.	Piège à éviter	243
10.4.7.	Les arbres CART, C5.0 et CHAID	244
10.4.8.	Avantages des arbres de décision	250
10.4.9.	Inconvénients des arbres de décision	250
10.5.	La prédiction par arbre de décision	253
10.6.	Le classement par analyse discriminante	254
10.6.1.	Problématique	254
10.6.2.	L'analyse discriminante géométrique descriptive (analyse factorielle discriminante)	255
10.6.3.	L'analyse discriminante géométrique prédictive	260
10.6.4.	L'analyse discriminante probabiliste	263
10.6.5.	Mesures de la qualité du modèle	266
10.6.6.	Syntaxe de l'analyse discriminante dans le logiciel SAS	271
10.6.7.	Application au <i>credit scoring</i>	272
10.6.8.	L'analyse discriminante sur variables qualitatives (méthode DISQUAL)	273
10.6.9.	Avantages de l'analyse discriminante	274
10.6.10.	Inconvénients de l'analyse discriminante	274
10.7.	La prédiction par régression linéaire	276
10.7.1.	La régression linéaire simple et multiple	276
10.7.2.	Exemple de régression linéaire	285
10.7.3.	Compléments sur la syntaxe SAS de la régression linéaire	292
10.7.4.	Problèmes de colinéarité en régression linéaire multiple	295
10.7.5.	La régression PLS	297
10.7.6.	La régression robuste	298
10.7.7.	Le modèle linéaire général	303
10.8.	Le classement par régression logistique	304
10.8.1.	Principes de la régression logistique binaire	304
10.8.2.	Les régressions logistiques <i>logit</i> , <i>probit</i> et <i>log-log</i>	308
10.8.3.	Les odds-ratios	311
10.8.4.	Illustration du découpage en classes	312
10.8.5.	Estimation des paramètres	313
10.8.6.	Déviance et mesure de qualité d'un modèle	315
10.8.7.	Séparation complète en régression logistique	318
10.8.8.	Tests statistiques de la régression logistique	320
10.8.9.	Effet du découpage en modalités et du choix de la modalité de référence	324
10.8.10.	Effet de la colinéarité	324
10.8.11.	Influence de l'échantillonnage sur la régression <i>logit</i>	325
10.8.12.	Syntaxe de la régression logistique dans le logiciel SAS	326
10.8.13.	Exemple de modélisation par régression logistique	328
10.8.14.	Avantages de la régression logistique	340
10.8.15.	Avantages du modèle <i>logit</i> sur le modèle <i>probit</i>	340
10.8.16.	Inconvénients de la régression logistique	340

10.9.	Développements de la régression logistique.....	340
10.9.1.	La régression logistique sur individus de poids différents	340
10.9.2.	La régression logistique sur données corrélées	341
10.9.3.	La régression logistique ordinale	343
10.9.4.	La régression logistique multinomiale.....	344
10.9.5.	La régression logistique PLS	345
10.9.6.	Le modèle linéaire généralisé	346
10.9.7.	La régression de Poisson	349
10.9.8.	Le modèle additif généralisé	354
10.10.	Le classement et la prédiction par réseaux de neurones	354
10.10.1.	Avantages des réseaux de neurones	354
10.10.2.	Inconvénients des réseaux de neurones	355
10.11.	Le classement par « support vector machines » (SVM).....	357
10.11.1.	Introduction aux SVM.....	357
10.11.2.	Exemple	359
10.11.3.	Avantages des SVM	363
10.11.4.	Inconvénients des SVM	363
10.12.	La prédiction par algorithmes génétiques	363
10.12.1.	Génération aléatoire des règles initiales	364
10.12.2.	Sélection des meilleures règles	364
10.12.3.	Génération de nouvelles règles	364
10.12.4.	Fin de l'algorithme.....	365
10.12.5.	Applications des algorithmes génétiques	365
10.12.6.	Inconvénients des algorithmes génétiques	366
10.13.	Le classement par systèmes experts	366
10.14.	L'amélioration des performances d'un modèle prédictif	367
10.15.	Bootstrap et agrégation de modèles	369
10.15.1.	Le bootstrap	369
10.15.2.	L'agrégation par bagging	372
10.15.3.	L'agrégation par boosting	374
10.15.4.	Quelques applications.....	376
10.15.5.	Conclusion	376
10.16.	Mise en œuvre des techniques de classement et prédiction	378
10.16.1.	Le choix des techniques de modélisation	378
10.16.2.	La phase d'apprentissage d'un modèle	380
10.16.3.	L'inférence des refusés.....	383
10.16.4.	La phase de test d'un modèle.....	385
10.16.5.	Courbe ROC, courbe de lift et indice de Gini	387
10.16.6.	La table de classification d'un modèle	396
10.16.7.	La phase de validation d'un modèle	398
10.16.8.	La phase d'application d'un modèle	398
11.	Une application du data mining : le scoring	401
11.1.	Les différents types de scores.....	401
11.2.	L'utilisation des scores d'appétence et de risque	403
11.3.	La méthodologie	405
11.3.1.	Détermination des objectifs	405

11.3.2.	Inventaire et préparation des données	405
11.3.3.	Constitution de la base d'analyse	406
11.3.4.	Élaboration d'un modèle prédictif	407
11.3.5.	Utilisation du score.....	407
11.3.6.	Déploiement du score	408
11.3.7.	Suivi des outils mis à disposition	408
11.4.	Mise en œuvre d'un score stratégique	409
11.5.	Mise en œuvre d'un score opérationnel	410
11.6.	Les différentes solutions de scoring pour une entreprise	411
11.6.1.	Score en interne ou en <i>credit bureau</i>	411
11.6.2.	Score générique ou personnalisé.....	413
11.6.3.	Résumé des différentes solutions possibles	414
12.	Les facteurs de succès d'un projet de data mining	415
12.1.	Le sujet.....	415
12.2.	Les hommes	416
12.3.	Les données	417
12.4.	L'informatique	417
12.5.	La culture d'entreprise	418
12.6.	Huit idées fausses sur le data mining.....	419
12.6.1.	Aucun <i>a priori</i> n'est nécessaire.....	419
12.6.2.	On n'a plus besoin de spécialistes du métier	420
12.6.3.	On n'a plus besoin de statisticiens (« Il suffit d'appuyer sur un bouton »)	420
12.6.4.	Le data mining permet de faire des découvertes incroyables	421
12.6.5.	Le data mining est révolutionnaire	421
12.6.6.	Il faut utiliser toutes les données disponibles	422
12.6.7.	Il faut toujours échantillonner	422
12.6.8.	Il ne faut jamais échantillonner	422
12.7.	Le retour sur investissement	422
13.	Les logiciels de statistique et data mining	427
13.1.	Typologie des logiciels de data mining et statistique	427
13.2.	Les caractéristiques importantes des logiciels	430
13.2.1.	Points de comparaison	430
13.2.2.	Méthodes implémentées.....	431
13.2.3.	Fonctions de préparation des données	432
13.2.4.	Autres fonctions	432
13.2.5.	Caractéristiques techniques	433
13.3.	Les principaux logiciels	433
13.3.1.	Vue d'ensemble	433
13.3.2.	SPSS.....	436
13.3.3.	SAS	437
13.3.4.	R.....	439
13.4.	Nous avons comparé pour vous : SAS et SPSS	447
13.5.	Pour diminuer les temps de traitement.....	461

14.	Le text mining	465
14.1.	Définition du text mining	465
14.2.	Les sources de textes utilisées	466
14.3.	Utilisation du text mining	466
14.4.	Recherche d'information	467
14.4.1.	Analyse linguistique	468
14.4.2.	Application de la statistique et du data mining	470
14.4.3.	Techniques applicables	471
14.5.	Extraction d'information	473
14.5.1.	Principe de l'extraction d'information	473
14.5.2.	Exemple d'application : transcription d'entretiens commerciaux.....	473
14.6.	Data mining multitype	474
15.	Le web mining	475
15.1.	Les objectifs du web mining	475
15.2.	Analyses globales	476
15.2.1.	À quoi servent-elles ?	476
15.2.2.	La structure du fichier « log »	476
15.2.3.	L'utilisation du fichier « log »	477
15.3.	Analyses individuelles	480
15.4.	Analyses nominatives.....	481
Annexe A :	Rappels de statistique	483
16.1.	Aperçu historique	483
16.1.1.	Quelques dates.....	483
16.1.2.	De la statistique... au data mining	484
16.2.	Rappels de statistique	485
16.2.1.	Caractéristiques statistiques.....	485
16.2.2.	Boîte à moustaches	487
16.2.3.	Les tests d'hypothèses	488
16.2.4.	Tests asymptotiques, exacts, paramétriques et non-paramétriques.....	489
16.2.5.	Intervalle de confiance d'une moyenne : le test de Student.....	490
16.2.6.	Intervalle de confiance d'une fréquence (ou proportion)	492
16.2.7.	Liaison entre deux variables continues : coefficient de corrélation linéaire	493
16.2.8.	Liaison entre deux variables numériques ou ordinales : coefficient de corrélation des rangs de Spearman et tau de Kendall	495
16.2.9.	Liaison entre n ensembles de plusieurs variables continues ou binaires : l'analyse de corrélation canonique	496
16.2.10.	Liaison entre deux variables nominales : le test du χ^2	497
16.2.11.	Exemple d'utilisation du test du χ^2	498
16.2.12.	Liaison entre deux variables nominales : le coefficient de Cramer	499
16.2.13.	Liaison entre une variable nominale et une variable numérique : le test de la variance (ANOVA à 1 facteur).....	500
16.2.14.	Modèle de survie semi-paramétrique de Cox.....	502
16.3.	Tables statistiques.....	504
16.3.1.	Table de la loi normale centrée réduite	504

16.3.2. Table de la loi de Student	506
16.3.3. Table du χ^2	506
16.3.4. Table de la loi de Fisher-Snedecor au seuil de probabilité 0,05	506
Annexe B : Data mining, informatique et libertés	511
17.1. Les textes	511
17.2. Les traitements soumis à autorisation préalable	512
17.3. Les traitements soumis à déclaration	514
17.4. Les pouvoirs de la CNIL	516
17.5. Les droits des personnes physiques	517
17.6. Les spécificités des traitements de data mining	517
17.6.1. Spécificités du scoring de risque	517
17.6.2. Spécificités de la segmentation de clientèle	519
17.6.3. Ce qu'il faut déclarer à la CNIL	519
17.6.4. Conclusion	520
Bibliographie	521
18.1. Sur la statistique et l'analyse des données	521
18.2. Sur le data mining	523
18.3. Sur le text mining	524
18.4. Sur le web mining	524
18.5. Sur le logiciel SAS	524
18.6. Quelques articles	525
18.7. Sites Internet	525
Index	529