Table des matières

Préface ..... **V**

Avant-propos ..... **VII**

Table des matières ..... **XI**

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 5
1.2.2. Le data mining par type d'application 9
1.3. Data mining et statistique 12
1.4. Data mining et informatique 14
1.5. Data mining et protection des données individuelles 20
1.6. Mise en œuvre du data mining 24

Le déroulement d'une étude de data mining **27**

2.1. Définition des objectifs 27
2.2. Inventaire des données existantes 28
2.3. Collecte des données 29
2.4. Exploration et préparation des données 32
2.5. Segmentation de la population 34
2.6. Élaboration et validation des modèles prédictifs 36
2.7. Synthèse des modèles prédictifs des différents segments 38
2.8. Itération des étapes précédentes 38
2.9. Déploiement des modèles 39
2.10. Formation des utilisateurs des modèles 40
2.11. Suivi des modèles 40
2.12. Enrichissement des modèles 42
2.13. Remarques 43
2.14. Cycle de vie d'un modèle 43
2.15. Charges pour un projet pilote 44

L'exploration et la préparation des données 47

3.1. Les différents types de données 47
3.2. L'examen de la distribution des variables 48
3.3. La détection des valeurs rares ou manquantes 49
3.4. La détection des valeurs aberrantes 54
3.5. La détection des valeurs extrêmes 56
3.6. Les tests de normalité 58
3.7. Homoscédasticité et hétéroscédasticité 62
3.8. La détection des variables les plus discriminantes 63
3.8.1. Variables explicatives qualitatives, discrètes ou découpées en classes 64
3.8.2. Variables explicatives continues 67
3.8.3. Précisions sur les tests non-paramétriques à un facteur 72
3.8.4. ODS et automatisation de la sélection des variables discriminantes 76
3.9. La transformation des variables 79
3.10. Le choix des tranches de valeurs des variables continues 83
3.11. La création de nouvelles variables 87
3.12. La détection des interactions 89
3.13. La sélection automatique des variables 91
3.14. La détection de la colinéarité 93
3.15. L'échantillonnage 96
3.15.1. L'utilisation de l'échantillonnage 96
3.15.2. Les méthodes d'échantillonnage aléatoire 97

L'utilisation des données commerciales 99

4.1. Les données utilisées dans les applications commerciales 99
4.1.1. Données sur les transactions et données RFM 99
4.1.2. Données sur les produits et contrats 100
4.1.3. Anciennetés 100
4.1.4. Données sur les canaux 102
4.1.5. Données relationnelles, attitudinales et psychographiques 102
4.1.6. Données sociodémographiques 103
4.1.7. Quand on manque de données 103
4.1.8. Données techniques 104
4.2. Des données particulières 104
4.2.1. Le géomarketing 104
4.2.2. La rentabilité 109
4.3. Les données utilisées par secteur d'activité 110
4.3.1. Les données utilisées dans la banque 110
4.3.2. Les données utilisées dans l'assurance 111
4.3.3. Les données utilisées dans la téléphonie 112
4.3.4. Les données utilisées dans la VPC 113

Les logiciels de statistique et data mining 115

5.1. Typologie des logiciels de data mining et statistique 115
5.2. Les caractéristiques importantes des logiciels 117
5.2.1. Points de comparaison 117

5.2.2. Méthodes implémentées 119
5.2.3. Fonctions de préparation des données 120
5.2.4. Autres fonctions 120
5.2.5. Caractéristiques techniques 121
5.3. Les principaux logiciels 121
5.3.1. Vue d'ensemble 121
5.3.2. IBM SPSS 124
5.3.3. SAS 127
5.3.4. R 129
5.3.5. Un peu de langage R 138
5.4. Comparaison des logiciels SAS et IBM SPSS 140
5.5. Pour diminuer les temps de traitement 155

Aperçu sur les techniques de data mining 159

6.1. Un point de terminologie 159
6.2. Classification des techniques 160
6.3. Comparatif des techniques 160
6.4. Utilisations de ces techniques dans le domaine commercial 164

L'analyse factorielle 165

7.1. L'analyse en composantes principales 165
7.1.1. Principe de l'ACP 165
7.1.2. Représentation des variables 171
7.1.3. Représentation des individus 176
7.1.4. Utilisation de l'ACP 177
7.1.5. Choix du nombre d'axes factoriels 180
7.1.6. En bref 182
7.2. Les variantes de l'analyse en composantes principales 182
7.2.1. ACP avec rotation 182
7.2.2. ACP des rangs 184
7.2.3. ACP sur variables qualitatives 184
7.3. L'analyse factorielle des correspondances 185
7.3.1. Principe de l'AFC 185
7.3.2. Mise en œuvre de l'AFC avec le logiciel IBM SPSS Statistics 188
7.4. L'analyse des correspondances multiples 193
7.4.1. Principe de l'ACM 193
7.4.2. Récapitulatif sur l'AFC et l'ACM 198
7.4.3. Mise en œuvre de l'ACM et de l'AFC avec le logiciel SAS 199

Les réseaux de neurones 209

8.1. Généralités sur les réseaux de neurones 209
8.2. Structure d'un réseau de neurones 212
8.3. Choix de l'échantillon d'apprentissage 214
8.4. Quelques règles empiriques pour le dimensionnement d'un réseau 214
8.5. Normalisation des données 215

8.5.1. Variables continues....215
8.5.2. Variables discrètes....216
8.5.3. Variables qualitatives....216
8.6. Les algorithmes d'apprentissage....217
8.7. Les principaux réseaux de neurones....217
8.7.1. Le perceptron multicouches....217
8.7.2. Le réseau à fonction radiale de base....220
8.7.3. Le réseau de Kohonen....223

Les techniques de classification automatique....229

9.1. Définition de la classification....229
9.2. Applications de la classification....230
9.3. Complexité de la classification....231
9.4. Structures de classification....231
9.4.1. Structure des données à classer....231
9.4.2. Structure des classes obtenues....231
9.5. Quelques points méthodologiques....232
9.5.1. Le nombre optimum de classes....232
9.5.2. L'utilisation de certains types de variables....233
9.5.3. L'utilisation de variables illustratives....234
9.5.4. L'évaluation de la qualité de la classification....235
9.5.5. L'interprétation des classes obtenues....235
9.5.6. Les critères de bonne classification....237
9.6. Comparaison de l'analyse factorielle et de la classification....237
9.7. Inerties intraclasse et interclasse....238
9.8. Mesures de qualité d'une classification....239
9.8.1. Tous types de classification....240
9.8.2. Classifications hiérarchiques ascendantes....241
9.9. Les méthodes de partitionnement....242
9.9.1. La méthode des centres mobiles....242
9.9.2. Les *k-means* et les nuées dynamiques....243
9.9.3. Traitement des données qualitatives....244
9.9.4. Les *k-medoids* et leurs variantes....244
9.9.5. Avantages des méthodes de partitionnement....245
9.9.6. Inconvénients des méthodes de partitionnement....246
9.9.7. Sensibilité au choix des centres initiaux....247
9.10. La classification ascendante hiérarchique....249
9.10.1. Principe de la classification ascendante hiérarchique....249
9.10.2. Les principales distances utilisées....250
9.10.3. Les méthodes par estimation de densité....254
9.10.4. Avantages de la classification ascendante hiérarchique....257
9.10.5. Inconvénients de la classification ascendante hiérarchique....257
9.11. Les méthodes mixtes de classification....258
9.11.1. Principe....258
9.11.2. Illustration avec le logiciel SAS....258
9.12. La classification neuronale....269
9.12.1. Avantages de la classification neuronale....269

9.12.2. Inconvénients de la classification neuronale 269
9.13. La classification par agrégation des similarités 270
9.13.1. Principe de l'analyse relationnelle 270
9.13.2. Mise en œuvre de la classification par agrégation des similarités 271
9.13.3. Exemple d'utilisation du package *amap* de R 272
9.13.4. Avantages de la classification par agrégation des similarités 273
9.13.5. Inconvénients de la classification par agrégation des similarités 273
9.14. La classification de variables numériques 274
9.15. Vue d'ensemble des méthodes de classification 280

La recherche d'associations 283

10.1. Principes 283
10.2. Utilisation de la taxinomie 287
10.3. Utilisation de variables supplémentaires 288
10.4. Applications 288
10.5. Exemple d'utilisation 289

Les techniques de classement et de prédiction 297

11.1. Introduction 297
11.2. Techniques inductives et transductives 298
11.3. Vue d'ensemble des techniques de classement et de prédiction 300
11.3.1. Les qualités attendues d'une technique de classement et prédiction 300
11.3.2. Pouvoir de généralisation 301
11.3.3. Théorie de l'apprentissage de Vapnik 303
11.3.4. Sur-apprentissage 305
11.4. Le classement par arbre de décision 309
11.4.1. Principe de l'arbre de décision 309
11.4.2. Définitions – première étape de la construction de l'arbre 309
11.4.3. Critère de séparation 313
11.4.4. Répartition dans les nœuds – deuxième étape de la construction de l'arbre 314
11.4.5. Élagage – troisième étape de la construction de l'arbre 315
11.4.6. Piège à éviter 316
11.4.7. Les arbres CART, C5.0 et CHAID 317
11.4.8. Avantages des arbres de décision 323
11.4.9. Inconvénients des arbres de décision 324
11.5. La prédiction par arbre de décision 326
11.6. Le classement par analyse discriminante 328
11.6.1. Problématique 328
11.6.2. L'analyse discriminante géométrique descriptive (analyse factorielle discriminante) 329
11.6.3. L'analyse discriminante géométrique prédictive 335
11.6.4. L'analyse discriminante probabiliste 338
11.6.5. Mesures de la qualité du modèle 341
11.6.6. Syntaxe de l'analyse discriminante dans le logiciel SAS 346
11.6.7. L'analyse discriminante sur variables qualitatives (méthode DISQUAL) 349
11.6.8. Avantages de l'analyse discriminante 350

11.6.9. Inconvénients de l'analyse discriminante 350
11.7. La prédiction par régression linéaire 351
11.7.1. La régression linéaire simple 352
11.7.2. Régression linéaire multiple et régression régularisée 355
11.7.3. Les tests en régression linéaire 360
11.7.4. Les tests sur les résidus 367
11.7.5. Influence d'observations 371
11.7.6. Exemple de régression linéaire 374
11.7.7. Compléments sur la syntaxe SAS de la régression linéaire 381
11.7.8. Problèmes de colinéarité en régression linéaire : un exemple avec R 384
11.7.9. Problèmes de colinéarité en régression linéaire : diagnostic et solutions 392
11.7.10. La régression PLS 396
11.7.11. Traitement de la régression régularisée avec SAS et R 398
11.7.12. La régression robuste 430
11.7.13. Le modèle linéaire général 435
11.8. Le classement par régression logistique 439
11.8.1. Principes de la régression logistique binaire 439
11.8.2. Les régressions logistiques *logit*, *probit* et *log-log* 443
11.8.3. Les odds-ratios 446
11.8.4. Illustration du découpage en classes 447
11.8.5. Estimation des paramètres 448
11.8.6. Déviance et mesure de qualité d'un modèle 452
11.8.7. Séparation complète en régression logistique 455
11.8.8. Tests statistiques de la régression logistique 457
11.8.9. Effet du découpage en modalités et du choix de la modalité de référence 461
11.8.10. Effet de la colinéarité 462
11.8.11. Influence de l'échantillonnage sur la régression *logit* 463
11.8.12. Syntaxe de la régression logistique dans le logiciel SAS 464
11.8.13. Exemple de modélisation par régression logistique 466
11.8.14. Régression logistique avec R 479
11.8.15. Avantages de la régression logistique 481
11.8.16. Avantages du modèle *logit* sur le modèle *probit* 482
11.8.17. Inconvénients de la régression logistique 482
11.9. Développements de la régression logistique 482
11.9.1. La régression logistique sur individus de poids différents 482
11.9.2. La régression logistique sur données corrélées 483
11.9.3. La régression logistique ordinale 485
11.9.4. La régression logistique multinomiale 486
11.9.5. La régression logistique PLS 487
11.9.6. Le modèle linéaire généralisé 488
11.9.7. La régression de Poisson 491
11.9.8. Le modèle additif généralisé 496
11.10. Méthodes bayésiennes 496
11.10.1. Le classifieur bayésien naïf 497
11.10.2. Réseaux bayésiens 501
11.11. Le classement et la prédiction par réseaux de neurones 502
11.11.1. Avantages des réseaux de neurones 503

11.11.2. Inconvénients des réseaux de neurones 504
11.12. Le classement par « support vector machines » (SVM) 506
11.12.1. Introduction aux SVM 506
11.12.2. Exemple 511
11.12.3. Avantages des SVM 515
11.12.4. Inconvénients des SVM 515
11.13. La prédiction par algorithmes génétiques 515
11.13.1. Génération aléatoire des règles initiales 516
11.13.2. Sélection des meilleures règles 516
11.13.3. Génération de nouvelles règles 516
11.13.4. Fin de l'algorithme 517
11.13.5. Applications des algorithmes génétiques 517
11.13.6. Inconvénients des algorithmes génétiques 518
11.14. L'amélioration des performances d'un modèle prédictif 518
11.15. Bootstrap et agrégation de modèles 520
11.15.1. Le bootstrap 520
11.15.2. L'agrégation par bagging 523
11.15.3. L'agrégation par boosting 525
11.15.4. Quelques applications 527
11.15.5. Conclusion 528
11.16. Mise en œuvre des techniques de classement et prédiction 530
11.16.1. Le choix des techniques de modélisation 530
11.16.2. La phase d'apprentissage d'un modèle 531
11.16.3. L'inférence des refusés 534
11.16.4. La phase de test d'un modèle 536
11.16.5. Courbe ROC, courbe de lift et indice de Gini 538
11.16.6. La table de classification d'un modèle 547
11.16.7. La phase de validation d'un modèle 549
11.16.8. La phase d'application d'un modèle 549

Une application du data mining : le scoring 551

12.1. Les différents types de scores 551
12.2. L'utilisation des scores d'appétence et de risque 553
12.3. La méthodologie 555
12.3.1. Détermination des objectifs 555
12.3.2. Inventaire et préparation des données 555
12.3.3. Constitution de la base d'analyse 556
12.3.4. Élaboration d'un modèle prédictif 557
12.3.5. Utilisation du score 557
12.3.6. Déploiement du score 558
12.3.7. Suivi des outils mis à disposition 558
12.4. Mise en œuvre d'un score stratégique 559
12.5. Mise en œuvre d'un score opérationnel 560
12.6. Les différentes solutions de scoring pour une entreprise 561
12.6.1. Score en interne ou en *credit bureau* 561

12.6.2. Score générique ou personnalisé 563
12.6.3. Résumé des différentes solutions possibles 564
12.7. Un exemple de *credit scoring* (préparation des données) 565
12.8. Un exemple de *credit scoring* (modélisation par régression logistique) 592
12.9. Un exemple de *credit scoring* (modélisation par analyse discriminante DISQUAL) 604
12.10. Une brève histoire du *credit scoring* 615

Les facteurs de succès d'un projet de data mining 617

13.1. Le sujet 617
13.2. Les hommes 618
13.3. Les données 619
13.4. L'informatique 619
13.5. La culture d'entreprise 620
13.6. Huit idées fausses sur le data mining 621
13.6.1. Aucun *a priori* n'est nécessaire 621
13.6.2. On n'a plus besoin de spécialistes du métier 622
13.6.3. On n'a plus besoin de statisticiens (« Il suffit d'appuyer sur un bouton ») 622
13.6.4. Le data mining permet de faire des découvertes incroyables 623
13.6.5. Le data mining est révolutionnaire 623
13.6.6. Il faut utiliser toutes les données disponibles 624
13.6.7. Il faut toujours échantillonner 624
13.6.8. Il ne faut jamais échantillonner 624
13.7. Le retour sur investissement 625

Le text mining 629

14.1. Définition du text mining 629
14.2. Les sources de textes utilisées 630
14.3. Utilisation du text mining 630
14.4. Recherche d'information 631
14.4.1. Analyse linguistique 632
14.4.2. Application de la statistique et du data mining 634
14.4.3. Techniques applicables 635
14.5. Extraction d'information 637
14.5.1. Principe de l'extraction d'information 637
14.5.2. Exemple d'application : transcription d'entretiens commerciaux 637
14.6. Data mining multitype 638

Le web mining 639

15.1. Les objectifs du web mining 639
15.2. Analyses globales 640
15.2.1. À quoi servent-elles ? 640
15.2.2. La structure du fichier « log » 640
15.2.3. L'utilisation du fichier « log » 641
15.3. Analyses individuelles 644
15.4. Analyses nominatives 645

Annexe A : Rappels de statistique 647

16.1. Aperçu historique 647
16.1.1. Quelques dates 647
16.1.2. De la statistique… au data mining 649
16.2. Rappels de statistique 650
16.2.1. Caractéristiques statistiques 650
16.2.2. Boîte à moustaches 651
16.2.3. Les tests d'hypothèses 652
16.2.4. Tests asymptotiques, exacts, paramétriques et non-paramétriques 654
16.2.5. Intervalle de confiance d'une moyenne : le test de Student 654
16.2.6. Intervalle de confiance d'une fréquence (ou proportion) 656
16.2.7. Liaison entre deux variables continues : coefficient de corrélation linéaire 658
16.2.8. Liaison entre deux variables numériques ou ordinales : coefficient de corrélation des rangs de Spearman et tau de Kendall 660
16.2.9. Liaison entre *n* ensembles de plusieurs variables continues ou binaires : l'analyse de corrélation canonique 662
16.2.10. Liaison entre deux variables nominales : le test du χ^2 662
16.2.11. Exemple d'utilisation du test du χ^2 663
16.2.12. Liaison entre deux variables nominales : le coefficient de Cramer 664
16.2.13. Liaison entre une variable nominale et une variable numérique : le test de la variance (ANOVA à 1 facteur) 665
16.2.14. Modèle de survie semi-paramétrique de Cox 667
16.3. Tables statistiques 669
16.3.1. Table de la loi normale centrée réduite 669
16.3.2. Table de la loi de Student 671
16.3.3. Table du χ^2 671
16.3.4. Table de la loi de Fisher-Snedecor au seuil de probabilité 0,05 671
16.3.5. Table de la loi de Fisher-Snedecor au seuil de probabilité 0,10 671

Annexe B : Data mining, informatique et libertés 679

17.1. Les textes 679
17.2. Les traitements soumis à autorisation préalable 680
17.3. Les traitements soumis à déclaration 681
17.4. Les pouvoirs de la CNIL 683
17.5. Les droits des personnes physiques 683
17.6. Les spécificités des traitements de data mining 684
17.6.1. Spécificités du scoring de risque 684
17.6.2. Spécificités de la segmentation de clientèle 685
17.6.3. Ce qu'il faut déclarer à la CNIL 686
17.6.4. Conclusion 687

Bibliographie 689

18.1. Sur la statistique et l'analyse des données 689
18.2. Sur le data mining et l'apprentissage statistique 693
18.3. Sur le text mining 695
18.4. Sur le web mining 695

18.5. Sur le logiciel R 695
18.6. Sur le logiciel SAS 696
18.7. Sur le logiciel IBM SPSS 697
18.8. Sites Internet 697

Index 701