Statistique et analyse de données pour l'assureur : des outils pour la gestion des risques et le marketing

Gilbert Saporta
Chaire de Statistique Appliquée, CNAM

- L’approche statistique permet de combiner connaissance des risques et connaissance du client
- Problèmes multidimensionnels
- Modèles, fouille de données

Statistique et tarification

- Recherche de variables tarifaires: un problème multivarié
  - Insuffisance des études univariées
    - exemple RC auto: tarif de base 100
    - modèle mutiplicatif $100(1+a_i)(1+b_i)...$
      a_i, b_j ....coefficients de réduction majoration négatifs ou positifs en %
    - réduction de 20% pour les agriculteurs, de 15% pour les zones rurales soit 32% de réduction!
  - Tenir compte des corrélations et des interactions

- Corrélation
  - Colinéarite entre prédicteurs numériques ou dépendance entre prédicteurs qualitatifs (χ^2)
    - mieux vaut éviter de garder des variables corrélées
    - possibilités de « proxy »
- Interaction:
  - Action différenciée sur la réponse (sinistralité) d'une variable selon les valeurs d'une autre :
    - ex. l'effet de « permis de conduire depuis moins de 2 ans » s'atténue avec l'age au permis
    - améliore les modèles

- Modèles classiques
  - Risque binaire: scores obtenus par régression logistique ou analyse discriminante
  - Fréquence: régression de Poisson
  - Coût: modèles linéaires généralisés

- Non prise en compte d'une variable significative (eg le genre)
  - Situation fréquente mais variable selon les réglementations
  - Effets statistiques
    - estimateurs biaisés
    - perte d'efficacité: plus d'incertitude
    - tarif moyenné

- Tarifs commerciaux vs tarifs techniques
  - on peut toujours utiliser le genre pour modéliser la prime pure
  - calcul des risques ≠ tarification
    - utile pour la réassurance
    - analyse *a posteriori* du S/P par segments

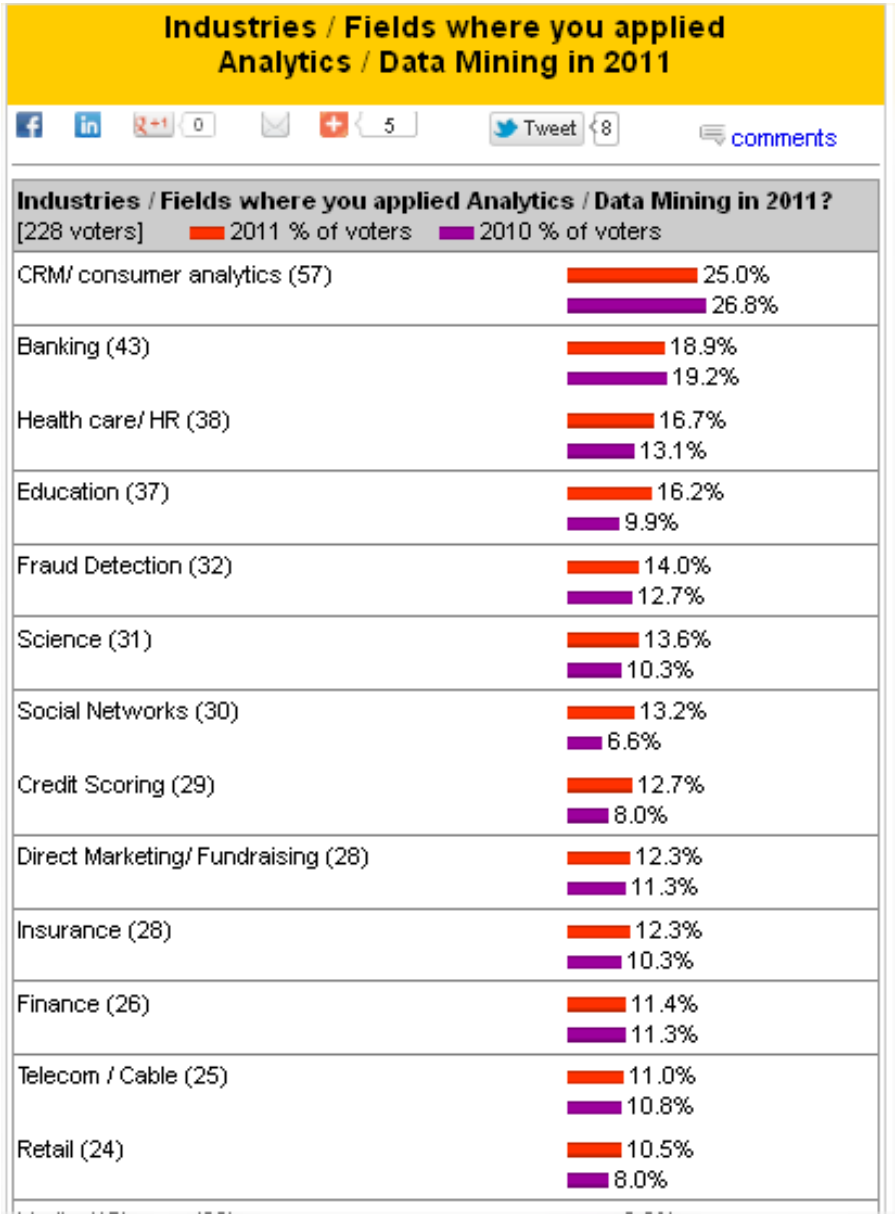
- Limites de la tarification *a priori* par segmentation
  - la plupart des modèles précédents reposent sur des historiques de sinistres d'assurés
  - absence fréquente de causalité (statut marital)
  - non-utilisation de variables pertinentes (non-respect du code de la route, alcoolisme, etc.) car absentes des dossiers

- Tarification *a posteriori*
  - corrige les défauts de la segmentation a priori
  - prise en compte du comportement réel
    - si les femmes sont moins risquées, leur prime diminuera dans le temps
  - Difficile en assurance de personnes

Data Mining

- Data Mining ou fouille de données
  - Exploration de modèles sans *a priori*
  - Nouvelle discipline apparue dans les années 90 à la frontière des bases de données, de l'IA, de la statistique
  - La métaphore du Data Mining signifie qu'il y a des trésors ou pépites cachés sous des montagnes de données

    « L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature. » (J.P.Benzécri 1973)



Industries / Fields where you applied Analytics / Data Mining in 2011

Industries / Fields where you applied Analytics / Data Mining in 2011?
[228 voters] 2011 % of voters 2010 % of voters

	2011 % of voters	2010 % of voters
CRM/ consumer analytics (57)	25.0%	26.8%
Banking (43)	18.9%	19.2%
Health care/ HR (38)	16.7%	13.1%
Education (37)	16.2%	9.9%
Fraud Detection (32)	14.0%	12.7%
Science (31)	13.6%	10.3%
Social Networks (30)	13.2%	6.6%
Credit Scoring (29)	12.7%	8.0%
Direct Marketing/ Fundraising (28)	12.3%	11.3%
Insurance (28)	12.3%	10.3%
Finance (26)	11.4%	11.3%
Telecom / Cable (25)	11.0%	10.8%
Retail (24)	10.5%	8.0%

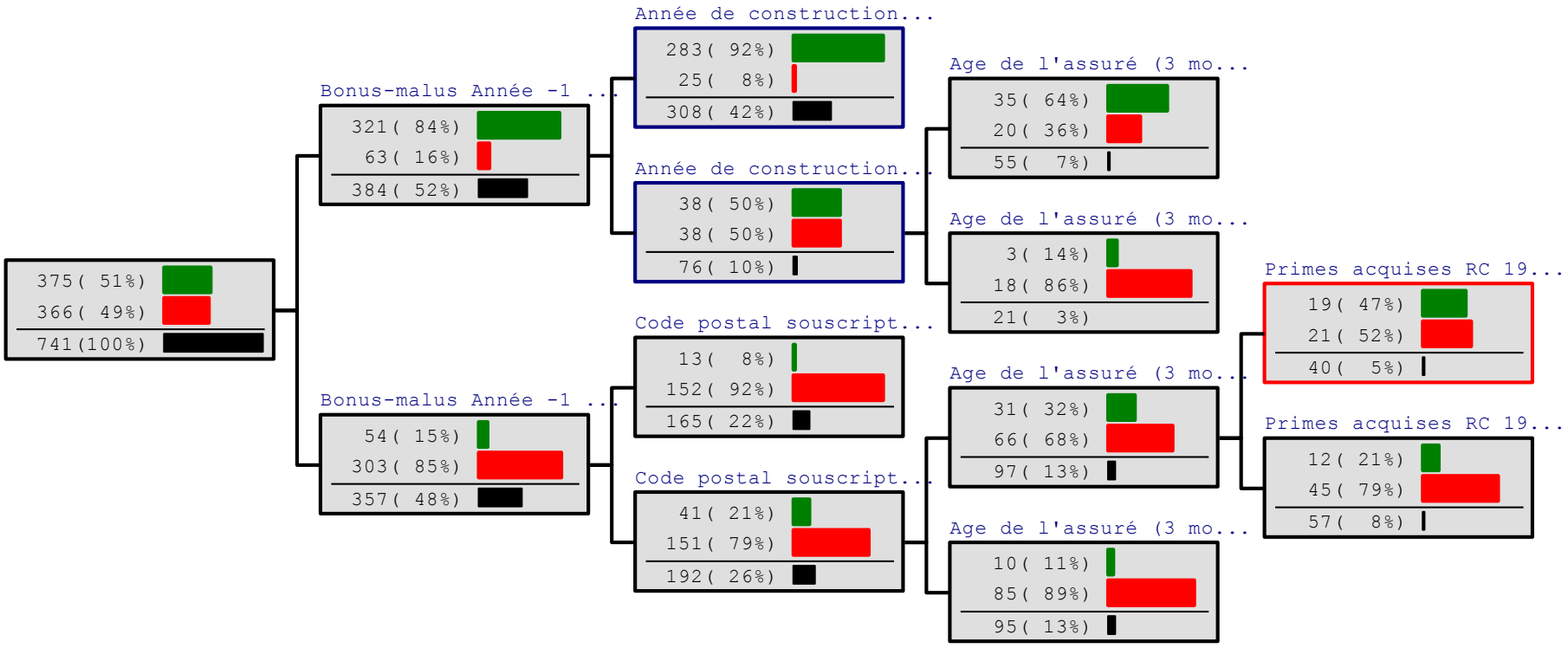
http://www.kdnuggets.com

n°1 marketing

n°10 assurance

- Comment choisir les bonnes variables?
  - Classiquement
    - à l'aide de tests statistiques
  - Inconvénients:
    - liste souvent restreinte à des indicateurs classiques
    - tests peu informatifs sur de grands jeux de données: tout est significatif !
  - Nécessité de nouvelles méthodes de validation
    - échantillons test; bootstrap

Arbre de décision



- Arbres de décision
  - grande lisibilité
  - souplesse d'utilisation: variables de types quelconques (numériques, catégorielles)
  - fournit des découpages en classe optimaux
  - détecte les intéractions (origine de la méthode AID)

- Bien d'autres méthodes algorithmiques de sélection

Algorithms for data analysis / data mining

comments

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

- Choix de variables, choix de modèles et théorie statistique de l'apprentissage
  - compromis entre complexité et capacité de prédiction
    - un modèle trop complexe s'ajuste bien mais prédit mal
  - mais plus les données disponibles sont nombreuses, plus on peut augmenter la complexité des modèles

- Une nouvelle façon de concevoir les modèles
  - Modèles pour comprendre ou modèles pour prévoir?
    - Compréhension des données et de leur mécanisme générateur à travers une représentation parcimonieuse
    - On peut prévoir sans comprendre: nul besoin d'une théorie du consommateur pour faire du ciblage
  - un modèle n'est qu'un algorithme
  - Combinaison de modèles: averaging, stacking

- Analyse des données et marketing
  - typologies et profils clientèle,ventes croisées, fidélisation vs acquisition
  - Outils
    - classification, analyse factorielles, scoring, règles d'association
    - statistique spatiale: données géolocalisées
    - Fusion de données de sources et de niveaux d'agrégation différents: fichiers clients, enquêtes, données administratives
  - élaboration de produits adaptés
    - prendre garde à l'antisélection

- Vers des offres segmentant hommes et femmes
  - aversion au risque plus grande chez les femmes
  - élaboration de contrats qui seront plus souvent choisis par les femmes et d'autres par les hommes
  - possibilité implicite de tarification différenciée

- **Conclusions**
  - la directive européenne prive les assureurs d'une variable de tarification mais pas d'une variable d'analyse technique
  - les méthodes de data mining peuvent mener à de nouveaux critères
  - mieux exploiter la convergence entre données marketing et sinistralité.
  - vers les « Big Data » en assurance