

CONSTRUCTION DE MODÈLES DE SCORING



BNP PARIBAS
PERSONAL FINANCE



Aurélien Marie, Data Scientist
aurelien.marie@bnpparibas-pf.com
Université de Bordeaux
MASTER IREF 2024-2025

ORGANISATION DU COURS

Quatre séances:

1^{ère} séance : Introduction au Scoring et à l'analyse de données

2^{ème} séance : Elaboration d'une grille de score à partir de la régression Logistique

3^{ème} séance : Initiation aux algorithmes de Machine Learning

4^{ème} séance: Accompagnement à la réalisation du projet

Type d'évaluation:

- Projet : Prédiction du risque de crédit à l'octroi
- Groupe de 3 à 4 étudiants
- Langage de programmation : Python
- Utilisation de notebooks (Jupyter)
- Librairies recommandées : pandas, numpy, scikit-learn, statsmodels
- Rapport PDF à rendre (maximum 20 pages)

SOMMAIRE

- 01** | Introduction au scoring
- 02** | Initialisation d'un projet de scoring
- 03** | Analyse des variables explicatives
- 04** | Modélisation
- 05** | Evaluation du modèle



1 - Introduction au scoring

Qu'est-ce qu'un score ?

- Littéralement : une note, une échelle d'évaluation.
- Plus précisément: une note reflétant la probabilité d'apparition d'un évènement à venir en fonction de diverses informations disponibles.
- En pratique :
 - une grille donnant la formule de calcul en fonction de diverses caractéristiques.
 - une note, calculée à partir de la grille, qui représente la probabilité d'occurrence de l'évènement à prédire.
- Un score permet de classer des individus et de créer des groupes ordonnés.
- Un score peut être le fruit d'une modélisation statistique ou de règles à dire d'expert.

1 - Introduction au scoring

Exemple de grille de score :

Caractéristique	Nombre de points attribués
Revenu	
< 1500€	10
1500€ - 3000€	50
> 3000€	60
Age	
< 25 ans	10
25 - 45 ans	30
46 - 60 ans	45
> 60 ans	55
Situation familiale	
Marié	40
Autre	10
Ancienneté bancaire	
< 1 an	10
1 - 5 ans	25
> 5 ans	45
Comportement de remboursement	
Connu, avec incidents de paiement	10
Inconnu	20
Connu, sans incident	50

Caractéristiques du dossier à noter :

- Revenu = 1600€
- Age = 50 ans
- Situation familiale = Marié
- Ancienneté bancaire = 2 ans
- Comportement de remboursement = *Connu de la banque et n'ayant eu aucun incident de paiement*



Note de score :

- Revenu → **50**
 - Age → **45**
 - Situation familiale → **40**
 - Ancienneté bancaire → **25**
 - Comportement de remboursement → **50**
- = **210**

1 - Introduction au scoring

Exemples de scores usuels :

- Risque de crédit :

- Score d'octroi : traduit la probabilité que le demandeur de crédit fasse défaut.
- Score fraude : traduit la probabilité que le demandeur de crédit soit fraudeur.

- Marketing :

- Score traduisant la probabilité de répondre à une sollicitation marketing.

- Assurance :

- Score reflétant la probabilité d'occurrence d'un sinistre.

- Média :

- Score d'influence.

1 - Introduction au scoring

Dans ce cours nous allons nous placer dans le cadre du risque de crédit.

Pourquoi les banques développent-elles des scores d'octroi ?

Exemple :

- Un organisme de crédit propose le financement d'une automobile.
- Un bon payeur (BP) rapporte en moyenne 400 euros.
- Un mauvais payeur (MP) coûte en moyenne 8 000 euros.

Le taux de mauvais payeurs observé par le passé est de 15 %.

- ✓ L'offre est-elle rentable pour l'organisme de credit ? Comment l'optimiser ?

1 - Introduction au scoring

Exemple :

- Quel taux de mauvais payeur cette banque peut-elle accepter en restant rentable ?

$$(1 - \text{taux}_{MP}) \times 400 - \text{taux}_{MP} \times 8000 \geq 0$$

$$\Leftrightarrow \text{taux}_{MP} \leq 1/_{21} \approx 4,76\%$$

- Comment s'assurer que le taux de mauvais payeurs soit inférieur à ce seuil de rentabilité ?

1 - Introduction au scoring

Le score est une solution qui permet d'identifier les bons des mauvais payeurs.

Un score donné reflète une probabilité d'être bon payeur.

- Plus la note est élevée, plus le risque est faible.
- Selon le risque que l'organisme de crédit pense pouvoir assumer, il fixera un seuil entre les clients qu'il peut accepter et les clients à refuser.

note min	note max	tranche	% population	Taux MP	Taux MP restant
0	304	1	9,5%	6,5%	15,0%
305	329	2	9,0%	3,3%	8,6%
330	349	3	10,2%	2,2%	5,3%
350	364	4	9,0%	1,2%	3,1%
365	379	5	9,8%	0,8%	1,9%
380	394	6	9,9%	0,5%	1,1%
395	414	7	12,4%	0,4%	0,6%
415	434	8	10,5%	0,2%	0,2%
435	459	9	9,5%	0,1%	0,1%
460	∞	10	10,2%	0,0%	0,0%
			100,00%	15,0%	

Dans cet exemple, la banque ne devrait accepter que les clients ayant un score supérieur à 350 pour être rentable.

1 - Introduction au scoring

Remarques sur le contenu du cours :

- La construction de modèles de score s'inscrit dans le cadre de l'apprentissage supervisé : la base d'apprentissage contient la variable à expliquer.
- Ce cours est à considérer comme un exemple de mise en œuvre de techniques statistiques orienté vers la pratique en environnement bancaire, ce qui impose quelques contraintes, qu'on peut ne pas retrouver dans d'autres domaines :
 - ✓ Classification binaire : deux sous-groupes (les bon payeurs BP et les mauvais payeurs MP).
 - ✓ Objectif : prédire l'appartenance d'un nouvel individu à l'un de ces groupes.
 - ✓ Contraintes de stabilité, de lisibilité, transparence et d'implémentation.

2 – Initialisation d'un projet de scoring

Etape 1: Formalisation du problème

Avant de se lancer dans la modélisation, il faut toujours se poser les questions suivantes :

Qu'est-ce qu'on veut modéliser ?

- Une probabilité de non-remboursement à un horizon précis ? Une probabilité de fraude ? ... etc.
- Comment définir ma variable cible : BP (ou MP) ?

A quel besoin répond mon modèle ?

- Prendre une décision à l'octroi du crédit ? Démarcher les clients les plus fiables ? Différencier le pricing ?

Sur quel périmètre va-t-il être appliqué ?

- Permet de choisir les individus à intégrer dans la base de modélisation; quels profils exclure?

De quel historique ai-je besoin pour construire mon modèle ?

- Un historique récent ? Y-a-t-il des périodes exceptionnelles à exclure ? Faut-il couvrir un cycle économique ?

2 – Initialisation d'un projet de scoring

Etape 1: Formalisation du problème

Avant de se lancer dans la modélisation, il faut toujours se poser les questions suivantes :

Qu'est-ce qu'on veut modéliser ?

- Une probabilité de non-remboursement à un horizon précis ? Une probabilité de fraude ? ... etc.
- Comment définir ma variable cible : BP (ou MP) ?

A quel besoin répond mon modèle ?

- Prendre une décision à l'octroi du crédit ? Démarcher les clients les plus fiables ? Différencier le pricing ?

Sur quel périmètre va-t-il être appliqué ?

- Permet de choisir les individus à intégrer dans la base de modélisation; quels profils exclure?

De quel historique ai-je besoin pour construire mon modèle ?

- Un historique récent ? Y-a-t-il des périodes exceptionnelles à exclure ? Faut-il couvrir un cycle économique ?

2 – Initialisation d'un projet de scoring

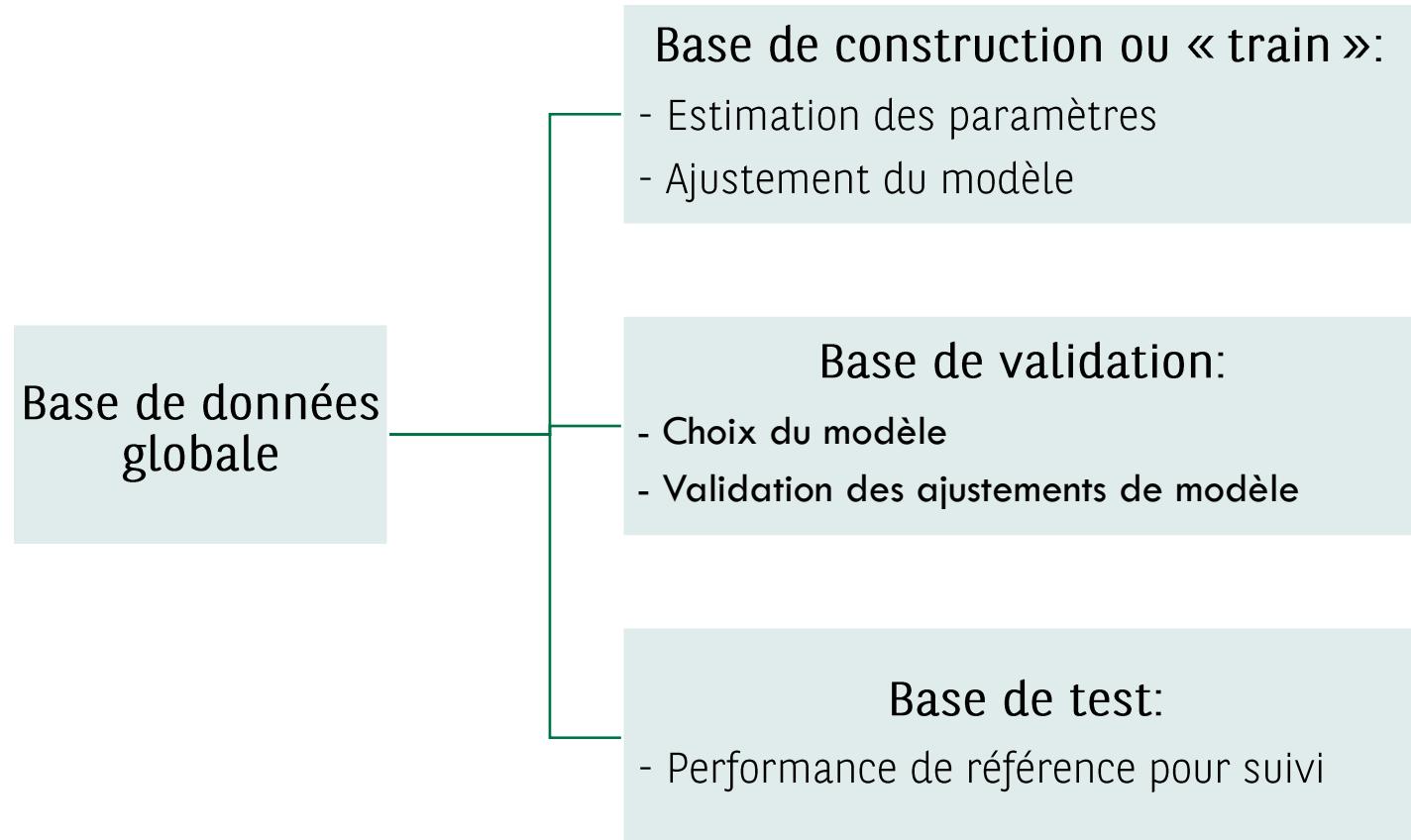
Etape 2: Construction de la base de données

Identifiant		Variables explicatives					Variable cible
A	61	Célibataire	Artisan	Locataire	22	1	
B	80	Marié	Artisan	Propriétaire	4	1	
C	22	Marié	Artisan	Propriétaire	5	1	
D	47	Marié	Commercant	Locataire	29	0	
E	71	Marié	Agriculteur	Propriétaire	30	1	
F	35	Marié	Commercant	Propriétaire	13	0	
G	48	Veuf	Agriculteur	Propriétaire	4	1	
H	51	Veuf	Medecin	Propriétaire	23	0	
I	67	Marié	Artisan	Locataire	2	1	
J	26	Célibataire	Commercant	Locataire	14	1	
...	



2 – Initialisation d'un projet de scoring

Etape 3 : Echantillonnage



3 – Analyse des variables explicatives

Problématique de la sélection de variables :

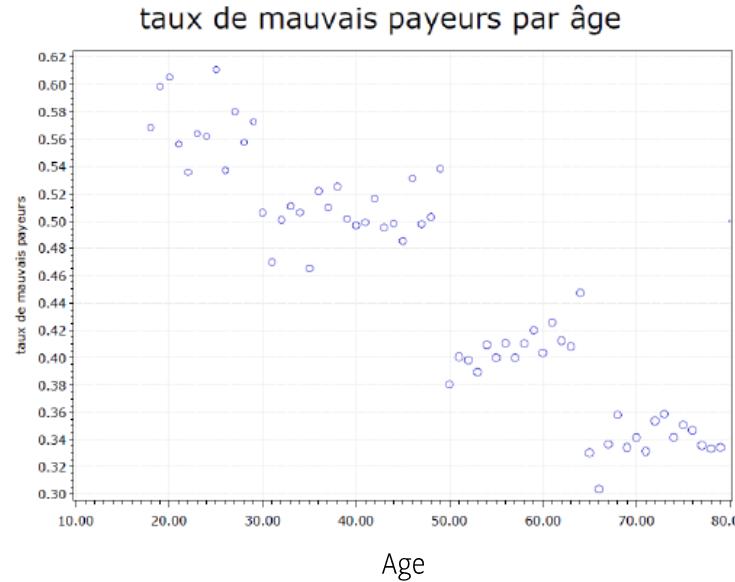
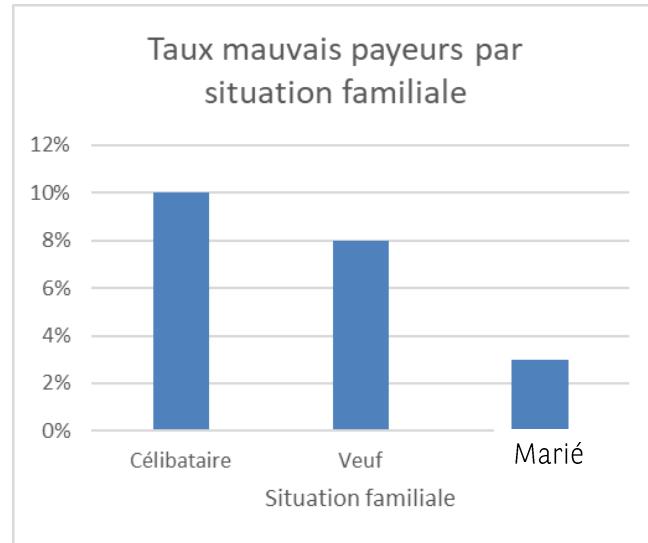
L'**usage**, lors de la construction d'un modèle de score de risque, est dans un premier temps d'identifier les variables explicatives les plus intéressantes, c'est-à-dire présentant des propriétés de :

- Discrimination
- Information
- Significativité (la discrimination n'est pas le fait du hasard)
- Interprétabilité (« bon sens »)
- Stabilité de l'effet discriminant dans le temps (on veut que le score fonctionne sur les demandes à venir)

3 – Analyse des variables explicatives

Approche intuitive :

Observer le taux de mauvais payeurs par modalité d'une variable :



Limites:

- L'utilisation du taux par modalité suppose une répartition de la variable en classes, ce qui n'est pas le cas des variables quantitatives.
- Ce type de visualisation graphique ne permet pas de voir la répartition de la population selon les classes de la variable.

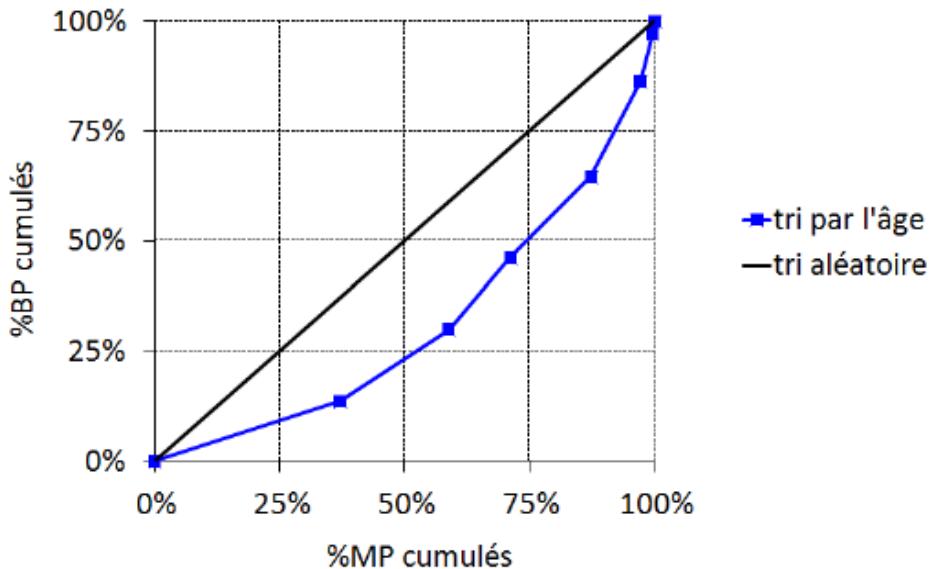
3 – Analyse des variables explicatives

Discrimination :

- Une variable a un bon pouvoir discriminant si elle permet de bien séparer la population des bons payeurs (BP) des mauvais payeurs (MP).

La courbe de concentration est un outil pour observer la discrimination d'une variable :

- Abscisse : % cumulé des MP
- Ordonnée : % cumulé des BP



3 – Analyse des variables explicatives

La courbe de concentration

Pour tracer cette courbe pour la variable « âge » :

- La population est triée par âge croissant.
- Les fonctions de répartition de l'âge sont calculées sur les populations de BP et MP (colonnes %BP cumulé et %MP cumulé).

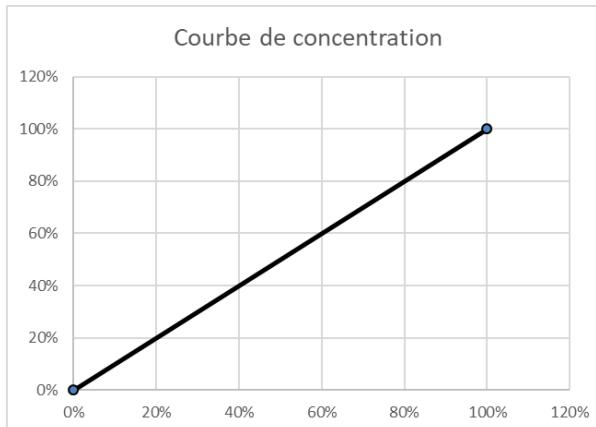
âge (années)	nb BP	nb MP	%BP	%MP	%BP cumulé	%MP cumulé
<20	250	300	14%	37%	14%	37%
20-29	300	175	16%	22%	30%	59%
30-39	300	100	16%	12%	46%	71%
40-49	340	130	18%	16%	65%	87%
50-59	400	80	22%	10%	86%	97%
60-69	200	20	11%	2%	97%	100%
≥ 70	50	3	3%	0%	100%	100%
total	1 840	808	100%	100%		

3 – Analyse des variables explicatives

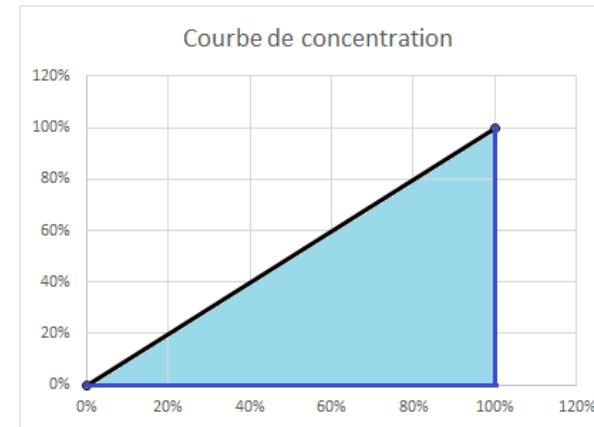
Lecture de la courbe de concentration :

- La courbe est croissante par définition.
- Cette courbe synthétise l'information sur la répartition de la population selon la variable ainsi que le taux de mauvais payeur par modalité.
- Plus la courbe est loin de la bissectrice, plus la variable est discriminante.

Discrimination nulle,
répartition aléatoire des MP



Discrimination parfaite,
les MP sont parfaitement isolés des BP



3 – Analyse des variables explicatives

L'indice de Gini :

- Permet de quantifier la discrimination représentée dans la courbe de concentration.

$$\text{Indice de Gini} = \frac{\text{aire rouge}}{\text{aire rouge+aire verte}} = 2 \times \text{aire rouge} = 1 - 2 \times \text{aire verte}$$

- L'indice de Gini est compris entre 0 et 1 :

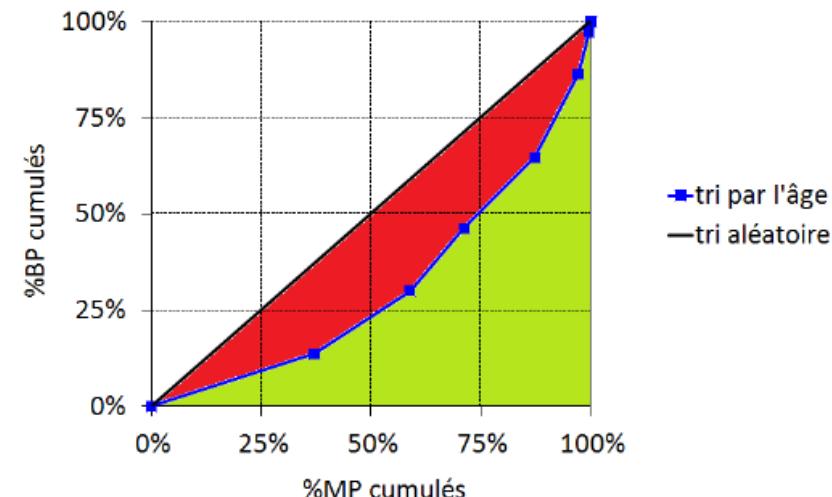
- Indice à 0 : la variable ne présente aucun caractère prédictif.
- Indice à 1 : la variable permet de séparer parfaitement les bons payeurs des mauvais.

Méthode de calcul du Gini :

Formule de Brown (méthode des trapèzes)

$$Gini = 1 - \sum_{k=0}^{n-1} (x_{k+1} - x_k)(y_{k+1} + y_k)$$

x étant % MP cumulés
 y étant % BP cumulés



3 – Analyse des variables explicatives

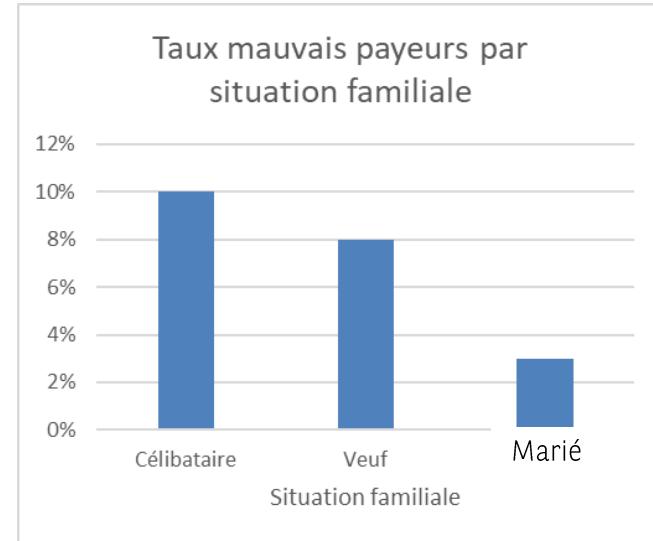
Exercice :

Cas 1 :

	Taux MP	%population	%MP cumulé	%BP cumulé
Célibataire	10%	40%	63%	38%
Veuf	8%	10%	76%	48%
Marié	3%	50%	100%	100%

Cas 2 :

	Taux MP	%population	%MP cumulé	%BP cumulé
Célibataire	10%	75%	86%	74%
Veuf	8%	10%	95%	84%
Marié	3%	15%	100%	100%



1. Calculer l'indice de Gini pour chacun des 2 cas.
2. Dans quel cas la variable « situation familiale » est-elle la plus discriminante ?

3 – Analyse des variables explicatives

Correction :

Cas 1 :

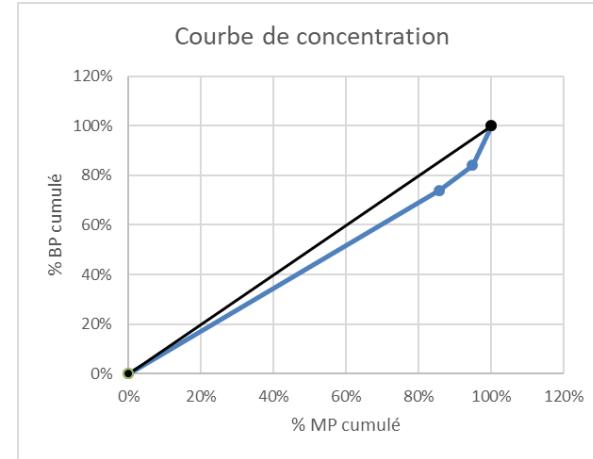
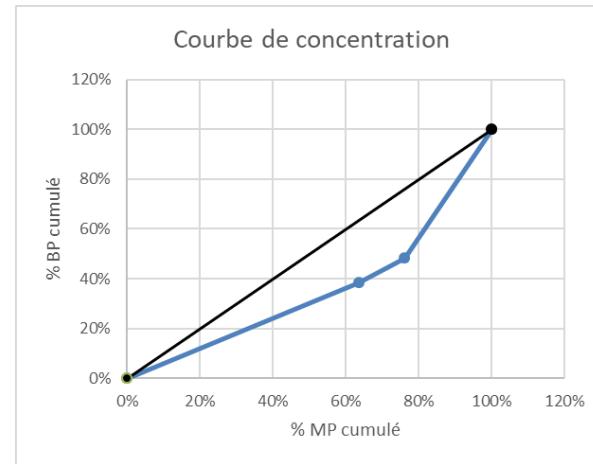
	Taux MP	%population	%MP cumulé	%BP cumulé
Célibataire	10%	40%	63%	38%
Veuf	8%	10%	76%	48%
Marrié	3%	50%	100%	100%

Gini = 29%

Cas 2 :

	Taux MP	%population	%MP cumulé	%BP cumulé
Célibataire	10%	75%	86%	74%
Veuf	8%	10%	95%	84%
Marrié	3%	15%	100%	100%

Gini = 13%



La variable « situation familiale » est plus discriminante dans le cas 1.

3 – Analyse des variables explicatives

Information :

- Une variable apporte de l'information sur la prédiction des mauvais payeurs, lorsque ses répartitions sur les bons et mauvais payeurs sont différentes.

➔ Comment quantifier l'information apportée par modalité ?

Exemple:

- Si on ne connaît rien d'un client, sa cote (odds) est :

$$\frac{P[BP=1]}{P[MP=1]} = \frac{\text{Nombre BP total}}{\text{Nombre MP total}} = 1,28$$

➔ Ça veut dire qu'un client pris au hasard a 1,28 fois plus de chance d'être BP que d'être MP

- Si on sait que le client a moins de 30 ans, sa cote est :

$$\frac{P[BP=1|\text{âge}<30]}{P[MP=1|\text{âge}<30]} = \frac{\text{Nombre BP}<30 \text{ ans}}{\text{Nombre MP}<30 \text{ ans}} = 0,75$$

➔ Un client qui a moins de 30 ans a plus de chances d'être mauvais payeur, le risque est plus fort dans cette catégorie que dans le reste de la population.

modalité	BP	MP	%BP	%MP
18-29 ans	2 527	3 368	11%	18%
30-49 ans	6 260	6 357	26%	34%
50-64 ans	6 837	4 709	29%	25%
>65 ans	8 187	4 204	34%	23%
total	23 811	18 638	100%	100%

3 – Analyse des variables explicatives

Exemple (suite) :

- Si on connaît l'âge du client, sa cote (odds) est : $\frac{P[BP=1|\text{âge}=A]}{P[MP=1|\text{âge}=A]}$

D'après le théorème de Bayes :

$$\frac{P[BP=1|\text{âge}=A]}{P[MP=1|\text{âge}=A]} = \frac{\cancel{P[\text{âge}=A|BP=1] \times P[BP=1]}}{\cancel{P[\text{âge}=A]}} \times \frac{\cancel{P[\text{âge}=A]}}{P[\text{âge}=A|MP=1] \times P[MP=1]}$$
$$\frac{P[BP=1|\text{âge}=A]}{P[MP=1|\text{âge}=A]} = \boxed{\frac{P[\text{âge}=A|BP=1]}{P[\text{âge}=A|MP=1]}} \times \frac{P[BP=1]}{P[MP=1]} \quad \text{Cote moyenne}$$

Information sur l'impact de la modalité sur la cote moyenne, indépendante des proportions de bons et mauvais payeurs dans la population globale

Définition: Poids d'évidence d'une modalité woe (weight of evidence):

- $woe_{\text{modalité}} = \ln\left(\frac{P[\text{modalité}|BP=1]}{P[\text{modalité}|MP=1]}\right)$
- $woe_{\text{modalité}} > 0 \Leftrightarrow$ le risque de la modalité est plus faible que celui de la population

3 – Analyse des variables explicatives

Définition: Valeur d'information IV (Information value):

- $IV_{modalité} = (P[modalité|BP = 1] - P[modalité|MP = 1]) \times woe_{modalité}$
- $IV_{modalité} \geq 0$
- $IV_{modalité} = 0 \Leftrightarrow$ la cote de la modalité égale celle de l'ensemble de la population
- IV d'une variable:

$$IV = \sum_{modalités} IV_{modalité}$$

- IV est un indicateur de dissimilarité entre les distributions de bons et mauvais payeurs.
- Permet de comparer (classer) le pouvoir discriminant de différentes variables.

Exercice :

Calculer les woe et IV pour les modalités de la variable âge (tableau ci-dessous) :

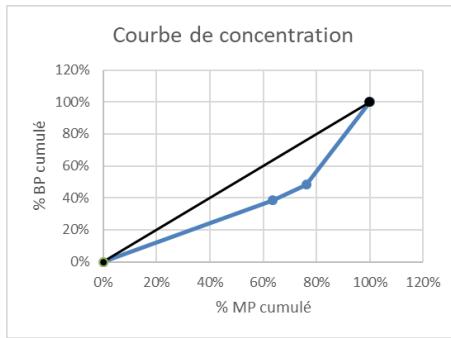
modalité	BP	MP	%BP	%MP
18-29 ans	2 527	3 368	11%	18%
30-49 ans	6 260	6 357	26%	34%
50-64 ans	6 837	4 709	29%	25%
>65 ans	8 187	4 204	34%	23%
total	23 811	18 638	100%	100%

woe	iv
-0.53	0.04
-0.26	0.02
0.13	0.00
0.42	0.05
	0.11

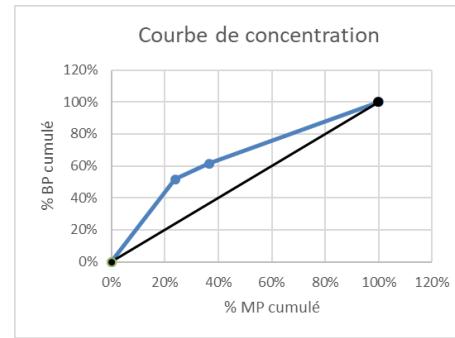
3 – Analyse des variables explicatives

Points d'attention :

La courbe de concentration (sur la base de laquelle est calculé l'indice de Gini) n'a de sens que si la variable a été préalablement ordonnée selon le risque :



Si les modalités sont classées de la plus risquée à la moins risquée: la courbe est en dessous de la bissectrice



Si les modalités sont classées de la moins risquée à la plus risquée: la courbe est au-dessus de la bissectrice

Il est souvent pertinent de mettre en classe les variables (quantitatives ou qualitatives) avant de se lancer dans la modélisation :

- l'indice de Gini, poids d'évidence (woe) et valeur d'information (VI) sont des indicateurs importants à regarder lors de cette mise en classes.
- Eviter d'agréger des petits groupes par proximité de poids d'évidence sans aucune interprétation.

3 – Analyse des variables explicatives

Découpages des variables :

Il est commun de discréteriser les variables quantitatives ou de faire des regroupements sur des variables qualitatives pour des raisons de lisibilité, prendre en compte des effets non linéaires, maîtriser l'effet des valeurs extrêmes ...etc

Objectif : Définir des groupes ayant des niveaux de risque différents tout en minimisant la perte d'information par rapport à la variable initiale.

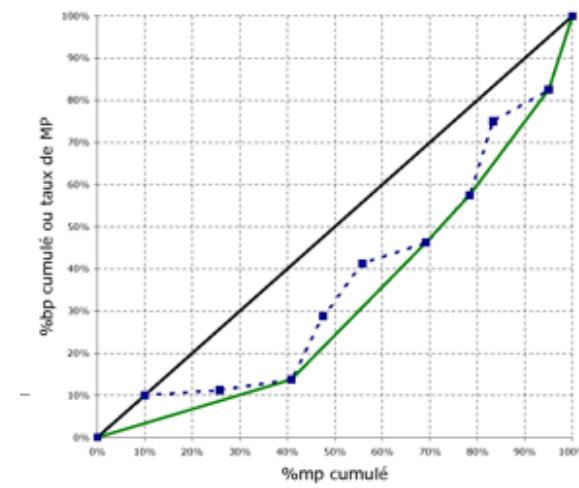
Les usages et bonnes pratiques en termes de découpage :

- Les classes doivent être différencierées en risque.
- Les volumes par classe doivent être suffisants.
- Le découpage doit avoir un sens métier interprétable.
- Les classes après découpage doivent être monotones en risque.

3 – Analyse des variables explicatives

Exemple :

X	nMP	nBP	%MP	%BP	Tx MP	%MP cumulés	%BP cumulés	woe
1	60	40	10.0%	10.0%	60.0%	10.0%	10.0%	0
2	95	5	15.8%	1.3%	95.0%	25.8%	11.3%	-2.5390
3	90	10	15.0%	2.5%	90.0%	40.8%	13.8%	-1.7918
4	40	60	6.7%	15.0%	40.0%	47.5%	28.8%	0.8109
5	50	50	8.3%	12.5%	50.0%	55.8%	41.3%	0.4055
6	80	20	13.3%	5.0%	80.0%	69.2%	46.3%	-0.9808
7	55	45	9.2%	11.3%	55.0%	78.3%	57.5%	0.2048
8	30	70	5.0%	17.5%	30.0%	83.3%	75.0%	1.2528
9	70	30	11.7%	7.5%	70.0%	95.0%	82.5%	-0.4418
10	30	70	5.0%	17.5%	30.0%	100.0%	100.0%	1.2528
Total	600	400	100.0%	100.0%	-	-	-	



- La découpe optimisant l'indice de Gini est obtenue en enveloppant la courbe brute.
- Avant la découpe : pas de monotonie sur les poids d'évidence.

3 – Analyse des variables explicatives

Exemple variable quantitative :

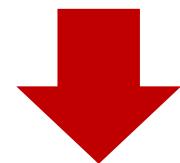
X	nMP	nBP	%MP	%BP	Tx MP	%MP cumulés	%BP cumulés	woe
1-3	245	55	40.8%	13.8%	81.7%	40.8%	13.8%	-1.0885
4-6	170	130	28.3%	32.5%	56.7%	69.2%	46.3%	0.1372
7	55	45	9.2%	11.3%	55.0%	78.3%	57.5%	0.2048
8-9	100	100	16.7%	25.0%	50.0%	95.0%	82.5%	0.4055
10	30	70	5.0%	17.5%	30.0%	100.0%	100.0%	1.2528
Total	600	400	100.0%	100.0%	-	-	-	

→ Après la découpe : monotonie sur les poids d'évidence.

3 – Analyse des variables explicatives

Exemple variable qualitative :

Situation habitation	#TOTAL	#BP	#MP	%TOTAL	%BP	%MP	woe	Taux de MP	Ivi
Propriétaire	2 464	2 349	115	36,9%	37,8%	24,5%	0,433	4,7%	0,058
Locataire	1 687	1 500	187	25,3%	24,1%	39,9%	-0,502	11,1%	0,079
Accédant	2 100	1 997	103	31,4%	32,1%	22,0%	0,381	4,9%	0,039
Logé famille	398	338	60	6,0%	5,4%	12,8%	-0,855	15,1%	0,063
Autre	32	28	4	0,5%	0,5%	0,9%	-0,638	12,5%	0,003
Total	6 681	6 212	469	100%	100%	100%		7,0%	0,2407



Attention: toujours se poser la question de la cohérence des regroupements pour les variables qualitatives.

Situation habitation	#TOTAL	#BP	#MP	%TOTAL	%BP	%MP	woe	Taux de MP	Ivi
1 Logé famille - Autre	430	366	64	6,4%	5,9%	13,6%	-0,840	14,9%	0,065
2 Locataire	1 687	1 500	187	25,3%	24,1%	39,9%	-0,502	11,1%	0,079
3 Propriétaire - Accédant	4 564	4 346	218	68,3%	70,0%	46,5%	0,409	4,8%	0,096
Total	6 681	6 212	469	100%	100%	100%		7,0%	0,2400

3 – Analyse des variables explicatives

Significativité :

Nous avons mesuré précédemment l'intensité de la liaison entre une variable et le fait d'être mauvais payeur (via l'indice de Gini, le woe, et la VI), nous allons maintenant analyser la significativité statistique de cette liaison (est-ce que le constat de liaison n'est pas le fruit du hasard).

Pour cela on fait un test d'hypothèse avec:

H_0 : Le fait d'être BP (ou MP) est indépendant la variable X

Statistique de test = la statistique du χ^2

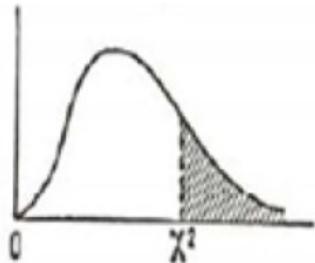
Propriété de la statistique du χ^2

Sous hypothèse d'indépendance, la statistique du χ^2 d'un tableau de contingence à N lignes et P colonnes suit une loi $\chi^2_{(N-1)(P-1)}$

3 – Analyse des variables explicatives

Table de χ^2 (*)

La table donne la probabilité α pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l.\alpha	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,155	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,282	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,221	5,353	7,175	8,349	10,520	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125

Rappel :

α correspond à l'erreur du risque de première espèce de ce test statistique.

→ Elle est égale à la probabilité de rejeter H_0 à tort.

→ $P_{H_0}(T > q_{1-\alpha}) < \alpha$

3 – Analyse des variables explicatives

Exercice :

Ci-dessous les effectifs observés sur la variable « situation familiale » :

Modalité	BP	MP	total	%colonne
Célibataire	360	40	400	40%
Veuf	92	8	100	10%
Marié	485	15	500	50%
Total	937	63	1000	100%
%ligne	94%	6%	100%	100%

1. Construire le tableau de contingence théorique (i.e. sous hypothèse d'indépendance).
2. Calculer la statistique du χ^2 de cette variable (T).
3. Cette statistique suit une loi du χ^2 à combien de degrés de liberté ?
4. Avec une erreur $\alpha=5\%$, quelle est la conclusion du test du χ^2 ?

3 – Analyse des variables explicatives

Correction:

1. Ci-dessous les effectifs sous hypothèse d'indépendance H_0 :

Modalité	BP	MP	total	%colonne
Célibataire	375	25	400	40%
Veuf	94	6	100	10%
Marrié	469	32	500	50%
Total	937	63	1000	100%
%ligne	94%	6%	100%	100%

$$375 = \frac{937}{1000} \times 400$$

2. $T = \frac{(360-375)^2}{375} + \frac{(92-94)^2}{94} + \frac{(485-469)^2}{469} + \frac{(40-25)^2}{25} + \frac{(8-6)^2}{6} + \frac{(15-32)^2}{32} = 18,99$
3. T suit une loi du χ^2 à 2 degrés de liberté.
4. Le quantile à 95% d'une χ^2 à 2 degrés de liberté est 5,99

Pour un test de niveau $\alpha=5\%$, H_0 est rejetée si $T>5,99$

Comme $T=18,99>5,99$ on rejette H_0

- Une autre approche consiste à calculer l'erreur de première espèce (p-value):

Sous H_0 : $P(T>18,99)=7,52 \times 10^{-5}<5\%$

→ Le lien entre situation familiale et risque est significatif.

3 – Analyse des variables explicatives

Remarques sur les tests :

- Bien distinguer significativité statistique et « intensité de la liaison » (écart de risque entre modalités) : à proportions données, **la significativité augmente mécaniquement avec le nombre d'observations.**
 - Le rejet de l'hypothèse d'indépendance ne veut pas dire que tous les écarts de risque entre modalités sont deux à deux distincts.
- ➔ **Ne pas interpréter aveuglément une statistique de significativité.**

4 – Modélisation : Régression Logistique

La partie précédente du cours présentait des outils permettant d'analyser les variables une à une en quantifiant l'intensité de leur lien avec la variable à prédire, ainsi qu'en testant la significativité de ce lien.

Il faut maintenant construire un modèle qui regroupe toute l'information contenue dans la base de données pour prédire au mieux la variable cible (ici BP).

Propriétés attendues du modèle :

- lisibilité
- simplicité (calibration et implémentation)
- performance
- robustesse (bonne capacité de généralisation)
- utilisation exhaustive de l'information

4 – Modélisation : Régression Logistique

La variable cible (BP) est une variable binaire (prend les valeurs 0 ou 1).

➤ Nous disposons de plusieurs modèles de classification pour pouvoir prédire MP à partir des variables explicatives contenues dans la base de données :

- Régression logistique
- Arbres de décision
- Random Forest
- Réseau de neurones ... etc

Dans ce cours nous allons utiliser la régression logistique.

Pourquoi ?

- C'est le modèle classique utilisé en scoring bancaire.
- C'est un modèle explicite et interprétable.
- En économétrie, c'est le modèle « canonique » dans le cas d'un critère cible binaire (au même titre que le modèle linéaire dans le cas d'un critère cible continu).

4 – Modélisation : Régression Logistique

Théorie :

Y = variable cible binaire

X = vecteur de variables explicatives (X_1, X_2, \dots, X_J)

β = vecteur des coefficients du modèle ($\beta_0, \beta_1, \dots, \beta_J$)

Le modèle de régression linéaire classique n'est pas adapté à l'explication d'une variable binaire.

Objectif : Adapter la modélisation linéaire à cette situation.

Idée : Expliquer $p = P(Y = 1)$ ou plutôt une transformation de celle-ci (via une fonction f), pour que la variable cible prenne des valeurs sur \mathbb{R} : ainsi nous pouvons utiliser un modèle linéaire de la forme :

$$f(p) = \beta X$$

➔ Pour la régression logistique nous utilisons la fonction **logit**:

$$f(x) = \text{logit}(x) = \ln\left(\frac{x}{1-x}\right) \text{ avec } f^{-1}(x) = \frac{1}{(1+e^{-x})}$$

4 – Modélisation : Régression Logistique

Théorie :

Modèle logistique :

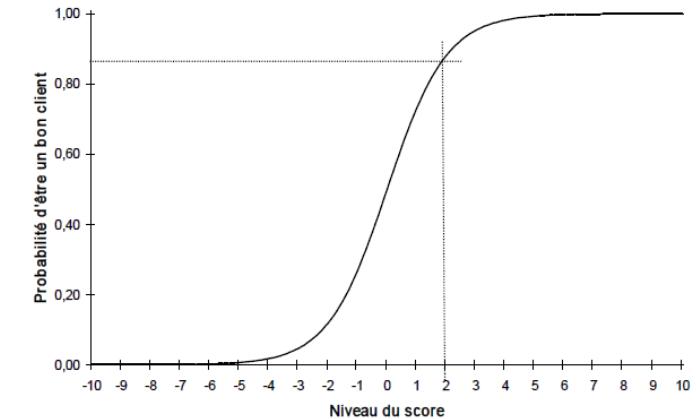
$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta X$$

$$P(Y = 1) = \frac{1}{(1 + e^{-\beta X})}$$

Score
brut

Estimation :

Le vecteur β peut être estimé par maximum de vraisemblance.



La vraisemblance (L) d'un échantillon (y_1, y_2, \dots, y_n) , n variables aléatoires binaires iid suivant une Bernoulli, est définie comme la probabilité d'observer cet échantillon

$$L = \prod_{i=1}^n P(y_i = 1)^{1_{y_i=1}} (1 - P(y_i = 1))^{1_{y_i=0}}$$

Idée derrière l'estimateur de maximum de vraisemblance : l'échantillon observé est celui qui était « le plus probable ».

4 – Modélisation : Régression Logistique

Mise en œuvre du maximum de vraisemblance pour une régression logistique :

$$L_\beta = \prod_{i=1}^n P(y_i = 1)^{\mathbb{1}_{y_i=1}} (1 - P(y_i = 1))^{\mathbb{1}_{y_i=0}}$$

$$\Leftrightarrow L_\beta = \prod_{i \in BP} \frac{1}{(1 + e^{\beta X_i})} \prod_{i \in MP} \frac{1}{(1 + e^{\beta X_i})}$$

$$\ln(L_\beta) = \sum_{i \in BP} -\ln(1 + e^{-\beta X_i}) - \sum_{i \in MP} \ln(1 + e^{\beta X_i})$$

$$\Leftrightarrow \ln(L_\beta) = \sum_{i \in BP} -\ln(e^{-\beta X_i}) - \sum_{i \in BP} \ln(1 + e^{\beta X_i}) - \sum_{i \in MP} \ln(1 + e^{\beta X_i})$$

$$\Leftrightarrow \ln(L_\beta) = \sum_{i \in BP} \beta X_i - \sum_{i \in [1:n]} \ln(1 + e^{\beta X_i})$$

$\beta_{EMV} = \operatorname{argmax}_\beta (\ln(L_\beta))$ est obtenu pour $\frac{\partial \ln(L_\beta)}{\partial \beta} = 0$

$$\frac{\partial \ln(L_\beta)}{\partial \beta_j} = \sum_{i \in BP} X_{ij} - \sum_{i \in [1:n]} \frac{e^{\beta X_i}}{(1 + e^{\beta X_i})} X_{ij} = 0$$

$$\Leftrightarrow \sum_{i \in BP} X_{ij} = \sum_{i \in [1:n]} \frac{1}{(1 + e^{-\beta X_i})} X_{ij}$$

L'estimateur de maximum de vraisemblance ne se calcule pas directement mais se résout par des méthodes numériques d'optimisation comme la descente de gradient ou la méthode de Newton-Raphson.

4 – Modélisation : Régression Logistique

Significativité des coefficients du modèle :

- La plupart des logiciels statistiques affichent, pour chaque coefficient estimé, le résultat du **test de Wald** (H_0 : le coefficient est nul).
- Sous H_0 , la statistique de Wald suit une loi du χ^2 à 1 degré de liberté.
- L'usage est souvent de ne retenir que les variables pour lesquelles l'hypothèse nulle (H_0 : le coefficient est nul) est rejetée.

➔ Est-il pertinent de tester cette hypothèse ?

➔ Pour cela il faut d'abord se demander ce que signifient les coefficients estimés.

4 – Modélisation : Régression Logistique

Coefficients β :

- Le coefficient β_i estimé (associé à la variable X_i) indique de combien varie en moyenne la valeur des log odds lorsque X_i augmente d'une unité.
- Cas de variables codées en indicatrices : le coefficient β_i , associé à l'indicatrice A_i de la variable A, indique de combien varie en moyenne la valeur des log odds lorsque A passe de la modalité de référence à la modalité représentée par A_i .

Exemple :

Soit A la variable situation familiale, A prend 3 modalités {célibataire, marié, veuf}. Si on veut coder A en indicatrices plusieurs options s'offrent à nous :

	Option 1		Option 2		Option 3	
A	A1	A2	A1	A2	A1	A2
célibataire	0	0	1	0	1	0
marié	1	0	0	1	0	0
veuf	0	1	0	0	0	1

Dans l'option 1 la modalité de référence est « célibataire ».

- Dans une régression logistique incluant A1 et A2 de l'option 1:
 - le score varie de β_1 si l'individu passe de **célibataire à marié**,
 - et de β_2 si l'individu passe de **célibataire à veuf**.

4 – Modélisation : Régression Logistique

Exemple :

Soit A et B deux variables ayant 3 modalités chacune A={ia1, ia2, ia3} et B={ib1, ib2, ib3}

On construit une régression logistique avec A et B codées en indicatrices :

Opt 1: ia3 et ib3 sont les modalités de référence:

Opt 2: ia1 et ib1 sont les modalités de référence:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0075	0.2119	17.5349	<.0001
ia1	1	-1.9017	0.2732	48.4546	<.0001
ia2	1	-1.3264	0.2412	30.2499	<.0001
ia3	0	0	.	.	.
ib1	1	-1.2980	0.2594	25.0377	<.0001
ib2	1	-0.7854	0.2480	10.0281	0.0015
ib3	0	0	.	.	.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3122	0.2925	62.4966	<.0001
ia1	0	0	.	.	.
ia2	1	0.3753	0.2860	4.0464	0.0443
ia3	1	1.9017	0.2732	48.4546	<.0001
ib1	0	0	.	.	.
ib2	1	0.5126	0.2721	3.3484	0.0396
ib3	1	1.2980	0.2594	25.0377	<.0001

→ On remarque que la significativité des coefficients change selon la modalité de référence choisie. En effet, selon la modalité de référence choisie, le sens des coefficients de la régression est différent.

Attention : la significativité des coefficients peut aussi changer selon le volume de la base (sur des volumes de bases plus faibles les résultats de tests de significativité seront détériorés).

Il ne faut pas interpréter aveuglément les tests de significativité !

4 – Modélisation : Régression Logistique

Coefficients β :

Modèle avec une seule variable explicative :

→ Les coefficients de la régression peuvent être calculés directement à partir des poids d'évidence des modalités de la variable.

Exercice:

Soit A = {célibataire, marié, veuf}

Tableau de poids d'évidence et valeur d'information :

sit.mat	BP	MP	BP/MP	%BP	%MP	woe	iv
1 :célibataire	766	2 638	0.2904	24.7%	38.3%	-0.4402	0.060
2 : autre	986	2 352	0.4192	31.7%	34.1%	-0.0730	0.002
3 : marié	1 356	1 902	0.7129	43.6%	27.6%	0.4580	0.073
total	3 108	6 892	0.4510	100.0%	100.0%		0.135

Calculer les coefficients d'un modèle de régression logistique contenant uniquement la variable A, codée en indicatrices simples comme suit :

A	A1	A2
célibataire	1	0
marié	0	0
autre	0	1

4 – Modélisation : Régression Logistique

Correction :

$$\ln\left(\frac{P(BP/A)}{P(MP/A)}\right) = \alpha + \beta_1 A_1 + \beta_2 A_2 \Leftrightarrow woe_A + \ln\left(\frac{P(BP)}{P(MP)}\right) = \alpha + \beta_1 A_1 + \beta_2 A_2$$

- Si A=Célibataire $\Rightarrow A_1 = 1, A_2 = 0 \Rightarrow woe_{A=\text{Célibataire}} + \ln\left(\frac{P(BP)}{P(MP)}\right) = \alpha + \beta_1$
 - Si A=autre $\Rightarrow A_1 = 0, A_2 = 1 \Rightarrow woe_{A=\text{autre}} + \ln\left(\frac{P(BP)}{P(MP)}\right) = \alpha + \beta_2$
 - Si A=marié $\Rightarrow A_1 = 0, A_2 = 0 \Rightarrow woe_{A=\text{marié}} + \ln\left(\frac{P(BP)}{P(MP)}\right) = \alpha$
-
- $\beta_1 = woe_{A=\text{Célibataire}} - woe_{A=\text{marié}} = -0,8982$
 - $\beta_2 = woe_{A=\text{autre}} - woe_{A=\text{marié}} = -0,5310$
 - $\alpha = woe_{A=\text{marié}} + \ln\left(\frac{\text{Effectif BP}}{\text{Effectif MP}}\right) = -0,3384$

4 – Modélisation : Régression Logistique

Indicatrices imbriquées :

Coder les variables en indicatrices imbriquées au lieu d'indicatrices simples permet de s'affranchir du choix d'une modalité de référence, et pouvoir interpréter plus sereinement les tests de significativités.

Qu'est-ce que des indicatrices imbriquées ?

- Ces indicatrices sont relatives à des ensembles inclus les uns dans les autres.

Exemple:

Soit A = {célibataire, marié, veuf}, un exemple de codage en indicatrices imbriquées :

	Indicatrices imbriquées	
A	A1	A2
célibataire	0	0
marié	1	0
veuf	1	1

- Dans une régression logistique incluant A1 et A2 du tableau:
 - le score varie de β_1 si l'individu passe de **célibataire à marié**,
 - et de β_2 si l'individu passe de **marié à veuf**.

4 – Modélisation : Régression Logistique

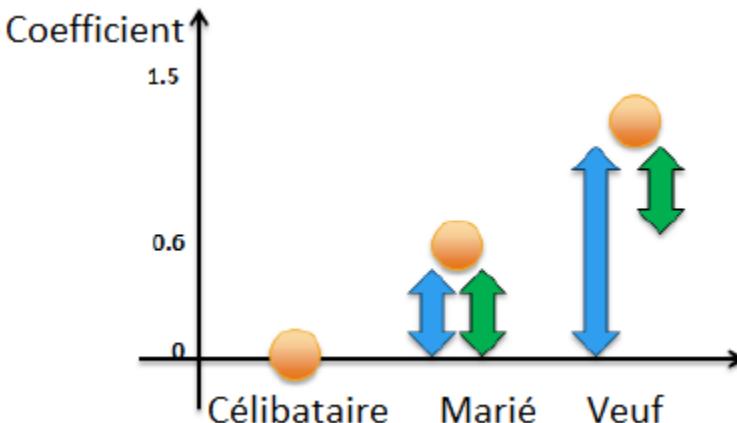
Indicatrices imbriquées :

Attention : Il existe plusieurs options pour coder une variable en indicatrices imbriquées. Dans le cas du scoring **on classe les modalités de la plus risquée à la moins risquée**, pour que tous les coefficients soient positifs.

Exemple:

	Indicatrices non imbriquées		Indicatrices imbriquées	
A	A1	A2	A1	A2
célibataire	0	0	0	0
marié	1	0	1	0
veuf	0	1	1	1

Modalité	Coefficient non imbriqué	Stat Wald	Coefficient imbriqué	Stat Wald
1	0	-	0	-
2	0.604	98.49	0.604	98.49
3	1.513	655.88	0.909	296.19



4 – Modélisation : Régression Logistique

Exercice :

Coder en indicatrices imbriquées la variable suivante :

Modalités	BP	MP	%MP par catégorie
Cadres	8187	4204	34%
Commerçant	6837	4709	41%
Artisan	6260	6357	50%
Etudiant	2527	3368	57%
Total	23811	18638	44%

4 – Modélisation : Régression Logistique

Exercice :

Coder en indicatrices imbriquées la variable suivante :

Modalités	BP	MP	%MP par catégorie
Cadres	8187	4204	34%
Commerçant	6837	4709	41%
Artisan	6260	6357	50%
Etudiant	2527	3368	57%
Total	23811	18638	44%

Correction :

Modalités	Indicatrices imbriquées		
	I1	I2	I3
Cadres	1	1	1
Commerçant	1	1	0
Artisan	1	0	0
Etudiant	0	0	0

La modalité la moins risquée est celle pour laquelle toutes les indicatrices sont égales à 1.

→ Pour que les coefficients de la régression soient positifs.

4 – Modélisation : Rappel

1. Bien définir sa problématique avant de se lancer dans la modélisation.
2. Commencer par analyser les variables une à une au moyen des woe, IV et indice de Gini (pour les variables numériques), cela permet de détecter les variables à fort potentiel.
3. Utiliser le Gini et les woe pour faire des mises en classes pertinentes.
4. Les coefficients d'une régression logistique sont obtenus par la méthode de maximum de vraisemblance.
5. Les coefficients d'une régression logistique sont homogènes à des woe, c'est-à-dire à des écarts de log odds.
6. Il est préférable de coder les variables en indicatrices imbriquées pour s'affranchir du choix d'une modalité de référence.
7. Toujours être attentif quand on interprète des tests de significativité.

4 – Modélisation : Construction d'un modèle

Méthodes de sélections de variables :

- Méthodes automatiques (stepwise) : méthodes ascendantes (ajout de variable) ou descendantes (retrait de variable).

➤ Ces méthodes optimisent un critère de significativité et non un critère d'intensité de relation.

Risque de sur-apprentissage : on trouve toujours quelque chose de significatif.

Méthode par récurrence utilisant l'information marginale :

➤ Point de départ : modèle « 0 » = modèle réduit à une constante.

➤ Etape n+1 :

1. Analyse marginale de chacune des variables candidates conditionnellement au modèle n (détailé dans les slides suivants)
2. Choix de la variable apportant le plus d'information (sous contrainte de significativité et d'interprétabilité)
3. Modèle n+1 = Modèle n + variable choisie à l'étape n (+contrôle de cohérence et de significativité)
4. Arrêt des itérations lorsqu'il n'existe plus d'information résiduelle significative

4 – Modélisation : Construction d'un modèle

Soit un modèle 1 construit avec les variables « âge », « CSP », « situation logement » et on se demande si la variable « situation familiale » apporte de l'information qui n'est pas déjà dans le modèle 1?

→ Pour répondre à cette question on compare les effectifs estimés et observés par modalité de « situation familiale » :

Situation familiale	BP réel	MP réel	BP prédit	MP prédit
Célibataire	766	2638	1068,03	2335,97
Autre	986	2352	1020,61	2317,39
Marié	1356	1902	1019,36	2238,64
Total	3108	6892	3108	6892

Les BP (et MP) prédits sont obtenus en sommant les probabilités estimées par le modèle 1 des individus de chaque modalité.

→ On calcule la cote estimée par le modèle vs la cote réelle:

$$\begin{aligned} \text{• Cote réelle : } & \frac{766}{2638} \approx 0,29 \\ \text{• Cote estimée : } & \frac{1068}{2336} \approx 0,46 \end{aligned}$$



Par quel facteur F doit on multiplier la cote estimée sur la modalité « célibataire » pour qu'elle soit égale à la cote réelle?

$$F = \frac{P(\text{Célibataire}/BP)}{P(\text{Célibataire}/MP)} / \frac{P(\text{Célibataire}/BP\text{prédit})}{P(\text{Célibataire}/MP\text{prédit})} \quad \Rightarrow \quad \ln(F) = woe_{célib} - \widehat{woe}_{célib}$$

4 – Modélisation : Construction d'un modèle

Poids d'évidence marginal d'une modalité:

$$dwoe_{mod} = woe_{mod} - \widehat{woe}_{mod}$$

- Le poids d'évidence marginal est l'écart entre le poids d'évidence de la modalité et son poids d'évidence « vu par le modèle ».
- $dwoe_{mod} > 0 \Leftrightarrow$ la cote de la modalité est supérieure à celle estimée par le modèle.

Valeur d'information marginale:

- *div d'une modalité:*

$$div_{mod} = (P[mod/BP] - P[mod/MP]) \times dwoe_{mod}$$

- *dIV d'une variable:*

$$dIV = \sum_{mod} div_{mod}$$

→ dIV s'interprète comme l'information de la variable non prise en compte dans le modèle.

4 – Modélisation : Construction d'un modèle

Exercice :

Soit un modèle 1 construit avec les variables « âge », « CSP », « situation logement » et on se demande si la variable « situation familiale » apporte de l'information qui n'est pas déjà dans le modèle 1?

1. Calculer le poids d'évidence marginal (dwoe) et la valeur d'information marginale (div) de la variable « situation familiale » :

Situation familiale	BP réel	MP réel	BP prédit	MP prédit
Célibataire	766	2638	1068,03	2335,97
Autre	986	2352	1020,61	2317,39
Marié	1356	1902	1019,36	2238,64
Total	3108	6892	3108	6892

2. Cette variable doit-elle être introduite dans le modèle ?

4 – Modélisation : Construction d'un modèle

Correction :

1. Calculer le poids d'évidence marginal (dwoe) et la valeur d'information marginale (div) de la variable « situation familiale » :

Situation familiale	%BP réel	%MP réel	%BP prédit	%MP prédit	woe	woe predict	dwoe	div
Célibataire	24,6%	38,3%	34,4%	33,9%	-0,44	0,01	-0,45	0,06
Autre	31,7%	34,1%	32,8%	33,6%	-0,07	-0,02	-0,05	0,00
Marié	43,6%	27,6%	32,8%	32,5%	0,46	0,01	0,45	0,07
Total	100%	100%	100%	100%				0,13

2. Cette variable doit-elle être introduit dans le modèle?

Comme $\text{div} = 0,13 > 0$, la variable « situation familiale » apporte de l'information additionnelle, il faut l'introduire dans le modèle.

4 – Modélisation : Construction d'un modèle

Significativité marginale:

On veut tester pour une variable donnée, si l'écart entre les cotes observées et estimées par le modèle est significatif :

H_0 = la distribution des BP et MP est celle évaluée par le modèle

→ On effectue un test du χ^2 (comme détaillé slide 26).

4 – Modélisation : Construction d'un modèle

Remarque sur la corrélation :

- La corrélation est l'existence d'une liaison affine entre deux variables.

Exemple : âge et salaire peuvent être très corrélés.

Ce n'est pas pour autant qu'il faut se restreindre à une seule variable : à âge donné le salaire peut être prédictif du risque, et inversement !

- ➔ Préférer la notion d'information marginale.
- ➔ Pas de règle générale liant corrélation, poids d'évidence et coefficients dans le modèle.

4 – Modélisation : Normalisation du score

Rappel : Le score brut correspond au $\beta X = \ln(\text{odds})$ estimé par le modèle logistique :

- Les valeurs du score brut sont d'usage incommodes (chiffres après la virgule, valeurs négatives ...), il est donc courant de normaliser le score pour le rendre plus lisible.

Pour Normaliser un score nous avons besoin de 4 paramètres:

- Odd ratio de référence (ex:1)
- Valeur de référence du score qui correspond au odd ratio de référence (ex: 300)
- Facteur multiplicatif (ex: 2)
- Nombre de points pour lequel les odds sont multipliés par le facteur multiplicatif (ex: 20)

$$\text{Score Normé} = \text{Valeur de référence} + \frac{\text{Points de variation}}{\ln(\text{Facteur multiplicatif})} \times (\text{Score brut} - \ln(\text{ratio de référence}))$$

$$\text{Ex: Score Normé} = 300 + \frac{20}{\ln(2)} \beta X$$

4 – Modélisation : Normalisation du score

Exemple:

Modalités		Coef bruts	Coef Normés	Coef centrés
Constante		-2,068	240	0
sit. Logement	autre	0	0	120
	locataire	0,604	17	137
sit matrimoniale	propriétaire	1,513	44	164
	celibataire	0	0	120
	autre	0,438	13	133
	marié	0,979	28	148

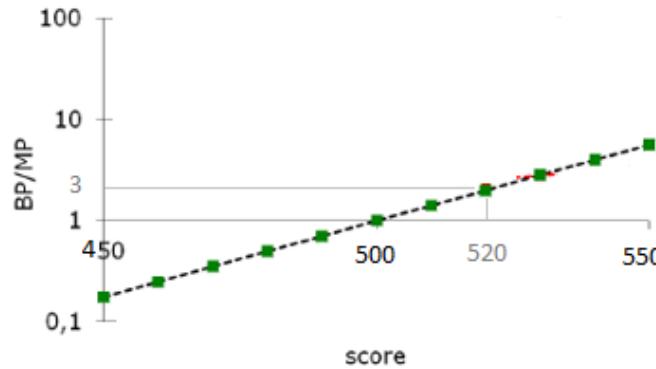
1. $cst_{normé} = 300 + \frac{20}{\ln(2)} cst$ et $\beta_i_{normé} = \frac{20}{\ln(2)} \beta_i$

2. Les coefficients centrés sont obtenus en répartissant la constante entre les variables :

sit.logement	autre	locataire	propriétaire
points	120	138	164
sit. Matrimoniale	célibataire	autre	marié
points	120	133	148

4 – Modélisation : Normalisation du score

1. En s'appuyant sur le graphe, retrouver les conventions qui ont été prises sur les odds pour normer le score.

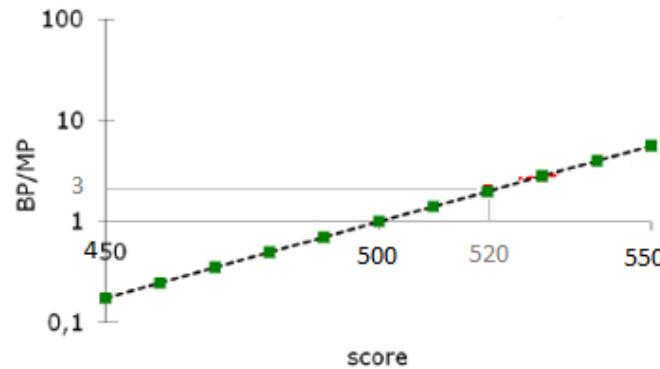


2. Calculer les coefficients normés de la régression suivante, selon les conventions de normalisation trouvées en 1.

	Coef Bruts
Constante	-0,3384
Marié	0
Célibataire	-0,8982
Autre	-0,531

4 – Modélisation : Normalisation du score

- En s'appuyant sur le graphe, retrouver les conventions qui ont été prises sur les odds pour normer le score.



- Ratio de référence = 1
- Note de référence = 500
- Facteur multiplicatif = 3
- Nombre de points = 20

- Calculer les coefficients normés de la régression suivante, selon les conventions de normalisation trouvées en 1.

	Coef Bruts	Coef Normés	Coef centrés
Constante	-0,3384	494	0
Marié	0	0	494
Célibataire	-0,8982	-16	477
Autre	-0,531	-10	484

4 – Modélisation : Normalisation du score

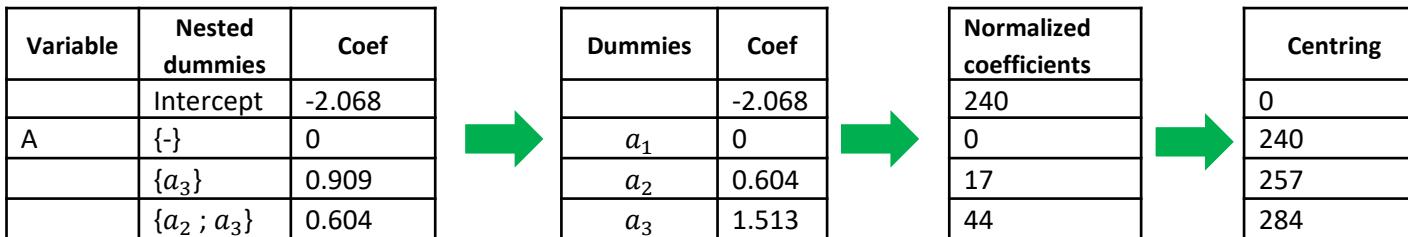
Exercice :

1. Ci-dessous les coefficients du modèle avec la variable A = {a1, a2, a3} codée en indicatrices imbriquées. Retrouvez les coefficients de ce modèle en cas d'usage d'indicatrices simples (modalité de référence = « logé en famille ») ?

Variable	Indicatrices imbriquées	Coef
	Intercept	-2.068
A	{-}	0
	I2	0.909
	I1	0.604

	Indicatrices imbriquées	
A	I1	I2
Propriétaire (a3)	1	1
Locataire (a2)	1	0
Logé famille (a1)	0	0

Correction:



5 – Evaluation du modèle

Critères souhaités pour le modèle :

➤ Le modèle doit être généralisable

Eviter le sur-apprentissage

➤ Performant

Indicateur de performance (indice de Gini, précision...etc)

➤ Robuste

Cohérence et significativité des coefficients

➤ Précis

Relation score – log(odds)

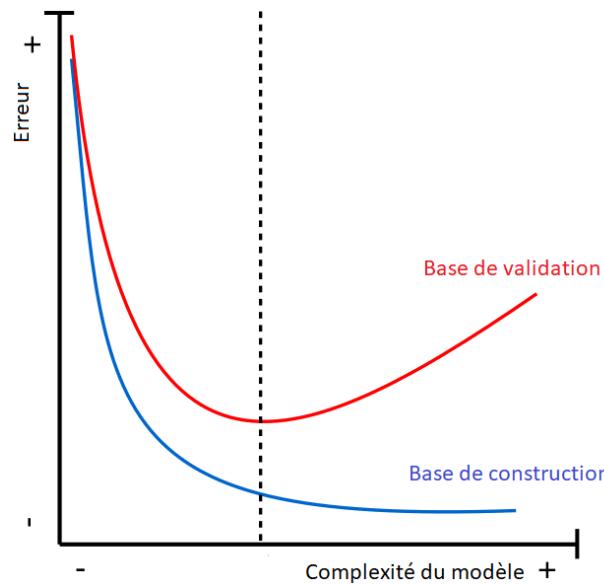
➤ Utilise l'exhaustivité de l'information

Toutes les variables non intégrées dans le modèle ont un dIV très faible

5 – Evaluation du modèle

Le modèle est estimé sur l'échantillon de construction et on teste ses performances sur un échantillon de validation ce qui nous permet d'éviter le surapprentissage.

→ En effet le modèle à choisir est celui qui minimise l'erreur sur l'échantillon de validation et non sur l'échantillon de construction (ou maximise la performance sur l'échantillon de validation).



En pratique :

On optimise le Gini sur la base de construction et on vérifie que le Gini sur la base de validation est dans l'intervalle de confiance du Gini de la base de construction :

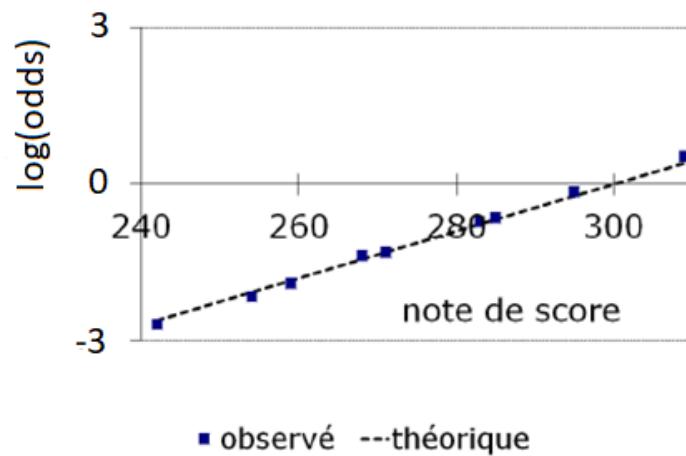
- Si c'est ok: on garde le modèle.
- Sinon on considère que le modèle surapprend (« over-fit ») donc on essaye de baisser la complexité du modèle.

5 – Evaluation du modèle

Qualité d'ajustement :

Vérifier la qualité d'ajustement d'un score *construit via une régression logistique* :

- Vérifier graphiquement que $\log\left(\frac{\text{effectif BP}}{\text{effectif MP}}\right)$ par note de score sont bien alignés et sont bien sur la droite théorique définie par le modèle.



En effet, par construction il existe une relation linéaire entre la note de score normé et les log(odd) :

$$\ln\left(\frac{P(BP/\text{score})}{P(MP/\text{score})}\right) = a \text{ score} + b$$

5 – Evaluation du modèle

Pseudo-R carré:

R-carré en régression linéaire → proportion de variance expliquée

➤ Pseudo coefficient de détermination [0,1] basé sur la vraisemblance du modèle

Modèle nulle → absence de variable indépendante L0

R-carré Cox/Snell $R_{CS}^2 = 1 - \left(\frac{L_0}{L_1} \right)^{2/n}$

R-carré Nahelkerkes $R_N^2 = \frac{1 - \left(\frac{L_0}{L_1} \right)^{2/n}}{1 - L_0^{2/n}}$

5 – Evaluation du modèle

AIC Critère d'information Akaike: $AIC = 2k - 2 \ln(L)$

- Méthode de sélection de modèle, similaire au R-carré en régression linéaire
- Pénalisation de la complexité des modèles

BIC Critère d'information bayésien: $BIC = k \ln(n) - 2 \ln(\hat{L})$.

- Méthode pénalisant plus fortement la complexité des modèles
- Probabilité de sélection du modèle le plus proche de la réalité augmente en fonction de la taille de l'échantillon
- Sélection de modèle trop simpliste si les échantillons sont trop petits et moins représentatifs

5 – Evaluation du modèle

- La performance d'un modèle est habituellement mesurée par l'indice de Gini du score

Attention à ne pas faire d'interprétation absolue de l'indice de Gini, celui-ci dépend de la base de données sur laquelle il est calculé.

Autres indicateurs de performance d'un modèle de classification :

- Matrice de confusion (précision, rappel ...)

-Courbe ROC

-Courbe Lift

... etc

➔ Le choix de l'indicateur dépend du besoin auquel répond le score.

(par exemple pour les scores de fraude on a besoin de connaître la « précision » du modèle)

Références :

Bibliographie (incomplète, subjective)

- Gerard Scallan - Scoreplus (www.scoreplus.com)
- Ricco Rakotomalala (*site web université Lyon 2*)
- Hosmer, Lemeshow, Sturdivant (*Applied Logistic Regression*)
- Naeem Siddiqi (*Credit Risk Scorecards*)
- AI access (www.aiaccess.net)
- Stéphane Tufféry (*Data Mining et statistique décisionnelle, data.mining.free.fr*)
- Olivier Decourt (www.od-datamining.com)
- Damien Jacomy (damien.jacomy@gmail.com)
- Gilbert Saporta (*Probabilités, analyses des données et statistiques*)
- Société Française de Statistique (*Modèles statistiques pour données qualitatives*)



BNP PARIBAS
PERSONAL FINANCE

