Optimisation du process prédictif de sélection médicale en prévoyance individuelle

Application au télémarketing

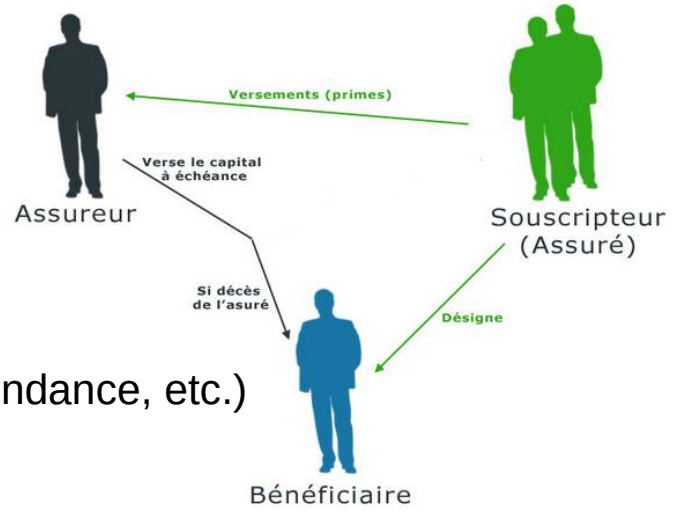
07 novembre 2016

Soutenance de mémoire d'actuariat

Jury Institut des Actuaires: Michaël CHOUKROUN & Carole MENDY
Correspondant ENSAE: Pierre PICARD
Maître de stage AXA: Sami FAYE-CHELLALI

Eve TITON

Introduction (1/3)



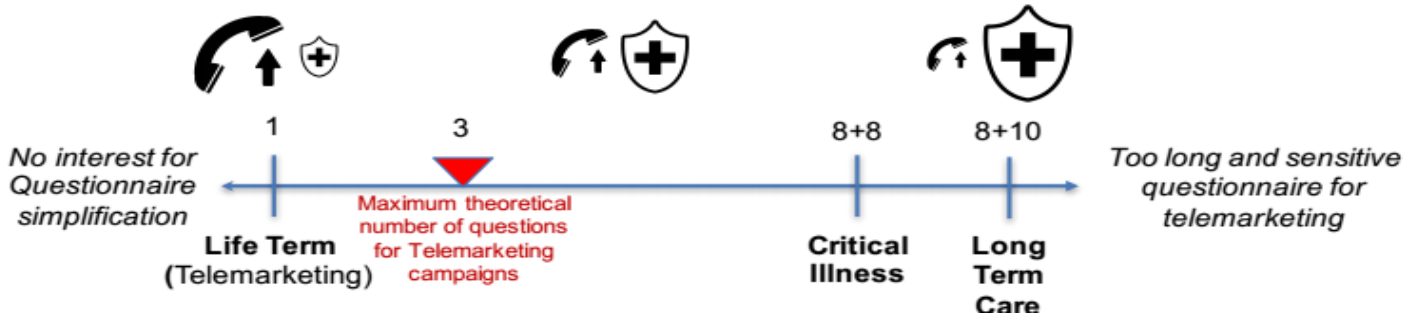
- Assurance prévoyance individuelle
  - Un contrat entre 2 parties: l'assuré et l'assureur.
  - Risques couverts: accidents de la vie (**décès**, dépendance, etc.)

- Importance de la sélection médicale
  - Le rôle de l'assurance est de protéger les clients des **aléas.**
    - Vérifier la présence d'aléa (risque modéré)
    - Avoir un portefeuille de contrats homogène (mutualisation des risques)
  - Le phénomène d'anti-sélection met en danger ces deux critères
  - La sélection médicale permet d'évaluer le risque réel des prospects.

- **Problème:** le processus de sélection médicale est très long, intrusif
  - Ventes manquées
  - Pas de diversité des canaux de distribution (difficile en télémarketing)

Introduction (2/3)

- Télémarketing
  - Vente par téléphone
  - Idéalement, on ne devrait poser aucune question (médicale ou autre) par téléphone.



- **Objectif**: trouver un moyen de réduire drastiquement le nombre de questions sans affecter l'évaluation du risque réel

- Les modèles d'apprentissage automatique, en permettant la reconnaissance de liens non-linéaires entre les variables, permettraient de prédire le risque des individus en réduisant le nombre de questions qui leur sont posées.

Introduction (3/3)

- **Avantages d'un modèle prédictif pour la sélection médicale à la souscription**
  - Réduction de la longueur et de la pénibilité de la souscription → amélioration de l'expérience client
  - Moins de travail pour les souscripteurs qui peuvent alors passer plus de temps sur les cas délicats
  - Augmentation des ventes (et du taux de conversion en télémarketing)
  - Possibilité de *cross-selling*

- **Etat des lieux – travaux sur ce sujet**

Sommaire

- **La sélection médicale en prévoyance individuelle chez AXA-MPS**
  - Deux produits de prévoyance individuelle chez AXA MPS
  - Différents process de sélection médicale à la souscription
  - Données
- **Prédiction de la décision de souscription médicale**
  - Quels modèles tester?
  - Présentation du principe des modèles retenus
  - Comparaison des modèles
- **Amélioration du questionnaire médical**
  - Questionnaire global – réduction du nombre de variables
  - Construction d'un meilleur questionnaire - personnalisé
- **Analyse des résultats**
  - Impacts potentiels sur les prix
  - Impacts potentiels sur les ventes
  - Bénéfices

La sélection médicale en prévoyance individuelle chez AXA-MPS

- Deux produits de prévoyance individuelle chez AXA MPS
- Différents process de sélection médicale à la souscription
- Données

Deux produits de prévoyance individuelle d'AXA-MPS

	Vita Sicura Plus	Pronto Vita
Accès	Vente directe	Uniquement par téléphone
Somme assurée	Au choix du client (sous réserve de condition de risque)	2 montants au choix (40000€ ou 60000€)
Souscription médicale	Longue	Très simplifiée (voire quasi-inexistante)
Garanties	Similaires (clauses d'exclusions différentes)	

Deux produits aux garanties similaires, seul le déroulement de la sélection médicale diffère significativement.

Zoom sur les types de sélection médicale à la souscription
- pratiques actuelles

	Sélection médicale automatisée AURA	Sélection médicale en télémarketing
Process	• Questionnaire informatique • Arbre de décision statique élaboré par des experts • Examens médicaux demandés pour moins de 20% des prospects	• Population sélectionnée: liste des clients de la banque MPS, filtrée sur certains critères d'âge, de profession, etc. • **Ces filtres ne prennent pas en compte le risque de mortalité** • **Information médicale: déclaration de bonne santé générale.**
Décision	Accepté avec prime standard (1) → Accepté Accepté avec surprime (2) Référé à un souscripteur (3) → Refusé Refusé (4)	
Avantages	Bonne précision dans l'évaluation du risque Process standardisé Moins long que la sélection médicale traditionnelle	Pas d'entrave à la vente télémarketing Process court
Inconvénients	Intrusif Reste long pour les cas (3) et (4) Impossible à mettre en œuvre dans le cadre du télémarketing	Pas de sélection médicale précise Présuppose que la population ciblée d'un point de vue commercial présente un bon risque

Prédiction à l'aide de 16 questions uniquement

Augmentation du nombre de questions – nombre max 4

Données utilisées: questionnaire AURA

- Cas soumis à AURA entre mars 2015 et mars 2016 : 45437 lignes.
- Informations en entrée
  - 8 informations générales
    - Informations physiologiques: sexe, âge, poids, taille (BMI)
    - Statut fumeur/non-fumeur
    - Emploi
    - Pratique sportive
    - Vente déjà refusée par le passé
  - 8 informations médicales
    - Traitements médicaux récents et passés
    - Antécédents familiaux

- Sortie obtenue: 4 catégories de risques

Décision AURA	Répartition	Cible pb telemarketing
1 – Accepté avec prime standard	86,27%	86,27%
2 – Accepté avec surprime	1,85%	13,73%
3 – Référé	11,64%	
4 – Refusé	0,24%	

Prédiction de la décision du souscripteur

Idée générale: prédire la décision AURA grâce à un ***algorithme d'apprentissage automatique*** *bien choisi.*

- Quels modèles tester?
- Présentation du principe des modèles retenus
- Comparaison des modèles

Quel(s) modèle(s) choisir?

Problème:

- Classification supervisée
- Variable cible:
  - Variable multi-classes non ordonnées mono-label
  - Classes déséquilibrées
  - Complexe
- Variables explicatives corrélées entre elles

Critères d'un algorithme adapté au problème

Prédiction: trouver un bon compromis entre l'erreur d'apprentissage et l'erreur de généralisation

Modèle non-linéaire

Rééquilibrer les classes

Simple à comprendre, à interpréter et à mettre en place

Modèles à priviliégier

- ✓ Modèles d'*asembling*
  - *Bagging* (forêts aléatoires)
  - *Boosting* (*adaptative boosting*)
- ✓ Modèles de base simples (Arbres de décision)
- ✓ Bien paramétrés

Qu'est-ce qu'une forêt aléatoire?

- Modèle d'*asembling* par *bagging*: combinaison de classifieurs simples pour obtenir un classifieur plus performant.

- Algorithme: supposons que l'on veut prédire une variable binaire *y* qui vaut 0 ou 1.

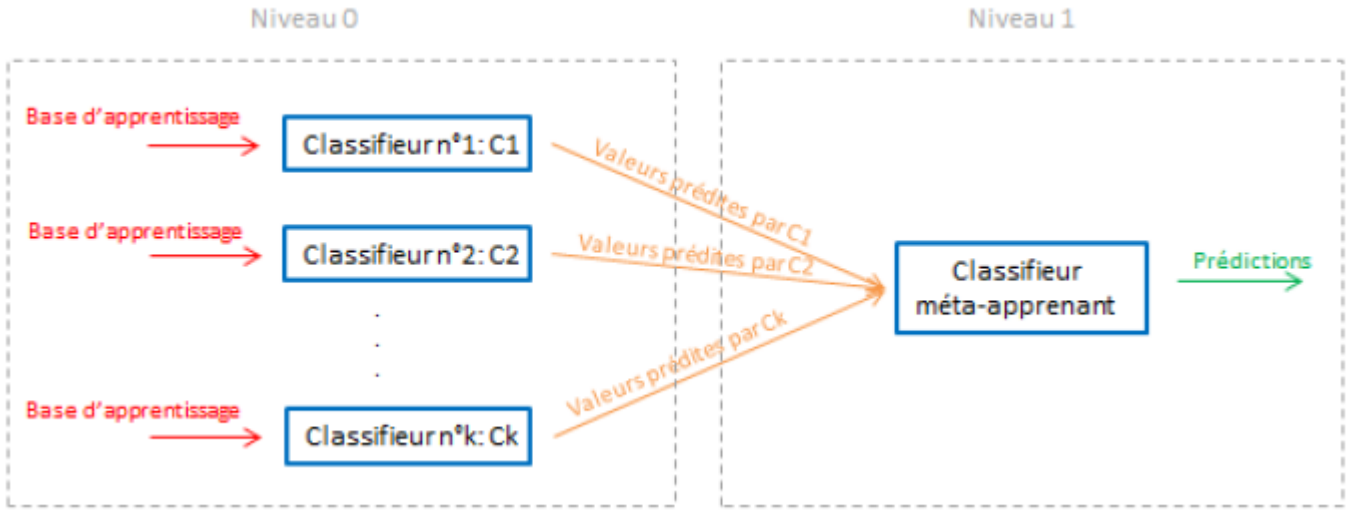
On veut prédire y_0, la décision associée à x_0.
On a un échantillon d'apprentissage $\mathbf{z} = ((x_1, y_1), \ldots, (x_n, y_n))$
Pour b allant de 1 à **B** faire:
    Tirer un échantillon aléatoire z^*.
    Estimer un arbre sur cet échantillon $\mathbf{z}^*$ avec randomisation des variables: la recherche de chaque nœud optimal est précédé d'un tirage aléatoire d'un sous-ensemble de **m** prédicteurs
Fin
Calculer l'estimation moyenne $\varphi_B(x_0) = \frac{1}{B}\sum_{b=1}^{B} \varphi_{z^*}(x_0)$;
Si cet estimation est plus grande que **α**, alors la décision finale est 1; sinon, c'est 0.

Permet de traiter le fait que les classes sont déséquilibrées

→ Permet de décorréler les arbres de décision générés
→ Paramètres
→ Bagging

Le *boosting* adaptatif et le *stacking*

- *Adaptative Boosting*
  - Construction de la famille de classifieurs faibles de façon différente par rapport au *bagging* chaque classifieur est une version adaptative du précèdent, en donnant plus de poids aux observations mal prédites.
  - Puis, on agrège ces classifieurs faibles par une méthode de vote.
  - Cet algorithme cible donc les observations les plus difficiles à prédire, tout en évitant l'overfitting (grâce a l'agrégation de tous les modèles).
- *Stacking*



Evaluation des modèles: mesure de risque adaptée

- ***Matrice de confusion*** générale associée à un problème de prédiction binaire, que l'on généralise à un problème de classification multi-classes.

MATRICE DE CONFUSION		Decision réelle	
		0	1
Décision prédite	0	Vrais positifs	Faux positifs
	1	Faux négatifs	Vrais négatifs

→

Matrice de confusion		Decision réelle			
		1	2	3	4
Décision prédite	1	# Individus 1 classés en 1	# Individus 2 classés en 1	# Individus 3 classés en 1	# Individus 4 classés en 1
	2	# Individus 1 classés en 2	# Individus 2 classés en 2	# Individus 3 classés en 2	# Individus 4 classés en 2
	3	# Individus 1 classés en 3	# Individus 2 classés en 3	# Individus 3 classés en 3	# Individus 4 classés en 3
	4	# Individus 1 classés en 4	# Individus 2 classés en 4	# Individus 3 classés en 4	# Individus 4 classés en 4

- ***Pondération des erreurs***: reflète le risque additionnel du portefeuille.

COEFFICIENTS du *MORTALITY SLIPPAGE*			Decision réelle			
			1	2	3	4
			80%	120%	250%	750%
Décision prédite	1	80%	1,00	1,50	3,13	9,38
	2	120%		1,00	2,08	6,25
	3	—			1,00	1,00
	4	750%				1,00

Méthodologie

1. Calculer la proportion d'individus correctement prédits et mal prédits, au sein de l'ensemble des prédits acceptés.
2. Multiplier ces proportions par les coefficients de *mortality slippage* associés; puis les sommer.
3. On obtient le *mortality slippage* global.

Comparaison des modèles

		Forêt aléatoire	Boosting adaptatif	Stacking
Résultats chiffrés	Mortality slippage	+++	+	+++
	Perte de business (faux négatifs)	++	+++	--
Propriétés théoriques	Simplicité	+++	++	+
	Action sur les paramètres	+++	++	++
	Consistance	+++	+++	+++
	Traitement des classes les plus risquées	++	+++	++

+++ très bon ; ++ bon; + satisfaisant; - peu satisfaisant; -- mauvais

→ Les forêts aléatoires semblent proposer le meilleur équilibre entre la performance chiffrée sur nos données et la simplicité du modèle.

Mortality slippage de 100.5%

Taux de faux négatifs de 0.75%

Amélioration du questionnaire médical

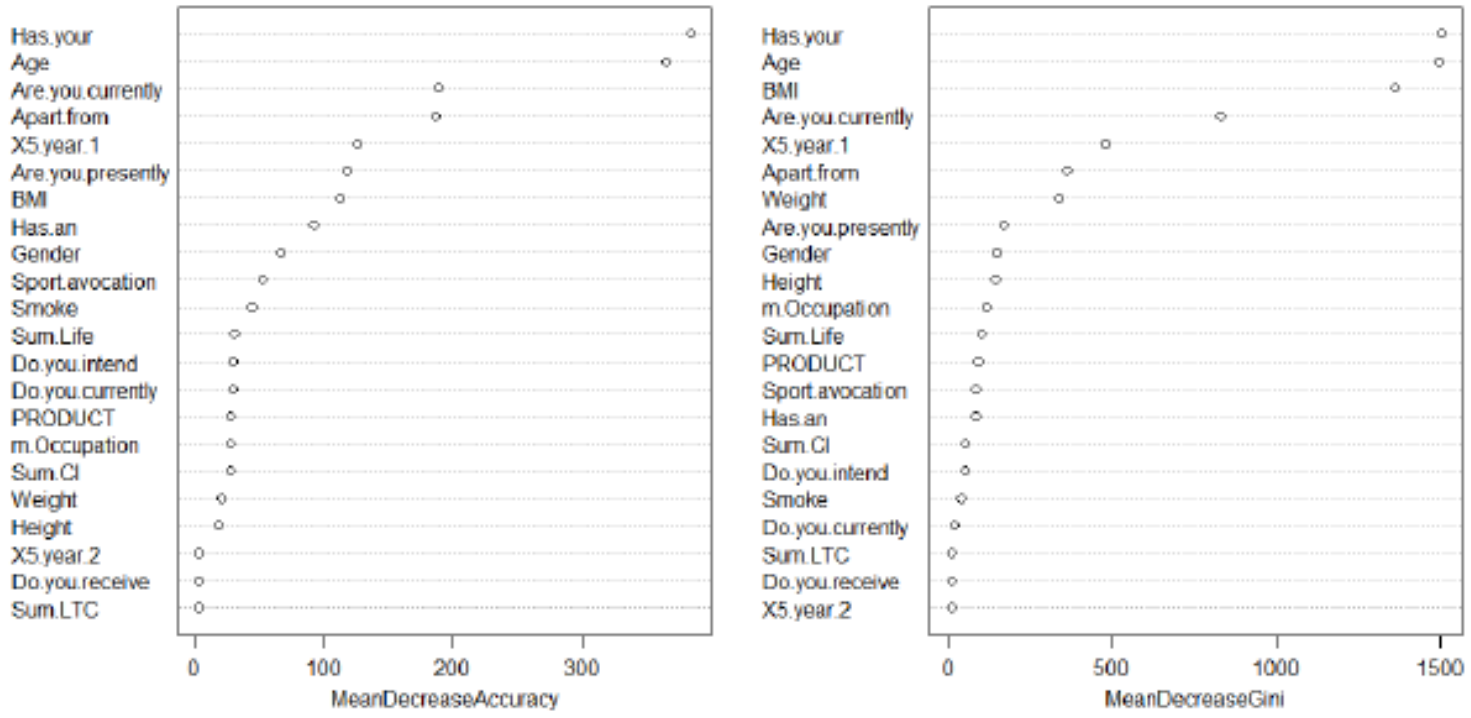
Idée générale: ***réduire le nombre de questions*** *en construisant un questionnaire contenant* ***les questions les plus informatives possible.***

Contexte: télémarketing

- Questionnaire global
- Questionnaire personnalisé

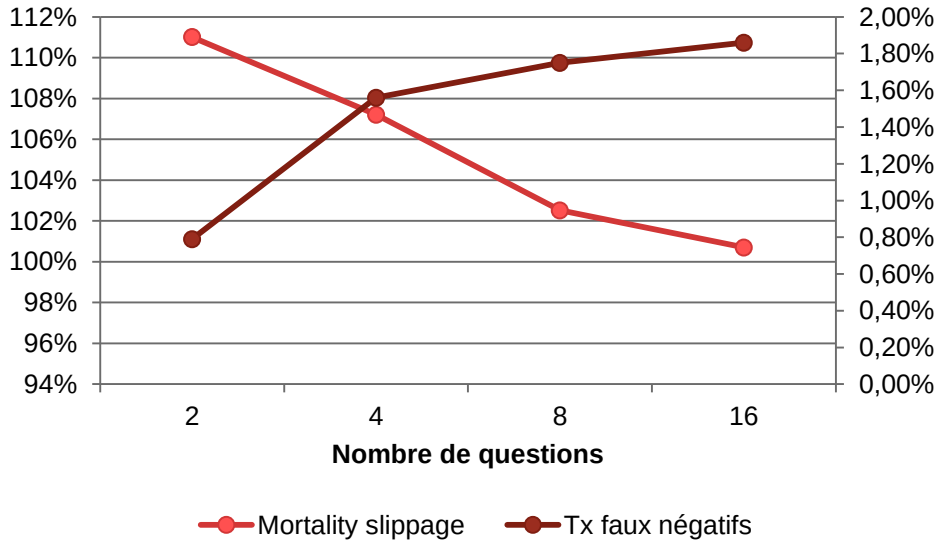
Première approche: conservation des k questions les plus informatives

- Contexte: télémarketing (cible binaire)
- Sortie de la forêt aléatoire
- Deux critères d'importance des variables: *Mean Decrease Accuracy, Mean Decrease Gini*
- Prise en compte des deux critères: utilisation d'une somme pondérée de la statistique de rang



Résultats

Nb questions	Questions posées	Mortality slippage	Tx faux négatifs
2 questions	Age + antécédents familiaux	111,01%	0,79%
4 questions	+ IMC (taille, poids)	107,21%	1,56%
8 questions	+ traitements médicaux pour certaines maladies graves + traitements médicaux prescrits pour une durée supérieure a 5 jours + chirurgie au cours des 5 dernières années + statut d'invalidite/incapacité de travail.	102,52%	1,75%
16 questions	'Toutes'	100,7%	1,86%



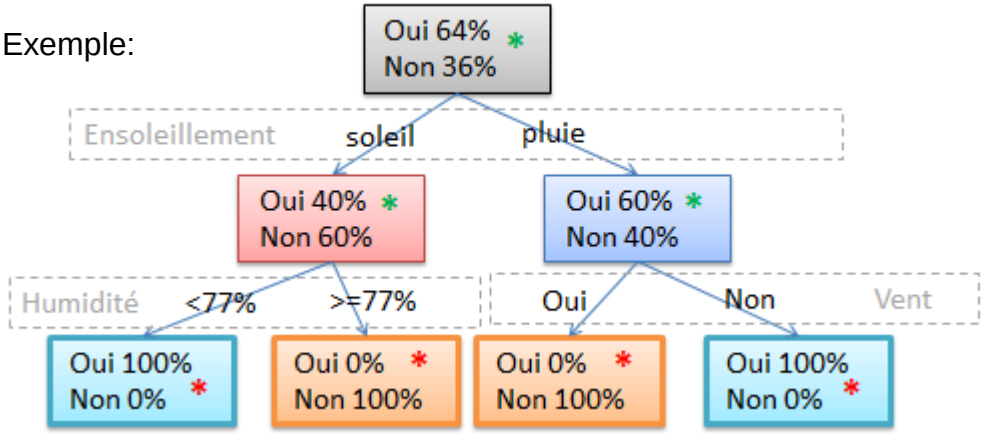
- Saut entre 2 et 4 questions
- Pente plus faible à partir de 4 questions
- 111,01% de MS: élevé.

Deuxième approche: questionnaire personnalisé

1ère étape: construction récursive d'un arbre « pur »

```
Procédure Construction_arbre(nœud X)
        * Si (tous les points de X appartiennent à la même classe)
        {créer une feuille portant le nom de cette classe}
        * Sinon{
        Choisir la meilleure variable pour créer un nœud**
        Le test associé à ce nœud sépare X en deux parties Xg et Xd
        Construction_arbre(Xd)
        Construction_arbre(Xg)}
        Fin si
Fin procédure
```

Exemple:



** Pour choisir la meilleure variable: critère de Gini ou information mutuelle

2ème étape: élagage de l'arbre

Adaptation du problème télémarketing

Rappel: le but est de trouver, à chaque noeud, la meilleure question à poser, sachant les réponses précédentes.

On veut savoir, pour chaque modalité (réponse) de chaque variable (question), quelle est la meilleure question à poser juste après. Chaque réponse doit donc correspondre à une variable.

La réponse à une question peut donner plus d'information que celle qui est utilisée à un noeud

Un arbre de décision classique peut réutiliser la même question plusieurs fois

Un simple arbre de décision n'est pas assez consistant, le choix des splits risque d'être trop sensible aux données

Toutes les variables sont binarisées au préalable.

Procédure Construction_arbre_questions(nœud X, Classifieur)

Si (tous les points de X appartiennent à la même classe) {créer une feuille portant le nom de cette classe}

Sinon{
1. Faire tourner Classifieur sur la population X.
2. Choisir la meilleure variable pour créer un nœud: pour chaque variable non binarisée, additionner la VarImp de toutes les variables binarisées associées.
3. Le test associé à ce nœud sépare X en autant k parties (k nombre de modalité dans la variable non binarisée)
4. Supprimer toutes les variables binarisées associées à la variable de test.

Construction_arbre_questions(X1, Classifieur)
.
.
.
Construction_arbre_questions(Xk, Classifieur)}

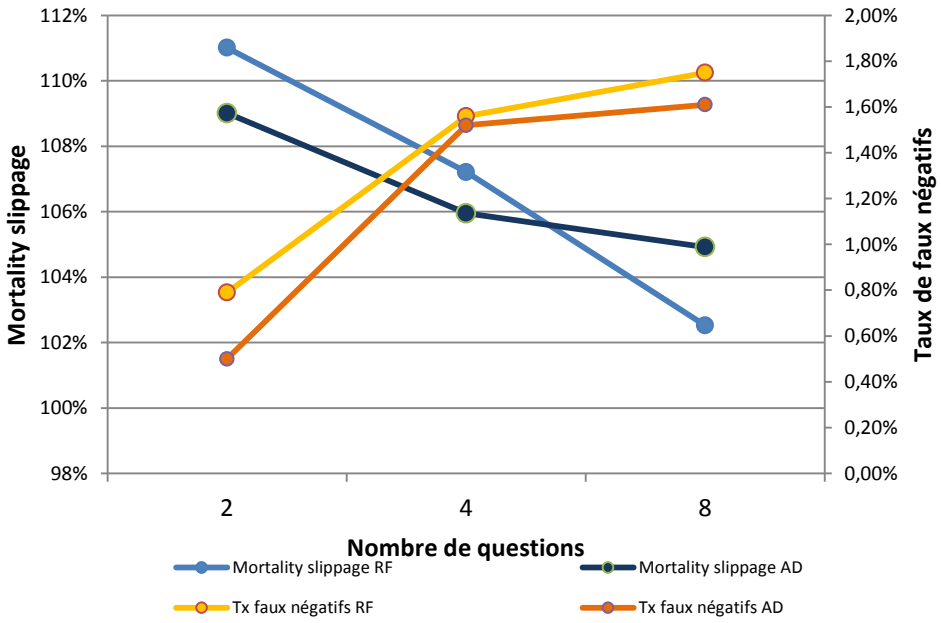
Fin si
Fin procédure

En sortie, on récupère un arbre de questions qui fournit un questionnaire qui s'adapte au fur et à mesure des réponses fournies.

La phase d'élagage n'est pas nécessaire ici, car on coupe l'arbre en fonction du nombre de questions posées.

Résultats

Nb questions	Mortality slippage	Tx faux négatifs
2 questions	109,01%	0,50%
4 questions	105,95%	1,52%
8 questions	104,92%	1,61%



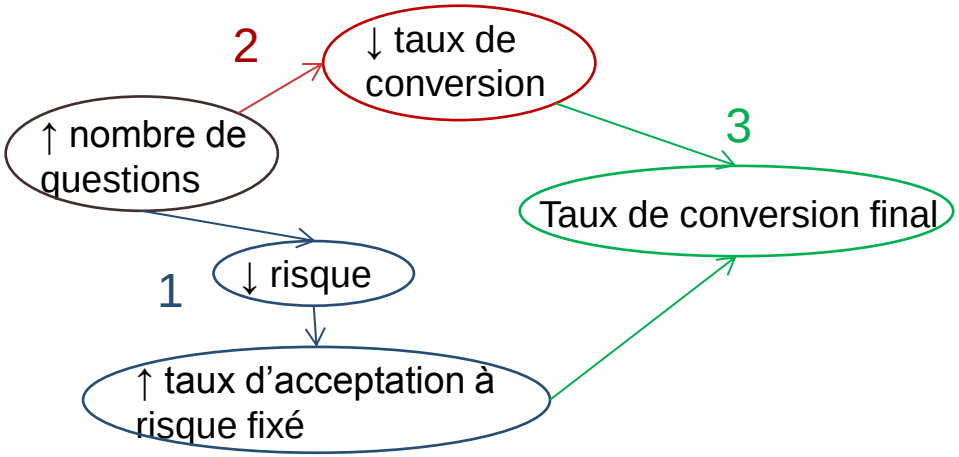
- Sur les questionnaires à 2 et 4 questions, l'arbre de questions construit sur le modèle d'un arbre de décision donne de meilleurs résultats que les forêts aléatoires.
- Cependant, le *mortality slippage* augmente très rapidement lorsque le nombre de questions diminue.

Hypothèse:
ajouter des variables externes, qui ne correspondent pas à des questions à poser, mais à de l'information supplémentaire disponible, pourrait améliorer les résultats.

Impacts potentiels sur le *business*

Comprendre l'impact d'un changement de process de sélection médicale à la souscription dans le cadre télémarketing d'un point de vue des prix, des ventes et de la marge effectuée.

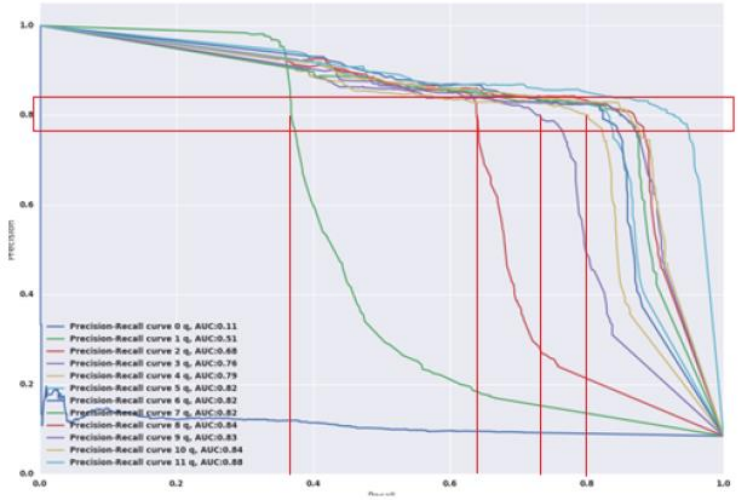
Impact potentiel sur les ventes



3. Taux de conversion final

Nb questions	Tx acceptation	Tx conversion	Tx final
0	3%	7,0%	0,21%
1	38%	3,5%	1,33%
2	63%	1,8%	1,10%
3	75%	0,9%	0,66%
4	79%	0,4%	0,35%
5	83%	0,2%	0,18%
6	83%	0,1%	0,09%
7	85%	0,1%	0,05%
8	85%	0,0%	0,02%
9	86%	0,0%	0,01%
10	86%	0,0%	0,01%
11	95%	0,0%	0,00%

2. Taux d'acceptation à risque fixé



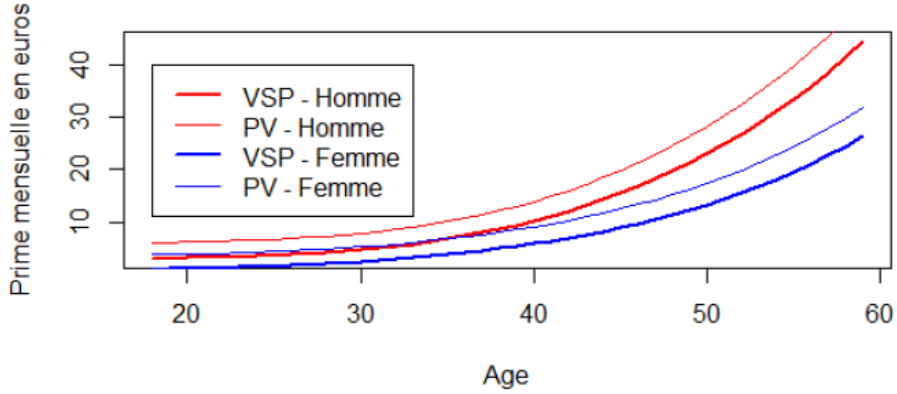
2. Hypothèse sur le taux de conversion

Décroissance exponentielle avec le nombre de questions

$$Taux\ de\ conversion = 7\% \times y\%^{Nb\ questions}$$

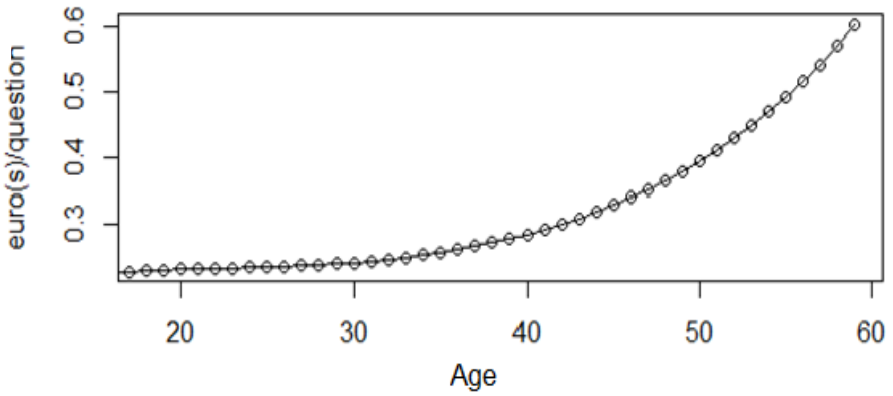
Impact potentiel sur les prix

- Calcul des primes pures dues selon le type de sélection médicale: 0 questions vs. 16 questions.
- Hypothèse: évolution linéaire du prix avec le nombre de questions.
- Calcul des droites Prix=a*Nb questions+b en fonction de l'âge



$$\sum_{t=m}^{m+T} \frac{PP}{(1+r)^{t-m}} \prod_{k=m}^{t} (1-(q_k+IL)) = \sum_{t=m}^{m+T} \frac{S(1+L)(q_t+IL)}{(1+r)^{t-m}} \prod_{k=m}^{t-1} (1-(q_k+IL))$$

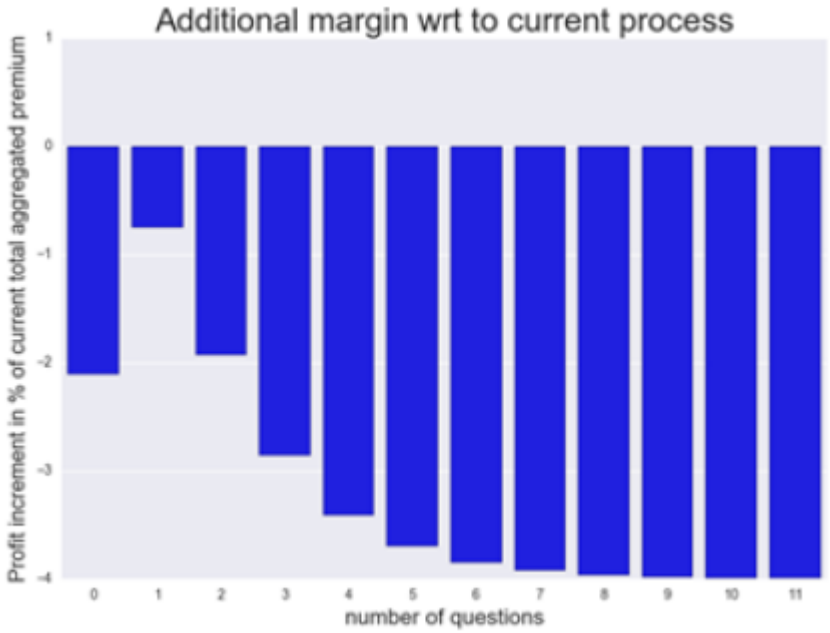
Exemple: Pour un assuré de 50 ans, on augmente la prime pure de 40 centimes par question et par mois, soit environ 5€ par an pour chaque question additionnelle.



Bénéfices effectués

$$Benefice = [M + (1 - MS/MS_0)] \times C \times X \times P$$

- ✓ M: taux de marge
- ✓ 1- MS/MS0: gain implicite sur la prime que l'on fait sachant que la population sélectionnée est moins risquée
- ✓ C×X: taux de conversion final
- ✓ P: prime pure



- ➢ Poser plus de questions permet de cibler des clients moins risqués, ce qui génère une réduction implicite de prime.
- ➢ Cependant, cela réduit également le taux de conversion et la taille de la population ciblée.

Conclusion

- Résultats obtenus satisfaisants, mais **perfectibles**
- **Axes de développement**
  - → Enrichir la base de données
    - Pression sur la banque MPS pour extraire les données supplémentaires nécessaires
    - Permettrait d'améliorer le taux d'acceptation à risque fixé et donc d'obtenir une marge positive
  - → Adapter le modèle au processus télémarketing en accord avec les équipes produit et marketing
    - Par exemple, pondérer les questions de l'arbre en fonction de leur caractère intrusif
    - Vérifier les hypothèses effectuées sur le taux de conversion
  - → Segmenter en amont suivant les variables dont on dispose

- Utilisation d'un tel modèle dans d'autres contextes: meilleure segmentation des risques, proposition d'une procédure de sélection médicale plus courte, même en vente directe.

- **Problématique de l'implémentation d'un modèle prédictif dans un cadre où l'on ne peut pas quantifier les sinistres rapidement.**

MERCI DE VOTRE ATTENTION

Questions/Réponses