TECHNIQUES DE SCORING

Jean-Philippe KIENNER

INTRODUCTION

PRINCIPES DU SCORING

CONSTRUCTION DE LA BASE D'ÉTUDE

MODÉLISATION

EXPLOITATION DU SCORE

CONCLUSION

PRÉSENTATION DE L'ENSEIGNEMENT

- L'objectif est de présenter l'une des méthodologies phares dans les études statistiques : le ***scoring***

- Le module est centré autour d'un projet « fil rouge » ayant pour objectif de lancer une ***campagne marketing de fidélisation*** auprès des clients d'un opérateur de téléphonie mobile

- Cela permettra
  - De présenter ce qui amène à construire un score
  - De tester la performance de plusieurs méthodologies de ciblages et de voir les résultats s'améliorer au fur et à mesure de l'utilisation de méthodes statistiques avancées
  - De dérouler la construction du score de A à Z en insistant sur les aspects opérationnels, les problèmes rencontrés en entreprise, la manière de présenter le projet à des non statisticiens, …

PRÉSENTATION DE L'ENSEIGNEMENT

- La présentation est structurée de la manière suivante (1/2) :

  - Principes du scoring
    - Intérêt d'un score : d'un ciblage simple à un ciblage scoré
    - Définition d'un score
    - Etapes d'un projet de scoring

  - Construction de la base d'étude
    - Identification de la population éligible
    - Définition de l'évènement à étudier
    - Détermination de la période d'étude
    - Construction des variables explicatives
    - Constitution des échantillons d'apprentissage et de validation(s)
    - Optimisation des variables explicatives

  - Modélisation
    - Construction des modèles
    - Évaluation des modèles
    - Interprétation des modèles

PRÉSENTATION DE L'ENSEIGNEMENT

- La présentation est structurée de la manière suivante (2/2) :
  - Exploitation du score
    - Application du score
    - Evaluation opérationnelle du score
    - Industrialisation du score
    - Suivi de la performance du score
  - Prise de recul
    - Facteurs clés de réussite ou d'échec d'un score
    - Ce que permet et ce que ne permet pas un score

PANORAMA DES ÉTUDES RÉALISÉES

ACTIVITÉS RÉCURRENTES

ÉTUDES CONNAISSANCE CLIENT

MESURER

ANALYSER

PRÉDIRE

Comptages

Tableaux de bord

Études

Segmentations

Scores

Prévisions

BIBLIOGRAPHIE

- L'ensemble de ce document et des informations fournies en séance n'engagent que moi, et regroupent l'essentiel de mes connaissances en scoring acquises tout au long de mon expérience professionnelle

- Néanmoins cette expérience s'est forgée et enrichie sur la base de lectures et d'échanges avec de nombreux autres experts, dont je me permets de citer ici les principaux contributeurs
  - Olivier DECOURT : « http://od-datamining.com »
  - Stéphane TUFFERY : « Data Mining et statistique décisionnelle » et « Étude de cas en statistique décisionnelle »
  - Philippe BESSE : « http://www.math.univ-toulouse.fr/~besse/ »
  - Ricco RAKOTOMALALA : « http://ricco-rakotomalala.blogspot.com/ »

INTRODUCTION

PRINCIPES DU SCORING

CONSTRUCTION DE LA BASE D'ÉTUDE

MODÉLISATION

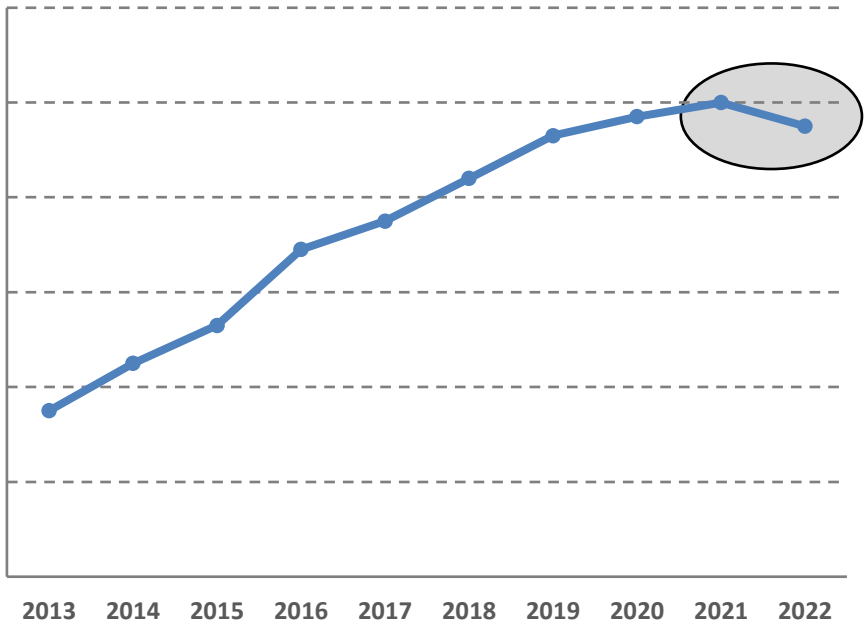
EXPLOITATION DU SCORE

CONCLUSION

EXEMPLE INTRODUCTIF

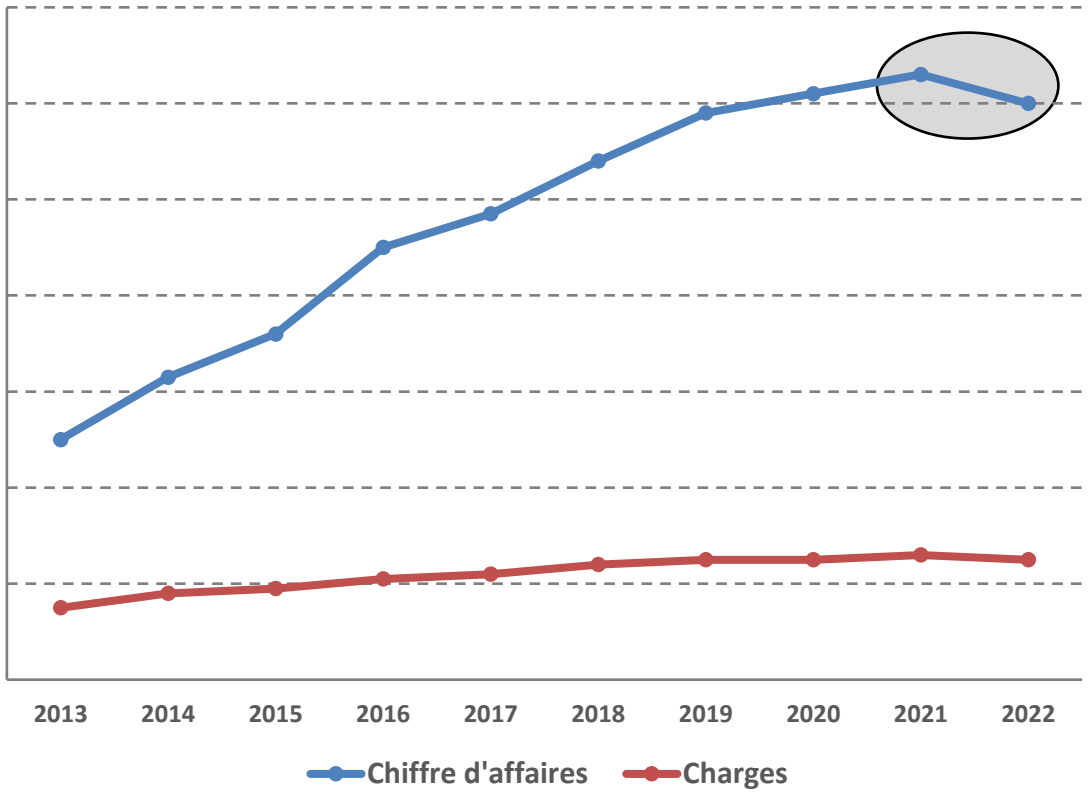
- Cet exemple introductif, pris sur une problématique d'un opérateur de téléphonie mobile, a pour objectif de montrer la démarche qui peut conduire à l'intérêt de construire un score

- Pour la première fois depuis la création de cet opérateur, l'année vient de se clôturer par une baisse du résultat net de l'entreprise



EXEMPLE INTRODUCTIF

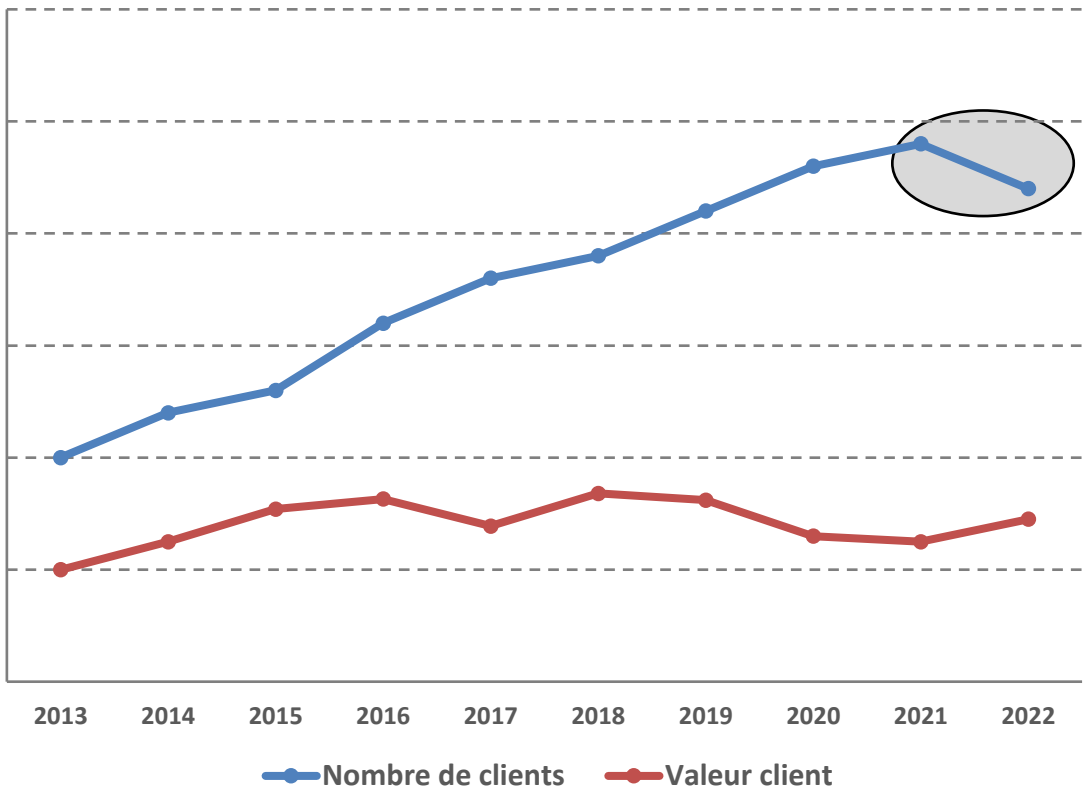
- Résultat = Chiffre d'affaires - Charges



La baisse du résultat vient principalement d'une baisse du chiffre d'affaires

EXEMPLE INTRODUCTIF

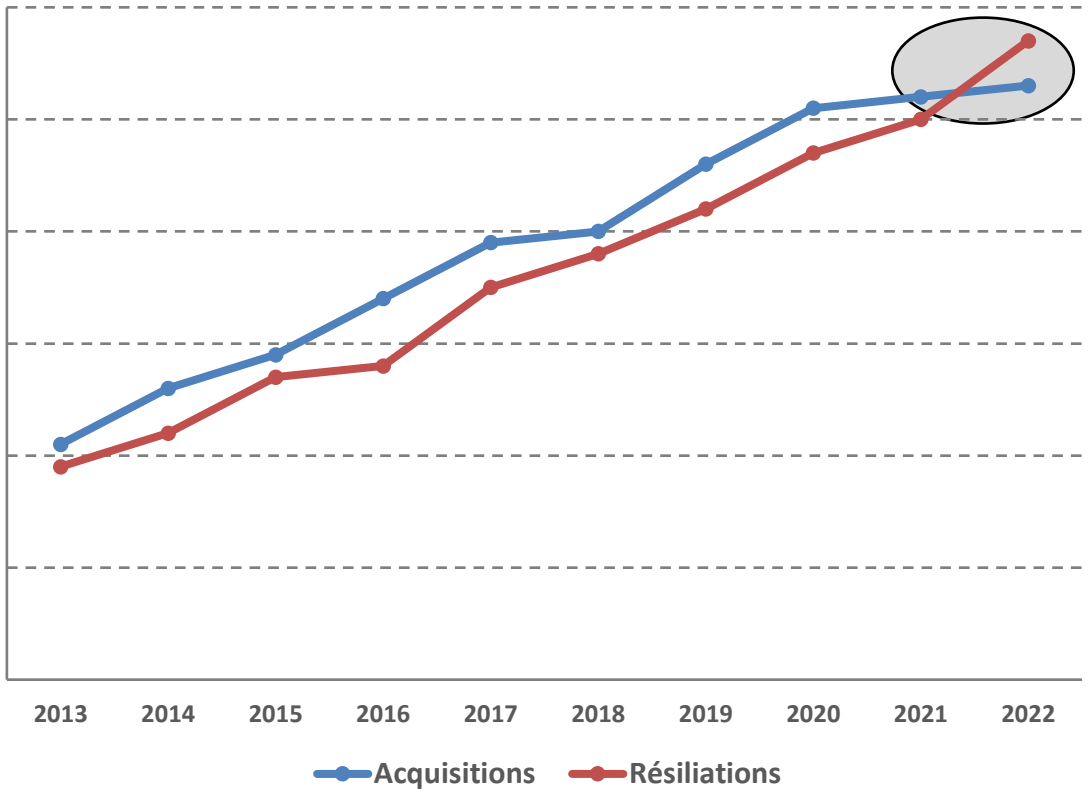
- Chiffre d'affaires = nombre de clients * valeur client



La baisse du CA vient principalement d'une baisse du nombre de clients

EXEMPLE INTRODUCTIF

- Nombre de clients = Fidèles + Acquisitions - Résiliations



La baisse du nombre de clients vient principalement d'une hausse des résiliations

EXEMPLE INTRODUCTIF

- Le ***tableau de bord*** de pilotage de l'activité montre clairement que c'est la hausse des résiliations qui est une des causes principales de la diminution du résultat de l'entreprise

- Ce phénomène ne peut évidemment continuer, une stratégie de ***fidélisation / rétention des clients*** est donc un enjeu majeur pour la nouvelle année

- Cette stratégie va prendre la forme d'une campagne marketing afin de proposer à 2 000 clients de se réengager en contrepartie d'une offre promotionnelle
  - Quel offre / service proposer ?
  - À quel tarif / remise ?
  - Par quel canal de communication ?
  - Auprès de quels clients ?

EXEMPLE INTRODUCTIF

- Le Data Scientist va contribuer principalement à identifier les clients à contacter / cibler, en collaboration étroite avec l'équipe marketing

- Plusieurs méthodes peuvent être utilisées, la première étant de réaliser un ***ciblage métier*** en faisant appel à la connaissance du produit et des comportements clients et marchés

- Exemple de ciblage
  - Clients âgés de 25 à 40 ans
  - Ayant au moins 1 an d'ancienneté
  - Détenteurs d'une offre de plus de 6H
  - Appartenant au segment A

EXEMPLE INTRODUCTIF

- Ce premier ciblage a l'avantage d'être rapide et peut être efficace lorsque l'on maîtrise très bien l'environnement et les comportements clients

- Il peut néanmoins être amélioré en analysant le profil des clients qui ont résilié dans le passé

- Des clients actuels ayant un profil similaire à celui de clients ayant déjà résilié ont en effet un risque de résiliation probablement plus important

- Il s'agit d'un ***ciblage profilé***

EXEMPLE INTRODUCTIF

- L'analyse du profil des résiliés a permis d'identifier les critères de ciblage suivants
  - Clients non engagés
  - Aucun réengagement
  - Forfait de courte durée (≤ 2H)
  - Faible volume d'appels (≤ 3H)

- Cette cible donne des résultats convenables mais présente plusieurs inconvénients
  - Liés à la méthodologie de construction du ciblage
    - Difficultés à ajuster la volumétrie souhaitée
    - Choix des variables / des modalités
    - Pondération des variables / des modalités
  - Liés aux données
    - Pas ou peu de nettoyage des données
    - Pas ou peu de nouveaux indicateurs
    - Corrélation entre les critères
    - Hétérogénéité de la population

EXEMPLE INTRODUCTIF

- Pour synthétiser
  - Il n'existe pas un profil unique de clients à risque
  - Au contraire la base clients est constituée d'une multitude de profils plus ou moins « risqués », constitués chacun par la combinaison des indicateurs
  - C'est donc la prise en compte simultanée de l'ensemble de ces profils, pondérés en fonction du risque que chacun représente, qui doit permettre d'identifier les clients à contacter
  - Et ceci pourra se faire grâce à une note globale de risque de résiliation calculée pour chaque client

- Cette méthodologie consiste à estimer un ***modèle de score***

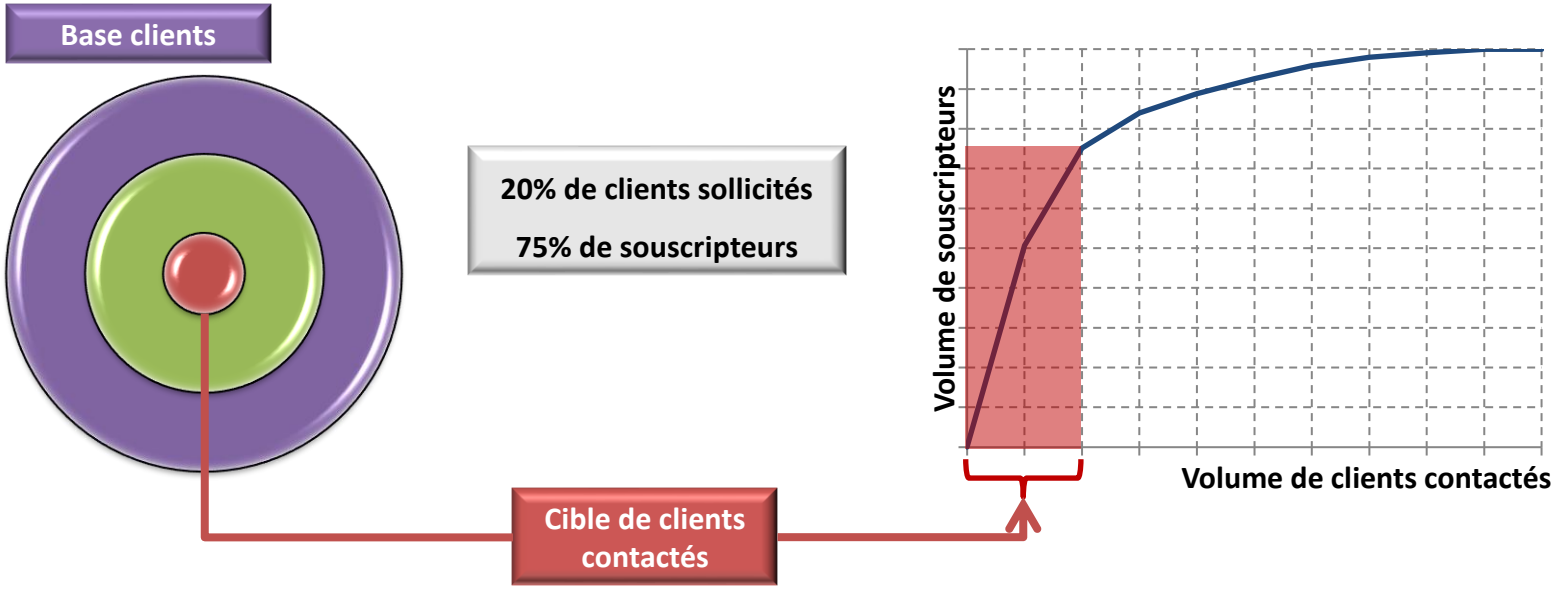
DÉFINITION ET OBJECTIF D'UN SCORE

- Un ***score*** permet d'attribuer à chaque individu une ***note*** estimant la ***probabilité de réalisation d'un évènement***

- De nombreuses applications se font dans l'univers marketing afin ***d'optimiser le ciblage*** d'une sollicitation commerciale
  - Les individus peuvent ainsi être hiérarchisés en fonction de leur chance / risque de réaliser cet évènement
  - La campagne marketing s'adressera alors en priorité aux individus les plus « appétants », c'est-à-dire ceux qui auront la meilleure note (et ne polluera pas ceux qui le sont moins)

DÉFINITION ET OBJECTIF D'UN SCORE

- L'utilisation d'un score permet donc d'obtenir
  - Des taux de retours plus importants
  - Avec un volume de sollicitations plus faible



DÉFINITION ET OBJECTIF D'UN SCORE

- De nombreux autres métiers / secteurs peuvent utiliser le scoring et tout type d'individu (au sens statistique) peut être scoré : personnes, entreprises, communes, points de vente, composants électroniques, équipes sportives, …
  - Appétence à un produit
  - Appétence à un canal de communication
  - Attrition
  - Octroi de crédit
  - Impayé
  - Fraude
  - Changement de segment
  - Survie
  - Survenue d'un accident
  - Recrutement
  - Dépôt de bilan
  - Victoire lors du prochain match
  - …

ÉTAPES D'UN PROJET DE SCORING

- Un projet de scoring va généralement se dérouler en 3 étapes
  - Construction de la base d'étude
  - Modélisation
  - Exploitation du score

Développement du score

Mise en production du score

INTRODUCTION

PRINCIPES DU SCORING

CONSTRUCTION DE LA BASE D'ÉTUDE

MODÉLISATION

EXPLOITATION DU SCORE

CONCLUSION

BASE D'ÉTUDE

- Un score correspond à un problème de modélisation supervisée, il s'agit donc de construire une variable à expliquer et un ensemble de variables explicatives
  - La construction de cette base d'étude est l'étape la plus longue d'un projet de scoring
  - En particulier pour un score marketing où cette construction est souvent plus complexe que dans d'autres domaines (ex : reconnaissance d'image)

- Cette étape est à réaliser très méticuleusement car tous les choix auront une influence directe sur la performance du modèle … et donc sur la rentabilité des campagnes marketing
  - En effet, contrairement à une étude descriptive classique qui se contente d'analyser un phénomène passé, le modèle de score a surtout pour objectif d'être appliqué dans le futur
  - C'est sur la base d'étude que le modèle statistique « apprendra » à différencier les individus réalisant l'évènement de ceux ne le réalisant pas
  - Le modèle construit sera ensuite appliqué pendant plusieurs mois / années sur de « nouvelles » bases de données

BASE D'ÉTUDE

- La ***base d'étude*** créée lors de la phase de construction du modèle doit obligatoirement ***reproduire les conditions réelles d'application du score***

- La construction de la base d'étude peut se décomposer en plusieurs étapes
  - Identification de la population éligible
  - Définition de l'évènement à étudier
  - Détermination de la période d'étude
  - Construction des variables explicatives
  - Constitution des échantillons d'apprentissage et de validation(s)
  - Optimisation des variables explicatives

BASE D'ÉTUDE

- Au final l'objectif sera de construire une base d'étude qui pourra prendre la forme suivante (à titre d'illustration) :

Population éligible	Évènement	Période d'étude	Variables explicatives	Échantillonnage

ID CLIENT	CIBLE	DATE ÉVÈNEMENT	DATE DE RÉFÉRENCE	ÂGE	SEXE	CSP	ÉCHANTILLON
A	0	.	2010_08	1	2	1	Apprentissage
B	1	2010_11	2010_08	3	1	4	Validation
C	0	.	2010_09	2	1	3	Apprentissage
D	0	.	2010_08	4	2	3	Apprentissage
E	1	2010_12	2010_09	3	1	4	Apprentissage
F	0	.	2010_09	2	2	2	Validation
G	1	2010_11	2010_08	4	2	3	Apprentissage

BASE D'ÉTUDE

- Population éligible
- Évènement à étudier
- Période d'étude
- Construction des variables explicatives
- Échantillons d'apprentissage et de validation(s)
- Optimisation des variables explicatives

POPULATION ÉLIGIBLE

- La ***population éligible*** est constituée de l'exhaustivité des individus pour lesquels on calculera un score

- Toute la base clients peut être éligible

- Dans certains cas il est nécessaire d'appliquer des filtres
  - Liés à des contraintes juridiques
    - Exclusion des mineurs, des interdits bancaires, …
  - Liés à des choix stratégiques (à partager avec les experts métiers)
    - Âge, Ancienneté, Segment, Détention de produits, …
  - Liés à l'évènement étudié
    - Intégrer uniquement les individus pour lesquels l'évènement peut se réaliser
    - Rendre comparable les populations étudiées (même périmètre de souscripteurs et de non-souscripteurs par exemple)

POPULATION ÉLIGIBLE

- Exemple : construction d'un score d'appétence à la souscription d'une carte de paiement

  - Une carte de paiement ne pouvant être souscrite qu'à partir de 16 ans, il ne faut pas inclure dans l'étude toute la base clients, les plus jeunes étant obligatoirement des non-souscripteurs

  - D'autres critères d'exclusion peuvent aussi être imaginés
    - Doit-on inclure les clients détenteurs d'une carte dans les 6 derniers mois ? (même s'ils l'ont résilié depuis)
    - Doit-on considérer un client qui a souscrit une carte et l'a résiliée 3 mois après comme ayant réalisé l'évènement ?
    - Doit-on inclure uniquement les clients déjà ciblés pour une campagne de souscription à la carte de paiement ? (méthode pertinente et efficace !)

- Attention, il ne s'agit pas ici d'utiliser des variables discriminantes, mais bien d'exclure les individus pour lesquels on ne souhaite même pas calculer le score

BASE D'ÉTUDE

- Population éligible
- Évènement à étudier
- Période d'étude
- Construction des variables explicatives
- Échantillons d'apprentissage et de validation(s)
- Optimisation des variables explicatives

ÉVÈNEMENT À ÉTUDIER

- Dans tout projet de scoring, un des premiers éléments fondamentaux à définir est ***l'évènement que l'on souhaite scorer***

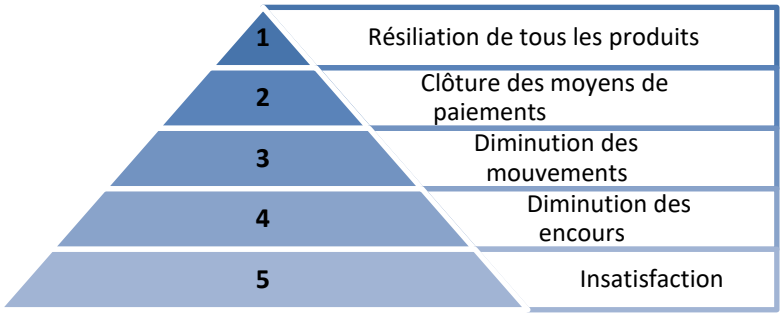
- Cette étape permet la construction de la ***variable à expliquer***

- En général la création d'un score repose sur la modélisation d'une variable à expliquer « Y » binaire qui permet de comparer (discriminer) 2 sous-populations
  - Les individus ayant réalisé l'évènement auront la valeur « 1 »
  - Les individus n'ayant pas réalisé l'évènement auront la valeur « 0 »

- Il faut donc définir de manière très précise les individus allant dans la catégorie « 0 » et ceux allant dans la catégorie « 1 »
  - Pour rappel, l'ensemble de ces individus constitue la population éligible

ÉVÈNEMENT À ÉTUDIER

- Exemple de l'attrition bancaire
  - De nombreuses questions se posent pour bien définir le départ du client
  - Le plus simple serait de considérer que son départ est caractérisé par la résiliation de tous ses produits
    - Mais ce processus de résiliation est souvent étalé sur plusieurs mois
    - Une variable à expliquer construite à partir de cet évènement va poser problème
      - Y = 1 : « le client résilie son dernier produit »
        - Pb : ces clients ont probablement déjà pris leur décision depuis longtemps
      - Y = 0 : « le client n'a pas résilié son dernier produit »
        - Pb: une partie de ces clients est déjà dans un processus de résiliation
    - Même si le modèle statistique peut donner de bons résultats, la campagne de rétention associée serait inefficace
  - Autres possibilités (non exhaustif) : 5 évènement différents donc 5 scores potentiellement très différents !



ÉVÈNEMENT À ÉTUDIER

Équilibrage

- Si l'évènement étudié a une répartition déséquilibrée, certaines méthodes statistiques peuvent avoir des difficultés à bien discriminer les sous-populations
  - Exemple : 0,5% de souscripteurs à un produit
  - Certains modèles statistiques se contenteront de prédire les clients en non-souscripteurs et obtiendront un taux de bien classés excellent (99,5% !)
  - Cela dépend aussi du contexte, des méthodes et de la volumétrie

- Il peut être utile dans ce cas d'équilibrer la base (ce n'est pas anodin, à ne pas appliquer de manière systématique !)
  - Under-sampling : échantillonner les individus ne réalisant pas l'évènement
    - Attention à la représentativité
    - Des algorithmes tels que TOMEK suppriment des individus ne réalisant pas l'évènement proches des individus réalisant l'évènement
  - Over-sampling : répliquer les individus réalisant l'évènement
    - Attention au sur-apprentissage
    - Des algorithmes tels que SMOTE simulent de nouveaux individus réalisant l'évènement proches des individus réels

ÉVÈNEMENT À ÉTUDIER

Stratification

- Il est fréquent que la population étudiée soit très hétérogène en étant constituée de sous-populations ayant des comportements très différents par rapport à l'évènement à scorer
  - Exemple : score d'attrition en téléphonie mobile mélangeant des clients engagés et libres d'engagement

- Laisser ces sous-populations dans un même modèle entrainerait un poids trop important de certaines variables au détriment d'autres indicateurs

- La solution consiste à séparer la population initiale et construire un score pour chacune des strates
  - Les strates de clients doivent être constituées à partir de variables simples et partagées avec les experts métiers : âge, segment, détention de produit, contrainte règlementaire …
  - Le nombre de strates ne doit pas être trop élevé

BASE D'ÉTUDE

- Population éligible
- Évènement à étudier
- Période d'étude
- Construction des variables explicatives
- Échantillons d'apprentissage et de validation(s)
- Optimisation des variables explicatives

PÉRIODE D'ÉTUDE

- Un score doit permettre de prédire un évènement futur grâce à des données passées

- Il est donc nécessaire de définir les ***périodes d'analyse***
  - La période au cours de laquelle on observe le comportement de l'individu (variables explicatives)
    - Date de référence connue lors de la phase de construction du score et lors de la phase d'application du score
  - La période au cours de laquelle on observe si l'individu réalise l'évènement (variable à expliquer)
    - Date de l'évènement connue uniquement lors de la phase de construction du score

- Important : le modèle devra être construit sur une base d'étude dont les données seront calées temporellement sur les mêmes hypothèses que lors de l'application du modèle
  - De nombreuses contraintes opérationnelles doivent donc être prises en compte

PÉRIODE D'ÉTUDE

Exemple initial

- Exemple : le modèle de score est déjà en production et nous souhaitons constituer la liste de clients à contacter le 24/10

  - Cette extraction se fera sur les données disponibles à date qui dépendent de la fréquence de mise à jour des bases de données
    - MAJ en temps réel : date des données = 24/10
    - MAJ quotidienne : date des données = 23/10
    - MAJ hebdomadaire : date des données = 20/10
    - MAJ mensuelle : date des données = 30/09

  - La campagne doit ensuite se dérouler sur une période qui dépend du canal
    - Fenêtre sur internet en temps réel : date de la campagne = 24/10
    - E-mail ou SMS : date de la campagne = 25/10
    - Courrier : date de la campagne = 28/10 au 03/11
    - Emission d'appels : date de la campagne = 28/10 au 10/11
    - Rendez-vous agence : date de la campagne = 28/10 au 24/11

  - Une fois la sollicitation reçue, le client peut aussi prendre du temps pour se décider avant que l'évènement soit retranscrit dans les bases de données

PÉRIODE D'ÉTUDE

Exemple initial

- Exemple (suite) :
  - Au final on peut donc avoir un écart plus ou moins important selon ces paramètres
  - Si tout se fait en temps réel
    - Date de référence = 24/10
    - Date de l'évènement = à partir du 24/10
  - Si la fréquence de mise à jour des bases et la durée de campagne sont importantes
    - Date de référence = 30/09
    - Date de l'évènement = à partir du 01/12
- Ces éléments sont donc à prendre en compte lors de la construction de la base d'étude, sous peine de solliciter le client après que sa décision soit prise !

PÉRIODE D'ÉTUDE

Synthèse

- Le délai entre date de référence et date d'évènement dépend donc de plusieurs paramètres, spécifiques à chaque score et chaque entreprise
  - Délai pour obtenir des données actualisées à la date de référence
  - Délai pour calculer le score
  - Délai pour transmettre le fichier à l'entité opérationnelle (exemple : outil de gestion de campagnes)
  - Délai pour envoyer le courrier
  - Délai pour contacter le client
  - Délai de réflexion du client
  - …
- Ce délai est souvent appelé « délai de carence » mais peut aussi se retrouver sous d'autres dénominations (délai de latence, période de gel, horizon de prévision, …)

PÉRIODE D'ÉTUDE

Synthèse

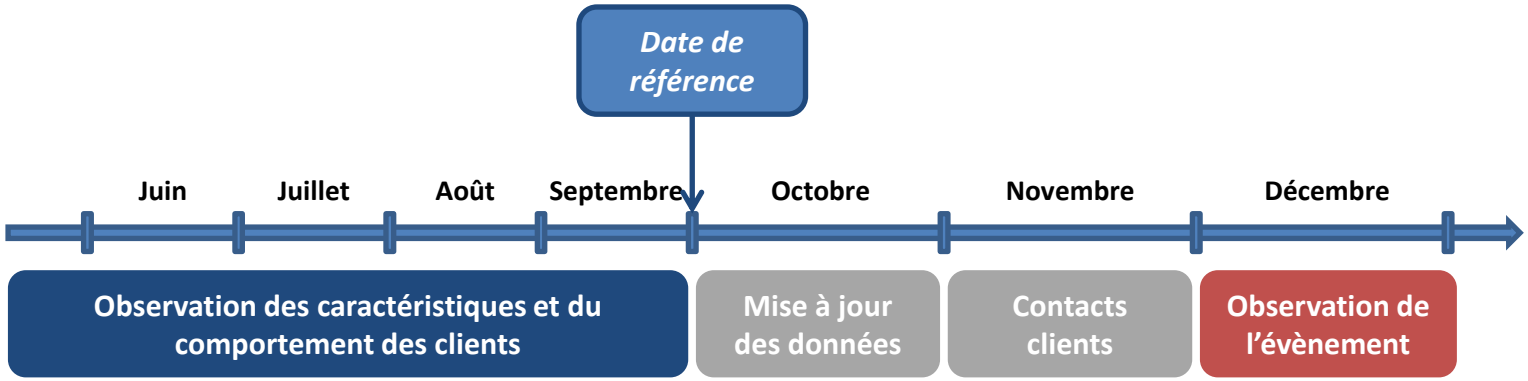
- Pour bien calibrer le délai de carence il est donc obligatoire de s'interroger sur le processus détaillé de l'application du score en analysant toutes les composantes listées précédemment

- Une fois cette analyse réalisée on en déduit les éléments à prendre en compte pour construire une base d'étude qui permettra au modèle statistique d'apprendre dans les conditions les plus proches de la réalité
  - Période pour la variable à expliquer (date d'évènement)
  - Période pour les variables explicatives (date de référence)

PÉRIODE D'ÉTUDE

Synthèse

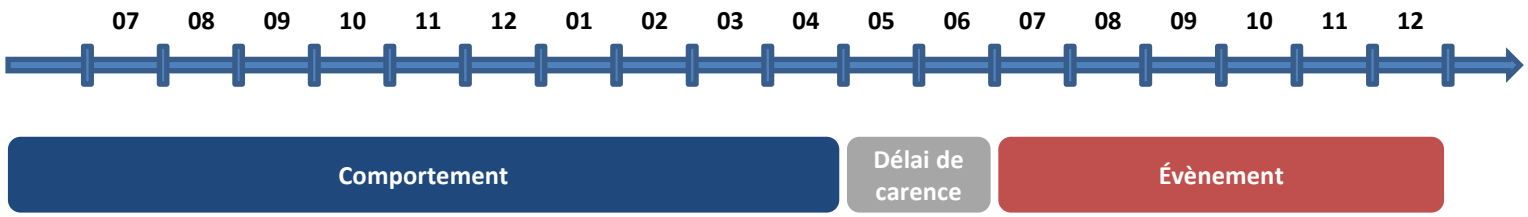
- Pour notre exemple sur les résiliations de l'abonnement télécom
  - Variable à expliquer (évènement à prédire)
    - 1 = le client a résilié en décembre
    - 0 = le client n'a pas résilié en décembre
  - Délai de carence = 2 mois, lié à des contraintes incompressibles
    - Les données du mois M sont disponible à M + 20 jours
    - La durée de la campagne marketing est estimée à 3 semaines
  - Date de référence = 30 septembre
  - Variables explicatives (comportement prédictif)
    - « Âge » calculé au 30 septembre
    - « Moyenne des appels sur 6 mois » calculée d'avril à septembre
    - …



PÉRIODE D'ÉTUDE

Complément

- Il est utile d'allonger la période pendant laquelle l'évènement est réalisé afin d'avoir une volumétrie suffisante et diminuer les effets de saisonnalité
  - Quelle date de référence choisir ?
  - En effet, plus cette période est longue plus l'écart entre la date où l'évènement est réalisé et la date où le comportement est analysé peut être important
    - Exemple : on souhaite prendre des résiliations de juillet à décembre
    - Date de référence = avril
    - On peut donc avoir pour certains clients jusqu'à 8 mois d'écart entre comportement prédictif et évènement à prédire (vs 3 mois « seulement » pour d'autres clients)

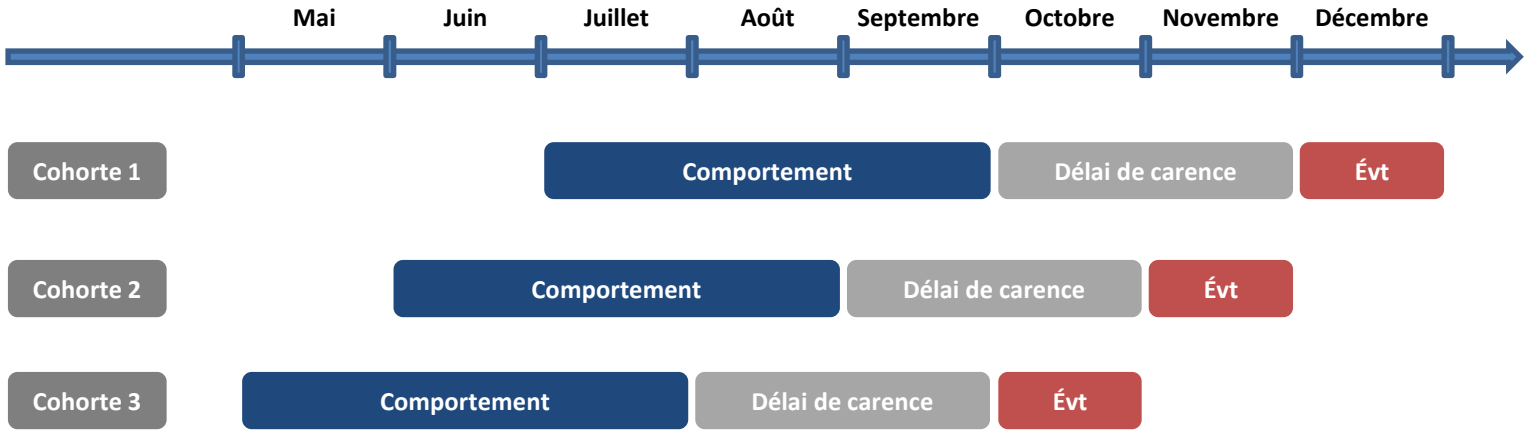


PÉRIODE D'ÉTUDE

Complément

- Une solution consiste à découper nos populations en cohortes et d'affecter à chacune une date de référence « glissante » selon la date de réalisation de l'évènement

- Dans notre exemple, en prenant 3 cohortes
  - Évènement réalisé en décembre ⇒ date de référence = 30 septembre
  - Évènement réalisé en novembre ⇒ date de référence = 31 août
  - Évènement réalisé en octobre ⇒ date de référence = 31 juillet



PÉRIODE D'ÉTUDE

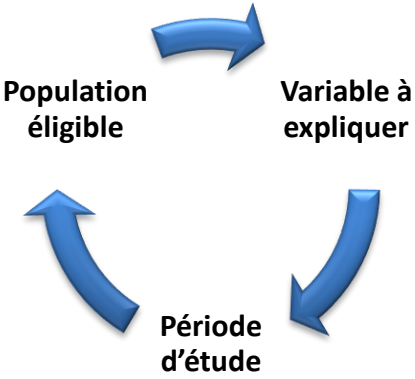
Complément

- En cumulant les périodes d'études, un problème se pose
  - Les sous-populations de résiliés sont distinctes, en revanche une grande partie des actifs d'octobre seront aussi actifs en novembre et en décembre
  - Afin de ne pas biaiser l'analyse, une solution consiste à leur affecter une date de référence aléatoire sur l'ensemble des périodes
    - 1/3 des clients auront comme mois de référence juillet, 1/3 des clients en août, et 1/3 des clients en septembre

- Quelques remarques
  - Plus l'évènement est rare, plus il faut de mois d'historique afin d'arriver à une volumétrie suffisante pour bien modéliser le comportement des clients
  - Plus l'historique est long, plus la saisonnalité est prise en compte
  - Plus l'historique est long, plus il peut y avoir d'éléments perturbateurs
  - Plus l'historique est long, plus les traitements sont lourds
    - En temps de programmation
    - En temps de calcul

BASE D'ÉTUDE

- Remarque importante sur ces 3 premières étapes de la construction de la base d'étude
  - Population éligible, variable à expliquer et période d'étude sont très liées
  - Il n'est pas rare d'avoir à réitérer et modifier certains traitements au fur et à mesure que la réflexion avance sur ces étapes
  - Par exemple on peut se rendre compte en construisant la variable à expliquer qu'une sous-population n'a jamais souscrit au produit
    - Il est alors raisonnable de s'interroger sur la pertinence d'intégrer cette sous-population à la population éligible
    - Tous ces choix doivent être réfléchis au moment de la construction du score, et leur validité seront à analyser lorsque le score sera en production



BASE D'ÉTUDE

- Population éligible
- Évènement à étudier
- Période d'étude
- Construction des variables explicatives
- Échantillons d'apprentissage et de validation(s)
- Optimisation des variables explicatives

VARIABLES EXPLICATIVES

- Les ***variables explicatives*** (ou « features ») correspondent à tous les indicateurs pouvant avoir un lien avec le phénomène étudié

- Dans cette partie, 3 étapes peuvent être distinguées
  - Rattachement des variables explicatives brutes
  - Fiabilisation des variables
  - Calcul des nouveaux indicateurs

VARIABLES EXPLICATIVES

Rattachement des variables de la BDD

- Une première série d'indicateurs est issue des variables brutes de la base de données
  - Ces variables doivent être correctement rattachées aux individus de la base d'étude en fonction de la date de référence
  - Exemple : âge du client
    - Date de référence en septembre ⇒ âge du client en septembre
    - Date de référence en juin ⇒ âge du client en juin

VARIABLES EXPLICATIVES

Nettoyage des données

- Chaque variable doit ensuite être fiabilisée via une analyse exploratoire

- Cette analyse doit permettre de détecter et si besoin traiter les points suivants
  - Variables inutiles
  - Valeurs manquantes
  - Valeurs aberrantes
  - Valeurs extrêmes
  - Modalités à faibles effectifs
  - Incohérence entre variables
  - Colinéarité entre variables

VARIABLES EXPLICATIVES

Nettoyage des données

- L'analyse exploratoire repose principalement sur une analyse univariée de la distribution de la variable
  - Variable quantitatives
    - Indicateurs de tendance centrale
    - Indicateurs de dispersion
    - Histogramme, boîte à moustache
  - Variables qualitatives
    - Répartition des modalités

- Certains problèmes seront en revanche détectés avec une analyse bivariée (exemples : retraités âgés de 20 ans, colinéarité entre le nombre de paiement et le montant des paiements), voire multivariée

VARIABLES EXPLICATIVES

Nettoyage des données

- Pour chaque anomalie, il faut comprendre d'où elle peut venir et décider de la meilleure stratégie à adopter
  - Supprimer les individus (⇔ population éligible)
  - Supprimer la variable
  - Remplacer la valeur par une estimation
- Remplacer la valeur par une estimation est une solution fréquemment utilisée
  - Moins de perte d'information
  - Mais il ne faut pas oublier que l'on modifie les données !
- Attention, certaines anomalies peuvent trouver une justification et ne doivent donc pas être traitées sans réflexion préalable
  - Une valeur manquante pour l'encours du Livret A peut être simplement causée par le fait que le client ne détient pas ce produit
  - Cette valeur a donc une signification et peut former une classe à part entière

VARIABLES EXPLICATIVES

Nettoyage des données

- Quelques exemples de transformations
  - Calculer une estimation logique en fonction des autres variables
  - Remplacer par la moyenne / médiane / mode
    - Calculés sur toute la population
    - Ou calculés sur un groupe de clients « proches » issu d'une typologie
  - Remplacer par un jumeau
  - Remplacer toutes les valeurs inférieures ou supérieurs à un quantile par la valeur du quantile (« écrêtage », « winsorisation »)
  - Remplacer par des valeurs permettant une distribution inchangée
  - Transformer en classes (« discrétiser »)
  - Créer un modèle statistique qui explique la variable posant problème en fonction d'autres variables (méthode plus complexe et induisant des colinéarités)
  - …

- Attention à ne pas systématiser les traitements : plusieurs méthodes d'imputations doivent être mises en concurrence pour traiter les anomalies

VARIABLES EXPLICATIVES

Construction de nouveaux indicateurs

- En plus de ces indicateurs bruts, il est indispensable de construire de nouveaux indicateurs
  - Moyenne / médiane de plusieurs variables
  - Ratio
  - Taux d'accroissement
  - Évolutions entre plusieurs dates
  - Moyenne / médiane sur une période
  - Croisement de variables
  - …

- Quelques exemples
  - Âge calculé à partir de la date de naissance
  - Nombre de produits par catégorie
  - Souscriptions et résiliations de produits
  - Volume d'appels moyen sur 6 mois
  - Évolution des encours sur 6 mois
  - …

VARIABLES EXPLICATIVES

Multi-colinéarité

- Dans les bases de données de nombreuses variables présentent des liens (multi-colinéarité)
  - Nombre de mouvements – montant des mouvements
  - Volume d'appels aux mois M6 – M5
  - …
- Beaucoup de méthodes de modélisation, en particulier la régression logistique, sont sensibles à ce problème : des variables explicatives trop corrélées ne devront donc pas être intégrées simultanément dans le modèle
- La détection de ces colinéarités se fait avec les méthodes classiques d'analyse du lien entre variables :
  - Tests d'hypothèses
  - Analyse factorielle
  - Classification de variables

VARIABLES EXPLICATIVES

Multi-colinéarité

- Il n'est pas conseillé de supprimer les variables corrélées entre elles lors de la construction de la base d'étude

- On ne connaît en effet pas encore la variable la plus pertinente à conserver : c'est la phase de modélisation qui permettra d'éliminer assez naturellement les variables n'apportant pas d'information supplémentaire par rapport aux variables déjà présentes dans le modèle

- Si de fortes colinéarités existent encore au sein du modèle construit il sera alors temps de traiter ce problème
  - Conserver une seule des variables via l'expertise métier, ou en privilégiant la plus forte intensité de la liaison avec la variable à expliquer
  - Construire un nouvel indicateur combinant les variables trop corrélées

BASE D'ÉTUDE

- Population éligible
- Évènement à étudier
- Période d'étude
- Construction des variables explicatives
- Échantillons d'apprentissage et de validation(s)
- Optimisation des variables explicatives

APPRENTISSAGE ET VALIDATION(S)

- Dans la partie « Modélisation » nous verrons que la construction du score sera réalisée en deux étapes
  - Estimation du modèle
  - Évaluation du modèle

- Afin de ne pas biaiser l'évaluation, il est indispensable que cette étape soit faite sur des données n'ayant pas servi à la construction du modèle
  - L'estimation du modèle se fera sur un échantillon appelé « apprentissage »
  - L'évaluation du modèle se fera sur un ou plusieurs échantillons appelés « validation(s) »

- D'une manière générale l'échantillon d'apprentissage contient 70% des données, et l'échantillon de validation contient les 30% de données restantes
  - La validation croisée est aussi une bonne alternative

APPRENTISSAGE ET VALIDATION(S)

- Un premier échantillon de validation peut être construit à partir de l'ensemble des données sur la même période d'étude
  - En gardant la sur-représentation des souscripteurs si les données ont nécessité un équilibrage de la variable à expliquer
  - Utile pour évaluer la stabilité du modèle en comparant les performances obtenues sur l'apprentissage avec celles obtenues sur la validation
- Un deuxième échantillon de validation peut être construit à partir de l'ensemble des données sur la même période d'étude
  - En utilisant la proportion réelle de souscripteurs
  - Utile pour évaluer des indicateurs avec une volumétrie réelle
- Enfin, un troisième échantillon de validation peut être réalisé avec des données d'une période plus récente (« out of time »)
  - Utile pour tester la robustesse du modèle sur de nouvelles données potentiellement troublées par une autre saisonnalité ou des phénomènes exogènes (lancement de produit, campagnes marketing, …)

BASE D'ÉTUDE

- Population éligible
- Évènement à étudier
- Période d'étude
- Construction des variables explicatives
- Échantillons d'apprentissage et de validation(s)
- Optimisation des variables explicatives

OPTIMISATION DES VARIABLES EXPLICATIVES

- L'objectif est de construire le meilleur modèle, c'est-à-dire celui qui discriminera au mieux les clients ayant réalisé l'évènement de ceux ne l'ayant pas réalisé

- Un modèle statistique s'écrit sous la forme d'une équation
  Variable à expliquer = fonction (variables explicatives)

- Il faut donc déterminer la meilleure combinaison de variables qui rentrera dans le modèle
  - Par conséquent ***chaque variable doit être la plus discriminante possible***
  - C'est l'objectif de l'optimisation des variables explicatives

OPTIMISATION DES VARIABLES EXPLICATIVES

- L'optimisation des variables explicatives consiste à les discrétiser
  - Ce n'est pas une obligation !
  - Mais de très nombreux scores en entreprises sont construits sur une base de variables discrétisées

- Principaux avantages de cette méthode
  - Elle permet une bonne prise en compte d'une liaison non-linéaire entre la variable à expliquer et la variable explicative
    - Si la liaison est linéaire, la discrétisation est moins pertinente (mais peut quand même être choisie pour des contraintes métiers et opérationnelles)
  - Elle peut résoudre des anomalies dans les données, valeurs manquantes ou individus atypiques, qui peuvent être regroupées dans une classe
    - Le modèle sera plus robuste (exemples : plus de 60 ans, plus de 100 000 €, …)
  - Cette transformation permet de synthétiser l'information apportée par la variable et rend plus facile la communication du modèle
    - Avec quelques classes bien choisies, on peut identifier des comportements caractéristiques des individus et l'interprétation en sera d'autant plus simple
    - Exemple : la facture discrétisée en trois classes représentera les clients ayant un montant à payer faible, moyen et élevé

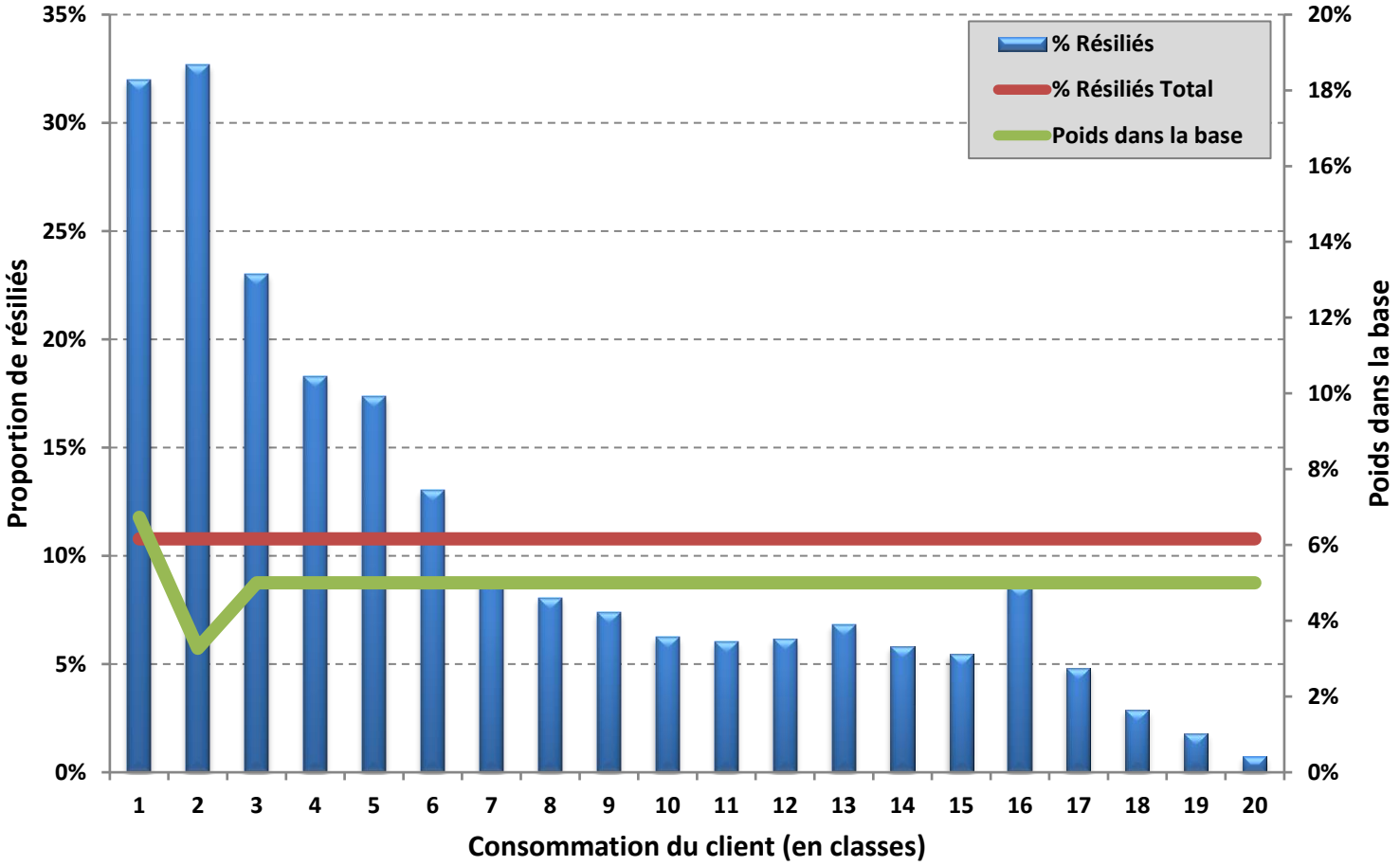
OPTIMISATION DES VARIABLES EXPLICATIVES

- La ***discrétisation optimale*** consiste à découper une variable quantitative ou à regrouper des modalités d'une variable qualitative de manière à ce que les ***classes*** obtenues ***discriminent*** au mieux ***l'évènement*** étudié
  - À l'intérieur d'une classe, les individus doivent être homogènes du point de vue de la variable à expliquer
  - D'une classe à l'autre, les individus doivent être très différents du point de vue de la variable à expliquer

- La variable discrétisée sera construite en regroupant les modalités dont le taux de cible ($\Leftrightarrow$ évènement) est similaire
  - S'il y a un ordre dans les valeurs de la variable initiale, il est préférable de conserver cet ordre en effectuant le regroupement
  - Pour une variable quantitative il faut dans un premier temps découper la variable en classes assez fines (une vingtaine de classes par exemple)
    - Seuils de classes logiques
    - Seuils de classes issus d'un découpage automatique (déciles, vingtiles, …)

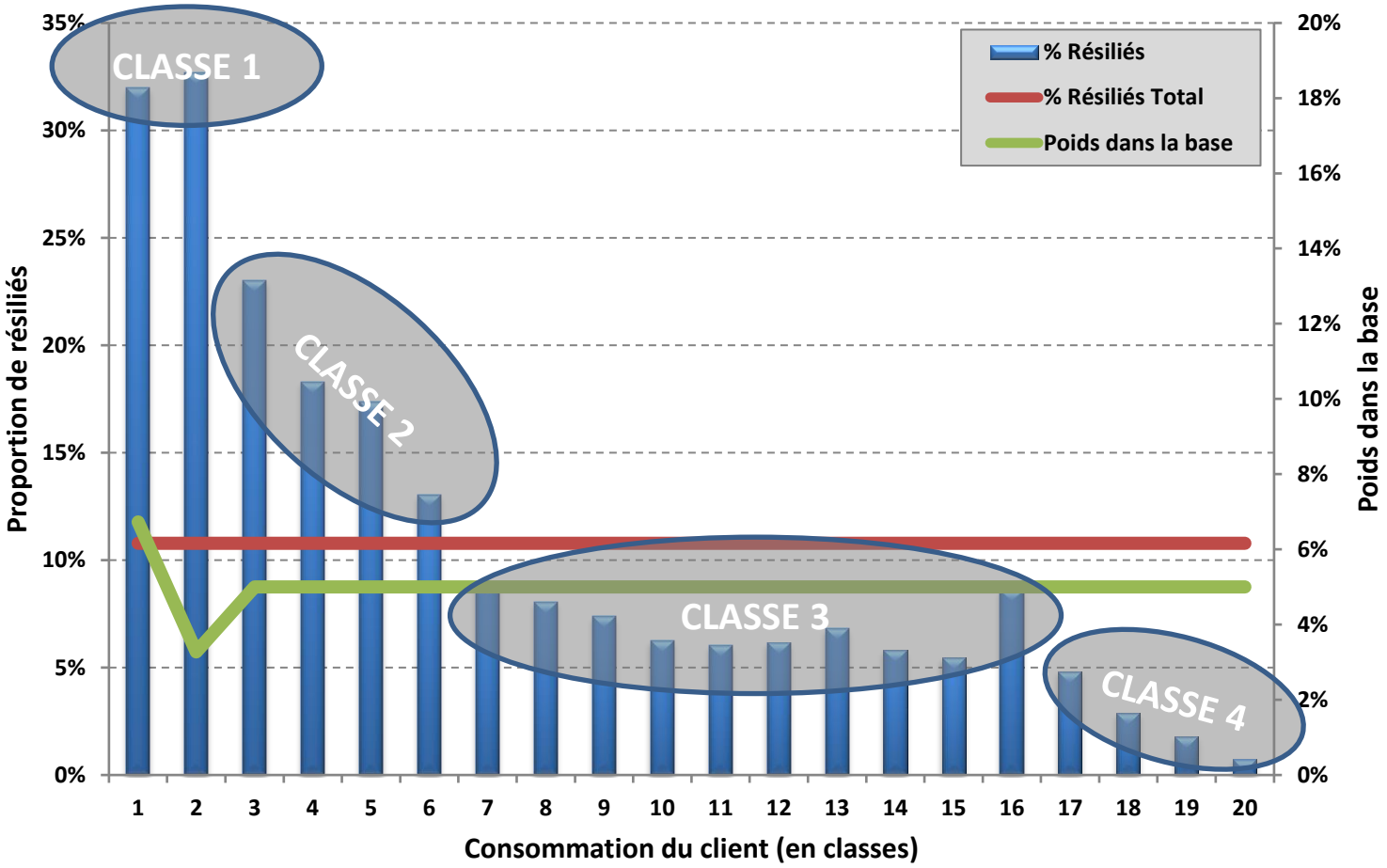
OPTIMISATION DES VARIABLES EXPLICATIVES

- Exemple : consommation du client
  - La variable quantitative est découpée en 20 classes



OPTIMISATION DES VARIABLES EXPLICATIVES

- Exemple : consommation du client
  - La variable quantitative est découpée en 20 classes



OPTIMISATION DES VARIABLES EXPLICATIVES

- La variable finale est construite avec ces 4 classes optimales
  - Le taux de résiliés est très différent selon les classes, cette nouvelle variable sera donc plus apte à bien discriminer les « Y = 1 » des « Y = 0 »
  - Il y a suffisamment d'individus dans chaque classe (éviter les classes avec moins de 5% des individus)

Classes optimales	Classes initiales	Taux de résiliés	Poids dans la base
1	1 à 2	32,3%	10%
2	3 à 6	17,1%	20%
3	7 à 16	7,4%	50%
4	17 à 20	2,1%	20%

OPTIMISATION DES VARIABLES EXPLICATIVES

- Remarques
  - Le nombre de classes ne doit pas être trop important
    - Environ 5 classes est un bon seuil, mais une 6ème classe peut être pertinente
    - Il sera toujours possible d'effectuer ensuite des regroupements de modalités lors de la modélisation si nécessaire
  - Si plusieurs regroupements sont en concurrence, l'expertise métier et l'intensité statistique du lien avec la variable à expliquer permettent de faire un choix
  - Au-delà de cette méthode manuelle qui peut être fastidieuse sur des bases contenant beaucoup de variables, des algorithmes de regroupements automatiques ou l'utilisation d'un arbre de décision fonctionnent très bien

BASE D'ÉTUDE

- À l'issue de cette phase de préparation des données, la base d'étude pourra prendre la forme suivante

ID CLIENT	CIBLE	DATE ÉVÈNEMENT	DATE DE RÉFÉRENCE	ÉCHANTILLON	ÂGE	SEXE	CSP	…
A	0	.	2010_08	Apprentissage	1	2	1	…
B	1	2010_11	2010_08	Validation	3	1	4	…
C	0	.	2010_09	Apprentissage	2	1	3	…
D	0	.	2010_08	Apprentissage	4	2	3	…
E	1	2010_12	2010_09	Apprentissage	3	1	4	…
F	0	.	2010_09	Validation	2	2	2	…
G	1	2010_11	2010_08	Apprentissage	4	2	3	…

- Elle peut contenir plusieurs centaines de variables explicatives
  - Ce n'est pas un problème, il n'est pas pertinent d'essayer de supprimer des variables à ce stade
  - Ce sont les méthodes statistiques qui les sélectionneront dans le modèle !

INTRODUCTION

PRINCIPES DU SCORING

CONSTRUCTION DE LA BASE D'ÉTUDE

MODÉLISATION

EXPLOITATION DU SCORE

CONCLUSION

PRINCIPE DE LA MODÉLISATION

- La première étape du scoring a permis d'établir une base d'étude, contenant des variables prêtes à la modélisation
  - Chaque variable est la plus liée possible à la variable à expliquer
- L'objectif est maintenant de construire le ***modèle statistique***, c'est-à-dire déterminer la meilleure fonction qui exprime la variable à expliquer en fonction des variables explicatives
- Cette phase de modélisation peut se décomposer en 3 parties
  - Construction de plusieurs modèles
  - Choix du « meilleur » modèle
  - Interprétation du modèle choisi

MODÉLISATION

Construction des modèles

Évaluation des modèles

Interprétation des modèles

CONSTRUCTION DES MODÈLES

- Un modèle de score peut être construit avec de nombreuses méthodes de modélisation statistique / algorithmes de machine learning
  - Régression logistique
  - Analyse discriminante
  - Arbres de décision
  - Forêts aléatoires
  - K plus proches voisins
  - Bagging
  - Boosting
  - Réseaux de neurones
  - Support Vector Machines
  - Algorithmes génétiques
  - …

CONSTRUCTION DES MODÈLES

- La méthode la plus utilisée a longtemps été la ***régression logistique*** qui représente un bon compromis en termes de performances et d'interprétation du modèle

- Certaines problématiques nécessitent des méthodes d'apprentissage plus évoluées afin de privilégier la performance
  - Comme souvent cela dépend du contexte, des objectifs et de la maturité de l'entreprise
  - Plusieurs méthodes peuvent être mises en concurrence afin d'identifier celle adaptée à la problématique et aux données

CONSTRUCTION DES MODÈLES

Sélection des variables

- En cumulant les variables brutes issues des systèmes sources avec les indicateurs calculés lors de la construction de la base d'étude, le nombre de variables explicatives peut parfois s'élever à plusieurs centaines !

- On ne peut évidemment pas construire un modèle avec autant de variables, tant d'un point de vue statistique qu'opérationnel ou même métier

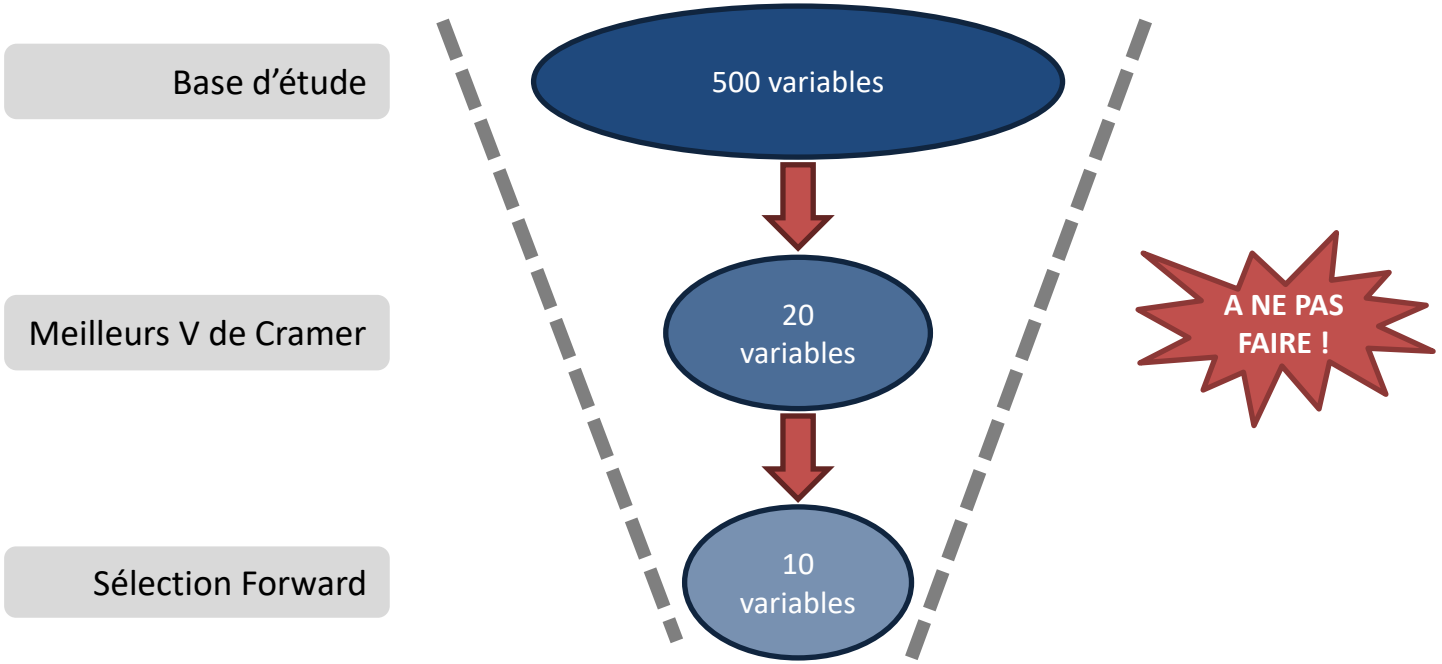
- La plupart des méthodes nécessitent donc de ***sélectionner les « meilleures » variables explicatives*** en utilisant
  - Les algorithmes de sélection automatique
  - La mesure de l'importance des variables dans le modèle
  - Les études de profils : les analyses bivariées et multivariées donnent de bonnes indications sur les futurs prédicteurs
  - La connaissance métier

CONSTRUCTION DES MODÈLES

Sélection des variables

- Attention, l'objectif d'un score n'est pas d'avoir le moins de variables possibles, mais d'avoir la ***meilleure combinaison de variables*** pour obtenir un ***score performant***

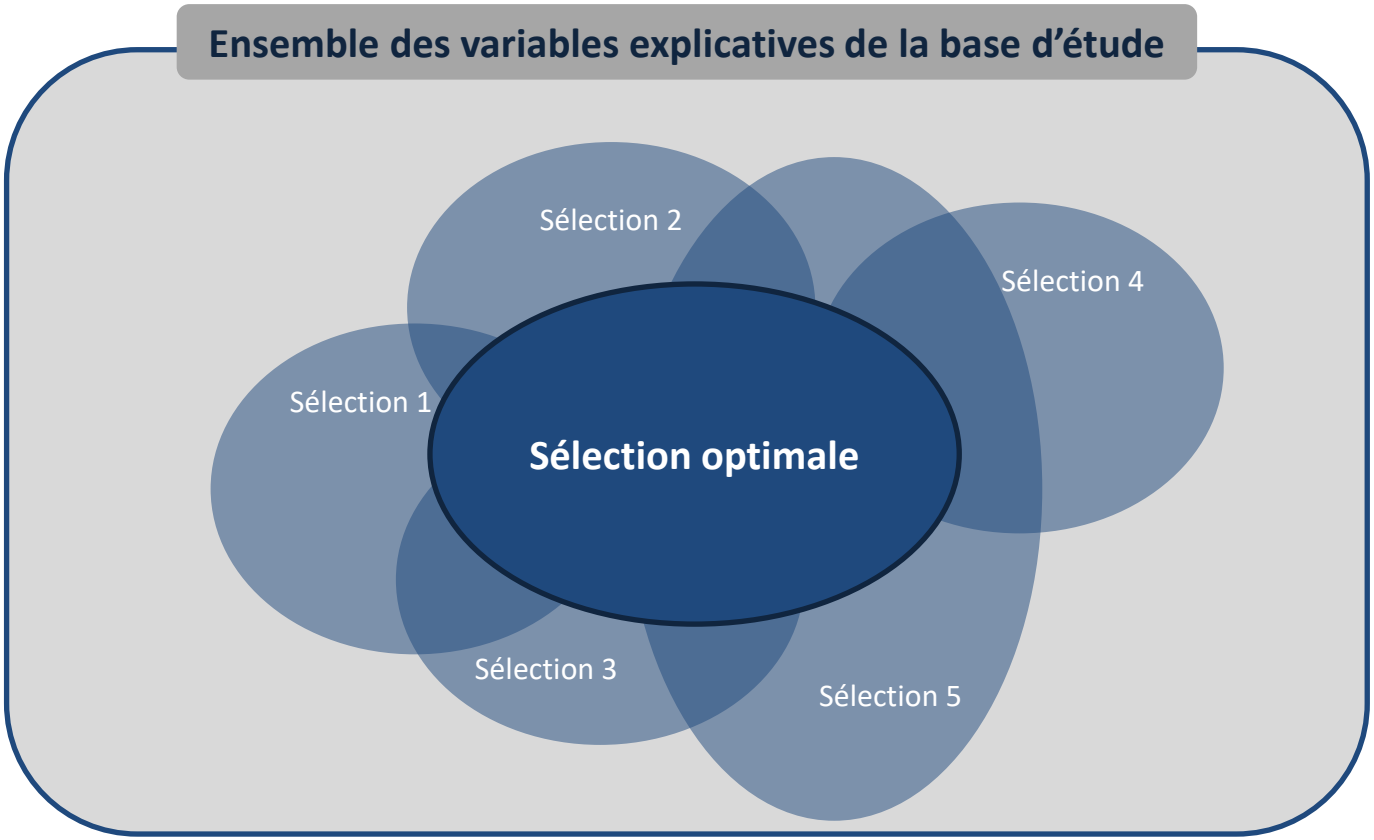
- Il ne faut donc pas raisonner en « entonnoir », en essayant d'enlever au fur et à mesure le maximum de variables



CONSTRUCTION DES MODÈLES

Sélection des variables

- Il est plutôt recommandé de ***combiner*** les méthodes de ***sélections automatiques ET manuelles*** à partir de l'ensemble des variables explicatives construites dans la base d'étude



CONSTRUCTION DES MODÈLES

Validité d'un modèle

- Les modèles construits doivent respecter des critères de validité (à adapter selon la méthode de modélisation)
  - Le coefficient estimé de chaque modalité des variables explicatives doit être significativement différent de 0
  - Le signe des coefficients doit être cohérent avec les analyses bivariées
  - La probabilité estimée ne doit pas trop dépendre d'une seule variable / modalité
  - Le modèle doit être stable
    - Stabilité de la significativité et de la valeur des coefficients du même modèle *réestimé* sur d'autres échantillons
    - Stabilité des performances du modèle *appliqué* sur d'autres échantillons
  - Le modèle ne doit pas être constitué de trop de variables explicatives
  - Les variables explicatives ne doivent pas être trop corrélées entre elles
  - Chaque variable explicative doit pouvoir s'expliquer facilement
  - Le modèle doit pouvoir s'appliquer facilement

- Et évidemment un modèle doit être performant ce qui nécessite des indicateurs pour l'évaluer

MODÉLISATION

Construction des modèles

Évaluation des modèles

Interprétation des modèles

ÉVALUATION DES MODÈLES

- Après avoir défini plusieurs modèles valides, il faut ***choisir le modèle final***

- Pour faire ce choix, de nombreux ***critères d'évaluation*** sont ***à comparer***
  - Critères de qualité d'ajustement
    - Tests Likelihood, Score, Wald
    - Critères AIC, BIC
    - R-Square
    - ...
  - Critères de qualité de prévision
    - Taux de Y=1 en fonction de la probabilité
    - Taux de bien classés / vrais positifs
    - Courbe ROC – AUC
    - Courbe de lift – Indice de Gini
    - ...
  - Critères métiers
  - Critères opérationnels

ÉVALUATION DES MODÈLES

Matrice de confusion – Courbe ROC – AUC

- La matrice de confusion et ses indicateurs associés est un outil privilégié par de nombreux utilisateurs

- Cette matrice se construit en ***comparant*** pour chaque individu la ***valeur*** de la variable à expliquer ***prédite*** par le modèle à sa ***valeur réelle***

- Cela nécessite donc d'utiliser un seuil permettant de découper la probabilité estimée
  - Si probabilité estimée < seuil $\Rightarrow$ Y estimé = 0
  - Si probabilité estimée ≥ seuil $\Rightarrow$ Y estimé = 1

- Le choix de ce seuil est donc essentiel
  - Il est souvent positionné à 0,5 par défaut mais ce n'est pas forcément pertinent, en particulier dans le cas d'un Y déséquilibré
  - Il est alors préférable de choisir un seuil correspondant au taux de Y=1

ÉVALUATION DES MODÈLES

Matrice de confusion – Courbe ROC – AUC

- Une fois le seuil choisi, la matrice de confusion prendra la forme suivante

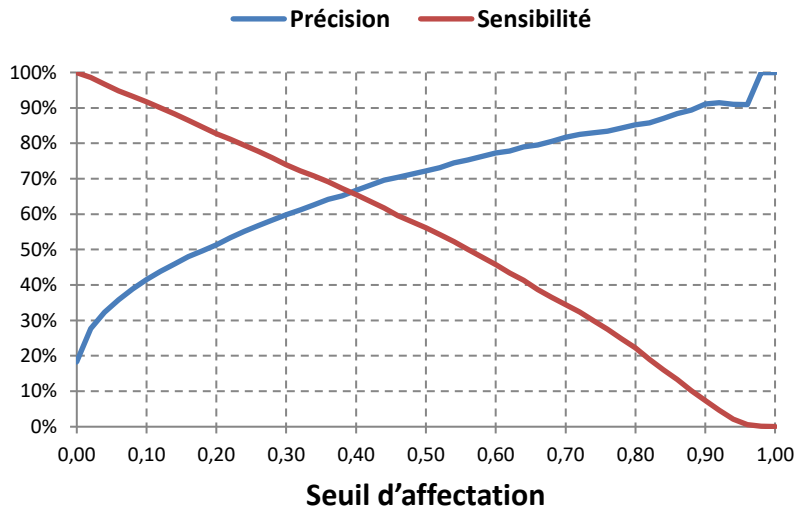
		VALEUR RÉELLE	
		1	0
VALEUR PRÉDITE	1	VP	FP
	0	FN	VN

- Principaux indicateurs associés
  - Taux de bien classés = Accuracy = (VP + VN) / (VP + FP + FN + VN)
  - Valeur prédictive positive = Précision = VP / (VP + FP)
  - Taux de vrais positifs = Sensibilité = Rappel = VP / (VP + FN)
  - Taux de vrais négatifs = Spécificité = VN / (FP + VN)
  - F1 score = 2 * (précision * sensibilité) / (précision + sensibilité)
  - Kappa = (accuracy – accuracy aléatoire) / (1 – accuracy aléatoire)
    - Accuracy aléatoire = taux « 1 » + taux « 0 » attendus si indépendance
      - Taux « 1 » = (VP+FP) / (VP+FP+FN+VN) * (VP+FN) / (VP+FP+FN+VN)
      - Taux « 0 » = (FN+VN) / (VP+FP+FN+VN) * (FP+VN) / (VP+FP+FN+VN)

ÉVALUATION DES MODÈLES

Matrice de confusion – Courbe ROC – AUC

- Les indicateurs précédents dépendent donc du seuil de probabilité choisi pour affecter un individu à une classe prédite et ne varient pas tous dans le même sens
  - Exemple : représentation de la précision et de la sensibilité en faisant varier le seuil d'affectation



Une diminution du seuil génère une cible plus grande

- On **privilégie alors la sensibilité** : on détecte beaucoup de Y=1 au sein de la cible
- Mais cela se fait **au détriment de la précision** : on intègre plus de Y=0 au sein de la cible (on prend plus de risques, le ciblage comporte plus d'erreurs)

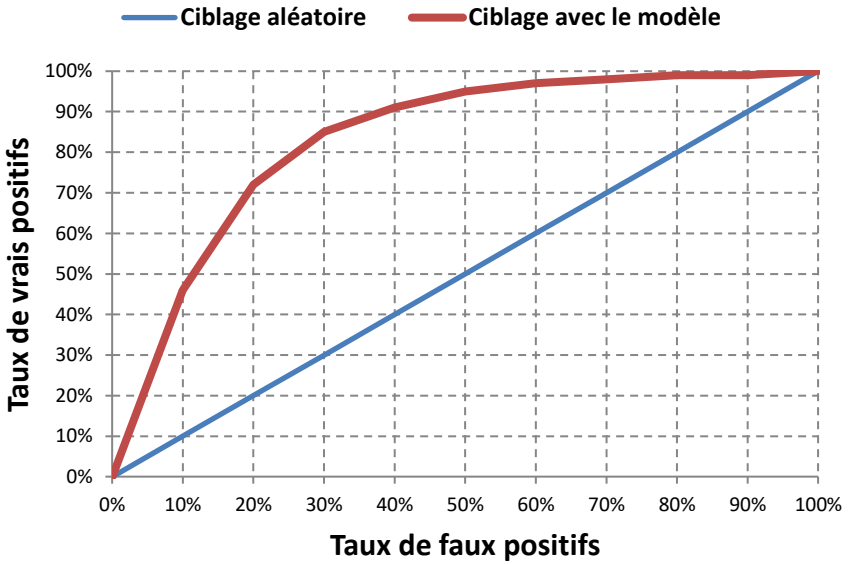
Une augmentation du seuil génère une cible plus petite

- On **privilégie alors la précision** : on intègre peu de Y=0 au sein de la cible (on prend moins de risques, le ciblage comporte peu d'erreurs)
- Mais cela se fait **au détriment de la sensibilité** : on détecte moins de Y=1

ÉVALUATION DES MODÈLES

Matrice de confusion – Courbe ROC – AUC

- Pour synthétiser les principaux indicateurs, la ***courbe ROC*** permet de représenter le taux de vrais positifs et le taux de faux positifs, en fonction de différents seuils



- L'aire sous la courbe ROC, appelée ***AUC***, représente la probabilité d'obtenir un score plus élevé pour un individu Y=1 que pour un individu Y=0
- C'est un bon indicateur de comparaison de modèles

ÉVALUATION DES MODÈLES

Lift - Gini

- Il n'est pas forcément toujours pertinent de vouloir affecter les individus à une classe grâce au modèle

- Dans le cadre d'un ***ciblage marketing***, on s'intéresse à ce que le modèle détecte correctement les futurs souscripteurs / résiliés parmi les individus que l'on contactera

- Les indicateurs et graphiques associés à la notion de ***lift*** permettent de quantifier la performance du modèle selon le niveau de probabilité estimée
  - On mesure la concentration d'individus ayant réalisé l'évènement étudié en fonction du nombre d'individus sélectionnés
  - Le lift s'intéresse donc à la qualité du modèle sur ce qui nous intéressera réellement, à savoir les individus avec une forte probabilité estimée (et non sur tous les individus de la base dont la majorité ne sera jamais contactée)

ÉVALUATION DES MODÈLES

Lift - Gini

- Exemple : 10 000 individus, les déciles sont calculés à partir de la probabilité estimée (D1 = probabilités les plus élevées)

Décile	Nb Y=0	Nb Y=1	TOTAL	% Effectif par décile	% Effectif cumulé	Concentration Y=1 par décile	Concentration Y=1 cumulé	Lift par décile	Lift cumulé
D1	610	390	1 000	10%	10%	39%	39%	3,90	3,90
D2	770	230	1 000	10%	20%	23%	62%	2,30	3,10
D3	880	120	1 000	10%	30%	12%	74%	1,20	2,47
D4	920	80	1 000	10%	40%	8%	82%	0,80	2,05
D5	940	60	1 000	10%	50%	6%	88%	0,60	1,76
D6	950	50	1 000	10%	60%	5%	93%	0,50	1,55
D7	970	30	1 000	10%	70%	3%	96%	0,30	1,37
D8	980	20	1 000	10%	80%	2%	98%	0,20	1,23
D9	990	10	1 000	10%	90%	1%	99%	0,10	1,10
D10	990	10	1 000	10%	100%	1%	100%	0,10	1,00
TOTAL	9 000	1 000	10 000	100%		100%			

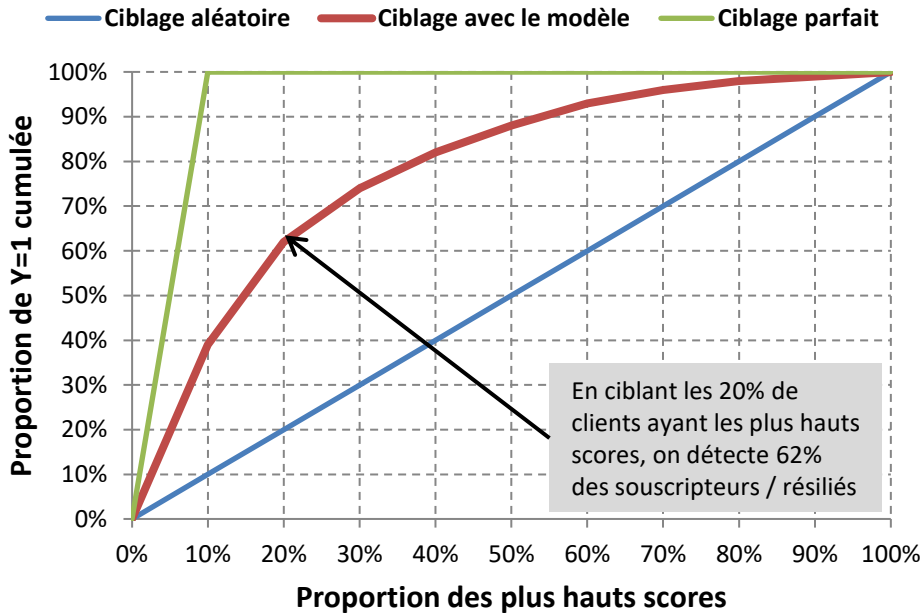
Courbe de lift

- La ***courbe de lift*** qui découle de ce tableau est un des critères les plus importants pour juger de la performance d'un modèle pour un ciblage de clientèle
  - En plus d'être statistiquement pertinente, la courbe de lift permet une interprétation métier et opérationnelle simple
  - Autres noms : courbe de concentration, courbe de gains cumulés, courbe de sélection

ÉVALUATION DES MODÈLES

Lift - Gini

- Exemple :



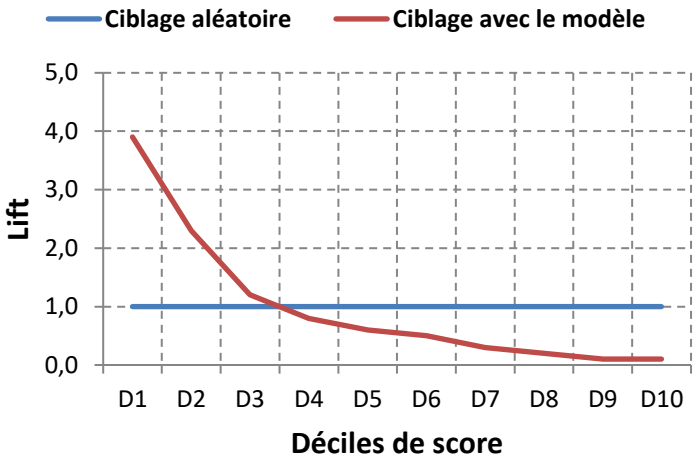
Remarque :
La courbe de lift et l'indice de Gini dépendent du taux de Y=1
Attention donc à ne pas comparer des modèles issus de populations dont l'évènement n'est pas distribué de la même manière

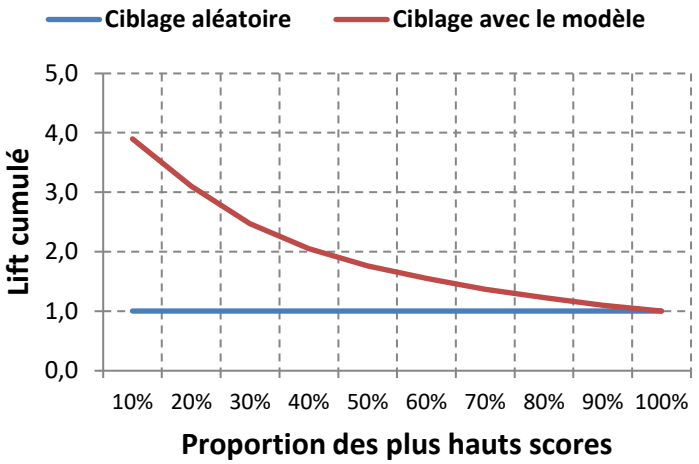
- L'indice de Gini, ou Accuracy ratio, correspond au rapport
  - De l'aire entre les courbes de lift de l'aléatoire et du modèle
  - Et de l'aire entre les courbes de lift de l'aléatoire et du parfait
- Gini = 2 * AUC - 1

ÉVALUATION DES MODÈLES

Lift - Gini

- D'autres représentations existent utilisant la valeur du lift

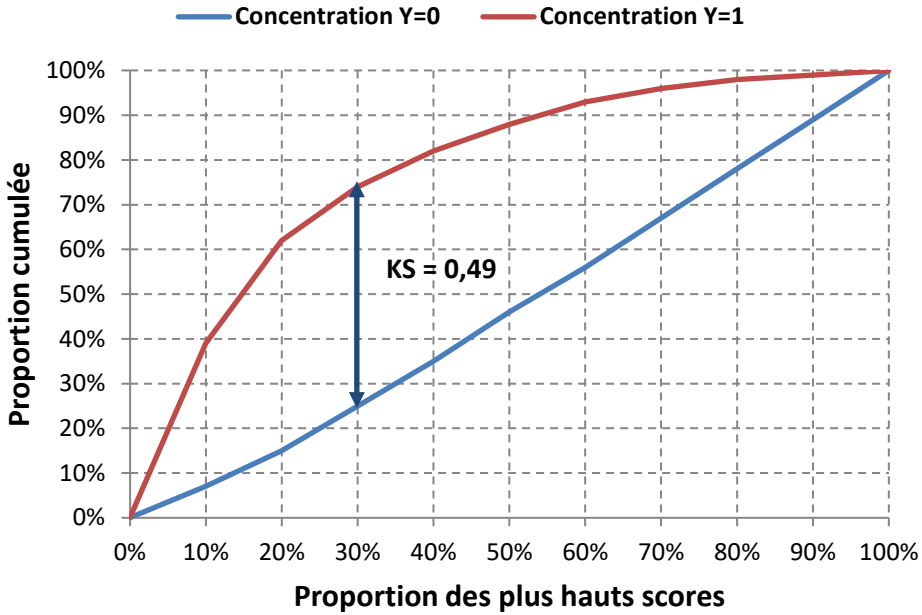




ÉVALUATION DES MODÈLES

Lift - Gini

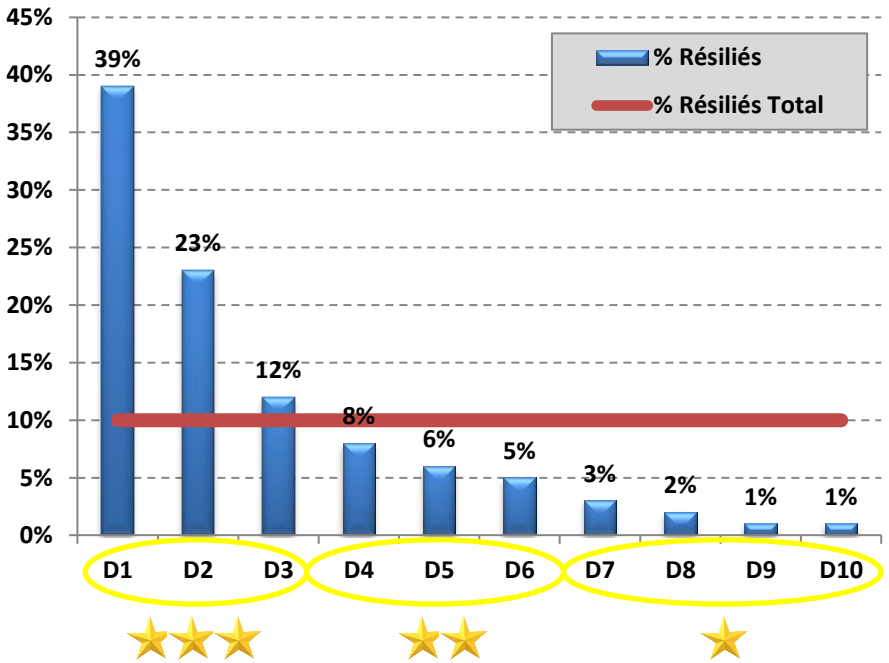
- Enfin une représentation de type Kolmogorov-Smirnov montre le niveau de séparation entre la détection des Y=1 et des Y=0



ÉVALUATION DES MODÈLES

Taux d'évènement

- On peut aussi représenter le taux de souscripteurs / résiliés par décile de score et vérifier que ces taux sont bien décroissants



- Le découpage de la population en plusieurs segments selon la valeur du score permet une utilisation opérationnelle simple

ÉVALUATION DES MODÈLES

- Il n'existe pas d'indicateur universel permettant d'évaluer et donc de comparer les modèles, il ne faut surtout pas se contenter du seul taux d'erreur par exemple

- Les indicateurs issus de la matrice de confusion doivent être utilisés avec précaution
  - Certains indicateurs sont totalement non pertinents en cas de données déséquilibrées (taux de bien classés par exemple)
  - Le coût de mauvais classement n'est souvent pas le même entre les 2 modalités de la variable à expliquer
  - Ces indicateurs dépendent du seuil choisi ce qui génère des variations de performance importantes et de sens inverse (précision vs sensibilité)

- Indicateurs à privilégier
  - Courbe ROC et indicateur AUC
  - Courbe de lift et indice de Gini (le plus pertinent dans le cadre d'un ciblage)

MODÉLISATION

Construction des modèles

Évaluation des modèles

Interprétation des modèles

INTERPRÉTATION DES MODÈLES

- Avoir un modèle performant est une nécessité, mais il est important de comprendre comment il est constitué, ou a minima de comprendre ce qui amène le résultat du modèle

- Cela permet
  - Au Data Scientist de vérifier que son modèle est pertinent, qu'il est cohérent par rapport aux analyses descriptives, qu'il ne comporte pas de biais, et donc de corriger des anomalies
  - Aux experts métiers d'avoir confiance en l'outil, de se l'approprier, et donc de l'utiliser afin d'élaborer une stratégie efficace
  - De respecter les normes et règlementations existantes

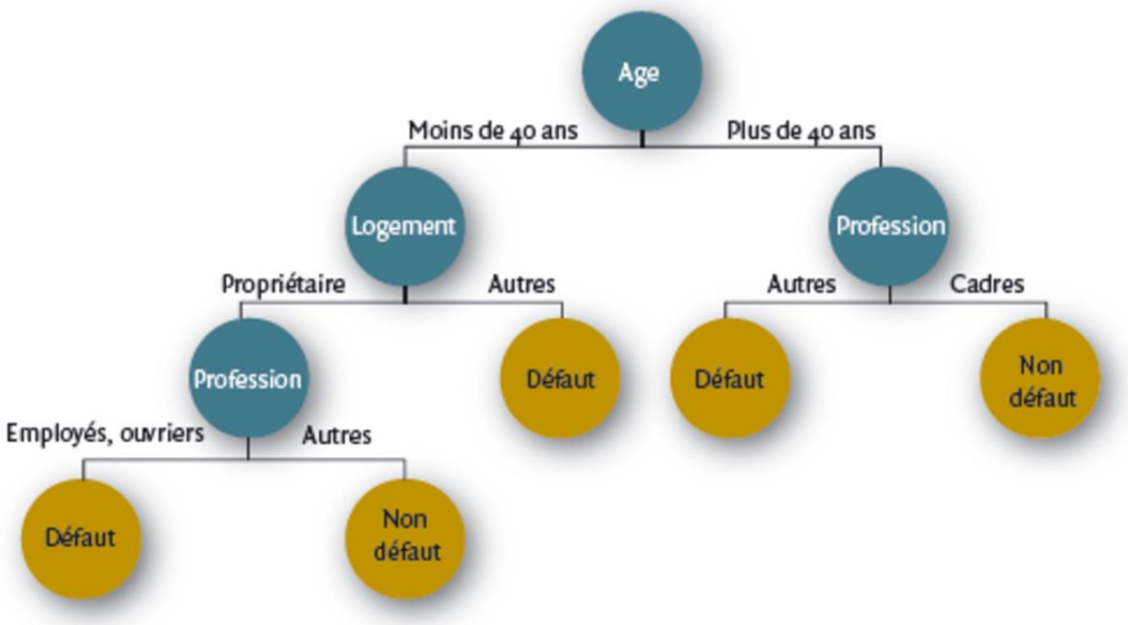
INTERPRÉTATION DES MODÈLES

- L'interprétation d'un modèle consiste principalement à répondre aux questions suivantes
  - Quelles variables ont le plus d'importance dans le modèle ?
  - Quelles valeurs ou plages de valeurs génèrent une probabilité estimée élevée ?

- Par exemple
  - L'âge joue-t-il un rôle important dans le score ?
  - Ciblera-t-on principalement des jeunes ou des personnes âgées ?
  - L'impact de cette tranche d'âge sur la probabilité estimée est-il plus important que détenir un forfait inférieur à 2H ?
  - Est-on capable de quantifier précisément l'apport de cette tranche d'âge sur la probabilité estimée ?

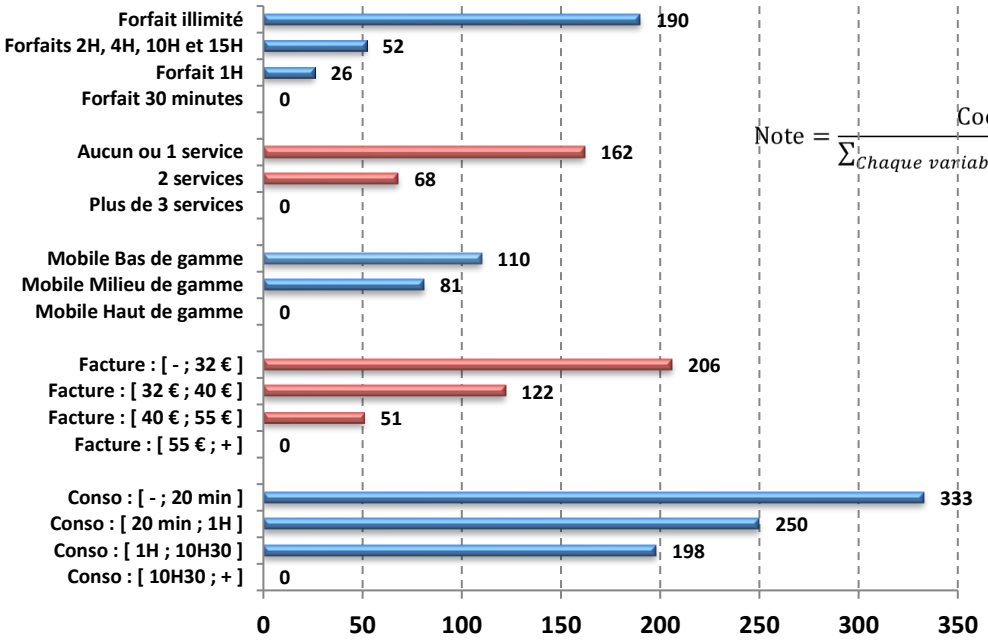
INTERPRÉTATION DES MODÈLES

- Les arbres de décisions fournissent une représentation directe des règles formant le modèle, l'interprétation est donc naturelle
  - Exemple pour un octroi de crédit (exemple fictif !)



INTERPRÉTATION DES MODÈLES

- La régression logistique permet aussi une représentation du modèle très facilement compréhensible par les experts métier
  - Il s'agit de calculer une ***grille de score*** avec une note par modalité
  - On peut ensuite additionner pour un client chaque note en fonction de son profil et on obtient au final un score « normalisé » entre 0 et 1000
  - L'ordre des clients fourni par ce score est exactement le même qu'avec la probabilité calculée avec le modèle



$$\text{Note} = \frac{\text{Coef}_{Modalité} - \text{Coef\_Min}_{Variable}}{\sum_{Chaque\ variable}(\text{Coef\_Max}_{Variable} - \text{Coef\_Min}_{Variable})} * 1000$$

Un client détenant un forfait 2H avec 1 service et un mobile haut de gamme, payant une facture mensuelle de 30 € et appelant 1H30 par mois aura un score de 618 sur 1000
(52 + 162 + 0 + 206 + 198)

INTERPRÉTATION DES MODÈLES

- Les méthodes traditionnelles (régression logistique, arbre de décision, analyse discriminante) sont donc très facilement interprétables

- L'interprétation des modèles issus d'autres méthodes d'apprentissage est plus complexe
  - La mesure de l'importance des variables est souvent intégrée dans les algorithmes
  - Mais pour comprendre finement le résultat du score selon tel ou tel profil d'individu, d'autres solutions doivent être envisagées
    - Analyse de la distribution de la probabilité estimée en fonction de chaque variable (analyses descriptives, PDP, analyses factorielles)
    - Application d'une méthode traditionnelle sur la prédiction estimée
    - Méthode LIME
    - Méthode SHAP

INTRODUCTION

PRINCIPES DU SCORING

CONSTRUCTION DE LA BASE D'ÉTUDE

MODÉLISATION

EXPLOITATION DU SCORE

CONCLUSION

QUELQUES RAPPELS

- L'objectif d'une « campagne marketing » est de solliciter les clients afin de leur proposer un produit, d'éviter qu'ils résilient …
- On peut difficilement contacter tous les clients : une cible restreinte de clients doit donc être définie
- Afin de maximiser la rentabilité de la campagne, il faut cibler les clients qui ont le plus de chance de souscrire au produit (ou de résilier)
- La définition de cette cible « idéale » peut être construite à partir
  - De règles métiers : « ciblage »
  - De règles statistiques : « scoring »

EXPLOITATION DU SCORE

- Une fois le modèle statistique construit, la dernière étape d'un projet de scoring consiste à exploiter / mettre en production le score

- Cette phase, à ne pas négliger, permet de livrer l'outil aux autres entités de l'entreprise
  - Application du score
  - Evaluation opérationnelle du score dans des campagnes test
  - Industrialisation du score
  - Suivi de la performance du score

EXPLOITATION DU SCORE

Application du score

Bilan de campagne

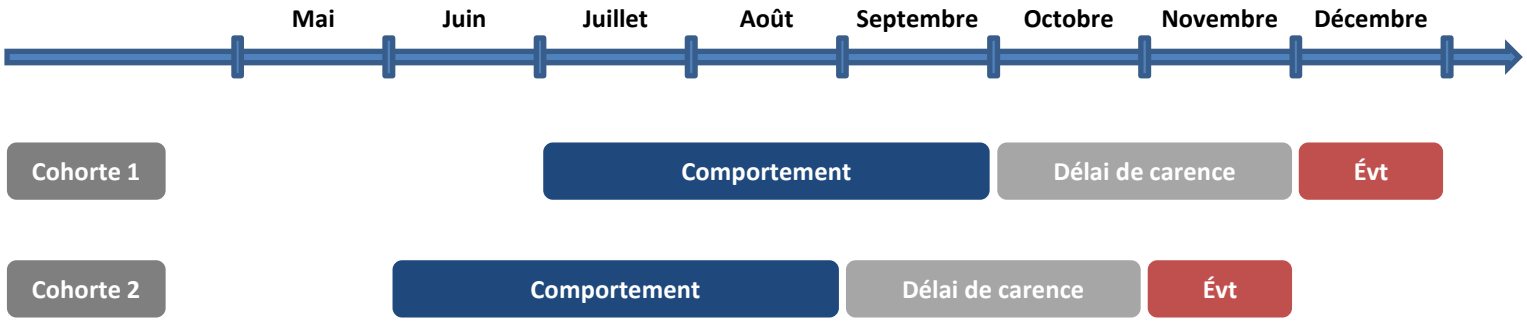
Industrialisation du score

Suivi du score

APPLICATION DU SCORE

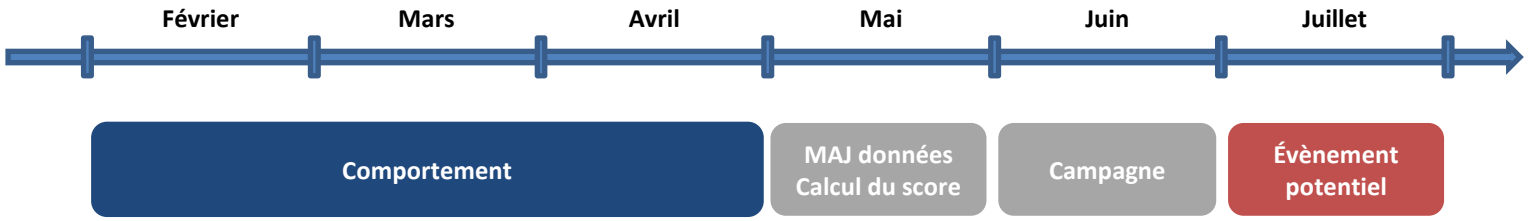
- Le scoring distingue 2 phases :
  - Estimation du modèle de score
    - À partir de clients ayant déjà réalisé l'évènement
    - Et de clients n'ayant pas réalisé l'évènement
  - L'application de ce modèle de score
    - Sur des clients n'ayant pas encore réalisé l'évènement

- Phase 1 : Estimation du modèle de score
  - Observation de l'évènement : novembre et décembre
  - Observation des caractéristiques des clients sur les 3 mois précédant la date de référence



APPLICATION DU SCORE

- Phase 2 : Application du score en vue d'une campagne marketing qui doit avoir lieu en juin de l'année suivante
  - Les données disponibles pour appliquer le modèle de score sont arrêtées à fin avril
    - Utiliser le même périmètre que lors de la construction du modèle de score (même « population éligible »)
    - Intégrer uniquement des clients n'ayant pas réalisé l'évènement
  - Le score est calculé en mai (le temps que les données à fin avril soient disponibles) grâce à la formule définie lors de la première phase
  - Les clients sont hiérarchisés en fonction de leur score (leur probabilité de réaliser l'évènement)
  - Seuls les X premiers clients constituent la cible de la campagne



EXPLOITATION DU SCORE

Application du score

Bilan de campagne

Industrialisation du score

Suivi du score

BILAN DE CAMPAGNE

- Une fois la campagne passée, on souhaite connaître ***l'efficacité*** du dispositif
  - Efficacité du score
  - Efficacité de la campagne marketing

- Il s'agit d'une « étude » dont la méthodologie consiste à comparer la population que l'on cherche à étudier et une population témoin

BILAN DE CAMPAGNE

- ***Un score est efficace*** si les clients sont bien hiérarchisés selon leur appétence ou leur risque
  - Cela revient à répondre à la question : est-ce que les clients ayant des scores élevés ont plus souscrit au produit (ou résilié) que les clients ayant des scores faibles ?
  - Il faut donc comparer 2 populations
    - Clients avec un score élevé
    - Clients avec un score faible
  - Et calculer le taux de souscription (ou taux de résiliation) pour chacune de ces populations

- Cela nécessite donc d'intégrer dans la campagne des clients peu appétants au produit proposé (ou peu risqués) …

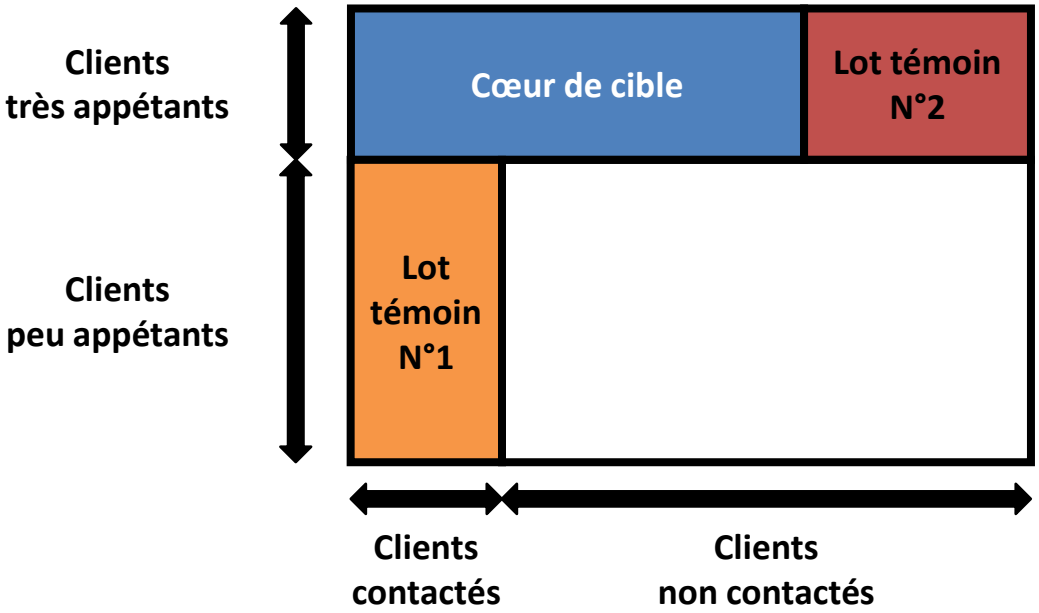
BILAN DE CAMPAGNE

- ***Une campagne est efficace*** si elle génère des souscriptions supplémentaires par rapport aux souscriptions naturelles
  - Cela revient à répondre à la question : est-ce que les clients sollicités ont plus souscrit au produit (ou résilié) que les clients non sollicités ?
  - Il faut donc comparer 2 populations
    - Clients contactés
    - Clients non contactés
  - Et calculer le taux de souscription (ou taux de résiliation) pour chacune de ces populations

- Cela nécessite donc de ne pas intégrer dans la campagne des clients très appétants au produit proposé (ou très risqués) …

BILAN DE CAMPAGNE

- En synthèse, pour tester les deux aspects de l'efficacité d'une campagne marketing scorée, il faut donc construire, en plus de la cible de clients très appétants qui sera contactée, 2 ***lots témoins*** (ou « lots aveugles », ou « groupes de contrôle »)
  - Efficacité score : clients peu appétants contactés (lot témoin N° 1)
  - Efficacité campagne : clients appétants non contactés (lot témoin N°2)



BILAN DE CAMPAGNE

- Le bilan de campagne s'effectuera principalement en comparant le taux d'évènement du groupe des clients très appétants contactés au taux d'évènement de chaque groupe témoin

- On peut ensuite élargir l'étude en fonction
  - Des différentes sollicitations soumises aux clients (offre, tarif, canal, …)
  - Des profils de clients
  - Des modifications de comportement du client suite à la campagne
  - …

- L'objectif final est de calculer un « ROI » du dispositif à partir des différents impacts positifs et négatifs de la campagne
  - Incrément de valeur client
  - Incrément de fidélisation
  - Coût de la campagne
  - …

BILAN DE CAMPAGNE

- Les taux de souscription / résiliation doivent évidemment être plus élevés pour les clients « cœur de cible » que pour les clients des lots témoins
  - Un résultat négatif pour la comparaison avec le lot témoin des clients avec des scores faibles montre que le modèle n'est pas assez performant
    - Des éléments n'ont pas été bien pris en compte dans tous les choix faits lors de la construction de la base d'étude et du modèle statistique
  - Un résultat négatif pour la comparaison avec le lot témoin des clients non contactés montre que la campagne n'est pas assez performante
    - Offre proposée, tarification, canal de communication sont probablement à revoir

- Ces enseignements seront utiles pour améliorer les futures campagnes et modifier le score si besoin
  - En particulier l'estimation d'un score basée sur les résultats d'une campagne est logiquement plus performant
  - On modéliserait ainsi la réponse du client suite à une sollicitation et non la simple souscription

EXPLOITATION DU SCORE

Application du score

Bilan de campagne

Industrialisation du score

Suivi du score

IMPLÉMENTATION DANS LE SI

- Une fois le score validé statistiquement et opérationnellement, il est nécessaire de l'industrialiser dans le système d'information : bases de données, outils de gestion de campagnes, …
  - Ce n'est pas au Data Scientist de mettre à jour manuellement le score à chaque nouvelle utilisation !

- Cette industrialisation nécessite un transfert aux équipes informatiques de toutes les règles définies lors de la construction du score
  - Population éligible
  - Construction des indicateurs rentrant dans le modèle
  - Calcul de la probabilité estimée par le modèle

- Une « recette » est ensuite nécessaire pour s'assurer que les valeurs calculées automatiquement sont bien celles attendues

EXPLOITATION DU SCORE

Application du score

Bilan de campagne

Industrialisation du score

Suivi du score

SUIVI DE LA QUALITÉ DU SCORE

- Le comportement des clients et l'environnement pouvant changer, la performance d'un score a pu se dégrader depuis sa mise en place

- Le ***suivi de la qualité*** d'un score consiste à comparer ses caractéristiques et sa performance entre la date de création et la date du suivi (au moins une fois par an)
  - Distribution du score
  - Distribution des variables du modèle
  - Nombre de souscripteurs
  - Taux de souscripteurs par décile de score
  - Courbe ROC, AUC, précision, recall
  - Courbe de lift, indice de Gini

- Remarque : les indicateurs par décile doivent être calculés avec les seuils identifiés lors de la création du score

INTRODUCTION

PRINCIPES DU SCORING

CONSTRUCTION DE LA BASE D'ÉTUDE

MODÉLISATION

EXPLOITATION DU SCORE

CONCLUSION

QUELQUES CONSEILS

- La construction d'un score nécessite plusieurs traitements à réaliser méticuleusement, en particulier la construction de la base en entrée du modèle
  - Des mauvais choix et / ou des erreurs dans certains traitements peuvent dégrader nettement la pertinence d'un score
  - Inversement certaines étapes bien réalisées permettent d'améliorer le score

• Construction de la base d'étude	
➢ Population éligible	↘ ↘
➢ Évènement à étudier	↘ ↘ ↘
➢ Période d'étude	↘ ↘
➢ Variables explicatives	↗ ↗ ↗
➢ Échantillonnage	↗
➢ Optimisation des variables	↗ ↗
• Modélisation	
➢ Méthodes de modélisation	↗ ↗
➢ Sélection des variables	↗ ↗
➢ Évaluation des modèles	↗
➢ Interprétation du modèle final	↗

QUELQUES CONSEILS

- Un score n'est pas une baguette magique qui va rendre toutes les campagnes marketing rentables !!!

- Un score ne crée pas de l'appétence
  - Un score sélectionne les clients les plus appétants
  - Un score ne garantit pas qu'il y ait beaucoup de clients appétants

- Un score hiérarchise des clients
  - Un score identifie les groupes de clients au sein desquels il y aura proportionnellement plus de probables souscripteurs (résiliés)
  - Un score ne sépare pas parfaitement les appétants des non-appétants ou les risqués des non-risqués

QUELQUES CONSEILS

- Un score reproduit des pratiques commerciales
  - Si un produit est systématiquement vendu à un profil de client, c'est ce profil qui ressortira dans le modèle de score
  - Il peut être pertinent de compléter un score par un « ciblage métier »
    - Études marketing (quantitatives, qualitatives)
    - Évolutions sociétales, observation des tendances comportementales des clients
    - Retours d'expériences des campagnes marketing

- La construction d'un score peut être largement automatisée
  - Attention à l'effet « boîte noire » !
  - Cela nécessite de nombreuses années d'expérience sur des projets de scoring
  - Et n'empêche pas un contrôle pointu de chaque étape

- Et enfin la qualité des données impacte très largement la qualité d'un modèle, quelle que soit la performance de l'algorithme

MERCI !

Jean-Philippe KIENNER