optimind.
manage risk, build your future

JOURNÉE IARD

UTILISATION DE DONNÉES EXTERNES EN MODÉLISATION

29 et 30 mars 2018

Marie-Catherine SARRAUDY
Partner Actuarial Services

Coralie LE PLAT
Manager IARD

Matthieu LAGADEC
Manager IARD

Sommaire

optimind.
manage risk, build your future

01

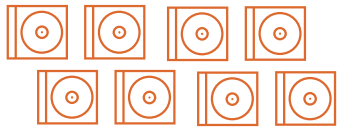
Data is everywhere !

Data is everywhere !

Quelques chiffres…

44 ZB

De volume de données d'ici 2020.



Une tour de DVD de 13 millions de kms de haut

300 h/min

De vidéos you tube

Réseaux sociaux un des 1er provider de données

50 Mds

D'objets connectés d'ici 5 ans

De plus en plus de données seront :

- ✓ *Collectées*
- ✓ *Analysées*
- ✓ *Objectif : partager les données*

300Mds$/an

D'économie santé au USA

+60%

De marge opérationnelle pour les distibuteurs

Créateur de valeur

Dans tous les domaines

< 0,5%

Du potentiel des données utilisé actuellement

Défis pour :

- ✓ *Intégrer*
- ✓ *Analyser*
- ✓ *Former*

Data is everywhere !

Quels types de données nous entourent ?

DONNEES STRUCTUREES

Données contenues dans les champs, identifiables facilement, avec un format prédéfinit, une dénomination explicite

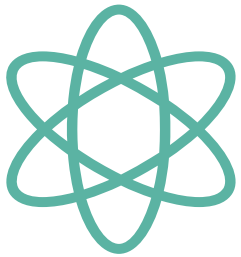
Exemples :

- *Base de données interne*
- *Informations externes requêtables*

DONNEES INTERNES

Bien souvent non partagées entre services

Notion de DARK DATA : données internes inexploitées

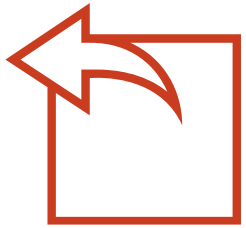


DONNEES NON STRUCTUREES

Données représentées ou stockées sans format prédéfini

Exemples :

- *Textuelles : courriels, les présentations PowerPoint, les documents Word,*
- *Non textuelles : images JPEG, les fichiers audio MP3*



DONNEES EXTERNES

- Données publiques (INSEE, data.gouv.fr)
- Données vendues (AAA Data)

- Attention à la réglementation (RGPD, propriété intellectuelle)
- Besoin de moyen technique (SI) et opérationels (équipes formées) pour les exploiter

Data is everywhere !

Comment les exploiter ? La chaîne de valeur autour de la valorisation de la donnée

01

STRATEGIE

Identification des **besoins**

02

SOURCING ET COLLECTE

Identifier les **sources de données**

Sélection et récolte des **données utiles**

03

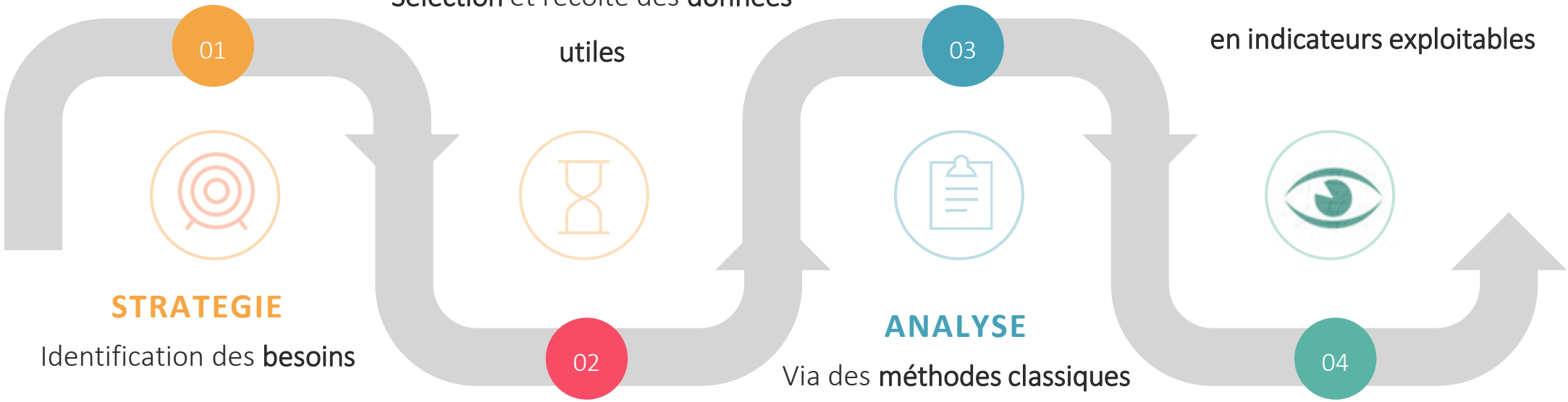
ANALYSE

Via des **méthodes classiques** ou **méthodes machine learning**

04

RESTITUTION

Transformation des données brutes **en indicateurs exploitables**



Data is everywhere !

Exemples d'applications du couple données externes / machine learning en assurance

GESTION SINISTRE

- Gestion de la fraude
- Amélioration des processus de gestion

MARKETING

- Comportement client
- Reduction du churn
- Géomarketing

ACTUARIAT

- Tarif (sinistres et frais)
- Actifs
- Réassurance
- CAT NAT / Modélisation des risques climatiques
- Indicateurs de pilotage

02

Mise en pratique de la valorisation de données externes en MRH

Contexte de l'étude - Objectif

Créer, à partir d'**arbre de classification**, un **outil de pilotage** de portefeuille permettant de **repérer des poches de contrats sous tarifés** au sein d'un portefeuille MRH.

MRH

Données externes

CART

Pilotage

Collecte et traitement des données

- Base tarifaire d'un produit MRH
- Ajout de données externes
- Prise en compte d'informations qui ne sont pas incluses dans le tarif

Modélisation

- Application de méthodes de Machine Learning
- Différente mais complémentaire aux GLM

Data visualisation

- Identification des poches de contrats sous-tarifées
- Outil de suivi des résultats

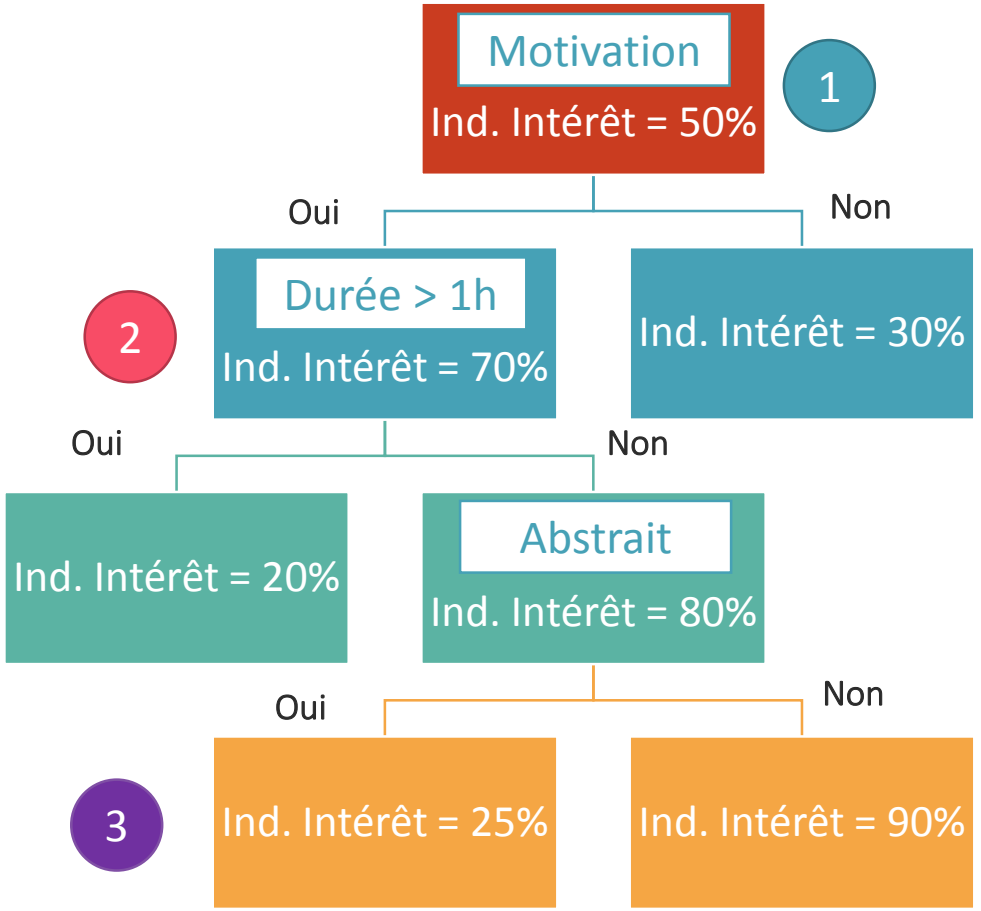
Principe de l'algorithme CART

Principe

1. On part :
   - **D'une variable à expliquer** : *« Etes vous intéressés par cette présentation ? »* ***Ind. intérêt***
   - De **variables explicatives** Variable
   - D'un **critère / d'une modalité** – *Exemple avec la durée*

2. L'objectif est de **segmenter en regroupant de façon binaire et par étapes** (nœuds) des critères homogènes ensemble (minimisation de la variance inter groupe).
   Deux réponses possibles : Oui / Non

3. De nombreuses questions autour de ces arbres :
   - Quand s'arrêter ?
   - Jusqu'à combien de variables explicatives puis-je introduire ?
   - Est-ce robuste ?



Contexte de l'étude – Algorithme CART

Avantages

- ✓ Sélection automatique des variables
- ✓ Identification de phénomènes non identifiés
- ✓ Visualisation simple de l'analyse multivariée
- ✓ Possibilité d'ajuster les métriques (choix parfois complexe)

Inconvénients

- × Un arbre de qualité peut nécessiter de nombreuses branches
- × Interprétation parfois complexe
- × Le modèle peut présenter une certaine instabilité

$\max\left(\left|\frac{S_1}{P_1} - \frac{S_2}{P_2}\right|\right)$: maximisation de l'écart de l'indicateur d'intérêt entre les nœuds fils

Contexte de l'étude – Base de données

BASE MRH

425 000
années polices

POPULATION DE RÉSIDENTS

30
variables explicatives

1 FORMULE

9 %
avec un sinistre à charge

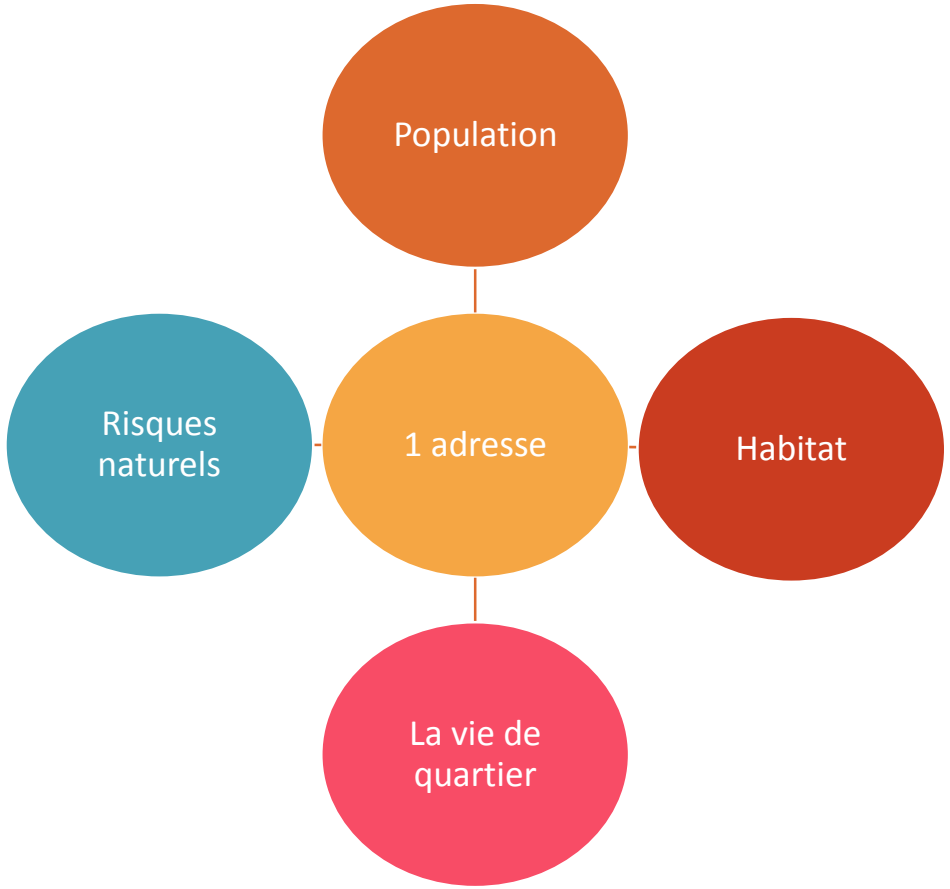
3 GARANTIES

100%
S/P indiciel

Sourcing – Données externes

Pour cette étude, la base a été enrichie de données externes :

- Structurées par **Code Postal/Département**
- Créant si possible des **distorsions géographiques**



Payants

Gratuits

Etapes de travail pour l'élaboration d'un CART

TRAITEMENT DES DONNEES

- Retraitement des variables
- Analyses univariées, corrélations, multivariées

PARAMETRAGE DES MODELES

- Choix des paramètres et méta-paramètres des modèles
- Utilisation de critères d'optimisations applicables

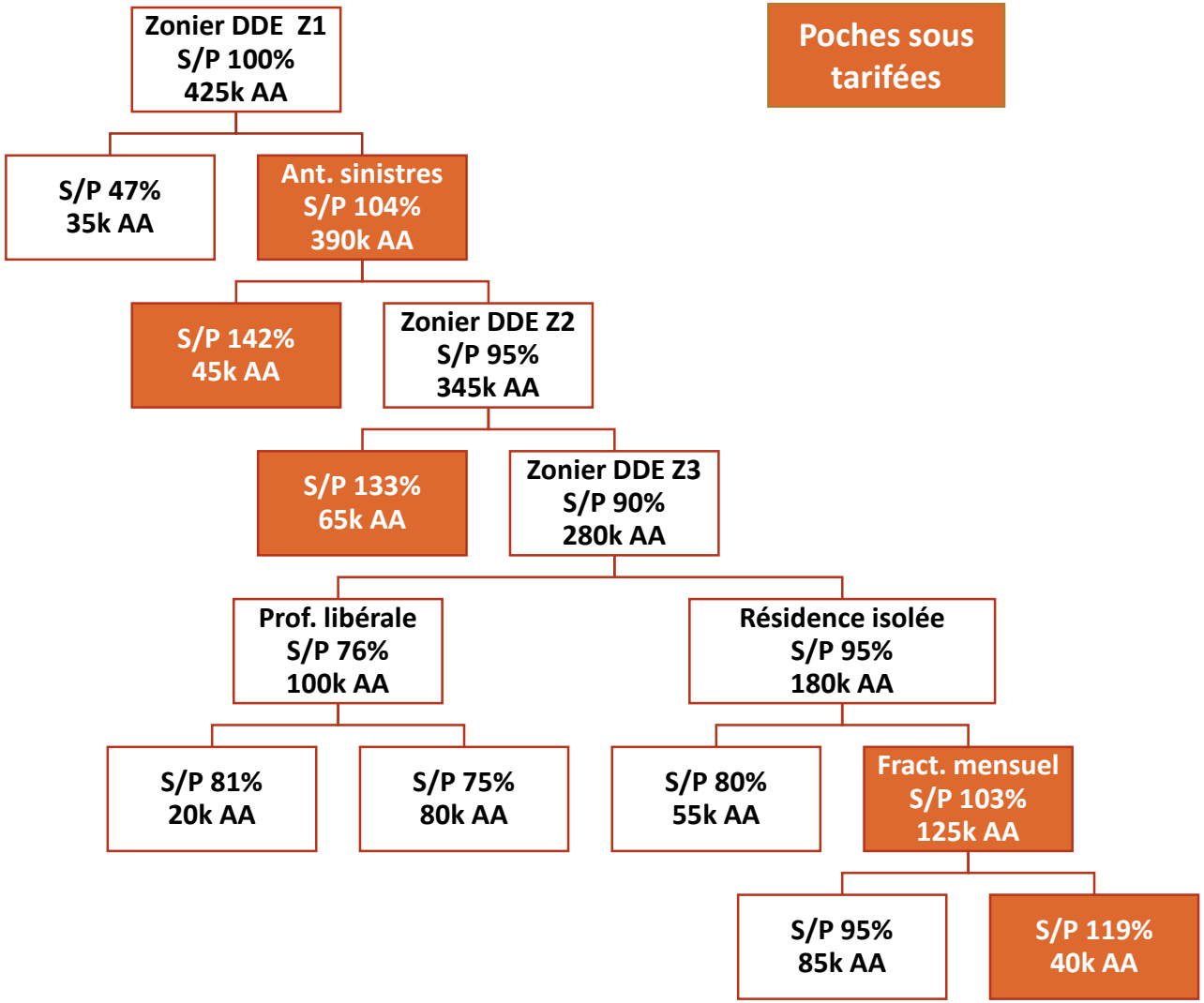
AJOUT DE DONNEES EXTERNES

- Présélection des données à utiliser
- Traitement des données sélectionnées
- Corrélations entre les données internes & externes

ANALYSE DES RESULTATS

- Analyse de l'impact des nouvelles données
- Interprétabilité des résultats

Le pilotage du portefeuille sans données exogènes



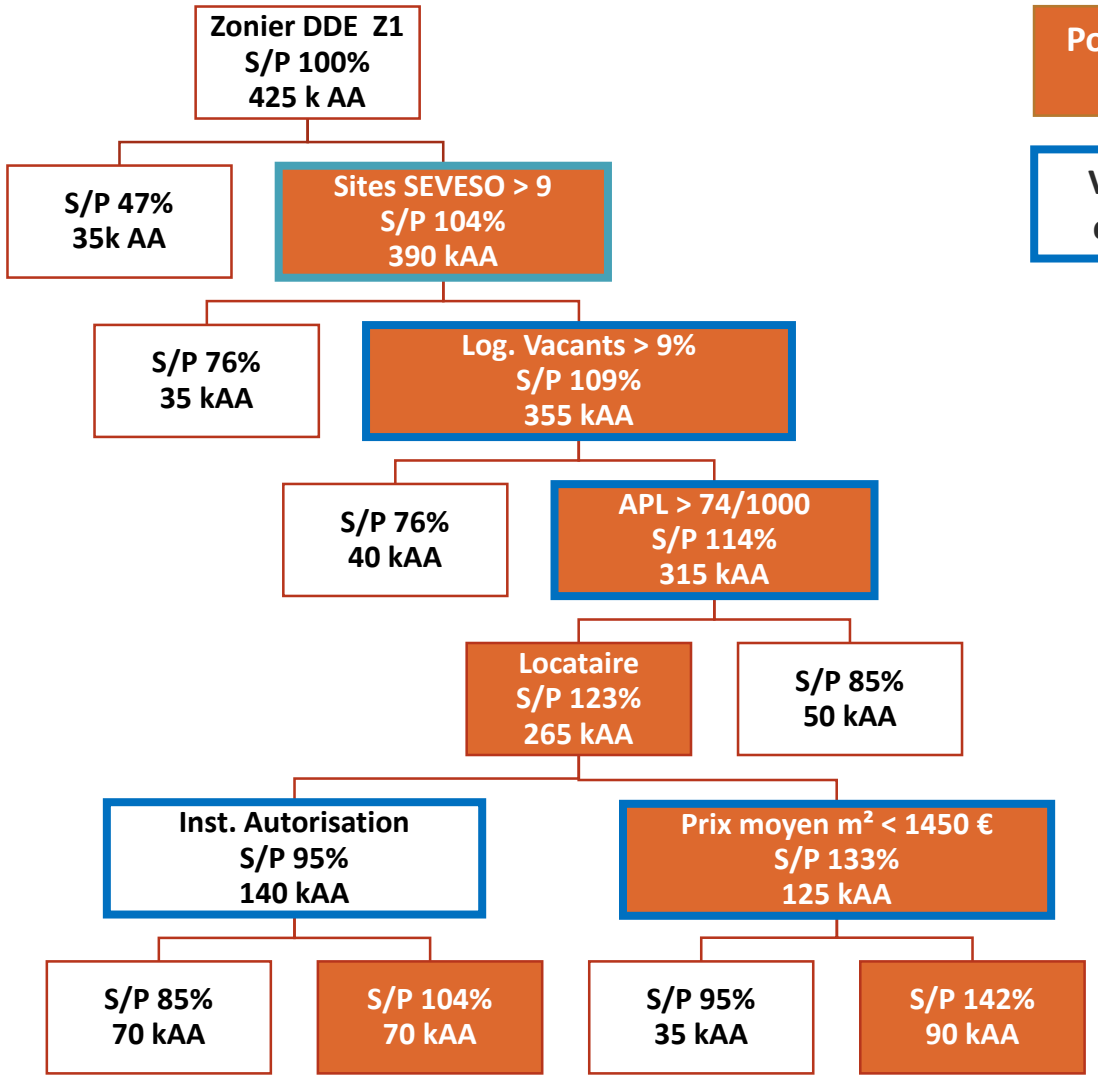
Segments déficitaires (5 poches)

- En dehors de la zone 1 et avec antécédents
- Dans la zone 2 et sans antécédent
- En dehors des zones 1, 2 et 3, dans une zone non isolée et avec une prime à fractionnement annuel

Observations

- **Un zonier** qui apparaît comme le premier élément expliquant la présence de poches sous-tarifées

Le pilotage du portefeuille avec données exogènes



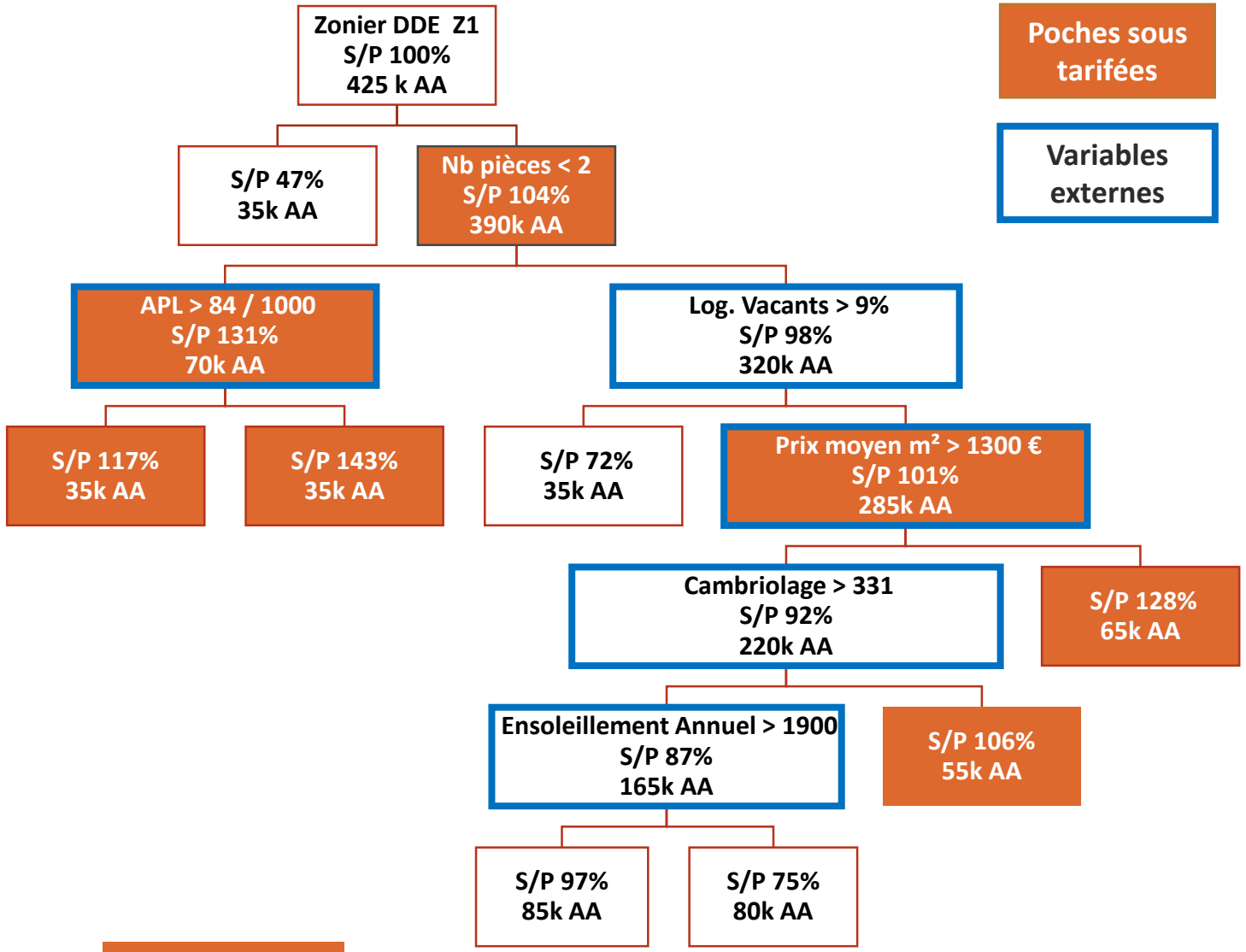
Segments déficitaires (7 poches)

- En dehors de la zone 1
- Avec un nombre de sites SEVESO ≤9
- Avec un nombre de logements vacants ≤9%
- Avec une proportion d'APL > 74/1000
- Pour les propriétaires
- Prix moyen m² < 1450 €

Observations

- Un **rôle prépondérant des données externes** dans l'explication des poches sous-tarifées (5 variables)
- **QUID de l'interprétation de la variable** SEVESO ? Installations soumises à autorisation ?
- Besoin d'**expertise métier** !

Le pilotage du portefeuille avec données exogènes & retraitements (expertise métier)



Variables retirées

- SEVESO
- Installations soumises à autorisation

Même nombres de segments déficitaires (7 poches)

Observations

- Même nouvelles variables externes + 2 nouvelles variables externes (dont une qui ne semble pas expliquer la sous-tarification)

Conclusion de l'étude

1. Les données externes permettent d'affiner le profil de risque

2. L'expertise métier reste indispensable pour comprendre les interactions de ces nouvelles données dans la modélisation du risque
   - L'algorithme n'est pas auto-suffisant et nécessite une validation des interactions observées entre les variables *a posteriori*

3. Besoin de temps pour tester la robustesse de l'approche et aller plus loin (création de nouveaux zoniers ?)
   - Instabilité en termes de classification lors de l'introduction ou de l'exclusion de nouvelles variables
   - Disponibilité des données externes

03

Quelles utilisations opérationnelles des données externes ?

Quelles utilisations opérationnelles ?

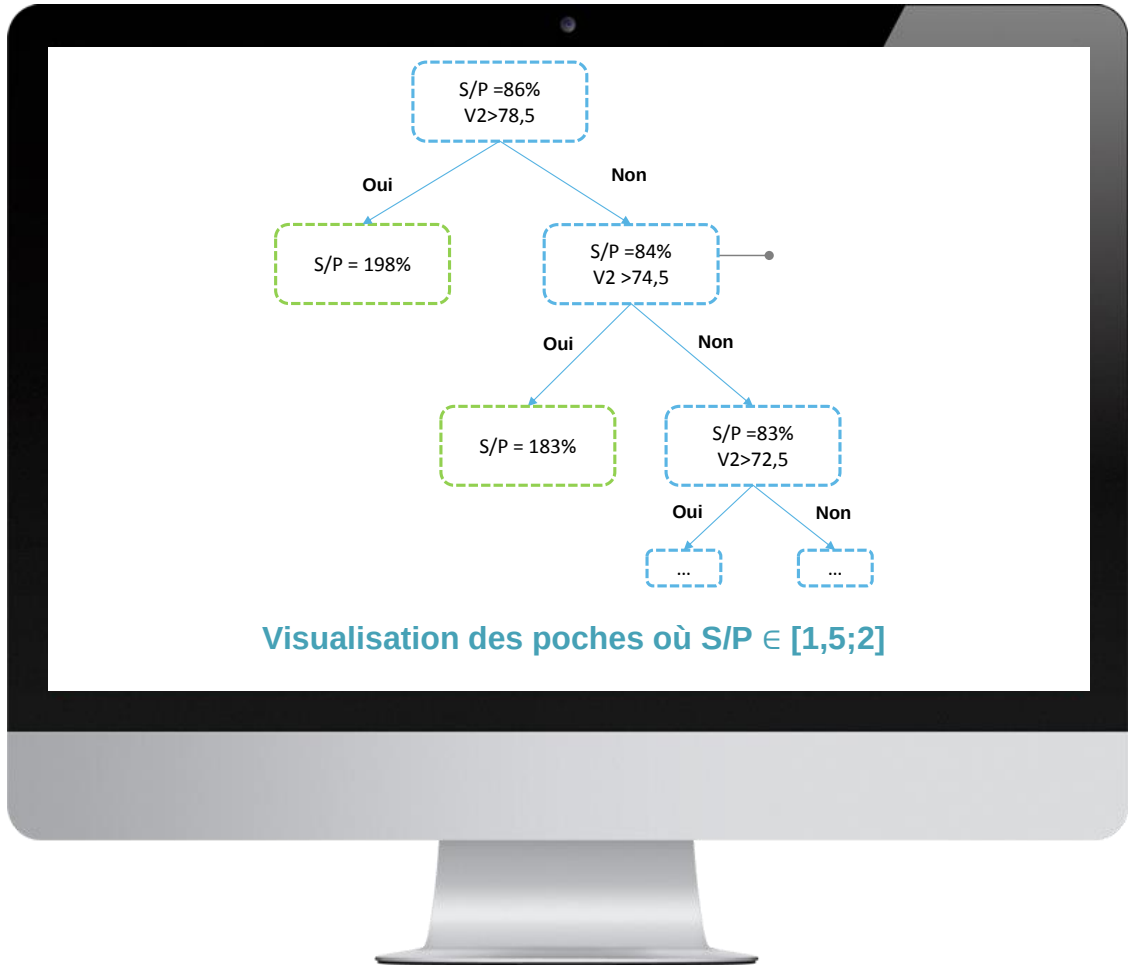
Data visualisation & suivi du risque

Visualisation de l'arbre et identification des **poches dont le S/P est compris entre 2 seuils à fixer**

Dynamique et en fonction de :

- La **granularité** souhaitée (garantie / Formule)
- De **données externes** sélectionnées
- Sur une **période** donnée

Permet de fournir **un reporting contrat par contrat dans les poches sous tarifées** paramétrées



Visualisation des poches où S/P ∈ [1,5;2]

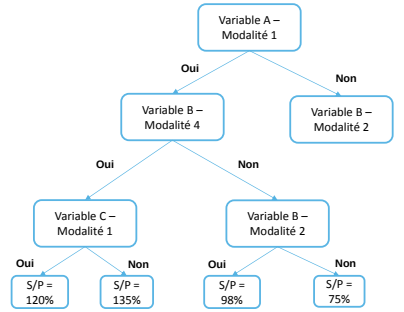
Quelles utilisations opérationnelles ?

Actions opérationnelles

Détermination des **variables les plus importantes dans la segmentation** du portefeuille et celles qui peuvent **expliquer les pertes**

- Revue des règles de souscription
- Revalorisation tarifaire
  - Règles d'action tarifaire si le S/P atteint un certain niveau
- Ajustements des algorithmes tarifaires
  - Intégration dans un GLM ou machine learning

Pilotage



Validation des intéractions observées

Ajustements

Actions tarifaires

Règles de souscription

Les ajustements ne doivent être réalisés qu'après la validation des intéractions observées sur une période donnée

Quelques bonnes pratiques *…pour intégrer au mieux les nouvelles données externes*

Fiabiliser l'existant

- **Inventorier les données disponibles** (internes/externes, structurées/non structurées) et traiter les aspects fonctionnels en étroite collaboration avec les métiers.
- **Fiabiliser les données** internes avant d'intégrer des données externes.

Investir dans des infrastructures et notamment des serveurs

- Prévoir un **stockage adapté** (pour des volumes de données en croissance exponentielle, les bases de données…).
- Anticiper les besoins en **performance** (pour le traitement en temps réel et l'accès aux données).

Tester sur un périmètre restreint

- Délimiter un périmètre (marketing, production…) pour réaliser rapidement un ou plusieurs **projets pilotes ou « proof of concept ».**

Echanger et challenger

- **Echanger avec les autres métiers** : gestion sinistre, la souscription

ANNEXE

Pour aller plus loin…

Collecte des données externes et analyse de données non structurées

Collecte de la données externes

Mise en place d'une collecte automatisée : Le Web Scraping

Le **Scraping** est un ensemble de techniques pour extraire le contenu d'un site Web

L'**objectif** est de **transformer les données récupérées** afin de les utiliser en base pour qu'elles soient **analysées.**

Collecte de la données externes

Les étapes pour la mise en place de web scraping

Définition de l'objectif

Réalisation des requêtes web scraping

Récupération des données et constitution d'une base

Stay compliant

Collecte de la données externes

Attention à la réglementation

Qu'est-il possible de faire aujourd'hui ?

- Scraper en vue de récupérer les données pour les « transformer »

Exemple

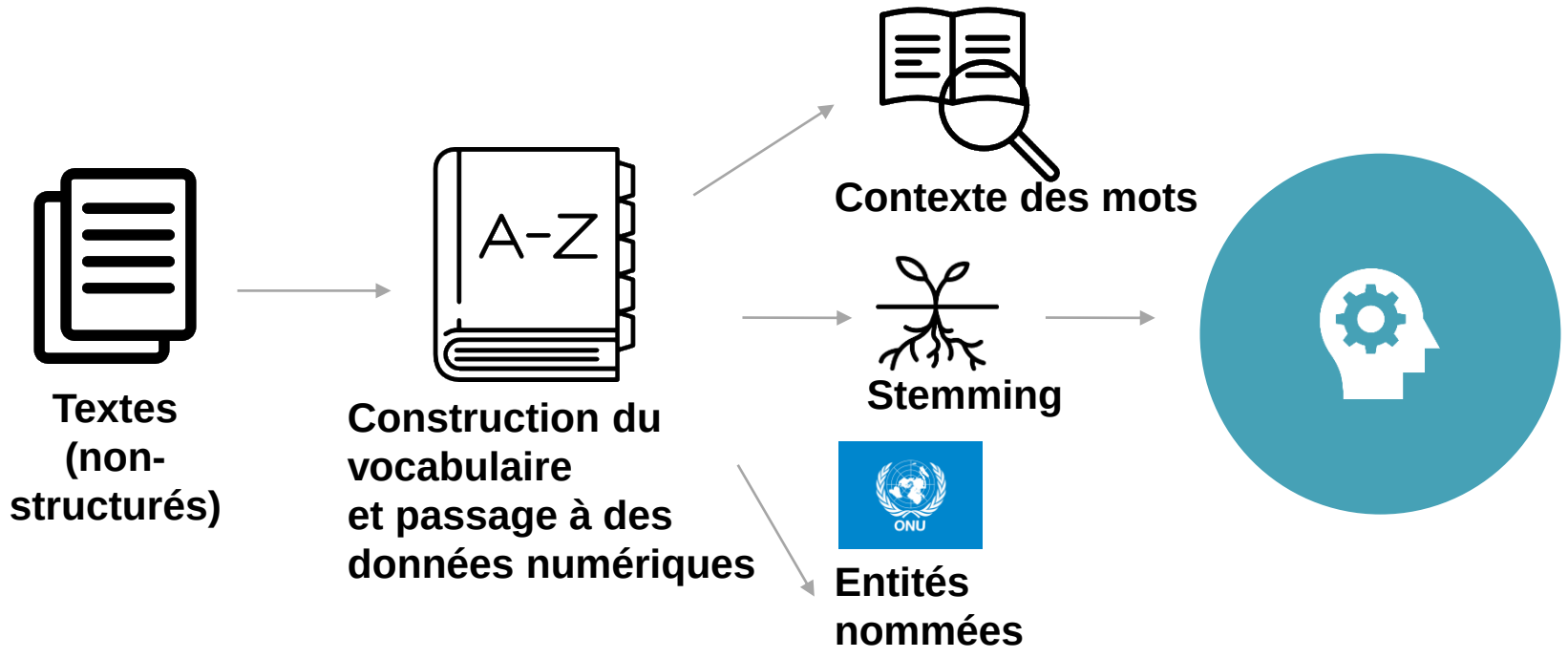
- Indice ou intervalle de score de marché pour un benchmark
- Transformation de données de plusieurs assureurs de place pour obtenir un zonier ou un algorithme de tarification marché

Les bonnes pratiques

- Lire les conditions générales d'utilisation
- Ne pas surcharger le site web scrapé
- Ne pas forcer l'accès à certaines données

Analyse de données externes non structurées

La technique du text mining



Préparation des données	Construction du vocabulaire	Lemmatization stemming	application du modèle

Analyse de données externes non structurées

Application dans l'analyse des actualités

News	Sentiment
La plus belle pluie d etoiles filantes de l annee aura lieu cette nuit	1,00
JeanMichel Blanquer a Millas on va etre tres attentifs a ce qui se passe pour les enfants ici	1,00
L echarpe plus indispensable qu il n y parait pour prevenir les crises d asthme	0,98
Serge Haroche Nobel de physique tres surpris et tres heureux	0,93
VIDEO SOCIETE Serge Haroche Nobel de physique tres surpris et tres heureux	0,93
…	
Une centaine d oeuvres du Caravage auraient ete retrouvees	0,00
L usine PSA d AulnaysousBois bloquee par les interimaires	0,00
Italie une centaine de dessins de jeunesse du Caravage trouves a Milan	0,00
Les pilotes de Canadair cessent leur greve	0,00
Exclu TV rencontre avec le pere d Amy Winehouse	0,00
…	
Prostitution le PS qualifie le texte adopte au Senat de regression scandaleuse	-0,80
Le Chilesaurus un dinosaure herbivore vraiment tres bizarre	-0,80
Philippot juge scandaleuse une fermeture de ligne de trains	-0,80
VIDEO SOCIETE Les soldes d hiver commencent tres mal	-1,00
Lyon discours tres agressif de Nicolas Sarkozy a l egard du PS	-1,00

Analyse de données externes non structurées

Application dans l'assurance

Surveillance

- ✓ Détection de fraude
- ✓ Détection des cyber attaques

Gestion / Relation client

- ✓ Gestion de recouvrement
- ✓ Gestion des sinistres
- ✓ Satisfaction client

Compliance

- ✓ Analyse de la jurisprudence

optimind.

manage risk, build your future

strategy. finance. risk. compliance. market.
human resources. digital transformation. data. bpo.