

DATA SCIENCE ET ANALYTICS EN SANTÉ

APPORT DES DONNÉES EXOGÈNES DANS LA COMPRÉHENSION DES RISQUES ASSURANTIELS ÉTUDE DE CAS

Data Science et Analytics en assurance

2

- La *data science* : un levier pour de nouvelles opportunités métiers
- De nouvelles expertises et de nouveaux outils
- Une démarche et une méthodologie spécifiques
- Construction du cas d'utilisation

Cas d'utilisation métier : évaluation de la consommation médicale à l'aide de données exogènes

5

- Identification, collecte et traitement de la donnée : une étape clé
- Modélisation : la puissance de la diversité
- Restitution des résultats : le retour au métier

La *data science* est une extraction de connaissances à partir de grands volumes de données structurées ou non structurées. Elle utilise de nombreuses techniques et outils empruntés à la physique, aux mathématiques, aux statistiques ou encore aux technologies de l'information : théorie du signal, algorithmique, datamining, modèles de probabilité, analyse prédictive, modèles de données.

Les assureurs sont par nature de grands collecteurs et analystes de données. Aujourd'hui, le secteur de l'assurance se transforme sous l'impulsion des nouvelles technologies. Pour tous les acteurs, l'enjeu est la valorisation du volume d'information à disposition. Ce terrain est très favorable pour utiliser les techniques et outils issus de la *data science*.

En effet, avec l'explosion des données disponibles et le développement de l'*Open Data*, de nouvelles perspectives s'ouvrent dans l'appréhension des risques. Mais la masse d'information - météo ou Insee par exemple - pouvant être croisée avec les données internes des organismes assureurs nécessite de dépasser les modèles traditionnellement utilisés en assurance.

Au travers d'un cas pratique sur la consommation médicale, Optimind Winter analyse pourquoi et comment les algorithmes de *machine learning* permettent d'améliorer la compréhension des risques assurantiels. Ce secteur à forts enjeux a besoin de se transformer, la connaissance et la mesure du risque en santé doivent être appréhendées.

Études de cas :
Modélisation de la
consommation santé
à l'aide de données
externes
*L'analyse
de notre expert*







Marc Dupuis, directeur métier en charge de l'offre Digital
Avec les contributions de Nicolas Le Berrigaud, actuaire, Practice Leader Santé
et Clarisse Guillou, actuaire.

Pour développer et pérenniser nos savoir-faire, favoriser le partage des connaissances
et la diffusion des meilleures pratiques professionnelles au sein de nos équipes,
Optimind Winter investit sur sa Practice.

La data science : un levier pour de nouvelles opportunités métiers

Exemples d'applications en assurance

DOMAINE	USAGE POTENTIEL	
 Marketing Distribution	<ul style="list-style-type: none"> • Améliorer les processus de souscription • Faciliter la veille concurrentielle et cibler ses opérations marketing 	AUTRES USAGES <ul style="list-style-type: none"> • Expliquer la sinistralité • Comprendre les comportements d'achat
 Produit	<ul style="list-style-type: none"> • Personnaliser les garanties et les tarifs • Pilotage de la rentabilité du produit et de son activité 	<ul style="list-style-type: none"> • Identifier les tendances de marché • Surveiller l'e-réputation de la société
 Gestion du risque	<ul style="list-style-type: none"> • Anticiper et mesurer les risques (lois comportementales) • Ajuster les modèles aux assurés en portefeuille 	<ul style="list-style-type: none"> • Obtenir des informations en quasi temps réel • Vision 360° du client
 Lutte anti-fraude	<ul style="list-style-type: none"> • Identifier les comportements anormaux en temps réel • Identifier les données volontairement biaisées 	

Source : Optimind Winter

- Assurance auto : passage du *Pay as you drive* au *Pay how you drive* via boîtiers embarqués/applications.
- Santé / prévoyance : appréhender le risque d'arrêt de travail des salariés.
- Lutte anti-fraude : croiser les déclarations récentes de sinistres avec une activité spécifique sur les réseaux sociaux et/ou des consultations sur des forums.

De nouvelles expertises et de nouveaux outils

Le rôle principal du *data scientist* est de créer de la valeur à partir des données. Curieux, communicatif, ayant du *leadership* et maîtrisant la gestion de projet, il doit aussi posséder un triptyque de compétences :

- Métiers : connaissance métier et sens *business* sont déterminants pour maîtriser les données, identifier les applications concrètes et imaginer de nouveaux usages.
- IT : une forte maîtrise de la programmation est indispensable.
- Statistiques : une connaissance des statistiques et des algorithmes de *machine learning* est nécessaire.

Algorithmes et logiciels utilisés

Outre le logiciel SAS, logiciel bien connu des actuaires et statisticiens, deux technologies sont aujourd'hui concurrentes... mais pourtant complémentaires.

Technologie Python

- Langage généraliste, universel et lisible avec un apprentissage simple.
- Un très grand nombre de modules disponibles dont Numpy (Vectorisation), Pandas (Gestion des *Dataframes*).

- *Scikit-Learn* librairie statistiques de référence optimisée qui dispose en natif des algorithmes de *machine learning*.

Technologie R

- Langage de référence en mathématiques et statistiques.
- Dispose en standard de la plupart des fonctions de *machine learning*.
- Plus approprié en phase exploratoire et en phase de restitution.



Attention l'utilisation de la licence R nécessite d'ouvrir son code en open source.

Exemples d'algorithmes :

- *Random Forest*
- *Gradient Boosting*
- *K-neighborhood*



Une démarche et une méthodologie spécifiques

Une démarche agile en 4 étapes : processus continu en mode projet Agile

1. Identification d'un cas d'utilisation

- Définition des enjeux et objectifs
- Formalisation analytique du besoin métier
- Établissement des indices clés de performance

2. Préparation des données

- Collecte des sources de données internes
- Identification des données externes nécessaires
- Préparation des données : mise en qualité
- Anonymisation

3. Application de modèles

- Enrichissement des données et génération de variables
- Sélection, implémentation des algorithmes - *machine learning*
- Combinaisons d'algorithmes

4. Généralisation

- Résultats métiers et data visualisation
- Définition et conception des applications ou services à destination des utilisateurs métier

Algorithmes et logiciels utilisés



La première étape de l'application des modèles est l'enrichissement des données internes via :

- la collecte de données externes (*Open Data*) ou *Text Mining*,
- la génération de données à partir des données collectées ; par exemple, la donnée « pluviométrie » permet de calculer la donnée « nombre de jours de pluie ».

Cette étape demande de la créativité, les modèles pouvant absorber un grand nombre de variables.

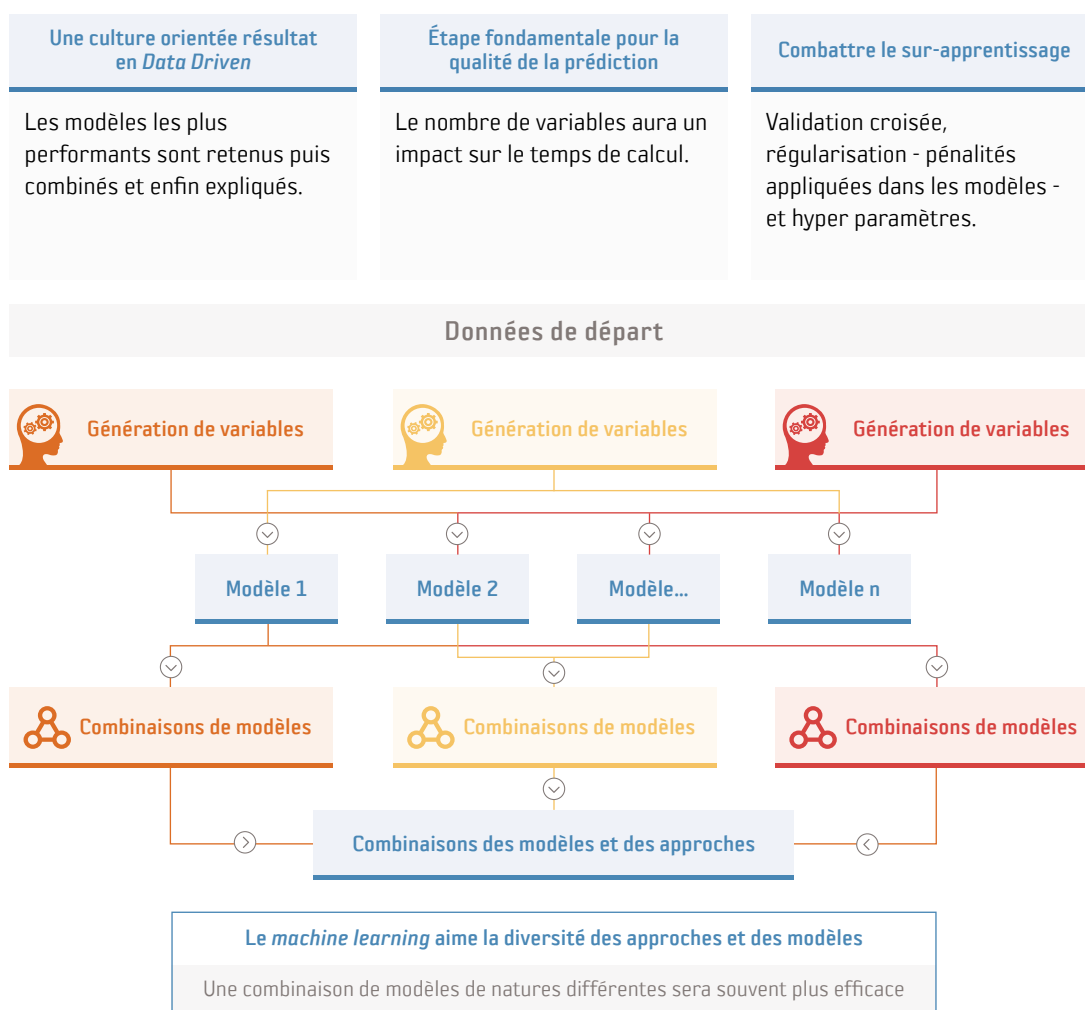
Pour certains modèles, il est également essentiel qu'il n'y ait pas de corrélations entre les variables.

La seconde étape consiste à optimiser les paramètres des modèles, à effectuer des validations croisées pour faire une estimation locale du modèle et se rendre compte de la nécessité de générer à nouveau des variables calculées.

Enfin, la dernière étape est la combinaison des différents modèles et des différentes approches, ce qui permet souvent d'obtenir de meilleurs résultats. Cette étape n'est pas forcément nécessaire, un modèle raffiné peut concurrencer les résultats d'une combinaison de modèle.



En cas de limite des capacités de calcul, une sélection des variables doit être effectuée avant d'appliquer les modèles.



Source : Optimind Winter

Il est recommandé de travailler à plusieurs sur cette étape d'application des modèles.

Dans un premier temps, à partir des données de départ, chaque intervenant travaille de manière indépendante et cloisonnée. Cela permet de laisser libre cours à la créativité de chacun pour générer des variables et déterminer les paramètres des modèles. En effet, le *machine learning* aime la

diversité : plus il y a de données et d'approches différentes, meilleures sont les estimations.

La phase d'amélioration des modèles peut commencer en cherchant à optimiser le modèle pour chaque type d'algorithme et en combinant éventuellement ces modèles. L'objectif est toujours d'obtenir le modèle optimal.

Construction du cas d'utilisation

Principes, objectifs et enjeux

Les organismes assureurs utilisent généralement un nombre restreint de variables explicatives : sexe, âge, zone géographique et régime social. Avec internet et

toutes les données qu'il véhicule en *Open Data* telles que les données environnementales, de nouvelles possibilités s'ouvrent aux assureurs dans le domaine de la compréhension de la consommation médicale.

Cas d'utilisation métier : évaluation de la consommation médicale à l'aide de données exogènes

2

Identification, collecte et traitement de la donnée : une étape clé

Recensement des données

Les sources externes d'informations sont nombreuses et de qualité différente.

- Les données étatiques : données neutres, structurées et soumises à des protocoles de collecte rigoureux.
- Les données d'intermédiaires et de fournisseurs de données.
- Les données collectées individuellement.

Les données disponibles à ce jour pour les assureurs et utilisées dans cette étude, sont les données classiques relatives aux bénéficiaires et aux prestations : âge, sexe, niveau de garantie souscrit, date de soins, acte concerné (pour cette étude un filtre a été effectué sur les consultations généralistes), montant de la prestation, ainsi que le code postal de la commune de résidence de l'assuré.

Les données externes utilisées sont des données environnementales : données INSEE et données météorologiques et climatologiques extraites du site Agri4Cast - données météo mises en libre-service par la Commission européenne.

Le lien entre les données internes et les données externes est réalisé grâce au code postal.

Traitement et qualité des données

Même structurées et fournies par un tiers de confiance, les données de base doivent faire l'objet de retraitements et de contrôles avant leur utilisation dans les modèles statistiques.

- Détection des anomalies
 - Tests de cohérence, requêtes, *data visualisation*
- Traitement
 - Suppression des données, imputation, gestion des valeurs extrêmes, discrétisation des variables continues
- Analyse des données
 - *Data visualisation*, données statistiques

La qualité peut être évaluée de plusieurs manières :

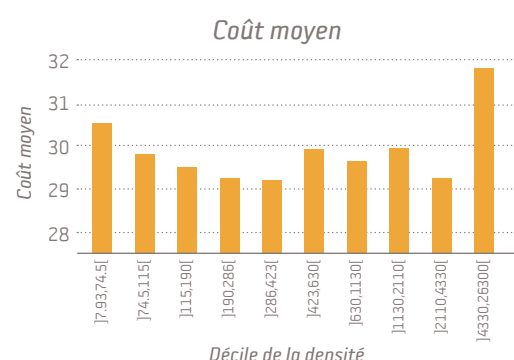
- Requêtes réalisées sur la base de données afin de ressortir les valeurs manquantes
- Contrôles de cohérence/statistiques
- *Data visualisation*

“

Le traitement des anomalies est primordial. La méthodologie choisie pouvant influencer fortement les résultats, notamment en cas de données restreintes.

Visualisation de données : une étape importante dans l'analyse

Exemple de *data visualisation* de la densité de population - *package leaflet* - et des coûts moyens en fonction de la densité.



* Uniquement réalisé sur les actes dépassant le tarif de convention

Source : Optimind Winter

Focus sur le traitement des données météo

Les données Agri4Cast sont des données journalières interpolées au travers d'un maillage de la France par des carrés de 25 km par 25 km.

Chaque coin des carrés représente une mesure de données météorologiques. Leur positionnement dans

l'espace est exprimé en Lambert. Afin de rattacher la météorologie aux différents codes postaux, il est nécessaire :

- d'avoir le positionnement dans l'espace de chaque commune de France,
- de calculer la distance entre les communes et le point le plus proche.

Modélisation : la puissance de la diversité

Étape initiale

La première étape est la création d'une base d'apprentissage et de test. L'objectif est de **donner la même base** de travail et d'évaluation à tous les intervenants afin de pouvoir **comparer les approches**.



Base d'apprentissage : 70 % des données

Subdivisée en 10 sous-ensembles afin de faire une validation croisée des modèles utilisés avant d'évaluer l'erreur sur la base de test.

Base de test : 30 % des données

Premiers résultats

Une pré-étape à l'étude : une première phase de *micro-testing* dont l'objectif a été de valider l'intérêt du *use case* sur quelques données météorologiques et sociologiques :

- réalisation d'un GLM,
- réalisation d'un arbre de décision simple.

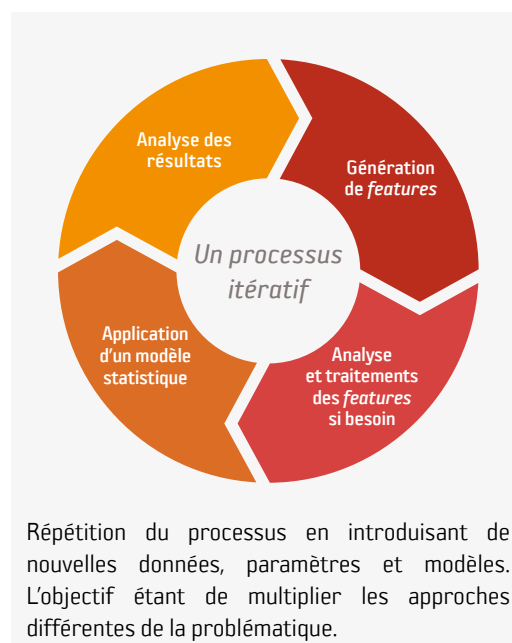
Résultats sur la fréquence :

- Certaines données intégrées ressortent comme significatives pour expliquer la fréquence de sinistres :
 - Météorologiques : nombre de jours de gel, nombre de jours de pluie et pluviométrie.
 - Sociologiques : part de bacheliers, taux d'activité et part de logements vacants.

Un premier constat : Arbre et GLM ne donnent pas nécessairement les mêmes variables explicatives.

“

Notre conseil : travailler au démarrage individuellement pour laisser libre cours à la créativité de chaque analyste



Répétition du processus en introduisant de nouvelles données, paramètres et modèles. L'objectif étant de multiplier les approches différentes de la problématique.

Source : Optimind Winter



Les modèles utilisés

GLM

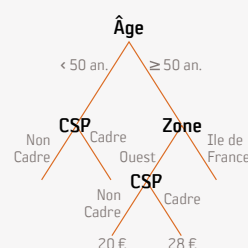
Modélisation paramétrique
à partir de variables explicatives

$$g(E[y]) = \sum_{i=1}^p \beta_i x_i$$

- La loi de la variable réponse y : log-normale, poisson, gamma
- La fonction lien g : fonction identité, fonction logarithme
- Une estimation des paramètres par maximum de vraisemblance

Arbre de décision

Représentation graphique d'un modèle prédictif



Division de l'arbre du
sommet vers les feuilles
en sélectionnant à chaque
étape une variable d'entrée
qui permet la meilleure
répartition de l'ensemble
d'objets.

- Feuilles : valeurs possibles de la variable cible
- Embranchements : combinaisons de variables d'entrée qui conduisent à ces valeurs

Random Forest

Création d'un ensemble d'arbres à agréger

- Construction de plusieurs *Bootstrap* chacun à la base d'un arbre



- Aléa supplémentaire : tirage aléatoire des prédicteurs
- Agrégation des arbres indépendants

Gradient boosting

Création d'un ensemble d'arbres à agréger



- Algorithme itératif
- Chaque arbre est une version adapté du précédent
- Application d'un poids plus fort aux segments les moins bien ajustés par le modèle sur l'arbre précédent

L'exposition aux risques

L'exposition aux risques est une donnée particulière à prendre en compte dans la modélisation de la fréquence. La fréquence des sinistres correspondant au rapport entre le nombre de sinistres et la durée d'observation des individus. Tous les modèles présents dans R ne permettent pas nativement de prendre en compte l'exposition aux risques.

- GLM
- Arbre de décision : méthode Poisson
- Random Forest : méthode Poisson
- GBM : distribution Poisson

Pour les modèles non adaptés, plusieurs possibilités :

- Travailler uniquement sur les individus ayant été présents durant toute la période d'étude : non optimal.
- Segmenter par durée d'exposition : risque de surestimation.
- Modifier les algorithmes afin de prendre en compte l'exposition aux risques.

Comparaison de modèles

Les résultats des modèles sur notre exemple démontrent que le modèle *Gradient Boosting* est généralement l'algorithme le plus performant. Ici c'est celui qui permet d'obtenir l'erreur la plus faible. Dans notre étude, le *Random Forest* est moins bon que le *Gradient Boosting* mais est meilleur que le GLM et l'arbre CART. L'algorithme est néanmoins souvent utilisé dans le cadre de travaux sur d'importants volumes de données car il est parallélisable, à l'inverse du *Gradient Boosting*.

Par construction, l'arbre CART est moins performant que le *Random Forest* et le *Gradient Boosting*.



Restitution des résultats : le retour au métier

Aujourd'hui, il est très difficile d'exploiter directement ce type de modèle pour faire de la tarification en raison de la capacité informatique et de la modélisation du remboursement du régime obligatoire.

Il est possible de créer un zonier fin en vue de son intégration dans un modèle plus classique de type GLM.

- Analyse des résidus du modèle créé à partir des variables les plus explicatives.
- Création de classes de scores sur ces résidus à l'aide de modèles d'apprentissage.
- Réintégration dans le modèle initial en tant que variables explicatives.

Le travail sur les données est très consommateur de temps : de 80% à 90% du temps passé. Il est nécessaire de travailler à 2 ou 3 personnes afin d'avoir des approches différentes du sujet. Idéalement les personnes doivent avoir des profils différents, 2 actuaire-experts métiers et 1 statisticien connaissant bien les données spatiales et météorologiques utilisées ou les autres domaines requis.

En phase de Recherche & Développement, il est préférable d'avoir à disposition des logiciels ouverts type R ou Python, ce qui permet ainsi d'avoir accès à un nombre important de fonctionnalités très rapidement.

Il est également recommandé d'investir dans des infrastructures et notamment des serveurs puissants. Les bases manipulées et les travaux réalisés avec les algorithmes de *machine learning* peuvent être très rapidement gourmands en ressources, notamment au niveau de la mémoire vive.

Pour utiliser ces nouvelles méthodes, outils et algorithmes, il est nécessaire de bousculer les habitudes et de faire preuve de créativité tout en privilégiant la rigueur dans le déroulé des travaux sur les données.



La mise en œuvre des projets en *data science* demande des méthodologies, des outils spécifiques, des expertises variées dont certaines peu ou pas présentes dans les structures d'assurance, une organisation et des projets collaboratifs mêlant ces différentes expertises. Il est important de savoir poser un regard différent sur les objectifs ou les problèmes à résoudre.

Les outils et techniques de la *data science* proposent de mettre en œuvre des méthodes d'analyses complémentaires à celles utilisées par les acteurs de l'assurance. Ces nouvelles approches sont d'autant plus intéressantes, qu'elle proposent d'utiliser des données jusqu'alors peu ou pas utilisées par ce secteur, pourtant grand consommateur d'informations sur la matière assurable.

Le périmètre et les objectifs de l'application sur le cas d'utilisation Santé ont été volontairement restreints, des extensions de périmètre sont nécessaires pour obtenir d'autres résultats. La méthodologie et les outils sont prêts pour adresser d'autres portefeuilles et d'autres données externes. L'objectif ne doit pas être centré uniquement sur les aspects tarifaires, en effet une meilleure compréhension du risque est la base de la prévention.



Leader de l'actuariat conseil et de la gestion des risques en France, Optimind Winter constitue l'interlocuteur de référence pour les organismes assureurs, banques et grandes entreprises qui souhaitent un partenaire métier de haut niveau les accompagnant dans leurs projets stratégiques.

Expertise, méthode, intégrité, engagement, pragmatisme, innovation, anticipation et disponibilité sont les valeurs clefs qui animent nos 180 collaborateurs, consultants experts pour la plupart, dont plus de 70 actuaires diplômés membres de l'Institut des Actuaire. Nos clients bénéficient ainsi des plus hautes expertises en gestion du risque associées à la qualité d'une signature de référence d'un des leaders européens en gestion des risques. Notre indépendance, garantie par un capital détenu uniquement par nos salariés et dirigeants, offre à nos clients la perspective d'une collaboration pérenne et engagée.

Optimind Winter vous apporte son expertise sur les métiers suivants :



Actuariat Conseil



Protection Sociale



Risk Management



Finance & Performance



Business Transformation

optimind winter. 
LOCAL OPTIMIZATION EUROPEAN MINDED

Pour plus d'informations, rendez-vous sur notre site www.optimindwinter.com

Vos Contacts /

Éric Gaubert / directeur du développement / eric.gaubert@optimindwinter.com

Marine de Pallières / responsable de la communication / marine.depallieres@optimindwinter.com

T / +33 1 48 01 91 66

